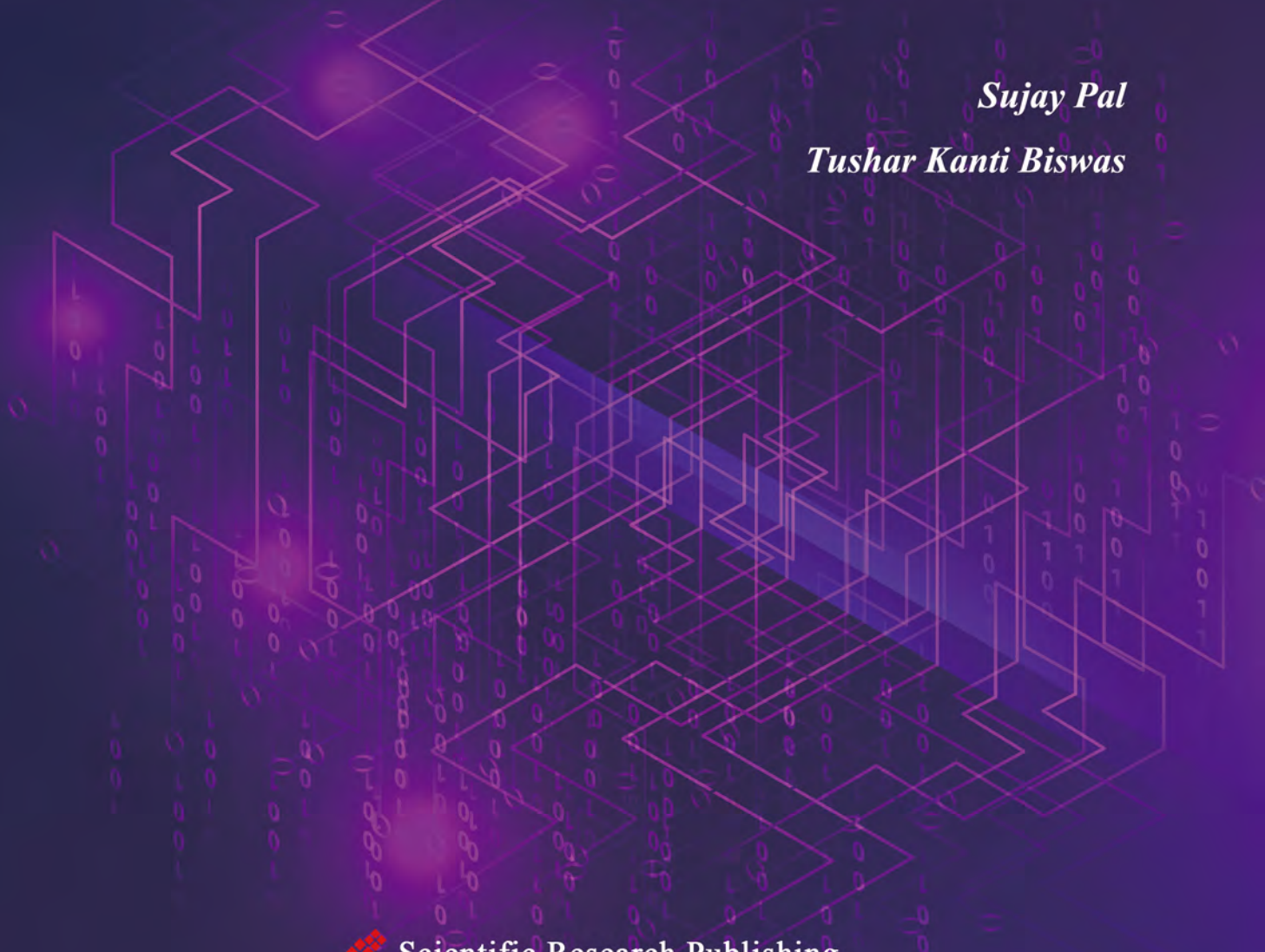
Advances in Modern and Applied Sciences

A Collection of Research Reviews on Contemporary Research (Volume 1)

Sujay Pal

Tushar Kanti Biswas



Scientific Research Publishing

Advances in Modern and Applied Sciences

A Collection of Research Reviews on
Contemporary Research
(Volume 1)

Sujay Pal
Tushar Kanti Biswas

Scientific Research Publishing

Advances in Modern and Applied Sciences
A Collection of Research Reviews on Contemporary Research (Volume 1)

Published by
Scientific Research Publishing, Inc.
ISBN: 978-1-64997-437-2
http://www.scirp.org

Advances in Modern and Applied Sciences

A Collection of Research Reviews on
Contemporary Research
(Volume 1)

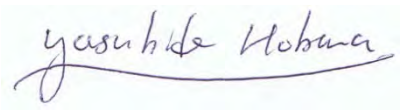
Editors
Sujay Pal
Tushar Kanti Biswas

Srikrishna College (Govt. Sponsored)
Affiliated to University of Kalyani
Bagula, Nadia, West Bengal, India 741502

Foreword

It is a great pleasure to write this foreword for the book "Advances in Modern and Applied Sciences, A Collection of Research Reviews on Contemporary Research (Volume 1)". When I heard about publishing a comprehensive book on science from one of the Editors, I was very much excited to see it. At last, the dream came true for Science Departments of Srikrishna College. Indeed, this book contains a wide variety of research topics in modern science presented in a systematic and engrossing way to acquire knowledge without knowing its cutting-edge technologies. The readers can experience the latest developments in the field of computer and material sciences. The field of atmospheric and space Sciences is of growing importance for our future life in view of Sustainable Developments Goals (SDGs). Then the articles regarding Astrophysics, Astronomy, and High Energy Physics attract our curiosity about how our universe works. All the above-mentioned topics are carefully and well documented in four chapters by many experts from their respective fields. This book must be suitable not only for scholars but also for students and researchers working in different research fields to widen their view of science. I am eagerly waiting for the next volume of this book.

With best wishes.



Yasuhide Hobara

Professor

Head, Center for Space Science and Radio Engineering

Graduate School of Informatics and Engineering

The University of Electro-Communications, Tokyo, Japan

Preface

This book *Advances in Modern and Applied Science* materializes our long-cherished dream of publishing a series of volumes consisting of review papers on contemporary research fields from a broad spectrum of basic sciences. The present volume, which is our first baby-step towards that fulfilment, includes a collection of twenty-five review articles contributed by about fifty researchers and scientists whose vocations are in diverse fields of science including astrophysics, astronomy, high energy physics, space science, atmospheric sciences, computer sciences to material sciences.

The main objective of this book is to provide an insight into the advances that modern day science has made and bring forth a better understanding of this vast and exhilarating discipline called *science*. Keeping this in mind the contributors have emphasized on the essemplastic power by incorporating both the quantitative and qualitative research outcomes in a very lucid manner that would dulcify the readers with its ineluctable pedagogy as put forward by esteemed personalities mostly through the review articles. We are certain that new graduates, Ph.D. scholars, teachers, and researchers from diverse fields will benefit from this volume, which can be considered as a stock-taking of the new developments in recent day science.

The editors have compiled and edited the articles duly to suit the purpose of the book and at the same time to keep a balance between diverse topics. We have organized our book into four specific chapters. To begin with, *Chapter 1* consists of nine articles from Astrophysics, Astronomy and High Energy Physics. In this section, the reader is provided with a brief overview on various topics focusing on Monte Carlo simulation of black hole imaging in X-ray domain, multi-messenger astronomy, properties of radio galaxies, galaxy rotation curves, neutron stars, radio study of the atmosphere in Saturn's moon and new quantum methods in Krein space. *Chapter 2* is devoted to Atmospheric and Space Sciences comprising eight articles. Articles on recent topics like extraordinary air pollution in New Delhi, atmospheric factors affecting transmission of Covid-19, impact of extreme weather events on agriculture, effects of ionospheric forcing from above and below due to solar flares, seismic events, cyclonic storms, polar stratospheric warming events and their experimental measurement techniques have enriched this Chapter. In *Chapter 3*, researchers have explored advances in modern Computer Science and Mathematics through five articles focusing on recent topics like routing and spectrum allocation scheme in elastic optical networks, Spectrally-spatially elastic optical networks technologies and their challenges considering different types of fibers and Verifiable Visual Cryptography for transmitting secret image over the internet and Alexander-Spanier cohomology theory. Finally, *Chapter 4* contains three articles from Material Sciences giving an overview of the current state of the art focusing on topics like data-based material designing using Machine Learning, thermo-electric devices, and applications of Chitin and Chitosan based composite as promising green energy resources.

The book has been sponsored by Srikrishna College, Bagula, West Bengal, India. In this regard, we take the opportunity to thank Dr. Sukdeb Ghosh, Principal of the College, and Mr. Anup Kumar Bhadra, President of Governing Body for their constant support and encouragement without which it would have been difficult for the book to see the light of day. The book can truly be said to act as a springboard in our endeavours to promote and further advance the culture of research environment in our institution..

We earnestly thank all the authors for their efforts and enthusiasm to submit their contributions to this volume and make this book a successful publication. Our sincere thanks reaches out to all the faculties and staff members of Srikrishna College, especially Dr. Bipul Mandal, Prof. Somnath Chakraborty, Prof. Anamika Chakraborty, Ujjal Kumar Das, Dr. Ankita Indra, Dr. Tushar Kanti Bose, Dr. Pranab Das, Supratick Adhikary, Dr. Paramita Hajra, Dr. Nabadyuti Barman whose constant support and succour made the book a success. Finally, we must thank the Publisher Scientific Research Publishing Inc., USA for agreeing to publish the book in a timely manner.

May, 2022

Sujay Pal
Tushar Kanti Biswas
Srikrishna College (Govt. Sponsored)
Affiliated to University of Kalyani
Bagula, Nadia, West Bengal, India 741502

Contents

Chapter 1: Astrophysics and Astronomy……1

1. Arka Chatterjee, Implications of Monte Carlo Simulations on the X-Ray Images and Temporal Properties of Stellar Mass Black Holes……2
2. Debabrata Adak, On the Galaxy Rotation Curves……14
3. Sourav Palit, Multi-Messenger Astronomy in the GW Era……...31
4. Netai Bhukta, Sabyasachi Pal, Sushanta K. Mondal, Properties of Giant Radio Galaxies……...38
5. Dusmanta Patra, 'Winged' Radio Sources: A Peculiar Subclass of Radio Galaxy……..49
6. Shobha Kumari, Sabyasachi Pal, Netai Bhukta, Sushanta K. Mondal, Winged Radio Galaxies: An Overview…….59
7. Manoj Mandal, Sabyasachi Pal, Cyclotron Resonant Scattering Features in Highly Magnetized Neutron Stars……76
8. Arijit Manna and Sabyasachi Pal, Millimeter Wavelength Studies of Complex Nitrile Species in the Atmosphere of Saturn's Moon Titan……90
9. Arindam Chakraborty, A Short Story of The New Quantum Methods in Krein Space: PT-Symmetry and Non-Hermiticity……..102

Chapter 2: Atmospheric and Space Sciences……108

10. T. Mukherjee, V. Vinoj, Post Monsoon Air Pollution Episodes over Megacity New Delhi……109
11. Souvik Manik, Sabyasachi Pal, Manoj Mandal, Impact of Environmental Factors on COVID-19 Transmission: An Overview……116
12. Javed Akhter, Manish Kumar Naskar and Shakil Hassan, Subrata Kumar Midya, Extreme Weather Events and Its Impact in Agriculture……..129
13. Sayak Chakraborty and Tamal Basak, Brief Review on the Lower Ionosphere and the Effects of Solar Flare Thereon……137
14. Bakul Das, P. K. Haldar, A Review of ELF/VLF Reception Techniques & Experiments……...…146
15. Kheyali Barman, B. Das, P. K. Haldar, S. Pal, Ionospheric Effects of Cyclonic Storms: A Brief Review……164
16. Suman Ray, Possible Precursory Effects of Seismic Events in VLF Radio Signals……171
17. Arnab Sen, S. K. Mondal, S. Pal, Effects of Sudden Stratospheric Warming (SSW) on the Upper Atmosphere……179

Chapter 3: Computer Science and Mathematics……..187

18. Imran Ahmed, Eiji Oki, and Bijoy Chand Chatterjee, Crosstalk-Avoided Resource Allocation in Spectrally-Spatially Elastic Optical Networks: An Overview……188
19. Abdul Wadud, Eiji Oki, Bijoy Chand Chatterjee, Performance of Non-Defragmentation and Batch Processing Based Proactive Fragmentation Management Scheme in Elastic Optical Networks……196
20. Tuhin Kumar Biswas, Kinsuk Giri, Applications of Computational Geometry in Clustering: A Review……...…202
21. Ujjal Kumar Das, Verifiable Visual Cryptography for Gray Scale Images with Meaningful Shares……..208
22. Tushar Kanti Biswas. Alexander-Spanier Cohomology Theory on Topological Spaces: A Review……...217

Chapter 4: Material Science……..……222

23. Sourav Ghosh, G. Ranga Rao, Tiju Thomas, Machine Learning in Electrochemical Energy Storage: Practice and Discussion……223
24. S. Sau, A. Jana, S. Mahakal, Diptasikha Das and K. Malik, Thermoelectric Device: Recent Status and Applications in Biomedical Instrument……232
25. Chinta Haran Majumder, Arpan Kool, Somtirtha Kool Banerjee, Krishanu Chatterjee, Advanced Electronic and Energy Applications of Chitin and Chitosan based Composite……...…246

Astrophysics & Astronomy

Implications of Monte Carlo Simulations on the X-Ray Images and Temporal Properties of Stellar Mass Black Holes

Arka Chatterjee[1]

[1]Department of Physics and Astronomy, University of Manitoba, Winnipeg, Manitoba, R3T 2N2, Canada

*Corresponding author: arka.chatterjee@umanitoba.ca, arka019icsp@gmail.com

Abstract

Theoretical X-Ray images of a black hole with an accretion disk have been studied for quite a long time. However, the spectral and temporal counterparts of such images are barely explored for Galactic black hole candidates. Spectra of such black hole candidates are primarily divided into two components: a disk radiation dominated black body component and a Comptonization dominated power-law component. We explore a disk geometry combined with a Keplerian disk which acts as seed photon source and a relativistic thick disks acting as a Compton cloud in our present study. Monte-Carlo technique has been employed for image generation. The spectra corresponding to the images are presented. Using Ray-Tracing process, we calculate the time lags of hard and soft X-Rays. The variations of lag properties with the size of Compton cloud and inclination angle are presented. Later, we investigate the implications of outflows on the timing properties of Galactic black holes using multi wavelength radio X-ray data and corroborate with our state-of-the art hydrodynamic simulations.

Keywords:*Accretion Disks, Black Hole Physics, Radiative Transfer, Ray-Tracing*

1 Introduction

Simulation of images of accretion disk started from 1979 with the pioneering work by J. P. Luminet [1]. Following his work, Fukue & Yokohama [2] first presented colored images of an accretion disk around a non-rotating black holes. Viergutz [3] extended the work including the spin in Kerr geometry. Marck [4], first employed Ray-Tracing process to get the images of accretion disk around black holes. In recent years, after commissioning of Event Horizon Telescope (EHT), this field of the study has received new boosts [5]. Bromley & Melia [6] studied polarization properties of accretion disk. The 2019 EHT observations [7, 8, 9, 10, 11, 12] of the super massive black hole at the centre of M87 galaxy put forward a new era of black hole observations in radio frequency. However, the images in X-ray domain are less explored due to interferometric challenges in shorter wavelength. For Galactic black hole candidates, Comptonization is a crucial physical process [13]. In X-ray regime, Chatterjee, Chakrabarti & Ghosh [14], simulated the theoretical images in presence of Comptonization confined within the thick disks. Three dimensional Ray-Tracing equations allow to generate images and spectra of any Compton cloud obtained from hydrodynamical solutions. Using the similar method, the time of arrival of photons on the plane of the observer are also calculated [15] which provide higher accuracy for inclination angle based studies of accretion physics.

X-ray time lags [16] provide deeper understanding of the disk geometry and physical processes in short timescales. Hard lags are found when harder photons delay over the softer energy band and are commonly explained by Comptonization model [16]. Soft/negative lag is produced when the hard photons reach the observer prior to the soft photons and can be interpreted by propagatory oscillation [17], [18] or hard X-ray reflection from the disk [19]. Often, Quasi-Periodic Oscillations are found in the light curves of Compact objects having broad range of mass [20], [21], [22]. The time lags around the QPOs provides the information of the length scale of various X-ray emitting regions on the accretion disk [23], [24]. They implicated that the gravitational bending could aid to the soft-lags as such cases are observed only for high inclination angle sources. For type-B QPOs, they suggested that the base of the jet acts as the oscillating region and are mostly observed when the source exhibits higher radio activity. Along with spectral properties, such as radio and X-ray flux correlations ($F^R \propto F^X$; see [25], [26], the disk-jet connections are also found from the optical-X-ray time lags of GX 339-4 [27]. Recently, Chatterjee et al. 2019 [28] discovered correlations between radio flux and soft-lags for XTE J1550-564 during its 1998 outburst. Later, Patra et al. 2019 [29] found similar correlations for three other high inclination GBHs. On the other hand, in case of low inclination sources, such as XTE J1650-500, the jet enhances the Comptonization which increases the time lag (see [30]). Thus, to gain deeper understanding of time lags through simulations, one has to encounter Comptonization, reflection, gravitational bending in the curved space-time, and disk-jet connections. To consider the inhomogeneity of the Compton cloud requires robust simulation approach utilizing Monte-Carlo technique (see [31] for details). Earlier theoretical works in the field includes (see [32], [33]) the jet model along with the Comptonization in the simulations to understand the disk-jet connections through timing properties of GBHs.

In this present context, we opt for Two Component Advective Flow (TCAF) model where advection dominated outflows can be produced self-consistently [34]. The model contains a low angular momentum and low viscous sub-Keplerian

component and a higher angular momentum and higher viscous Keplerian component. Both, these component bloats after reaching centrifugal barrier and creates a post-shock region which acts as Compton cloud. The region inside this barrier is known as CENtrifugal pressure supported BOundary Layer or CENBOL and acts as the Compton cloud. Molteni et al. (1994) [35] showed that post shock region behaves like thick disks that were earlier studied by [36], [37], [38], [39, 40]. Thus, the thick disks are safe assumptions in the context of modeling Compton clouds in TCAF paradigm. Later, [34] produced the spectral variations of accretion disks around black holes due to variation of physical parameters like disk rate, halo rate, shock location and shock strength. In this framework, it is much needed to investigate the images of TCAF where connection between source spectra and observer spectra can be established by considering curved geometry.

The structure of this article is the following. We briefly demonstrate the thick disk geometry (§2), Radiation profile of Keplerian disk (§3), Ray-Tracing process (§4) and Monte-Carlo Comptonization technique (§5). Then in Results (§6), we first show the source spectra after Comptonization. In §6.2, we show image of Keplerian disk, image of an accretion disk with thick disk acting as Compton cloud. We show the image of TCAF in presence outflows where post-shock region is obtained from hydrodynamic TVD code. In §6.2.4, we show the observed spectra and spectral hardening due to inclination angle. §6.3 deals with simulations of time lag properties and the implications of outflows on the lag-energy spectrum. Later in §7, we draw our conclusions.

2 Compton Cloud or Thick Disks

In case of barotropic ($p(\epsilon)$) process, Euler's equation for perfect fluid becomes,

$$W - W_{in} = \int_0^p \frac{dp}{p+\epsilon} = \int_{u_{tin}}^{u_t} \ln(u_t) - \int_{l_{in}}^{l} \frac{\Omega \nabla l}{(1-\Omega l)}, \tag{1}$$

where, von-Zeipel relation [41] is taken as $\Omega = \Omega(l) = c^{2/n} l^{1-2/n}$ with $\lambda = \frac{r\sin\theta}{(1-\frac{2}{r})^{1/2}}$ (C85 [39]). We have considered four velocity components as $[u_t, 0, 0, u_\phi]$.

From the intersection of Keplerian angular momentum and angular momentum of the thick disk at inner boundary (r_{in}) and the centre of disk (r_c), we can easily calculate c and n using algebraic equations. We considered length in the unit of $r_g = GM/c^2$, where G, M, c represents gravitational constant, mass of the black hole, and speed of the light respectively.

We take a polytropic equation of state, $p = K\rho^\gamma$ with p is pressure, ρ is matter density and K is measure of the entropy. γ is the polytropic index having value as $4/3$.

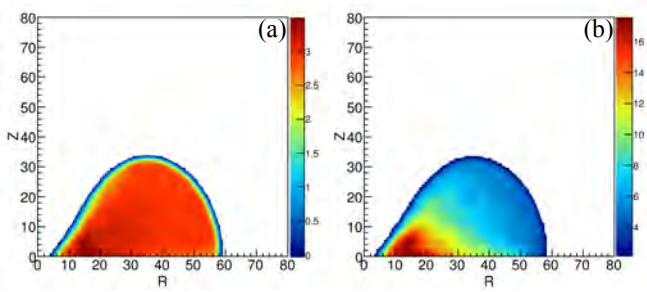


Figure 1: (a) Energy (in keV plotted in log scale) of escaping Photons, (b) number of scattering (dimensionless number) suffered by emergent photons from the Compton cloud are represented for a typical Compton cloud having the outer edge at around 60 r_g.

Typical meridional cross-section of the Compton cloud with outer boundary at 60 r_g is presented in Fig. 1 where the central temperature was around 200 keV. The energy and number of scattering distribution emergent photons suggests the correlation between higher electron density and temperature regions produces harder photons. Later, for dynamical studies, we employ the TVD (Ryu et al. 1997) [42] method where inflowing sub-Keplerian matter suffers through the shock before entering the black hole. The ubiquitous outflows are generated at the shock front.

3 Keplerian Disk

While simulating the geometry of the accretion disk, the Keplerian disk is placed around the Compton cloud and has much higher viscosity (see, Chakrabarti 1996 [43]; Chakrabarti & Das 2004 [44] and Nagarkoti & Chakrabarti 2016 [45]) with lower temperature than that of the Compton cloud region.

We consider truncated Keplerian disk with radiation profile adopted from Page & Thorne, 1974 [46]. Flux and temperature are given by,

$$F(r) = \frac{F_c(\dot{m}_d)}{(r-3)r^{5/2}} \times \left[\sqrt{r} - \sqrt{6} + \frac{\sqrt{3}}{3} log\left(\frac{(\sqrt{r}+\sqrt{3})(\sqrt{6}-\sqrt{3})}{(\sqrt{r}-\sqrt{3})(\sqrt{6}+\sqrt{3})}\right)\right] \quad and \quad T(r) = \left(\frac{F(r)}{\sigma}\right)^{1/4} \tag{2}$$

where, $F_c(\dot{m}_d) = \frac{3m\dot{m}_d}{8\pi r_g^3}$, with $\dot{m}_d$ being the disk accretion rate in Eddington unit, $\sigma = \frac{2\pi^5 k^4}{15h^3c^3}$ is the Stefan-Boltzmann constant.

Photon flux emitted from the Keplerian disk surface is considered same as reported in Garain, Ghosh & Chakrabarti, 2014 [47] and [14, 15].

4 Ray-Tracing Process

The general equation motion of particles is given by,

$$\frac{d^2x^\mu}{dp^2} + \Gamma^\mu_{\nu\lambda}\frac{dx^\nu}{dp}\frac{dx^\lambda}{dp} = 0, \tag{3}$$

with $\mu = [0, 1, 2, 3]$; $x^0 = t$, $x^1 = r$, $x^2 = \theta$ and $x^3 = \phi$, where p is an Affine parameter.

Introducing energy $P_t = E = (1 - \frac{2}{r})\frac{dt}{dp}$ and angular momentum $P_\phi = L = r^2 \sin^2\theta \frac{d\phi}{dp}$ (Chandrasekhar, 1983) [48], one can drop fourth equation of motion keeping all the generality of the trajectory of photons. So, we can write

$$\begin{aligned} \frac{d^2r}{dt^2} + \frac{3}{r(r-2)}\left(\frac{dr}{dt}\right)^2 - (r-2)\left(\frac{d\theta}{dt}\right)^2 \\ -(r-2)r\sin^2\theta\left(\frac{d\phi}{dt}\right)^2 + \frac{r-2}{r^3} = 0, \\ \frac{d^2\theta}{dt^2} + \frac{2r-6}{r(r-2)}\left(\frac{d\theta}{dt}\right)\left(\frac{dr}{dt}\right) - \sin\theta\cos\theta\left(\frac{d\phi}{dt}\right)^2 = 0 \text{ and} \\ \frac{d^2\phi}{dt^2} + \frac{2r-6}{r(r-2)}\left(\frac{d\theta}{dt}\right)\left(\frac{dr}{dt}\right) + 2\cot\theta\left(\frac{d\theta}{dt}\right)\left(\frac{d\phi}{dt}\right) = 0. \end{aligned} \tag{4}$$

Using tetrad formalism, velocities are connected from the curved to the flat spacetime. The three spatial velocity components can be expressed as,

$$v^{\hat{r}} = \frac{d\hat{r}}{dt} = \frac{r}{(r-2)}\frac{dr}{dt},\ v^{\hat{\theta}} = \frac{d\hat{\theta}}{dt} = \frac{r\sqrt{r}}{\sqrt{(r-2)}}\frac{d\theta}{dt} \quad \text{and } v^{\hat{\phi}} = \frac{d\hat{\phi}}{dt} = \frac{r\sqrt{r}\sin\theta}{\sqrt{(r-2)}}\frac{d\theta}{dt}. \tag{5}$$

Ray-Tracing process is similar to what has been earlier used in [14, 15, 49, 50].

5 Monte-Carlo Comptonization

Injected photons that are generated from Keplerian disk acquire random velocities from three random numbers. We have modulated the flux of the disk such that radiation from Keplerian disk maximize along the Z-axis and minimize along the equatorial plane. A scattering within the Compton cloud occurs when the optical depth of a photon crosses the critical optical depth τ_c.

We have considered the Klein-Nishina scattering cross section σ and is given by:

$$\sigma = \frac{2\pi r_e^2}{x}\left[\left(1 - \frac{4}{x} - \frac{8}{x^2}\right) ln\,(1+x) + \frac{1}{2} + \frac{8}{x} - \frac{1}{2\,(1+x)^2}\right], \tag{6}$$

where, x is given by,

$$x = \frac{2E}{mc^2}\gamma\left(1 - \mu\frac{v}{c}\right), \tag{7}$$

$r_e = e^2/mc^2$ is the classical electron radius and m is the mass of the electron.

Energy exchange between electron and photon occurs by following Comptonization mechanism. Gravitational redshift modulates the energy of photons in their entire path. The process is similar to what was used in Ghosh, Chakrabarti, Laurent, 2009 [51]; Ghosh, Garain, Chakrabarti, Laurent, 2010 [52]; Ghosh et al. 2011 [14], [15], [53].

After Comptonization, the last scattering point of photons are stored. Panel (a) shows the most energetic photons came from the centre of thick disk where density and temperature are at maximum. Thus, these photons suffer most scatterings (presented in Panel (b)) while leaving the central region.

6 Results

6.1 Spectra

To study the X-Ray imaging, we have considered outer boundary of the Compton cloud to be at 30 and this is the inner boundary of the Keplerian disk. Keplerian disk extends up to 50. This saves significant amount of computational time. The injected and emergent spectra are shown below.

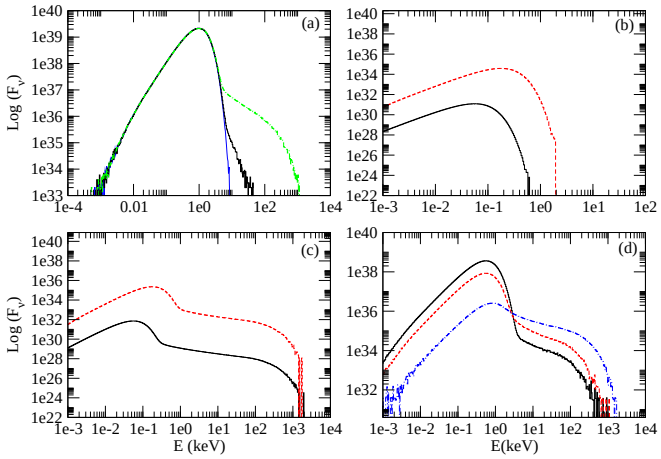


Figure 2: Variation of emergent spectrum with respect to (a) central temperature of Compton cloud keeping $\dot{m}_d = 1.0$ as fixed. The injected spectrum is shown by solid-blue line. The dashed-black and double-dash-dot-green spectra correspond to the central temperatures 50 and 140 keV, respectively. (b) Injected seed photons from truncated disks of accretion rates $\dot{m}_d = 0.00001\,(\mathrm{solid-black})$ and $\dot{m}_d = 0.001\,(\mathrm{dashed-red})$. The central temperature is 200 keV with corresponding emergent spectra in panel (c). (d) Direction dependent emergent spectra for $\dot{m}_d = 0.1$ with $0^\circ - 22.5^\circ$ (solid-black), $45^\circ - 67.5^\circ$ (dash-red) and $67.5^\circ - 90^\circ$ (dot-dash-blue).

Spectral hardening with increasing Compton cloud temperature is shown Fig. 2. Injected spectra for accretion rates of $\dot{m}_d = 0.00001\,(\mathrm{solid-black})$ and $\dot{m}_d = 0.001\,(\mathrm{dashed-red})$ are shown in Fig. 2b. Fig. 2c shows the corresponding Comptonized spectra for the central temperature fixed at 200 keV. The direction dependent source spectra is presented in Fig. 2d (see [14] for further details).

6.2 Image

6.2.1 Image of Keplerian Disk

Due to the increase in Doppler effect the Keplerian disk gets brighter in one side and dims on the other side. The image of the Keplerian disk seen from an inclination angle of 80° is shown in Fig. 3. Only primary photons are considered in this current image. Contribution of secondary photons are minimal compared to the primary photons. The Keplerian disk has the inner edge at 6 and outer edge at 50.

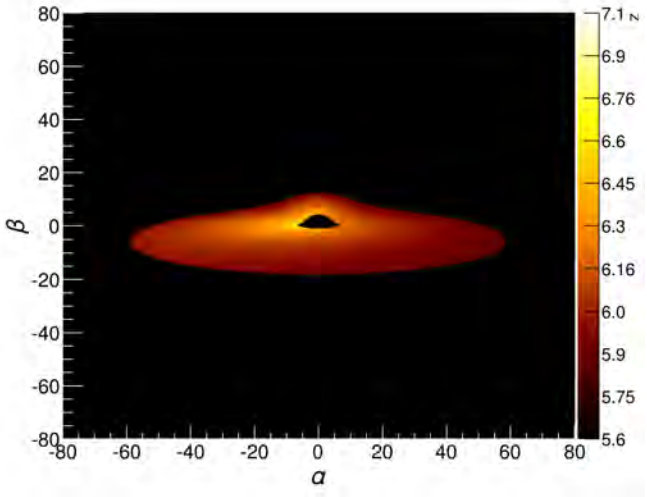


Figure 3: Keplerian disk with $r_{out} = 50.0$ seen from an inclination angle of $80°$. The color bar represents $log(T_{obs})$ of the Keplerian disk. The inner edge (r_{in}) of the Keplerian disk is at $6.0r_g$. Disk rate $\dot{m}_d = 10^{-1}$.

Using proper Ray-Tracing mechanism allows us to have the entire phase space information of the photons in its trajectory.

6.2.2 Image of an Accretion Disk

We constructed the image of an accretion disk on the plane of an observer by event simulation mechanism. Most of the earlier studies in case of X-Ray imaging were done from the reference frame of the observer. This method save computational time. But, it restricts most the information about the physical processes that are occurring in the accretion disk. The variation of images with inclination angle and accretion rates are presented in [14].

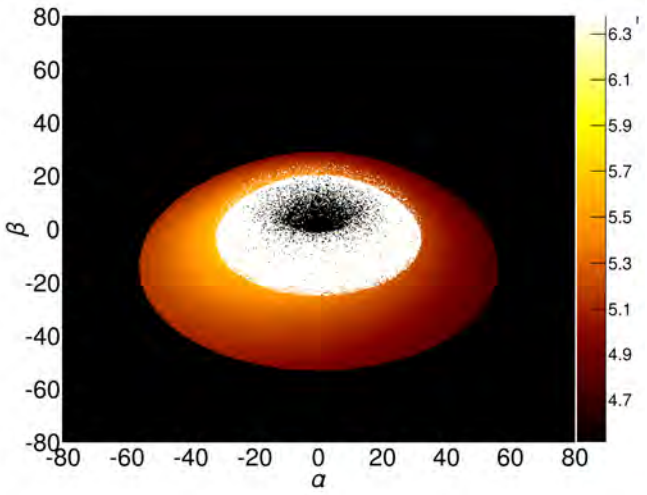


Figure 4: Image of an accretion disk seen from $50°$ having a disk accretion rate of $\dot{m}_d = 10^{-3}$. $log(T_{obs})$ is the color -bar with the scale $10^{4.5} - 10^{6.5}$ Kelvin. Outer edge of Compton cloud is at 30 and Keplerian disk extends up to 50.

As presented in Fig. 4, the central region of the disk glows in hot colors as the hard photons are generated from the central region. The colder geometrically thin but optically thin Keplerian disk exhibits the combined effect of gravitational and Doppler red-shift on the observer plane which distorts the symmetry of the disk on the observer plane.

6.2.3 Disk with Outflows

To study the images of accretion disk in presence of outflows, we have used the hydrodynamic code TVD (Ryu et al. 1997 [42]; Giri & Chakrabarti 2012 [54]). The outer boundary of the Keplerian disk is kept at 100 while the inner boundary is the shock location (around 45). Outflows are natural consequence of accretion. We have considered the reflection symmetry of the outflowing matter. The visual appearance or image changes is a little due to the Lorentz boosting of the Jets (see

Fig. 5). Inner region of the Compton cloud glows due to the hard photons that are produced via inverse Comptonization mechanism. Detailed analysis of outflows are reported in [49].

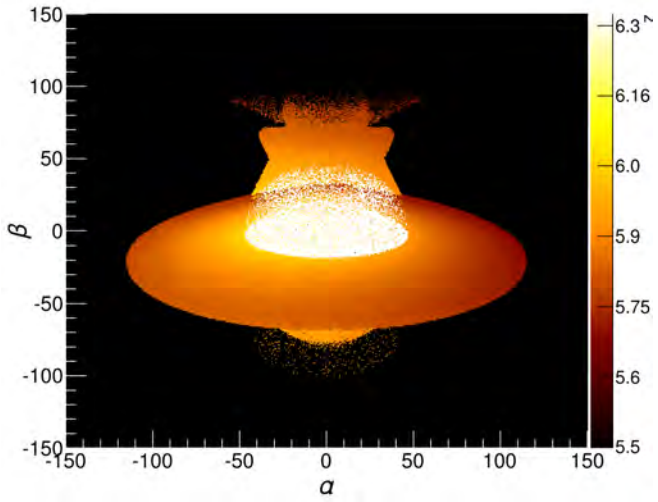


Figure 5: Image of TCAF seen from 70° having a disk accretion rate of $\dot{m}_d = 10^{-1}$. $log(T_{obs})$ is the color -bar with the scale $10^{5.5} - 10^{6.3}$ Kelvin. Outer edge of Compton cloud is at 45 and Keplerian disk extends up to 100.

6.2.4 Observed Spectra

Spectral variation with respect to inclination angle is shown in Fig. 6. The disk rate is kept constant. With increasing inclination angle more and more soft photons from Keplerian disk enters into the Comptonizing region. Gravitational lensing bends the trajectories of photons such that the number photons which are passing through the Compton cloud region increase.

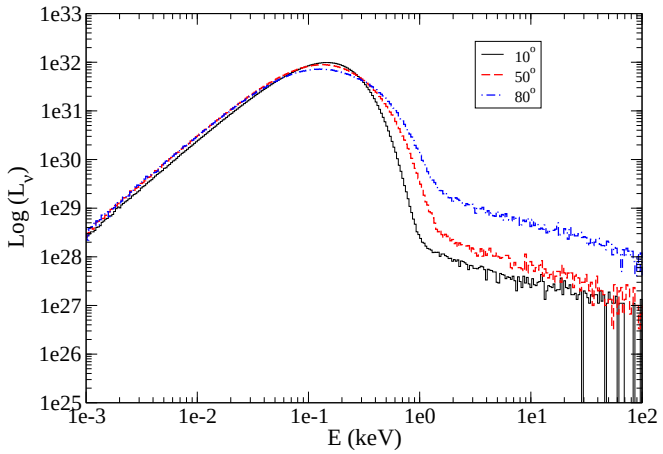


Figure 6: Variation of observed spectrum with inclination angles 10°, 50° and 80° where disk rate fixed at $\dot{m}_d = 10^{-3}$.

Due to gravitational bending of photons, we observe the spectral hardening in higher inclination angle objects (see Fig. 6). Spectra of 50° inclination angle corresponds to the image is shown at Fig. 4. Spectral hardening from our simulation matches with the observational survey that has been done by Heil et al. 2015 [55].

6.3 Time Lag

Time lag properties were first discovered by Miyamoto et al. 1988 [16]. Dutta & Chakrabarti, 2016 [23], explained the physical processes that contribute to the time lag properties. Comptonization, disk reflection, gravitational bending and the inclination angle of the observer. We have added all these features to simulate the variation of time lag properties with shock location X_s, energy of photons, inclination angles and compared our simulations with observational results. We simulate electron clouds of various size from 20 to 65 to study the time lag variation with shock location and QPO frequencies.

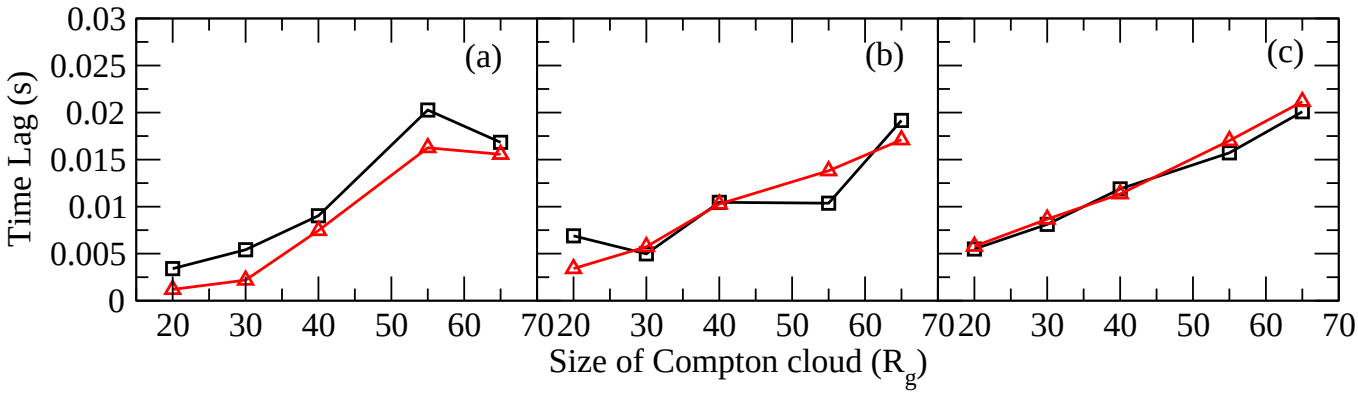


Figure 7: Time lag ($in\ sec$) between $5.0 - 100.0$ keV and $1.0 - 5.0$ keV is plotted with size of the Compton cloud (in r_g unit). (a) Black (square) and red (triangle) lines represent 0° & 10°, (b) Black (square) & red (triangle) lines represents 36° & 50° respectively and (c) Black (square) & red (triangle) lines represent 60° & 70° respectively.

6.3.1 Time Lag Variation with Size of Compton Cloud

To study time lag properties, we have varied the Compton cloud size 20 to 65 r_g. The soft energy band is considered 1 to 5 keV and the hard energy band is considered 5 to 100 keV (Fig. 7). With inclination angles, we have shown the variation of time lag for numerous sizes of Compton cloud (see [15] for further details).

6.3.2 Time Lag Variation with Energy

Soft photons originated from Keplerian disk are intercepted by the electron cloud where inverse Comptonization occurs. This enhances the energy of photons. The maximum energy of electrons can be found in the central region of Compton cloud. The more scattering a photon suffers within cloud the more energetic it becomes. Thus, it spends more time in the Compton cloud.

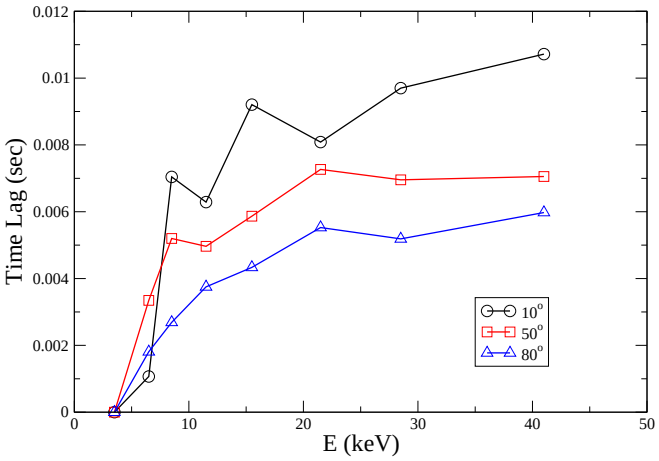


Figure 8: Time lag w.r.t $2.0 - 5.0$ keV is plotted with energy bins for shock locations at $55r_g$ for three different inclination angles. Lag magnitude decreases with inclination angle.

In Fig. 8, we show the time lags of photons of different energy bins having fixed size of the Comton cloud with various inclination angles. More soft photons from the opposite size of the disk are intercepted by the Comton cloud which increase the number of hard photons for an observer at high inclination. The unscattered soft photons from the other side which are arriving directly to the observer delayed even further due to bent trajectory. As a result, when the inclination increase, the lag magnitude decrease. Details can be seen in CCG17b. Earlier in RXTE era, the energy dependent time lag studies were possible up to 30 keV. With NuSTAR (Harrison et al. 2013) [56] and AstroSat (Singh et al. 2014) [57] missions, the time lag studies are being conducted at higher energies (Mishra et al. 2017 [58]; Middleton et al. 2021 [59]).

6.3.3 Comparison with Observation

Chakrabarti & Manickam 2000 [60], showed the QPO frequency is inversely proportional to the shock location. We considered the shock is moving inwards from day 0 with a speed of 500 cm/s.

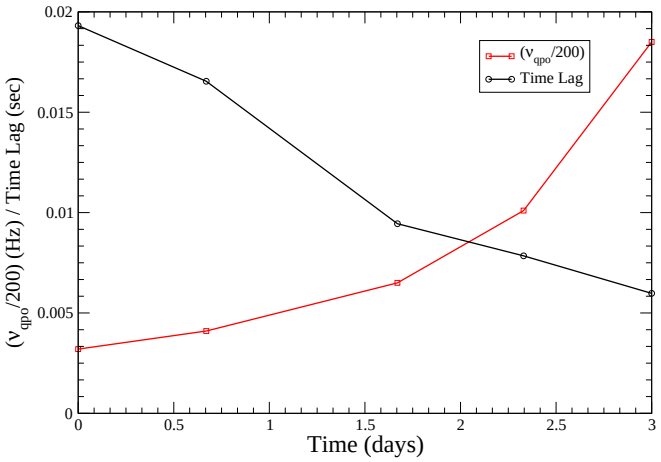


Figure 9: Time lag between $5.0 - 100.0$ and $0.1 - 5.0$ keV (solid-black-circle) photons is plotted for 50° inclination angle along with calculated $\nu_{qpo}/200$ Hz (solid-red-triangle). X-axis represents the day of evolution.

For 50° inclination angle, we show the time lag variation with $\nu_{qpo}/200$. From Fig. 9, we can see the smallest QPO produces largest time lag. This behavior is in complete agreement with the Fig. 4 of Dutta & Chakrabarti, 2016 [23], where the similar variation of time lag and QPO frequency was obtained for GX 339-4 [27], an object with inclination angle around 50° (Zdziarski et al. 1998) [61].

6.3.4 Lag-Energy Spectra in Presence of Outflows

We introduced the hydrodynamic simulation (Ryu et al. 1997 [42]; Giri & Chakrabarti 2012 [54]) to produce the dynamic images of an accretion disk in presence of outflows (Chatterjee et al. 2018) [49].

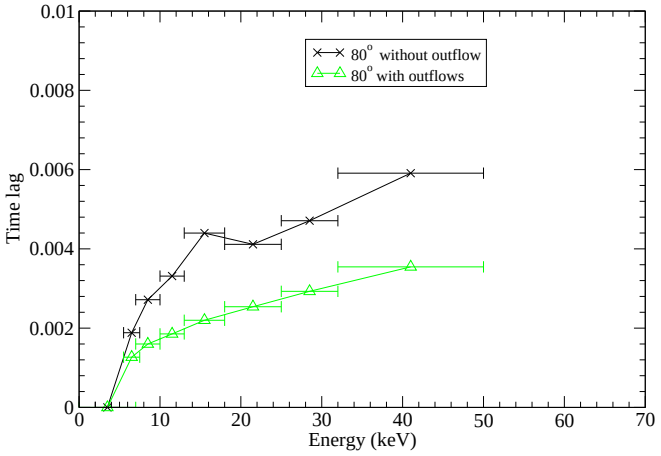


Figure 10: Time lag w.r.t $2.0 - 5.0$ keV is plotted with energy bins for shock locations at $45r_g$ for a source at an inclination angle of 80°. Lag magnitude decreases with the inclusion of the outflows.

Simulated energy dependent time lags with and without the effect of outflows are presented in Fig. 10 for a source at 80° inclination. The soft energy band, considered here, is 2.0-5.0 keV for simulated lag calculations. We find that in the presence of outflows, time lag magnitude decreases. The simulated results shows a general agreement with the observed radio flux soft-lag correlation (Chatterjee et al. 2019) [28]. Currently, the simulation boundary extends up to 100 r_g which is also the outer boundary of the Keplerian disk. The outer boundary of the Compton cloud is around 45 r_g. Studies on the effect of jets for low inclination angle sources will be reported elsewhere.

7 Conclusion

In this article, we presented the theoretical images of Two Component Advective Flow (Chakrabarti & Titarchuk 1995) [34] using Monte-Carlo simulations. The model has been successful to explain the various spectral states present in the outbursting or persistent black hole sources. Here, we have showed the images of modeled accretion disks around a non-rotating black hole in presence of Comptonization. Using Ray-Tracing method, we have showed that even a very simplified version of the model is able explain the observational time lag properties.

Monte-Carlo technique allowed us to investigate the energy transfer between electrons and photons in the inhomogeneous Compton cloud and its implications on the inclination dependent spectral and temporal properties of the Galactic black holes.

The signatures of disk-jet connections are found in the spectral (Merloni et al. 2003 [25]; Corbel et al. 2003 [26]) and temporal (Gandhi et al. 2008 [27]; Reig et al. 2018 [32]; Reig & Kylafis 2019 [33]; Chatterjee et al. 2019 [28]; Patra et al. 2019 [29]; Chatterjee et al. 2020 [30]) properties of the GBHs. Both high and low inclination sources manifest the jet interception of photons in different manner. For, high inclination, the feedback from the jet could downscatter photons causing soft lags (see Chatterjee et al. 2019) [28]. Whereas, for low inclination sources, the jet interception could introduce more Comptonization aiding to the hard lag (Chatterjee et al. 2020) [30]. In future, we expect to extend the numerical simulations on low inclination sources where disk-jet interactions will be explored through spectral and temporal features of the outbursting and persistent sources.

Acknowledgements

The work of AC has been supported by the Canadian Space Agency (CSA) and the Natural Sciences and Engineering Research Council of Canada (NSERC). AC acknowledges the suggestions from the Editors of the book.

References

[1] J. P. Luminet, "Image of a spherical black hole with thin accretion disk." *A&A*, vol. 75, pp. 228–235, May 1979.

[2] J. Fukue and T. Yokoyama, "Color photographs of an accretion disk around a black hole," *Publications of the Astronomical Society of Japan*, vol. 40, no. 1, pp. 15–24, Jan. 1988.

[3] S. U. Viergutz, "Image generation in Kerr geometry. I. Analytical investigations on the stationary emitter-observer problem," *A&A*, vol. 272, p. 355, May 1993.

[4] J.-A. Marck, "Short-cut method of solution of geodesic equations for Schwarzchild black hole," *Classical and Quantum Gravity*, vol. 13, no. 3, pp. 393–402, Mar. 1996.

[5] M. D. Johnson, V. L. Fish, S. S. Doeleman, and et al., "Resolved magnetic-field structure and variability near the event horizon of Sagittarius A*," *Science*, vol. 350, no. 6265, pp. 1242–1245, Dec. 2015.

[6] B. C. Bromley, F. Melia, and S. Liu, "Polarimetric Imaging of the Massive Black Hole at the Galactic Center," *ApJL*, vol. 555, no. 2, pp. L83–L86, Jul. 2001.

[7] Event Horizon Telescope Collaboration, K. Akiyama, A. Alberdi, W. Alef, and et al., "First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole," *ApJL*, vol. 875, no. 1, p. L1, Apr. 2019.

[8] Event Horizon Telescope Collaboration, K. Akiyama, A. Alberdi, W. Alef, and et al., "First M87 Event Horizon Telescope Results. II. Array and Instrumentation," *ApJL*, vol. 875, no. 1, p. L2, Apr. 2019.

[9] Event Horizon Telescope Collaboration, K. Akiyama, A. Alberdi, W. Alef, and et al., "First M87 Event Horizon Telescope Results. III. Data Processing and Calibration," *ApJL*, vol. 875, no. 1, p. L3, Apr. 2019.

[10] ——, "First M87 Event Horizon Telescope Results. IV. Imaging the Central Supermassive Black Hole," *ApJL*, vol. 875, no. 1, p. L4, Apr. 2019.

[11] ——, "First M87 Event Horizon Telescope Results. V. Physical Origin of the Asymmetric Ring," *ApJL*, vol. 875, no. 1, p. L5, Apr. 2019.

[12] ——, "First M87 Event Horizon Telescope Results. VI. The Shadow and Mass of the Central Black Hole," *ApJL*, vol. 875, no. 1, p. L6, Apr. 2019.

[13] R. A. Sunyaev and L. G. Titarchuk, "Comptonization of X-Rays in Plasma Clouds - Typical Radiation Spectra," *A&A*, vol. 86, p. 121, Jun. 1980.

[14] A. Chatterjee, S. K. Chakrabarti, and H. Ghosh, "Images and spectral properties of two-component advective flows around black holes: effects of photon bending," *MNRAS*, vol. 465, no. 4, pp. 3902–3912, Mar. 2017.

[15] ——, "Temporal evolution of photon energy emitted from two-component advective flows: origin of time lag," *MNRAS*, vol. 472, no. 2, pp. 1842–1849, Dec. 2017.

[16] S. Miyamoto, S. Kitamoto, K. Mitsuda, and T. Dotani, "Delayed hard X-rays from Cygnus X-l," *Nature*, vol. 336, no. 6198, pp. 450–452, Dec. 1988.

[17] M. Böttcher and E. P. Liang, "A New Model for the Hard Time Lags in Black Hole X-Ray Binaries," *ApJL*, vol. 511, no. 1, pp. L37–L40, Jan. 1999.

[18] D. Lin, I. A. Smith, E. P. Liang, and M. Böttcher, "Complex Phase Lag Behaviors of the 0.5-10 HZ Quasi-periodic Oscillations in GRS 1915+105," *ApJL*, vol. 543, no. 2, pp. L141–L144, Nov. 2000.

[19] J. Poutanen and A. C. Fabian, "Spectral evolution of magnetic flares and time lags in accreting black hole sources," *MNRAS*, vol. 306, no. 3, pp. L31–L37, Jul. 1999.

[20] R. A. Remillard and J. E. McClintock, "X-Ray Properties of Black-Hole Binaries," *Annual Review of Astronomy and Astrophysics*, vol. 44, no. 1, pp. 49–92, Sep. 2006.

[21] M. Gierliński, M. Middleton, M. Ward, and C. Done, "A periodicity of ~1hour in X-ray emission from the active galaxy RE J1034+396," *Nature*, vol. 455, no. 7211, pp. 369–371, Sep. 2008.

[22] W. N. Alston, M. L. Parker, J. Markevičiūtė, A. C. Fabian, M. Middleton, A. Lohfink, E. Kara, and C. Pinto, "Discovery of an $\sim$2-h high-frequency X-ray QPO and iron Kα reverberation in the active galaxy MS 2254.9-3712," *MNRAS*, vol. 449, no. 1, pp. 467–476, May 2015.

[23] B. G. Dutta and S. K. Chakrabarti, "Temporal Variability from the Two-Component Advective Flow Solution and Its Observational Evidence," *ApJ*, vol. 828, no. 2, p. 101, Sep. 2016.

[24] J. van den Eijnden, A. Ingram, P. Uttley, S. E. Motta, T. M. Belloni, and D. W. Gardenier, "Inclination dependence of QPO phase lags in black hole X-ray binaries," *MNRAS*, vol. 464, no. 3, pp. 2643–2659, Jan. 2017.

[25] A. Merloni, S. Heinz, and T. di Matteo, "A Fundamental Plane of black hole activity," *MNRAS*, vol. 345, no. 4, pp. 1057–1076, Nov. 2003.

[26] S. Corbel, M. A. Nowak, R. P. Fender, A. K. Tzioumis, and S. Markoff, "Radio/X-ray correlation in the low/hard state of GX 339-4," *A&A*, vol. 400, pp. 1007–1012, Mar. 2003.

[27] P. Gandhi, K. Makishima, M. Durant, A. C. Fabian, V. S. Dhillon, T. R. Marsh, J. M. Miller, T. Shahbaz, and H. C. Spruit, "Rapid optical and X-ray timing observations of GX 339-4: flux correlations at the onset of a low/hard state," *MNRAS*, vol. 390, no. 1, pp. L29–L33, Oct. 2008.

[28] A. Chatterjee, B. G. Dutta, D. Patra, S. K. Chakrabarti, and P. Nandi, "Discovery of Jet-Induced Soft Lags of XTE J1550-564 during Its 1998 Outburst," in *Recent Progress in Relativistic Astrophysics*, C. Bambi and S. Nampalliwar, Eds., May 2019, p. 9.

[29] D. Patra, A. Chatterjee, B. G. Dutta, S. K. Chakrabarti, and P. Nandi, "Evidence of Outflow-induced Soft Lags of Galactic Black Holes," *ApJ*, vol. 886, no. 2, p. 137, Dec. 2019.

[30] A. Chatterjee, B. G. Dutta, P. Nandi, and S. K. Chakrabarti, "Time-domain variability properties of XTE J1650-500 during its 2001 outburst: evidence of disc-jet connection," *MNRAS*, vol. 497, no. 4, pp. 4222–4230, Oct. 2020.

[31] L. A. Pozdnyakov, I. M. Sobol, and R. A. Syunyaev, "Comptonization and the shaping of X-ray source spectra - Monte Carlo calculations," *Astrophysics and Space Physics Reviews*, vol. 2, pp. 189–331, Jan. 1983.

[32] P. Reig, N. D. Kylafis, I. E. Papadakis, and M. T. Costado, "The photon-index-time-lag correlation in black hole X-ray binaries," *MNRAS*, vol. 473, no. 4, pp. 4644–4652, Feb. 2018.

[33] P. Reig and N. D. Kylafis, "Inclination effects on the X-ray emission of Galactic black-hole binaries," *A&A*, vol. 625, p. A90, May 2019.

[34] S. Chakrabarti and L. G. Titarchuk, "Spectral Properties of Accretion Disks around Galactic and Extragalactic Black Holes," *ApJ*, vol. 455, p. 623, Dec. 1995.

[35] D. Molteni, G. Lanzafame, and S. K. Chakrabarti, "Simulation of Thick Accretion Disks with Standing Shocks by Smoothed Particle Hydrodynamics," *ApJ*, vol. 425, p. 161, Apr. 1994.

[36] M. Abramowicz, M. Jaroszynski, and M. Sikora, "Relativistic, accreting disks." *A&A*, vol. 63, pp. 221–224, Feb. 1978.

[37] M. Kozlowski, M. Jaroszynski, and M. A. Abramowicz, "The analytic theory of fluid disks orbiting the Kerr black hole." *A&A*, vol. 63, no. 1-2, pp. 209–220, Feb. 1978.

[38] M. J. Rees, M. C. Begelman, R. D. Blandford, and E. S. Phinney, "Ion-supported tori and the origin of radio jets," *Nature*, vol. 295, no. 5844, pp. 17–21, Jan. 1982.

[39] S. K. Chakrabarti, "The natural angular momentum distribution in the study of thick disks around black holes," *ApJ*, vol. 288, pp. 1–6, Jan. 1985.

[40] ——, "Analytic structure of cosmic radio jets - A preliminary investigation," *ApJ*, vol. 288, pp. 7–13, Jan. 1985.

[41] ——, "Von-Zeipel Surfaces," *MNRAS*, vol. 245, p. 747, Aug. 1990.

[42] D. Ryu, S. K. Chakrabarti, and D. Molteni, "Zero-Energy Rotating Accretion Flows near a Black Hole," *ApJ*, vol. 474, no. 1, pp. 378–388, Jan. 1997.

[43] S. K. Chakrabarti, "Grand Unification of Solutions of Accretion and Winds around Black Holes and Neutron Stars," *ApJ*, vol. 464, p. 664, Jun. 1996.

[44] S. K. Chakrabarti and S. Das, "Properties of accretion shock waves in viscous flows around black holes," *MNRAS*, vol. 349, no. 2, pp. 649–664, Apr. 2004.

[45] S. Nagarkoti and S. K. Chakrabarti, "Upper Limit of the Viscosity Parameter in Accretion Flows around a Black Hole with Shock Waves," *ApJ*, vol. 816, no. 1, p. 7, Jan. 2016.

[46] D. N. Page and K. S. Thorne, "Disk-Accretion onto a Black Hole. Time-Averaged Structure of Accretion Disk," *ApJ*, vol. 191, pp. 499–506, Jul. 1974.

[47] S. K. Garain, H. Ghosh, and S. K. Chakrabarti, "Quasi-periodic oscillations in a radiative transonic flow: results of a coupled Monte Carlo-TVD simulation," *MNRAS*, vol. 437, no. 2, pp. 1329–1336, Jan. 2014.

[48] S. Chandrasekhar, *The Mathematical Theory of Black Holes*, 1998.

[49] A. Chatterjee, S. K. Chakrabarti, H. Ghosh, and S. K. Garain, "Images and spectra of time-dependent two-component advective flow in presence of outflows," *MNRAS*, vol. 478, no. 3, pp. 3356–3366, Aug. 2018.

[50] A. Chatterjee and S. K. Chakrabarti, "Impacts of photon bending on observational aspects of two component advective flow," in *Fourteenth Marcel Grossmann Meeting - MG14*, M. Bianchi, R. T. Jansen, and R. Ruffini, Eds., Jan. 2018, pp. 1066–1071.

[51] H. Ghosh, S. K. Chakrabarti, and P. Laurent, "Monte Carlo Simulations of the Thermal Comptonization Process in a Two-Component Accretion Flow around a Black Hole," *International Journal of Modern Physics D*, vol. 18, no. 11, pp. 1693–1706, Jan. 2009.

[52] H. Ghosh, S. K. Garain, S. K. Chakrabarti, and P. Laurent, "Monte Carlo Simulations of the Thermal Comptonization Process in a Two-Component Accretion Flow around a Black Hole in the Presence of AN Outflow," *International Journal of Modern Physics D*, vol. 19, no. 5, pp. 607–620, Jan. 2010.

[53] H. Ghosh, S. K. Garain, K. Giri, and S. K. Chakrabarti, "Effects of Compton cooling on the hydrodynamic and the spectral properties of a two-component accretion flow around a black hole," *MNRAS*, vol. 416, no. 2, pp. 959–971, Sep. 2011.

[54] K. Giri and S. K. Chakrabarti, "Hydrodynamic simulations of viscous accretion flows around black holes," *MNRAS*, vol. 421, no. 1, pp. 666–678, Mar. 2012.

[55] L. M. Heil, P. Uttley, and M. Klein-Wolt, "Inclination-dependent spectral and timing properties in transient black hole X-ray binaries," *MNRAS*, vol. 448, no. 4, pp. 3348–3353, Apr. 2015.

[56] F. A. Harrison, W. W. Craig, F. E. Christensen, and et al., "The Nuclear Spectroscopic Telescope Array (NuSTAR) High-energy X-Ray Mission," *ApJ*, vol. 770, no. 2, p. 103, Jun. 2013.

[57] K. P. Singh, S. N. Tandon, P. C. Agrawal, and et al., "ASTROSAT mission," in *Space Telescopes and Instrumentation 2014: Ultraviolet to Gamma Ray*, ser. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, T. Takahashi, J.-W. A. den Herder, and M. Bautz, Eds., vol. 9144, Jul. 2014, p. 91441S.

[58] R. Misra, J. S. Yadav, J. Verdhan Chauhan, P. C. Agrawal, H. M. Antia, M. Pahari, V. R. Chitnis, D. Dedhia, T. Katoch, P. Madhwani, R. K. Manchanda, B. Paul, and P. Shah, "AstroSat/LAXPC Observation of Cygnus X-1 in the Hard State," *ApJ*, vol. 835, no. 2, p. 195, Feb. 2017.

[59] M. J. Middleton, D. J. Walton, W. Alston, and et al., "NuSTAR reveals the hidden nature of SS433," *MNRAS*, vol. 506, no. 1, pp. 1045–1058, Sep. 2021.

[60] S. K. Chakrabarti and S. G. Manickam, "Correlation among Quasi-Periodic Oscillation Frequencies and Quiescent-State Duration in Black Hole Candidate GRS 1915+105," *ApJL*, vol. 531, no. 1, pp. L41–L44, Mar. 2000.

[61] A. A. Zdziarski, J. Poutanen, J. Mikolajewska, M. Gierlinski, K. Ebisawa, and W. N. Johnson, "Broad-band X-ray/gamma-ray spectra and binary parameters of GX 339-4 and their astrophysical implications," *MNRAS*, vol. 301, no. 2, pp. 435–450, Dec. 1998.

On the Galaxy Rotation Curves

Debabrata Adak *
Government General Degree College, Singur, Hooghly-712409
* Corresponding author: debabrata.adak.sinp@gmail.com

Abstract

This review on the galaxy rotation curve provides the reader a brief historical account of observational discoveries of galaxies and their properties. The galaxy rotation curve as indicated by Keplerian dynamics deviates for the outward stars as their distances increase from the centre of the galaxy. This deviation may account for the presence of dark matter in the galaxy. However there are other promising alternatives which can also explain these deviations. In this article we review all the possible solutions so far in this context for the explanation of the galaxy rotation curves.

Keywords: Galaxy Rotation; Dark Matter; MOND.

1 Introduction

The galaxies were first identified by the French astronomer Charles Messier in the $17^{\rm th}$ century [1, 2]. Having no knowledge about what they are, he made a catalogue, known as Messier Catalogue, of 110 star clusters and spiral nebulae. In this Catalogue, each astronomical object was named by the letter 'M' followed by a numeric digit. For instances, M1 is identified as a supernova remnant which is also known as the Crab Nebula. M31 was identified as Andromeda Galaxy which was first observed around the year 964 by the Persian astronomer Abd al-Rahman al-Sufi [3]. Until 1900, astronomers observed numerous faint, fuzzy objects in sky but could not gather much knowledge about them. These fuzzy objects were believed by some astronomers to be the "island universes" beyond Milky Way at that time. There were other astronomers who had a thought about these objects as being the planetary systems in the early stage of evolution like our solar system [4]. The American astronomer Vesto Slipher, in 1912, was the first to measure the radial velocities of the galaxies [5, 6] and his study on twenty five spiral nebulae and globular clusters led to the discovery that the distant galaxies are moving away from us or the Milky Way [4]. Despite being known about the receding of the galaxies from Slipher's observations, those days it was highly controversial to get rid of the other classes of thoughts regarding those nebulae i.e., whether or not they were island universes outside our own galaxy Milky Way. In this context this is worth mentioning about the "Great Debate" in 1920 between Harlow Shapley and Heber Doust Curtis [7, 8, 9, 10, 11]. In the debate Shapley argued that the spiral nebulae are the parts of Milky Way galaxy. On the contrary, Curtis took stand by the idea that the spiral nebulae are the separate galaxies or the "island universes" outside our own galaxy Milky Way and they are comparable in size and nature to our own galaxy. This debate about the spiral nebulae continued until another American astronomer Edwin Hubble published his results of the study of distances of the galaxies in 1929. His observations on M31 (known as Andromeda Nebula) and M33 (known as Triangulum Galaxy) in 1924, concluded that those galaxies were too distant to be a part of our Milky Way [12, 13]. Hence they are separate galaxies outside Milky Way. Moreover his results revealed that the velocity of a galaxy is proportional to its distance i.e., the distant galaxies move faster than the nearby ones. This is known as the Hubble's law and given by

$$\begin{aligned} v &\propto d\,, \\ v &= H_0 d\,, \end{aligned} \tag{1}$$

where v and d are the radial velocity and the distance of the galaxy and H_0 is the proportionality constant also known as the Hubble's constant which he measured to have a value around 65 Km.s^{-1}.Mpc^{-1} [14]. Hubble's results for the first time conclusively argued that the Universe was expanding [1].

Slipher's investigations on Virgo in 1913 showed that the spectral lines were not only shifted towards red (implication of receding of the nebula) but also were slightly inclined which was the indication of the rotational motion of the nebula [15]. Later in 1917, Francis Pease's observation on the Andromeda galaxy revealed that the central region of it rotates with almost constant angular velocity [16]. This observed rotational velocity of Andromeda was used to calculate its mass and mass-to-light ratio by several astronomers and found to be consistent with the results that were obtained by Hubble [17] and Oort [18]. In 1937, Fritz Zwicky claimed in his work [19] that the lower limits on the masses of galaxies can be put from the observed luminosities, however the calculation of masses of the galaxies is not possible from the observed rotational velocities alone. Later in 1939, Horace Babcock published the results of his observations on M31 in his PhD dissertation [20] however the results are not consistent with the modern day observations. Revolution in observational astronomy began around 1950's when the first radio astronomical observations of rotation curve of M31 galaxy (2°away from its centre) was

[1] Slipher's results also had the cosmological implication of expanding Universe [5, 6] before Hubble.

published by H. C. van de Hulst, Jean Jacques Raimond, and Hugo van Woerden in 1957 [21]. In a publication [22], Maartin Schmidt explained the results of [21] on the basis of constant mass to light ratio of M31. Since the study was confined in the region of 2°away from the galactic centre, it was not possible to concluded anything about the central or outer regions of Andromeda [22]. Despite these extensive studies of rotation of galaxies, there was almost no sense of crisis among the astronomers that the observations of rotation curves are in conflict with the present day understanding of galaxies as argued by Robert H. Sanders in his book [23].

However the situation began to change drastically in the early 1970s when Kent Ford and Vera Rubin published their spectroscopically observed rotation curve on Andromeda galaxy (M31) [24] with the image tube spectrograph designed by Kent Ford himself. Their observations included the stars near the edge of the Andromeda galaxy. Together with this the other studies of Rubin with K. W. Ford and N. Thonnard [25, 26] suggested that the orbital velocities of most of the stars in the spiral galaxies are nearly same which was the implication of growing mass of the galaxies with the radial distances. D. Rogstad and G. Shostak [27] in 1972, analysed the rotation curves of five galaxies namely M33, NGC 2403, IC 342, M101 and NGC 6946 by the radio telescope at the Owens Valley Radio Observatory. They found, from their analyses, the rotation curves to become flat at the largest observed radius. Morton Roberts was the first to realise the implications of the observed flatness of the galaxy rotation curves at the large radius. M. Roberts in 1972 published his research work [28] on galaxy M31 with R. Whitehurst and extended further study on M81 and M101 in 1973 with Arnold Rots [29]. The conclusion that was put forward from these analyses was pointing towards the existence of significant matter density at the large distances from the galactic centre which is responsible for the flatness of rotation curves at the large radius. In his PhD thesis (1978), Albert Bosma [30] put forward the results of radio observation of 25 galaxies. He also found the flat rotation curve for the largest radius observed which was larger than the optical size of the galaxy. Hence the interesting conclusion was that the masses of the galaxies continue to grow beyond the visibly seen region occupied by stars and gas constituting that galaxy. Rubin, Ford and Norbert Thonnard also in 1978, put forward the results of optical observation of rotation curves in the optical region for ten spiral galaxies of considerably high luminosity [25]. The results of these observations were similar to that they found earlier, i.e., the rotation curves flattens out for the outermost measured radius. This work is widely accepted in the literature. However the fact that is to be noticed is that the optical measurements cannot meet the distances as large as those that are probed by radio observations.

The aim of this article is to provide the readers with a brief historical account of the observations of the galaxy rotation curves and their corresponding mass measurements in chronological order. In the Sec. 2, we describe the first radio observation of van de Hulst, Jean Jacques Raimond, and Hugo van Woerden [21] and mass measurement of Maartin Schmidt [22]. The Sec. 3 next to it, describes the results of observations obtained by Vera Rubin, Kent Ford and N. Thonnard [24, 25, 26]. In Sec. 4, the physical implications of galaxy rotation curve is discussed in the Newtonian gravity background. Here we discuss how nonluminious matter dubbed dark matter is important in explaining the galaxy rotation curves. The Sec. 5 is dedicated for the current status of dark matter. Other possible alternatives for explaining the galaxy rotation curves are discussed in Secs. 6, 7 and eventually bring about the conclusions in Sec. 8.

2 Radio Observations of 1950s

In 1957, the first radio measurements of Andromeda [21] was put forward by H. C. van de Hulst, E. Raimond and H. van Woerden with the observations that the rotational velocity of Andromeda is very similar to that of the Milky way galactic system with a time period of revolution of $T = 6.12(R/v) \times 10^9$ years, where R (expressed in kpc) and v (expressed in km/sec) are the distance form galactic centre and orbital velocity respectively. Fig. 1, adopted from the Ref. [21], shows a comparison between the rotational velocities in Andromeda Nebula with the Milky Way galaxy. The dashed curve shows the rotational velocities as estimated by Schwarzschild in his work [31]. The total mass in the first and final model of the galactic system was computed by Schmidt in 1956 [32]. Several authors before Schmidt put forward models for determination of mass distribution of Milky Way. It is worth mentioning about Oort's 1952's work [33] which consisted of seven homogeneous spheroids having discontinuity in the derivative of the force at the boundary. Aller and Mayall [34] studied the rotation of M33 in 1942 and Wyse and Mayall before Oort, in the same year for the estimation of mass distribution considered a model of M31 and M33 galaxy as a plane circular disc where the density changes gradually outwards [35]. Combining two aforesaid models, Perek [36] introduced in literature, the non-homogeneous spheroid model of Milky Way galaxy with the arbitrary density profile. However assuming the ellipsoidal velocity distribution everywhere in the galactic system, Parenago [37, 38] derived a model in 1950. Schmidt [32] in 1956 used this non-homogeneous spheroids model for determining the mass distribution in galactic system and later in 1957 for Andromeda nebula [22]. Considering M31 not exactly planner, but a set of two non-homogeneous spheroids the required mass distribution for giving rise to those rotation curves were estimated by Schmidt [22]. Schmidt argued that in the first spheroid, inner part has the density is given by,

$$\rho_1 \;=\; -29.324 + \frac{30.888}{a} \times 10^{-24}\,\mathrm{gm.cm}^{-3}\,,\;\; (\text{inner part of Andromeda nebula}\,,) \tag{2}$$

where a is the semi-axis of the spheroid and is given by,

$$a \;=\; \sqrt{x^2 + \Big(\frac{z}{0.07}\Big)^2}\,, \tag{3}$$

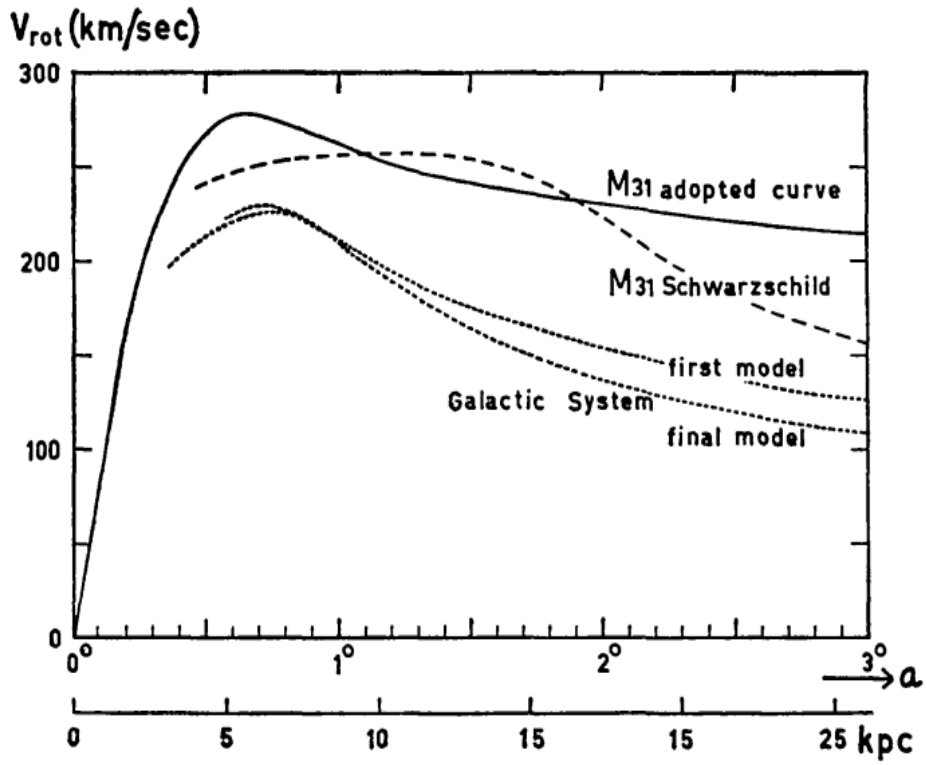


Figure 1: Rotational velocity of stars in the Andromeda nebula and in the Milky Way galaxy [21].

in which x is the distance from the axis and z is the distance from the plane of the galaxy. The factor 0.07 ($= 1 - e^2$, e being the eccentricity of the spheroid) is the axial ratio for the Milky way galaxy [32] and also used arbitrarily for Andromeda nebula. For the outer part the density in the spheroid is given by,

$$\rho_1 \quad = \quad \frac{3.807}{a^4} \times 10^{-24}\,\mathrm{gm.cm}^{-3}\,, \;\; (\text{outer part of Andromeda nebula},) \tag{4}$$

Inner and outer part of the first spheroid is separated at $a = 0.79°$. The density profile for the second spheroid is

$$\rho \quad = \quad -0.494 + \frac{2.447}{a}\,, \tag{5}$$

where the boundary between two spheroids is given by $a = 4.957°$. This is worth mentioning here that $1°$ corresponds to 11 kpc. The total mass of Andromeda with this density profile turns out to be $M = 3.38 \times 10^{11}\ M_\odot$. Later in 1957, Hulst, Raimond and Woerden derived the density of hydrogen (Fig. 2) in Andromeda as well as in Milky Way from the observations of 21-cm line [21]. From the Fig. 2, it is evident that the density of Andromeda nebula is similar to that of Milky Way galaxy.

3 1970's Revolution: Rubin's Observations

In 1970, Rubin and Ford furnished results of the spectral emissions sixty seven regions emitting Hα lines from M31 in distance range 3 to 24 kpc from the galactic centre with the DTM image-tube spectrograph developed by Ford himself [24]. In the Fig. 3 the rotation curve as obtained by Rubin and Ford is shown. The minimum near 2 kpc distance from the center of galaxy corresponds to a narrow N II emission region of M31. They used this rotational velocity curve to make an estimation of mass of M31 using the mass formula given by Kuzmin [39] and Brandt [40]

$$M \quad = \quad \frac{2}{G\pi}\int_0^R \frac{v^2 a da}{\sqrt{R^2 - a^2}}\,, \tag{6}$$

where G is the Newton's gravitational constant, v is the rotational velocity and M is the mass out to $a = R$. The integration was carried out on IBM 1130 numerically with the substitution $a = R\sin\theta$. Their results showed that near $r = 2$ kpc the density is very low and for the region $R > 4$ kpc the total mass of the Andromeda increases almost linearly upto $R = 14$ kpc. Thereafter the rate of increment total mass slows down. The total mass estimated upto $R = 24$ kpc is $M = (1.85 \pm 0.1) \times 10^{11} M_\odot$, half of which is confined in the region $R \leq 9$ kpc. Having been continuing the study, later in 1978, a paper [25] came up where Vera C. Rubin, W. Kent Ford and Norbert Thonnard extended their study to 10 spiral galaxies of high luminosity. To the utter surprise, they found all the rotation curves to be approximately flat upto a distance of 50 kpc from the galactic centres. Their results are adopted in Figs. 4 and 5. Needless to mention that the rotation curves

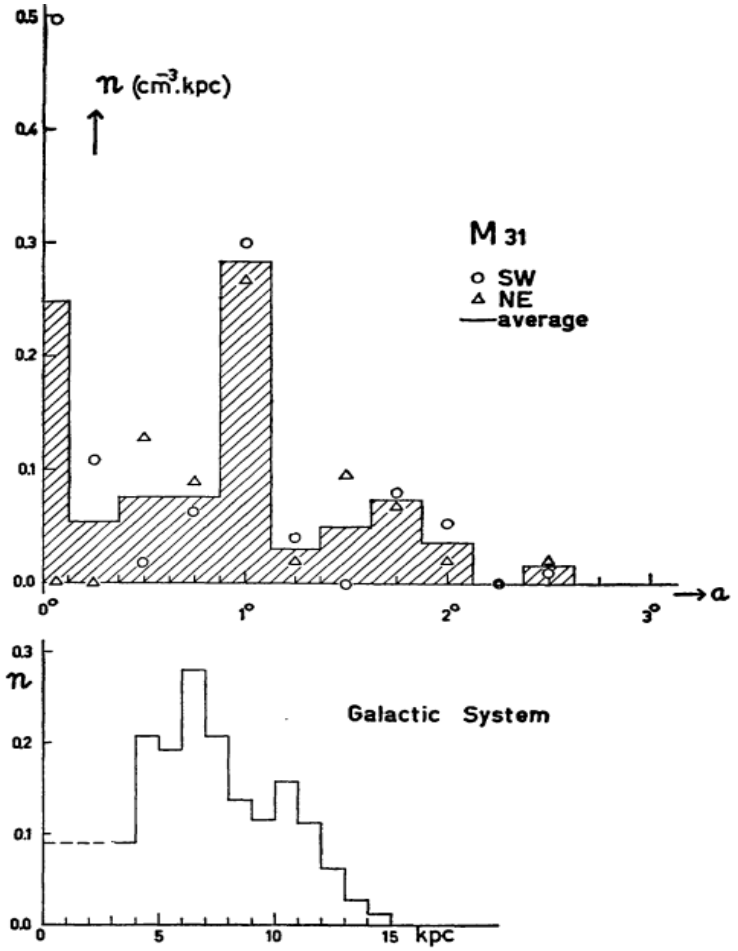


Figure 2: Density of hydrogen in the Andromeda nebula and in the Milky Way galaxy [21].

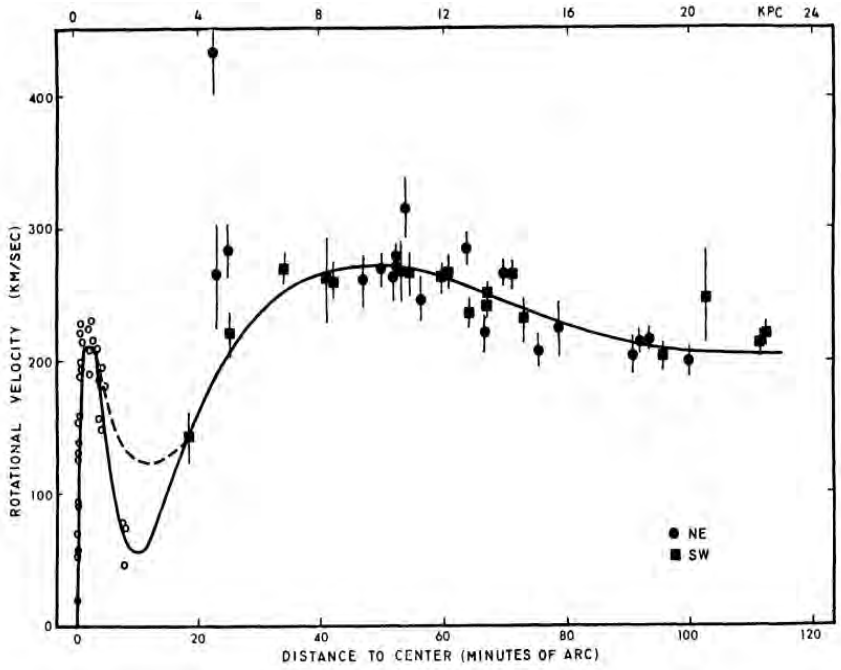


Figure 3: Rotation curves M31 [24].

for all the galaxies in Figs. 4 and 5 flattens out to large radius. From their analysis, they argued that the linear increase in the mass of the galaxies with the radius results in the flatness of the rotation curves. In another plot (Fig. 6) they showed the mass gradient remains linear for all the galaxies they studied, however the slope differs in case of different galaxies. Moreover they suggested that there should be a correlation between mass and luminosity of the galaxies as expected from

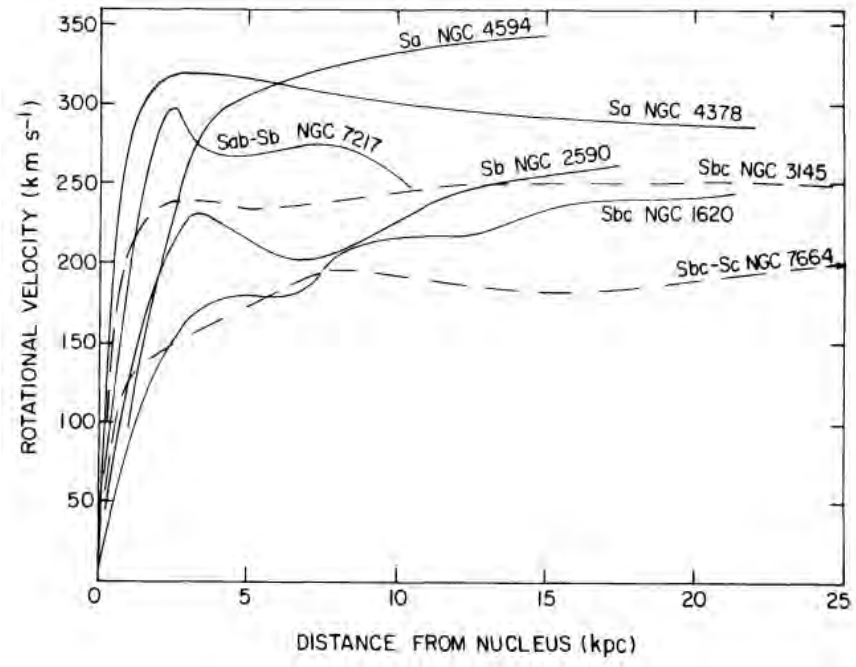


Figure 4: Rotation curves for seven galaxies are shown here. [25].

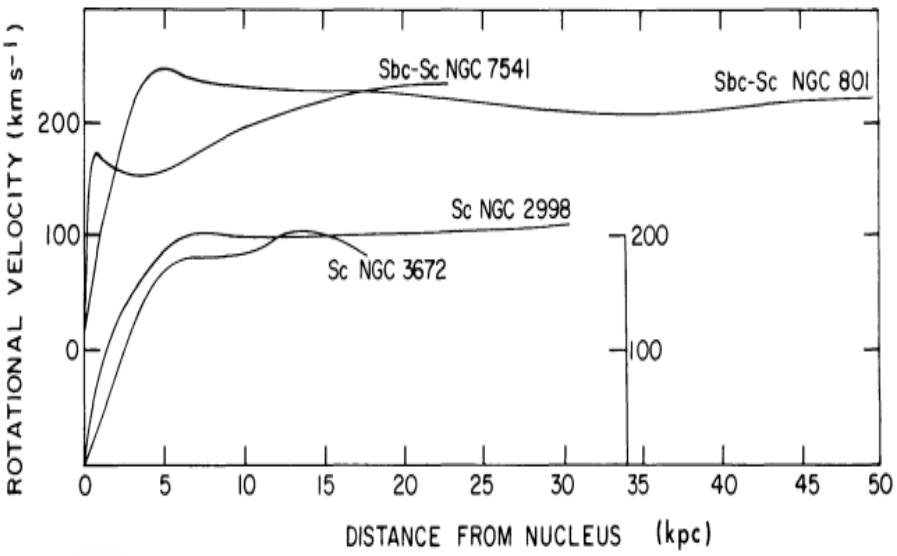


Figure 5: Rotation curves for four galaxies which implies the lack of Tully-Fisher relation (see Sec. A) [25].

the flatness of the rotation curves such as the Holmberg relation of luminosity and the radius [41] given by

$$L = r^{2.4}. \tag{7}$$

However they preferred to keep it open and did not reach any conclusion. In 1980, in another work [26] on 21 Sc galaxies in the luminosity, mass and radius range $3 \times 10^9 - 2 \times 10^{11} L_\odot$, $10^{10} - 2 \times 10^{12} M_\odot$ and $4 - 122$ kpc respectively, they furnished the results of flat rotation curves for very large galaxies. The implication of this results was that the mass of the galaxies is neither centrally concentrated nor converges to a limit at the optical boundary of the galaxies but increases with radius linearly. They argued that beyond the optical edge of the galaxies there exists the non-luminous matter. In the meantime, 1972's other radio observations on five Scd galaxies (M33, NGC 2403, IC 342, M101 and NGC6946) of Rogstad and Shostak [27] argued that the total mass of the galaxies remained unknown because of the flat rotation curves however they found the mass luminosity ratios around 20 at Holmberg radius of the galaxies. In a work, M. S. Roberts and A. H. Rots in 1973 [29] first argued from their study of 3 spiral galaxies that significant amount of matter exists at the large distances and the galaxies are larger than their photometric measurements. In his PhD thesis, Albert Bosma in 1978 argued from radio observations of 25 galaxies that the galaxies exhibit flat rotation curves and hence their mass keeps on growing beyond region occupied by the stars and gas [30].

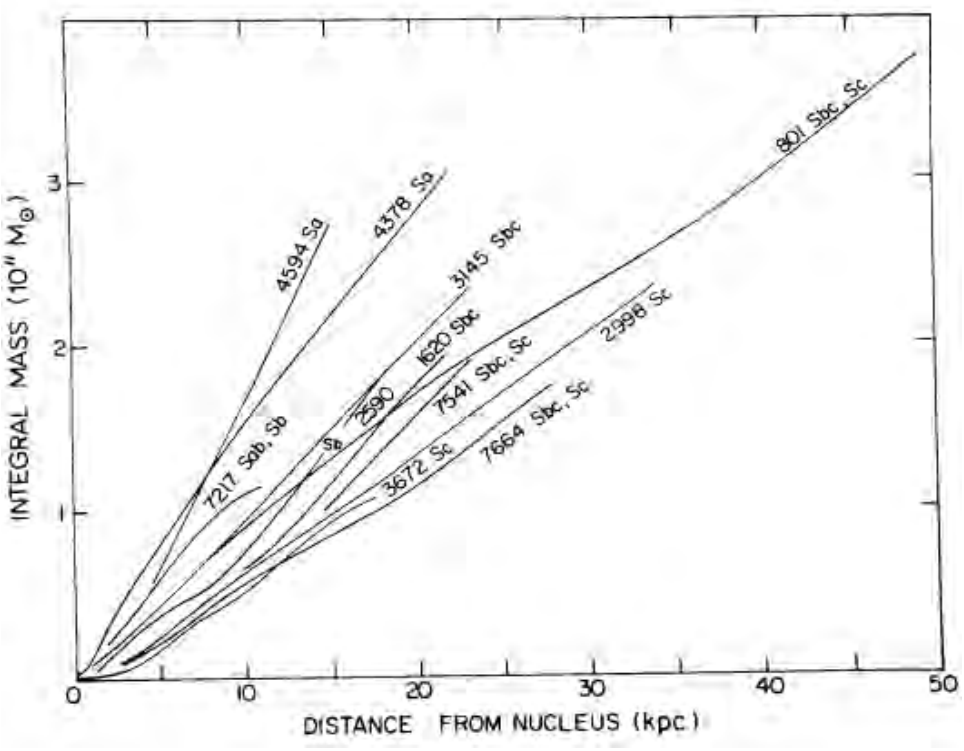


Figure 6: Mass of 11 galaxies are shown with the diatances from the galactic centre. The linear increase of mass with the distance results from the observed flatness of the rotation curves [25].

4 Newtonian Version of Keplerian Dynamics

The argument that the mass of the galaxy increases at the large distances from the galactic center for the flat portion of the rotation curves is very simple to understand and is described below. If the cross-radial velocity of a star of mass m is v, the centripetal force for its rotation around the galactic centre is provided by the gravitational force between the galaxy and the star.

$$\frac{mv^2}{R} = \frac{GMm}{R} ,$$

$$v^2 = \frac{GM}{R} , \tag{8}$$

where M is the mass of the galaxy upto the star of mass m distant at a radius R from galactic centre. Eq. (8) suggest that the orbital velocity of the stars decreases as their distances from the galactic centre increase. This is in general referred to as the "Keplerian" behavior. This is in contradiction with the results obtained by Rubin [24, 25, 26] since observational indications suggests that for the flat portion of rotation curve v is independent of distances of the stars from the galactic centre (shown in Fig. 7). From Eq. (8) we obtain $M = \frac{v^2}{G}R$. As G is Newton's constant of Gravitation, $M \propto R$ for the flat portion of rotation curve implying the fact that mass of the galaxy increases as radius increases. Needless to say that this matter is certainly not visible matter as the visible matter density was already been estimated by Schmidt for andromeda galaxy in Ref. [22].

5 Dark Matter

To explain the observational results obtained by Rubin [24, 25, 26] with theoretical results obtained from Newtonian gravity, it is evident from Eq. (8) that $M \propto r$ i.e., the mass of the galaxy should increase with galactic radius which in turn is in contradictory with the visible matter present in the galaxy. Hence other than luminous matter there should be present some non-luminous matter in the galaxy. That is known as dark matter which is optically invisible as well as does not undergo any known interactions present in standard model of particle physics.

In a work of 1996 [44], Julio F. Navarro, Carlos S. Frenk and Simon D. M. White considered a spherically symmetric density profile for dark matter halos given by

$$\rho(r) \propto \frac{1}{r\left(1+\frac{r}{r_s}\right)^2} \tag{9}$$

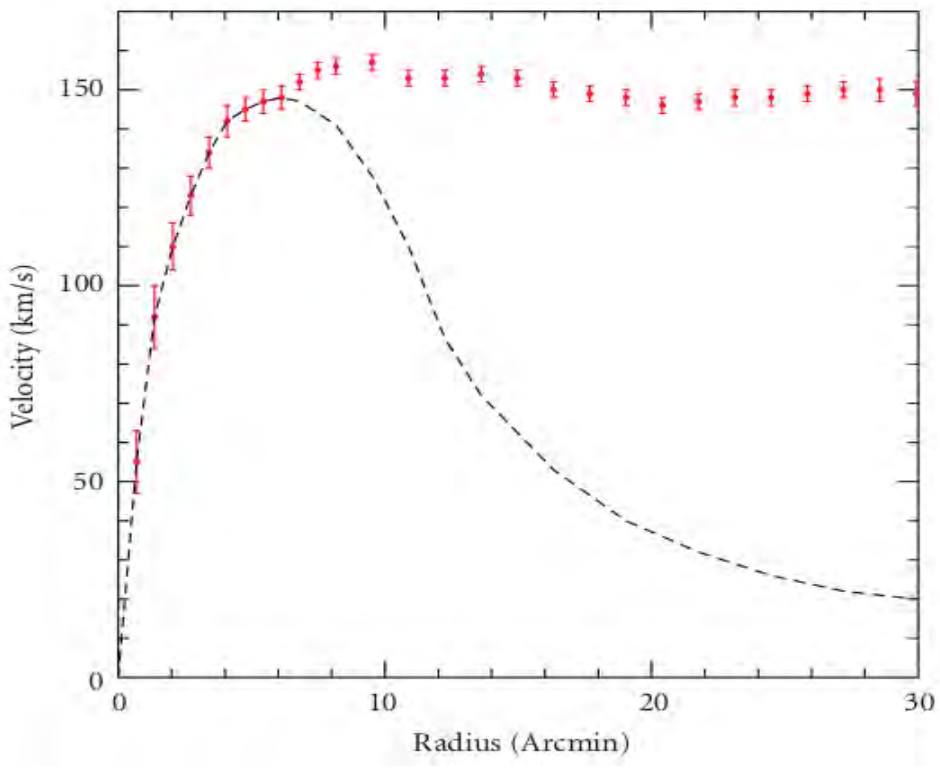


Figure 7: NGC 3198 galaxy rotation curves data points are shown red [42] and the rotation curves prediction from Keplerian dynamics are shown in dashed curves. The figure is adopted from the Ref. [43].

which they proposed in the context of X-ray galaxy clusters earlier in 1995 [45]. This is popularly known as NFW density profile of dark matter halos. They performed N-body simulations for looking into the structure of dark halos. They argued that the disk galaxies are embedded in the cold dark matter halo and structure of dark matter halos plays a significant role in context of rotation curves spiral galaxies. Their results are shown in the Fig. 8 where $\mathrm{r_{opt}}$ is the optical radius of the galaxies and the choice of parameters is also shown in the same figure. It is clear from this figure the that presence of the dark matter halo in the disk galaxies can explain the rotation curves very well. Massimo Persic, Paolo Salucci and Fulvio

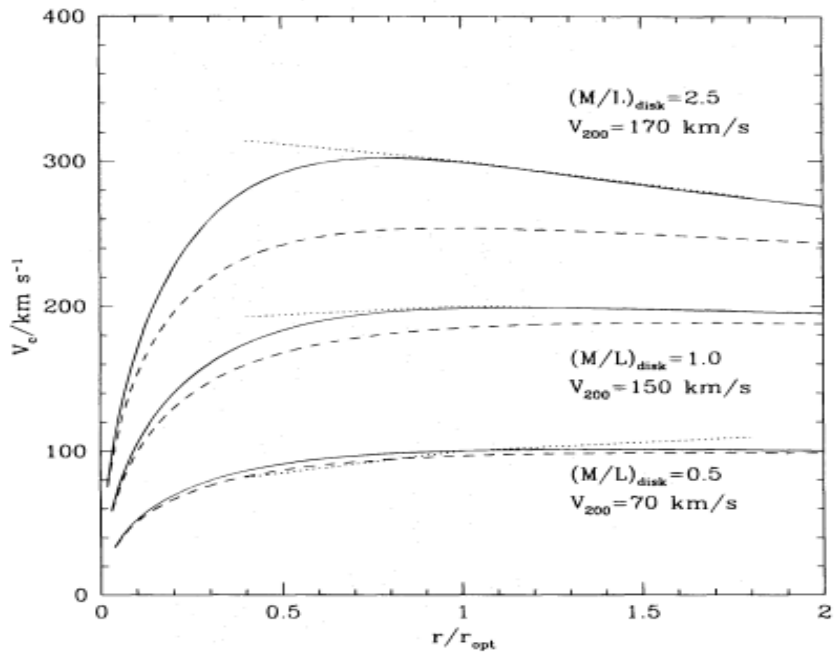


Figure 8: Rotation curves for spiral galaxies with cold dark matter halos (solid lines). The dashed lines corresponds to contribution from the dark matter halos [46].

Stel claimed in a work [46] from the investigations of a sample of 1100 optical and radio rotation curves that for high luminosity galaxies, the discrepancy between the observed rotations curves with those that are theoretically predicted from

luminous matter can be resolved by the presence of dark matter in the galaxies. For low luminosity galaxies the dark matter is the only dominant component. Their results are shown in the Fig. 9 where $\mathrm{R_{opt}}$ is the optical radius of the galaxies.

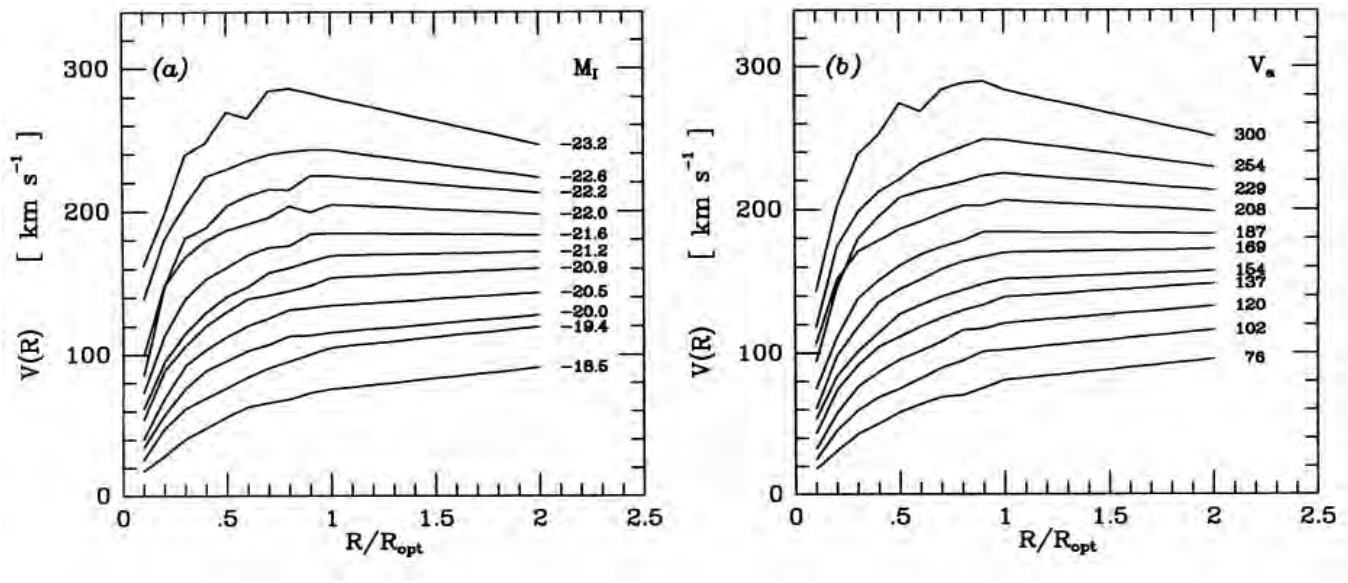


Figure 9: Universal rotation curves of spiral galaxies. $\mathrm{R_{opt}}$ is the optical radius of the galaxies [46].

5.1 Experimental Search for Dark Matter Particle

As the astrophysical observations of rotation curve provides the existence of dark matter in galaxies, it is therefore obvious that we would be able to detect it in our own galaxy Milky way. As the density is uniform throughout the galaxy, in principle the dark matter particles in the form of WIMPs (Weakly Interacting Matter Particles) can be detected in the laboratories on Earth. Here we will describe few experiments that are engaged in direct detection of WIMPs.

Direct detection methodology: The underlying principle behind the direct detection of WIMPs is very simple. The detector are so designed that the WIMPs collide with the target nuclei placed in the detector and the recoils of the target nuclei are observed to estimate the mass of WIMPs (Fig. 10). For example, WIMP mass of 10-1000 GeV/c^2 (c is the speed of light in free space) will produce the recoil energy of the target nucleus around 1-100 KeV [47].

Throughout the world, many research collaborations such as DAMA/LIBRA [48], CRESST [49], XENON [50, 51, 52, 53], CDMS [54] and CoGeNT [55] are engaged in direct detection of WIMPs. The direct detection techniques of WIMPs are discussed in details in recent work by Schumann [56]. It is in principle also possible to detect the particles of dark matter indirectly in the laboratories (Fig. 11).

Indirect detection methodology: There are many theoretical model for dark matter in the extended sector of Standard Model of particle physics. These theoretical models explain the self interactions and scattering of dark matter particles or decay of dark matter particles into Standard Model particles. These processes produce a particle flux that can be measured experimentally. The products of the possible annihilation channels may further decay into photons and neutrinos which are basically detected in this case of indirect detection of dark matter.

To date no significant signals from dark matter annihilations has been observed and upper limits are derived by the collaborations MAGIC [58, 59], HESS [60, 61] and VERITAS [62, 63]. There are also satellite based instruments such as Fermi-LAT [64, 65, 66, 67, 68] for detecting low-energy gamma rays of energy 20 MeV- 300 GeV. For the detection of neutrinos as a signal from dark matter annihilations, large neutrino detectors such as Ice Cube [69], ANTARES [70] or Super Kamiokande [71] are in function. However to this date, no signals as the evidence of dark matter from these detectors has yet been observed and thus these experiment results only in putting constraints on the annihilation cross-sections. For a detailed discussions on indirect detection of dark matter consider the Tasi Lectures by Tracy R. Slatyer [72].

6 Modified Newtonian Dynamics

Milgrom in 1983 [73] put forward a proposal that the observed flatness of rotation curves may be explained without the presence of non-luminous dark matter in the spiral galaxies. Rather he proposed a suggestion to modify Newton's second law which in case of small accelerations would result in a stellar motion independent of its radius. Modified version of Newton's second law as proposed by Milgrom is given by,

$$\Rightarrow F \quad = \quad m\mu\left(\frac{a}{a_0}\right) \Rightarrow a\,, \tag{10}$$

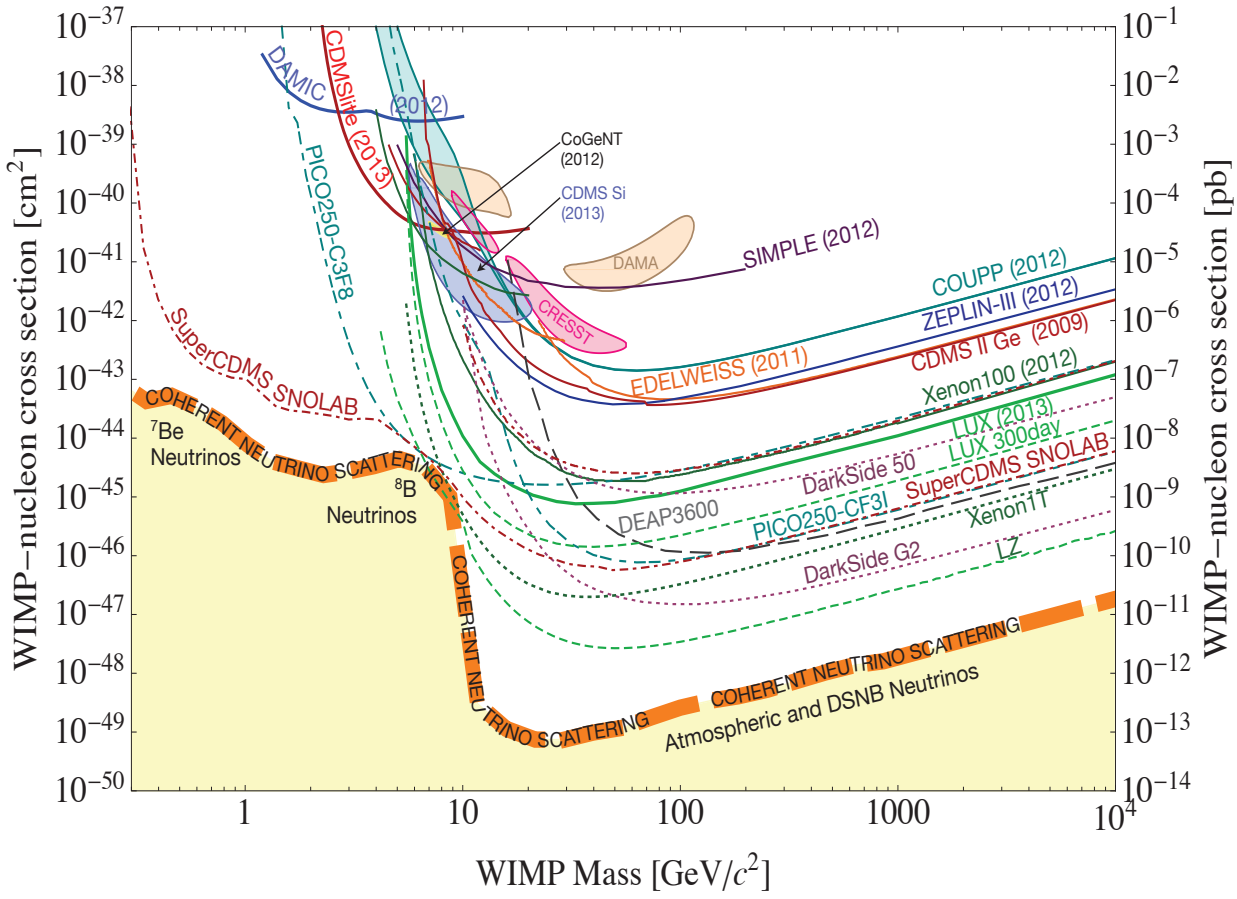


Figure 10: The present constraints from all the experimental search for spin-independent WIMP nucleon scattering cross section is shown here against the WIMP mass. The parameter space above the lines is excluded by the experiments at 90% confidence level. Figure courtesy: J. Cooley [57]

where $a_0 \simeq 2 \times 10^{-8}$ cm.sec^{-2} is a constant acceleration determined by Milgrom himself and $\mu\left(\frac{a}{a_0}\right)$ is a function given by,

$$\mu\left(\frac{a}{a_0}\right) \simeq 1\,, \quad \text{for a} >> \text{a}_0 \quad \text{and} \quad \mu\left(\frac{a}{a_0}\right) \simeq \frac{a}{a_0}\,, \quad \text{for a} << \text{a}_0\,. \tag{11}$$

For the accelerations a we encounter in our daily life are greater than a_0 and hence Eq. (10) reduces to Newton's second law motion. For the celestial objects far away from the centre of galaxies a is very small and in this situation a_0 becomes important.

$$F = \frac{GMm}{r^2} = m\mu\left(\frac{a}{a_0}\right)a\,, \tag{12}$$

where G is the Newton's universal gravitational constant, m and M are the masses of the star at a distance e from the centre of galaxy and the galaxy respectively. For $a << a_0$ from Eq.(12) we obtain

$$\frac{GM}{r^2} = \frac{a^2}{a_0}\,, \tag{13}$$

where a is given by the centripeatal acceleration of rotation and is given by,

$$a = \frac{v^2}{r}\,, \tag{14}$$

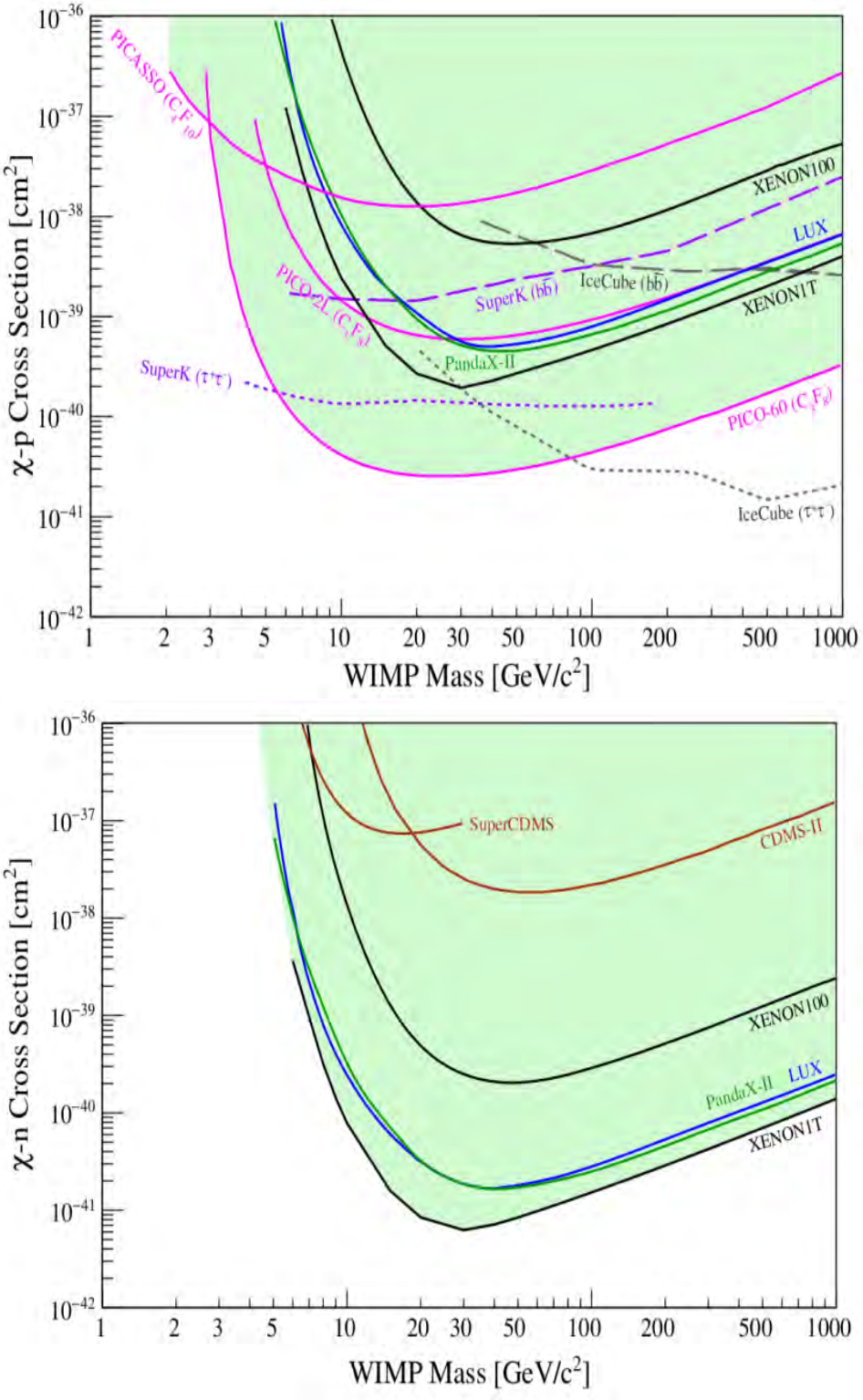


Figure 11: The present constraints from all the experimental search for spin-dependent WIMP proton (Top) and WIMP neutron (Bottom) scattering cross section is shown here against the WIMP mass. Figure is adopted from [56].

where v is the rotational velocity. From Eqs. (13) and (14), we obtain

$$\begin{aligned} \frac{GM}{r^2} &= \frac{v^4}{a_0 r^2}\,, \\ v &= (GMa_0)^{1/4}\,. \end{aligned} \tag{15}$$

Eq. (15) shows that the rotational velocity is independent of the distance from the galactic centre for large distances. Thus MOND successfully explains the observed flatness of rotation curves.

7 Metric Skew Tensor Gravity

The other alternative to explain galaxy rotation curves without postulating the presence of dark matter was suggested by J. W. Moffat [74] in 2004. He put forward a simple theory of Einstein's general relativity coupled to a skew symmetric tensor of rank three known as metric-skew-tensor-gravity (MSTG) theory. Here in this MSTG model, the radial acceleration experienced by a particle in a static and spherically symmetric gravitational field is given by

$$a(r) = -\frac{G_\infty M}{r^2} + \sigma \frac{\exp(-r/r_0)}{r^2}\left(1+\frac{r}{r_0}\right), \tag{16}$$

where G_∞ is the effective gravitational constant at infinity and is given by

$$G_\infty = G_0\left(1+\sqrt{\frac{M_0}{M}}\right). \tag{17}$$

Here G_0 is the Newton's gravitational constant (bare) and M_0 is the coupling parameter. σ is the coupling constant for repulsive Yukawa force and is given by

$$\sigma = G_0\sqrt{M_0 M}\,. \tag{18}$$

Hence Eq. (16) reduces to

$$a(r) = -\frac{G(r)M}{r^2}, \tag{19}$$

where $G(r)$, the running gravitational constant, is given by,

$$G(r) = -G_0\left(1+\sqrt{\frac{M_0}{M}}\left[1-\exp\left(-\frac{r}{r_0}\right)\left(1+\frac{r}{r_0}\right)\right]\right). \tag{20}$$

Therefore the rotational velocity of a star in the galaxy, derived from the acceleration $a(r)$, is given by

$$\begin{aligned} v(r) &= \sqrt{a(r)r}\,, \\ &= \sqrt{\frac{G_0 M}{r}}\left(1+\sqrt{\frac{M_0}{M}}\left[1-\exp\left(-\frac{r}{r_0}\right)\left(1+\frac{r}{r_0}\right)\right]\right)^{1/2}, \end{aligned} \tag{21}$$

The parameters M_0 and r_0 are defined in such a way so as to obtain

$$a_0 = \frac{G_0 M_0}{r_0^2}, \tag{22}$$

with an assumption that $a_0 = cH_0$, where H_0 is the Hubble constant given by $H_0 = 100h$ km.sec^{-1}.Mpc^{-1} and $h = 0.71 \pm 0.07$. This leads to $a_0 = 6.90 \times 10^{-8}$ cm.sec^{-2}. For low surface brightness (LSB) and high surface brightness (HSB) galaxy data, good fits are obtained with the parameters $M_0 = 9.6 \times 10^{11} M_\odot$ and $r_0 = 13.92$ kpc. The authors in a work [75] argued to have remarkably good fits to rotation curves from MSTG model for 102 LSB and HSB galaxies including one elliptical galaxy (NGC 3379). Thus MTMG, being a consistent and viable relativistic gravitational theory proposed by Moffat [74, 76], can successfully fits the rotation curves of galaxies and the X-ray data of galaxy cluster [77] without having any necessity for dark matter.

8 Conclusion

Observations of a half century on galaxy reveals that the total mass of the luminous matter does not suffice to provide a satisfactory explanation for the rotation curves. Hence existence of non-luminous matter dubbed as dark matter is invoked to account for the galaxy rotation curves. This is worth mentioning here that the presence of dark matter is not only inferred from the galaxy rotation curves but also supported by independent observations of galaxy cluster formation [16, 78, 79], bullet cluster[80, 81] and gravitational lensing phenomenon [82, 83, 84, 85, 86]. Since till date there is no direct [48, 49, 50, 51, 52, 53, 54, 55] as well as indirect evidence [58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71] in favour of existance of dark matter particles in laboratories or in natural events, other possibilities are also explored by several authors. MOND is efficient in explaining galaxy rotaion curve. However there exist some serious issues that MOND fails to explain. In case of galaxy clusters, MOND cannot explain the requirement of unseen matter as well as the density and temperature profiles [87]. Moreover extreme low acceleration experiments suggest no departure from Newton's second law and hence MOND necessarily needs to be reduced to Newton's second law in laboratory experiments [88, 89]. Since MOND proposed by Milgrom [73] is a non-relativistic theory, it is unable to explain gravitational lensing phenomenon. Bekenstein in 2004 made an attempt to develop a relativistic theory of Milgrom's MOND in his work [90] to incorporate the gravitational lensing

in MOND. Several authors used galaxy-galaxy lensing data to put empirical constraints on the MONDian light deflection [91, 92, 93]. MSTG, being a viable model of gravitational theory, is also accepted as a fair alternative explanation to the galaxy rotation curves and has been studied in extensively in the works [74, 75, 76, 77, 94]. Despite being successful alternatives dark matter is still preferred over the MOND or MSTG to date in the context of cosmological missing mass problem in the Universe. However in the galactic length scale which is much less than the cosmological length scale, things may differ and MOND or MSTG may be a viable explanation. Hence being unbiased on the situtation we should be open to all the explanations and it is not the time to bring about any conclusion.

A Tully-Fisher Relation

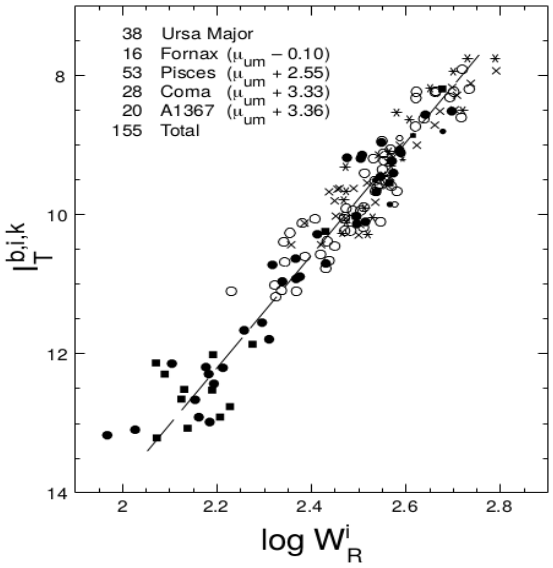


Figure 12: Apparent magnitude in near-infrared I band plotted against the maximum rotational velocity for five cluster. [95].

Most abundant element in galaxies or interstellar medium is the hydrogen. Due to the symmetry of hydrogen molecule (H_2), it does not have any permanent dipole moment. Hence hydrogen molecule does not emit any radiation in radio frequency range. However in the low density region of interstellar medium atomic hydrogen is abundant and spin interaction of proton and electron results in hyperfine structure where two energy levels are separated by energy difference corresponding to a wavelength of 21 cm. Hence whenever the spin state flips from a parallel to antiparallel state a photon of $\lambda = 21$ cm is emitted. Doppler effect due to the motion of the galaxies broadens this 21 cm line. Hence the line profile of a galaxy informs us about the motions in the galaxy. The observed 21 cm profile width ΔV corrected in accordance with the inclination is a measure of maximum rotational velocity v_m in the disk galaxies near its edge [96] and also a tool for mass measurement of the galaxies [97, 98, 99]. In 1977, R. Brent Tully and J. Richard Fisher developed the a tool for better understanding of the galactic structure on the basis of the global hydrogen profile widths versus absolute magnitude or diameters of galaxies [100]. From the observations on the Virgo cluster, they found an empirical correlation between the luminosity and the global profile width as

$$L \propto \Delta V^{2.5\pm0.3}\,. \tag{23}$$

The mass of the galaxy is related to the global profile width ΔV and the radius R of the galaxy by

$$M \propto \Delta V^2 R\,. \tag{24}$$

From Eqs. (23) and (24), it is clear that the ratio of mass of the galaxy to its luminosity (M/L) is almost constant which was earlier obtained by Morton Roberts [98]. A strong empirical constrain on Tully-Fisher relation was found from 11 spiral galaxies in Virgo and 18 in the Ursa Major cluster [96]. Here the authors put forward the correlation $L_H \propto v_m^4$ in the infrared region and concluded that M/L_H is a constant for late type galaxies but mass to blue light ration (M/L_B) varies as $L_H^{1/4}$ where L_H is the luminosity for 21 cm line. In Fig. 12, the relation between apparent magnitude of 155 galaxies in the near infrared I band and width of the 21 cm line is shown [95]. The parameter $\mathrm{W_R}$ in Fig 12 is defined [101] in such a way that it accounts for approximately twice the maximum rotational velocity of a disk galaxy. The five galaxy clusters and number of galaxies provided by them is also shown in the aforesaid figure. From the study of slopes of B, R, K' and I bands it was found that in the infrared region the luminosity and the maximum velocity of rotation of galaxies takes the form given by

$$L \propto v_{\mathrm{m}}^{3.4\pm0.1}\,. \tag{25}$$

where L and v_{m} are the luminosity and the maximum velocity of rotation respectively [95].

References

[1] C. Messier, "A Table of the Places of the Comet of 1764 Discovered at the Observatory of the Marine at Paris, the 3d of January, about 8 O'Clock in the Evening, in the Constellation of the Dragon, Concluded from Its Situation Observed with Regard to the Stars: By Monsieur Charles Messier, Astronomer at the Depot of the Plans of the Marine of France, at Paris," *Philosophical Transactions of the Royal Society of London Series I*, vol. 54, p. 68, Jan. 1764.

[2] ——, "Catalogue des Nébuleuses et des Amas d'Étoiles (Catalog of Nebulae and Star Clusters)," Connoissance des Temps ou des Mouvements Célestes, pp. 227–267, Jan. 1781.

[3] I. Hafez, "Abd al-Rahman al-Sufi and his book of the fixed stars: a journey of re-discovery," Ph.D. dissertation, James Cook University, Oct. 2010.

[4] W. G. Hoyt, *Vesto Melvin Slipher, 1875—1969, A Biographical Memoir*. National Academy of Sciences, 1980.

[5] V. M. Slipher, "The radial velocity of the Andromeda Nebula," *Lowell Observatory Bulletin*, vol. 1, pp. 56–57, Jan. 1913.

[6] ——, "Spectrographic Observations of Nebulae," *Popular Astronomy*, vol. 23, pp. 21–24, Jan. 1915.

[7] H. D. Curtis, "Modern Theories of the Spiral Nebulae," *The Journal of the Royal Astronomical Society of Canada*, vol. 14, p. 317, Oct. 1920.

[8] H. Shapley and H. D. Curtis, "The Scale of the Universe," *Bulletin of the National Research Council*, vol. 2, no. 11, pp. 171–217, May 1921.

[9] V. Trimble, "The 1920 Shapley-Curtis Discussion: Background, Issues, and Aftermath," *Publications of the Astronomical Society of the Pacific*, vol. 107, p. 1133, Dec. 1995.

[10] M. A. Hoskin, "The 'Great Debate': What Really Happened," *Journal for the History of Astronomy*, vol. 7, p. 169, Jan. 1976.

[11] R. W. Smith, *The expanding universe: astronomy's 'great debate' : 1900-1931*. Cambridge: Cambridge Univ. Press, 1982. [Online]. Available: https://cds.cern.ch/record/101439

[12] A. S. Sharov and I. D. Novikov, *Edwin Hubble, The Discoverer of the Big Bang Universe*. Cambridge University Press, 1993.

[13] M. Bartusiak, *The Day We Found the Universe*. Vintage Books, 2010. [Online]. Available: https://books.google.co.in/books?id=7XojzXh4_KEC

[14] E. P. Hubble, "A spiral nebula as a stellar system, Messier 31." *The Astrophysical Journal*, vol. 69, pp. 103–158, Mar. 1929.

[15] V. Slipher, "The detection of nebular rotation," *Lowell Observ. Bull*, vol. 2(62):66.

[16] G. Bertone and D. Hooper, "History of dark matter," *Rev. Mod. Phys.*, vol. 90, no. 4, p. 045002, 2018.

[17] E. P. Hubble, "Extragalactic nebulae." *The Astrophysical Journal*, vol. 64, pp. 321–369, Dec. 1926.

[18] J. H. Oort, "The force exerted by the stellar system in the direction perpendicular to the galactic plane and some related problems," *Bulletin of the Astronomical Institutes of the Netherlands*, vol. 6, p. 249, Aug. 1932.

[19] F. Zwicky, "On the Masses of Nebulae and of Clusters of Nebulae," *The Astrophysical Journal*, vol. 86, p. 217, Oct. 1937.

[20] H. W. Babcock, "The rotation of the Andromeda Nebula," *Lick Observatory Bulletin*, vol. 498, pp. 41–51, Jan. 1939.

[21] H. C. van de Hulst, E. Raimond, and H. van Woerden, "Rotation and density distribution of the Andromeda nebula derived from observations of the 21-cm line," *Bulletin of the Astronomical Institutes of the Netherlands*, vol. 14, p. 1, Nov. 1957.

[22] M. Schmidt, "The distribution of mass in M 31," *Bulletin of the Astronomical Institutes of the Netherlands*, vol. 14, p. 17, Nov. 1957.

[23] R. H. Sanders, *The Dark Matter Problem A Historical Perspective*. Cambridge University Press, 2010.

[24] V. C. Rubin and J. Ford, W. Kent, "Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions," *The Astrophysical Journal*, vol. 159, p. 379, Feb. 1970.

[25] V. C. Rubin, J. Ford, W. K., and N. Thonnard, "Extended rotation curves of high-luminosity spiral galaxies. IV. Systematic dynamical properties, Sa -> Sc." *The Astrophysical Journal Letters*, vol. 225, pp. L107–L111, Nov. 1978.

[26] ——, "Rotational properties of 21 SC galaxies with a large range of luminosities and radii, from NGC 4605 (R=4kpc) to UGC 2885 (R=122kpc)." *The Astrophysical Journal*, vol. 238, pp. 471–487, Jun. 1980.

[27] D. H. Rogstad and G. S. Shostak, "Gross Properties of Five Scd Galaxies as Determined from 21-CENTIMETER Observations," *The Astrophysical Journal*, vol. 176, p. 315, Sep. 1972.

[28] R. N. Whitehurst and M. S. Roberts, "High-Velocity Neutral Hydrogen in the Central Region of the Andromeda Galaxy," *The Astrophysical Journal*, vol. 175, p. 347, Jul. 1972.

[29] M. S. Roberts and A. H. Rots, "Comparison of Rotation Curves of Different Galaxy Types," *Astron. & Astrophys*, vol. 26, pp. 483–485, Aug. 1973.

[30] A. Bosma, "The distribution and kinematics of neutral hydrogen in spiral galaxies of various morphological types," Ph.D. dissertation, 1978, date submitted:2008 Rights: University of Groningen.

[31] M. Schwarzschild, "Mass distribution and mass-luminosity ratio in galaxies," *The Astronomical Journal*, vol. 59, p. 273, Sep. 1954.

[32] M. Schmidt, "A model of the distribution of mass in the Galactic System," *Bulletin of the Astronomical Institutes of the Netherlands*, vol. 13, p. 15, Jun. 1956.

[33] J. H. Oort, "Problems of Galactic Structure." *The Astrophysical Journal*, vol. 116, p. 233, Sep. 1952.

[34] N. U. Mayall and L. H. Aller, "The Rotation of the Spiral Nebula Messier 33." *The Astrophysical Journal*, vol. 95, p. 5, Jan. 1942.

[35] A. B. Wyse and N. U. Mayall, "Distribution of Mass in the Spiral Nebulae Messier 31 and Messier 33." *The Astrophysical Journal*, vol. 95, p. 24, Jan. 1942.

[36] L. Perek, "A model of the galaxy," *Contr. Astr. Inst. Masaryk Univ.*, vol. 1, no. 6, 1948.

[37] P. P. Parenago, *Astronom. J. USSR.*, vol. 27, no. 329, 1950.

[38] ——, *Astronom. J. USSR.*, vol. 29, no. 245, 1952.

[39] G. G. Kuzmin, *Pub. Astr. Obs. Tartu*, vol. 32, no. 211, 1952.

[40] J. C. Brandt, "On the Distribution of Mass in Galaxies. II. a Discussion of Tne Mass of the Galaxy." *The Astrophysical Journal*, vol. 131, p. 553, May 1960.

[41] E. Holmberg, *Magnitudes, Colors, Surface Brightness, Intensity Distributions Absolute Luminosities, and Diameters of Galaxies*, 1975, p. 123.

[42] K. G. Begeman, "HI rotation curves of spiral galaxies. I. NGC 3198." *Astron. & Astrophys*, vol. 223, pp. 47–60, Oct. 1989.

[43] K. Garrett and G. Dūda, "Dark matter: A primer," *Advances in Astronomy, Hindawi Publishing Corporation*, vol. 2011, no. 968283, p. 22 pages, Sep. 2010, doi:10.1155/2011/968283.

[44] J. F. Navarro, C. S. Frenk, and S. D. M. White, "The Structure of Cold Dark Matter Halos," *The Astrophysical Journal*, vol. 462, p. 563, May 1996.

[45] J. F. Navarro, C. S. Frenk, and S. D. M. White, "Simulations of X-ray clusters," *Monthly Notices of the Royal Astronomical Society*, vol. 275, no. 3, pp. 720–740, 08 1995. [Online]. Available: https://doi.org/10.1093/mnras/275.3.720

[46] M. Persic, P. Salucci, and F. Stel, "The universal rotation curve of spiral galaxies — I. The dark matter connection," *Monthly Notices of the Royal Astronomical Society*, vol. 281, no. 1, pp. 27–47, Jul. 1996.

[47] J. D. Lewin and P. F. Smith, "Review of mathematics, numerical factors, and corrections for dark matter experiments based on elastic nuclear recoil," *Astropart. Phys.*, vol. 6, pp. 87–112, 1996.

[48] R. Bernabei *et al.*, "New results from DAMA/LIBRA," *Eur. Phys. J.*, vol. C67, pp. 39–49, 2010.

[49] G. Angloher, M. Bauer, I. Bavykina, A. Bento, C. Bucci, C. Ciemniak, G. Deuter, F. von Feilitzsch, D. Hauff, P. Huff, C. Isaila, J. Jochum, M. Kiefer, M. Kimmerle, J.-C. Lanfranchi, F. Petricca, S. Pfister, W. Potzel, F. Pröbst, F. Reindl, S. Roth, K. Rottler, C. Sailer, K. Schäffner, J. Schmaler, S. Scholl, W. Seidel, M. v. Sivers, L. Stodolsky, C. Strandhagen, R. Strauß, A. Tanzke, I. Usherov, S. Wawoczny, M. Willers, and A. Zöller, "Results from 730 kg days of the CRESST-II Dark Matter search," *European Physical Journal C*, vol. 72, p. 1971, Apr. 2012.

[50] J. Angle *et al.*, "First Results from the XENON10 Dark Matter Experiment at the Gran Sasso National Laboratory," *Phys. Rev. Lett.*, vol. 100, p. 021303, 2008.

[51] E. Aprile, J. Angle, F. Arneodo, L. Baudis, A. Bernstein, A. Bolozdynya, P. Brusov, L. C. C. Coelho, C. E. Dahl, L. DeViveiros, A. D. Ferella, L. M. P. Fernandes, S. Fiorucci, R. J. Gaitskell, K. L. Giboni, R. Gomez, R. Hasty, L. Kastens, J. Kwong, J. A. M. Lopes, N. Madden, A. Manalaysay, A. Manzur, D. N. McKinsey, M. E. Monzani, K. Ni, U. Oberlack, J. Orboeck, D. Orlandi, G. Plante, R. Santorelli, J. M. F. dos Santos, P. Shagin, T. Shutt, P. Sorensen, S. Schulte, E. Tatananni, C. Winant, and M. Yamashita, "Design and performance of the XENON10 dark matter experiment," *Astroparticle Physics*, vol. 34, pp. 679–698, Apr. 2011.

[52] J. Angle *et al.*, "A search for light dark matter in XENON10 data," *Phys. Rev. Lett.*, vol. 107, p. 051301, 2011, [Erratum: Phys. Rev. Lett.110,249901(2013)].

[53] E. Aprile *et al.*, "Search for Electronic Recoil Event Rate Modulation with 4 Years of XENON100 Data," *Phys. Rev. Lett.*, vol. 118, no. 10, p. 101101, 2017.

[54] R. Agnese *et al.*, "Silicon Detector Dark Matter Results from the Final Exposure of CDMS II," *Phys. Rev. Lett.*, vol. 111, no. 25, p. 251301, 2013.

[55] C. E. Aalseth *et al.*, "Search for An Annual Modulation in Three Years of CoGeNT Dark Matter Detector Data," 2014.

[56] M. Schumann, "Direct Detection of WIMP Dark Matter: Concepts and Status," *J. Phys. G*, vol. 46, no. 10, p. 103003, 2019.

[57] J. Cooley, "Overview of Non-Liquid Noble Direct Detection Dark Matter Experiments," *Phys. Dark Univ.*, vol. 4, pp. 92–97, 2014.

[58] J. Aleksić *et al.*, "Optimized dark matter searches in deep observations of Segue 1 with MAGIC," *JCAP*, vol. 1402, p. 008, 2014.

[59] M. L. Ahnen *et al.*, "Limits to dark matter annihilation cross-section from a combined analysis of MAGIC and Fermi-LAT observations of dwarf satellite galaxies," *JCAP*, vol. 1602, no. 02, p. 039, 2016.

[60] A. Abramowski *et al.*, "Search for dark matter annihilation signatures in H.E.S.S. observations of Dwarf Spheroidal Galaxies," *Phys. Rev.*, vol. D90, p. 112012, 2014.

[61] H. Abdallah *et al.*, "Search for dark matter annihilations towards the inner Galactic halo from 10 years of observations with H.E.S.S," *Phys. Rev. Lett.*, vol. 117, no. 11, p. 111301, 2016.

[62] T. Arlen, T. Aune, M. Beilicke, W. Benbow, A. Bouvier, J. H. Buckley, V. Bugaev, K. Byrum, A. Cannon, A. Cesarini, L. Ciupik, E. Collins-Hughes, M. P. Connolly, W. Cui, R. Dickherber, J. Dumm, A. Falcone, S. Federici, Q. Feng, J. P. Finley, G. Finnegan, L. Fortson, A. Furniss, N. Galante, D. Gall, S. Godambe, S. Griffin, J. Grube, G. Gyuk, J. Holder, H. Huan, G. Hughes, T. B. Humensky, A. Imran, P. Kaaret, N. Karlsson, M. Kertzman, Y. Khassen, D. Kieda, H. Krawczynski, F. Krennrich, K. Lee, A. S. Madhavan, G. Maier, P. Majumdar, S. McArthur, A. McCann, P. Moriarty, R. Mukherjee, T. Nelson, A. O'Faoláin de Bhróithe, R. A. Ong, M. Orr, A. N. Otte, N. Park, J. S. Perkins, M. Pohl, H. Prokoph, J. Quinn, K. Ragan, L. C. Reyes, P. T. Reynolds, E. Roache, J. Ruppel, D. B. Saxon, M. Schroedter, G. H. Sembroski, C. Skole, A. W. Smith, I. Telezhinsky, G. Tešić, M. Theiling, S. Thibadeau, K. Tsurusaki, A. Varlotta, M. Vivier, S. P. Wakely, J. E. Ward, A. Weinstein, R. Welsing, D. A. Williams, B. Zitzer, C. Pfrommer, and A. Pinzke, "Constraints on Cosmic Rays, Magnetic Fields, and Dark Matter from Gamma-Ray Observations of the Coma Cluster of Galaxies with VERITAS and Fermi," *The Astrophysical Journal*, vol. 757, p. 123, Oct. 2012.

[63] S. Archambault *et al.*, "Dark Matter Constraints from a Joint Analysis of Dwarf Spheroidal Galaxy Observations with VERITAS," *Phys. Rev.*, vol. D95, no. 8, p. 082001, 2017.

[64] M. Ackermann *et al.*, "Search for Gamma-ray Spectral Lines with the Fermi Large Area Telescope and Dark Matter Implications," *Phys. Rev.*, vol. D88, p. 082002, 2013.

[65] ——, "Limits on Dark Matter Annihilation Signals from the Fermi LAT 4-year Measurement of the Isotropic Gamma-Ray Background," *JCAP*, vol. 1509, no. 09, p. 008, 2015.

[66] ——, "Updated search for spectral lines from Galactic dark matter interactions with pass 8 data from the Fermi Large Area Telescope," *Phys. Rev.*, vol. D91, no. 12, p. 122002, 2015.

[67] ——, "Searching for Dark Matter Annihilation from Milky Way Dwarf Spheroidal Galaxies with Six Years of Fermi Large Area Telescope Data," *Phys. Rev. Lett.*, vol. 115, no. 23, p. 231301, 2015.

[68] A. Albert *et al.*, "Searching for Dark Matter Annihilation in Recently Discovered Milky Way Satellites with Fermi-LAT," *Astrophys. J.*, vol. 834, no. 2, p. 110, 2017.

[69] M. G. Aartsen *et al.*, "Search for annihilating dark matter in the Sun with 3 years of IceCube data," *Eur. Phys. J.*, vol. C77, no. 3, p. 146, 2017.

[70] S. Adrian-Martinez *et al.*, "Limits on Dark Matter Annihilation in the Sun using the ANTARES Neutrino Telescope," *Phys. Lett.*, vol. B759, pp. 69–74, 2016.

[71] K. Choi *et al.*, "Search for neutrinos from annihilation of captured low-mass dark matter particles in the Sun by Super-Kamiokande," *Phys. Rev. Lett.*, vol. 114, no. 14, p. 141301, 2015.

[72] T. R. Slatyer, "Indirect Detection of Dark Matter," in *Theoretical Advanced Study Institute in Elementary Particle Physics: Anticipating the Next Discoveries in Particle Physics*, 2018, pp. 297–353.

[73] M. Milgrom, "A modification of the Newtonian dynamics as a possible alternative to the hidden mass hypothesis." *The Astrophysical Journal*, vol. 270, pp. 365–370, Jul. 1983.

[74] J. W. Moffat, "Gravitational theory, galaxy rotation curves and cosmology without dark matter," *Journal of Cosmology and Astroparticle Physics*, vol. 2005, no. 5, p. 003, May 2005.

[75] J. R. Brownstein and J. W. Moffat, "Galaxy rotation curves without nonbaryonic dark matter," *The Astrophysical Journal*, vol. 636, no. 2, pp. 721–741, jan 2006. [Online]. Available: https://doi.org/10.1086%2F498208

[76] J. Moffat, "Scalar-tensor-vector gravity theory," *JCAP*, vol. 03, p. 004, 2006.

[77] J. Brownstein and J. Moffat, "Galaxy cluster masses without non-baryonic dark matter," *Mon. Not. Roy. Astron. Soc.*, vol. 367, pp. 527–540, 2006.

[78] E. Hubble and M. L. Humason, "The Velocity-Distance Relation among Extra-Galactic Nebulae," *The Astrophysical Journal*, vol. 74, p. 43, Jul. 1931.

[79] F. Zwicky, "Die Rotverschiebung von extragalaktischen Nebeln," *Helvetica Physica Acta*, vol. 6, p. 110, 1933.

[80] M. Markevitch, A. H. Gonzalez, D. Clowe, A. Vikhlinin, L. David, W. Forman, C. Jones, S. Murray, and W. Tucker, "Direct constraints on the dark matter self-interaction cross-section from the merging galaxy cluster 1E0657-56," *Astrophys. J.*, vol. 606, pp. 819–824, 2004.

[81] D. Clowe, A. Gonzalez, and M. Markevitch, "Weak lensing mass reconstruction of the interacting cluster 1E0657-558: Direct evidence for the existence of dark matter," *Astrophys. J.*, vol. 604, pp. 596–603, 2004.

[82] F. W. Dyson, A. S. Eddington, and C. Davidson, "A Determination of the Deflection of Light by the Sun's Gravitational Field, from Observations Made at the Total Eclipse of May 29, 1919," *Philosophical Transactions of the Royal Society of London Series A*, vol. 220, pp. 291–333, 1920.

[83] X.-P. Wu and L.-Z. Fang, "Comparisons of cluster mass determinations by x-ray observations and gravitational lensing," *Astrophys. J.*, vol. 467, pp. L45–L48, 1996.

[84] R. Massey, T. Kitching, and J. Richard, "The dark matter of gravitational lensing," *Rept. Prog. Phys.*, vol. 73, p. 086901, 2010.

[85] A. N. Taylor, S. Dye, T. J. Broadhurst, N. Benitez, and E. van Kampen, "Gravitational lens magnification and the mass of abell 1689," *Astrophys. J.*, vol. 501, p. 539, 1998.

[86] P. Natarajan *et al.*, "Mapping substructure in the HST Frontier Fields cluster lenses and in cosmological simulations," *Mon. Not. Roy. Astron. Soc.*, vol. 468, no. 2, pp. 1962–1980, 2017.

[87] A. Aguirre, J. Schaye, and E. Quataert, "Problems for MOND in clusters and the Ly-alpha forest," *Astrophys. J.*, vol. 561, p. 550, 2001.

[88] A. Abramovici and Z. Vager, "Test of newtons second law at small accelerations," *Physical Review D*, vol. 34, no. 10, pp. 3240–3241, 1986.

[89] J. H. Gundlach, S. Schlamminger, C. D. Spitzer, K.-Y. Choi, B. A. Woodahl, J. J. Coy, and E. Fischbach, "Laboratory test of newton's second law for small accelerations," *Phys. Rev. Lett.*, vol. 98, p. 150801, Apr 2007. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.98.150801

[90] J. D. Bekenstein, "Relativistic gravitation theory for the modified newtonian dynamics paradigm," *Phys. Rev. D*, vol. 70, p. 083509, Oct 2004. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevD.70.083509

[91] D. J. Mortlock and E. L. Turner, "Gravitational Lensing and Modified Newtonian Dynamics," *Publ. Astron. Soc. Aust*, vol. 18, no. 2, pp. 189–191, Jan. 2001.

[92] H. S. Zhao, D. J. Bacon, A. N. Taylor, and K. Horne, "Testing Bekenstein's Relativistic MOND gravity with Gravitational Lensing," *Mon. Not. R. Astron. Soc.*, vol. 368, no. astro-ph/0509590, pp. 171–186, Sep 2005. [Online]. Available: https://cds.cern.ch/record/885699

[93] M. Milgrom, "Testing the mond paradigm of modified dynamics with galaxy-galaxy gravitational lensing," *Phys. Rev. Lett.*, vol. 111, p. 041105, Jul 2013. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.111.041105

[94] J. Moffat, "A Modified Gravity and its Consequences for the Solar System, Astrophysics and Cosmology," *Int. J. Mod. Phys. D*, vol. 16, pp. 2075–2090, 2008.

[95] R. B. Tully and M. J. Pierce, "Distances to Galaxies from the Correlation between Luminosities and Line Widths. III. Cluster Template and Global Measurement of H_0," *The Astrophysical Journal*, vol. 533, no. 2, pp. 744–780, Apr. 2000.

[96] M. Aaronson, J. Huchra, and J. Mould, "The infrared luminosity/H I velocity-width relation and its application to the distance scale." *The Astrophysical Journal*, vol. 229, pp. 1–13, Apr. 1979.

[97] M. S. Roberts, "The neutral hydrogen content of late-type spiral galaxies." *The Astronomical Journal*, vol. 67, pp. 437–446, Jan. 1962.

[98] ——, "Integral Properties of Spiral and Irregular Galaxies," *The Astronomical Journal*, vol. 74, pp. 859–876, Sep. 1969.

[99] N. Heidemann, "Masses and Angular Momenta of Galaxies," *Astrophysical Letters*, vol. 3, p. 153, Jan. 1969.

[100] R. B. Tully and J. R. Fisher, "Reprint of 1977A&A....54..661T. A new method of determining distance to galaxies." *Astron. & Astrophys*, vol. 500, pp. 105–117, Feb. 1977.

[101] R. B. Tully and P. Fouque, "The extragalactic distance scale. I - Corrections to fundamental observables." *The Astrophysical Journal Supplement Series*, vol. 58, pp. 67–80, May 1985.

Multi-Messenger Astronomy in the GW Era

Sourav Palit

Post-doctoral fellow, Dept. of Physics, IIT Bombay, India

"Equipped with his five senses, man explores the universe around him and calls the adventure Science." - Edwin Hubble

Abstract Astronomy and astrophysics, a multi-disciplinary subject has evolved into multi-messenger one as well. Various messengers, namely photons, Cosmic ray particles, neutrinos and others have been found to reveal different aspects of physical, chemical and even potential biological (or pre-biological) processes in space. Very recently, Gravitational waves (GW) have been detected from a few transient astrophysical sources and one of them is identified with a short Gamma ray burst (sGRB) resulting from collision of two neutron stars. The joint analysis of broadband multi-wavelength electromagnetic (EM) data and the GW signal of that single event have provided us enormous amount of information spanning fields as varied as compact object merger, jet physics, cosmology, and nucleosynthesis, helping us verifying many theoretical predictions. There have been extensive ongoing efforts on multi-messenger study of such events with existing and forthcoming instruments. The future of multi-messenger astronomy, especially involving EM and GW, is looking extremely bright.

1. Introduction

We, humans live in a tiny rock in space, famously called "The Pale Blue Dot" by Carl Sagan. With all our insignificance in the vastness of the universe, we have been pursuing relentlessly to comprehend it. In the backdrop of accelerating expansion of the universe, we are extremely fortunate to be existing in a time that has allowed us to receive messengers of information from distant parts of it. All our efforts in Astronomy then come down to extending the reach of our senses by constructing innovative instruments and detectors of those messengers and analysing them to extract the scientific information they are carrying. The success so far is overwhelming and should be counted at the top of the achievements of human intellect.

Astronomers have long been dependent on receiving and analysing photons of various energies in their research and discoveries. The other messengers have also been exploited time to time with tremendous efforts, culminating at the latest addition of the gravitational waves in the last decade. Simultaneous detection and coherent interpretation of various messengers from astronomical sources can reveal true nature of those that is not conceivable by analysing any single type of messenger. In this article I intend to present a brief overview of the human endeavour in multi-messenger astronomy and astrophysics starting from the historical perspective, describing the current scenario and future efforts. I also introduce and briefly discuss about a very ambitious Indian mission called "Daksh" aiming to play the leading role in the coming decade in multi-messenger astronomy regarding gravitational counterpart detection in high energy.

2. Astronomical Messengers: A Historical Perspective

There are mainly four types of messengers, namely, photons, neutrino, particles with mass and the gravitational waves representing different fundamental forces and interactions in the universe.

2.1. Astronomy with Photons or Electromagnetic (EM) Signal

From the earliest age of known history humans have wondered seeing light from numerous tiny dots in the night sky with their naked eye, and performed some calculation on the motion and brightness of those sources. They have gradually identified them as stars, planets, comets, asteroids, meteorites and many more. First astronomical observation with an instrument other than eye was done by Gelelio in the 15th century using his telescope. The visible light has been the only captured messenger from space till the 19th century. At the beginning of this century William Herschel separated the colours and hence energy of light through prism and laid the foundation of astronomical spectroscopy. By early 20th century many ground breaking astronomical observations were made and objects were discovered with optical telescope. The quantum theory recognized light as packets of energy or quanta and named them "Photon". The same theory provided the inner structure and working of atoms and help shedding new light on interpreting astronomical spectra. In 1930s radio signal from astronomical sources were discovered and radio astronomy was founded by Karl G. Jansky. Solar coronal X-ray was detected in 1948 by instrument onboard a German-made V-2 rocket from above the Earth's atmosphere. X-ray/gamma ray astronomy effectively started with the detection of the first X-ray source (Scorpios-X1) outside of our Solar System. Cosmic microwave background from the remnant of the very hot radiation from the Big

Bang was first detected in 1968 [1]. UV and infrared astronomy were developed in the meantime as distinguished branches alongside visible light astronomy. Right now, there are numerous grounds and space-based observatories spanned along the whole of the photon spectrum plowing the sky for various astronomical objects.

2.2. Neutrino and Particle Astronomy

First solar neutrinos was detected in the Homestake experiment in 1968 [2]. Since neutrino has been found and analysed for nuclear interaction in many astronomical objects ranging from solar and stellar core to supernova and many highly energetic transients. Cosmic ray was first detected in 1912 with balloon borne instrument by Victor Hess. Solar charged particles and neutrino have also been in the use for understanding the nuclear and electromagnetic physics going in the solar atmosphere.

2.3. Gravitational Waves (GW)

The theory of gravitational waves (GW) was formulated from the field equations of General Relativity [3]. The orbital decay of the binary pulsar (binary NS) PSR 1913+16 indirectly supported the existence of GW [4]. There were many efforts of Bar detection of GW by ALLEGRO, AURIGA, EXPLORER, NAUTILUS and NIOBE over the subsequent decades [5] [6]. It was predicted that interferometry could possibly achieve the required GW detection sensitivity with better fundamental noise source characterization and broadband operation.

Nearly a century after the prediction of GW, a signature emission of GW from a binary blackhole merger (BBH) event (GW150914) was detected on September 14, 2015 by the Advanced Laser Interferometer Gravitational Wave Observatory (LIGO) detectors [7]. GW signals from a total of 90 compact binary coalescence (CBC) candidates have been identified so far by the assembly of gravitational wave detectors, consisting of LIGO, Virgo and KAGRA, with (estimated) probability of astrophysical origin greater than 0.5 [8]. The advancement in GW science has allowed us to study extragalactic objects with a new perspective and finesse previously unknown to humankind.

3. The Multi-Messenger Approach: EMGW

On August 17, 2017 a GW event called GW170817 was detected jointly in the second observation run (O2) of the GW detector network comprising of advanced LIGO and Virgo detectors [9]. The event was identified as a merger of a binary neutron star (BNS) system, with the following observation of electromagnetic counterpart across a broad spectrum ranging from radio to gamma rays. The high energy X-ray/gamma ray space monitor Fermi detected a Gamma Ray Burst (GRB) near the location of the GW event obtained by LIGO/Virgo localization [10], just after 1.7 s of the GW detection. The INTEGRAL satellite also detected a gamma ray signal independently [11]. This is the most unique such astronomical achievement commencing a new era of multimessenger observation in astrophysics consisting of simultaneous detection of electromagnetic waves and gravitational waves (EMGW).

3.1. Compact Object Merger as Source of EMGW

Mergers post collision of two compact objects, one of which is not a black hole, are considered to be most genuine and prominent source of GW emission. This is the reason, the merger events of stellar-mass BHs and NSs have been the primary sources of interest for GW detection. Before the collision when the two compact objects spiral around each other, gravitational waves are generated lasting from a few seconds up to few minutes. The signal chirps upwards in frequency peaking when the two objects merge. There are three possible pairs of merging compact objects. They are black hole-black hole (BH-BH), black hole-neutron star (NS-BH) and neutron star-neutron star (NS-NS or BNS).

When two black holes merge, they produce a new black hole as the final product. EM counterpart should not be expected from this merging, as black holes have no matter capable of producing EM signals. There are some theories that the interaction of the accretion discs around the BHs may produce some photons (see [12] for example). The NS-BH mergers are among the possible sources of EMGW emission. A black hole should always swallow the entire neutron star in a NS-BH merger with large masses, so there should not be any EM signature from those events. For lighter NS-BH systems, the tidal disruption of the NS before plunging into the BH [13] can release some amount of mass-producing EM counterpart. In Figure 1 the sequence of the merger of two compact objects (NS) and corresponding process of kilonova, formation of final compact object and finally the emanation of jet are demonstrated.

3.2. EMGW from BNS Merger and Its Physical Significance

For a long time BNS mergers are considered to be the most promising candidates of short GRBs. They are also considered a strong site of heavy r-process nucleosynthesis producing heavy elements. Following a BNS merger the remaining object created possess $\geq 90\%$ of the total masses of the two individual objects. The GW signals from the whole process of merging should last for relatively longer period of time than other such candidates. Observing simultaneous emission of GW and EM counterpart from such events can provide significant insights into the progenitors (e.g. [14]) of short GRBs. In a joint observation of such merger event the GW signal can be used as standard siren to infer the distance of the source, while the source redshift can be obtained by joint EM observation. This ideal combination has the potential of providing unambiguous measurement of the Hubble constant [15] bearing immense significance for Cosmology.

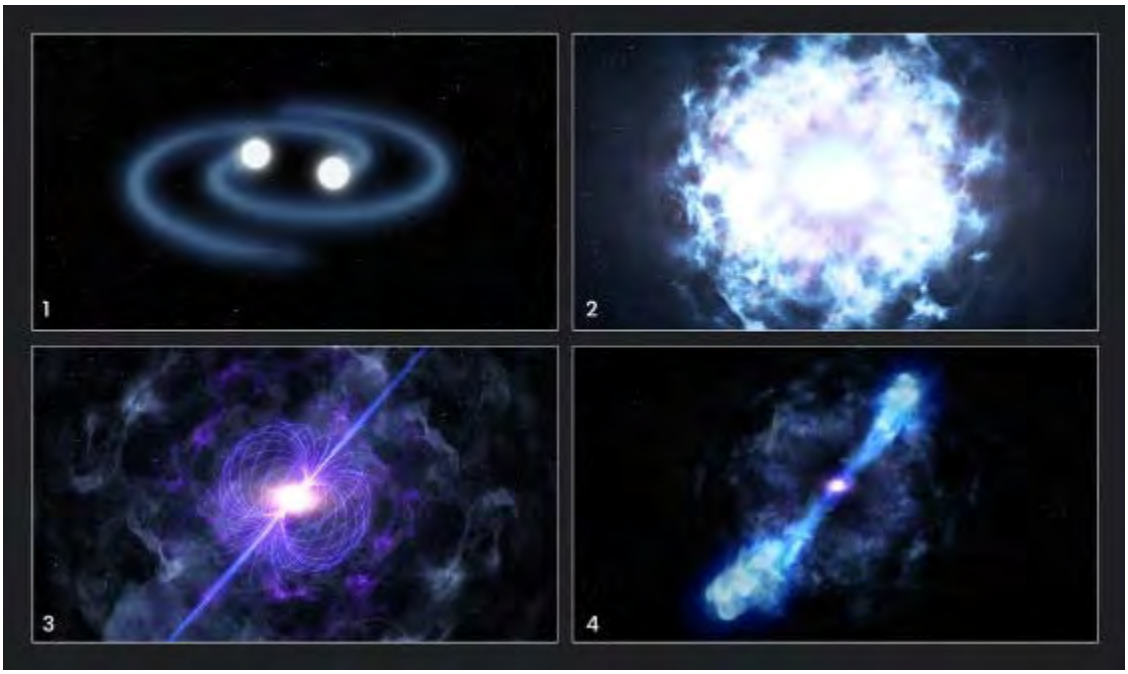


Figure 1. *A pictorial representation of the collision/merger process of two compact object through inspiral (1), followed by a kilonova (2), formation of a final spinning neutran star or magnetar with extraordinarily high magnetic field and emanation of jet (4). (Courtesy: sci-news.com)*

The tidal deformations of Neutron stars are governed by their Equation of State (EoS). These deformations enhance the GW emission and accelerate the decay of the quasi-circular in-spiral. The study of GW from the merger thus can infer the dynamics of the tidal deformation and hence infer a great deal of physics of the NS EOS.

GW and EM observation of BNS merger, when done in combination can unravel significant information related to the physics of such energetic events. They can provide us a great deal of knowledge on the emission process, particle acceleration mechanism, jet emanation, propagation, and the nature of the progenitor itself.

4. GW170817: Electromagnetic Counterpart Detections

GW170817 (see Figure 2) is the only astrophysical event so far observed both with EM and GW signals [16]. There has been a worldwide collaborative effort on the electromagnetic study of this event post GW detection. The observation spans a large band of spectrum, starting from radio waves to X-ray and gamma ray allowing the scientific community to study various physical and cosmological aspects of the event.

4.1. Radio Observations

During the merger the interaction of expanding outflows and the surrounding material produces shocks, inside which the Synchrotron radiation of accelerated electrons produces radio emission. A radio afterglow with a few week timescale [17] should also be produced from the sGRB originated by the merger, that can be observed in an small viewing angle. This has already been detected by Fong. To probe the structure of corresponding merger ejecta regardless of the observing geometry, radio detection plays the key tole and can put constrain on the structure of magnetic fields via measurement of polarization.

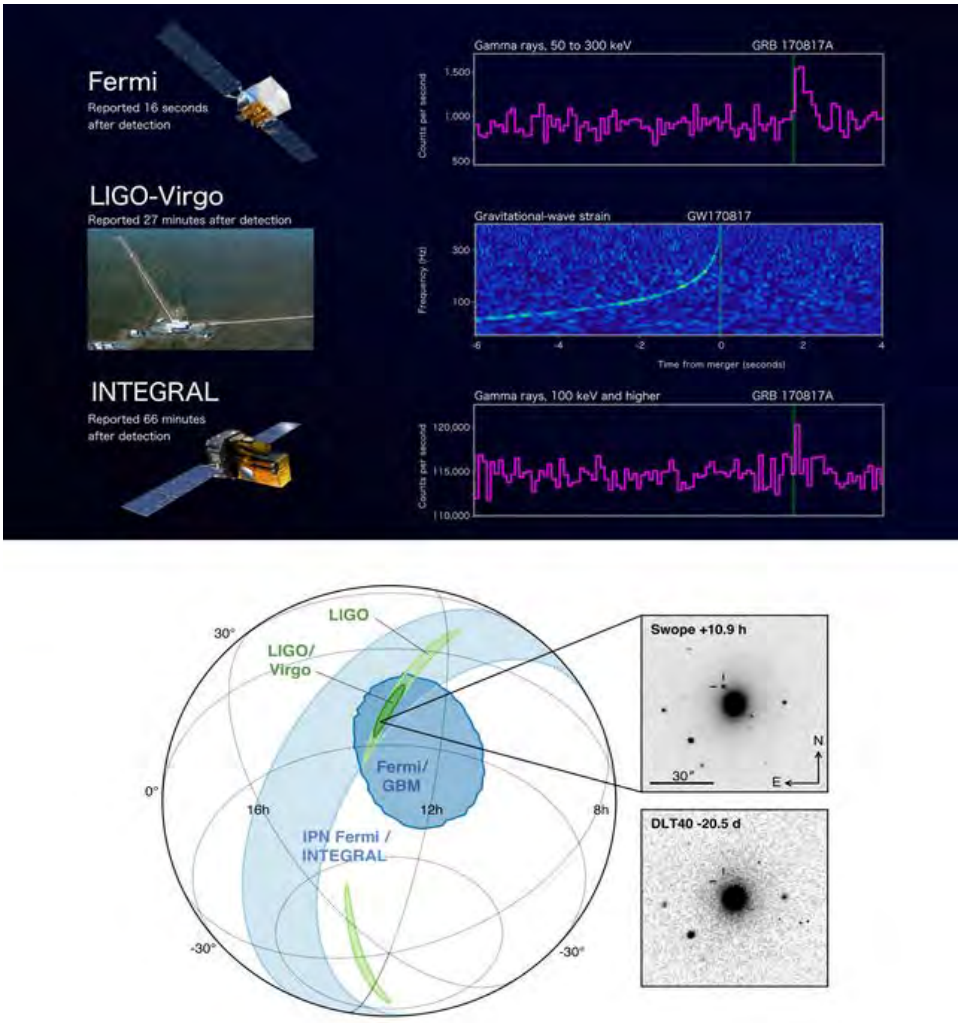


Figure 2. *Top: The detected GW signal by LVC and hard x-ray/gamma ray signal by instruments on board Fermi and INTEGRAL satellites from the BNS merger producing GW170817 and GRB170817A. Bottom: Localization of the GW, gamma-ray, and optical signals. (Courtesy: NASA GSFC & Caltech/MIT/LIGO Lab (Top), [10] (Bottom))*

4.2. Optical Studies

For merger lead sGRB, the interaction of the relativistic jet with the surrounding medium produces afterglow in the optical band. Also, the kilonova, powered by the radioactive decay of the sub-relativistic material ejecta produces optical counterpart just after the merger. When the associated jet axis is oriented at a small angle to the observer, as in the case for GW170817, it should be detected as sGRB during multi-wavelength afterglows in optical and other spectral ranges. For this particular event the viewing angle is slightly misaligned (10 - 30 deg) with respect to the polar axis [18]. Numerous telescope facilities and teams including the Antarctic Survey Telescopes (AST3), GRAvitational Wave Inaf TeAm (GRAWITA), Swope supernova survey telescope, ESO-VLT Survey Telescope (VST), Dark Energy Camera (DECam), DLT40, REM-ROS2, HST, etc. [19] has been surveying the sky for optical counterpart of GW. The AST3 Dome A detected the object and located it at NGC 4993, an S0 galaxy at a distance of 40 mega-parsecs. It measured the brightness and time evolution of optical properties and characterized it as merging binary neutron star [20]. The Swope supernova survey telescope detected the kilonova [21] corresponding to the merger and located it at the same galaxy. The optical counterpart detection of GW170817 using VLT and REM telescopes by the GRAWITA team has also been reported [16]. The spectroscopic follow-up in optical/infrared of the related kilonova AT2017gfo provided the first compelling observation establishing BNS mergers as the dominant sites for r-process heavy elements production site.

4.3. X-Ray/Gamma Ray Detections

X-ray/gamma ray sky surveys play a significant role in the multi-messenger astronomical observation. EM counterpart from the mergers involving neutron stars producing sGRBs are the prime source of interest as gravitational counterpart in high energy. The prompt emission from GW170817 was detected by hard X-ray/soft gamma-ray with Fermi-GBM detectors [22], on-board Fermi and SPI-ACS [11] on-board the INTEGRAL satellite. It consists of an emission spike of about 0.5s followed by a softer signal with a t90 of 2 ± 0.5 s as observed by Fermi. T90 is the time period when the burst is observed to emit between 5% to 95% of total measured counts in X-ray/gamma ray. The source was found to have isotropic equivalent gamma-ray energy three orders of magnitude lower than the faintest sGRB observed so far. The probable off-axis configuration of the sGRB and resulting suppressed emission along the viewing angle of the jet was assumed to be the cause of this low energy observation. Apart from the first confirmation of compact object merger as predicted source of sGRBs, there are further implications of the X-ray/gamma ray observation of the event. The delay

between the GWs and photons detected from the BNS merger provides valuable information of the jet mechanism, such as jet injection time, breakout time of the jet/cocoon etc. [23]. The essential role of a structured outflow outside the jet core in gamma-ray emission observed by Kasliwal [24] has also been established. Two competing models have been proposed as the probable cause of the observed high energy emission. The first one is the emission inside a less energetic wide-angle jet "wings" around the jet core. The second one is the shock breakout emission of the cocoon emerging from the ejecta.

MAXI telescope got the first observation of soft X-ray in the limit of 2-10 keV after ~5 hours. Swift, NuSTAR and CXO followed the afterglow observation in X-rays from hours to days following the trigger and corresponding to deeper upper limits. X-ray afterglow continuously rose to the peak up to 130 days and then went through a rapid fall, a pattern showing a clear signature of sGRB observed off-axis. From the single case of GW170817 the estimated rate of BNS merger is found to be consistent with local rate of sGRB ($\sim$1000 $Gpc^{-3}yr^{-1}$). It is of utmost necessity to have a larger statistical sample of the BNS/NS-BH mergers needing more and more future endeavour to further constrain the proposed models.

The detection of the sGRB associated with the GW event (GW170817) has been proved to be of immense importance and established many 'firsts'. This consists of the discovery of the association of EM signals with GW, first localization of BNS merger in the local universe, definitive detection of a kilonova, first observation of a structured relativistic jet observed from the side and so forth. The efforts in observing this particular event in high energy has led the astronomical community towards clear comprehension of the future need on the specifications of the detectors/instruments.

5. Current Facilities and Future Opportunities

Surveys of GW objects with various bands of electromagnetic waves have been performed during the LIGO and Virgo Collaboration (LVC) scientific runs O_2 and O_3 by numerous telescope facilities and science teams.

5.1. Existing and Proposed Facilities of EMGW Observation

Currently, radio telescopes like Karl G. Jansky Very Large Array (VLA), The Square Kilometre Array (SKA) and the next-generation Very Large Array (ngVLA) etc. will be employed for such GW counterpart detection in radio. Kasliwal [25] presents the important results related to optical counterpart observation of 13 GW triggers, which involve at least one neutron star. ZTF survey helped finding the upper limits to constrain the associated kilonova luminosity function. Though there has not been fruitful simultaneous detection of EM counterpart of GW events, the non-defections help us constraining the GRB rate and observation parameters [26]. Among many optical telescopes worlwide the GROWTH-India telescope or GIT [27] situated at the Indian astronomical observatory (IAO), Hanle, ~4500 m above the sea level is going to play a significant role in study of EMGW events during O_3 and further runs of LVC.

Other than the detection of GRB170817A by Fermi and INTEGRAL as GW counterpart of GW170817, attempt has been made to search for possible similar sGRB signals in past data of similar observations in the Fermi-GBM catalogue. Thirteen probable candidates have been identified. Out of these, for 12 GRBs no redshift information exists, and so, no further conclusion on GW association could be made. For the remaining one, namely, GRB150101B, analyses of prompt and afterglow emission have been done extensively. It is found that for this one prompt and several afterglow emission properties matched those of GRB170817A. A similar cocoon breakout model has been used successively to explain the corresponding emission mechanism [28]. Recently, significant progress has been made in the attempt of searching for sub-threshold GW events, coincident with Fermi and Swift GRB detections. Real time detection of such events can be made through ongoing and future cross searches between GWs and GRBs and would be of tremendous importance in constraining the emission models. Efforts of further such studies have been underway using both existing and future high energy space missions. Existing instruments, like those in Chandra, XMM, NuSTAR, Fermi, and IXPE observatories are actively looking for such events to make follow up EM survey of GW events. Many new instruments have also been launched and proposed. GECAM [23] already launched by Chinese academy of science (CAS); SVOM, a small X-ray telescope satellite [29] to be launched jointly by CAS and French Space Agency; Daksha, an all-sky monitor to be launched by IIT Bombay with an ISRO rocket are some notable such missions. Some proposed small high energy satellites (cubesats) such as Burstcube [30], Camelot [31], BlackCAT [32] are also waiting to join the venture. Existing AstroSat CZTI instrument [33] and proposed Polar 2 mission intend to investigate the polarization property of prompt and afterglow emission from such events to lift the degeneracy in spectral property and emission mechanism imposed by spectral observations alone.

5.2. Daksha: A High Energy Transient Monitor from India

The sGRB, GRB170817A has very low isotropic energy (E_{iso}) and only been detected by high energy instrument onboard Fermi and INTEGRAL satellites. If the sGRB were fainter even by 30% it would have not been detected at all. Due to the fact that it was occulted by the earth as seen from other important instruments like Swift-BAT and AstroSat-CZTI, the event was missed by them entirely.

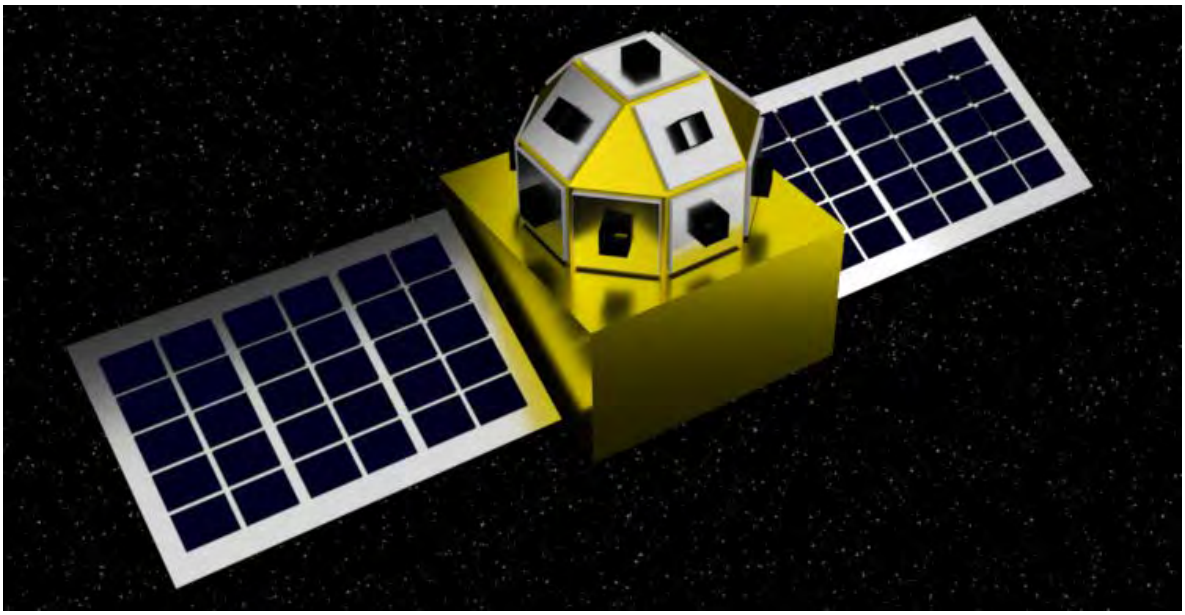


Figure 3 : A schematic of one of the pair of Daksha satellites (Courtesy: Daksha collaboration)

These facts point out the shortcomings of the existing observational facilities in detecting such events. Higher sensitivity and better sky coverage are the two most essential criteria for the upcoming instruments for finding and observing such EMGW counterparts. Proposed Indian mission 'Dakshs' has such capabilities. Daksha (Figure 3) proposed by IIT Bombay and to be launched by ISRO in few years. It will consist of two identical satellites in two exactly opposite positions in a Low Earth Orbits (LEO). It will have detectors in each of the low energy (~1-20 keV), medium energy (~20-1000 keV) and high energy (above 1 MeV) detectos spanning the possibility of observation in the whole range of X-ray to soft gamma ray (~1 keV-2 MeV) range. It will have the effective area of ~1300 cm^2 for the medium energy part of the spectrum. At any point of time, it will have a sky coverage of ~87% and will have detection sensitivity of 4×10^{-8} erg $cm^{-2}\,s^{-1}$ or 0.6 ph $cm^{-2}\,s^{-1}$. Daksha with its large energy range will have the ability to observe Comptonised spectrum of all kinds of GRBs, such as, classical (on-axis) long and short GRBs, fainter high redshifted classical GRBs, higher off-axis GRBs. The high energy coverage of Daksha will allow constraining the spectral (Band) parameters (β, E_{peak}), that is very difficult with the existing instruments.

On the localization front, large sensitive area detector plates, aligned at certain angles with respect to each other will allow precise (upto ~5^o) and prompt localization of GW counterparts. In addition, for high fluence sources, Compton imaging will be used to localize more precisely, with an angular resolution of ~1^o. Currently it is realized that the unambiguous identification of the emission process from such sources cannot be possible with spectral observation alone. X-ray/soft gamma ray polarization may play an important role in proper understanding of the physics and geometry of emission, jet launching and propagation geometry, configuration of matter and magnetic field etc. Daksha with its medium-high energy detector arrays will take an important part in achieving the goal.

With all the above-mentioned capabilities, Daksha will be extremely effective in detecting high energy transients. Its wide field of view, high sensitivity and polarisation capabilities will make it able to observe ~10 GW counterparts (sGRBs) and thousands of classical GRBs in a year, making it one of the most important missions to study prompt emission from GRBs as well as electromagnetic counterparts of the GW in the coming years.

6. Conclusion

We are standing in an exciting period of Astronomy. The culmination of human endeavour in Astronomy has reached in the form of the acquired ability to detect fundamental ripple of space time by most sophisticated instrument. It has enabled us to look at the universe with a new eye. Gradual and tremendous development of the instrumentation to capture and analyse electromagnetic and other messengers has strengthen the ability to comprehend extreme nature of the universe at the same time. With the establishment of the new LIGO-India facility the localization by GW observation alone is going to be more precise, enabling the astronomical community in more prompt and efficient observation in all EM spectrum. We are looking forward to upcoming runs of LVC to detect more and more exotic events like GW170817. In association with the next level high energy telescopes, like Daksha, giant optical telescopes, like VLT and TMT and radio detection facilities, like VLA and SKA, we are hoping to explore the physics of nature at its extremity, It will expand our knowledge and ability to comprehend the universe manifold.

References

[1] Penzias, A. A., Wilson. R. W., "A Measurement of Excess Antenna Temperature at 4080 Mc/s", *Astrophysical Journal Letters*, 142: 419–421. Bibcode:1965ApJ...142..419P. Doi:10.1086/148307, 1965.

[2] Davis, R., "A review of the homestake solar neutrino experiment", Volume 32, Pages 13-32, ISSN 0146-6410, https://doi.org/10.1016/0146-6410(94)90004-3, 1994.

[3] Einstein, A., Die Grundlage der allgemeinen Relativitätstheorie, *Annalen der Physik*, 354, 769-822. http://dx.doi.org/10.1002/andp.19163540702

[4] Hulse, R. A., & Taylor, J. H., "Discovery of a pulsar in a binary system", *Astrophysoical Journal Letter*, 195, L51, doi:10.1086/181708, 1975.

[5] Collins, H., "Gravity's shadow: The search for gravitational waves", *The University of Chicago Press*, 2004.

[6] Saulson, P. R., "Fundamentals of interferometric gravitational wave detectors", *World Scientific*, https://doi.org/10.1142/10116, 1995.

[7] Abbott, B. P., Abbott, R., Abbott, T. D. et al., "Observation of Gravitational Waves from a Binary Black Hole Merger", *Phys. Rev. Lett.*, 116, 061102, doi:10.1103/PhysRevLett.116.061102, 2016.

[8] Abbott, R., Abbott, T. D.,Abraham, S., et al., "GWTC-2: Compact Binary Coalescences Observed by LIGO and Virgo during the First Half of the Third Observing Run", *Physical Review X*, 11, 021053, doi:10.1103/PhysRevX.11.021053, 2021a.

[9] Abbott, B. P., Abbott, R., Abbott, T. D., et al., "GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral", *Phys. Rev. Lett.*, 119, 161101, doi:10.1103/PhysRevLett.119.161101, 2017a.

[10] Abbott, B. P., Abbott, R., Abbott, T. D., et al., "Gravitational Waves and Gamma-Rays from a Binary Neutron Star Merger: GW170817 and GRB 170817A", *Astrophysical Journal Letter*, 848, L13, doi:10.3847/2041-8213/aa920c, 2017b.

[11] Savchenko, V., Ferrigno, C., Kuulkers, E., et al., "INTEGRAL Detection of the First Prompt Gamma-Ray Signal Coincident with the Gravitational-wave Event GW170817", *Astrophysical Journal Letter*, 848, L15, doi:10.3847/2041-8213/aa8f94, 2017.

[12] Kelly, B. J., Baker, J. G., Etienne, Z. B., Giacomazzo, B., & Schnittman, J., "Prompt electromagnetic transients from binary black hole mergers", *Phys. Rev. D*, 1096, 123003, doi:10.1103/PhysRevD.96.123003, 2017.

[13] Foucart, F., Hinderer, T., & Nissanke, S., "Remnant baryon mass in neutron star-black hole mergers: Predictions for binary neutron star mimickers and rapidly spinning black holes", *Phys. Rev. D*, 98, 081501, doi:10.1103/PhysRevD.98.081501, 2018.

[14] Burns, E., "Neutron star mergers and how to study them", *Living Reviews in Relativity*, 23, doi:10.1007/s41114-020-00028-7, 2020.

[15] Schutz, B. F., "Determining the Hubble constant from gravitational wave observations", *Nature*, 323, 310, doi:10.1038/323310a0, 1986.

[16] Grado, A., "The optical electromagnetic counterpart of the gravitational wave event GW170817", *Nuclear and Particle Physics Proceedings*, 306-308, 42, doi:https://doi.org/10.1016/j.nuclphysbps.2019.07.006, 2019.

[17] Hotokezaka, K., Nissanke, S., Hallinan, G., et al., "Radio Counterparts of Compact Binary Mergers Detectable in Gravitational Waves: A Simulation for an Optimized Survey", *The Astrophysical Journal*, 831, 190, doi:10.3847/0004-637x/831/2/190, 2016.

[18] Mooley, K. P., Nakar, E., Hotokezaka, K., et al., "A mildly relativistic wide-angle outflow in the neutron-star merger event GW170817", *Nature*, 554, 207–210, doi:10.1038/nature25452, 2017.

[19] Abbott, B. P., Abbott, R., Abbott, T. D. et al., "Multi-messenger Observations of a Binary Neutron Star Merger", *Astrophysical Journal Letter*, 848, L12, doi:10.3847/2041-8213/aa91c9, 2017c.

[20] Hu, L., Wu, X., Andreoni, I., et al., "Optical observations of LIGO source GW 170817 by the Antarctic Survey Telescopes at Dome A, Antarctica", *Science Bulletin*, 62, 1433–1438, doi:10.1016/j.scib.2017.10.006, 2017.

[21] Coulter, D. A., Foley, R. J., Kilpatrick, C. D., et al., "Swope Supernova Survey 2017a (SSS17a), the optical counterpart to a gravitational wave source", *Science*, 358, 1556–1558, doi:10.1126/science.aap9811, 2017.

[22] Goldstein, A., Veres, P., Burns, E., et al., "An Ordinary Short Gamma-Ray Burst with Extraordinary Implications: Fermi-GBM Detection of GRB 170817A", *Astrophysical Journal Letter*, 848, L14, doi:10.3847/2041-8213/aa8f41, 2017.

[23] Zhang, D., Li, X., Xiong, S., et al., "Energy response of GECAM gamma-ray detector based on LaBr3:Ce and SiPM array", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 921, 8, doi: https://doi.org/10.1016/j.nima.2018.12.032, 2019.

[24] Kasliwal, M. M., Korobkin, O., Lau, R. M., Wollaeger, R., & Fryer, C. L., "Infrared Emission from Kilonovae: The Case of the Nearby Short Hard Burst GRB 160821B", *Astrophysical Journal Letter*, 843, L34, doi:10.3847/2041-8213/aa799d, 2017.

[25] Kasliwal, MJ., Anand, S., Ahumada, T., et al., "Kilonova Luminosity Function Constraints Based on Zwicky Transient Facility Searches for 13 Neutron Star Merger Triggers during O3", *The Astrophysical Journal*, 905, 145, doi:10.3847/1538-4357/abc335, 2020.

[26] Andreoni, I., Coughlin, M. W., Kool, E. C., et al., "Fast-transient searches in real time with ZTFReST: Identification of three optically discovered gamma-ray burst afterglows and new constraints on the kilonova rate", *The Astrophysical Journal*, 918, 63, doi:10.3847/1538-4357/ac0bc7, 2021.

[27] Prabhu, T. P., "Indian Astronomical Observatory, Hanley", *Bullentin of the Astronomical Society of India"*, 28, 233, 2000.

[28] Troja, E., Ryan, G., Piro L. et al., "A luminous blue kilonova and an off-axis jet from a compact binary merger at z = 0.1341; Nat. Commun 9, 4089. https://doi.org/10.1038/s41467-018-06558-7; 2018.

[29] Paul, J., Wei, J., Basa, S., & Zhang, S.-N., "The Chinese–French SVOM mission for gamma-ray burst studies", *Comptes Rendus Physique*, 12, 298, doi: 10.1016/j.crhy.2011.01.009, 2011.

[30] Racusin, J., Perkins, J. S., Briggs, M. S., et al., "BurstCube: A CubeSat for Gravitational Wave Counterparts", https://arxiv.org/abs/1708.09292, 2017.

[31] Werner, N., Ripa, J., Pal, A., et al., "CAMELOT: Cubesats Applied for Measuring and Localising Transients-Mission Overview", https://arxiv.org/abs/1806.03681, 2018.

[32] Falcon, T., A. D., Burrows, D. N., Fox, D. B., & Palmer, D., "BlackCAT CubeSat: a soft x-ray sky monitor, transient finder, and burst detector for high-energy and multimessenger astophysics", *Proceedings of SPIE – The International Society for Optical Engineering*, 10699, doi:10.1117/12.2314274, 2018.

[33] Bhalerao, D. Bhattacharya, A. Vibhute, P. Pawar, A. R. Rao, M. K. Hingar, Rakesh Khanna, A. P. K. Kutty, J. P. Malkar, M. H. Patil, Y. K. Arora, S. Sinha, P. Priya, Essy Samuel, S. Sreekumar, P. Vinod, N. P. S. Mithun, S. V. Vadawale, N. Vagshette, K. H. Navalgund, K. S. Sarma, R. Pandiyan, S. Seetha & K. Subbarao, "The Cadmium Zinc Telluride Imager on AstroSatV", *Astrophys. Astr. 38:31 Indian Academy of Sciences*, doi 10.1007/s12036-017-9447-8, 2017.

Properties of Giant Radio Galaxies

Netai Bhukta[1], Sabyasachi Pal[2,*], Sushanta Kumar Mondal[1]

[1]Department of Physics, Sidho Kanho Birsha University, Purulia, 723104, India
[2]Midnapore City College, Kuturia, Bhadutala, Paschim Medinipur, West Bengal, 721129, India
*Corresponding author: sabya.pal9@gmail.com

Abstract

One of the largest astrophysical radio structures in the Universe is the Giant Radio Galaxies (GRGs), with a linear projected size of nearly 0.7 Mpc or more. GRGs are rare objects that were once used to grow in low-density environments. Morphologically, the majority of GRGs display bright hot-spots at the edges of their radio lobes (FR-II). Over the course of about 45 years, a few hundred GRGs were discovered. In this review, we study different radio and physical properties of GRGs (total radio flux density, angular and projected linear size, jet kinetic power, radio power, the mass of black holes, black hole spin, and classification of GRGs, etc.).

Keywords: *Radio Galaxies, Giant Radio Galaxies, Active Galactic Nuclei (AGN), Radio Jets.*

1 Introduction

In the Universe, giant radio galaxies are the largest singular structures. These sources are rare and useful probes to study intergalactic space. In the 1950s, it was discovered that many galaxies dominantly emit at radio wavelengths through synchrotron radiation [1] [2] [3] [4]. These galaxies later came to be known as radio galaxies (RGs) because their radio emission frequently extends beyond the visual limit of galaxies at optical wavelengths. There are double linear formations of two diffuse portions of radio emission, known as lobes, which are located on both sides of a host optical galaxy. Between the lobes, a small, luminous centre (known as the core) is frequently detected, coinciding with the host galaxy. The core is related to one or both lobes via straight, narrow jets. Often, a hotspot is visible on the outer edge of a lobe where a jet approaches the end of the lobe. A hotspot is a tiny area of stronger emission within the lobe. Based on the distribution of luminosity inside their lobes [5], there are two kinds of double sources: (1) FR-I, 2) FR-II [6]. The core region of FR-I samples is edge-darkened, with a greater density of radio flux compared to the edges of the lobes. But for FR-II sources, they are edge-brightened, where the outside edges of lobes have higher flux density compared to the interior a reas. The core areas of an FR-II source may be invisible depending on the observation limit of flux d ensity. In general, FR-II sources have a stronger luminosity (total power of radio emission) [6]. Radio galaxies with an active supermassive black hole (SMBH) at the centre are called active galactic nuclei (AGNs), which are observable from radio frequencies to gamma frequencies. AGNs that mainly radiate at radio frequencies are known as radio-loud AGNs (RLAGNs). Quasars are called radio-loud quasars (RQs) when they emit radiation at radio wavelengths.

GRGs are active galaxies that form primarily from SMBHs. Massive amounts of mass are accreted under the impact of the gravitational field of SMBH. From different morphological studies, for most GRGs, the radio core is less luminous than two edge-bright hot spots, and hence they are classified as FR-II radio g alaxies. The core engine consists of a super massive black hole (SMBH) with a typical mass of $10^8 - 10^{10}\ M_\odot$. Two collimated, relativistic radio jets are produced in the perpendicular direction to the accretion disc [7]. In the Universe, a small fraction of AGNs display high-power radio emission and bright luminosity. The radio jets are directed through the low density intercluster medium ($10^{-5} - 10^{-6}$ per cubic centimetre), resulting in a projected linear size ranging from a thousand kiloparsec (kpc) to a megaparsec (Mpc). Such types of GRG sources are not detected in the high-frequency surveys since radio lobes are not always visible in that case. Low-frequency radio surveys are ideal for detecting GRGs [8].

The main criterion of a GRG is a large angular size. A very extensive 4.69 Mpc GRG is detected in the NVSS survey [9]. Most GRGs are found below redshift $z \sim 0.5$, and their projected linear sizes are limited to 2 Mpc. Usually, bright elliptical galaxies are the optical hosts of GRG galaxies, but in some rare cases, two

tremendously massive rapid rotating spiral galaxies host the two relativistic jets and lobes, and they elongate on Mpc scales [10]. There are many studies to realise the cause of the long projected linear sizes of GRGs. The study of the dynamical age of GRGs indicates these have evolved over a long interval of time [11]. However, the spectral age of GRGs is comparable to that of normal-sized radio galaxies [12]. Hence, different scientists believe that the lengthy projected linear sizes of GRGs are probably due to the lower density of the intergalactic medium surrounding these sources. Many authors have also investigated the role of AGN power in creating these giant radio sources. The giant radio galaxies have similar core strengths compared to the rest of the radio galaxies of normal sizes, having comparable luminosity [13]. It is also interesting to note that several of the giant radio sources show recurrent jet activity [14].

2 Discovery of GRGs in Different Surveys

Millions of RGs have been discovered and catalogued over the last six decades, but only a few hundred have been discovered to have linear sizes of Mpc-scale. Since GRGs' discovery in the 1970s, various names have been given to this uncommon sub-class of RGs, such as 'giant radio sources' (GRSs), 'large radio galaxies' (LRGs), and 'giant radio galaxies' (GRGs) [15]. To get rid of confusion and retain uniformity, we assign to this giant sub-class of RGs the name giant radio galaxies [16] [17] [18] [19] [20].

From the 1970s until the early 2000s, many RGs were classified as GRGs depending on their projected linear size, which was computed by using the Hubble constant (H_0). At that time values of H_0 were assumed in between 50 to 100 km $s^{-1}Mpc^{-1}$. The outcome of the projected linear sizes of these sources were often above or under-estimated, which led to the misleading statistical result of their population.

The value of H_0 was settled to $\sim$ 68 km $s^{-1}Mpc^{-1}$ with the increasing precision of cosmology measurements using the cosmic microwave background radiation seen with the Wilkinson Microwave Anisotropy Probe (WMAP; [21]) and Planck mission [22]. The first two GRGs (3C 236 and DA240) were discovered by Willis [15], both of which have a linear size longer than 2 Mpc, and as a result, they did not clearly define a minimum size requirement for RGs to be classified as G RGs. In later research, a 1 Mpc linear size was used as a lower limit of GRGs, assuming the value of H_0 nearly to 50 km $s^{-1}Mpc^{-1}$ [13] [15] [16]. Recent research [17] [18] [19] has adopted that 0.7 Mpc linear size is the lower limit of GRGs. This limiting value is computed from the updated H_0 value.

Different deep and large sky radio surveys like Faint Images of the Radio Sky at Twenty cm (FIRST; [23], NRAO VLA Sky Survey (NVSS; [9]), Westerbork Northern Sky Survey (WENSS; [24]) and Sydney University Molonglo Sky Survey (SUMSS; [25]) were conducted to search for different RGs. Millions of RGs have been discovered, and a significant portion of them have been thoroughly investigated. Only a few hundred of these RGs, however, have been classified as giant radio sources, or GRGs. Several difficulties in the identification resulted in the detection of a small number of GRGs.

The lobes of GRGs are dominated by steep spectral indices, and hence they are clearly detected at low radio frequencies. The low-frequency 7C survey first discovered a sample of GRGs at 151 MHz [26]. Using a better resolution of the WENSS survey in comparison to the 7C survey, 47 GRGs were discovered [16]. Four large low-frequency surveys have been conducted in recent years: (1) $119-158$ MHz Multifrequency Snapshot Sky Survey (MSSS; [27]), (2) 150 MHz TIFR GMRT Sky Survey-Alternate data release-1 (TGSS-ADR1; [28]), (3) $72-231$ MHz GaLactic and Extragalactic All-sky Murchison Widefield Array (GLEAM) survey [29], (4) $120-168$ MHz LOFAR Two-metre Sky Survey (LoTSS; [30] [31]). Over the course of about 45 years, about six hundred GRGs were discovered [18].

3 Formation Model of GRGs

Giant radio galaxies have extreme sizes, being bigger than other radio galaxies. The giant radio sources are the last episodes of radio galaxy evolution in existing models [11]. Radio galaxies begin as small and confined sources with a size not greater than 10 kiloparsecs (kpc), which is shorter than the optical host galaxy. When radio jets are extended and begin to inflate, then the radio sources are transformed into normal FR-I or FR-II RGs depending on the jet kinetic power. The two radio jets are eventually closed down, and the radio jet kinetic power reduces as the electrons miss out on the energy, but the radio lobes continue to spread. Radio galaxies may have grown large enough to become giant radio galaxies after $\sim 10^7$ years. Radiation cutoff energy in giant radio galaxies can be explained by the loss of energy in the ageing lobes. In Figure 1, we presented a sample of GRGs taken from the TGSS ADR 1 survey.

The models of the mechanisms in GRGs are in their infant stage. These are very general theories, and many details need to be filled in and checked. Till now, it is unclear what percentage of FR-I and FR-II sources become giant radio sources and which factors influence the possibility of such a change. The lifetimes of giant radio galaxies are not very different from the ages of other small-sized radio galaxies, which suggests that other elements may play a role in their giant size. The fact that GRGs are more likely to be FR-II suggests that substantial beginning energies are required to form a 0.70 Mpc projected linear size of lobe. The lobes of FR-I sources are hard to identify at first because they are less energetic, and as they lose strength with age, they become more difficult to identify [13]. Another possible scenario is that the density of the intergalactic medium (IGM) is lower near GRGs. As the Universe grows, the density and pressure of the IGM should decrease. If the production of GRGs is highly influenced by ambient pressure, we should expect to detect a lower number of GRGs at higher redshifts.

4 Radio Properties of GRGs

4.1 Radio Flux Density

Radio emission seen from extragalactic populations observed in radio surveys, AGN, and star-forming galaxies, originates mainly from synchrotron radiation. This radiation is generated by relativistic electrons in the presence of magnetic fields and produces the electromagnetic energy spectrum that is emitted by charged particles (relativistic and ultra-relativistic). When charged particles move at a velocity near the velocity of light, the trajectory of particles is changed by the magnetic field. This radiation is mainly dependent on the mass of the charged particle. This process is accountable for the radio spectrum generated from GRGs. It is also responsible for the optical spectra of the non-thermal process.

The observed radio spectrum (1 cm – 100 m) of extragalactic sources are well described with a power law: $F_\nu \propto \nu^{-\alpha}$, where F_ν is the observed monochromatic energy flux at the specific frequency ν, and α is the spectral index ($\alpha = -\frac{p-1}{2}$, where p is the power-law index of the initially injected electron energy distribution). A pure power-law spectrum is a clear signature of synchrotron radiation. Synchrotron self-absorption (photons emitted via the synchrotron mechanism can scatter off an electron) is observed at the low-frequency end of radio-emitting plasma, where power follows as $F_\nu \propto \nu^{\frac{5}{2}}$.

4.2 Angular and Linear Size

The conventional physical projected linear size of GRG $\geq$ 0.7 Mpc. The linear size of GRGs can be computed using the following formula:

$$d(\text{Mpc}) = \frac{\delta \times D_{co}}{(1+z)} \times \frac{\pi}{10800} \tag{1}$$

where D_{co} is a comoving distance of the galaxies, δ is angular size between two radio lobes, z is the redshift, and d is projected linear size [32]. Throughout this review, the flat ΛCDM cosmological model is adopted based on the Planck results (H_0 = 67.8 s^{-1} Mpc^{-1}, Ω_m = 0.308, Ω_Λ = 0.692, and $\Omega_k = 0$ [22].

In the updated catalogue of giant radio sources, the projected linear size of GRGs is larger than 0.7 Mpc, which extends up to 4.7 Mpc with the median size of 1.14 Mpc [33]. In the catalogue of GRGs using LOFAR, GRGs have projected linear sizes of between 0.7 Mpc – 3.5 Mpc, with a median and a mean size of 0.89 Mpc and 1.02 Mpc, respectively [18]. In the TGSS GRG catalogue, 54 GRGs have linear sizes in the range of 0.7 Mpc to 1.82 Mpc, with a median size of 1.02 Mpc and a mean size of 1.09 Mpc [34]. With matched redshifts of samples of GRGs and GRQs, the extension of the projected linear size of GRQs is not remarkably separate from that of GRGs.

4.3 Spectral Index

The spectral index ($alpha$) of a radio source represents the energy distribution of relativistic electrons, and its spectrum should ideally extend over a wide wavelength range [35]. The radio flux density changes with frequency as $F_\nu \propto \nu^{-\alpha}$, resulting in a two-point spectral index measurement, given as follows:

$$\alpha = \frac{\log F_{\nu_1} - \log F_{\nu_2}}{\log \nu_2 - \log \nu_1} \tag{2}$$

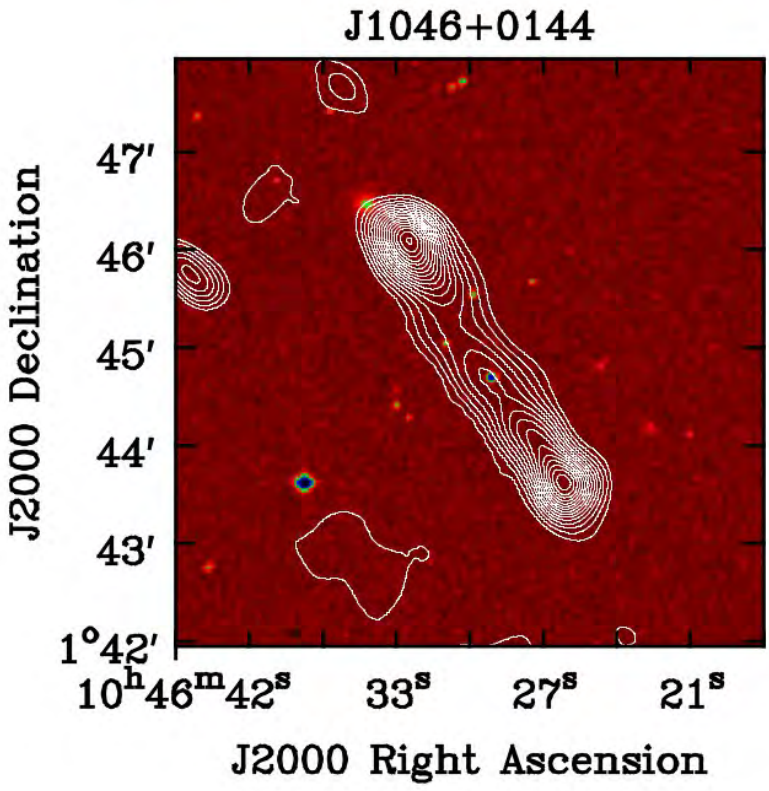


Figure 1: Giant radio galaxy J1046+0144 in TGSS as observed in 150 MHz [34].

The uncertainty of spectral index measurements due to flux density uncertainty [36] is

$$\Delta\alpha = \frac{1}{\ln\frac{\nu_1}{\nu_2}}\sqrt{\left(\frac{\Delta F_1}{F_1}\right)^2 + \left(\frac{\Delta F_2}{F_2}\right)^2} \tag{3}$$

where F_{ν_1} and F_{ν_2} refers to flux densities at two different frequencies. For example, the flux density accuracy in TGSS ADR1 and NVSS are $\sim$ 10 per cent [28] and $\sim$ 5 per cent [9]. Using the above equation, the spectral index uncertainty ($\Delta\alpha$) is 0.05. Different surveys (Westerbork Synthesis Radio Telescope (WSRT) at 325 MHz and 1400 MHz towards the Lynx field [37]; Australia Telescope Compact Array (ATCA) at 1.4 GHz and 2.4 GHz towards the Marano field [38]; Molonglo Radio Catalogue sources with the VLA at L and S bands [39]; 150 MHz band of the Giant Metrewave Radio Telescope and other archival data from GMRT at 610 MHz and 325 MHz in the LBDS-Lynx field [40], and LOFAR 150 MHz and WSRT 1.4 GHz in the Lockman Hole field [36]) were conducted to study the spectral properties of RGs. All the above surveys found a spectral index in the range of $\sim 0.7 - 0.8$ for most RGs, and as a result, $\alpha \sim 0.75$ is commonly considered to be the average spectral index for RGs.

GRGs and GRQs are found in the LOFAR survey. For GRGs, the mean and median values of the spectral index are 0.79 and 0.77, respectively, and for GRQs they are 0.76 and 0.78, respectively [18]. The average value of the spectral indexes of GRGs and GRQs is analogous to those of RGs. The aforementioned result indicates that RGs and GRGs have similar spectral index properties and that the majority of GRGs are active phases of RGs rather than remnants or dead RGs.

4.4 Radio Power

The luminosity distance (D_L) is dependent on various cosmological assumptions, specifically, the density of matter as a fraction of critical density Ω_m, the cosmological constant as a fraction of critical density Ω_Λ, the deceleration parameter $q_0 = \frac{1}{2}\Omega_m - \Omega_\Lambda$ and the Hubble constant H_0.

For a Universe with $\Omega_\Lambda = 0$, the luminosity distance is defined by [41]

$$D_L = \frac{c}{q_0^2 H_0^2}[q_0 z + (q_0 - 1)(\sqrt{1 + 2q_0 z} - 1)] \tag{4}$$

The radio power (P_{rad}) of GRGs can be calculated (using spectroscopic and photometric z value) using standard formula [42]

$$P_{rad} = 4\pi {D_L}^2 F_0 (1 + z)^{\alpha - 1} \tag{5}$$

where z is the redshift of the radio galaxy, α is the spectral index ($S \propto \nu^{-\alpha}$), D_L is luminosity distance to the source (Mpc), F_0 is the flux density (Jy) at a given frequency.

The overall radio power of FR-I sources is, in general, less than that of FR-II sources [43]. The distribution of radio power of the FR-I type of GRGs has a limited range of 10^{24} to 10^{26} W/Hz, whereas the FR-II displays a broad range of radio powers (10^{24} – 10^{28} W/Hz). In the SAGAN GRGs and GRQs catalogue at 1400 MHz (for matched and not matched redshift samples), GRQs have more radio power than GRGs [18]. Core engines of GRQs can generate massive, powerful radio jets, resulting in more radio-luminous sources than GRGs. In the GRG catalogue from TGSS, sources of radio power at 150 MHz are in the order of 10^{27} W Hz^{-1}, which is similar to a typical giant radio galaxy. The average value of $Log\ P$ [W Hz^{-1}] for GRGs is 27.25 (1σ standard deviation $= 1.61$, median $= 27.12$). The detected GRGs from TGSS ADR 1 at 150 MHz are more luminous compared to those detected from the NVSS survey at 1400 MHz and the low-frequency survey LOFAR Two-metre Sky Survey.

4.5 Black Hole Mass Estimation of GRG

The central black hole mass of the host of the GRGs can be computed using two methods, which are as follows: 1) $M_{\text{BH}} - \sigma$ relation, 2)$M_{\text{BH}} - L_{K,bulge}$ relation.

Information about the central velocity dispersion for optical hosts is available via fibre spectroscopy in SDSS DR15 [44]. One can compute the central black hole mass using the well known $M_{\text{BH}} - \sigma$ [45].

$$\log(M_{\text{BH}}/M_\odot) = \alpha + \beta\ log(\sigma/200\ \text{kms}^{-1}) \tag{6}$$

where $\alpha = 8.32 \pm 0.05$ and $\beta = 5.64 \pm 0.32$.

The central black hole mass of GRGs was computed using the K band magnitude of the (bulge dominated) elliptical host galaxy ($L_{K,bulge}$) and using the correlation $M_{\text{BH}} - L_{K,bulge}$ [46] [47]

$$\log(M_{\text{BH}}/M_\odot) = (0.95 \pm 0.15) \log \frac{L_{K,sph}}{10^{10.91 L_{K,\odot}}} + 8.26(\pm 0.11) \tag{7}$$

where $\frac{L_{K,sph}}{L_{K,\odot}}$ is the K band luminosity of the spheroid component of the galaxy (i.e., the bulge or the elliptical galaxy itself) in solar units. A broad-band luminosity in K band is calculated from [48]

$$L_K = 4\pi D_L^2 \alpha_K \nu_K 10^{-0.4 \times m_k}, \tag{8}$$

where α_K is the zero-magnitude flux (Jy) for given frequency band, m_k is the observed source magnitude, ν_K is the central frequency (Hz), and D_L is the luminosity distance at specified redshift (z).

4.6 Jet Kinetic Power

The jet kinetic power (P_{jet}) can be derived from the low-frequency radio observations, which allow us to study various characteristics of the radio-loud AGN system. High radio frequencies (~ 1 GHz) are ideal for studying the nuclear radio jet components due to their flatter spectral nature. Due to the high velocities of these components, relativistic effects, including the Doppler enhancement effects, are significant. Therefore, low radio frequencies are more appropriate for probing the jet's kinetic power. At low frequencies, the contribution from Doppler boosting is negligible. The following relation from the simulation-based analytical model can be used to estimate the jet kinetic power [49]:

$$L_{150} = 3 \times 10^{27} \frac{P_{\text{jet}}}{10^{38} W} \text{WHz}^{-1}, \tag{9}$$

where L_{150} is the radio luminosity at 150 MHz (W/Hz). In GRGs, the radio luminosity of the sources reduces with linear size for different jet kinetic powers [49].

4.7 Accretion Rate of SMBH

The accretion rate of the SMBH of GRGs can be measured by the dimensionless Eddington ratio (λ_{EDD}). This ratio is defined as the AGN bolometric luminosity to the maximum Eddington luminosity, i.e.,

$$\lambda_{\rm EDD} = \frac{L_{\rm bol}}{L_{\rm EDD}}, \tag{10}$$

where the term $L_{\rm bol}$ is the bolometric luminosity, and $L_{\rm EDD}$ is the Eddington luminosity. The bolometric luminosity of AGN is sometimes approximated by its optical luminosity. From the optical [OIII] emission line spectra, $L_{\rm bol}$ is calculated as $L_{bol} = 3500 \times L_{[III]}$ [50]. The highest luminosity of a black hole is known as the Eddington luminosity. For pure ionised hydrogen plasma, the expression of the Eddington luminosity is $L_{EDD} = 1.3 \times 10^{38} \times \left(\frac{M_{\rm BH}}{M_\odot}\right)$ erg/s.

4.8 Black Hole Spin

The spin (a) is a fundamental property of the black hole together with the mass. This can assist in the understanding of the history of mergers and accretion activity in the central engine during the past billion years [51] [52] [53] [54]. In the B-Z model, the combined impact of rotation and the accumulating magnetic field surrounding the black hole fed matter through the rotating accretion disc [55] [56] [57].

According to the B-Z model, the relationship between the jet power ($P_{\rm jet}$), the black hole mass ($M_{\rm BH}$), the black hole dimensionless spin ($a = Jc/GM^2$, where J is the angular momentum), and the poloidal magnetic field (B) threading the accretion disc and ergosphere take the following form:

$$P_{jet} \propto B^2 M_{\rm BH} a^2 \tag{11}$$

where $P_{\rm jet}$ is in units of 10^{34} erg/s, B is in units of 10^4 Gauss, $M_{\rm BH}$ is in units of $10^8 M_\odot$. The zero and unit value of dimensionless spin parameter (a) represents a non-rotating black hole and a maximally spinning black hole. In this model, the proportional constant of the above equation is taken $\sim \sqrt{0.5}$. In order to compute the spin of a black hole, there is difficulty in predicting the magnetic field in the surroundings of the black hole.

We consider the Eddington magnetic field strength ($B_{\rm Edd}$) [54] [57] is

$$B \sim B_{EDD} \approx 6 \times 10^4 (\frac{M_{\rm BH}}{10^8 M_\odot})^{-\frac{1}{2}} \quad \text{Gauss} \tag{12}$$

This is the maximum strength of the magnetic field near the centre engine. There is an assumption that the overall energy density of the accreting plasma with an Eddington luminosity radiation field is balanced by the magnetic field energy density.

5 Physical Properties of GRGs

5.1 Morphology of GRGs

GRGs can be classified into four categories using various kinds of radio sky surveys (VLASS, FIRST, NVSS, and TGSS): FR-I, FR-II [6], HyMoRS [58] [59] [60], and DDRG. HyMoRS, or hybrid radio galaxies, are radio galaxies having FR-I and FR-II morphologies on opposite sides of the radio core. Higher-resolution radio maps are required to evaluate the morphology of the HyMoRS candidate GRGs. About 93 per cent (50/54) of the GRGs in the TGSS survey [34] show an FR-II type (edge-brightened hotspots within radio lobes) radio morphology, whereas only 4 out of 54 GRGs (7 per cent) show an FR-I type radio morphology. The radio power (P_{150}) for FR-I type GRG ranges from $\sim 0.3 \times 10^{27} WHz^{-1}$ to $\sim 1.3 \times 10^{27}$ W Hz^{-1}, and for FR-II type GRGs the range extends from $\sim 3.0 \times 10^{27} WHz^{-1}$ to $\sim 1.21 \times 10^{30}$ W Hz^{-1}.

5.2 GRGs and GRQs

Giant radio galaxies which acquire the radio powered by radio-loud AGN and quasars are known as GRGs and GRQs, respectively. Normally, GRGs and GRQs have different redshift distributions since GRQs tend to be born at higher redshifts than GRGs. The estimated linear size distributions of GRQs and GRGs are not significantly

different for a particular redshift evolution range. The radio power of GRQs is greater than that of GRGs at 1400 MHz and 150 MHz for matching redshift samples [18] [34]. The central engines of GRQs are capable of producing stronger radio jets and luminosity than GRGs. The jet kinetic power of GRQs is found to be larger than that of GRGs. It is suggested that they may host more massive black holes that accrete at a greater Eddington rate. Till now, the GRQs with the strongest radio power have been observed in the TGSS field with $P_{\text{rad}} \sim 1.21 \times 10^{30}$ WHz^{-1}, at $z \sim 4.30$ [34]. The jet kinetic power of GRQs is greater than that of GRGs. The jet kinetic power of GRGs was reported in the range of 10^{42} ergs^{-1} to 10^{44} ergs^{-1} [19]. There is a negative correlation between jet power (P_{jet}) and dynamical age (t_{age}), i.e., $P_{\text{jet}} \propto \frac{1}{t_{age}^2}$ [61]. This implies that, if their sizes are identical, the more powerful radio jets of GRQs would scale Mpc length faster than the GRGs (if they belong to a similar environment of ambient density). The spectral index distributions of GRGs and GRQs are identical [18] [34].

5.3 Comparison of HEGRGs and LEGRGs

Depending on various accretion mechanisms, radio-loud AGNs can be divided into two classes: low-excitation giant radio galaxies (LEGRGs) and high-excitation giant radio galaxies (HEGRGs) [62]. The colour-colour plot in four mid-IR bands (W1, W2, W3, and W4) is used to separate not only LEGRGs and HEGRGs, but also AGNs from starforming and inactive galaxies. For HEGRGs, W1–W2 > 0.5, W2–W3 < 5.1, and for LEGRGs, W1–W2 < 0.5, $0 <$ W2–W3 < 1.6.

Considering the projected linear sizes, the natural tendency of HEGRGs is found to be higher than LEGRGs. HEGRGs have a stronger radio power compared with LEGRGs because they are in high accretion condition. As a result, GRGs have gained high radio power [63]. The higher median and mean values of jet kinetic power of HEGRGs compared to LEGRGs strongly suggest that GRG-AGNs with high-excitation types are accountable for launching maximum-powered radio jets [18]. For LEGRGs, optical r-band magnitude (M_r) is higher compared to HEGRGs. LEGRGs are primarily hosted by bright, massive elliptical galaxies with an aged star population [64]. A comparative study of the black hole mass (M_{BH}) distribution in LEGRGs and HEGRGs is very important for understanding the main factors controlling the two distinct accretion modes of AGN. The mean and median M_{BH} of LEGRGS are higher than those of HEGRGs in a redshift-matched sample of LEGRGs and HEGRGs [18]. LEGRGs exhibit lower Eddington ratios than HEGRGs, which signifies the inefficient radiative state with a low accretion rate of LEGRGs. HEGRGs, in contrast to LEGRGs, are rare and have a relatively high λ_{Edd} that shows a more radiatively efficient mode and a fast rate of accretion of these sources.

5.4 P-D Diagram

A P-D diagram depicts the relationship between the linear size (D) and radio power (P) of radio galaxies at a given frequency [65] [66]. Many giant radio sources appear to have low radio power at high projected linear sizes, as the P-D diagram clearly shows. The radiative energy loss and adiabatic expansion help in the reduction of radio luminosity. In certain situations, the radio-loud process may be switched off. The power of the radio lobes reduces sharply as the GRG size increases to a longer extent after the first rapid increment [67]. As the sources become older, they become less powerful and larger in scale. Figure 2 shows the P-D diagram of giant radio galaxies at 150 MHz.

5.5 GRGs in Galaxy Cluster

The huge, extensive linear size of GRGs is found in less dense galaxy cluster environments. GRGs are found in brightest galaxy cluster (BCG), which possibly implies the tracer of inhomogeneities in the intergalactic gas. This could be one of the important factors for the evolution and growth of these giant sources. The intergalactic gas controls the ongoing propagation of radio jets and also back-flow from hot-spots in the lobes of G RGs. These galaxies grow to a linear size of 0.71 to 0.88 Mpc despite being in a dense galaxy cluster [33]. A detailed study is needed to fully understand their basic and spectroscopic evolution.

A very small number of massive-GRGs are detected in BCGs [53] [68] [69]. This study shows the necessity of further study and investigation into alternative parameters in the external surroundings of GRGs.

5.6 High Redshift GRGs

In the TGSS GRG catalogue, 9 GRG candidates have a redshift ≥ 1, where the most distant GRG is J1612+5945 at a redshift of 4.30 [34]. In the previous GRG catalogue with an NVSS survey at 1.4 GHz, [33] inspected 27

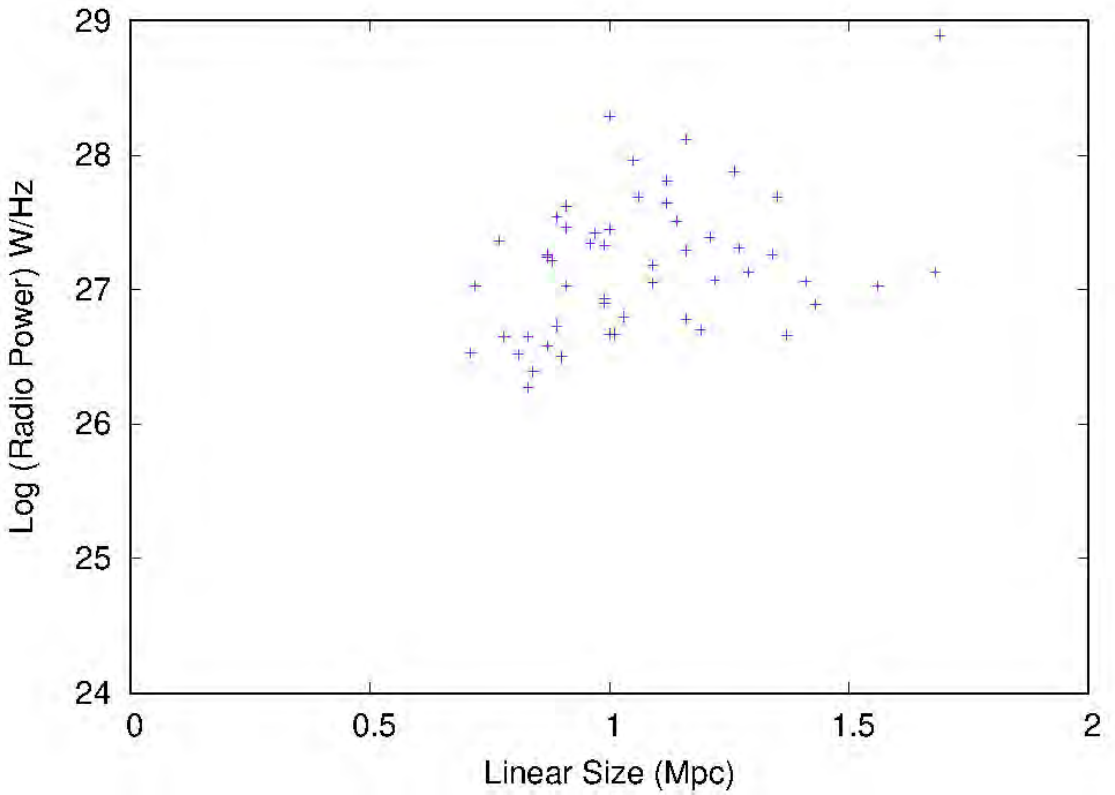


Figure 2: PD diagram of giant radio galaxies.

GRGs with a redshift of ≥ 1 where the maximum value of redshift was ~ 3.2200. The density (ρ_{medium}) of the intergalactic medium (IGM) increases with redshift as $(1+z)^4$. At high redshifts, the growth of the radio lobe can be obstructed significantly [70]. Moreover, the increasing value of redshift can affect the brightness of the surface as it shrinks as $(1+z)^{-4}$. This makes identification of GRG radio lobes difficult. Nevertheless, high resolution and sensitive radio surveys are very helpful to find high-redshift GRGs.

6 Summary

GRGs and GRGQs are both thought to be rare. Data from various radio surveys, such as the FIRST, TGSS ADR1, LOFAR, WENSS, and VLASS, were critical in identifying and studying a large number of GRGs. A summary of giant radio source properties is given below.

• The finding of a large number of new GRGs from current all-sky surveys gives the chance to make statistical studies of different properties of these enigmatic sources and helps to understand the causes of their enormous size and rarity.

• In the radio galaxy family, GRGs and GRQs have lobe-dominated extended structures, indicating a steep spectral energy distribution ($\alpha \sim 0.70$). The particle injection index and projected linear size of GRGs and GRQs are in identical patterns.

• The maximum GRGs have an FR-II morphology as observed in different radio catalogues (92 and 93 per cent in SGS and TGSS catalogue), and the remaining GRGs belong to FR-I, HyMoRS, and DDRG morphology.

• The jet kinetic power and radio power of GRQs are superior to those of GRGs. This clearly demonstrates that central engines of GRQs are more potent than GRGs. The minimum values of radio power at 1400 MHz and 150 MHz are 10^{23} WHz^{-1} and 10^{25} WHz^{-1} respectively.

• From the study of GRGs-AGN at mid-IR and optical frequencies, HEGRGs have higher jet kinetic power, radio power, and mass accretion rate but lower black hole mass.

• The natural tendency of giant radio sources is to avoid the galaxy cluster environment. It is uncommon to discover a giant radio sources in a extremely rich galaxy cluster (the optical mass $M_{200} > 2\times10^{14} M_{\odot}$). In LOFAR

and TGSS catalogues, only 21 (21/240) and one (1/54) GRGs are found in rich galaxy cluster environments.

References

[1] Jennison R. C. and Das Gupta, M. K., "Fine structure of the extra-terrestrial radio source Cygnus", Nature, 1953, 172, 996

[2] Baade, W. and Minkowski, R., "Identification of the radio sources in Cassiopeia, Cygnus A, and Puppis A", Astrophysical Journal, 1954, 119, 206

[3] Shklovskii, I. S., "On the discovery of extragalactic radio sources", Astronomicheskij Zhurnal, 1955, 32, 215

[4] Burbidge, G. R., "On Synchrotron Radiation from Messier 87", Astrophysical Journal, 1956, 124, 416

[5] De Young, D.S., "Extended extragalactic radio sources", Ann. Rev. of Astron. & Astrop, 1976, 14, 447–473

[6] Fanaroff, B.L. & Riley, J.M., "The morphology of extragalactic radio sources of high and low luminosity", Monthly Notices of the Royal Astronomical Society, 167, 31–35

[7] Lynden-Bell D., "Galactic nuclei as collapsed old quasars", Nature, 1969, 223, 690–694

[8] Schuch N., "The Ursa Major supercluster–I. The optical field and the 5C10 radio survey", Monthly Notices of the Royal Astronomical Society, 1981, 196, 695

[9] Condon J. J., Cotton W. D., Greisen E. W., Yin Q. F., Perley R. A., Taylor G. B., Broderick J. J., "The NRAO VLA sky survey", Astronomical Journal, 1998, 115, 1693–1716

[10] Bagchi J., et al., "Megaparsec relativistic jets launched from an accreting supermassive black hole in an extreme spiral galaxy", Astrophysical Journal, 2014, 788, 174

[11] Kaiser, C. R., and Alexander, P., "A self-similar model for extragalactic radio sources", Monthly Notices of the Royal Astronomical Society, 1997, 286, 215–222

[12] Mack, K.-H., Klein, U., O'Dea, C. P., Willis, A. G., and Saripalli, L. 1998, "Spectral indices, particle ages, and the ambient medium of giant radio galaxies", Astronomy and Astrophysics, 329, 431–442

[13] Ishwara-Chandra, C.H. and Saikia, D.J., "Giant radio sources", Monthly Notices of the Royal Astronomical Society, 1999, 309, 100–112

[14] Konar, C., and Hardcastle, M. J., "Particle acceleration and dynamics of double–double radio galaxies: theory versus observations", Monthly Notices of the Royal Astronomical Society, 2013, 436, 1595–1614

[15] Willis, A. G., Strom, R. G., and Wilson, A. S., "3C236, DA240; the largest radio sources known", Nature, 1974, 250, 625–630

[16] Schoenmakers, A. P., de Bruyn, A. G., Röttgering, H. J. A., and vander Laan, H., "A new sample of giant radio galaxies from the WENSS survey-I. Sample definition, selection effects and first results", Astronomy and Astrophysics, 2001, 374, 861–870

[17] Dabhade, P., Gaikwad, M., Bagchi, J., et al., "Discovery of giant radio galaxies from NVSS: radio and infrared properties", Monthly Notices of the Royal Astronomical Society, 2017, 469, 2886-2906

[18] Dabhade, P., Röttgering, H. J. A., Bagchi, J., et al., "Giant radio galaxies in the LOFAR Two-metre Sky Survey-I. Radio and environmental properties", Astronomy and Astrophysics, 2020, 635, A5

[19] Ursini, F., Bassani, L., Panessa, F., et al., "Hard X-ray-selected giant radio galaxies–I. The X-ray properties and radio connection", Monthly Notices of the Royal Astronomical Society, 2018, 481, 4250–4260

[20] Lara, L., Cotton, W. D., Feretti, L., et al., "A new sample of large angular size radio galaxies-I. The radio data", Astronomy and Astrophysics, 2001, 370, 409–425

[21] Hinshaw, G., Larson, D., Komatsu, E., et al., "Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: cosmological parameter results", Astrophysical Journal Supplement Series, 2013, 208, 19

[22] Planck Collaboration, Ade, P. A. R., Aghanim, N., et al., "Planck 2015 results-xiii. cosmological parameters", Astronomy and Astrophysics, 2016, 594, A13

[23] Becker, R. H., White, R. L., and Helfand, D. J., "The FIRST survey: faint images of the radio sky at twenty centimeters", Astrophysical Journal, 1995, 450, 559

[24] Rengelink, R. B., Tang, Y., de Bruyn, A. G., et al., "The Westerbork Northern Sky Survey (WENSS)-I. A 570 square degree Mini-Survey around the North Ecliptic Pole", Astronomy and Astrophysics, 1997, 124, 259

[25] Bock, D. C.-J., Large, M. I., and Sadler, E. M., "SUMSS: a wide-field radio imaging survey of the southern sky. I. Science goals, survey design, and instrumentation", Astronomical Journal, 1999, 117, 1578-1593

[26] McGilchrist, M. M., and Riley, J. M., "The 7c Survey of Radio Sources at 151-MHZ-a Search for Low-Frequency Variability", Monthly Notices of the Royal Astronomical Society, 1990, 246, 123

[27] Heald, G. H., Pizzo, R. F., Orrú, E., et al., "The LOFAR Multifrequency Snapshot Sky Survey (MSSS)-I. Survey description and first results", Astronomy and Astrophysics, 2015, 582, 214.02

[28] Intema, H. T., Jagannathan, P., Mooley, K. P., and Frail, D. A., "The GMRT 150 MHz all-sky radio survey-First alternative data release TGSS ADR1", Astronomy and Astrophysics, 2017, 598, A78

[29] Hurley-Walker, N., Callingham, J. R., Hancock, P. J., et al., "GaLactic and Extragalactic All-sky Murchison Widefield Array (GLEAM) survey - I. A low-frequency extragalactic catalogue", Monthly Notices of the Royal Astronomical Society, 2017, 464, 1146

[30] Shimwell, T. W., Röttgering, H. J. A., Best, P. N., et al., "The LOFAR Two-metre Sky Survey. I. Survey description and preliminary data release", Astronomy and Astrophysics, 2017, 598, A104

[31] Shimwell, T. W., Tasse, C., Hardcastle, M. J., et al., “The LOFAR Two-metre Sky Survey-II. First data release”, Astronomy and Astrophysics, 2019, 622, A1.

[32] Kellermann K. I., Verschuur G. L., “Galactic and extragalactic radio astronomy”, Galactic and extragalactic radio astronomy (2nd edition), 1988, 109, 163

[33] Kuźmicz, A., Jamrozy, M., Bronarska, K., et al., “An Updated Catalog of Giant Radio Sources”, The Astrophysical Journal Supplement Series, 2018, 238, 9

[34] Bhukta, N., Pal, S., Mondal, S. K., “Search for Megapersec Giant radio sources from TGSS”, arXiv:2201.12353, 2022

[35] Scheuer, P. A. G. and Williams, P. J. S. 1968, “Radio spectra”, Annual Review of Astronomy and Astrophysics, 6, 321–350

[36] Mahony, E. K., Morganti, R., Prandoni, I., et al., “ The Lockman Hole project: LOFAR observations and spectral index properties of low-frequency radio sources”, Monthly Notices of the Royal Astronomical Society, 2016b, 463, 2997–3020

[37] Oort, M. J. A., Steemers, W. J. G., and Windhorst, R. A., “The SPECFIND V2.0 catalogue of radio cross-identifications and spectra”, Astronomy and Astrophysics, 1998. 73, 103

[38] Gruppioni, C., Zamorani, G., de Ruiter, H. R., et al., “Radio observations of the Marano Field and the faint radio galaxy population”,Monthly Notices of the Royal Astronomical Society, 1997, 286, 470–482

[39] Kapahi, V. K., Athreya, R. M., van Breugel, W., et al., “The Molonglo Reference Catalog 1 Jy Radio Source Survey. II. Radio Structures of Galaxy Identifications”, Astrophysical Journal Supplement Series, 1998, 118, 275–326

[40] Ishwara-Chandra, C. H., Sirothia, S. K., Wadadekar, Y., et al., “ultrasteep spectrum radio source”, Monthly Notices of the Royal Astronomical Society, 2010, 405, 436–446

[41] McVittie, G.C., “Gravitational motions of collapse or of expansion in general relativity”, General Relativity and Cosmology, 1965, 15, 41

[42] Donoso E., Best P. N., Kauffmann G., “Evolution of the radio-loud galaxy population”, Monthly Notices of the Royal Astronomical Society, 2009, 392, 617–619

[43] Ledlow, M. J., and Owen, F. N., “A 20 cm VLA survey of Abell clusters of galaxies vi. radio/optical luminosity functions”, Astronomical Journal, 1996, 112, 9

[44] Aguado, D. S., Romina, A.; Andrés, A., et al., “The Fifteenth Data Release of the Sloan Digital Sky Surveys: First Release of MaNGA-derived Quantities, Data Visualization Tools, and Stellar Library”, Astrophysical Journal Supplement Series, 2019, 240, 23

[45] McConnell N. J., Ma C.-P., “Revisiting the scaling relations of black hole masses and host galaxy properties”, Astrophysical Journal, 2013, 764, 184

[46] Marconi A., Hunt L. K., “The relation between black hole mass, bulge mass, and near-infrared luminosity”, Astrophysical Journal, 2003, 589, L21-L24

[47] Graham A. W., “The black hole mass–spheroid luminosity relation”, Monthly Notices of the Royal Astronomical Society, 2007, 379, 711-722

[48] Kozlowski, “Empirical Conversions of Broad-Band Optical and Infrared Magnitudes to Monochromatic Continuum Luminosities for Active Galactic Nuclei”, Acta Astronomica, 2015, 65, 251–265

[49] Hardcastle, M. J., “A simulation-based analytic model of radio galaxies”, Monthly Notices of the Royal Astronomical Society, 2018b, 475, 2768–2786

[50] Heckman, T. M., Kauffmann, G., Brinchmann, J., et al., “Present-day growth of black holes and bulges: The Sloan Digital Sky Survey perspective”, Astrophysical Journal, 2004, 613, 109-118

[51] Hughes, S. A., and Blandford, R. D., “Black hole mass and spin coevolution by mergers”, Astrophysical Journal, 2003, 585, L101–L104

[52] Volonteri, M., Sikora, M., and Lasota, J., “Black Hole Spin and Galactic Morphology”, The Astrophysical Journal, 2007, 667, 704–713

[53] King, A. R., Pringle, J. E., and Hofmann, J. A., “The evolution of black hole mass and spin in active galactic nuclei”, Monthly Notices of the Royal Astronomical Society, 2008, 385, 1621–1627

[54] Daly, R. A., “Estimates of black hole spin properties of 55 sources”, Monthly Notices of the Royal Astronomical Society, 2011, 414, 1253–1262

[55] Blandford, R. D. and Znajek, R. L., “Electromagnetic extraction of energy from Kerr black holes”, Monthly Notices of the Royal Astronomical Society, 1977, 179, 433–456

[56] Blandford, R. D., R. D. Blandford, H. Netzer, L. Woltjer, T. J. L. Courvoisier, and M. Mayor, “Active Galactic Nuclei”, Astronomy-Astrophysics, 1990, 161–275

[57] Beskin, V. S., “Magnetohydrodynamic models of astrophysical jets”, Uspekhi Fizicheskikh Nauk, 2010, 53, 1199–1233

[58] Kumari, S., Pal, S., “Search for hybrid morphology radio galaxies from the FIRST survey at 1400 MHz”, Monthly Notices of the Royal Astronomical Society, 2022, 514, 4290-4299

[59] Kumari, S., Pal, S., “A catalogue of newly discovered Hybrid Morphology Radio (HyMoR) galaxies from the VLA FIRST survey”, Conference: 21st National Space Science Symposium; DOI: http://dx.doi.org/10.13140/RG.2.2.30624.66568, 2022, org; IISER Kolkata

[60] Kumari, S., Pal, S., Bhukta, N., Mondal, S. K., “Winged Radio Galaxies: An Overview”, Scientific Research Publishing, Advances in Modern and Applied Sciences, A Collection of Research Reviews on Contemporary Research, 2022, 1

[61] Ito, H., Kino, M., Kawakatu, N., Isobe, N., and Yamada, a., “The estimate of kinetic power of jets in FR II radio galaxies: existence of invisible components?”, The Astrophysical Journal, 2010, 685, 828–838

[62] Gürkan, G., Hardcastle, M. J., and Jarvis, M. J., "The Wide-field Infrared Survey Explorer properties of complete samples of radio-loud active galactic nucleus", Monthly Notices of the Royal Astronomical Society, 2014, 438, 1149–1161

[63] Kozieł-Wierzbowska, D. and Stasińska, G., "FR II radio galaxies in the Sloan Digital Sky Survey: observational facts", Monthly Notices of the Royal Astronomical Society, 2011, 415, 1013–1026

[64] Hardcastle, M. J., Evans, D. A., and Croston, J. H., "Hot and cold gas accretion and feedback in radio-loud active galaxies", Monthly Notices of the Royal Astronomical Society, 2007, 376, 1849–1856

[65] Baldwin J. E., Heeschen D. S., Wade C. M., eds, Extragalactic Radio Sources, 1982, 97, 21–24

[66] Bhukta, N., Pal, S., Mondal, S. K., "Search for X/Z Shaped Radio Sources from TGSS ADR 1", Monthly Notices of the Royal Astronomical Society, 2022, 4308-4323

[67] Blundell K. M., Rawlings S., "The inevitable youthfulness of known high-redshift radio galaxies", 1999, Nature, 399, 330–332

[68] Bagchi J., Kapahi V. K., "A VLA 20 and 90 centimetre radio survey of distant A-bell clusters with central cD galaxies", Journal of Astrophysics and Astronomy, 1994, 15, 275–308

[69] Best P. N., von der Linden A., Kauffmann G., Heckman T. M., Kaiser C. R., "On the prevalence of radio-loud active galactic nuclei in brightest cluster galaxies: implications for AGN heating of cooling flows", Monthly Notices of the Royal Astronomical Society, 2007, 379, 894–908

[70] Kapahi, V.K., "Redshift and luminosity dependence of the linear sizes of powerful radio galaxies", Astronomical Journal, 1989, 97, 1

‘Winged’ Radio Sources: A Peculiar Subclass of Radio Galaxy

Dusmanta Patra[1*]

[1]Department of Astrophysics and Cosmology, S. N. Bose National Centre for Basic Sciences,
Block-JD, Sector-III, Salt Lake, Kolkata-700106, India

*Dusmanta Patra: dusmanta.phy@gmail.com

Abstract

‘Winged’ or ‘X’-shaped radio galaxies (XRGs) are a small subclass of extra-galactic radio sources. These sources exhibit a pair of secondary low surface brightness radio lobes or wings oriented at an angle to the ‘active’, or primary high surface brightness lobes, resulting in the complete source an ‘X’ shape. In some cases, the set of secondary lobes comes from the edges of the primary lobes, giving a ‘Z’-symmetry. We provide an overview of the observational results that provide us with the census of these peculiar ‘winged’ radio sources. Very Large Array Faint Images of the Radio Sky survey at Twenty centimeter is one of the promising database to detect these sources. Along with the radio emission, optical and X-ray emissions are detected from these sources. The nature of ‘X’-shaped sources is a matter of considerable debate. Several authors have prescribed different models to explain the mechanism for the peculiar morphology of these sources. Backflow of plasma from the hot spots of the active lobes into the surrounding medium is one of the strong possibilities for such exotic structures. The merger event of two supermassive black holes is another possible cause for the origin behind ‘winged’ radio sources. They have even been proposed to provide evidence for black hole mergers/spin reorientation, and therefore constrain the rate of strong gravitational wave events. We briefly summarize the few models and discuss their drawbacks. A more detailed morphological and spectral results on milliarcsecond-scales is needed to provide a crucial test of these model.

Keywords: *Active Galactic Nuclei (AGN), Jets, Quasars, Radio Continuum Emission, Catalogs, Surveys*

1 Introduction

Radio galaxies (RGs) are the largest observable astronomical object in the sky. The overall linear sizes of RGs range from less than a few tens of parsecs to over a Mpc distance. RGs are lobe-jet structure objects that harbor active galactic nuclei (AGN) at the center between the lobes. There is narrow collimated features called jets connecting the core to the extended components and the outer lobes, which are the signatures of the beams carrying energy from the core to the outer lobes. For a typical double-lobed radio galaxy, the jets are directed opposite to the central AGN. Based on the radio brightness distribution, radio galaxies are categorized into two major classes; one is Fanaroff-Riley class I (FR-I), where the central region of the radio galaxy is brighter than both the edges and another is Fanaroff-Riley class II (FR-II), where brightness increases from the center towards the edges [1]. FRI RGs have higher radio flux density in the center than the outer edges of the lobes. On the other hand, FR II sources have edge-brighten morphology, i.e., the outer edges have higher radio flux than the inner regions.

‘Winged’ or ‘X’-shaped radio galaxies (XRGs) are a small subclass of extra-galactic radio sources that exhibit a pair of secondary low surface brightness radio lobes or wings oriented at an angle to the ‘active’, or primary high surface brightness lobes, resulting in the total source an ‘X’ shape (see Figure 1) [2, 3, 4, 5]. In some cases, the set of secondary lobes comes from the edges of the primary lobes, gives a ‘Z’-symmetry [5], and these sources are known as ‘Z’-shaped radio galaxies (ZRGs). There is another class called Head-Tail or ‘C’-shaped radio sources [6, 7], where two jets are bent in the same direction. Morphologically RGs can be classified into two types: FR I and FR II [1]. Some radio galaxy obeys some mixed type of Fanaroff-Riley dichotomy, known as Hybrid Morphology Radio Sources (HyMoRS) [8], having FR-I morphology in one side of the active nucleus and FR-II morphology on another side.

3C 272.1 is the first reported ‘winged’ radio galaxy [9] , showing a ‘Z’-like structure. Ekers et al. 1978 found that NGC 326 also showed a ‘Z’-like structure with possible precessing beams. Later Kotanyi (1990) identified NGC 3309 as an S-shaped source. A source with an ‘X’-shaped structure was first classified by [2]. Cheung (2007) searched for XRG candidates using the Very Large Array (VLA) Faint Images of the Radio Sky at Twenty-centimeters (FIRST) [10] survey and identified 100 candidates. A color thumbnail image of the sample of 100 FIRST ‘X’-shaped candidates adopted from Cheung (2007) is presented in Figure 2. Proctor (2011) also identified 156 XRG candidates from the FIRST survey. Later Yang et al. (2019) found 290 ‘winged’ radio sources from the FIRST survey. Bera et al. (2020) also continued the search for ‘winged’ radio sources from the FIRST survey and found 296 ‘winged’ radio sources, out of which 161 are XRGs, and 135 are ZRGs.

The reason behind the peculiar morphology of ‘winged’ radio is a debatable subject. Several authors have prescribed different models to explain the mechanism for the peculiar morphology of these sources. Leahy & Williams 1984 and Capetti et al. 2002 have proposed the backflow of plasma model. The merger event of two supermassive black holes

(SMBHs) is another possible cause for the origin behind ‘winged’ radio sources [11, 12]. Reference [13] suggested that the realignment of a central SMBH accretion disk system is also responsible for forming secondary lobes in XRGs. The precession of twin jets models [14] may be another probable reason for the formation of these sources. A summary regarding all of the above models is presented in Section 5. However, none of these models are self-sufficient to explain each property of all reported XRGs.

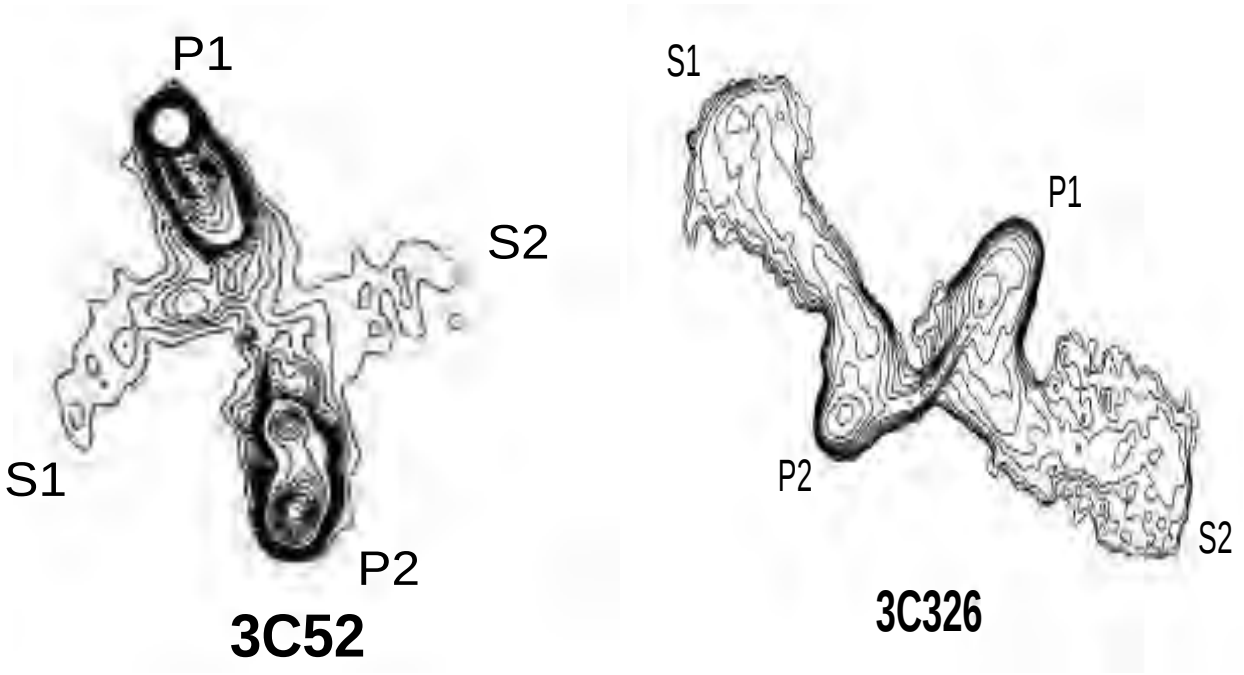


Figure 1: VLA radio maps of two prominent XRGs are shown: 3C 52 (Leahy & Williams 1984) in the left panel and NGC 326 (Murgia et al. 2001) in the right panel. The primary lobes are marked with P1 and P2, while secondary lobes are marked with S1 and S2.

2 The Observation of ‘Winged’ Radio Sources

This section provides an overview of the observational results that provide us with the census of ‘winged’ radio sources. Different authors carry several observations from radio to X-ray wavelength to understand the nature of these sources.

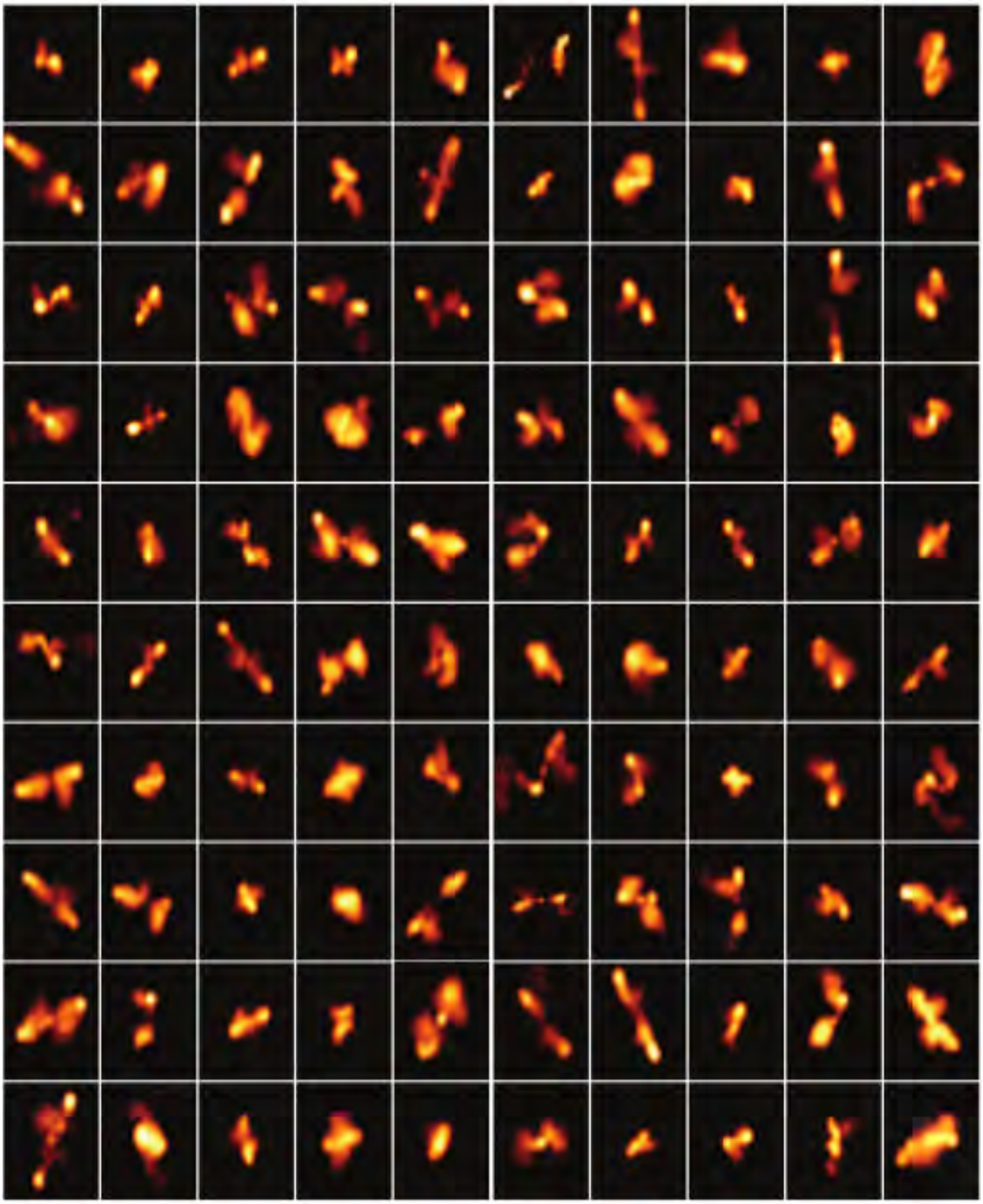


Figure 2: Color thumbnail images of the 100 FIRST 'X'-shaped radio source candidates from Cheung et al. (2007) showing the diffuse emission from the objects.

2.1 A Brief History of the Searches

3C 272.1 is the first reported 'winged' radio galaxy (Riley 1972), showing a 'Z'-like structure at 2.7 and 5.0 GHz maps. Ekers et al. 1978 found that NGC 326 also showed a 'Z'-like structure at a 5 GHz radio map with possible precessing beams. Reference [15] found that radio maps of some of their objects show that two-thirds of the bridges are deformed near the central galaxy. Among 39 samples of their study, 3C315, 3C136.1, and 3C52 showed 'X'-shaped morphology. Later [16] identified NGC 3309 with an S-shaped morphology. Reference [2] reported 11 sources with wings and classified them as XRGs. Reference [3] did a literature survey for XRGs and reported another eight XRGs, increasing the sample to 19 XRGs. A systematic study was done by [3] to look for 'winged' radio galaxies and identified 100 new XRG candidates using the FIRST survey. The samples of 100 sources by [3] are presented in Figure 2. Out of the 100 XRGs, they found optical identifications for 94 candidates. Reference [17] presented the lowest-frequency images of 11 known 'X'-shaped

sources at 240 and 610 MHz using Giant Metrewave Radio Telescope (GMRT) and studied the spectral indices properties across the sources. Reference [18] identified 156 XRG candidates using an automated morphological classification method from the FIRST radio survey. Twenty-one of these were previously reported in [3], and one source, 3C 315, is a well-known XRG, giving 134 new XRG candidates. Reference [4] presented a catalog of 290 XRGs extracted from the FIRST survey. Out of the 290 candidates, they classified 106 as strong XRG candidates and 184 as probable XRG candidates. These probable XRGs need to be verified by further observations. Another search for XRGs from the FIRST survey was done by [5] , and they cataloged 296 'winged' radio sources, out of which 161 are XRGs, and 135 are ZRGs.

Reference [19] studied the 'X'-shaped RG, SDSS J1130+0058 in optical wavelength. They found double-peaked low-ionization broad emission lines (see Figure 3), supporting the twin-AGN model formation scenario. The first time optical spectral properties of XRGs using a sample of $\sim$ 50 sources was studied by [20]. They found that the XRGs population is composed about equally of sources with weak and strong emission-line spectra, which means that they typically have radio powers between those of FR I and FR II types RGs. Reference [20] also found that the nuclear regions of XRGs are relatively hot. Reference [21] identified the optical host galaxy and found their position for a sample XRGs. The orientation of the wings shows a strong connection with the optical axis, providing support for a hydro-dynamic origin of the formation of the radio wings [21].

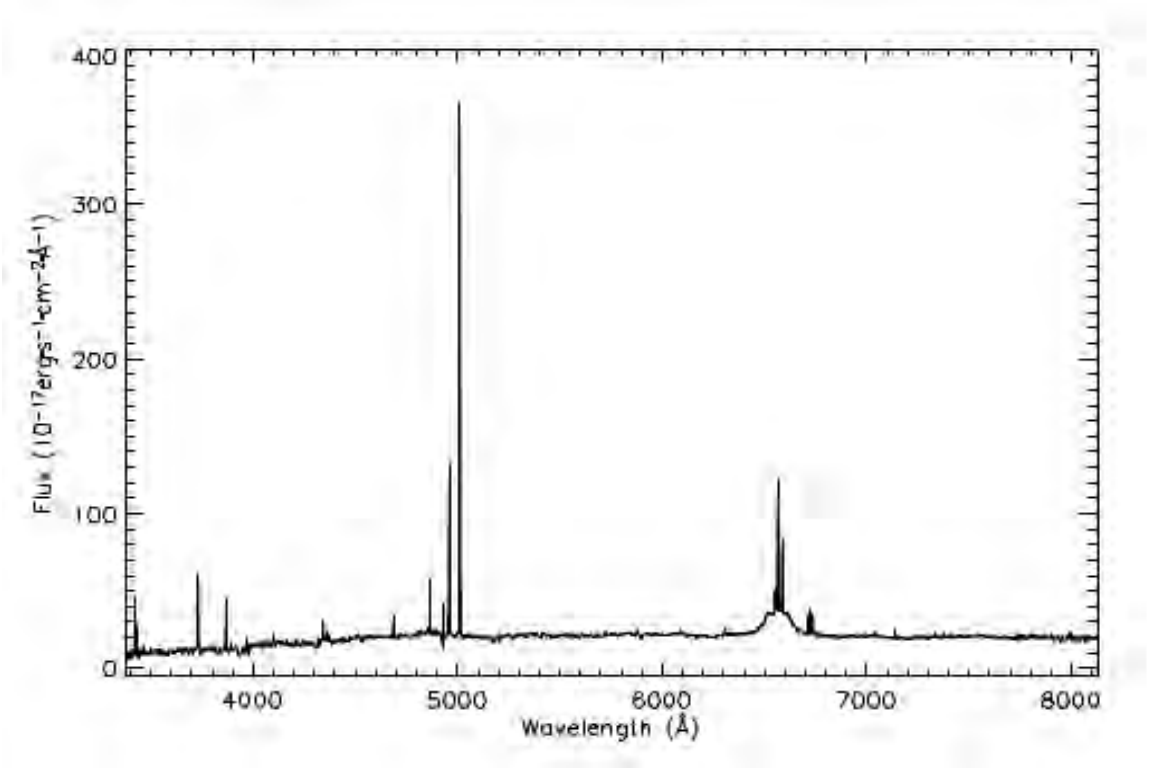


Figure 3: The observed spectrum of SDSS J1130+0058 (Zhang et al., 2007).

2.2 X-Ray Observation of 'Winged' Radio Sources

X-ray emissions from RGs are a crucial factor in studying their gaseous medium and helps to understand better the mechanisms related to the dynamics of the radio structures. Observations of radio galaxies by Chandra and XMM-Newton have revealed non-thermal X-ray emission from their jets and hot spots, as well as the emission from thermal coronae. Reference [22] presented the first Chandra observation of an 'X'-shaped radio galaxy 3C 403 to determine the relationship between the gas emitting from the X-ray and the 'X'-shaped radio morphology.

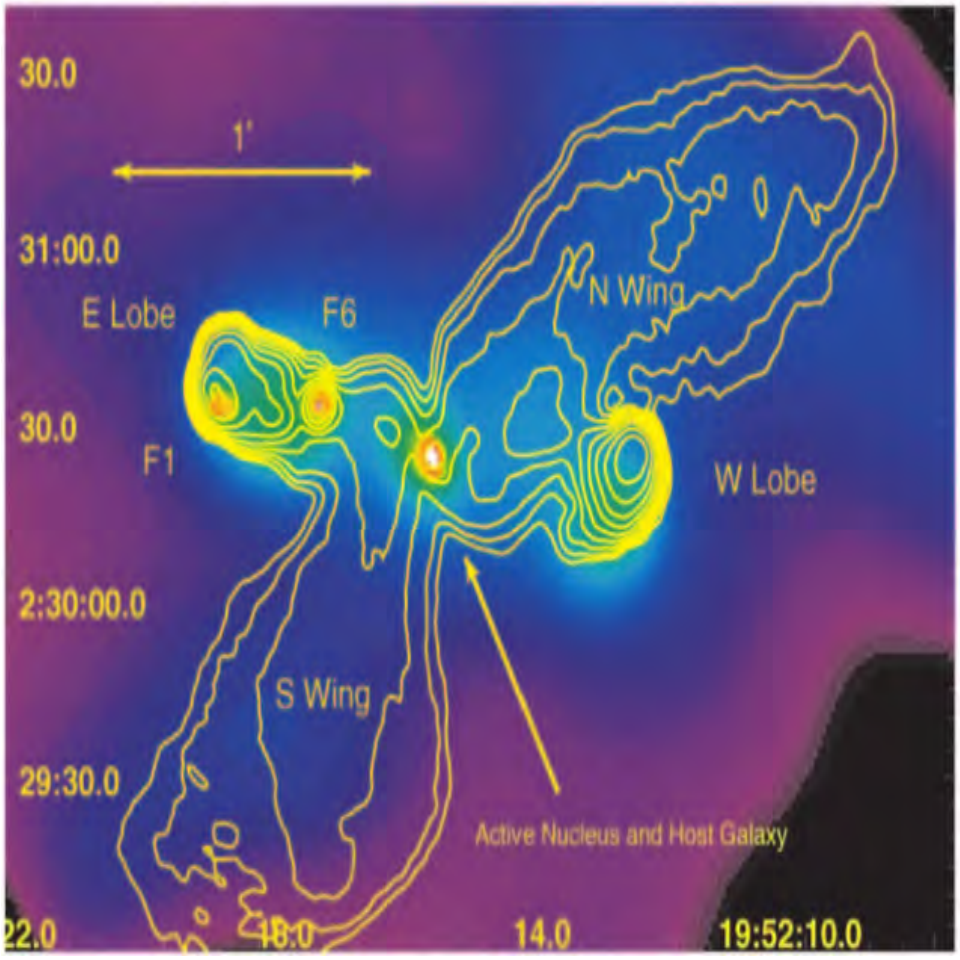


Figure 4: A Chandra ACIS-S image of 3C 403 in the 0.5–2.0 keV band overlaid with 8.4 GHz radio contours (Kraft et al., 2005). Emission from the active nucleus, the hot ISM, compact radio components, and diffuse emission from the lobes and wings are detected.

Chandra ACIS-S image of the XRG 3C 403 in the 0.5–2.0 keV band overlaid with VLA 8.4 GHz radio contours of 'X'-shaped radio galaxy 3C 403 is shown in Figure 4 [22]. They found diffuse emission from the lobes and wings and emission from the active nucleus, the hot ISM, and several of the compact radio components. They have also detected X-ray and optical emissions from radio knots and hot spots of the source. The X-ray emission flux detected from the radio wings is consistent with inverse Compton scattering of cosmic microwave background (CMB). According to the radio/optical/X-ray spectra, they attribute the emission in all cases to the synchrotron emission rather than inverse Compton scattering of CMB. Reference [23] presented Chandra X-ray observations of XRGs within redshift 0.1 and compared the result with the sample of standard double-lobed FR I and II RGs. They found that the ellipticity and position angle of the hot gas follows that of the stellar light distribution for radio galaxy hosts by fitting elliptical distributions to the observed diffuse hot X-ray emitting atmospheres. Reference [24] examine NGC 326, one of the most prominent 'X'- or 'Z'-shaped radio galaxies, with a 100 ks Chandra X-Ray Observatory exposure and find several features associated with the RG.

3 Morphology of 'X'-Shaped Radio Sources

About 10% of the FR II radio galaxies have an unusual morphology of two misaligned pairs of radio lobes, and these objects are classified as 'winged' radio sources (e.g., [2]). In Figure 1, color thumbnail images of 100 'winged' radio sources are presented by [3] showing the diffuse emission from the objects. These misaligned pairs of radio wings in radio galaxies appear as the secondary lobes. These two pairs of misaligned radio lobes are approximately equal in linear extent. These secondary radio lobes are aligned at a reasonably large angle from the primary radio axis defined by the primary lobe pair. VLA radio images of two prominent XRGs are shown in Figure 1. In left panel 3C 52 [15] and, in the right panel, NGC 326 [25] are shown. The two primary lobes of these XRGs are marked with 'P1' and 'P2', while the secondary lobes are marked with 'S1' and 'S2'. These types of radio galaxies have low radio luminosities, generally lying near the FR I/FR II division. The maximum number of these sources are of FR II type [1], and the remaining are either FR I or mixed [12].

Depending on the position of the wings, 'winged' radio sources are classified into XRGs and ZRGs [5]. XRGs exhibit a pair of secondary low surface brightness radio lobes or wings oriented at an angle to the 'active' or primary high surface brightness lobes, resulting in the complete source an 'X' shape. In this scenario, the wings are coming out from near the central region of the primary lobe. In some cases, the set of secondary lobes or wings from the edges of the primary lobes gives a 'Z'-symmetry, and these sources are ZRGs [26]. In some cases, the latter group of wing sources shows a 'S'-like morphology also.

4 Spectral Index Distribution

Studies of the radio spectral index along the lobes and wings and from comparisons their value between the lobes and wings, provides us a better insights to understand their physical properties. The variation of radio spectral index is typically observed to be steeper in the wings than in the primary lobes, suggesting that the wings have older radio emission than the primary lobes. (e.g. [11, 13, 17, 25]. The spectral-index map of the 'X'-shaped RG PKS 2014–55 is shown in Figure 5 [27], showing the variation of spectral index is steeper in the wings than in the primary lobes . Reference [28] studied sample of 28 'X'-shaped radio sources using the GMRT at 610 and 240 MHz and presented the spectral index map of these sources. Based on the spectral index distribution they categorized these sample into three groups:
(i) the wings have flatter spectra as compared to the primary active lobes,
(ii) the wings and primary lobes have comparable spectral index,
(iii) the wings have steeper spectra compare to the primary lobe.

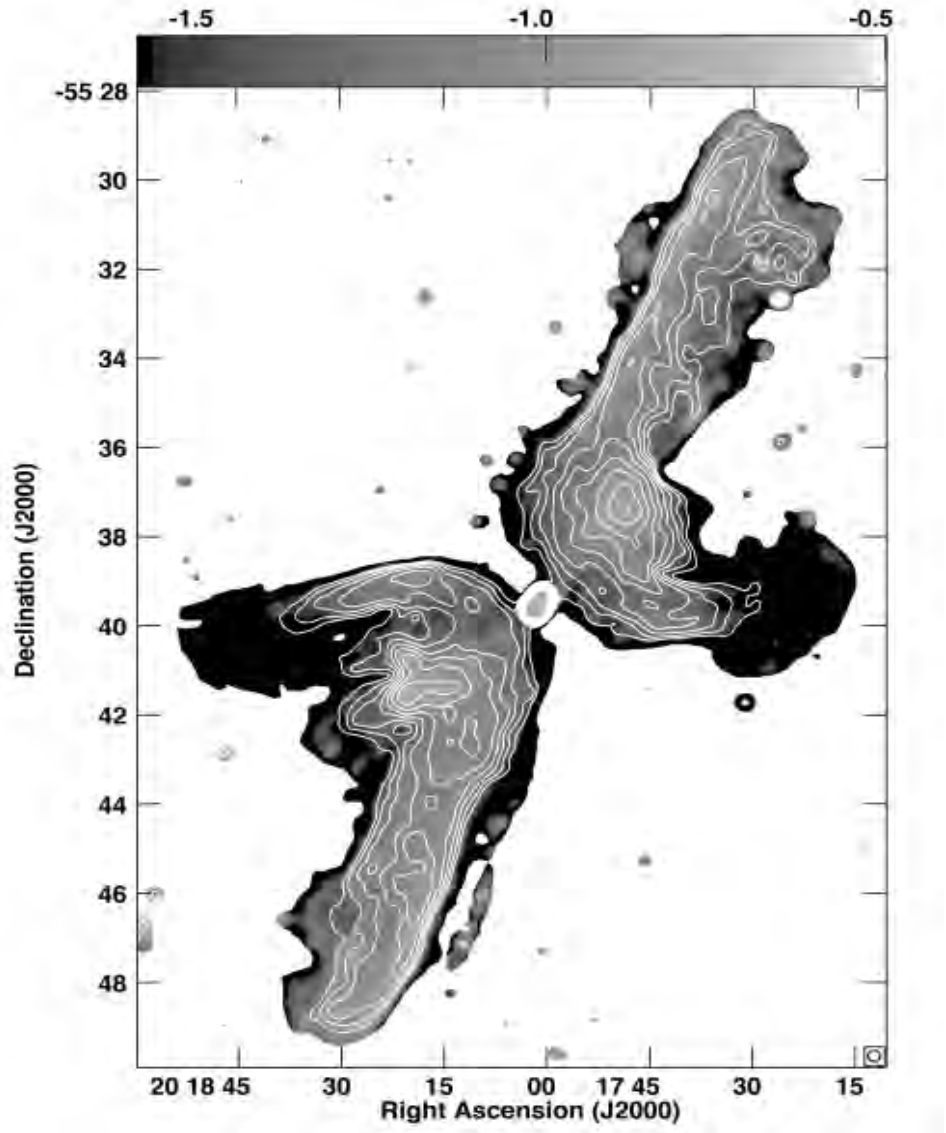


Figure 5: The greyscale image shows the spectral index variation of the XRG PKS 2014–55 with values given by the top scale bar (Cotton et al. 2020).

5 Formation of 'X'-Shaped Radio Sources

The origin behind these peculiar structures is still not clear to us. There are few models available in different literature which explain the possible reason for these peculiar morphologies. In this section we briefly discuss about the models, that have been proposed for the XRG phenomenon.

5.1 Backflow

One possibility is that the wings emerge from the backflow of plasma from the hot spots of the active lobes into the surrounding medium [15, 29]. Backflow is produced by jet material released by the hotspots and flows back towards the

host galaxy. In their model, [15] proposed that this back-flowing jet material remains straight until it collides with the opposite backflow and spread out laterally in the perpendicular direction to the radio lobe axis (see Figure 6). They also proposed two symmetry-breaking processes, which might cause the bending of backflow in the opposite direction. The first one is that the back-flowing material is deflected by a spheroidal gas distribution misaligned with the radio axis. The second mechanism proposed that an existing cavity in the interstellar medium provides a path for backflow to flow preferentially.

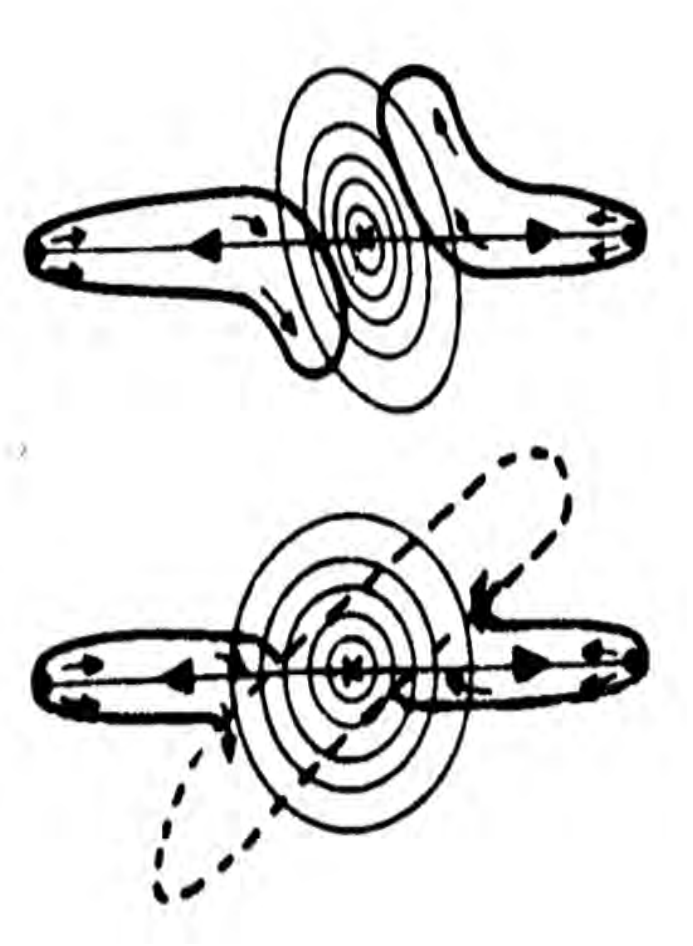


Figure 6: Sketch of two symmetry breaking mechanisms proposed by Leahy & Williams (1984).

Hodges-Kluck & Reynolds (2011) Reference [30] have conducted hydrodynamic simulations of light, hypersonic jets to study the viability of the backflow model for the formation of wings in 'X'-shaped radio galaxies. Their simulation showed that the wings are produced considerably by the deflection of the back-flowing plasma from the hot spots. They also argued that their models show that lobe interaction with the hot atmosphere may have a significant influence in defining the morphology of these radio galaxies.

5.2 Buoyancy

Usually, the lobes of a radio galaxy have less density compare to the surrounding ISM/IGM [31]. As a result, buoyancy may be responsible for the large-scale shape of the radio lobes. Sometimes buoyant forces are responsible for the bending of the lobes towards the ISM/IGM that maintain the density equilibrium between the lobe and the surrounding medium. Therefore, buoyancy can form such 'X'-shaped structures [30, 32, 33]. [33] could not explain the development of 'X'-shaped radio galaxies NGC 326 using the buoyancy model. The primary and secondary lobes of an 'X'-shaped source have an angle $\pm 10^{o}$ between them, and if the buoyancy is the dominating creation process, one would expect a more random distribution of these angles [11]. Reference [17] concluded that buoyancy may act upon the large-scale morphology of radio galaxies near the dense cluster environments as is seen for head-tail radio sources in a cluster of galaxies, and it is unlikely that buoyancy, without a suitable configuration of the ISM or IGM, would influence the morphology of XRGs.

5.3 Twin-AGN Jet Model

NGC 326 is a well-known XRG associated with a dumbbell galaxy, i.e., two roughly equally luminous ellipticals within a common envelope [33, 34]. This NGC 326 phenomenon inspired the twin-AGN model concept. In this model, it is considered that the set of twin radio lobes to be two distinct radio doubles associated with a nearby pair of active SMBHs inside a merging pair of giant ellipticals [17]. In this illustration, the interacting neighbor could be in a very eccentric

orbit or a circular orbit. In the first scenario, an inversion-symmetric distortion of the jet pair happens due to an impulsive gravitational interaction, leading to an 'X'-shaped or 'Z'-shaped morphology. In the second situation, the continued tidal interaction can generate periodic inversion-symmetric wiggles in the radio jet pair, leading to a helical radio structure. A warped and slanted accretion disc is often expected to form in SMBH binaries, resulting in an 'X'-shaped radio morphology [35]. Zhang et al. 2007 discovered double-peaked emission lines in the optical spectra of a few XRGs, which supports this twin-AGN scenario also.

The twin-AGN model appears especially promising if the dynamical friction is minimal, causing the two SMBHs to delay their approach. On the other hand, such delay is unusual [36]. Furthermore, because this scenario requires both central engines to launch jets simultaneously, it has an extremely low probability. This twin-AGN jet model cannot explain why none of the XRGs have both FR II type lobe pairs. Furthermore, it does not explain why the primary lobe prefers to be oriented along the major axis of the optical host galaxy [29, 32]. In addition, it is difficult in this scenario to understand the 'Z'-symmetric morphology of the secondary lobes [26]. On these arguments, it is clear that, while this model may account for a small percentage of XRGs, it is incapable of being defended for the vast majority of XRGs.

5.4 Reorientation of Jet

A natural way to understand the existence of the wings in XRGs is a change of jet axis at a past time during the lifespan of the source. The wings in XRGs are the relic emission of the previously active radio lobe developed during a previous phase of nuclear activity once the jets were pointing in that direction. After that, the jet orientation changed quickly, probably in an abrupt flip, resulting in the presently feeding primary lobes pair. Precession or other realignment mechanisms have been proposed in different models to explain jet reorientation [13, 37, 38, 39].

Reference [26] argued that the spin-flip scenario could also easily account for the 'Z'-symmetry morphology of these sources.provide an MWE for your .tex file and .bib file. They also demonstrated how it might occur before the spin-flip as the jets propagated through the ISM of the huge elliptical. A study of NGC 3801, a 'Z'-shaped FR I RG, finds solid evidence for a recent merger and a large rapidly rotating gas disk interacting with the jets [40].

6 Summary

This article gives brief information about the peculiar class of 'X'-shaped radio galaxies. In the above sections, we have tried to understand the XRGs through the observation aspects. XRGs are a class by themselves, and the formation mechanism is unlike normal RGs. There are many 'winged' radio sources in the sky, and detection effects keep us to find more of them. From different literature surveys, we give an idea about the census of 'winged' radio sources till now. FIRST survey is one of the promising database to detect these sources.

We have discussed the morphology of these sources as observed by the different authors. These types of radio galaxies have low radio luminosities, generally lying near the FR I/FR II division. The definition of XRGs and ZRGs is discussed based on morphology. XRGs exhibit a pair of secondary low surface brightness radio lobes or wings oriented at an angle to primary high surface brightness lobes, resulting in the complete source an 'X' shape. In this scenario, the wings are coming out from near the central region of the primary lobe. On the other hand, the wings pair emerges from the edges of the primary lobes, gives a 'Z'-symmetry, and these sources are known as ZRGs. Diffuse X-ray emission from the lobes and wings of XRG 3C 403 and emission from the active nucleus, the hot ISM, and several of the compact radio components are detected [22]. In general, the variation of radio spectral index was noticed to be steeper in the wings than in the primary lobes. This typical variation suggests that the wings are older than the primary lobes.

We also compare various available clues from different literature to arrive at a better understanding of the mechanism responsible for the XRGs. We briefly discuss the following formation models of XRGs. Backflow, buoyancy, twin-AGN jet model, and reorientation of the jet axis are discussed. We present strong evidence against these models based on the different literature surveys. The drawbacks are also discussed based on the previous study. The orientation of the wings shows a strong connection with the optical axis of the host galaxy, providing support for a hydro-dynamic origin of the formation of the radio wings. Further milliarcsecond-scales radio observation (e.g. VLBA observation) and optical and X-ray observation are needed to understand the formation scenario of these 'winged' radio sources.

Acknowledgements

The author acknowledges the post-doctoral fellowship of S. N. Bose National Centre for Basic Sciences.

References

[1] B. Fanaroff and J. M. Riley, "The morphology of extragalactic radio sources of high and low luminosity," *MNRAS*, vol. 167, no. 1, pp. 31–36, 1974.

[2] J. P. Leahy and P. Parma, "Multiple outbursts in radio galaxies."

[3] C. C. Cheung, "First "winged" and x-shaped radio source candidates," *AJ*, vol. 133, no. 5, pp. 2097–2121, 2007.

[4] X. e. a. Yang, "Extended catalog of winged or x-shaped radio sources from the first survey."

[5] B. S., P. S., S. T. K., and M. S., "First winged radio galaxies with x and z symmetry," *ApJS*, vol. 251, p. 9, 2020.

[6] L. Rudnick and F. N. Owen, "Head-tail radio sources in clusters of galaxies."

[7] P. D., P. S., K. C., and C. S. K., "Multi-frequency properties of an interacting narrow-angle tail radio galaxy j0037+18."

[8] Gopal-Krishna and P. J. Wiita, "Extragalactic radio sources with hybrid morphology: implications for the fanaroff-riley dichotomy."

[9] J. M. Riley, "Observations of 3c 272.1 at 2.7 and 5.0 ghz."

[10] R. H. Becker, R. L. White, and D. J. Helfand, "The first survey: Faint images of the radio sky at twenty centimeters," *ApJ*, vol. 450, p. 559, 1995.

[11] H. Rottmann, "Jet-reorientation in x-shaped radio galaxies."

[12] D. Merritt and R. D. Ekers, "Tracing black hole mergers through radio lobe morphology."

[13] J. Dennett-Thorpe, P. A. G. Scheuer, and R. A. e. a. Laing, "Jet reorientation in active galactic nuclei: two winged radio galaxies," *MNRAS*, vol. 330, no. 3, pp. 609–620, 2002.

[14] K. H. Mack, L. Gregorini, P. Parma, and U. Klein, "High-frequency radio continuum observations of radio galaxies with low and intermediate luminosity. ii. sources with sizes 4' to 5'."

[15] J. P. Leahy and A. G. Williams, "The bridges of classical double radio sources."

[16] C. Kotanyi, "Ngc 3309: an s-shaped radio galaxy in a nearby cluster."

[17] D. V. Lal and P. R. A., "Giant metrewave radio telescope observations of x-shaped radio sources."

[18] D. D. Proctor, "Morphological annotations for groups in the first database."

[19] X.-G. Zhang, D. Dultzin-Hacyan, and T.-G. Wang, "Sdss j1130+0058 an x-shaped radio source with double-peaked low-ionization emission lines: a binary black hole system?"

[20] H. Landt, C. C. Cheung, and S. E. Healey, "The optical spectra of x-shaped radio galaxies."

[21] M. Gillone, A. Capetti, and P. Rossi, "Origin of x-shaped radio-sources: further insights from the properties of their host galaxies," *A&A*, vol. 587, pp. A25–A38, 2016.

[22] R. P. Kraft, M. J. Hardcastle, D. M. Worrall, and S. S. Murray, "A chandra study of the multicomponent x-ray emission from the x-shaped radio galaxy 3c 403."

[23] E. J. Hodges-Kluck, C. S. Reynolds, C. C. Cheung, and M. C. Miller, "The chandra view of nearby x -shaped radio galaxies."

[24] E. J. Hodges-Kluck and C. S. Reynolds, "A chandra study of the radio galaxy ngc 326: Wings, outburst history, and active galactic nucleus feedback."

[25] M. Murgia, P. Parma, and e. a. de Ruiter, H. R., "A multi-frequency study of the radio galaxy ngc 326."

[26] Gopal-Krishna, P. L. Biermann, and P. J. Wiita, "The origin of x-shaped radio galaxies: Clues from the z-symmetric secondary lobes," *ApJ*, vol. 594, no. 2, p. L103, 2003.

[27] C. et al., "Hydrodynamical backflow in x-shaped radio galaxy pks 2014–55," *MNRAS*, vol. 495, pp. 1271–1283, 2020.

[28] D. V. Lal, B. Sebastian, C. C. Cheung, and P. R. A., "Gmrt low-frequency imaging of an extended sample of x-shaped radio galaxies."

[29] A. Capetti, S. Zamfir, and P. e. a. Rossi, "On the origin of x-shaped radio-sources: New insights from the properties of their host galaxies," *A&A*, vol. 394, no. 1, pp. 39–45, 2002.

[30] E. J. Hodges-Kluck and C. S. Reynolds, "Hydrodynamic models of radio galaxy morphology: Winged and x-shaped sources."

[31] A. G. Williams, "Numerical simulations of radio source structure."

[32] L. Saripalli and R. Subrahmanyan, "The genesis of morphologies in extended radio sources: X-shapes, off-axis distortions, and giant radio sources."

[33] D. M. Worrall, M. Birkinshaw, and R. A. Cameron, "The x-ray environment of the dumbbell radio galaxy ngc 326."

[34] P. Battistini, F. Bonoli, and S. e. a. Silvestro, "The orientation of radiosources associated with elliptical galaxies," *A&A*, vol. 85, no. 1-2, pp. 101–105, 1980.

[35] M. C. Begelman, R. D. Blandford, and M. J. Rees, "Massive black hole binaries in active galactic nuclei," *Nature*, vol. 287, pp. 307–309, 1980.

[36] L. A. Gergely and P. L. Biermann, "The spin-flip phenomenon in supermassive black hole binary mergers," *ApJ*, vol. 697, pp. 1621–1633, 2009.

[37] R. D. Ekers, R. Fanti, C. Lari, and P. Parma, "Ngc326- a radio galaxy with a precessing beam?" *Nature*, vol. 276, pp. 588–590, 1978.

[38] U. Klein, K. H. Mack, L. Gregorini, and P. Parma, "High-frequency radio continuum observations of radio galaxies with low and intermediate luminosities. iii. spectral indices and particle ages."

[39] D. Falceta-Gonçalves, A. Caproni, Z. Abraham, D. M. Teixeira, and E. M. de Gouveia Dal Pino, "Precessing jets and x-ray bubbles from ngc 1275 (3c 84) in the perseus galaxy cluster: A view from three-dimensional numerical simulations," *ApJ*, vol. 713, no. 1, pp. L74–L78, 2010.

[40] A. Hota, J. Lim, Y. Ohyama, D. J. Saikia, D. v Trung, and J. H. Croston, "A multiwavelength study of a young, z-shaped, fr i radio galaxy ngc 3801."

Winged Radio Galaxies: An Overview

Shobha Kumari[1,*], Sabyasachi Pal[1], Netai Bhukta[2], Sushanta K. Mondal[2]

[1]Midnapore City College, Kuturia, Bhadutala, Paschim Medinipur, West Bengal, 721129, India
[2]Department of Physics, Sidho Kanho Birsha University, Purulia, 723104, India
*Corresponding author: shobhakumari@mcconline.org.in

Abstract

Winged radio galaxies are known to be a mysterious and weird set of radio galaxies with complex structures. In winged radio galaxies, there are a pair of secondary lobes (wings) misaligned with the primary lobes. Depending on the location of wings in the structure, these radio galaxies are classified as *X*-shaped radio galaxies (XRGs) and *Z*-shaped radio galaxies (ZRGs). In XRGs, secondary lobes (wings) appear to emerge from near the centre of the galaxy and, most of the time, move nearly perpendicular to the primary lobes. In ZRGs, secondary lobes (wings) appear to emerge from near the edge of the primary lobes by making some angle with the primary lobes. To explain the formation mechanism of wing structures in winged radio galaxies, various models have been proposed. According to these models, the origin of winged radio galaxies may be due to the backflow of plasma, jet re-orientation due to the SMBH merger, and dual active galactic nuclei (AGN) present in the core of the galaxies, etc. Each source of winged radio galaxies has its own mystery, and none of these models is capable of explaining the structures of all winged radio galaxies. Statistical and physical properties of winged radio galaxies in comparison to normal radio galaxies are studied here

Keywords: *Radio Galaxies, Winged Radio Galaxies, Active Galactic Nuclei (AGN).*

1 Introduction

Radio galaxies are well known to have a compact radio nucleus coincident with the core of the host galaxy from which a pair of opposite jets emerge. This jet pair travels a large distance in the order of a few kpc to Mpc, which makes the structure of an extended radio galaxy. It is believed that a supermassive black hole (SMBH) in the galactic nucleus is the energy source of the radio galaxy from which the jets seem to emerge. The SMBH actively accumulates the gas and dust from its surroundings. This accretion of gas and dust helps to launch high-energy jet streams, which have the power to accelerate charged particles away from the supermassive black hole almost to the speed of light. Radio wavelengths allow for a clear observation of these jet streams. The jets of radio galaxies can be many times larger than the optical galaxy (see Figure 1). Radio galaxies were discovered in the era of the 1940s by some radar handling engineers, but it took another decade to better understand them. The first radio galaxy to be identified was Cygnus A, which is still among the brightest radio sources discovered in the sky.

In 1974, Fanaroff-Riley (FR) classified radio galaxies into two groups: Fanaroff-Riley type I (FR I) and Fanaroff-Riley type II (FR II) [24]. Initially, radio galaxies were classified on the basis of the FR index (the ratio of the separations of the two brightest hotspots on either side of the host galaxy to the farther extension of the source). The sources are called FR I if this FR index value is less than 0.5 and FR II if this ratio is greater than 0.5. It is found that FR I radio galaxies are very prominent and have clear jets emerging from the centre of the host galaxy. FR I-type radio galaxies are centre-oriented and have low power with no hotspot near the edge of the structures. Lobes in FR I galaxies diffuse after achieving ~ 1 parsec distance. FR II-type radio galaxies are edge-dominated (lobe-dominated) with compact hotspots at the edges of radio lobes. The critical radio luminosities of the majority of FR I sources are measured as $L_{178\,\mathrm{MHz}} < 2 \times 10^{25}$ W Hz^{-1}sr^{-1} and that of FR II sources, $L_{178\,\mathrm{MHz}} > 2 \times 10^{25}$ W Hz^{-1}sr^{-1} [24].

The life of radio galaxies can be understood by their different stages, just like human life. The changes that occur in radio galaxies can be visualised by their sizes. The variation in the radio luminosity with their linear sizes is depicted in Figure 2. With their increasing linear sizes, they become older and fainter with a decrease in their radio luminosity. The sizes of radio galaxies in their early stages are observed to be smaller, and these

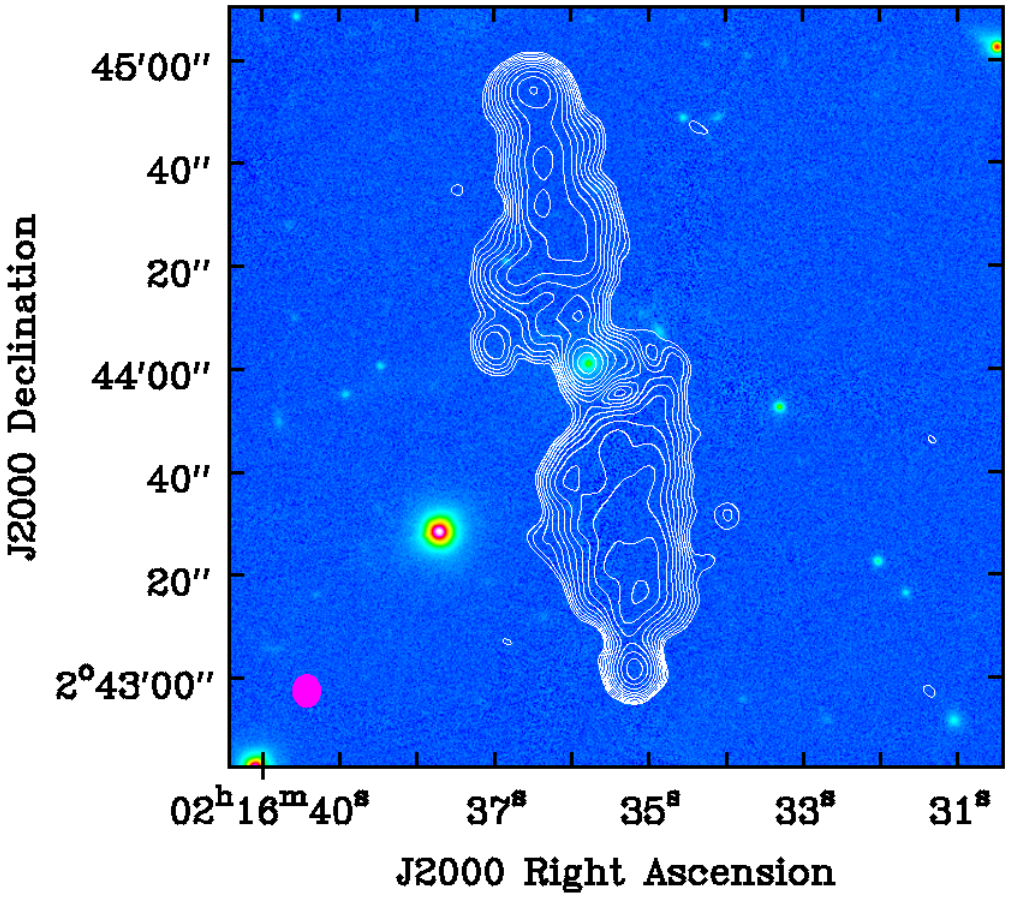


Figure 1: **J0216+0244: The FR radio galaxy with large angular size as observed by the VLA FIRST survey.**

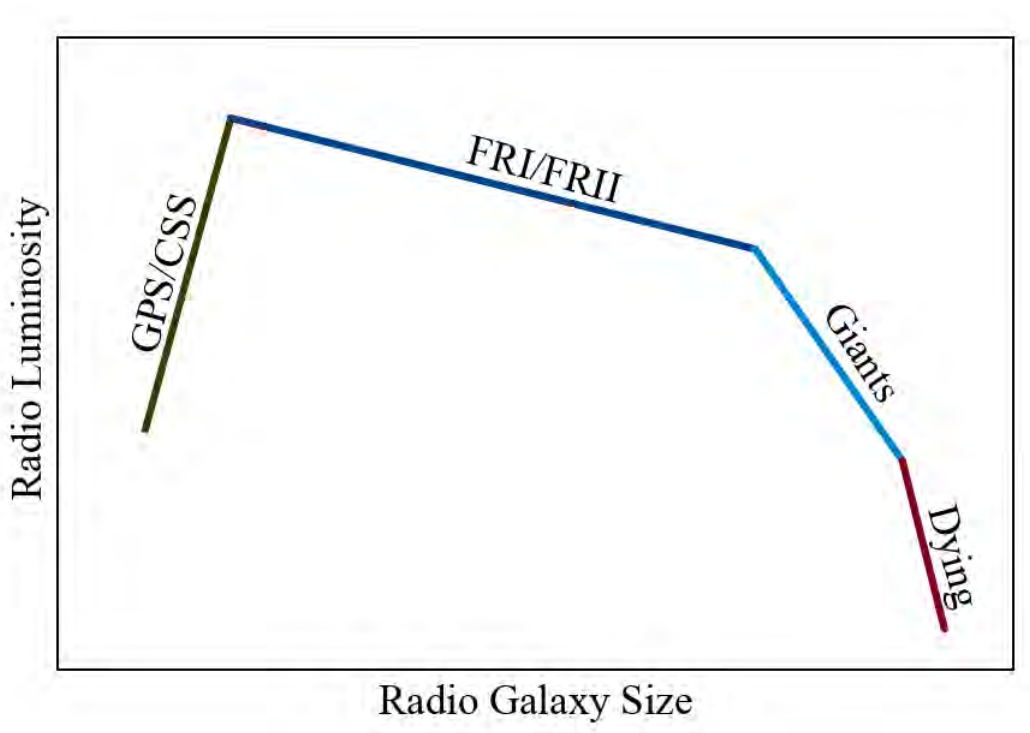


Figure 2: **A schematic diagram of the variation of radio galaxy size with radio luminosity.**

sources are known as CSS (Compact Steep Spectrum) and GPS (Gigahertz Peaked Sources) [11]. These compact radio sources reside only a few kiloparsecs away from the core. After this stage, radio galaxies grow and become powerful radio sources like FR I/FR II. It is believed that the life of giant radio galaxies began when their sizes were very large in comparison to normal FR I/FR II radio galaxies. After passing through the giant stage, radio galaxies enter the dying stage. The radio luminosity of dying radio galaxies becomes very faint. Because of their extremely low radio luminosity, they are difficult to detect even with advanced radio telescopes.

Various radio galaxies with irregular shapes and sizes have been observed in the last two decades, in addition to normal FR I/FR II radio galaxies. The morphology of these radio galaxies depends on the alignment and structure of radio jets, and they are classified as follows:

i. Winged Radio Galaxies: Winged radio galaxies are described as a mysterious sub-class of radio galaxies. In these radio galaxies, despite the main lobes (primary lobes), extra secondary lobes (wings) originate near the center or edge of the primary lobes. Because of the extra wings (secondary jets) in these radio galaxies, they are known as winged radio galaxies. A detailed discussion of these sources is described in section 2.

ii. Head–tailed radio sources: Head–tail (HT) radio galaxies are known as a subclass of radio galaxies with jets in two opposite directions that are bent in an apparent common direction. This distortion possesses 'HT' morphology, in which the primary jets are twisted rearward to form a tail, while the bright galaxy serves as the 'head' [62]. The bent morphology of jets in HT galaxies is thought to be based on two possible mechanisms relating to the intra-cluster medium (ICM) and intra-galactic medium. For the first mechanism, the host galaxy has a higher than predicted velocity that leads to the ram pressure on the primary jets, which distorts the galaxy morphology [60] [74]. For the second mechanism, cluster weather is responsible for the bending of jets. There may be dynamic interactions (cluster-cluster merger, galaxy merger) in the cluster that cause the strong winds in the ICM and are responsible for the bending of the jets [9]. HT radio sources are commonly found in the environments of rich clusters of galaxies [10].

Because high-sensitive radio observations can detect radio sources up to a high redshift, the HT radio galaxies can be used as a tracer of galaxy clusters for more distant clusters. The peculiar morphology of HT radio galaxies

indicates that radio jets interact strongly with their intracluster medium. These galaxies are classified as Wide Angle Tailed (WAT) radio sources or Narrow-Angle Tailed (NAT) radio sources based on their distorted angle and radio luminosity [65]. WATs are the sources with a bending angle greater than or equal to 90 degrees. NAT sources are those that have an angle of less than 90 degrees. 3C 465 is the first known WAT [22] [32] and NGC 1265 is the first known NAT radio source [75] [63].

Using the NRAO Very Large Array (VLA; [80]) Faint Images of the Radio Sky at Twenty cm (FIRST; [3]) survey, 614 new HT radio sources (of which 398 are WAT and 216 are NAT sources) are identified [78]. Using the TIFR GMRT Sky Survey, 268 (189 WATs and 79 NATs) sources are identified [8]. Recently, 459 bent-tailed radio galaxies [61] have been detected by the LOFAR Two-Metre Sky Survey (LoTSS). Using the same survey (LoTSS), fifty new HT radio sources (of which forty-five are WATs and five are NATs) are identified [67]. An interacting narrow-angle tail radio galaxy J0037+18A has been discovered in the Cygnus constellation [68].

iii. Hybrid Morphology Radio Sources: Hybrid Morphology Radio Sources (HyMoRS) are sources with mixed FR morphology. It is seen that these sources exhibit FR I and FR II lobes together on the opposite side of the structure [30] [27]. HyMoRS are found to be an extremely rare sources in the sky ($<$1% of radio galaxies fall into this category), and only a small number of sources have been discovered so far [27] [39] [44]. Though the exact cause of HyMoRS is unknown, it is thought to be the result of asymmetric jet interactions with the interstellar medium (ISM) or due to jet orientation, [30] [39]. These sources will demonstrate the relationship between the morphology of radio galaxies and the nature of the core engine, its surroundings, and the constitution of jets [30] [44] [18]. Hybrid morphology radio sources may be extremely useful to understand the origin of the FR dichotomy, which has been a topic of intense debate for more than four decades [44]. The asymmetric structure of HyMoRS may also be caused by the properties of host galaxies like pressure, accretion rate, black hole mass, etc., which can affect jet propagation as well as jet orientation.

From VLBI observations on 5 HyMoRS with a 10 kpc jet length, it is indicated that the jet orientation may not be responsible for the morphology of HyMoRS [13]. The supermassive black hole (SMBH) spin may also play a crucial role in the asymmetric structure in HyMoRS [44]. With the help of the Giant Metrewave Radio Telescope, six candidate HyMoRS were detected [30]. Three certain and two possible HyMoRS were discovered from the list of 1700 sources observed by the Faint Images of the Radio Sky at Twenty-cm (FIRST) survey, which proves that these sources are very rare [3] [27]. Previously, there was a very small number of known HyMoRS and, among them, only a small percentage had been thoroughly investigated. Recently, using the same NRAO Very Large Array (VLA; [80]) FIRST; [3] data, Kumari and Pal (2022) discovered a large number of thirty-three HyMoRS that significantly increased the sample size of known HyMoRS [44] [43].

iv. Giant Radio Galaxies (GRGs): Giant radio galaxies (for a detailed review, see Bhukta, Pal and Mondal (2022) of the present volume [6]) are much larger objects compared to most of the other objects in the universe. They are thought to be very rare objects that grow in low-density environments [37]. We can use a giant radio galaxy to understand the evolution of the life of a radio galaxy. It is also useful to investigate the distribution of gas in the inter-galactic medium [79] [69]. A large number of GRGs were discovered using TGSS at 150 MHz [5].

2 Winged Radio Galaxies

Winged radio galaxies are a subclass of radio galaxies that have two extra secondary lobes (wings) in addition to the primary lobes. These secondary lobes (wings) extend at an angle from the centre or from the edge to a distance that is nearly equal to or less than the length of the active lobes and sometimes more than the primary lobe. Winged radio galaxies are classified as *X*-shaped radio galaxies (XRGs) or *Z*-shaped radio galaxies (ZRGs) based on the location of their secondary lobes (ZRGs). In *X*-shaped radio galaxies (XRGs), secondary lobes emerge near the central region of the host galaxy at some angle to the primary jets, giving rise to the *X*-shaped structure. In *Z*-shaped radio galaxies, secondary lobes emerge from the edge of the primary lobe, making an angle mostly perpendicular to the primary lobes [31]. It is found that for the majority of XRGs, the primary lobe pair are FR II type and the secondary wings are always FR I type [55].

For most XRGs, wing sizes are at least 80% of the size of the primary lobes [55]. Due to the projection effect, some of the wings may look much shorter than their actual size. Because of the lack of hotspots near the edges of wings, detection of wings is difficult and the size of wings is highly dependent on the depth and spatial

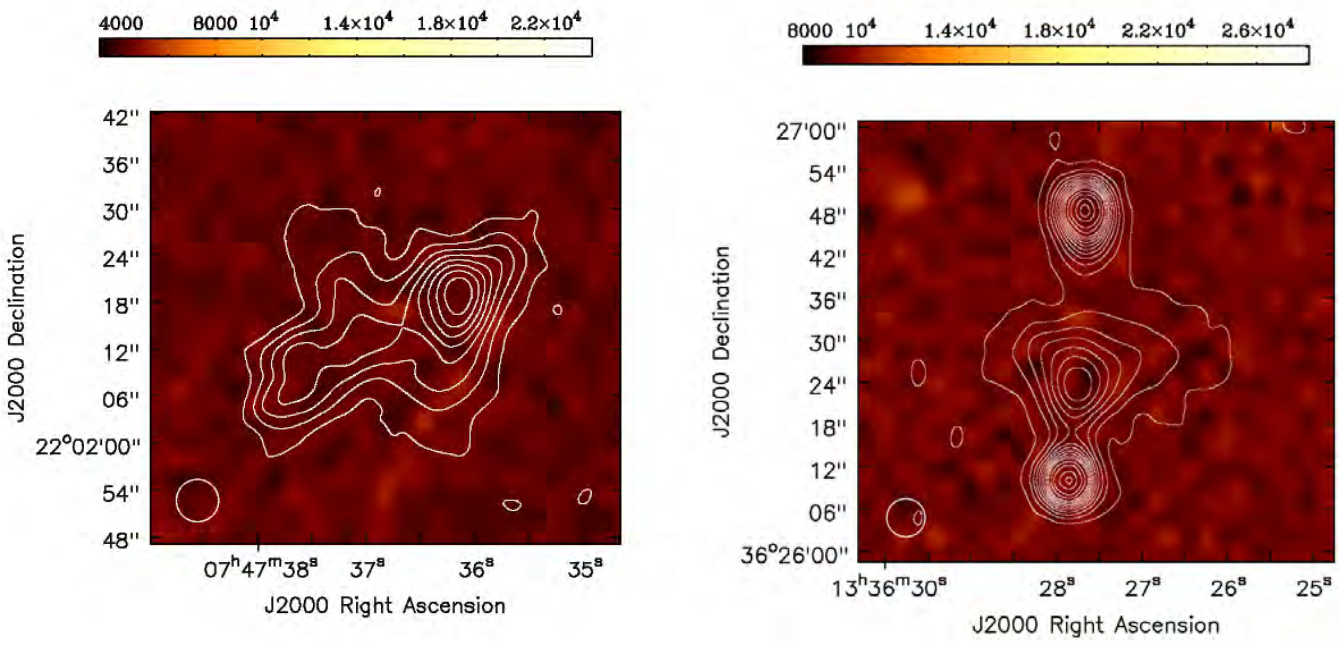


Figure 3: Examples of *X*-shaped radio galaxies from FIRST survey [4].

resolution of the radio map. Identifying the optical counterparts is very important to understanding the nature of winged radio galaxies. After knowing the optical counterparts, the classification of winged radio galaxies can be easily understood. For searching for optical counterparts and redshifts of sources, various versions of SDSS[1] and NASA/IPAC Extragalactic Database (NED) are used. In Figure 3 and 4, we presented images of two XRG sources taken from FIRST and LOFAR survey with optical counterpart at the center of the galaxy [4] [66]. In Figure 5 and 6, we presented a sample of two ZRG sources taken from FIRST and LOFAR survey with optical counterpart at the center of the galaxy [4] [66].

Discoveries of a large number of XRGs and ZRGs will help to understand the nature of XRGs in comparison to normal FR II radio galaxies. The detailed study of 388 FR II radio galaxies with 106 strong XRG candidates was done [35] [41] [86]. The mass of SMBH is estimated using the well-known tight relation [28] [26] [81] with the equation 1. It is found that in comparison to normal radio galaxies, the central black holes of XRGs are somewhat less massive (average masses of XRGs and FR II are $\log\ M_{BH}$ $\sim$8.81$M_{\odot}$ and 9.07 $M_{\odot}$, respectively).

$$\frac{M_{BH}}{10^9 M_{\odot}} = (0.310^{+0.037}_{-0.033}) \left(\frac{\sigma_*}{200 \text{ km s}^{-1}} \right)^{4.38 \pm 0.29} \quad (1)$$

Wide-field Infrared Survey Explorer (WISE) counterparts were searched with the help of WISE archival data for 25 XRGs taken from Yang and Cheung ([86] [14]) and 388 FR II radio galaxies [35]. With measured infrared colours for three mid-IR bands of the WISE survey for each host galaxy, it is found that XRGs are more infrared than FR IIs and often have a substantial fraction of $\sim$80% [35]. This shows that XRGs have more cool ISM than FR II radio galaxies, presumably caused by recent merger activity.

2.1 Statistical Properties

2.1.1 Spectral Index

The two-point spectral index is calculated between two frequencies by assuming $S_\nu \propto \nu^{-\alpha}$, where S_ν indicates the flux density at the corresponding frequency ν and α denotes the spectral index. The spectral indices are determined by measuring flux density at both frequencies over the same aperture.

[1] http://www.sdss.org.

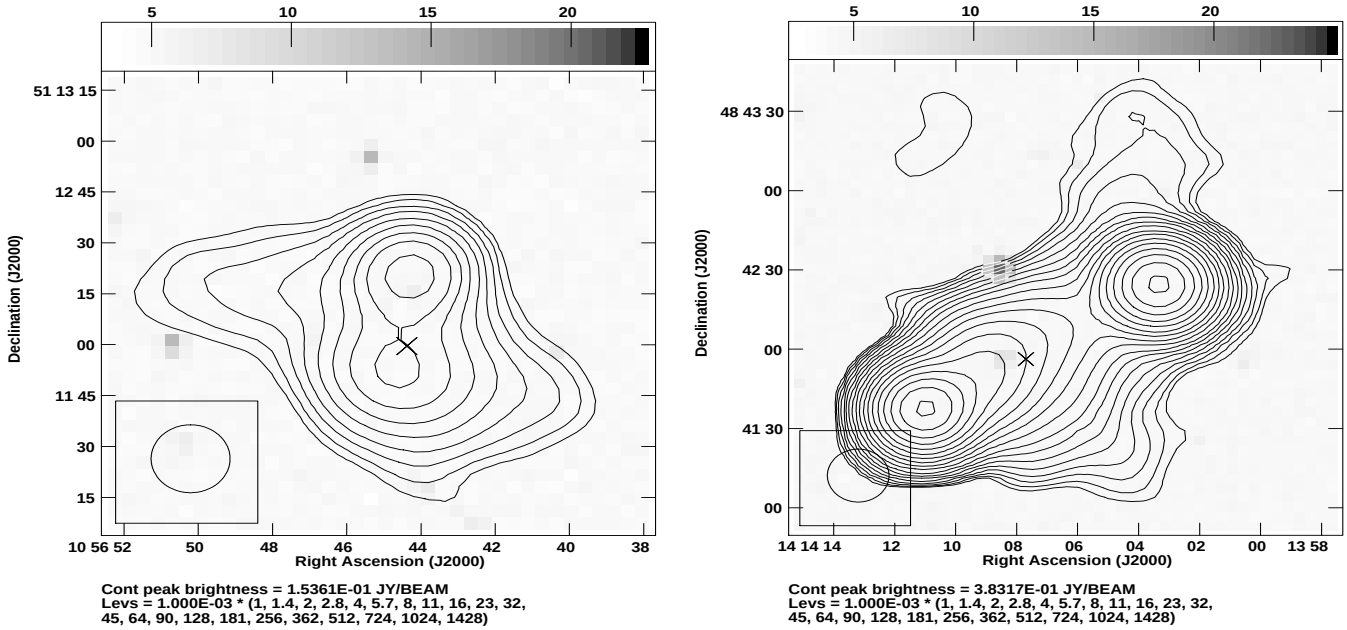


Figure 4: Examples of *X*-shaped radio galaxies from LOFAR survey [66].

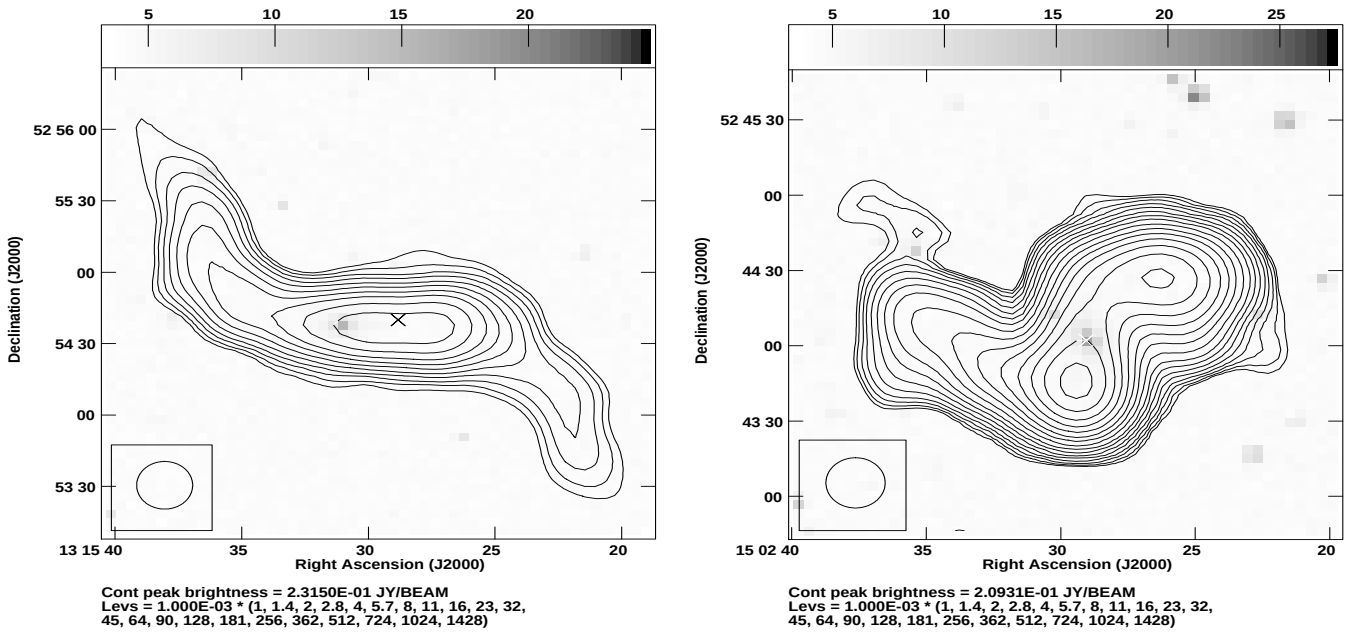


Figure 5: Examples of *Z*-shaped radio galaxies from LOFAR survey [66]

.

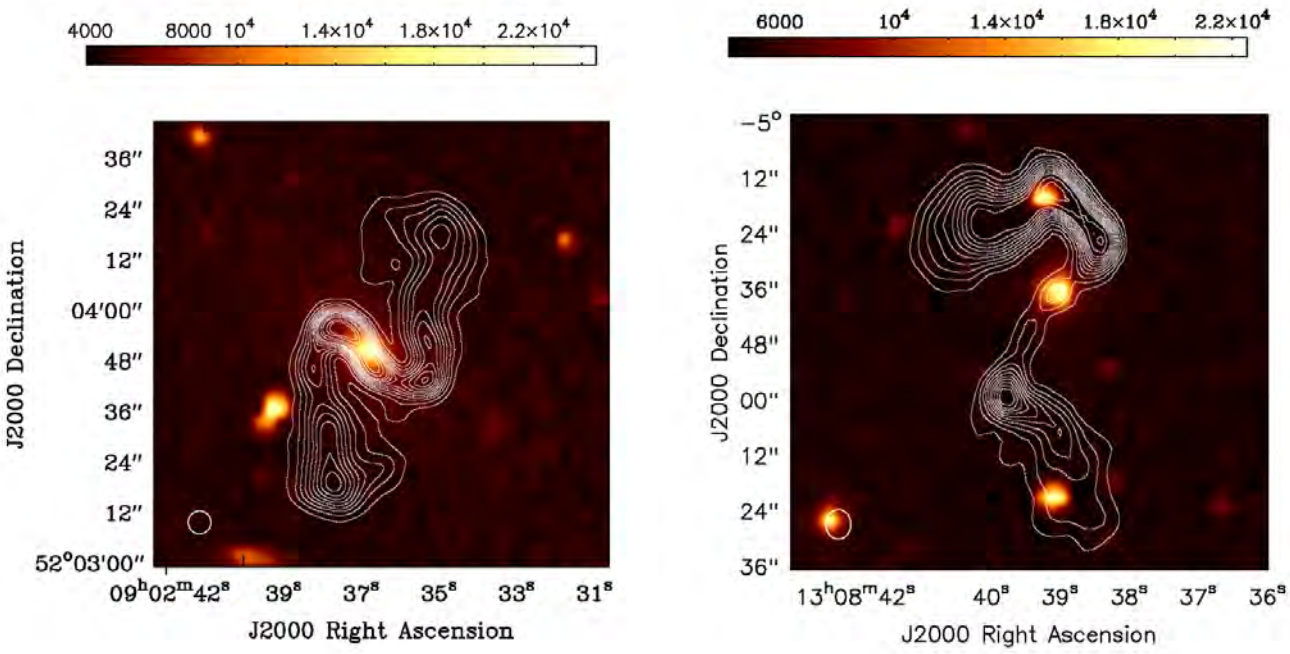


Figure 6: Examples of *Z*-shaped radio galaxies from FIRST survey [4].

Cheung (2007) calculated spectral indexes (α) for 100 winged radio sources between frequencies of 365 MHz and 1.4 GHz and between 1.4 GHz and 4.9 GHz [14]. Among these winged radio sources, 90–94 sources showed a steep radio spectrum ($\alpha > 0.5$). A spectral index was calculated in the frequency range 150 MHz – 5 GHz and steep radio spectra were detected for almost all strong XRG candidates in the catalogue of Yang et al., (2019) [86] using FIRST survey. There are only a very few winged sources that have flat or inverted radio spectra ($\alpha < 0.5$) [86] [14]. This result helps to probe the dual-AGN scenario for the formation of *X*-shaped morphology [86] [49] [50]. The spectral index of regular radio galaxies is normally found in the range of 0.7 to 0.8 [58] [64] [38] [34] and the average spectral index for all identified winged radio sources is also in the range of 0.7–0.8 [4] [66] [86] [14] [7]. The histogram in Figure 7 shows the distribution of the spectral index of XRGs and ZRGs using all major surveys. The histogram shows the peak of the spectral index nearly at 0.7–0.75 for all identified winged radio galaxies (XRGs and ZRGs). This value of the spectral index of winged sources suggests that, in terms of spectral index, these sources are comparable to normal radio galaxies.

2.1.2 Radio Luminosities

Radio luminosity is an important parameter for understanding the nature of radio galaxies. The below formula can be used for calculating the radio luminosity of winged sources.

$$L_{\text{rad}} = 4\pi D_L^2 S_\nu (1+z)^{(\alpha-1)} \tag{2}$$

where S_ν is the radio flux at a frequency ν and D_L is the luminosity distance (in meter), here $(1+z)^{(\alpha-1)}$ is the standard k-correction term. In this equation, the spectral index (α) is assumed to follow $S_\nu \propto \nu^{-\alpha}$. It is expected that X-shaped radio sources will have radio luminosities close to the FR I/FR II division of $L_{178} \sim 2{\times}10^{25}$ W Hz^{-1} [55] [20]-[51]. The connection between FR I and FR II radio galaxy divisions is not well understood, and winged galaxies may be the missing links between these two classes of galaxies. The average radio luminosity of 32 (out of 100) identified winged radio galaxies (from Cheung (2007); [14]) was close to the FR I-FR II divide ($\log L_{\text{rad}} \sim 25.49$ W Hz^{-1}). For all discovered winged radio galaxies, including some quasars, $\log L_{\text{rad}}$ lies in the range of 24.56 to 27.09 [66] [86] [14] [7]. The distribution of radio luminosity ($\log L_{rad}$) with redshift (z) for all reported winged radio sources in surveys like FIRST [4] [86] [14], TGSS [7] and LoTSS [66] is shown in Figure 8 with various colors. The majority of XRG sources reported in these surveys were average luminous sources. Some of the sources reported in FIRST were more extended and powerful winged sources compared with other

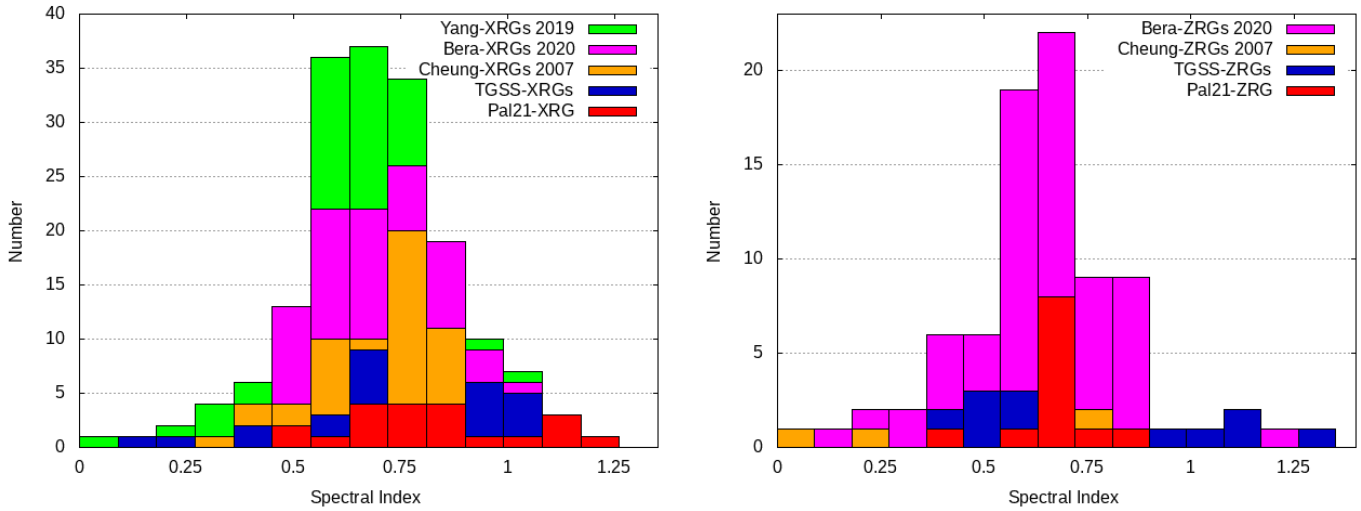


Figure 7: Spectral index distribution of XRGs and ZRGs. Here sources are taken from surveys based on LOFAR [66], TGSS [7] and FIRST [4] [86] [14].

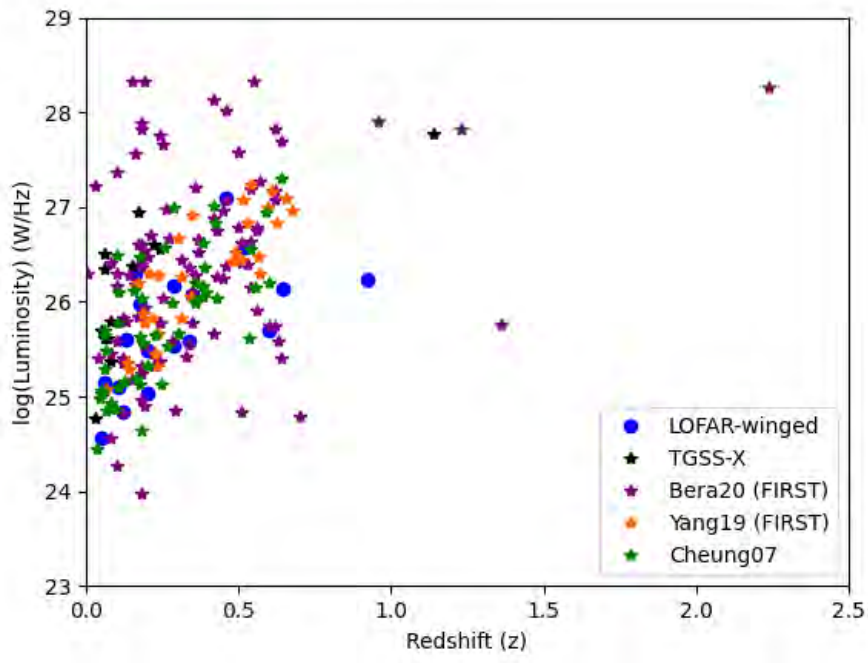


Figure 8: Plot of radio luminosities ($\log L_{rad}$) with redshifts (z) for XRGs and ZRGs. Blue filled circles represented sources from LoTSS DR1; purple, yellow and green color stars represented sources from FIRST [4] [86] [14]; black color stars represented sources from TGSS-XRGs (FR II) [7].

surveys. It is found that the redshift limit of winged radio sources reported in these surveys ranges from 0.3–2.245 [4] [66] [86] [15].

2.2 P-D Diagram

The projected linear size (pc) with log(radio luminosity) (W Hz^{-1}) of available sources discovered in TGSS [7], FIRST [4] and LOFAR [66] surveys was shown in Figure 9. This diagram depicted the dynamic evolution of radio galaxies from CSO–CSS–FR I/FR II-(XRGs) for high radio-powered sources to low radio-powered and very faint sources. The radio power of sources, the local environment of the host galaxy, and the evolutionary age of sources have a great impact on the structural and spectral properties of radio sources [36] [1]. The compact steep spectrum sources (CSSs), and compact symmetric objects (CSOs) are very small and represent the earliest stages of radio source growth. The radio power varies greatly with source size and synchrotron losses are dominates by adiabatical expansion inside the ISM–intergalactic medium (IGM) transition point, where the radial dependence (β) of the ISM density changes. This distance falls in the range of 1–3 kpc, but it can also be significantly larger. They reach a distance in the range of kpc from the central black hole (a few thousand light-years) [1]. The lifetime of CSO and CSS sources is very short, which means that within a very short time, these sources are spotted in the sky. As the brightness of these sources is high enough to be detected easily, some sources are detected as compact. As the source grows in size, the radio luminosity of the radio galaxy starts to decrease. XRG sources (mostly FR II) belong to an adult class of radio sources with a larger size and appreciably medium radio luminosity. Figure 9 showed the distribution of radio power (P_{rad}) with the linear sizes (D) of discovered winged sources from various surveys such as LoTSS DR1-XRGs [66], TGSS-XRG (FR I-FR II) sources [7], FIRST-XRGs and ZRG sources [4]. Here we also included low and high-power compact symmetric objects (CSOs), compact steep-spectrum sources (CSSs), GHz peaked-spectrum (GPS) sources [46]-[57], and FR I giant radio galaxy (GRG) sources [54]. The orange dashed line and the black dashed line represent the evolution paths for low and high radio power sources. Three black vertical dashed lines represent the range of medium-sized symmetric objects (MSOs). From Figure 9, it can be seen that some XRGs belong to the group of large symmetric objects (LSOs) with high as well as low radio power and linear sizes of the order of mega-parsec (Mpc). CSO and CSS sources are the first stage, and LSOs are the final stage of radio source evolution. The well-grown FR I and FR II sources of larger linear size (greater than 10 kpc) were also found in the last stage of evolution. The variation of radio power with linear size follows the relation $P_{rad} \propto D^{-1.6}$ [1] [7]. Here, radio power decreases as the linear size of radio sources increases because inverse Compton losses from the cosmic microwave background (CMB) take precedence over synchrotron losses [1] [7]. This leads to the hypothesis that XRGs are undergoing a transformation into giant radio galaxies (GRGs) with linear sizes of $\geq$ 2 Mpc.

2.3 Wing Structure

The wings that originated from the centre and edges of the primary jets are different in shape, size, point of ejection, and direction of propagation, or more than one such condition. Thus, morphologically, the sources look asymmetric. There is a variation of wing structures in XRGs. Based on the properties of wings, XRGs are further classified into four sub-classes [7]:
(1) XRGs with small wings : The primary lobes in these XRGs are larger than the wings. The faint and small wing structure may point to either the beginning or end of the source activity. A study with high resolution and better sensitivity radio images of these types of structures is required to circumvent this limitation.
(2) XRGs with long-wings: The wings of these XRGs are larger than the primary lobes. We must keep in mind that the effect of projection may have an impact on primary jets. Long wings indicate the maximum backflow activity [15].
(3) Single-wing XRGs: These XRGs have a small or missing wing on one side.
(4) Z-symmetry XRGs: A Z-symmetric structure is formed by the centre AGN because of the distinct misalignment of the opposite jets of the galaxies. Here it is noticeable that ZRGs that have wings that seem to emerge from the edge of the primary jets are different from Z-symmetric XRGs.

2.4 Different Models

In the last five decades, several mechanisms (models) have been proposed for the formation of *X* or *Z*-shaped radio galaxies. Some widely discussed models are summarised below:

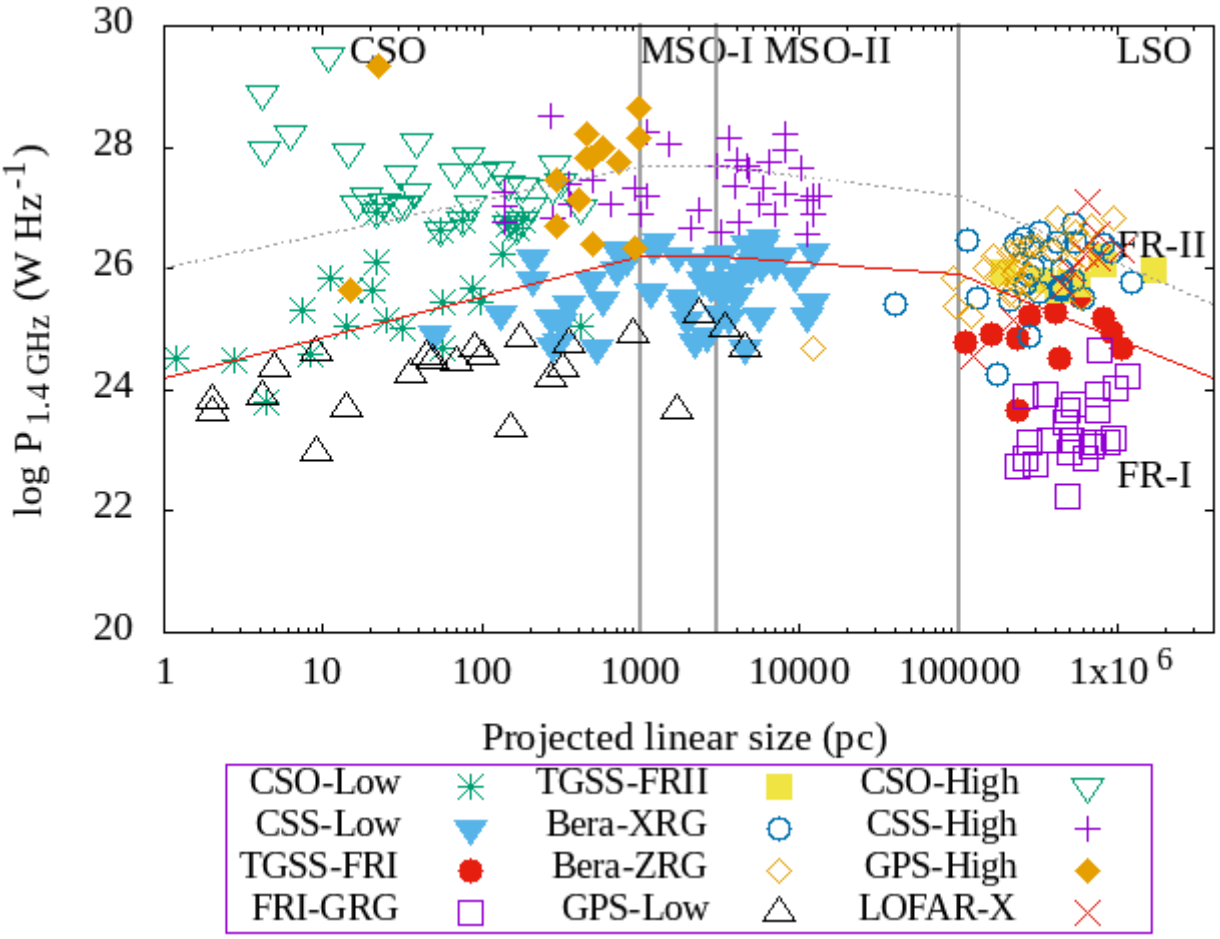


Figure 9: Distribution of radio luminosity with the linear size of mentioned sources; green stars and green reverse triangles represent low and high power compact symmetric objects (CSOs) [46] [45]; filled blue reverse triangles and magenta pluses represent low and high power compact steep-spectrum sources (CSSs) [16] [17] [25]; red filled circles and yellow filled squares represent TGSS-XRGs (FR I) and FR II sources [7], blue open circles and orange open diamonds represent FIRST-XRGs and ZRG sources [4], black open triangles and orange filled diamonds represent low and high power GHz peaked-spectrum (GPS) [19] [85] [57], magenta open squares represent FR I giant radio galaxy (GRG) [54] and orange crosses represent LoTSS DR1-XRGs [66]. Here, orange and black dashed lines represent the evolution paths for low and high radio power sources.

- **Backflow and Buoyancy Model:** This model suggests that the formation of wings in winged radio galaxies is due to the backflow of plasma near the central host galaxy. The dense ISM of the host galaxy plays a crucial part in the backflow diversion concept for the formation of secondary wings by applying buoyancy pressure to the back-flowed synchrotron plasma [56] [42]. In this approach, there is a significant propensity for the wings to align with the optical minor axis of the elliptical host galaxy. This hydrodynamic backflow hypothesis, however compelling for its simplicity [20], is contradicted by the crucial finding that in some of the XRGs the wings are noticeably longer than the primary lobes [4] [86] [7] [76]. In contrast to the wings, which are designed to grow subsonically, it is expected to move forward supersonically for the external medium (whose ram pressure generates a glowing hotspot at the edge of the lobes) [55].

- **Black hole Merger (**[20] [59] [29]**):** According to this model, when two radio galaxies merge, their central supermassive black holes also merge. Due to the merger of two black holes together, their axis of rotation gets disturbed. If one of the black holes is launching a jet along its spin axis before the merger, and if the resulting merged black hole also launches a jet, then the jet will appear to change its direction, which may cause the wing structures to shape. This re-orientation of the jet axis in XRGs is well explained by this model. Previously, Merritt & Ekers (2002) [59] proposed that XRGs are indicators of such mergers, and they predicted that the prevalence of such sources will be useful to detect the magnitude of the gravitational wave background (GWB) radiation.

- **Twin active galactic nuclei (AGNs):** The occurrence of elliptical galaxies with twin active galactic nuclei, such as NGC 326 [84] [2], served as motivation for this model. According to this model, twin radio lobes are thought to be two separate radio doubles connected to a pair of nearby active SMBHs that are merging into two massive ellipticals [49]. Furthermore, because it requires both of the centre engines to be concurrently launching jets, this scenario has an extremely low likelihood. The conceptual simplicity of this "twin AGN" model is advantageous, but it is unable to explain the XRGs that possess FR II type lobe pairs. It also does not explain the reason for pointing the primary lobe along the major axis of the host galaxy [12] [77] [33].

2.5 Identified Winged Radio Galaxies in the Last Ten Years

3C 272.1 is the first discovered radio source [71] that has Z-shaped morphology. NGC 326 and NGC 3309 are classified as winged radio sources [23] [21] [40]. For the first time, a list of 11 sources as an X-shaped radio galaxies was reported [55]. Cheung (2007) [14] used the NRAO Very Large Array (VLA; [80] Faint Images of the Radio Sky at Twenty cm (FIRST; [3]) data to identify a list of 100 X-shaped sources. Recently, using the VLA FIRST survey, 296 winged sources (of which 161 were identified as XRGs and 135 as ZRGs) were detected using FIRST [4]. A list of 290 XRG candidates using the same survey is also presented in the literature [86]. Using the TIFR GMRT Sky Survey (TGSS) at 150 MHz, 58 winged sources are discovered, out of which 40 are XRGs and 18 are ZRGs [7]. By the automated morphology detection process, a list of 156 XRG sources from the FIRST is presented [70]. A detailed study of the properties of XRG sources detected in Cheung (2007) was carried out using follow-up VLA observations [14] [76] [72] [73]. Recently, using the LOFAR Two-metre Sky Survey First Data Release (LoTSS DR1), Pal and Kumari discovered twenty-nine winged radio galaxies [66]. Some of the quasars are seen in the samples of XRGs in various surveys. In the Cheung (2007) sample, 12 sources are confirmed as quasars [14] [76]. The detection of X-shaped quasars are quite rare; the first reported X-shaped quasar was (4C+01.30 at $z = 0.132$) [83]. According to further study, one more X-shaped quasar, WGA J2347+0852, was reported in the literature [53]. It is found that 4C+01.30 exhibits a double-peaked (broad) emission line [87]. In the catalogue of Yang et al. (2019) of XRG candidates, three sources of a similar kind have been detected that show double-peaked narrow emission lines [86]. In Figure 10, the location of winged sources observed from survey-based on LOFAR [66], TGSS [7] and FIRST [4] [86] [14] are shown. SMBH pair/binaries can be searched from these sources with the help of VLBI high-resolution observations because the appearance of double-peaked emission lines in XRGs may be interpreted as a piece of evidence preferring the binary or dual blackhole model [50] [47].

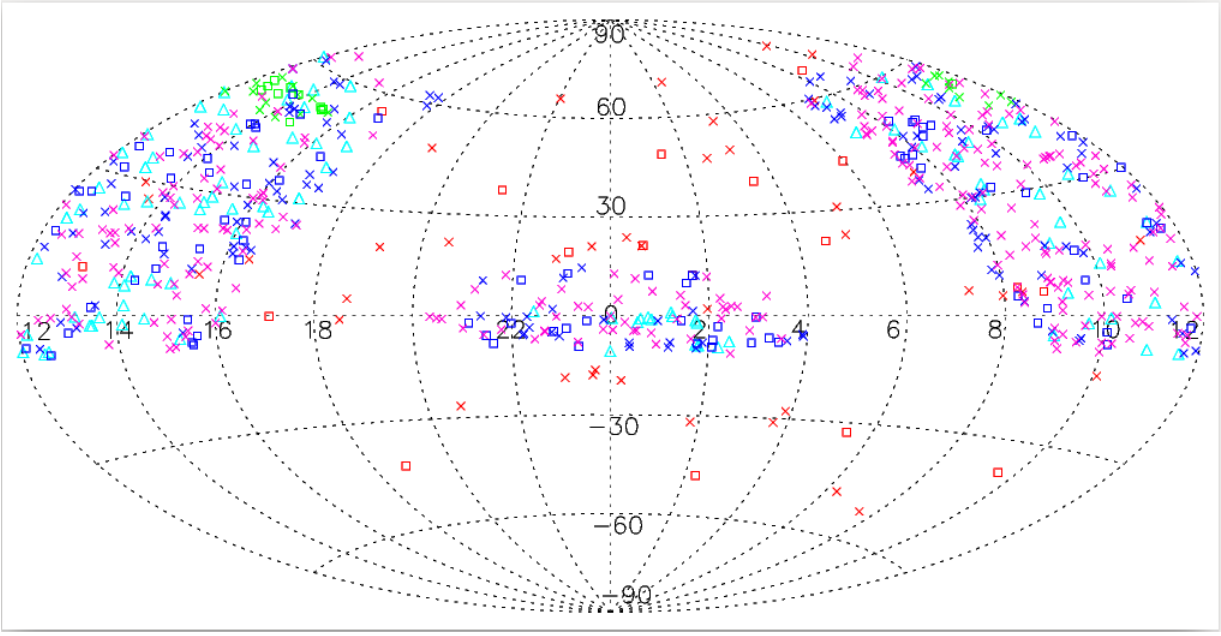


Figure 10: The distribution of winged radio sources. The 'X' symbols represent X-type and the square symbols represent Z-type sources. Here the sources are taken from surveys based on LOFAR [66], TGSS [7] and FIRST [4] [86] [14], indicated by green, red, magenta, blue and cyan colours respectively.

3 Multi-Wavelength View of Winged Radio Galaxies

3.1 High Frequency Radio Observation

With the help of MeerKAT observation advanced telescope data in African Radio Astronomy Observatory (SARAO) situated in the Northern Cape of South Africa, a study on the well-developed *X*-shaped radio galaxy PKS 2014–55 at 1.28 GHz frequency was carried out. It is observed that this source is well consistent with the hydrodynamical flow model in terms of its morphology, spectra, and magnetic field st ructure. This source has a spectral index of 0.8 and radio luminosity is in the order of 2×10^{25} W Hz^{-1}. According to this study, AGN activity has recently restarted in this source (PKS 2014–55) with the resurrected jets pointing in the same general direction as the primary lobe. As a result, it is extremely unlikely that the secondary wings were caused by a shift in the spin axis of SMBH.A host galaxy called PGC 064440 is located at the centre of the galaxy, and it has the appropriate position angle and virial halo to effectively redirect backflows from the large primary jets in the direction of the observation.

3.2 Optical Observation

After studying high-frequency radio observations and Chandra-X-ray observations, optical observations are also very important to understanding the nature and origin of winged radio galaxies. Optical imaging revealed that XRGs are typically found in weak clusters and are mostly found in unaltered, highly elliptical galaxies [20]-[15] [82]. It is found that for some of the sources, there exist optical spectra with narrow and broad emission lines [66] [15]. Broad emission lines are primarily visible in winged sources with quasar-like behaviour. These spectra show the Balmer emission lines with the inclusion of [O III] λ5007 and [O II] λ3727 emission lines [51]. In the work of Landt, Cheung and Healey (2010) [51], they calculated the flux ratio between the sum of the oxygen line [O III] λ4363 and the oxygen doublet [O III] $\lambda\lambda$ 4959, 5007 when available. This flux ratio is an effective indicator of electron temperature (temperature reduces with increasing its value). They detected [O III] λ4363 line for nineteen sources out of twenty-eight sources, and these nineteen sources were categorised as strong-lined radio-loud AGNs. With the help of these spectra, the power of the emission line is calculated and, considering the general relationship between emission line and radio power, the evidence of recent merger activity can be searched and it can also probe the nuclear environment [51]. In these spectra, most sources have narrow emission lines, while only a few (quasars) have broad emission lines. A transition point in radio and narrow emission line power

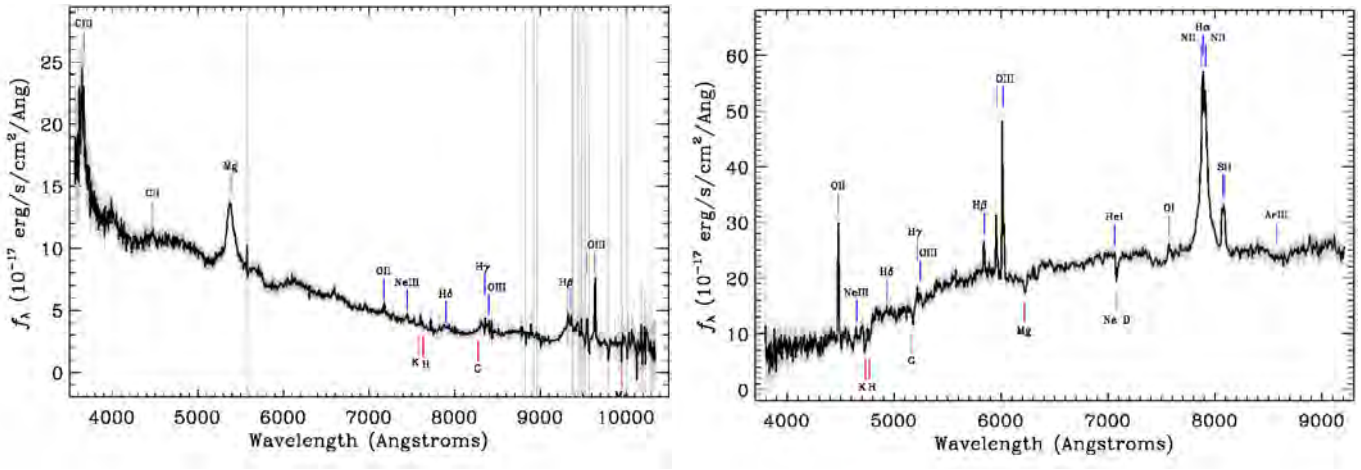


Figure 11: Optical spectra of winged radio galaxies J1054+5537 and J1239+5314 [66] from SDSS DR16.

is pointed out at $L_{\rm rad} \sim 10^{25.6}$ and $L_{NLR} \sim 10^{35}$ W Hz^{-1} [51]. Calculations are also made for the Ca II break value (placed at wavelength nearly 4000 Å) in the spectrum of f_λ versus λ. The Ca II break value is calculated using the formula $C = 1 - (f_-/f_+)$. Here, f_+ and f_- indicate the fluxes in the wavelength range (4050–4250) Å and (3750–3950) Å respectively. Ca II break value, which measures the angle between the line of sight of the observer and radio jets, is used to determine the orientation of AGN [51] [52].

In Figure 11, optical spectra of two winged sources are presented, one for the XRG source (J1054+5537) and the other for the ZRG source (J1239+5314) [66]. Both the sources are quasars with redshifts of $z = 0.924$ and $z = 0.20$. With the measured fluxes of [O III](λ5007) and H $_\alpha$ lines, the flux ratio of [O II I](λ5007)/H$_\alpha$ of the source J1239+5314 is found to be > 0.2 (suggesting that this source is high-excitation radio galaxy (HERG)).

Any unique broad emission lines that might reveal details about large-separation binary black holes or dusty nuclear environments as explored by the narrow emission lines are looked for as indicators of a recent galaxy merger. Searches are also being conducted for indications of accelerated star formation in the host galaxy. Further research revealed that all three approaches (galaxies merging, dusty nuclear environments, and accelerated star formation regions in the host galaxy) produced unfavourable findings, suggesting that some sort of merger had to have taken place a long time ago, and thus the pair of wings can be attributed to relic radio emission [51]. The electron densities and nuclear environment temperatures of XRGs were estimated with the analysis of narrow emission lines. It is found that the estimated temperatures for the majority of sources are relatively high ($T > 15,000$ K). This result is in favour of the hypothesis that the formation of XRGs may be caused by overly pressured conditions rather than recent mergers [51].

4 Discussion and Conclusion

Primary lobes usually have an edge-brightened morphology, which is characterised by a hot spot near the edge of the lobe. Only a small proportion of radio galaxies ($< 10\%$) have wings [55]. The formation of wings is linked with the properties of the host galaxy. Usually, XRGs possess high ellipticity ($\epsilon \geq 0.2$) galaxies. All the wings form in the minor axis direction of the host elliptical galaxy [12]. Wings usually have a steep spectral index, suggesting that the wings are comparably older than the lobes. Some XRGs also possess a flat spectral index in the wings compared to the lobes [50] [48]. Usually, the radio power of FR II is higher compared to the radio power of FR I sources [24]. XRGs radio luminosity is comparable to the FR I and FR II division luminosity $L_{1.4}$=10^{25} W Hz^{-1} at 1.4 GHz [4] [20] [15]. Winged radio galaxies may be the missing link in the transition between FR I and FR II sources. The structure of wings varies in XRGs and ZRGs. As discussed in Bhukta et al., (2022) [7], sources differ in their wing structure. For the sources taken from LoTSS DR1 [66], the secondary jets (wings) were longer than the primary jets (wings) for four sources (J1056+5111, J1129+5407, J1414+4842, and J1336+4900). Longer wing XRGs were also found in the previous study [4] [86] [7] [76]. The longer size of the wings can indicate the stage of evolution of the radio source. Long wings can also indicate maximum backflow activity in a hydrodynamic origin, whereas short wings can indicate either the beginning or end of source activity.

The morphology of one of the sources J1109+5314 in Pal and Kumari (2021) [66] was complicated. It has

only one wing, but a clear, bright hotspot was visible in the wing, which had never been seen before in any XRG. It is possible that this elongation was due to the presence of a background source. A follow-up deep observation must be conducted to understand the nature of this source. For the XRG source J1112+4755, secondary jets had a sign of a Z-symmetry structure [66]. The Z-symmetry of the wings may be easily comprehended using the spin-flip model [31] [20]. This model is characterised by the merging of two radio galaxies in which both possess a supermassive black hole at their core. A significant amount of gas is emitted into the ISM as the smaller galaxy spirals toward the common centre. The axial spin vector of the heavier active galaxy will be required to move towards the orbital angular momentum vector of gas of the confined black hole [59] [88]. A significant spin re-orientation of the heavier active radio sources in a particular direction may occur at a distance of $\sim$10 kpc from the AGN due to adequate ram pressure resulting the Z-symmetry secondary wings [31]. In the source J1442+5043, the primary jets and secondary jets both have a sign of Z-symmetry in their structure [66]. The Backflow scenario appears to be the most likely explanation for the evolution of *X*-structures, where a combination of buoyancy, high cocoon pressure, and galactic winds may promote the construction of wings. More systematic study is needed to look for the signature of other models, such as galaxy mergers or dual AAGN. t is clear that although there are various models for describing the origin of winged radio galaxies, there is no such unique model that fits the structure of all winged radio galaxies. To confirm the exact nature of these unique sources, a detailed study with multi-frequency high-resolution radio observations with deeper optical and X-ray imaging of XRGs and radio spectral mapping of these sources is encouraged. We anticipate that future deeper high-resolution surveys, such as SKA, high-resolution LOFAR, and MeerKAT, will allow discoveries because they can detect faint, diffused light, which previous telescopes were unable to do. Additionally, it will also reveal any spectral changes between the wings and lobes, as well as any spectral signature, regarding the age of the wings and the life-time of sources.

References

[1] T An and A. W Baan, *The Dynamic Evolution of Young Extragalactic Radio Sources*, The Astrophysical Journal **760** (2012).

[2] P Battistini, F Bonoli, S Silvestro, R Fanti, I. M Gioia, and G Giovannini, *The orientation of radio sources associated with elliptical galaxies*, The Astronomy and Astrophysics **85** (1980), no. 1-2, 101–105.

[3] R.H Becker, R.L White, and D.J Helfand, *The FIRST survey: Faint Images of the Radio Sky at Twenty Centimeters*, The Astrophysical Journal **450** (1995), 559.

[4] S Bera, S Pal, T. K Sasmal, and S Mondal, *FIRST winged radio galaxies with X and Z symmetry*, The Astrophysical Journal Supplement Series **251** (2020), 15.

[5] N Bhukta, S Pal, and S. K Mondal, *Giant radio sources in TGSS*, submitted (2022).

[6] ———, *Properties of giant radio galaxies*, Vol. 1, Scientific Research Publishing, Advances in Modern and Applied Sciences, A Collection of Research Reviews on Contemporary Research, 2022.

[7] ———, *Search for X/Z shaped radio sources from TGSS ADR 1*, Monthly Notices of the Royal Astronomical Society **512** (2022), 4308–4323.

[8] ———, *Tailed radio galaxies from the TIFR GMRT sky survey*, Monthly Notices of the Royal Astronomical Society, in press, DOI: 10.1093/mnras/stac2001 (2022).

[9] J. O Burns, *Stormy weather in galaxy clusters*, Science **280** (1972), 400–404.

[10] ———, *The radio properties of cD galaxies in Abell clusters. I. an X-ray selected sample*, The Astronomical Journal **99** (1990), 14.

[11] J. R Callingham et al., *Extragalactic peaked-spectrum radio sources at low frequencies*, The Astrophysical Journal **836** (2017), 174.

[12] A Capetti, S Zamfir, P Rossi, G Bodo, C Zanni, and S Massaglia, *On the origin of X-shaped radio-sources: New insights from the properties of their host galaxies*, The Astronomy and Astrophysics **394** (2002), 39.

[13] M Ceglowski, M. P Gawronski, and M Kunert-Bajraszewska, *Orientation of the cores of hybrid morphology radio sources*, The Astronomy and Astrophysics **557** (2013), A75.

[14] C. C Cheung, *FIRST winged and X-shaped radio source candidates*, The Astronomical Journal **133** (2007), 2097.

[15] C. C Cheung, S. E Healey, H Landt, G Verdoes Kleijn, and A Jordan, *FIRST "Winged" and X-shaped radio source candidates. II. New redshifts*, The Astrophysical Journal Supplement Series **181** (2009), 548–556.

[16] D Dallacasa, C Fanti, S Giacintucci, C Stanghellini, R Fanti, L Gregorini, and M Vigotti, *The B3–VLA CSS sample III. EVN & MERLIN images at 18 cm*, Astronomy & Astrophysics **389** (2002).

[17] D Dallacasa, S Tinti, C Fanti, R Fanti, L Gregorini, C Stanghellini, and M Vigotti, *The B3–VLA CSS sample II. VLBA images at 18 cm*, Astronomy & Astrophysics **389** (2002).

[18] de Gasperin F, *Multifrequency study of a new Hybrid Morphology Radio Source*, Monthly Notices of the Royal Astronomical Society **467** (2017), 2234.

[19] N de Vries, I. A. G Snellen, R. T Schilizzi, K. H Mack, and C. R Kaiser, *VLBI observations of the CORALZ sample: young radio sources at low redshift*, Astronomy & Astrophysics **498** (2009), 641–659.

[20] J Dennett-Thorpe, P. A. G Scheuer, R. A Laing, A. H Bridle, G. G Pooley, and W Reich, *Jet reorientation in active galactic nuclei: Two winged radio galaxies*, Monthly Notices of the Royal Astronomical Society **330** (2002), no. 3, 609–620.

[21] J. L. E Dreyer, *A new general catalogue of nebulae and clusters of stars, being the catalogue of the late Sir John F. W. Herschel, Bart, revised, corrected, and enlarged*, Memoirs of the Royal Astronomical Society **49** (1888), 1.

[22] J. A Eilek, J. O Burns, C. P O'Dea, and F. N Owen, *What bends 3C 465?*, The Astrophysical Journal **278** (1984), 37.

[23] R. D Ekers, R Fanti, C Lari, and P Parma, *NGC 326-a radio galaxy with a precessing beam?*, Nature **276** (1978), 588–590.

[24] B. L Fanaroff and J. M Riley, *The morphology of extragalactic radio sources of high and low luminosity*, Monthly Notices of the Royal Astronomical Society **167** (1974), 31–36.

[25] C Fanti, F Pozzi, D Dallacasa, R Fanti, L Gregorini, C Stanghellini, and M Vigotti, *Multi-frequency VLA observations of a new sample of CSS/GPS radio sources*, Astronomy & Astrophysics **369** (2001), 380–420.

[26] L Ferrarese, R. W Pogge, B. M Peterson, D Merritt, A Wandel, and C. L Joseph, *Supermassive black holes in active galactic nuclei. I. The consistency of black hole masses in quiescent and active galaxies*, The Astrophysical Journal Letters **555** (2001), L79.

[27] M. P Gawronski, A Marecki, M Kunert-Bajraszewska, and A. J Kus, *Hybrid morphology radio sources from the FIRST survey*, The Astronomy and Astrophysics **447** (2006), 63–70.

[28] K Gebhardt et al., *A relationship between nuclear black hole mass and galaxy velocity dispersion*, The Astrophysical Journal Letters **539** (2000), L13.

[29] B. P Gong, Y. P Li, and H. C Zhang, *The first kinematic determination of million-year precession period of active galactic nuclei*, The Astrophysical Journal Letters **734** (2011), L32.

[30] Gopal-Krishna and P. J Wiita, *Extragalactic radio sources with hybrid morphology: Implications for the Fanaroff-Riley dichotomy*, The Astronomy and Astrophysics **363** (2000), 507–516.

[31] P. L Gopal-Krishna Biermann and P. J Wiita, *The origin of X-shaped radio galaxies: Clues from the Z-symmetric secondary lobes*, The Astrophysical Journal Letters **594** (2003), L103.

[32] M.J Hardcastle, I Sakelliou, and D.M Worrall, *A Chandra and XMM–Newton study of the wide-angle tail radio galaxy 3C 465*, Monthly Notices of the Royal Astronomical Society **359** (2005), no. 3, 1007–1021.

[33] E. J Hodges-Kluck, C. S Reynolds, C. C Cheung, and M. C Miller, *The chandra view of nearby X-shaped radio galaxies*, The Astrophysical Journal **710** (2010), 1205.

[34] C. H Ishwara-Chandra, S. K Sirothia, Y Wadadekar, S Pal, and R Windhorst, *Deep GMRT 150-MHz observations of the LBDS-Lynx region: Ultrasteep spectrum radio sources*, Monthly Notices of the Royal Astronomical Society **405** (2010), 436–446.

[35] Ravi Joshi et al., *X-shaped radio galaxies: Optical properties, large-scale environment, and relationship to radio structure*, The Astrophysical Journal **887** (2019), no. 2, 266.

[36] C. R Kaiser and P. N Best, *Luminosity function, sizes and FR dichotomy of radio-loud AGN*, Monthly Notices of the Royal Astronomical Society **381** (2007).

[37] C. R Kaiser, J Dennett-Thorpe, and P Alexander, *Evolutionary tracks of FRII sources through the P-D diagram*, Monthly Notices of the Royal Astronomical Society **292** (1997), 723–732.

[38] V. K Kapahi, R. M Athreya, Wil van Breugel, P. J McCarthy, and C. R Subrahmanya, *The Molonglo reference catalog 1 Jy radio source survey. II. Radio structures of galaxy identifications*, The Astrophysical Journal Supplement Series **118** (1998), 275.

[39] A. D Kapińska et al., *Radio galaxy zoo: A search for hybrid morphology radio galaxies*, The Astronomical Journal **154** (2017), no. 6, 253.

[40] C Kotanyi, *NGC 3309: An S-shaped radio galaxy in a nearby cluster*, Revista Mexicana de Astronomía y Astrofísica **21** (1990), 173–176.

[41] D Kozieł-Wierzbowska and G Stasińska, *FR II radio galaxies in the Sloan Digital Sky Survey: Observational facts*, Monthly Notices of the Royal Astronomical Society **415** (2011), 1013–1026.

[42] R. P Kraft, M. J Hardcastle, D. M Worrall, and S. S Murray, *A chandra study of the multicomponent X-Ray emission from the X-shaped radio galaxy 3C 403*, The Astrophysical Journal **622** (2005), no. 1, 149–159.

[43] S Kumari and S Pal, *A catalogue of newly discovered Hybrid Morphology Radio (HyMoR) galaxies from the VLA FIRST survey*, Conference: 21st national space science symposium; doi: http://dx.doi.org/10.13140/rg.2.2.30624.66568, 2022.

[44] _____, *Search for hybrid morphology radio galaxies from the FIRST survey at 1400 MHz*, Monthly Notices of the Royal Astronomical Society **514** (2022), 4290–4299.

[45] M Kunert-Bajraszewska, M. P Gawroski, A Labiano, and A Siemiginowska, *A survey of low-luminosity compact sources and its implication for the evolution of radio-loud active galactic nuclei – I. Radio data*, Monthly Notices of the Royal Astronomical Society **408** (2010).

[46] M Kunert-Bajraszewska, A Marecki, and P Thomasson, *FIRST-based survey of compact steep spectrum sources IV. Multifrequency VLBA observations of very compact objects*, Astronomy & Astrophysics **450** (2006).

[47] D. V Lal, M. J Hardcastle, and R. P Kraft, *'Normal' Fanaroff–Riley type II radio galaxies as a probe of the nature of X-shaped radio sources*, Monthly Notices of the Royal Astronomical Society **390** (2008), 1105.

[48] D. V Lal and A. P Rao, *Spectral structure of X-shaped radio sources*, Bulletin of the Astronomical Society of India **32** (2004), 247.

[49] _____, *Giant Metrewave Radio Telescope observations of X-shaped radio sources*, Monthly Notices of the Royal Astronomical Society **374** (2007), 1085–1102.

[50] D. V Lal, B Sebastian, C. C Cheung, and A Pramesh Rao, *GMRT low-frequency imaging of an extended sample of X-shaped radio galaxies*, The Astronomical Journal **157** (2019), 195.

[51] H Landt, C. C Cheung, and S. E Healey, *The optical spectra of X-shaped radio galaxies*, Monthly Notices of the Royal Astronomical Society **408** (2010), 1103–1112.

[52] H Landt, P Padovani, and P Giommi, *The classification of BL Lacertae objects: the Ca H&K break*, Monthly Notices of the Royal Astronomical Society **336** (2002), 945–956.

[53] H Landt, E. S Perlman, and P Padovani, *VLA observations of a new population of blazars*, The Astrophysical Journal **637** (2006), 183.

[54] L Lara, W. D Cotton, L Feretti, G Giovannini, J. M Marcaide, z I Marque, and T Venturi, *A new sample of large angular size radio galaxies*, Astronomy & Astrophysics **370** (2001), 409–425.

[55] J. P Leahy and P Parma, *Multiple outbursts in radio galaxies*, Extragalactic Radio Sources: From Beams to Jets, 1992, pp. 307.

[56] J. P Leahy and A. G Williams, *The bridges of classical double radio sources*, Monthly Notices of the Royal Astronomical Society **210** (1984), 929–951.

[57] X Liu, L Cui, W. F Luo, W. Z Shi, and H. G Song, *VLBI observations of nineteen GHz-peaked-spectrum radio sources at 1.6 GHz*, Astronomy & Astrophysics **470** (2007), 97–104.

[58] E. K Mahony et al., *The Lockman Hole project: LOFAR observations and spectral index properties of low-frequency radio sources*, Monthly Notices of the Royal Astronomical Society **463** (2016), 2997–3020.

[59] D Merritt and R. D Ekers, *Tracing black hole mergers through radio lobe morphology*, Science **297** (2002), 1310–1313.

[60] G. K Miley, G. C Perola, P. C Van Der Kruit, and H Van Der Laan, *Active galaxies with radio trails in clusters*, Nature **237** (1972), 269–272.

[61] B Mingo et al., *Revisiting the Fanaroff–Riley dichotomy and radio-galaxy morphology with the LOFAR Two-Metre Sky Survey (LoTSS)*, Monthly Notices of the Royal Astronomical Society **488** (2019), 2701–2721.

[62] Y. M Minnie, J. H Melanie, B. S Jamie, and J. W Simon, *Head–tail Galaxies: beacons of high-density regions in clusters*, Monthly Notices of the Royal Astronomical Society **392** (2009), 1070–1079.

[63] C. P O'Dea and F. N Owen, *Multifrequency VLA observations of the proto typical narrow-angle tail radio source, NGC 1265*, The Astrophysical Journal **301** (1986), 841.

[64] Windhorst R. A Oort M. J. A. Steemers W. J. G., *A deep 92 CM survey of the Lynx area*, Astronomy and Astrophysics Supplement Series **73** (1988), 103.

[65] F. N Owen and L Rudnick, *Radio sources with wide-angle tails in Abell clusters of galaxies*, The Astrophysical Journal **205** (1976), L1.

[66] S Pal and S Kumari, *Winged radio sources from LOFAR Two-metre Sky Survey First Data Release (LoTSS DR1)*, arXiv:2104.00410 (2021).

[67] ———, *A new catalogue of head-tail radio galaxies from LoTSS DR1*, Journal of Astrophysics and Astronomy, in press (2022).

[68] D Patra, S Pal, C Konar, and S. K Chakrabarti, *Multi-frequency properties of an interacting narrow-angle tail radio galaxy J0037+18*, Astrophysics and Space Science **364** (2019), 1–8.

[69] B Peng, R. R Chen, and R Strom, *Giant radio galaxies as probes of the ambient WHIM in the era of the SKA*, Advancing Astrophysics with the Square Kilometre Array (AASKA14) **109** (2015).

[70] D. D Proctor, *Morphological annotations for groups in the first database*, The Astrophysical Journal Supplement Series **194** (2011), 33.

[71] J. M Riley, *Observations of 3C 272.1 at 2·7 and 5·0 GHz*, Monthly Notices of the Royal Astronomical Society **157** (1972), 349–357.

[72] D. H Roberts, J. P Cohen, J Lu, L Saripalli, and R Subrahmanyan, *The abundance of X-shaped radio sources. I. VLA survey of 52 sources with off-axis distortions*, The Astrophysical Journal Supplement Series **220** (2015), 7.

[73] D. H Roberts, L Saripalli, K. X Wang, M. S Rao, R Subrahmanyan, C. C KleinStern, C. Y Morii-Sciolla, and Simepson L, *What are "X-shaped" radio sources telling us? I. Very Large Array imaging of a large sample of candidate XRGs*, The Astrophysical Journal **852** (2018), 47.

[74] L Rudnick and F. N Owen, *Head-tail radio sources in clusters of galaxies*, The Astronomical Journal **203** (1976), L107–L111.

[75] M Ryle and M. D Windram, *The radio emission from galaxies in the perseus cluster*, Monthly Notices of the Royal Astronomical Society **138** (1968), no. 1, 1–21.

[76] L Saripalli and D. H Roberts, *What are "X-shaped" radio sources telling us? II. Properties of a sample of 87*, The Astrophysical Journal **852** (2018), 48.

[77] L Saripalli and R Subrahmanyan, *The genesis of morphologies in extended radio sources: X-shapes, off-axis distortions, and giant radio sources*, The Astrophysical Journal **695** (2009), 156.

[78] T. K Sasmal, S Bera, S Pal, and S Mondal, *A New Catalog of Head–Tail Radio Galaxies from the VLA FIRST Survey*, The Astrophysical Journal Supplement Series **259** (2022), 31.

[79] R Subrahmanyan, L Saripalli, V Safouris, and R. W Hunstead, *WHIM environment of giant radio galaxies*, The Astrophysical Journal **677** (2008), 63.

[80] A.R Thompson, B.G Clark, C.M Wade, and P.J Napier, *The Very Large Array*, The Astrophysical Journal Supplement Series **44** (1980), 151–167.

[81] S Tremaine et al., *The slope of the black hole mass versus velocity dispersion correlation*, The Astrophysical Journal **574** (2002), 740.

[82] M. H Ulrich and J Rönnback, *The host of B2 0828+32, a radio galaxy with two sets of radio lobes*, The Astronomy and Astrophysics **313** (1996), 750–754.

[83] T. G Wang, H. Y Zhou, and X. B Dong, *4C +01.30: An X-shaped radio source with a quasar nucleus*, The Astronomical Journal **126** (2003), 113.

[84] D. M Worrall, M Birkinshaw, and R. A Cameron, *The X-ray environment of the dumbbell radio galaxy NGC 326*, The Astrophysical Journal **449** (1995), 93.

[85] L Xiang, C Reynolds, R. G Strom, and D Dallacasa, *European VLBI network observations of fourteen GHz-peaked-spectrum radio sources at 5 GHz*, Astronomy & Astrophysics **454** (2001), 729–740.

[86] Xiaolong Yang et al., *Extended catalog of winged or X-shaped radio sources from the FIRST survey*, The Astrophysical Journal Supplement Series **245** (2019), no. 1, 17.

[87] X. G Zhang, D Dultzin-Hacyan, and T. G Wang, *SDSS J1130+0058 an X-shaped radio source with double-peaked low-ionization emission lines: a binary black hole system?*, Monthly Notices of the Royal Astronomical Society **377** (2007), 1215.

[88] C Zier and P. L Biermann, *Binary black holes and tori in AGN. I. Ejection of stars and merging of the binary*, Astronomy & Astrophysics **377** (2001), 23–43.

Cyclotron Resonant Scattering Features in Highly Magnetized Neutron Stars

Manoj Mandal[1,*] and Sabyasachi Pal[1]

[1]Midnapore City College, Kuturia, Bhadutala, Paschim Medinipur, West Bengal, 721129, India
[*]Corresponding author: manojmandal@mcconline.org.in

Abstract

The cyclotron resonant scattering feature (CRSF) is important for estimating the magnetic fields of neutron stars directly. We have summarized results from different sources where the CRSFs have been discovered. Since their discovery, cyclotron lines have produced plenty of important results, which this review explores and summarizes. This review aims at summarising the importance of CRSFs, particularly in the context of studying the magnetic field and luminosity pulse phase dependence of the CRSFs. The evolution of cyclotron line parameters with luminosity has given a probe into how accretion geometry changes with accretion rate fluctuations. We also review the impact of CRSFs on different timing properties like pulse profile, pulse fraction, and emission geometry. Near the cyclotron line energy, pulse profile, pulse fraction, and beaming pattern show significant evolution. We have discussed the impact of the cyclotron line on different timing parameters for different pulsars.

Keywords: *Cyclotron Lines, Pulsars, Accretion, Accretion Disks, Magnetic Fields, High Energy Astrophysical Phenomena.*

1 Introduction

In X-ray astronomy, the study of cyclotron lines in accreting X-ray binaries or isolated neutron stars has grown into its own discipline. This is a thriving field, both from the perspective of theoretical and observational research. The study of Cyclotron Resonant Scattering Features (CRSF) is focused not only on the finding of new lines but also on several recently discovered features of these lines, such as the dependence of line energy, line width, and depth on X-ray luminosity. CRSF was first discovered in the spectrum of Her X-1 [108]. The majority of CRSFs are found in the 10–60 keV energy range. Important information on the emission geometry and many physical characteristics, such as the electron temperature and optical depth, are provided by CRSFs. The accretion physics of highly magnetized neutron stars can be explored using CRSF.

A few dozen X-ray pulsars (XRPs) have CRSFs, which are used as clear indications of a strong magnetic field at the surface of accreting neutron stars (NSs). The absorption-like characteristics in the energy spectrum are caused by the continuum photons' resonant scattering with electrons, which quantizes into Landau levels in the presence of strong magnetic fields [72]. Cyclotron lines are known to vary with accretion luminosity for a few XRPs. This is expected that variations in the geometry and dynamics of the accretion flow over the magnetic poles of NS are related to variations in the cyclotron line scattering features that have been seen. Due to accretion from hot spots at the magnetic poles, a positive correlation between the line centroid energy and luminosity is seen for the sub-critical regime of XRPs. In super-critical XRPs with large mass accretion rates, where radiation pressure sustains accretion columns above the star surface, the negative correlation is mostly seen.

The accretion rate, accretion geometry, and other factors all have a significant impact on the precise location of the X-ray emission region. It is also uncertain whether the CRSF region is consistent with the shock region or the emission zone of the X-ray continuum. In order to determine the magnetic field from the observed CRSF, it is typically assumed that the CRSF is formed somewhere in the accretion column, near to the neutron star surface. However, point to the CRSF generating region being close to the shock region. The luminosity dependence of CRSF properties is used to determine the plasma deceleration mechanism and the transition from subcritical to supercritical accretion regimes [5]. These accretion regimes are dependent on the "critical luminosity" L_{crit}, which strongly depends on the magnetic field of the neutron star [5], [7]. For $L > L_{crit}$, an accretion column

forms in the super-critical regime, and the falling plasma is slowed down by a radiation shock that originates at a specific distance from the neutron star surface. In this instance, when luminosity rises, emission height rises as well. In the sub-critical regime, for $L < L_{crit}$, the in-falling matter is presumably decelerated by Coulomb interaction forming a region whose height decreases with increasing luminosity. At even lower luminosity, the description of the deceleration process of falling material is not very conclusive.

The goal of this review is to provide an overview of the significance of CRSFs, particularly in the context of study on the magnetic field, luminosity, and pulse phase dependency of the CRSFs for various NSs. Section 2 describes the origin of CRSFs. Section 3 reviews the evolution of the cyclotron line for different sources. Sections 4 and 5 summarize results on the luminosity and pulse phase dependence of line energies, respectively. Section 6 reviews the magnetic field strength estimated from different sources. The impact of CRSFs on the timing properties is discussed in Section 7, and we have summarized conclusions in Section 8.

2 Origin of CRSFs in Different Neutron Stars

The most common cause of CRSF is the inelastic scattering of photons off electrons in a strong magnetic field and quantized the electron momentum perpendicular to the magnetic field [95], [14], [4]. The magnetic field of the neutron star can be directly calculated using this CRSF and the energy of the fundamental line and the distance between harmonics are exactly proportional to the magnetic field strength [76]. Sometimes, in a few instances, the cyclotron line can be created as a result of X-rays reflecting off the neutron star [85], [57].

The detected CRSF can be described using different spectral models. Here we have summarized the most used spectral models to explain the CRSFs.

1. `XSPEC` multiplicative model `cyclabs`

$$exp\left(\frac{-\tau_{cycl}(E/E_{cycl})^2\sigma_{cycl}^2}{(E-E_{cycl})^2+\sigma_{cycl}^2}\right) \tag{1}$$

where E_{cycl}, σ_{cycl} and τ_{cycl} are the line central energy, width and depth, respectively [74].

2. `XSPEC` multiplicative model `gabs`

$$exp\left[\left(\frac{-\tau_{cycl}}{\sqrt{2\pi}\sigma_{cycl}}\right)\ exp\left(\frac{-(E-E_{cycl})^2}{2\sigma_{cycl}^2}\right)\right] \tag{2}$$

where E_{cycl}, τ_{cycl}, and σ_{cycl} are the line central energy, depth, and width respectively (see equations 6 and 7 in [19]).

The normalization corresponds to the line depth is associated to the optical depth, which at the line center is given by

$$\tau = \frac{norm}{\sqrt{2}\pi\sigma} \tag{3}$$

3. A simple additive Gaussian with a negative normalization.

3 Evolution of Cyclotron Line for Different Pulsars

Sometimes the position of the cyclotron line for different pulsars evolved with time. The parameters related to CRSF also show variations with time for different sources. Table 1 summarizes different CRSF sources and their corresponding magnetic fields. The CRSF line energy evolves with time in sources like Her X–1 [98] and 4U 1538–522 [38]. In the pulse phase averaged spectra from 1996 to 2012, Her X—1 indicated evidence of a long-term decline in the centroid energy of the CRSF [99].

Here we have discussed a few sources, where the presence of CRSF was discovered or CRSF was observed recently in an outburst.

1. **1A 0535+262:** This is a well known pulsar with a spin period ∼104 s [92] and an orbital period ∼110.3 d [29]. During the 1989 outburst, a fundamental cyclotron resonance feature was detected near 50 keV and the first harmonic was found near 100 keV [46]. This source went through a giant outburst in 2020 [67], [82] and the flux

Table 1: A summary of sources with detected cyclotron lines. The sources are arranged in increasing order of cyclotron line energy. The sources with the mark "" imply that the CRSFs have claimed but need more confirmation.

	Source	**Type**	**E_{cyc} (keV)**	**Ref.**	**Magnetic Field (10^{12}) G**
1	GRO J1744–28	LMXB	4.7, 10.4, 15.8	[23]	0.53
2	XMMU J054134.7–682550	HMXB	9	[68]	1.0
3	Swift J1626.6–5156	Be pers	10, 18	[20], [24]	1.1
4	GRO J2058+42	HMXB	10, 20, 30	[75]	2.0
5	4U 0115+63	Be trans.	12, 24, 36, 48, 62	[117]	1.2
6	KS 1947+300	Be trans.	12	[31]	1.2
7	NGC300 ULX1	Be HMXB	13	[116]	–
8	XTE J1829–098	Be trans.	15	[97]	1.7
9	4U 1700–37	HMXB	16	[3]	2.1
10	IGR J17544–2619	Be trans.	17	[9]	2.0
11	4U 1907+09	HMXB	18, 36	[59], [90]	2.1
12	IGR J18179–1621	HMXB	21.5	[55]	2.4
13	4U 1538–52	HMXB	22, 47	[17], [15]	2.5
14	IGR J18027–2016	HMXB	23	[58]	3.0
15	2S 1553–542	Be trans.	23.5	[113]	3.0
16	Vela X–1	HMXB	24, 52	[59], [51]	2.6
17	SMC X–2	HM trans.	27	[41]	2.3
18	V 0332+53	Be trans.	28, 51, 74	[60]	2.5
19	IGR J16393–4643	HMXB	29	[11]	2.5
20	4U 0352+309 (X-Per)	Be XRB	29	[18]	3.3
21	4U 2206+54	HMXB	29–35	[107], [70]	3.3
22	Cen X–3	HMXB	30	[12]	3.5
23	Cep X–4	Be trans.	30, 45	[73], [40]	2.6
24	4U 1901+03	HMXB	30	[8]	3.5
25	IGR J16493–4348	SG HMXB	30	[22]	3.7
26	RX J0520.5–6932	Be trans.	31	[105]	2.0
27	RX J0440.9+4431	Be trans.	32	[109]	3.2
28	4U 1822–371	LMXB	0.7, 33	[39], [94]	2.8
29	MXB 0656–072	HMXB	33	[71]	3.6
30	XTE J0658–073	Be trans.	35	[37]	–
31	EXO 2030+375	Be trans.	36	[86]	–
32	XTE J1946+274	Be trans.	36	[36]	3.8
33	IGR J19294+1816	Be trans.	36	[93]	4.1
34	Her X–1	LMXB	37	[108]	3.8
35	GX 301–2	HMXB	37	[59], [102]	4.1
36	4U 1626–67	LMXB	37	[79]	3.8
37	4U 1909+07	HMXB	44	[43]	3.8
38	MAXI J1409–619	HMXB	44, 73, 128	[80]	3.8
39	1A 0535+262	Be trans.	45, 100	[51], [46]	5.0
40	XTE J1858+034	HMXB	48	[114], [62]	5.2
41	GX 304–1	Be trans.	54	[120]	4.8
42	1A 1118–616	Be trans.	55	[26]	4.8
43	GRO J1008–57	Be trans.	78	[96], [119]	6.6

reached a record high of ∼12 Crab as measured by Swift/BAT (15–50 keV). A transition from sub-critical to super-critical accretion regime happened [82]. During the transition, the study of the evolution of CRSF parameters is significant and interesting. The source was studied using NuSTAR observations and the presence of cyclotron lines was detected in 40–46 keV during several days of the outburst [65]. The fundamental cyclotron line and first harmonic was found near ∼46 keV and ∼100 keV respectively [65], [47], [33], [101]. About five days before the peak of a typical (type-I) outburst, during the pre-outburst flare, a higher CRSF line energy of sim50 keV was discovered [13], [84]. The cyclotron line energy varied significantly with luminosity during the giant outburst of 2020. The mass accretion rate changed significantly during the transition, and the beaming pattern and emission geometry also evolved significantly. The luminosity dependence and evolution of cyclotron energy during the outburst were studied during the different phases of the outburst.

Figure 1 represents a cyclotron absorption feature in the spectrum during a NuSTAR observation [65]. The feature was observed near 43 keV during this observation, and the line energy varied between 40–46 keV during the outburst. Most of the NuSTAR observations were performed when the source was in the super-critical regime with high mass accretion rates. A negative correlation between the line energy and luminosity is observed in the supercritical regime [65].

2. **XTE J1858+034:** In 1998, the All-Sky Monitor on the RXTE observatory discovered the transient X-ray pulsar (XRP) XTE J1858+034 [88]. During the same outburst, pulsation with a period of ∼221 s was found in the RXTE/PCA data [104]. A cyclotron absorption feature was found at ∼48 keV during the 2019 outburst in both the pulse phase averaged and resolved spectra, [114], [62]. The associated magnetic field was calculated to be 5.2×10^{12} G. At the same time, the presence of a low quasi-periodic oscillation was also detected at ∼196 mHz [63]. Due to a single NuSTAR observation, it was not possible to study the luminosity dependence of the CRSFs. So, the correlation study between the line energy and luminosity in different accretion regimes is still not done yet, which may provide important information related to the accretion phenomenon. The phase dependence of the CRSF was investigated. The single-peaked, sine-like shape of the XTE J1858+034 pulse profile is energy independent, and an indication of phase lag is seen with the soft profile trailing the hard one. Variations of the spectral continuum over the pulse phase are also consistent with a pencil-beam emission [114].

3. **2S 1553–542:** The accreting X-ray pulsars with Be optical companions subclass includes the transient source 2S 1553–542 [1]. 2S 1553–542 was discovered by the SAS-3 observatory in 1975 during the Galactic Plane study. Later, the transient character of the source was confirmed, and significant pulsations with a period of 9.3 s and an amplitude of approximately 80% were discovered [45] and an orbital period of nearly 30 d [81]. Phase-averaged and phase-resolved spectra both showed an absorption feature at sim23.5 keV, which was identified as the cyclotron resonant scattering feature corresponding to the magnetic field strength of the neutron star B $\sim 3 \times 10^{12}$ G [113]. Recently, the source went through an outburst in 2021 [44], [66], [54]. The luminosity dependence of cyclotron line parameters studied by [61] during the 2021 outburst and a state transition is reported from the subcritical to supercritical accretion regime above 4×10^{37} erg s^{-1}. [64] studied the luminosity dependence of the pulse fraction and photon index for this source using NICER and NuSTAR observations and also looked for the impact of the cyclotron line on timing parameters during the recent outburst in 2021. The pulse profile and pulse fraction showed an evolution near the cyclotron line energy. The pulsed fraction variation can be explained by a local minimum of about 25 keV overlaid on the progressive energy increase that is typical of most X-ray pulsars [56].

4. **GRO J2058+42:** During a type-II (giant) outburst in 1995 September, the Burst and Transient Source Experiment (BATSE) on board the Compton Gamma-Ray Observatory (CGRO) found GRO J2058+42, a slowly revolving transient XRP [118]. During the type-II outburst of 2019, the fundamental cyclotron absorption line was discovered at a sim10 keV energy, and the associated magnetic field can be estimated to be $B \sim 10^{12}$ G [75]. From the NuSTAR spectrum, two harmonics of the cyclotron line were discovered near sim20 keV and sim30 keV, respectively [75].

5. **XTE J1829–098 :** XTE J1829–098 was discovered by the RXTE observatory during scans of the Galactic plane in July 2004. It has been identified as a transient X-ray pulsar with a pulse period of ∼7.8 s [69]. A strong absorption feature was detected at an energy of 15 keV in the source spectrum during the outburst of August 2018. This feature was interpreted as a cyclotron resonant scattering line and the corresponding magnetic field strength of the neutron star surface was estimated to be $B = 1.7 \times 10^{12}$ G [97]. The cyclotron line is considerably detected at all phases of the pulse and its energy and other properties change across the pulse period, according to the pulse

phase-resolved spectroscopy. The pulsed fraction changes as energy changes, and this includes a local rise around the cyclotron line.

6. **IGR J19294+1816:** The INTEGRAL observatory discovered IGR J19294+1816 on March 27, 2009 [115]. Prior to this, the RXTE data was used to report a tentative detection of the cyclotron absorption feature at ∼35.5 keV based on the source spectrum's deviation from a power law at higher energies [93]. The archival Swift/XRT data of this region showed a relatively bright source with signs of pulsations at 12.4 s [91]. A cyclotron absorption line was detected in the energy spectrum of the bright state at $E_{cyc} \sim 42.8$ keV, and the corresponding magnetic field was calculated to be 5×10^{12} G [112]. Phase-resolved spectroscopy was used to investigate how the cyclotron line energy and optical depth rely on the pulse phase. Although the depth of the cyclotron line appears to vary significantly, the cyclotron line does not show a significant pulse phase dependence.

7. **IGR J18027–2016:** IGR J18027–2016, a faint persistent hard X-ray source, was discovered using the INTEGRAL observatory during deep scans of the Galactic Center [89]. The spin period was found to be ∼139.8 s and the spin-down rate was $\simeq 6 \times 10^{-10}$ s s^{-1} [58]. The NuSTAR energy spectrum revealed a possible cyclotron absorption feature at energy near 23 keV [58]. This energy corresponds to the magnetic field $B \simeq 3 \times 10^{12}$ G at the surface of the neutron star. Near the cyclotron line centroid energy, it was found that the pulse profile and pulse fraction significantly influenced on energy.

8. **IGR J16393–4643:** IGR J16393–4643 was discovered by ASCA [103] and was re-discovered in high energies by INTEGRAL [10]. A consistent pulse period of 911 s indicative of a slowly revolving, magnetized neutron star was discovered by XMM-Newton and INTEGRAL data. Later, the pulsation from this source was confirmed using RXTE, Chandra, and Suzaku observations [106]. NuSTAR observed the source in 2014 with an exposure of 50 ks. The pulsation was found at ∼904 s [11]. The source has undergone a long spin-up trend since 2006 at a rate of -2×10^{-8} s s^{-1} [11]. NuSTAR spectra were used to discover a cyclotron resonant scattering feature with centroid energy of ∼29.3 keV [11]. The magnetic field corresponding to the cyclotron line was estimated to be B = 2.5 × 10^{12} G.

9. **SMC X-2:** The transient source SMC X-2 was discovered during an outburst in 1977 in the Small Magellanic Cloud [16] using SAS 3. SMC X-2 was observed several times using different satellites in multi-wavelength observations. The source was classified as an X-ray pulsar of period 2.37 s using ASCA and RXTE [21], [121]. The presence of a cyclotron line was detected near 27 keV using NuSTAR during the 2015 outburst [41]. A negative correlation between the cyclotron line energy and luminosity was found for SMC X-2 [41]. The change in shock height or the line-forming zone with luminosity was proposed as the cause of the negative correlation between cyclotron line energy and luminosity. The corresponding magnetic field related to the CRSF feature was estimated to be 2.3 ×10^{12} G [41]. The pulse phase dependence of CRSF features was studied and variability was observed. The pulse phase dependence of the CRSF parameters was explained as because of either a complex structure of the pulsar magnetic field or due to the effect of emission geometry [41].

4 Luminosity Dependence of Cyclotron Line

It is interesting to study the variation of cyclotron line energy with luminosity to probe the accretion geometry. Cyclotron energy and luminosity are correlated in some accretion-powered pulsars. The cyclotron line energy and luminosity are likely to correlate negatively in the supercritical regime and positively in the subcritical regime [7]. The mass accretion rate changes significantly with different phases of an outburst, which has an impact on the evolution of CRSF parameters. We have summarized different sources for which luminosity dependence of the cyclotron line energies was observed and tried to understand the mechanism behind it.

The change in height of the accretion column, which is sustained by radiative pressure and only becomes evident above a certain critical luminosity, is generally thought to be responsible for the reported anti-correlation at high mass accretion rates [5]. It is found that the cyclotron line energy positively correlates with the luminosity below the critical luminosity, either as a result of the Doppler effect [77] or as a result of a change in the atmosphere height above the NS surface brought on by the in-falling material ram pressure [100].

The X-ray pulsar 1A 0535+262 was observed several times with NuSTAR during the giant outburst of 2020. There is an opportunity to study the luminosity dependence of the CRSF at a high luminosity for the first time. During the giant outburst, a transition from a subcritical to a supercritical accretion regime took place. Most of

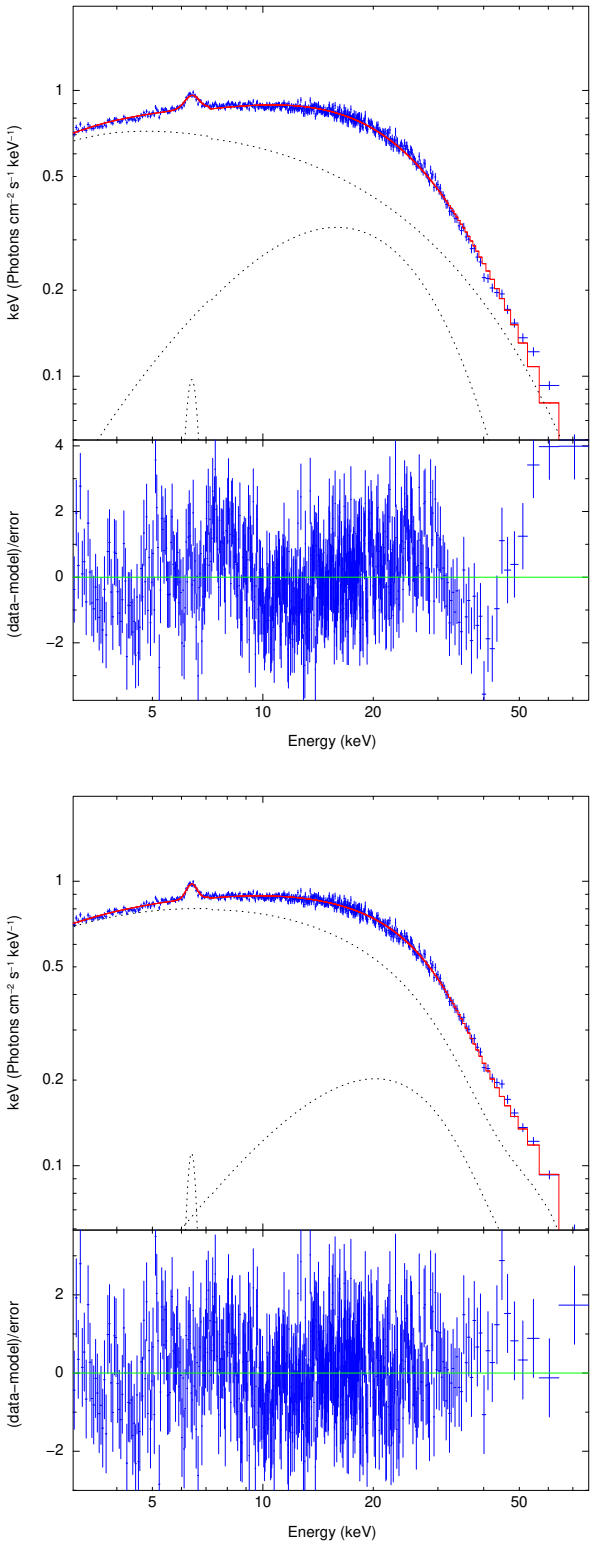


Figure 1: The left-side image represents the unfolded spectrum of 1A 0535+262 during the giant outburst of 2020 [65] and the right-side image shows the spectrum with the best-fitted model with gabs along with other model components (absorbed cut-off power law, a blackbody component, and an iron emission line at 6.4 keV). The left-side figure shows a Gaussian absorption feature near 43 keV, which is absent in the right-side figure.

the NuSTAR observations for this source were conducted when the mass accretion rate was high. A negative correlation between the cyclotron line energy and luminosity was reported above the critical luminosity, which is shown in Figure 2, and the estimated critical luminosity was found $\sim 6 \times 10^{37}$ erg s^{-1} [65]. The cyclotron line energy and luminosity are observed to correlate positively in the subcritical regime (shown in Figure 2). For 1A 0535+262 a transition from the subcritical to the supercritical regime is being discovered for the first time. The hardness intensity diagram (HID) further supported the state transition. During the 2020 giant outburst, the HID exhibited a change from the horizontal to the diagonal branch [65].

Earlier, the correlation between the cyclotron line energy and luminosity was observed for several sources. A positive correlation was observed between line energy and luminosity for the sources like Vela X–1 [32], Her X–1 [100], Cep X–4 [30], GX 304–1 [48], and Swift J1626.6–5156 [24] . An anti-correlation between the E_{cyc} and the X-ray luminosity was confirmed for the sources V 0332+53 [27], [111], [110], 4U 0115+6415 [78], and SMC X–2 [41].

The critical luminosity (L_{crit}) directs if the radiation pressure of emitting plasma is capable of decelerating the accretion flow and helps to determine two different accretion regimes.If $LX\ Lcrit$ (sub-critical regime), then the accreted material enters the neutron star surface via nuclear collisions with atmospheric protons (pencil beam pattern) or coulomb collisions with thermal electrons. The radiation pressure is high enough to stop the accreting matter at a distance above the neutron star when $LX > Lcrit$ (super-critical regime). This results in the formation of a radiation-dominated shock (fan beam pattern) [50], [49], [6]. These accretion regimes can also be investigated by observing changes in cyclotron line energy, pulse profiles, and spectral shape [87], [83].

5 Pulse Phase Dependence of the Cyclotron Absorption Features

For rotating neutron stars, pulse phase-resolved spectroscopy of the cyclotron parameters is an important tool for examining the emission geometry at various viewing angles. It can also be utilized to map the neutron star magnetic field geometry. The emission geometry or the beaming pattern can be traced using the pulse phase dependence of the CRSFs. In a simple "fan-beam" radiation pattern, a deeper CRSF is seen close to the peak of the profile and a shallower one at its decreasing edges. For a fan beam pattern, the CRSF centroid energy is also higher close to the peaks. In this case, the phases suggesting narrower and deeper CRSFs at higher centroid energies might be referring to the region influenced by fan-beaming patterns, and the phases denoting broader and shallower CRSFs at lower energies might be referring to the region dominated by "pencil-beam" patterns [95].

The Landau levels are not fully equidistant in the relativistic regime. The angular dependency of the cyclotron line energy can be expressed as [95], [35]

$$E = m_e c^2 \times \frac{\sqrt{1 + 2n\frac{B}{B_{crit}} sin^2\theta} - 1}{sin^2\theta} \times (1+z)^{-1} \tag{4}$$

where c is the speed of light, m_e is the electron mass, θ is the direction between the incident photon and the magnetic field vector, and B_{crit} is the critical magnetic field strength $\sim 44 \times 10^{12}$ G.

An uneven spacing results from the CRSF fundamental and harmonic energies dependency on θ. The observed spectral features are affected by the plasma geometry and the intrinsic alignment of the magnetic field as these energies depend on the photon propagation angle with respect to the field.

There are several sources where variations of the cyclotron centroid line energy was found with the pulse phase, e.g., 4U 1538–52 [17], Her X-1 [34], Vela X-1 [52], and GX 301–2 [53]. A strong dependence on the CRSF feature was observed for the source 4U 1901+03. The pulse-phase spectroscopy revealed the 30 keV absorption feature, and E_{cyc} varied up to $\sim$60% [8]. Such variation can be attributed to non-dipolar geometry, complicated accretion geometry, the large gradient in the magnetic field or sampling at various heights of the line-forming area with pulse phase [101].

6 Estimation of Magnetic Field from Cyclotron Line

The CRSFs are used to measure the magnetic fields of neutron stars precisely. Landau levels are discrete energies that electrons have when moving perpendicular to the magnetic field. Resonant photon scattering on these electrons results in photon scattering at the resonance energy, resulting in resonant absorption characteristics in the X-ray spectrum. The fundamental energy is equal to the difference in energy between neighboring Landau levels,

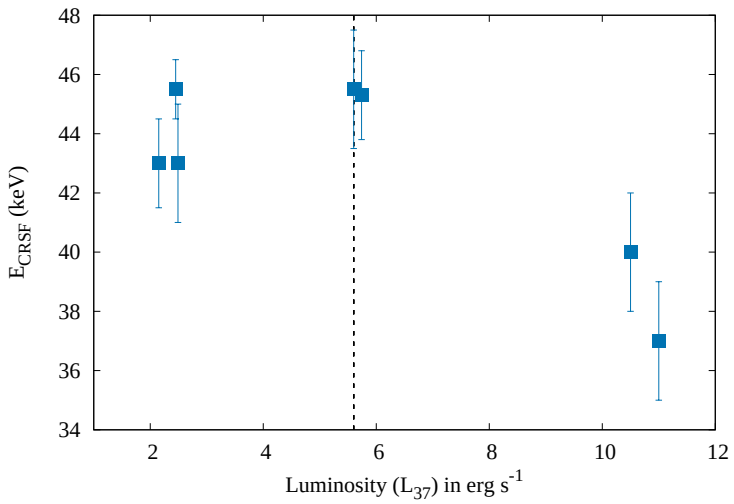


Figure 2: An anti-correlation between the cyclotron line energy and luminosity is observed in the super-critical regime for 1A 0535+262 during the giant outburst of 2020. The vertical dotted line represents the critical luminosity. Below the critical luminosity, the correlation is positive. We have used the value of E_{CRSF} and luminosity from [65].

which is given by $\hbar\omega$, where

$$\omega = \frac{eB}{m_e c} \tag{5}$$

is the cyclotron frequency, B is the magnetic field in the scattering region, e is the electron charge, c is the speed of light and m_e is the mass of the electron. The magnetic field strength is linearly related to the Landau levels (to first order equidistant). The relativistic generalization can be written as

$$\omega = \frac{eB}{\gamma m_e c} \tag{6}$$

where γ is the relativistic factor. If r_L is the radius of the electron trajectory perpendicular to the magnetic field and $v_\perp$ is the perpendicular velocity of the electron in the magnetic field B then the Larmor radius is found from

$$2\pi r_L = v_\perp \frac{2\pi}{\omega} \tag{7}$$

which gives

$$r_L = v_\perp \frac{m_e c}{eB} \tag{8}$$

If B_{crit} is the critical magnetic field strength ($\sim 44 \times 10^{12}$ G), where the Gyration radius is below the de Broglie wavelength (non-relativistic case). For $B << B_{crit}$, the magnetic field of the neutron star is related to the cyclotron energy, which can be expressed from Equation 5 in a relevant form to X-ray astronomy (*12-B-12-rule*)

$$E_{CRSF} = \hbar\omega = 11.57 \times B_{12} \frac{n}{(1+z)} \tag{9}$$

where, z is the gravitational red-shift ($z \sim$0.3 for typical neutron star) [95], [25], B_{12} is the magnetic field in 10^{12} G, and n = 1 for the fundamental line and $n \geq 2$ for the harmonics. For the fundamental line (n = 1), scattering has occurred from the ground level to the first excited Landau level and for $n \geq 2$, the lines are known as harmonics. We have summarized CRSF sources and their corresponding magnetic fields in Table 1.

7 Impact of CRSFs on Timing Properties

The emission geometry and beaming pattern affect near the CRSF line energy for different sources. The pulse profile, pulse fraction, and emission pattern evolve near the energy close to the CRSF. A trend of increase in the

pulse fraction is observed at high energies (20–50 keV), however, there is a noticeable change in the pulse profile and pulse fraction of some sources, such as Her X–1, GX 301–2, and 1A 1118–61. Surprisingly, all of these sources exhibit a deep CRSF in their energy spectra. Sources like 4U 1907+09, Vela X–1, and 4U 1626–67 are featureless near their CRSF fundamental energies.

Recently, a significant variation in pulse profile and pulsed fraction near the cyclotron line energy was observed for 1A 0535+262 [65] and pulse fraction showed a local maxima near this feature during the giant outburst of 2020. The pulse profile showed a double peak feature in low energy ranges and evolved to a single peak feature in high energy near the CRSF fundamental energy. Earlier, several studies have found that the significant change of pulse profiles occurs near the cyclotron line energies for different sources, like V 0332+53 [110], 4U 0115+63 [28], and 4U 1901+03 [8]. The scattering cross-sections are believed to be changed significantly near the CRSF energy [2], which may affect the change of the beaming patterns of the source.

Earlier, 4U 0115+63 exhibited notable changes close to the pulse peak at the CRSF fundamental and its related harmonics [28]. In addition, V 0332+53 revealed an asymmetrical single-peaked structure close to its fundamental CRSF energy, in contrast to the profile's other double-peaked features [110]. Integral observations were used to study the energy and luminosity dependence of the pulse profiles in the hard X-ray region (20-100 keV) [56]. The main outcome of this study was a general trend of an increase in the pulse fractions with energy for most of the sources, with local maxima for some sources near the cyclotron line energy. The pulse profile of GX 304–1 revealed a phase shift close to the cyclotron line energy [42].

8 Conclusion

We have summarized the X-ray sources for which the CRSFs have been observed. We have discussed sources for which CRSFs were discovered recently or CRSFs were detected during recent outbursts. A study into the change in accretion geometry with variations in the accretion rate has been made possible by the variation of cyclotron line parameters with luminosity. The X-ray pulsar 1A 0535+262 recently went through a giant outburst and we have found that at the super-critical regime the source showed an anti-correlation between the cyclotron line energy and luminosity. For the pulsar 1A 0535+262, the critical luminosity was found to be $\sim 6 \times 10^{37}$ erg s^{-1} above which a transition from sub-critical to super-critical state occurred. The transition from the subcritical regime to the supercritical regime is also observed from the hardness intensity diagram for 1A 0535+262 during the giant outburst of 2020. A transition from a horizontal to a diagonal branch indicates a state transition. The distribution of the plasma temperature, magnetic field, and optical depth in the line-forming region, as well as the beaming pattern and the accretion geometry, have all been revealed by studying the CRSF from various viewing angles. We have also summarized different sources for which strong phase dependence of CRSF was observed.

The luminosity of the sources affects the beaming pattern. A "pencil beam"-shaped X-ray emission is produced when material falls directly onto the neutron star surface while the source luminosity is below the critical luminosity. The material generates a radiation-dominated shock in the accretion column, where it is slowed down and gradually sinks to the neutron star surface if the source luminosity is greater than the critical luminosity. In this case, X-rays are released as a "fan-beam" from the sides of the accretion column. The shock zone goes up higher in the accretion column, where the magnetic field is weaker, as the accretion rate increases. For a CRSF emission region close to the shock region, the CRSF centroid energy falls with rising luminosity.

References

[1] K. M. V. Apparao, H. V. Bradt, R. G. Dower, R. E. Doxsey, J. G. Jernigan, and F. Li, *Positions of galactic X-ray sources: 320° ¡ l^{II} ¡340°*, Nature **271** (January 1978), no. 5642, 225–228.

[2] Rafael A. Araya and Alice K. Harding, *Cyclotron Line Features from Near-critical Magnetic Fields: The Effect of Optical Depth and Plasma Geometry*, The Astrophysical Journal **517** (May 1999), no. 1, 334–354.

[3] Suman Bala, Jayashree Roy, and Dipankar Bhattacharya, *Possible detection of a new cyclotron feature in 4U 1700-37*, Monthly Notices of the Royal Astronomical Society **493** (April 2020), no. 2, 3045–3053.

[4] Ralf Ballhausen, Katja Pottschmidt, Felix Fürst, Jörn Wilms, John A. Tomsick, Fritz-Walter Schwarm, Daniel Stern, Peter Kretschmar, Isabel Caballero, Fiona A. Harrison, Steven E. Boggs, Finn E. Christensen, William W. Craig, Charles J. Hailey, and William W. Zhang, *Looking at A 0535+26 at low luminosities with NuSTAR*, Astronomy & Astrophysics **608** (2017), A105.

[5] M. M. Basko and R. A. Sunyaev, *The limiting luminosity of accreting neutron stars with magnetic fields.*, Monthly Notices of the Royal Astronomical Society **175** (1976), 395–417.

[6] P. A. Becker, D. Klochkov, G. Schönherr, O. Nishimura, C. Ferrigno, I. Caballero, P. Kretschmar, M. T. Wolff, J. Wilms, and R. Staubert, *Spectral formation in accreting X-ray pulsars: bimodal variation of the cyclotron energy with luminosity*, Astronomy & Astrophysics **544** (August 2012), A123.

[7] Becker, P A, Klochkov, D, Schönherr, G, Nishimura, O, Ferrigno, C, Caballero, I, Kretschmar, P, Wolff, M T, Wilms, J, and Staubert, R, *Spectral formation in accreting x-ray pulsars: bimodal variation of the cyclotron energy with luminosity*, Astronomy & Astrophysics **544** (2012), A123.

[8] Aru Beri, Tinku Girdhar, Nirmal K. Iyer, and Chandreyee Maitra, *Evolution of timing and spectral characteristics of 4U 1901+03 during its 2019 outburst using the Swift and NuSTAR observatories*, Monthly Notices of the Royal Astronomical Society **500** (January 2021), no. 1, 1350–1365.

[9] Varun Bhalerao, Patrizia Romano, John Tomsick, Lorenzo Natalucci, David M. Smith, Eric Bellm, Steven E. Boggs, Deepto Chakrabarty, Finn E. Christensen, William W. Craig, Felix Fuerst, Charles J. Hailey, Fiona A. Harrison, Roman A. Krivonos, Ting-Ni Lu, Kristin Madsen, Daniel Stern, George Younes, and William Zhang, *NuSTAR detection of a cyclotron line in the supergiant fast X-ray transient IGR J17544-2619*, Monthly Notices of the Royal Astronomical Society **447** (March 2015), no. 3, 2274–2281.

[10] A. J. Bird, E. J. Barlow, L. Bassani, A. Bazzano, A. Bodaghee, F. Capitanio, M. Cocchi, M. Del Santo, A. J. Dean, A. B. Hill, F. Lebrun, G. Malaguti, A. Malizia, R. Much, S. E. Shaw, J. B. Stephen, R. Terrier, P. Ubertini, and R. Walter, *The First IBIS/ISGRI Soft Gamma-Ray Galactic Plane Survey Catalog*, The Astrophysical Journal Letters **607** (May 2004), no. 1, L33–L37.

[11] Arash Bodaghee, John A. Tomsick, Francesca M. Fornasini, Roman Krivonos, Daniel Stern, Kaya Mori, Farid Rahoui, Steven E. Boggs, Finn E. Christensen, William W. Craig, Charles J. Hailey, Fiona A. Harrison, and William W. Zhang, *NuSTAR Discovery of a Cyclotron Line in the Accreting X-Ray Pulsar IGR J16393-4643*, The Astrophysical Journal **823** (June 2016), no. 2, 146.

[12] L. Burderi, T. Di Salvo, N. R. Robba, A. La Barbera, and M. Guainazzi, *The 0.1-100 KEV Spectrum of Centaurus X-3: Pulse Phase Spectroscopy of the Cyclotron Line and Magnetic Field Structure*, The Astrophysical Journal **530** (February 2000), no. 1, 429–440.

[13] I. Caballero, A. Santangelo, P. Kretschmar, R. Staubert, K. Postnov, D. Klochkov, A. Camero-Arranz, M. H. Finger, I. Kreykenbohm, K. Pottschmidt, R. E. Rothschild, S. Suchy, J. Wilms, and C. A. Wilson, *The pre-outburst flare of the A 0535+26 August/September 2005 outburst*, Astronomy & Astrophysics **480** (March 2008), no. 2, L17–L20.

[14] V. Canuto and J. Ventura, *Quantizing Magnetic Fields in Astrophysics*, Fund. Cosmic Phys. **2** (1977).

[15] G. Clark, R. Doxsey, F. Li, J. G. Jernigan, and J. van Paradijs, *On two new X-ray sources in the SMC and the high luminosities of the Magellanic X-ray sources.*, The Astrophysical Journal Letters **221** (April 1978), L37–L41.

[16] ———, *On two new X-ray sources in the SMC and the high luminosities of the Magellanic X-ray sources.*, The Astrophysical Journal **221** (April 1978), L37–L41.

[17] George W. Clark, Jonathan W. Woo, Fumiaki Nagase, Kazuo Makishima, and Taro Sakao, *Discovery of a Cyclotron Absorption Line in the Spectrum of the Binary X-Ray Pulsar 4U 1538-52 Observed by GINGA*, The Astrophysical Journal **353** (April 1990), 274.

[18] W. Coburn, W. A. Heindl, D. E. Gruber, R. E. Rothschild, R. Staubert, J. Wilms, and I. Kreykenbohm, *Discovery of a Cyclotron Resonant Scattering Feature in the Rossi X-Ray Timing Explorer Spectrum of 4U 0352+309 (X Persei)*, The Astrophysical Journal **552** (May 2001), no. 2, 738–747.

[19] W. Coburn, W. A. Heindl, R. E. Rothschild, D. E. Gruber, I. Kreykenbohm, J. Wilms, P. Kretschmar, and R. Staubert, *Magnetic Fields of Accreting X-Ray Pulsars with the Rossi X-Ray Timing Explorer*, The Astrophysical Journal **580** (November 2002), no. 1, 394–412.

[20] W. Coburn, K. Pottschmidt, and R. Rothschild, Bulletin of the Amer-ican Astronomical Society **38** (2006).

[21] R. H. D. Corbet, F. E. Marshall, M. J. Coe, S. Laycock, and G. Handler, *The Discovery of an Outburst and Pulsed X-Ray Flux from SMC X-2 Using the Rossi X-Ray Timing Explorer*, The Astrophysical Journal Letters **548** (February 2001), no. 1, L41–L44.

[22] A. D'Aì, G. Cusumano, V. La Parola, A. Segreto, T. di Salvo, R. Iaria, and N. R. Robba, *Evidence for a resonant cyclotron line in IGR J16493-4348 from the Swift-BAT hard X-ray survey*, Astronomy & Astrophysics **532** (August 2011), A73.

[23] A. D'Aì, T. Di Salvo, R. Iaria, J. A. García, A. Sanna, F. Pintore, A. Riggio, L. Burderi, E. Bozzo, T. Dauser, M. Matranga, C. G. Galiano, and N. R. Robba, *GRO J1744-28: an intermediate B-field pulsar in a low-mass X-ray binary*, Monthly Notices of the Royal Astronomical Society **449** (June 2015), no. 4, 4288–4303.

[24] Megan E. DeCesar, Patricia T. Boyd, Katja Pottschmidt, Jörn Wilms, Slawomir Suchy, and M. Coleman Miller, *The Be/X-Ray Binary Swift J1626.6-5156 as a Variable Cyclotron Line Source*, The Astrophysical Journal **762** (January 2013), 61.

[25] V. Doroshenko, A. Santangelo, R. Doroshenko, I. Caballero, S. Tsygankov, and R. Rothschild, *XMM-Newton observations of 1A 0535+262 in quiescence*, Astronomy & Astrophysics **561** (2014), A96.

[26] V. Doroshenko, S. Suchy, A. Santangelo, R. Staubert, I. Kreykenbohm, R. Rothschild, K. Pottschmidt, and J. Wilms, *RXTE observations of the 1A 1118-61 in an outburst, and the discovery of a cyclotron line*, Astronomy & Astrophysics **515** (June 2010), L1.

[27] Victor Doroshenko, Sergey S. Tsygankov, Alexander A. Mushtukov, Alexander A. Lutovinov, Andrea Santangelo, Valery F. Suleimanov, and Juri Poutanen, *Luminosity dependence of the cyclotron line and evidence for the accretion regime transition in V 0332+53*, Monthly Notices of the Royal Astronomical Society **466** (2017), no. 2, 2143–2150.

[28] C. Ferrigno, M. Falanga, E. Bozzo, P. A. Becker, D. Klochkov, and A. Santangelo, *4U 0115+63: phase lags and cyclotron resonant scattering*, Astronomy & Astrophysics **532** (August 2011), A76.

[29] M. H. Finger, R. B. Wilson, and B. A. Harmon, *Quasi-periodic Oscillations during a Giant Outburst of A0535+262*, The Astrophysical Journal **459** (March 1996), 288.

[30] F. Fürst, K. Pottschmidt, H. Miyasaka, V. Bhalerao, M. Bachetti, S. E. Boggs, F. E. Christensen, W. W. Craig, V. Grinberg, C. J. Hailey, F. A. Harrison, J. A. Kennea, F. Rahoui, D. Stern, S. P. Tendulkar, J. A. Tomsick, D. J. Walton, J. Wilms, and W. W. Zhang, *Distorted Cyclotron Line Profile in Cep X-4 as Observed by NuSTAR*, The Astrophysical Journal Letters **806** (2015), no. 2, L24.

[31] Felix Fürst, Katja Pottschmidt, Jörn Wilms, Jamie Kennea, Matteo Bachetti, Eric Bellm, Steven E. Boggs, Deepto Chakrabarty, Finn E. Christensen, William W. Craig, Charles J. Hailey, Fiona Harrison, Daniel Stern, John A. Tomsick, Dominic J. Walton, and William Zhang, *NuSTAR Discovery of a Cyclotron Line in KS 1947+300*, The Astrophysical Journal Letters **784** (April 2014), no. 2, L40.

[32] Felix Fürst, Katja Pottschmidt, Jörn Wilms, John A. Tomsick, Matteo Bachetti, Steven E. Boggs, Finn E. Christensen, William W. Craig, Brian W. Grefenstette, Charles J. Hailey, Fiona Harrison, Kristin K. Madsen, Jon M. Miller, Daniel Stern, Dominic J. Walton, and William Zhang, *NuSTAR Discovery of a Luminosity Dependent Cyclotron Line Energy in Vela X-1*, The Astrophysical Journal **780** (2014), no. 2, 133.

[33] J. E. Grove, M. S. Strickman, W. N. Johnson, J. D. Kurfess, R. L. Kinzer, C. H. Starr, G. V. Jung, E. Kendziorra, P. Kretschmar, M. Maisack, and R. Staubert, *The Soft Gamma-Ray Spectrum of A0535+26: Detection of an Absorption Feature at 110 keV by OSSE*, The Astrophysical Journal Letters **438** (January 1995), L25.

[34] D. E. Gruber, W. A. Heindl, R. E. Rothschild, W. Coburn, R. Staubert, I. Kreykenbohm, and J. Wilms, *Stability of the Cyclotron Resonance Scattering Feature in Hercules X-1 with RXTE*, The Astrophysical Journal **562** (November 2001), no. 1, 499–507.

[35] Alice K. Harding and Joseph K. Daugherty, *Cyclotron Resonant Scattering and Absorption*, The Astrophysical Journal **374** (June 1991), 687.

[36] W. A. Heindl, W. Coburn, D. E. Gruber, R. E. Rothschild, I. Kreykenbohm, J. Wilms, and R. Staubert, *Discovery of a Cyclotron Resonance Scattering Feature in the X-Ray Spectrum of XTE J1946+274*, The Astrophysical Journal Letters **563** (December 2001), no. 1, L35–L39.

[37] William Heindl, Wayne Coburn, Ingo Kreykenbohm, and Joern Wilms, *Cyclotron Line in XTE J0658-073*, The Astronomer's Telegram **200** (October 2003), 1.

[38] Paul B. Hemphill, Richard E. Rothschild, Diana M. Cheatham, Felix Fürst, Peter Kretschmar, Matthias Kühnel, Katja Pottschmidt, Rüdiger Staubert, Jörn Wilms, and Michael T. Wolff, *The First NuSTAR Observation of 4U 1538-522: Updated Orbital Ephemeris and a Strengthened Case for an Evolving Cyclotron Line Energy*, The Astrophysical Journal **873** (2019), 62.

[39] R. Iaria, T. Di Salvo, M. Matranga, C. G. Galiano, A. D'Aí, A. Riggio, L. Burderi, A. Sanna, C. Ferrigno, M. Del Santo, F. Pintore, and N. R. Robba, *A possible cyclotron resonance scattering feature near 0.7 keV in X1822-371*, Astronomy & Astrophysics **577** (May 2015), A63.

[40] Gaurava K. Jaisawal and Sachindra Naik, *Detection of fundamental and first harmonic cyclotron line in X-ray pulsar Cep X-4*, Monthly Notices of the Royal Astronomical Society **453** (October 2015), no. 1, L21–L25.

[41] ———, *Detection of cyclotron resonance scattering feature in high-mass X-ray binary pulsar SMC X-2*, Monthly Notices of the Royal Astronomical Society **461** (September 2016), L97–L101.

[42] Gaurava K. Jaisawal, Sachindra Naik, and Prahlad Epili, *Suzaku view of the Be/X-ray binary pulsar GX 304-1 during Type I X-ray outbursts*, Monthly Notices of the Royal Astronomical Society **457** (April 2016), no. 3, 2749–2760.

[43] Gaurava K. Jaisawal, Sachindra Naik, and Biswajit Paul, *Possible Detection of a Cyclotron Resonance Scattering Feature in the X-Ray Pulsar 4U 1909+07*, The Astrophysical Journal **779** (December 2013), no. 1, 54.

[44] Peter Jenke, Colleen Wilson-Hodge, and Christian Malacaria, *Fermi/GBM detects a new outburst from the transient Be/X-ray binary 2S 1553-54*, The Astronomer's Telegram **14301** (January 2021), 1.

[45] R. L. Kelley, S. Rappaport, and S. Ayasli, *Discovery of 9.3s X-ray pulsations from 2S 1553-542 and a determination of the orbit.*, The Astrophysical Journal **274** (November 1983), 765–770.

[46] E. Kendziorra, P. Kretschmar, H. C. Pan, M. Kunz, M. Maisack, R. Staubert, W. Pietsch, J. Truemper, V. Efremov, and R. Sunyaev, *Evidence for cyclotron line features in high energy spectra of A 0535+26 during the March/April 1989 outburst.*, Astronomy & Astrophysics **291** (November 1994), L31–L34.

[47] ———, *Evidence for cyclotron line features in high energy spectra of A 0535+26 during the March/April 1989 outburst.*, Astronomy & Astrophysics **291** (November 1994), L31–L34.

[48] D. Klochkov, V. Doroshenko, A. Santangelo, R. Staubert, C. Ferrigno, P. Kretschmar, I. Caballero, J. Wilms, I. Kreykenbohm, K. Pottschmidt, R. E. Rothschild, C. A. Wilson-Hodge, and G. Pühlhofer, *Outburst of GX 304-1 monitored with INTEGRAL: positive correlation between the cyclotron line energy and flux*, Astronomy & Astrophysics **542** (2012), L28.

[49] U. Kraus, S. Blum, J. Schulte, H. Ruder, and P. Meszaros, *Analyzing X-Ray Pulsar Profiles: Geometry and Beam Pattern of Centaurus X-3*, The Astrophysical Journal **467** (August 1996), 794.

[50] U. Kraus, H. P. Nollert, H. Ruder, and H. Riffert, *Analyzing X-Ray Pulsar Profiles: Asymmetry as a Key to Geometry and Beam Pattern*, The Astrophysical Journal **450** (September 1995), 763.

[51] P. Kretschmar, H. C. Pan, E. Kendziorra, M. Kunz, M. Maisack, R. Staubert, W. Pietsch, J. Truemper, V. Efremov, and R. Sunyaev, *Absorption features in the hard X-ray spectra of PSR A 0535+26 and VELA X-1.*, Astronomy & Astrophysics **120** (November 1996), 175–178.

[52] I. Kreykenbohm, W. Coburn, J. Wilms, P. Kretschmar, R. Staubert, W. A. Heindl, and R. E. Rothschild, *Confirmation of two cyclotron lines in Vela X-1*, Astronomy & Astrophysics **395** (November 2002), 129–140.

[53] I. Kreykenbohm, J. Wilms, W. Coburn, M. Kuster, R. E. Rothschild, W. A. Heindl, P. Kretschmar, and R. Staubert, *The variable cyclotron line in GX 301-2*, Astronomy & Astrophysics **427** (December 2004), 975–986.

[54] V. A. Lepingwell, M. Fiocchi, J. Chenevez, A. Bazzano, A. J. Bird, V. Sguera, L. Natalucci, P. Ubertini, A. Bodaghee, and E. Kuulkers, *INTEGRAL detection of 2S 1553-542*, The Astronomer's Telegram **14335** (January 2021), 1.

[55] J. Li, S. Zhang, D. F. Torres, A. Papitto, Y. P. Chen, and J. M. Wang, *INTEGRAL and Swift/XRT observations of IGR J18179-1621*, Monthly Notices of the Royal Astronomical Society **426** (October 2012), no. 1, L16–L20.

[56] A. A. Lutovinov and S. S. Tsygankov, *Timing characteristics of the hard X-ray emission from bright X-ray pulsars based on INTEGRAL data*, Astronomy Letters **35** (July 2009), no. 7, 433–456.

[57] A. A. Lutovinov, S. S. Tsygankov, V. F. Suleimanov, A. A. Mushtukov, V. Doroshenko, D. I. Nagirner, and J. Poutanen, *Transient X-ray pulsar V 0332+53: pulse-phase-resolved spectroscopy and the reflection model*, Monthly Notices of the Royal Astronomical Society **448** (2015), 2175–2186.

[58] Alexander A. Lutovinov, Sergey S. Tsygankov, Konstantin A. Postnov, Roman A. Krivonos, Sergey V. Molkov, and John A. Tomsick, *NuSTAR observations of the supergiant X-ray pulsar IGR J18027-2016: accretion from the stellar wind and possible cyclotron absorption line*, Monthly Notices of the Royal Astronomical Society **466** (April 2017), no. 1, 593–599.

[59] K. Makishima and T. Mihara, *Frontiers O X-ray Astronomy*, Universal Academy Press Inc, Tokyo **23** (1992).

[60] K. Makishima, T. Mihara, M. Ishida, T. Ohashi, T. Sakao, M. Tashiro, T. Tsuru, T. Kii, F. Makino, T. Murakami, F. Nagase, Y. Tanaka, H. Kunieda, Y. Tawara, S. Kitamoto, S. Miyamoto, A. Yoshida, and M. J. L. Turner, *Discovery of a Prominent Cyclotron Absorption Feature from the Transient X-Ray Pulsar X0331+53*, The Astrophysical Journal Letters **365** (December 1990), L59.

[61] C. Malacaria, Y. Bhargava, J. B. Coley, L. Ducci, P. Pradhan, R. Ballhausen, and F. Fuerst, *Accreting on the Edge: A Luminosity-dependent Cyclotron Line in the Be/X-Ray Binary 2S 1553-542 Accompanied by Accretion Regimes Transition*, ApJ **927** (March 2022), 194.

[62] C. Malacaria, P. Kretschmar, K. K. Madsen, C. A. Wilson-Hodge, Joel B. Coley, P. Jenke, Alexander A. Lutovinov, K. Pottschmidt, Sergey S. Tsygankov, and J. Wilms, *The X-Ray Pulsar XTE J1858+034 Observed with NuSTAR and Fermi/GBM: Spectral and Timing Characterization plus a Cyclotron Line*, The Astrophysical Journal **909** (March 2021), no. 2, 153.

[63] Manoj Mandal and Sabyasachi Pal, *Detection of Low-Frequency QPO From X-ray Pulsar XTE J1858+034 During Outburst in 2019 with NuSTAR*, arXiv e-prints (January 2021), arXiv:2101.09250.

[64] ———, *Multi-Wavelength Study of X-ray Pulsar 2S 1553-542 During Outburst in 2021*, arXiv e-prints (February 2021), arXiv:2103.00603.

[65] ———, *Study of Timing and Spectral Properties of the X-ray Pulsar 1A 0535+262 During the Giant Outburst in 2020 November-December*, Monthly Notices of the Royal Astronomical Society **511** (January 2022), 1121–1130.

[66] Manoj Mandal, Sabyasachi Pal, and Mangal Hazra, *Swift/MAXI follow-up of recent outburst from X-ray pulsar 2S 1553-542*, The Astronomer's Telegram **14308** (January 2021), 1.

[67] Manoj Mandal, Sabyasachi Pal, Mangal Hazra, Arindam Jana, Bikram Bhunia, and Arnab Ghanta, *Swift/BAT detection of flaring activity from X-ray binary pulsar A 0535+262*, The Astronomer's Telegram **14157** (November 2020), 1.

[68] A. Manousakis, R. Walter, M. Audard, and T. Lanz, *Pulsed thermal emission from the accreting pulsar XMMU J054134.7-682550*, Astronomy & Astrophysics **498** (April 2009), no. 1, 217–222.

[69] C. B. Markwardt, J. Halpern, and J. H. Swank, *XTE J1829-098 Predicted for Another Outburst in Early April*, The Astronomer's Telegram **2007** (April 2009), 1.

[70] N. Masetti, D. Dal Fiume, L. Amati, S. Del Sordo, F. Frontera, M. Orlandini, and E. Palazzi, *A look with BeppoSAX at the low-luminosity Galactic X-ray source 4U 2206+54*, Astronomy & Astrophysics **423** (August 2004), 311–319.

[71] V. A. McBride, J. Wilms, M. J. Coe, I. Kreykenbohm, R. E. Rothschild, W. Coburn, J. L. Galache, P. Kretschmar, W. R. T. Edge, and R. Staubert, *Study of the cyclotron feature in MXB 0656-072*, Astronomy & Astrophysics **451** (May 2006), no. 1, 267–272.

[72] Peter Meszaros, *High-energy radiation from magnetized neutron stars*, 1992.

[73] T. Mihara, K. Makishima, S. Kamijo, T. Ohashi, F. Nagase, Y. Tanaka, and K. Koyama, *Discovery of a Cyclotron Resonance Feature at 30 keV from the Transient X-Ray Pulsar Cepheus X-4*, The Astrophysical Journal Letters **379** (October 1991), L61.

[74] T. Mihara, K. Makishima, T. Ohashi, T. Sakao, and M. Tashiro, *New observations of the cyclotron absorption feature in Hercules X-1*, Nature **346** (July 1990), no. 6281, 250–252.

[75] S. Molkov, A. Lutovinov, S. Tsygankov, I. Mereminskiy, and A. Mushtukov, *Discovery of a Pulse-phase-transient Cyclotron Line in the X-Ray pulsar GRO J2058+42*, The Astrophysical Journal Letters **883** (September 2019), no. 1, L11.

[76] D. Müller, D. Klochkov, I. Caballero, and A. Santangelo, *A 0535+26 in the April 2010 outburst: probing the accretion regime at work*, Astronomy & Astrophysics **552** (2013), A81.

[77] Alexander A. Mushtukov, Sergey S. Tsygankov, Alexander V. Serber, Valery F. Suleimanov, and Juri Poutanen, *Positive correlation between the cyclotron line energy and luminosity in sub-critical X-ray pulsars: Doppler effect in the accretion channel*, Monthly Notices of the Royal Astronomical Society **454** (2015), 2714–2721.

[78] M. Nakajima, T. Mihara, K. Makishima, and H. Niko, *A Further Study of the Luminosity-dependent Cyclotron Resonance Energies of the Binary X-Ray Pulsar 4U 0115+63 with the Rossi X-Ray Timing Explorer*, The Astrophysical Journal **646** (August 2006), no. 2, 1125–1138.

[79] M. Orlandini, D. Dal Fiume, F. Frontera, S. Del Sordo, S. Piraino, A. Santangelo, A. Segreto, T. Oosterbroek, and A. N. Parmar, *BEPPOSAX Observation of 4U 1626-67: Discovery of an Absorption Cyclotron Resonance Feature*, The Astrophysical Journal Letters **500** (June 1998), no. 2, L163–L166.

[80] Mauro Orlandini, Filippo Frontera, Nicola Masetti, Vito Sguera, and Lara Sidoli, *BeppoSAX Observations of the X-Ray Pulsar MAXI J1409-619 in Low State: Discovery of Cyclotron Resonance Features*, The Astrophysical Journal **748** (April 2012), no. 2, 86.

[81] Mayukh Pahari and Sabyasachi Pal, *RXTE observation of recent flaring activity from the transient X-ray pulsar 2S 1553-542*, Monthly Notices of the Royal Astronomical Society **423** (July 2012), no. 4, 3352–3359.

[82] Sabyasachi Pal and Manoj Mandal, *X-ray pulsar A 0535+262 reached 3 Crab: update with Neil Gehrels Swift Observatory*, The Astronomer's Telegram **14170** (November 2020), 1.

[83] A. N. Parmar, N. E. White, and L. Stella, *The Transient 42 Second X-Ray Pulsar EXO 2030+375. II. The Luminosity Dependence of the Pulse Profile*, The Astrophysical Journal **338** (March 1989), 373.

[84] K. Postnov, R. Staubert, A. Santangelo, D. Klochkov, P. Kretschmar, and I. Caballero, *The appearance of magnetospheric instability in flaring activity at the onset of X-ray outbursts in A0535+26*, Astronomy & Astrophysics **480** (March 2008), no. 2, L21–L24.

[85] Juri Poutanen, Alexander A. Mushtukov, Valery F. Suleimanov, Sergey S. Tsygankov, Dmitrij I. Nagirner, Victor Doroshenko, and Alexander A. Lutovinov, *A Reflection Model for the Cyclotron Lines in the Spectra of X-Ray Pulsars*, The Astrophysical Journal **777** (2013), 115.

[86] P. Reig and M. J. Coe, *X-ray spectral properties of the pulsar EXO 2030+375 during an outburst*, Monthly Notices of the Royal Astronomical Society **302** (February 1999), no. 4, 700–706.

[87] P. Reig and E. Nespoli, *Patterns of variability in Be/X-ray pulsars during giant outbursts*, Astronomy & Astrophysics **551** (March 2013), A1.

[88] R. Remillard, A. Levine, T. Takeshima, R. H. D. Corbet, F. E. Marshall, J. H. Swank, and D. Chakrabarty, *XTE J1858+034*, IAUC **6826** (February 1998), 2.

[89] M. G. Revnivtsev, R. A. Sunyaev, D. A. Varshalovich, V. V. Zheleznyakov, A. M. Cherepashchuk, A. A. Lutovinov, E. M. Churazov, S. A. Grebenev, and M. R. Gilfanov, *A Hard X-ray Survey of the Galactic-Center Region with the IBIS Telescope of the INTEGRAL Observatory: A Catalog of Sources*, Astronomy Letters **30** (June 2004), 382–389.

[90] Elizabeth Rivers, Alex Markowitz, Katja Pottschmidt, Stefanie Roth, Laura Barragán, Felix Fürst, Slawomir Suchy, Ingo Kreykenbohm, Jörn Wilms, and Richard Rothschild, *A Comprehensive Spectral Analysis of the X-Ray Pulsar 4U 1907+09 from Two Observations with the Suzaku X-ray Observatory*, The Astrophysical Journal **709** (January 2010), no. 1, 179–190.

[91] J. Rodriguez, M. Tuerler, S. Chaty, and J. A. Tomsick, *Swift archival observations of the field around the new INTEGRAL source IGR J19294+1816*, The Astronomer's Telegram **1998** (March 2009), 1.

[92] F. D. Rosenberg, C. J. Eyles, G. K. Skinner, and A. P. Willmore, *Observations of a transient X-ray source with a period of 104 S*, Nature **256** (August 1975), 628–630.

[93] Jayashree Roy, Manojendu Choudhury, and P. C. Agrawal, *Timing and spectral study of igr j19294+1816 with the rxte: The discovery of cyclotron features*, The Astrophysical Journal **848** (2017Oct), no. 2, 124.

[94] Makoto Sasano, Kazuo Makishima, Soki Sakurai, Zhongli Zhang, and Teruaki Enoto, *Suzaku view of the neutron star in the dipping source 4U 1822-37*, Publications of the Astronomical Society of Japan **66** (April 2014), no. 2, 35.

[95] G. Schönherr, J. Wilms, P. Kretschmar, I. Kreykenbohm, A. Santangelo, R. E. Rothschild, W. Coburn, and R. Staubert, *A model for cyclotron resonance scattering features*, Astronomy & Astrophysics **472** (2007), 353–365.

[96] C. R. Shrader, F. K. Sutaria, K. P. Singh, and D. J. Macomb, *High-Energy Spectral and Temporal Characteristics of GRO J1008–57*, The Astrophysical Journal **512** (February 1999), no. 2, 920–928.

[97] A. E. Shtykovsky, A. A. Lutovinov, S. S. Tsygankov, and S. V. Molkov, *Discovery of a cyclotron absorption line in the transient X-ray pulsar XTE J1829-098*, Monthly Notices of the Royal Astronomical Society **482** (January 2019), no. 1, L14–L18.

[98] R. Staubert, D. Klochkov, F. Fürst, J. Wilms, R. E. Rothschild, and F. Harrison, *Inversion of the decay of the cyclotron line energy in Hercules X-1*, Astronomy & Astrophysics **606** (2017), L13.

[99] R. Staubert, D. Klochkov, J. Wilms, K. Postnov, N. I. Shakura, R. E. Rothschild, F. Fürst, and F. A. Harrison, *Long-term change in the cyclotron line energy in Hercules X-1*, Astronomy & Astrophysics **572** (December 2014), A119.

[100] R. Staubert, N. I. Shakura, K. Postnov, J. Wilms, R. E. Rothschild, W. Coburn, L. Rodina, and D. Klochkov, *Discovery of a flux-related change of the cyclotron line energy in Hercules X-1*, Astronomy & Astrophysics **465** (2007), no. 2, L25–L28.

[101] R. Staubert, J. Trümper, E. Kendziorra, D. Klochkov, K. Postnov, P. Kretschmar, K. Pottschmidt, F. Haberl, R. E. Rothschild, A. Santangelo, J. Wilms, I. Kreykenbohm, and F. Fürst, *Cyclotron lines in highly magnetized neutron stars*, Astronomy & Astrophysics **622** (February 2019), A61.

[102] Slawomir Suchy, Felix Fürst, Katja Pottschmidt, Isabel Caballero, Ingo Kreykenbohm, Jörn Wilms, Alex Markowitz, and Richard E. Rothschild, *Broadband Spectroscopy Using Two Suzaku Observations of the HMXB GX 301-2*, The Astrophysical Journal **745** (February 2012), no. 2, 124.

[103] Mutsumi Sugizaki, Kazuhisa Mitsuda, Hidehiro Kaneda, Keiichi Matsuzaki, Shigeo Yamauchi, and Katsuji Koyama, *Faint X-Ray Sources Resolved in the ASCA Galactic Plane Survey and Their Contribution to the Galactic Ridge X-Ray Emission*, The Astrophysical JournalS **134** (May 2001), no. 1, 77–102.

[104] T. Takeshima, R. H. D. Corbet, F. E. Marshall, J. Swank, and D. Chakrabarty, *XTE J1858+034.*, IAUC **6826** (February 1998), 1.

[105] Shriharsh P. Tendulkar, Felix Fürst, Katja Pottschmidt, Matteo Bachetti, Varun B. Bhalerao, Steven E. Boggs, Finn E. Christensen, William W. Craig, Charles A. Hailey, Fiona A. Harrison, Daniel Stern, John A. Tomsick, Dominic J. Walton, and William Zhang, *NuSTAR Discovery of a Cyclotron Line in the Be/X-Ray Binary RX J0520.5-6932 during Outburst*, The Astrophysical Journal **795** (November 2014), no. 2, 154.

[106] Thomas W. J. Thompson, John A. Tomsick, Richard E. Rothschild, J. J. M. in't Zand, and Roland Walter, *Orbital Parameters for the X-Ray Pulsar IGR J16393-4643*, The Astrophysical Journal **649** (September 2006), no. 1, 373–381.

[107] J. M. Torrejón, I. Kreykenbohm, A. Orr, L. Titarchuk, and I. Negueruela, *Evidence for a Neutron Star in the non-pulsating massive X-ray binary 4U2206+54*, Astronomy & Astrophysics **423** (August 2004), 301–309.

[108] J. Truemper, W. Pietsch, C. Reppin, W. Voges, R. Staubert, and E. Kendziorra, *Evidence for strong cyclotron line emission in the hard X-ray spectrum of Hercules X-1.*, The Astrophysical Journal Letters **219** (1978), L105–L110.

[109] S. S. Tsygankov, R. A. Krivonos, and A. A. Lutovinov, *Broad-band observations of the Be/X-ray binary pulsar RX J0440.9+4431: discovery of a cyclotron absorption line*, Monthly Notices of the Royal Astronomical Society **421** (April 2012), no. 3, 2407–2413.

[110] S. S. Tsygankov, A. A. Lutovinov, E. M. Churazov, and R. A. Sunyaev, *V0332+53 in the outburst of 2004-2005: luminosity dependence of the cyclotron line and pulse profile*, Monthly Notices of the Royal Astronomical Society **371** (September 2006), no. 1, 19–28.

[111] S. S. Tsygankov, A. A. Lutovinov, and A. V. Serber, *Completing the puzzle of the 2004-2005 outburst in V0332+53: the brightening phase included*, Monthly Notices of the Royal Astronomical Society **401** (January 2010), no. 3, 1628–1635.

[112] Sergey S. Tsygankov, Victor Doroshenko, Alexander A. Mushtukov, Alexander A. Lutovinov, and Juri Poutanen, *Study of the X-ray pulsar IGR J19294+1816 with NuSTAR: Detection of cyclotron line and transition to accretion from the cold disk*, Astronomy & Astrophysics **621** (January 2019), A134.

[113] Sergey S. Tsygankov, Alexander A. Lutovinov, Roman A. Krivonos, Sergey V. Molkov, Peter J. Jenke, Mark H. Finger, and Juri Poutanen, *NuSTAR discovery of a cyclotron absorption line in the transient X-ray pulsar 2S 1553-542*, Monthly Notices of the Royal Astronomical Society **457** (March 2016), 258–266.

[114] Sergey S. Tsygankov, Alexander A. Lutovinov, Sergey V. Molkov, AnlAug A. Djupvik, Dmitri I. Karasev, Victor Doroshenko, Alexander A. Mushtukov, Christian Malacaria, Peter Kretschmar, and Juri Poutanen, *X-Ray Pulsar XTE J1858+034: Discovery of the Cyclotron Line and the Revised Optical Identification*, The Astrophysical Journal **909** (March 2021), no. 2, 154.

[115] M. Turler, J. Rodriguez, and C. Ferrigno, *INTEGRAL discovers the new hard X-ray source IGR J19294+1816*, The Astronomer's Telegram **1997** (March 2009), 1.

[116] D. J. Walton, M. Bachetti, F. Fürst, D. Barret, M. Brightman, A. C. Fabian, B. W. Grefenstette, F. A. Harrison, M. Heida, J. Kennea, P. Kosec, R. M. Lau, K. K. Madsen, M. J. Middleton, C. Pinto, J. F. Steiner, and N. Webb, *A Potential Cyclotron Resonant Scattering Feature in the Ultraluminous X-Ray Source Pulsar NGC 300 ULX1 Seen by NuSTAR and XMM-Newton*, The Astrophysical Journal **857** (April 2018), no. 1, L3.

[117] W. A. Wheaton, J. P. Doty, F. A. Primini, B. A. Cooke, C. A. Dobson, A. Goldman, M. Hecht, S. K. Howe, J. A. Hoffman, A. Scheepmaker, E. Y. Tsiang, W. H. G. Lewin, J. L. Matteson, D. E. Gruber, W. A. Baity, R. Rotschild, F. K. Knight, P. Nolang, and L. E. Peterson, *An absorption feature in the spectrum of the pulsed hard X-ray flux from 4U0115+63*, Nature **282** (November 1979), no. 5736, 240–243.

[118] C. A. Wilson, S. N. Zhang, M. H. Finger, R. B. Wilson, M. Scott, T. Koh, D. Chakrabarty, B. VAughan, and T. A. Prince, *GRO J2058+42*, IAUC **6238** (September 1995), 1.

[119] Takayuki Yamamoto, Tatehiro Mihara, Mutsumi Sugizaki, Motoki Nakajima, Kazuo Makishima, and Makoto Sasano, *Firm detection of a cyclotron resonance feature with Suzaku in the X-ray spectrum of GRO J1008-57 during a giant outburst in 2012*, Publications of the Astronomical Society of Japan **66** (June 2014), no. 3, 59.

[120] Takayuki Yamamoto, Mutsumi Sugizaki, Tatehiro Mihara, Motoki Nakajima, Kazutaka Yamaoka, Masaru Matsuoka, Mikio Morii, and Kazuo Makishima, *Discovery of a Cyclotron Resonance Feature in the X-Ray Spectrum of GX 304-1 with RXTE and Suzaku during Outbursts Detected by MAXI in 2010*, Publications of the Astronomical Society of Japan **63** (2011nov), no. sp3, S751–S757.

[121] Jun Yokogawa, Ken'ichi Torii, Takayoshi Kohmura, and Katsuji Koyama, *ASCA Identification of SMC X-2 with the 2.37-s Pulsar Discovered by RXTE*, Publications of the Astronomical Society of Japan **53** (April 2001), no. 2, 227–231.

Millimeter Wavelength Studies of Complex Nitrile Species in the Atmosphere of Saturn's Moon Titan

Arijit Manna[1,*] and Sabyasachi Pal[1]

[1]Midnapore City College, Kuturia, Bhadutala, Paschim Medinipur, West Bengal, 721129, India
[*]Corresponding author: arijitmanna@mcconline,org.in

Abstract

In our solar system, Saturn is the second-largest gas planet after Jupiter. The ring gas planet Saturn has the highest number of (a total of eighty-two) moons, and Titan is the largest moon of Saturn. In this review, we described the composition of the complex nitrile-bearing molecules in the atmosphere of Titan. The atmosphere of Titan is mainly composed of a large amount of nitrogen-rich clouds and it also consists of minor amounts of methane (CH_4) and ethane (C_2H_6). Titan is the only moon in the solar system after the Earth that has a dense atmosphere with clear evidance of liquid flowing over the surface. In this review, we describe the detection of complex nitrile-bearing molecules such as ethyl cyanide (C_2H_5CN), vinyl cyanide (C_2H_3CN), hydrogen cyanide (HCN), and methyl cyanide (CH_3CN) at millimeter-wavelength using the Atacama Large Millimeter/Submillimeter Array (ALMA). We discuss the possible formation mechanism of the detected nitrile-bearing molecules in the atmosphere of Titan. Additionally, we also discuss the possible formation pathways of the production of the simplest amino acid glycine (NH_2CH_2COOH) on the surface of Titan using the Strecker synthesis reaction.

Keywords: *Solar Planets, Planetary atmosphere, Millimeter wavelength, Astrochemistry; Astrobiology.*

1 Introduction

Titan is the second-largest moon in our solar system and it is the largest moon of the gaseous ring planet Saturn. Titan is the only natural satellite which consists of a deep atmosphere. The atmosphere of Titan is about $\sim$1.6 times denser than the atmosphere of Earth. Except for Earth, Titan is the only moon in our solar system where clear evidence exists for the flow of surface liquid. Titan was discovered by the Dutch astronomer Christiaan Huygens in 1665. Titan is surrounded by a dense layer of organic gases and it is the only satellite in our solar system that consists of a thick atmosphere (1.45 bar). In the solar system, humans can survive without pressure suits only on Earth and Titan because these two planetary objects naturally consist of solid surfaces [1]. A large amount of N_2 (94.2%) and a small amount of CH_4 (5.65%) and H_2 (0.099%) are found in the lower atmosphere of Titan [2]. The atmosphere of Titan consists of various types of hydrocarbon related complex organic molecules like C_2H_6, C_4H_2, C_2H_2, C_3H_8 and Polycyclic Aromatic Hydrocarbons (PAHs) [3]. The atmosphere of Titan also consists of HC_3N, HCN, CO_2, CO, CH_3CN, Ar, and He [4]. The presence of N_2 and CH_4 in the thick atmosphere of Titan creates many complex organic molecules using photochemical pathways, and it also participates in creating hazes that look like the typical brown or orange colour.

The study of the atmosphere of Titan is one of the interesting topics to investigate the origin and properties of the complex organic molecules in our solar system [5] [6]. The effective temperature (T_{eff}) of Titan is $\sim$82 K which implies it is too cold than Earth. The surface pressure of Titan is $\sim$1.5 bar, which implies the surface condition is near the triple point of CH_4. As a result, the liquid CH_4 flows on the surface and gaseous CH_4 survives in the atmosphere of Titan.

The Cassini 10th mission in 2006 detected CH_4 from the atmosphere of Titan at 3.3 μm band using the Visible and Infrared Mapping Spectrometer (VIMS) [7]. The VIMS can observe the planetary nadir, limb, and occultation. The observed chemical mechanism of CH_4 at 3.3 μm in the atmosphere of Titan is very similar to the observed chemical mechanism in the Interstellar Medium (ISM) [8] [9]. The feature of CH_4 was also observed by VMS during the stellar occultation in the atmosphere of Saturn [10] [11]. Scientists are also analyzing the observed optical depth of the detected CH_4 at 3.3 μm spectra in the atmosphere of Saturn which shows the broad feature between

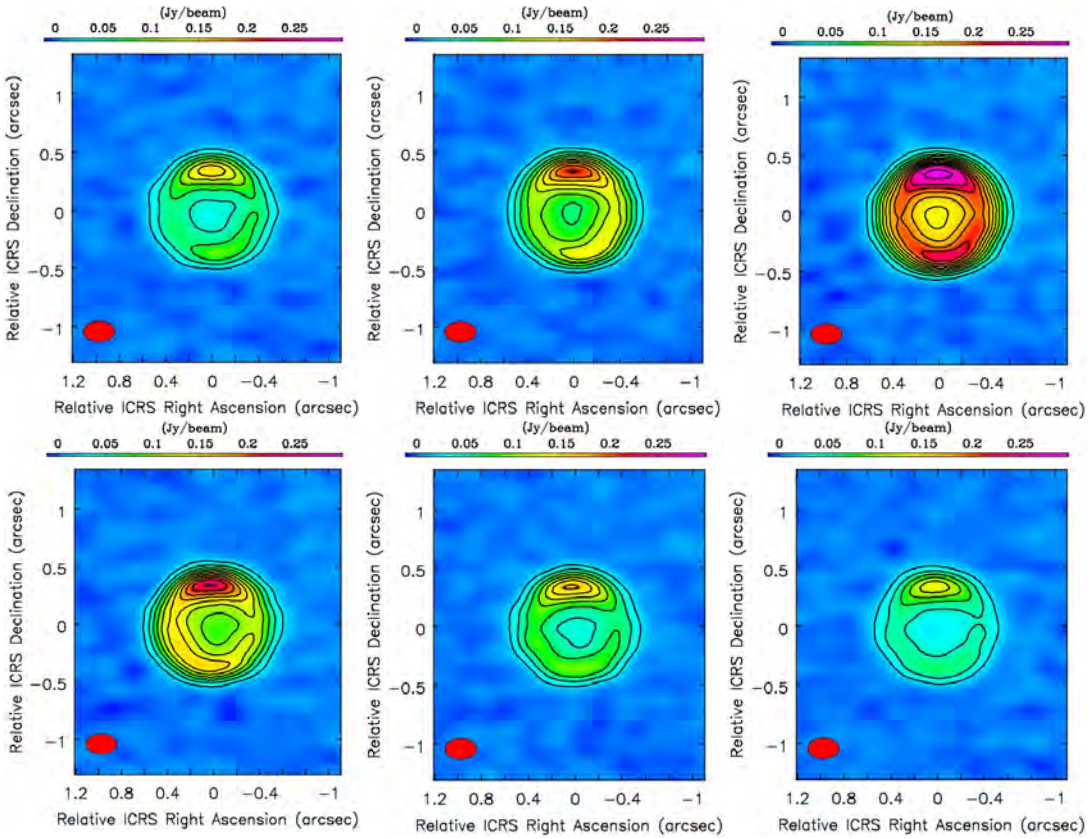


Figure 1: Channel maps of C_2H_5CN in the atmosphere of Titan. From these channel maps, it was evident that the C_2H_5CN was coming from the stratosphere and upper atmosphere of Titan. The red elliptical circles indicated the synthesized beams of the channel maps.

the pressure levels 0.0150 and 0.0018 mbar [11]. The aliphatic C–H stretching bands of solid-state hydrocarbons, such as C_6H_{12}, C_5H_{12}, C_7H_{14}, and C_6H_{14} were ascribed to these 3.4 μm spectral characteristics [11]. Planetary scientists also found tentative evidence of the narrow absorption line at 3.28 μm which indicated the existence of PAHs in the stratosphere of Titan [12]. Earlier, scientists also claimed the presence of aromatic hydrocarbons in the upper atmosphere of Titan ($\sim$650 to 1300 km). The strong absorption lines at 3.2 and 3.5 μm are connected to the C$-$H bonds. These C-H bonds can be observed in simple molecules, icy hydrocarbons, and complex PAHs. The situ measurement indicated the presence of a trace of hydrocarbons and nitrogen-bearing molecules which were created in the atmosphere of Titan via gas-phase chemical reactions and cosmic ray ionizations [13].

After a few successful detections of the molecular lines using various space missions (mostly Cassini), ground-based radio telescopes are used regularly to detect various complex organic molecules from the atmosphere of Titan. In this review, we mainly discussed the detection of complex nitrogen-bearing molecules from the atmosphere of Titan using the Atacama Large Millimeter/Submillimeter Array (ALMA).

2 Complex Nitrogen-Bearing Molecules in the Atmosphere of Titan

2.1 Ethyl Cyanide (C_2H_5CN)

The complex nitrile molecule ethyl cyanide (C_2H_5CN) is also known as propionitrile. In the ISM, the emission lines of C_2H_5CN are mainly found in the high mass star formation regions, hot molecular cores, and solar-like protostars, and they were created via the reaction between CH_3CN and CH_3. Recently, the rotational emission lines of C_2H_5CN were detected in the atmosphere of Titan using the ALMA band 6 between the frequency range of 221–241 GHz [13]. The vertical column density of C_2H_5CN was $(1–5)\times10^{14}\,cm^{-2}$ in the atmosphere of

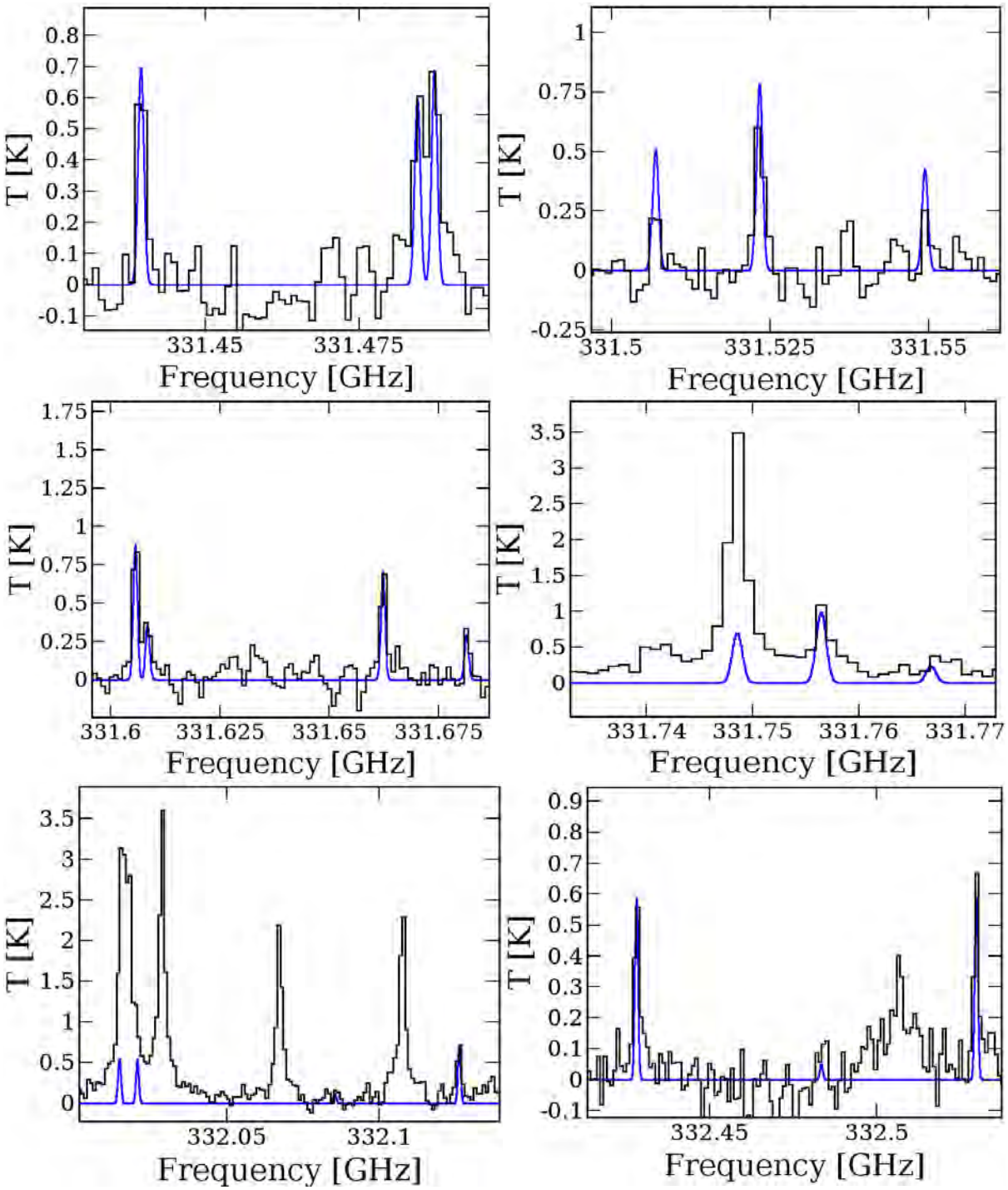


Figure 2: Rotational emission spectrums of C_2H_5CN with different transitions were detected in the frequency range of 330.75–332.75 GHz. In the emission spectrums, the black line indicated the millimeter wavelength spectra of Titan, and the blue line indicated the best fit Local Thermodynamic Equilibrium (LTE) model over the original spectrum. After fitting the LTE model, we estimate the column density of C_2H_5CN was 7.0×10^{14} cm^{-2}.

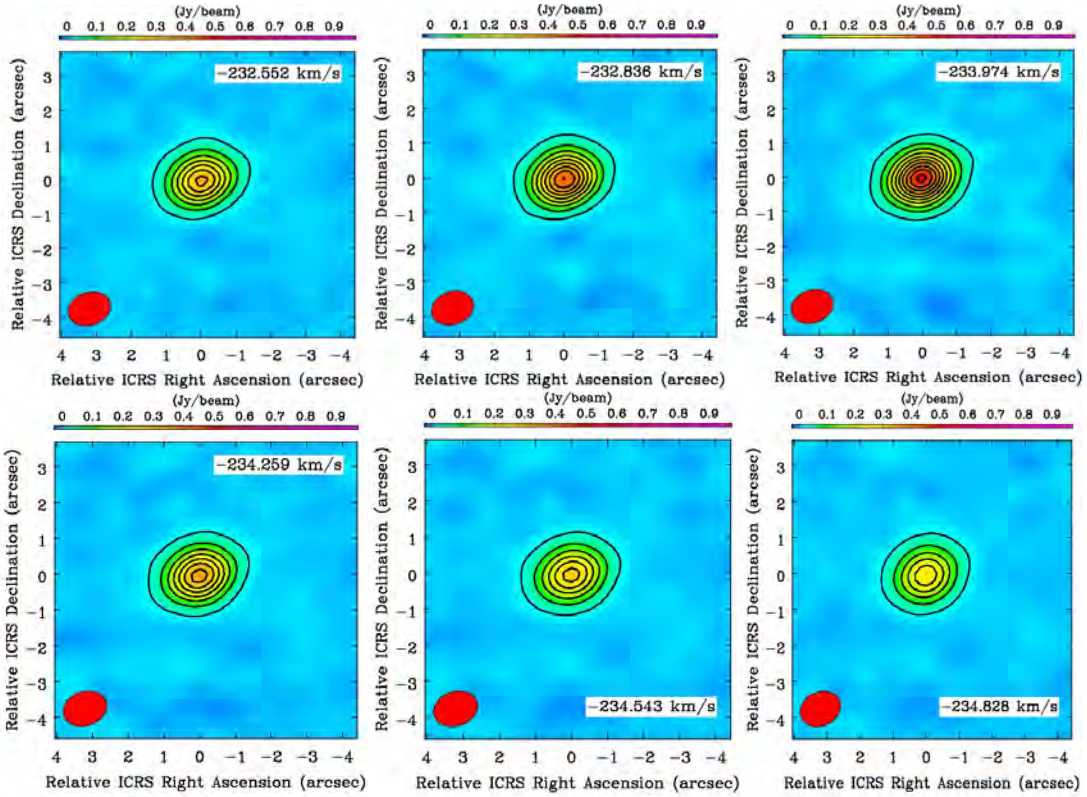


Figure 3: Channel maps of C_2H_3CN in the atmosphere of Titan. From these channel maps, it was evident that the C_2H_3CN was coming from the stratosphere of Titan. The red elliptical circles indicated the synthesized beams of the channel maps.

Titan, which was estimated by Non-linear optimal Estimator for MultivariatE spectral analySIS (NEMESIS)[1] atmospheric radiative transfer model. The radiative transfer model indicated that the C_2H_5CN is presented at altitudes $\gtrsim$ 200 km, which indicates that the C_2H_5CN molecule is produced predominantly in the stratosphere and above. The photochemical reaction network indicated that the C_2H_5CN is one of the most abundant nitrile-bearing molecules in the atmosphere of Titan [13]. Additionally, we detected new transition lines of C_2H_5CN in the atmosphere of Titan between the frequency range of 330.75–332.75 GHz using the ALMA band 7. After the detection of the new transition lines of C_2H_5CN, we used the Local Thermodynamic Equilibrium (LTE) model to estimate the column density of C_2H_5CN. The resultant channel maps and rotational emission spectrums of C_2H_5CN were shown in Figure 1 and 2.

2.2 Vinyl Cyanide (C_2H_3CN)

The vinyl cyanide (C_2H_3CN) is one of the complex organic molecules which is also known as Acrylonitrile. It is an important molecule in the atmosphere of Titan for the formation of the cell-like membranes that are known as 'azotosomes' [14]. Earlier, many studies suggested that the complex molecule C_2H_3CN may exist in the atmosphere of Titan, but many observations failed to detect this molecule [15]. A laboratory experiment suggested the atmosphere of Titan can produce the C_2H_3CN [16]. Using the Infrared Space Observatory (ISO), scientists could not find any evidence of the presence of C_2H_3CN in the atmosphere of Titan [17]. The upper limit column density of C_2H_3CN in the order of 2×10^{-9} in the stratosphere of Titan which was estimated using the Institut de Radio Astronomie Millimétrique (IRAM) 30-m single-dish radio telescope [18] [19]. After many failures finally, the rotational emission lines of C_2H_3CN were detected from the atmosphere of Titan using the ALMA band 6 [15]. Using the NEMESIS atmospheric radiative transfer model, scientists estimated the vertical column density of C_2H_3CN which is found in the range of 3.7×10^{13} to 1.4×10^{14} cm^{-2} [15]. The photochemical model in the atmosphere of Titan indicated that C_2H_3CN is formed via the reaction with C_2H_4 and CN in the upper atmosphere

[1] https://nemesiscode.github.io/pdf/Nemesis_Revised_Accepted.pdf

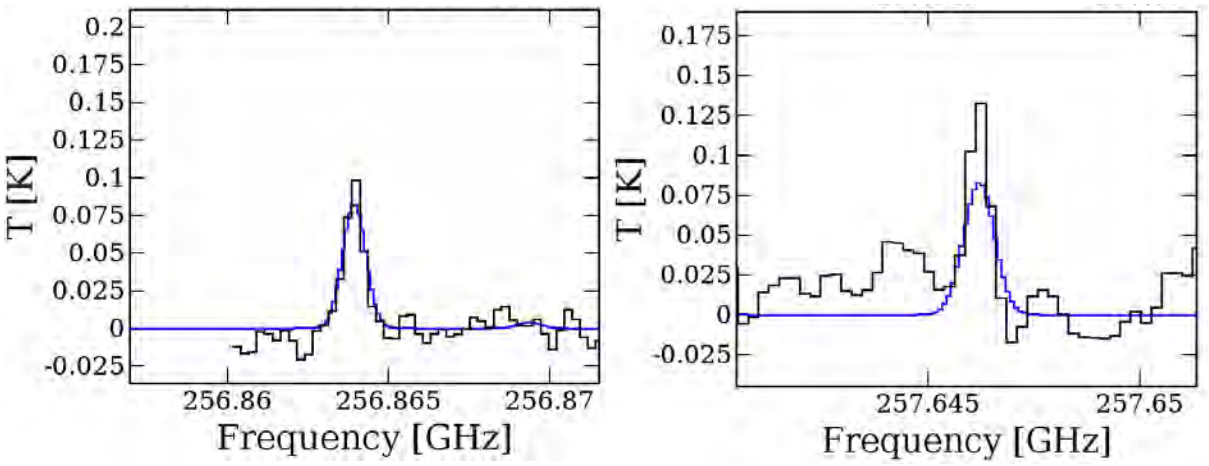


Figure 4: Rotational emission spectrums of C_2H_3CN between the frequency range 256.86–257.80 GHz. In the emission spectrums, the black line indicated the millimeter wavelength spectra of Titan, and the blue line indicated the best fit Local Thermodynamic Equilibrium (LTE) model over the observed spectrum. After fitting the LTE model, we estimate the column density of C_2H_3CN was 1.2×10^{14} cm^{-2}.

of Titan [15]. In the lower stratosphere, C_2H_3CN will be produced via the reaction of HCN and C_2H_3 [15]. Additionally, we also detected another two rotational transition lines of C_2H_3CN between the frequency ranges of 256.86–257.80 GHz using the ALMA band 6. The channel maps and rotational emission spectrums of C_2H_3CN between the frequency range 256.86–257.80 GHz is shown in Figure 3 and 4.

2.3 Hydrogen Cyanide (HCN)

In the planetary atmosphere, hydrogen cyanide (HCN) is one of the important molecule for the formation of other prebiotic molecules [20]. The HCN molecule has an important role in creating the amino acids in the planetary atmosphere, star-formation regions, hot molecular cores, and solar-like protostars via the Strecker synthesis reactions. In the planetary atmosphere, the organic molecule HCN is created via the dissociation of CH_4 and N_2 [21]. In our planetary system, Saturn's moon Titan has the most HCN-rich atmosphere. Earlier, the Cassini spacecraft measured the mixing ratio of HCN in the atmosphere of Titan in the order of 0.1–10 ppm in the lower atmosphere ($\leq$ 600 km) and 0.1-5% in the upper atmosphere ($\geq$ 700 km) [22] [23] [24] [25]. UV light is primarily responsible for the formation of radical species in Titan's upper and lower atmospheres by dissociating CH_4 and N_2 [26] [27]. Earlier, the HCN molecule was also detected in the atmosphere of Neptune and Pluto with an abundance of 1 ppb and 40 ppm [28] [29]. Many simulations show the HCN molecule produced in the atmosphere of Titan via the reaction of H_2CN and H and HCN are destroyed via the photolysis reaction, (HCN+hν $\longrightarrow$CN+H) [26] [30] [31] [32]. Recently, the rotational emission lines of HCN were detected in the atmosphere of Titan at frequencies of 354.505 and 354.460 GHz using the ALMA band 7 observation [33]. The channel maps and rotational emission spectrum of HCN is shown in Figure 5 and 6 [33]. The channel map shows that the emission line of HCN mainly coming from the eastern and western limbs of Titan [33]. The detection of HCN has indicated the atmosphere of Titan has the ability to create the simplest amino glycine (NH_2CH_2COOH) via hydrolysis of aminoacetonitrile (NH_2CH_2COOH) in the gas phase using the Strecker synthesis reactions. The atmosphere of Titan has a low amount of H_2O in the order of 0.5–8 ppb [34] [35]. Recently, scientists have discovered the emission lines of HCN in the stratosphere of Saturn at a frequency of 354.505 GHz using ALMA band 7 observation [20]. Scientists have claimed that the detection of HCN in the atmosphere of Saturn opens a new window to the synthesis of other complex bio-molecules via HCN [20].

2.4 Acetonitrile/Methyl Cyanide (CH_3CN)

Methyl cyanide (CH_3CN) is one of the complex organic molecule which is observed in particular hot molecular cores [36]. Recently, ALMA detected many rotational transition lines of CH_3CN from the atmosphere of Titan using the band 7 observations [33] [37]. In the atmosphere of Titan, HCN and CH_3 are responsible for the production of CH_3CN. The channel maps of CH_3CN with transition J = 19–18 and emission spectrums of CH_3CN

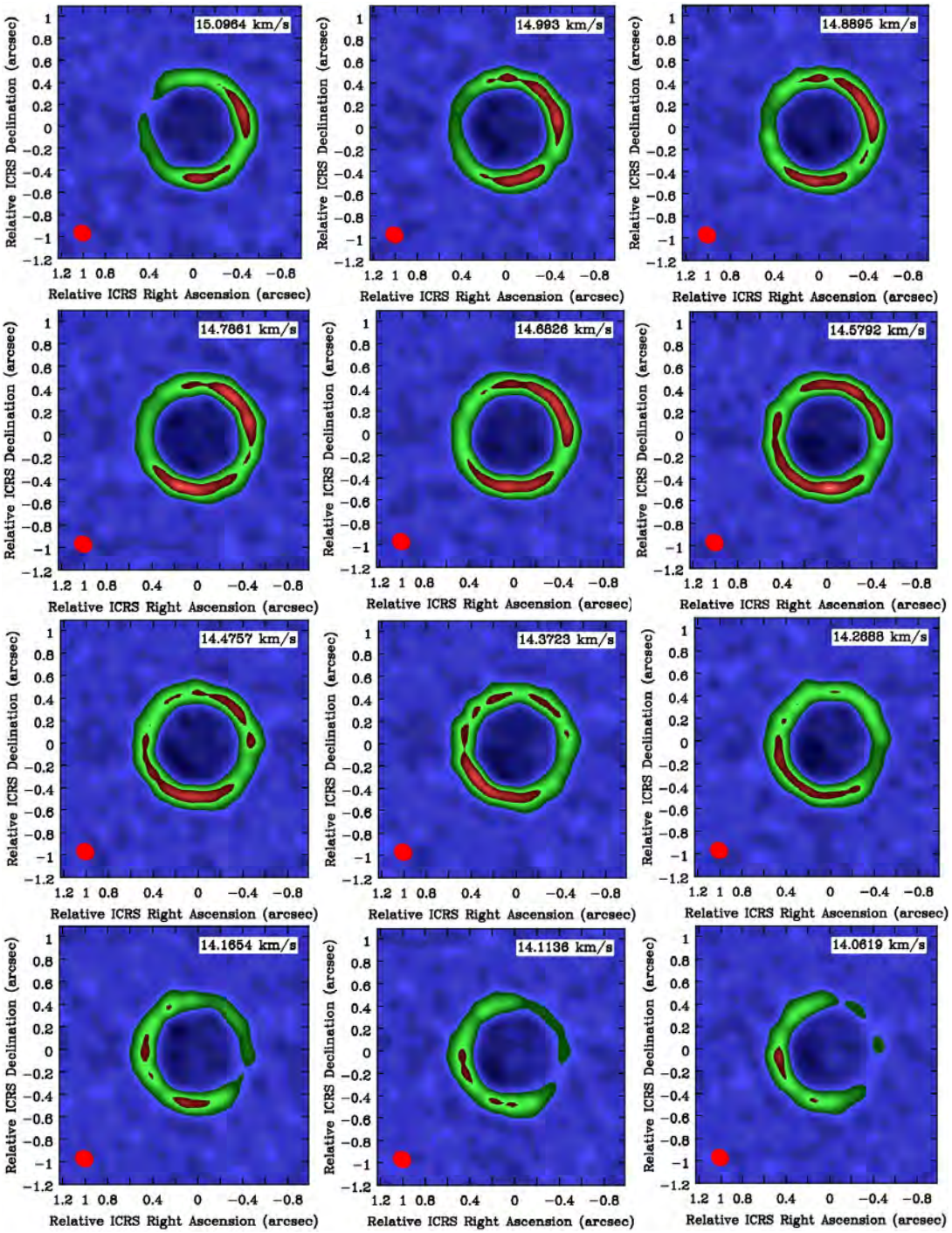


Figure 5: Channel maps of HCN in the atmosphere of Titan at frequency 354.505 GHz. From these channel maps, it was clearly evident that the HCN was coming from the eastern and western limbs of Titan [33]. The red elliptical circles indicated the synthesized beams of the channel maps.

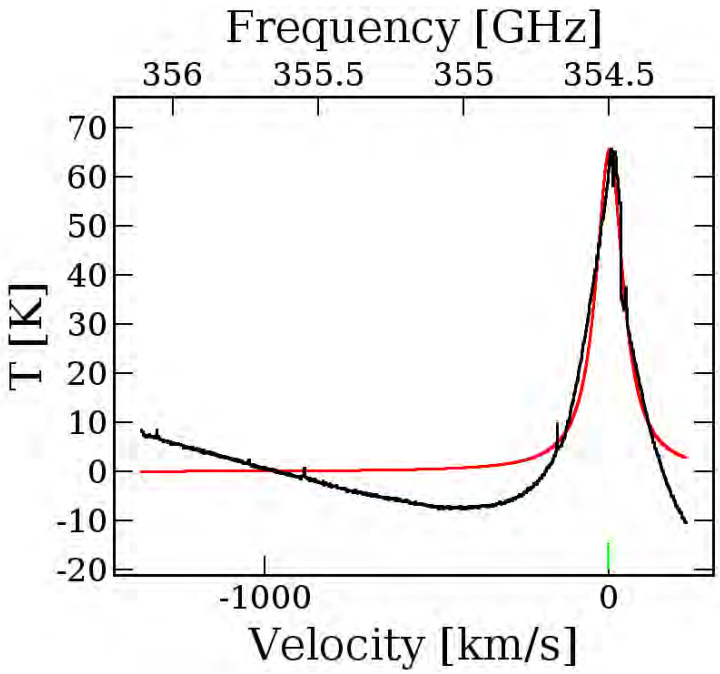


Figure 6: Rotational emission spectrum of HCN at frequency 354.505 GHz [33]. In the emission spectrum, the black spectrum indicated the original transition and the red spectrum indicated the best fit gaussian model over the original spectrum. The green line indicated the peak position of HCN.

between the frequency range of 349.21–349.45 GHz are shown in Figure 7 and 8 [33] [37]. The channel maps show that the emission lines of CH_3CN are mainly coming from the eastern and western limbs of Titan [33]. Recently, the emission lines of CH_3CN have been detected in the high-mass star-formation region IRAS 18566+0408 with Submillimeter Array (SMA) telescope, and this complex molecule has an important role in creating the C_2H_5CN [36]. After a detailed search, we detected another rotational emission line of CH_3CN with transition J = 14–16 with a different K ladder (K = 0–9) in ALMA band 6 observation between the frequency range of 256.86–257.80 GHz, which is shown in Figure 9.

3 Possible Formation Mechanism of Nitrile-Bearing Molecules in the Atmosphere of Titan

The CH_3CN molecule was known as one of the best thermometers to understand the gas temperature of the planetary atmosphere and hot molecular cores. The emission lines of CH_3CN are mainly found in the hot molecular cores and hot corinos. In our solar system, the emission lines of CH_3CN were first detected in the atmosphere of Titan. The CH_3CN molecule records the gas temperature because its numerous K ladder spectral signatures become thermalized at their physical conditions. The CH_3CN molecule was known as a symmetric top molecule, which was created via radiative association of CH_3^+ with HCN in the atmosphere of Titan [38]. The CH_3CN molecule has an important role in the production of C_2H_5CN in the atmosphere of Titan [38]. The complex molecule C_2H_5CN has an important role in enhancing the nitrile chemistry in the atmosphere of Titan. In the gas phase, the C_2H_5CN will produce in the atmosphere of Titan with the help of CH_3^+ and CH_3CN. The possible chemical reaction is:

$CH_3^+ + CH_3CN \longrightarrow C_2H_5CNH^+ + h\nu$ (1) [38]
$C_2H_5CNH^+ + e^- \longrightarrow C_2H_5CN + H$ (2) [39]

In reaction 1, the protonated methane (CH_3^+) reacts with methyl cyanide (CH_3CN) to create $C_2H_5CNH^+$ via radiative association reaction. The C_2H_5CN molecule is produced after the dissociative recombination of $C_2H_5CNH^+$ which is shown in reaction 2. Similarly, the C_2H_3CN molecule was created via the cosmic ray-induced photo re-

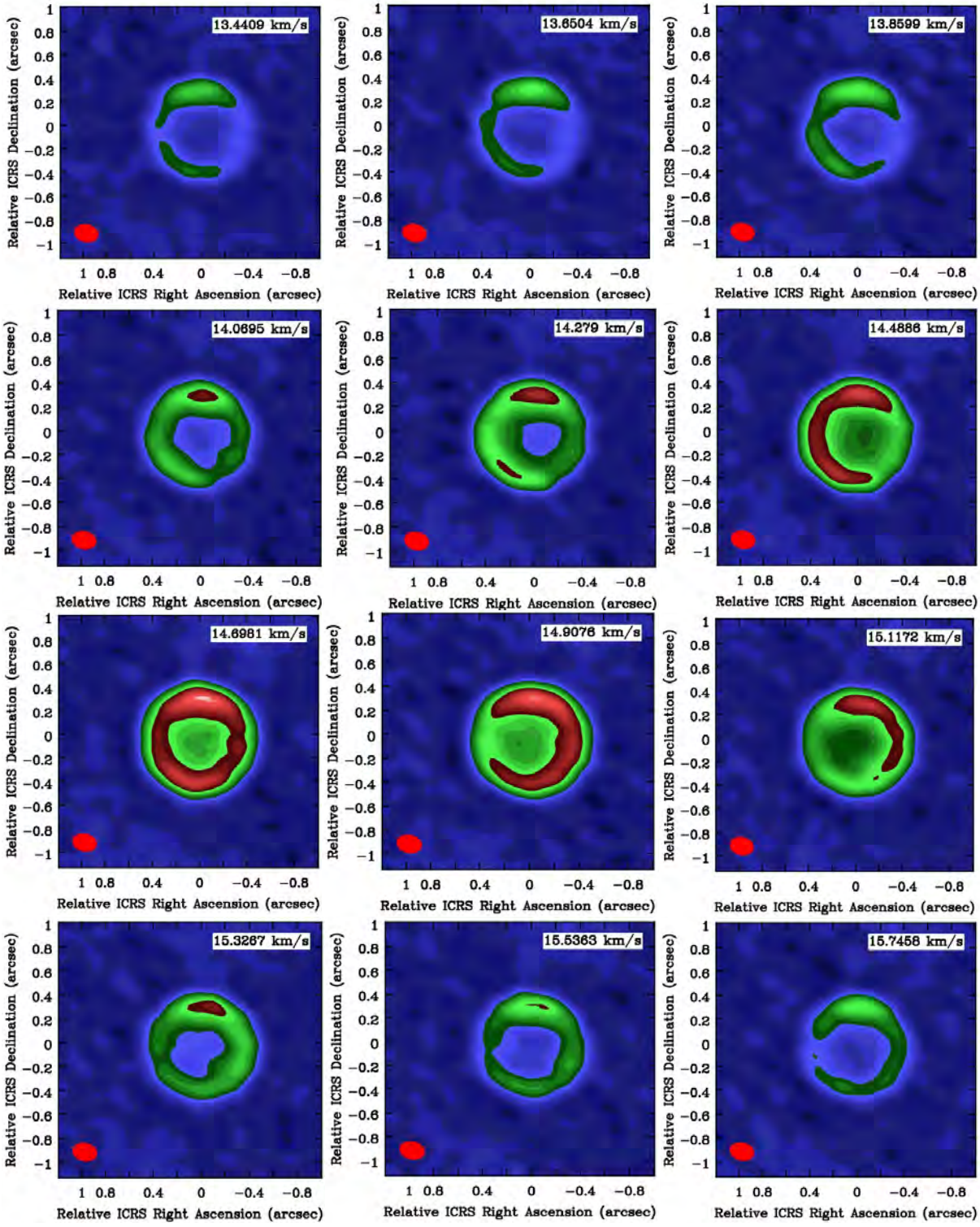


Figure 7: Channel maps of CH_3CN in the atmosphere of Titan with transition J = 19–18. From these channel maps, it was clearly evident that the CH_3CN was coming from the eastern and western limbs of Titan [33] [37]. The red elliptical circles indicated the synthesized beams of the channel maps.

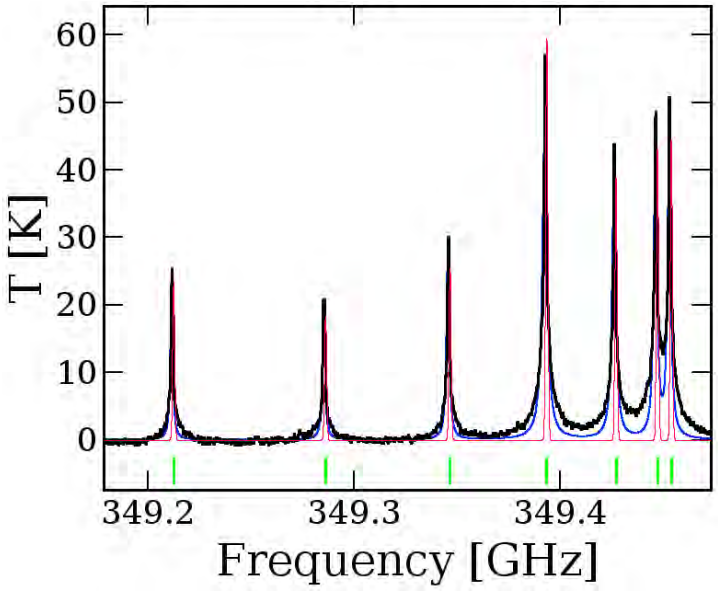


Figure 8: Rotational emission spectrum of CH_3CN between the frequency range 349.21–349.45 GHz with transition J=19–18 [33] [37]. In the emission spectrum, the black line indicated the original transition of CH_3CN, blue spectra indicated the best fit gaussian model, and the red line indicated the best fit Local Thermodynamic Equilibrium (LTE) model over the original spectrum. After fitting the LTE model, we estimate the column density of CH_3CN was 7.0×10^{14} cm^{-2}.

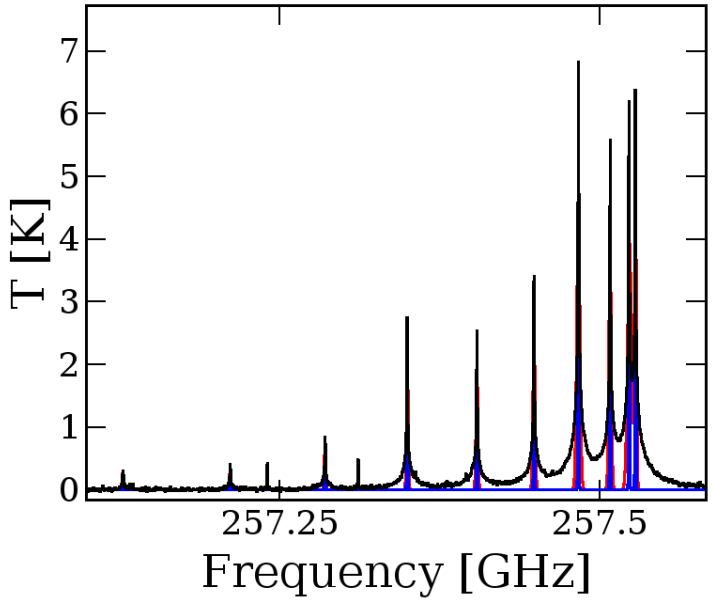


Figure 9: Rotational emission spectrum another transition lines of CH_3CN between the frequency range of 256.86–257.80 GHz with transition J = 14–16. In the emission spectrum, the black line indicated the spectra of Titan, red spectra indicated the best fit gaussian model, and the blue line indicated the best fit Local Thermodynamic Equilibrium (LTE) model over the original spectrum. After fitting the LTE model, we estimate the column density of CH_3CN was $7.5\times10^{14}\,cm^{-2}$.

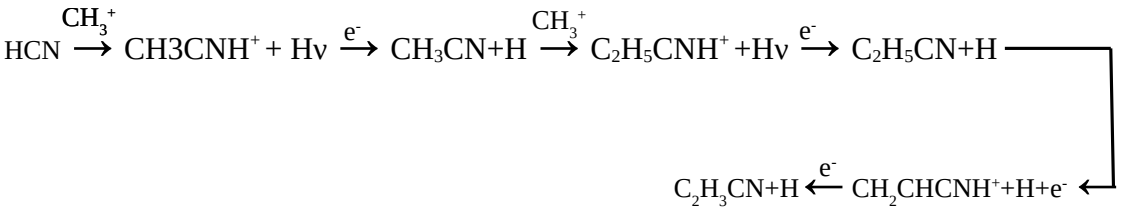


Figure 10: Possible formation mechanism of CH_3CN, C_2H_5CN, and C_2H_3CN and chemical link up with HCN.

action and dissociative recombination of C_2H_5CN. We created a chemical link-up formation route of detected nitrogen-bearing molecules with respect to HCN, which is shown in Figure 10. That reaction network indicated that CH_3CN, C_2H_5CN, and C_2H_3CN are dependent on HCN in the atmosphere of Titan.

The detection of CH_3CN, C_2H_3CN, and C_2H_5CN indicated more complex nitrogen-bearing molecules like propyl cyanide (C_3H_7CN) may be craeated in the upper atmosphere of Titan via the following chemical reactions:

$H_2O+h\nu \longrightarrow H+OH$ (3) [40]
$OH+CH_3CN \longrightarrow CH_2CN+H_2O$ (4) [41]
$CH_2CN+C_2H_6 \longrightarrow C_3H_7CN+H$ (5) [41]

Reaction 3 indicated that the hydroxyl (OH) radical may be created due to the photolysis of H_2O molecule (water vapour) in the upper atmosphere of Titan. Reaction 4 indicated the generated OH may be reacted with CH_3CN in the upper atmosphere of Titan via gas-phase reaction and created cyanomethyl radical (CH_2CN). Now the generated CH_3CN when reacted with C_2H_6 in the upper atmosphere of Titan, they created C_3H_7CN. The observation of the emission lines of C_3H_7CN was needed using the ALMA to understand the nitrile chemistry in the upper atmosphere of Titan.

4 Possible Formation Mechanism of Simplest Amino Acid Glycine (NH_2CH_2COOH) in the Titan

In the upper atmosphere of Titan, the formaldehyde (HCHO) reacts with ammonia (NH_3) and hydrogen cyanide (HCN) to produce the possible glycine precursor molecule aminoacetonitrile (NH_2CH_2CN) using the Strecker synthesis reactions [40]:

$HCHO+NH_3+HCN \rightarrow NH_2CH_2CN$ (6) [40]

In the upper atmosphere of Titan, HCHO was created when atomic oxygen and hydroxyl radical reacted with CH_3 ($CH_3+O \rightarrow HCHO+H$) [40]. Similarly, NH_3 molecule was created in the upper atmosphere of Titan via dissociative electron recombination of $NH_4{}^+$ ion [40]. The NH_2CH_2CN molecule was alternatively known as glycine nitrile [42], and that molecule was found in the hot molecular cores G10.47+0.03 and Sgr B2(N) [42] [43]. The NH_2CH_2CN molecule eventually lands on the surface of Titan from the upper atmosphere during the hydrocarbon rain. Now, the landing NH_2CH_2CN molecule on Titan's surface reacts with hot liquid water in Titan's pond, and NH_2CH_2COOH is formed as a result of the hydrolysis of NH_2CH_2CN as follows:

$NH_2CH_2CN+H_3O^+ + heat \rightarrow NH_2CH_2COOH$ (7) [40]

The hydrolysis of NH_2CH_2CN on the surface of Titan can occur when pure liquid water i.e. auto-ionization of H_2O molecules produces hydronium ions (H_3O^+) [40]. The landed NH_2CH_2CN molecule will freeze on the surface of Titan and they will come to react with hot liquid water for hydrolysis to create NH_2CH_2COOH [40]. This chemical reaction occurs during some surface activity like cryovolcanism (volcano that erupts with ice, water, methane, and ammonia). The mixture of hot liquid water and NH_3 erupts from the surface via cryovolcano activity and creates an impact pond which is filled with underneath NH_3 and hot liquid water [40]. That hypothesis indicated the simplest amino acid NH_2CH_2COOH may be created in the pond via the hydrolysis of NH_2CH_2CN

with hot water, and most likely possible life would be created in the pond of Titan. More chemical simulations were nedded to solve the puzzale of NH_2CH_2CN and NH_2CH_2COOH in the surface of Titan.

5 Summary

In this review, we summarised the detection of complex nitrile-bearing molecules from the atmosphere of Saturn's largest moon, Titan, using the Atacama Large Millimeter/Submillimeter Array (ALMA). Here we discussed the early detection of ethyl cyanide (C_2H_5CN), vinyl cyanide (C_2H_3CN), hydrogen cyanide (HCN), and methyl cyanide (CH_3CN) [13] [15] [33] [37]. We also described the vertical column density of the detected species in the atmosphere of Titan using the atmospheric radiative transfer and LTE models. Here we pointed out the emission spectrum of HCN at frequency 354.505 GHz and CH_3CN between the frequency range 349.21–349.45 GHz with transition J = 19–18 arising from the eastern and western limbs of the Titan [33] [37]. We discussed the possible formation mechanism of the nitrile-bearing molecules including propyl cyanide (C_3H_7CN) and we created a chemical link with HCN. Additionally, we also discuss the possible formation mechanism of the simplest amino acid glycine (NH_2CH_2COOH) on the surface of Titan via the hydrolysis of aminoacetonitrile (NH_2CH_2CN) and hot water. We recommend the planetary science community to look for more complex nitrile-bearing molecules as well as propyl cyanide (C_3H_7CN) and aminoacetonitrile (NH_2CH_2CN) in the upper atmosphere of Titan using the ALMA to uncover the mystery of the atmospheric composition and the hypothesis of the formation of life on the surface of Titan.

Acknowledgements

To reproduce the emission spectrum of C_2H_3CN, HCN, and CH_3CN, we used the following ALMA data: ADS /JAO.ALMA#2015.1.01023.S and 2012.1.00198.S. ALMA is a partnership of ESO (representing its member states), NSF (USA), and NINS (Japan), together with NRC (Canada), MOST and ASIAA (Taiwan), and KASI (Republic of Korea), in co-operation with the Republic of Chile. The Joint ALMA Observatory is operated by ESO, AUI/NRAO, and NAOJ.

References

[1] Brown, R. H., Lebreton, J. P., Waite, J. H., "Titan from Cassini-Huygens", Berlin: Springer, 2009.

[2] Catling, D. C., Kasting, J. F., "Atmospheric Evolution on Inhabited and Lifeless Worlds (1 ed.)", Cambridge University Press, ISBN 978-0521844123, 2017.

[3] Cours, T., Cordier, D., Seignovert, B., Maltagliati, L., Biennier, L., "The 3.4 μm absorption of the Titan's stratosphere: Contribution of ethane, propane, butane, and complex hydrogenated organics", Icarus, 339, 2019.

[4] Niemann, H., Atreya, S. K., Bauer, S., Carignan, G. R., Demick, J. E. et al., "The abundances of constituents of Titan's atmosphere from the GCMS instrument on the Huygens probe", Nature, 438, 779–84, 2006.

[5] Johnson, R. E., Tucker, O. J., Volkov, A. N., "Evolution of an early Titan atmosphere", Icarus, 271, 202–206, 2016.

[6] Charnay, B. et al., "Titan's past and future: 3D modeling of a pure nitrogen atmosphere and geological implications", Icarus, 241, 269–279, 2014.

[7] Bellucci, A. et al., "Titan solar occultation observed by Cassini/VIMS: Gas absorption and constraints on aerosol composition", Icarus, 201(1), 198–216, 2009.

[8] Sandford, S. A. et al., "The interstellar C-H stretching band near 3.4 microns - Constraints on the composition of organic material in the diffuse interstellar medium". The Astrophysical Journal, 371, 607–620, 1991.

[9] Pendleton, Y. J., Allamandola, L. J., "The Organic Refractory Material in the Diffuse Interstellar Medium: Mid-Infrared Spectroscopic Constraints". The Astrophysical Journal Supplement Series, 138, 75–98, 2002.

[10] Nicholson, P. D. et al., "Probing Saturn's Atmosphere with Procyon", In American Astronomical Society, 38th DPS Meeting, 2006.

[11] Kim, S. J. et al., "The three-micron spectral feature of the Saturnian haze: Implications for the haze composition and formation process", Planetary and Space Science, 65, 122–129, 2012.

[12] Maltagliati, L. et al., "Titan's atmosphere as observed by Cassini/VIMS solar occultations: CH_4, CO and evidence for C_2H_6 absorption", Icarus, 248, 1–24, 2015.

[13] Cordiner, M., Palmer, M., Nixon, C., Irwin, P. et al., "Ethyl Cyanide On Titan: Spectroscopic Detection and Mapping Using ALMA". The Astrophysical Journal. 800, L14, 2014.

[14] Stevenson, J., Lunine, J., Clancy, P., "Membrane alternatives in worlds without oxygen: Creation of an azotosome", Science Advances, 2015.

[15] Palmer, M., Cordiner, M., Nixon, C., Charnley, S. et al., "ALMA detection and astrobiological potential of vinyl cyanide on Titan", Science Advances, 3, e1700022, 2017.

[16] Clarke, D., Joseph, J., Ferris, J., "The Design and Use of a Photochemical Flow Reactor: A Laboratory Study of the Atmospheric Chemistry of Cyanoacetylene on Titan", Icarus. 147. 282-291, 2000.

[17] Coustenis, A., Salama, A., Schulz, B., Ott, S., Lellouch, E., Encrenaz, T., Gautier, D., Feuchtgruber, H., "Titan's atmosphere from ISO mid-infrared spectroscopy", Icarus, 161, 383–403, 2003.

[18] Hidayat, T., "Observations heterodynes millimetriques et submillimetriques de Titan: Etude de la composition chimique de son atmosphere", thesis, Universite de Paris-Meudon, 1997.

[19] Marten, A., Hidayat, T., Biraud, Y., Moreno, R., "New millimeter heterodyne observations of Titan: Vertical distributions of nitriles HCN, HC_3N, CH_3CN, and the isotopic ratio 15N/14N in its atmosphere", Icarus, 158, 532–544, 2002.

[20] Manna. A., and Pal, S., "ALMA detection of hydrogen cyanide and carbon monoxide in the atmosphere of Saturn", arXiv:2104.10474, 2021.

[21] Catling, D., Kasting, J. F., "Planetary atmospheres and life", ed. W. T. Sullivan, III & J. A. Baross (Cambridge: Cambridge University Press), 91–116, 2007.

[22] Vinatier, S., B*é*zard, B., Nixon, C. A. et al., "Analysis of Cassini/CIRS limb spectra of Titan acquired during the nominal mission: I. Hydrocarbons, nitriles and CO_2 vertical mixing ratio profiles", Icarus, 205, 559, 2010.

[23] Adriani, A., Dinelli, B. M., López-Puertas, M. et al., "Distribution of HCN in Titan's upper atmosphere from Cassini/VIMS observations at 3 μm, Icarus, 214, 584, 2011.

[24] Koskinen, T. T., Yelle, R. V., Snowden, D. S. et al., "The mesosphere and lower thermosphere of Titan revealed by Cassini/UVIS stellar occultations", Icarus, 216, 507, 2011.

[25] Magee, B. A., Waite, J. H., Mandt, K. E. et al., "INMS-derived composition of Titan's upper atmosphere: Analysis methods and model comparison", P&SS, 57, 2009.

[26] Vuitton, V., Yelle, R. V., Klippenstein, S. J., H*ö*rst, S. M., Lavvas, P., "Simulating the density of organic species in the atmosphere of Titan with a coupled ion-neutral photochemical model", Icarus, 324, 120, 2019.

[27] Gronoff, G., Lilensten, J., Desorgher, L., Fl*ü*ckiger, E., "Ionization processes in the atmosphere of Titan", A&A, 506, 955, 2009.

[28] Lellouch, E., Gurwell, M., Butler, B. et al., "Detection of CO and HCN in Pluto's atmosphere with ALMA", Icarus, 286, 289, 2017.

[29] Marten, A., Gautier, D., Owen, T. et al., "First observations of CO and HCN on Neptune and Uranus at millimeter wavelengths and the implications for atmospheric chemistry", ApJ, 406, 285, 1993.

[30] H*é*brard, E., Dobrijevic, M., Loison, J. C., Bergeat, A., Hickson, K. M., "Neutral production of hydrogen isocyanide (HNC) and hydrogen cyanide (HCN) in Titan's upper atmosphere", A&A, 541, A21, 2012.

[31] Loison, J. C., H*é*brard, E., Dobrijevic, M. et al., "The neutral photochemistry of nitriles, amines and imines in the atmosphere of Titan", Icarus, 247, 218, 2015.

[32] Willacy, K., Allen, M., Yung, Y., "A new astrobiological model of the atmosphere of Titan", ApJ, 829, 79, 2016.

[33] Lellouch, E., Gurwell, M. A., Moreno, R., et al., "An intense thermospheric jet on Titan", Nature Astronomy, 37, 614, 2019.

[34] Catling, D. C., "Planetary Atmospheres", ed. G. Schubert, Treatise on Geophysics, 2nd Ed. (Oxford: Elsevier), 429–472, 2015.

[35] H*ö*rst, S. M., "Titan's atmosphere and climate", JGRE, 122, 432, 2017.

[36] Silva, A., Zhang, Q., Sanhueza, P. et al., "SMA Observations of the Hot Molecular Core IRAS 18566+0408", The Astrophysical Journal. 847. 2017.

[37] Iino, T., Sagawa, H., Tsukagoshi, T., "$^{14}N/^{15}N$ Isotopic Ratio in CH_3CN of Titan's Atmosphere Measured with ALMA". The Astrophysical Journal. 890, 2020.

[38] Anicich, V.G., "Evaluated bimolecular ion-molecule gas phase kinetics of positive ions for use in modeling planetary atmospheres, cometary comae, and interstellar clouds". Journal of Physical and Chemical Reference Data, 22, 1469, 1993

[39] Vigren, E., Hamberg, M., Zhaunerchyk, V., et al., "Dissociative recombination of protonated propionitrile, $CH_3CH_2CNH^+$: implications for Titan's upper atmosphere". The Astrophysical Journal, 722, 847, 2010.

[40] Saxena, P. P., "On the Possibility of Gly and Ala Amino Acids on Titan's Surface", Earth Moon and Planets, 106, 113, 2010.

[41] Belloche, A., Garrod, R. T., Müller, H. S. P., et al., "Detection of a branched alkyl molecule in the interstellar medium: iso-propyl cyanide", Science, 345, 6204, 1584, 2014.

[42] Manna, A., and Pal, S., "Identification of interstellar amino acetonitrile in the hot molecular core G10.47+0.03: Possible glycine survey candidate for the future", Life Sciences and Space Research, 34, 9, 2022.

[43] Belloche, A., Menten, K. M., Comito, C., et al., "Detection of amino acetonitrile in Sgr B2(N)", Astronomy & Astrophysics, 482, 179, 2008.

A Short Story of The New Quantum Methods in Krein Space: $\mathcal{PT}$-Symmetry and Non-Hermiticity

Arindam Chakraborty

Department of Physics, Heritage Institute of Technology, Kolkata-700107, India
email: **arindam.chakraborty@heritageit.edu**

Abstract

The present article briefly reviews $\mathcal{PT}$-symmetric operators and and their relevance in non-hermitian quantum methods. The adoption of Krein Space as an indefinite inner product space becomes natural in this context since the symmetry operator automatically demands so. Furthermore, as opposed to conventional quantum theory where a pre-assigned inner product space determines the possibility of unitary-hermitian frame work, in non-hermitian theory it is the system that dictates the choice of suitable inner product space in a problem specific way. Examples have been constructed at every stage of discussion starting from basic priciples to symmetry operators in krein Space. Some future scopes of investigatuon in this area have also been discussed.

Keywords:*Non-hermitian operator, $\mathcal{PT}$-symmetry, Krein Space, canonical symmetry, conjugation*

"What we observe is not nature itself, but nature exposed to our method of questioning." Werner Heisenberg

1 Introduction

In a typical text-book level quantum mechanics [1, 2, 3] the time evolution of a quantum state is represented by the the equation $|\psi(t)\rangle = U(t)|\psi(0)\rangle$, where $U(t) = e^{-iHt}$ is a **unitary** operator, H is the hamiltonian of the system and $|\psi(t)\rangle$ represents the quantum state of the system at any time t. The operator H has to be **hermitian** (or more technically **self-adjoint** relative to a preassigned inner product space (**Hilbert Space**)) for obvious reason. It is called the **generator of time evolution** and popularly known as the **Hamiltonian** of the system. The unitarity of U ensures the probability conservation whereas the hermitian operator H necessarily gives real eigen values and orthogonal eigenvectors when time independent Schrödinger equation is taken into consideration. The reality of the eigenvalues speaks in favour of real values of energy and orthogonality property determines the stationary states with no spontaneous transition from one such state to another. A dynamical quantity represented by an operator A qualifies as an **observable** iff A is hermitian and the commutator $[H, A] = 0$. These fact ensures that the operators A and H are simultaneously measurable and both can be expanded in terms of the **Orthogonal Projection Operators** in the respective eigensubspaces. The last criterion is known as **Spectrsl Representation** which allows projective measurement of a physical quantity. In every quantum system there exits a minimal set of mutually commuting observables which can completely determine the quantum state of the system. Such a set is called **Complete Set of Commuting Observables (CSCO)**. The prime objective of canonical quantum method can thus be summerized into following steps : (i) Finding eigenvalues and eigenstates of the hamiltonian, (ii) Determining the CSCO to obtain the set of uniquely specified quantum states. This has been condidered as the **Dirac Quantization**. Though this method has not been beyond question on several grounds even since the time of Dirac our present discussion is in consonance with the Dirac formalism.

The study of quantum mechanical formalism involving non-hermitian operators has gained much attention for last couple of decades engendering a major epistemological break in the practice of conventional quantum mechanics. Since the identification of non-hermitian operators with real spectrum and Bender and Boettcher's [4, 5, 6] attribution of this possibility to space-time reflection symmetry ($\mathcal{PT}$-symmetry), numerous studies have been undertaken in a variety of contexts relating to discrete symmetries [7, 8, 9]. Non-hermitian formulation shows its relevance when complications arise in the usual hermitian formalism (cases with complex potentials), in the study of so called resonance phenomena associated with nuclear, molecular or atomic systems or even with nano-structured materials or condensates, in understanding systems which are not so quantum mechanical in sense but their physical behavior is quite amenable to quantum language (for example classical statistical mechanical systems, biological systems with diffusion, light propagation in wave guides) and many other fields where even the conventional quantum mechanics has already shown success [10, 11].

It is to be noted that the notion of unitary-hermitian framework is always associated to the notion of inner product space. As a consequence one can alter the definition of inner product to make sense of an otherwise non-hermitian operator to a hermitian one. Such a possibility has been emphasized in [8, 9, 16] and leads us to the notion of pseudo-hermiticity, a concept introduced in the 1940s by Dirac and Pauli and later discussed by Lee,Wick and Sudarshan, who were trying to resolve the problems that arose in quantization of electrodynamics and other quantum field theories in which negative-norm states used to appear as a consequence of re-normalization [12, 13, 14, 15].

The present discussion makes a brief review of the possibility of a new quantum method based on $\mathcal{PT}$-symmetry and Non-hermitian operators. As opposed to conventional quantum mechanics where the operators are defined in relation to a preassigned inner product, a system admitting non-hermitian hamiltonian itself determines the inner-product space (and hence the relevant Hilbert space) suitable to it. We shall discuss this possibility in view of the actions of one or more symmetry operators which in turn are capable to ward off the troubles with non-hermiticity. A number of concrete examples will also be introduced. The notion of $\mathcal{PT}$-symmetry comes into existence in this context. Let us define an operator $\mathcal{P}$ whose action has the effect:$x \to -x$, $p \to -p$ and an operator $\mathcal{T}$ whose action has the effect : $x \to x$, $p \to -p$ and $i \to -i$. A quantum hamiltonian is said to be $\mathcal{PT}$**-symmetric** if it remains to be the same under the simultaneous action of $\mathcal{P}$ and $\mathcal{T}$. In quantum language this is equivalent to saying $[H, \mathcal{PT}] = 0$. For example the hamiltonian of 1-D Harmonic Oscillator $H = \frac{1}{2}(P^2 + X^2)$ is $\mathcal{PT}$-symmetric. A hamiltonian is called pseudo-hermitian under $\mathcal{PT}$-transformation if $(\mathcal{PT})H(\mathcal{PT}) = H^\dagger$. The action of $\mathcal{PT}$-transformation on eigenfunctions of a hamiltonian may or may not render them unchanged corresponding to $\mathcal{PT}$**-invariant** and $\mathcal{PT}$**-broken** states respectively.

2 Basics of Indefinite Inner Product Space

Definition 1. *An inner product on a complex vector space $\mathcal{V}$ is a complex valued function $\langle\cdot|\cdot\rangle : \mathcal{V} \times \mathcal{V} \to \mathbf{C}$ and defined for all pairs of vectors $|u\rangle, |v\rangle \in \mathcal{V}$ with the properties (i)$\langle\alpha u + \beta v|w\rangle = \alpha\langle u|w\rangle + \beta\langle v|w\rangle$, $\forall\ \alpha, \beta \in \mathbf{C}$ and (ii) $\langle u|v\rangle = \overline{\langle v|u\rangle}$. A vector space equipped with an inner product is called an* ***Inner product Space****.[17]*

Remark 1. *(i) The above definition implies the an tilinear relation $\langle w|\alpha u + \beta v\rangle = \overline{\alpha}\langle w|u\rangle + \overline{\beta}\langle w|v\rangle$*

(ii) When the quantity $\langle u|u\rangle >, <= 0$ u is said to be positive, negative, neutral respectively. $\mathcal{V}$ becomes an ***indefinite inner product space****.*

(iii) Two vectors $|u\rangle, |v\rangle \in \mathcal{V}$ are said to be orthogonal if $\langle u|v\rangle = 0$.

example : $\{\mathbf{R^2}, \langle\cdot|\cdot\rangle_M\}$ is an indefinite inner product space where $\langle u|v\rangle_M = (u_1 \quad u_2)\begin{pmatrix} 1 & 1 \\ 1 & -3 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$. Where, M stands for the matrix $\begin{pmatrix} 1 & 1 \\ 1 & -3 \end{pmatrix}$. It is immediate to see $\langle u|u\rangle = (u_1 + u_2)^2 - 4u_2^2$. This means $\langle u|u\rangle >, =, < 0$ when $\left(1 + \frac{u_1}{u_2}\right)^2 >, =, < 4$ respectively. As for example $\begin{pmatrix} 2 \\ 1 \end{pmatrix}$ is a positive vector $\begin{pmatrix} 1 \\ 3 \end{pmatrix}$ is a negative vector and $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} 3 \\ -1 \end{pmatrix}$ are neutral vectors. The vector $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ is the trivial neutral vector. One can also verify that the vectors $|r\rangle = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$ and $|s\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ are mutually orthogonal (or M-orthogonal)vectors in $\{\mathbf{R^2}, \langle\cdot|\cdot\rangle_M\}$. We write $|r\rangle[\perp]|s\rangle$.

Let us introduce the following notations

$$\begin{aligned}
\mathcal{V}^+ &= \{|u\rangle \in \mathcal{V} : \langle u|u\rangle \geq 0\} \\
\mathcal{V}^- &= \{|u\rangle \in \mathcal{V} : \langle u|u\rangle \leq 0\} \\
\mathcal{V}^{++} &= \{|u\rangle \in \mathcal{V} : \langle u|u\rangle > 0\} \\
\mathcal{V}^{--} &= \{|u\rangle \in \mathcal{V} : \langle u|u\rangle < 0\} \\
\mathcal{V}^{00} &= \{|u\rangle \in \mathcal{V} : \langle u|u\rangle = 0\}
\end{aligned}$$

The above sets are called **lineals** (linear subsets) of $\mathcal{V}$. They are not subspaces since for example the vectors $|p\rangle = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$ and $|q\rangle = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$ are positive vectors in $\{\mathbf{R^2}, \langle\cdot|\cdot\rangle_M\}$ but $|p\rangle - |q\rangle$ is a negative vector in the sense of relevant inner product.

Lemma 1. *An indefinite inner product space contains non-zero neutral vectors.*

Proof : Let us consider $|u\rangle \in \mathcal{V}^{++}$ and $|v\rangle \in \mathcal{V}^{--}$ and set $|w\rangle = |u\rangle + \lambda|v\rangle$ so that λ is a real solution of the equation

$$\langle u|u\rangle + 2\lambda Re\langle u|v\rangle + \lambda^2\langle v|v\rangle = 0 \tag{1}$$

The above equation is nothing but the claim $\langle w|w\rangle = 0$. Now if $|w\rangle = \mathbf{0}$, $|u\rangle = -\lambda|v\rangle$ or $\langle u|u\rangle = \lambda^2\langle v|v\rangle$ contradicting our premise. q. e. d.

Note : If $\mathcal{V}$ is an inner product space for a pair of vectors $|u\rangle, |v\rangle \in \mathcal{V}$ one can replace each value $\langle u|v\rangle$ by $\langle u|v\rangle' = -\langle u|v\rangle$ to obtain the so called **anti-space** $\mathcal{V}'$ of $\mathcal{V}$.

Definition 2. *Two lineals $\mathcal{M}, \mathcal{N} \in \mathcal{V}$ are said to be orthogonal to each other if $|q\rangle[\perp]|r\rangle\ \ \forall\ |q\rangle \in \mathcal{M}$ and $\forall\ |r\rangle \in \mathcal{N}$.*

Definition 3. *A vector $|k\rangle \in \mathcal{V}$ is called an* ***isotropic vector*** *if $|k\rangle[\perp]\mathcal{V}$. Correspondingly an* ***isotropic lineal*** *$\mathcal{V}^0$ is the set of all vectors such that $\mathcal{V}^0 = \mathcal{V} \cap \mathcal{V}^\perp$.*

Remark 2. *(i)If $\mathcal{V}^0$ contains only the trivial neutral vector it is called* ***non-degenerate*** *otherwise* ***degenerate****. It is easy to appreciate that every definite lineal is nondegenerate.*

(ii)It has been claimed [18] that not all non-degenerate lineals $\mathcal{V}$can be decomposed into the direct sum $\mathcal{V} = \mathcal{V}^+ \dot{+} \mathcal{V}^-$ of a positive lineal ($\mathcal{V}^+$) and negative lineal ($\mathcal{V}^-$) the converse is true. This means if $\mathcal{V} = \mathcal{V}^+ + \mathcal{V}^-$ where, $\mathcal{V}^+ \subset \mathcal{V}^{++} \cup \{\mathbf{0}\}$ and $\mathcal{V}^- \subset \mathcal{V}^{--} \cup \{\mathbf{0}\}$ then $\mathcal{V}$ is non-degenerate.

Definition 4. *Let the linear space $\mathcal{K}$ has a decomposition into the direct sum of positive and negative lineals i.e.; $\mathcal{K} = \mathcal{K}^+ \dot{+} \mathcal{K}^-$. Such a decomposition is called* ***canonical decomposition*** *if $\mathcal{K}^+[\perp]\mathcal{K}^-$.*

2.1 Krein Space

Definition 5. *A linear space $\mathcal{K}$ is called a* ***Krein Space*** *when the above mentioned canonical decomposition holds and the the lineals $\mathcal{K}^+$ and $\mathcal{K}^-$ are complete in the norms $||u|| = \langle u|u\rangle^{\frac{1}{2}}$ when $|u\rangle \in \mathcal{K}^+$ and $||u|| = -\langle u|u\rangle^{\frac{1}{2}}$ when $|u\rangle \in \mathcal{K}^-$ respectively. This makes the lineals* ***Hilbert Spaces****.*

The notion of Krein Space is very crucial in the context of $\mathcal{PT}$-symmetry. The following objects will be necessary in this regard.

Definition 6. *Let $\{\mathcal{H},\ \langle\cdot|\cdot\rangle\}$ be a Hilbert Space. A* ***fundamental/canonical symmetry*** *J is an operator with the following properties : (i) $J^2 = \mathbf{1}$ and (ii) $\langle J\psi|J\phi\rangle = \langle\psi|\phi\rangle$, (iii) $\langle\psi|J\phi\rangle = \langle J\psi|\phi\rangle$.*

If J is a nontrivial fundamental symmetry (i.e.;$J \neq \pm\mathbf{1}$) Then the Hilbert space equipped with the indefinite inner product $\langle\psi|\phi\rangle_J = \langle J\psi|\phi\rangle$ becomes a Krein Space : $\{\mathcal{H},\ \langle\cdot|\cdot\rangle_J\} = \mathcal{K}$. It is also obvious from the definition that $J = J^\dagger = J^{-1}$ and the operator introduces a **fundamental decomposition** of

$$\{\mathcal{H},\ \langle\cdot|\cdot\rangle_J\} = \mathcal{H}_+ \dot{+} \mathcal{H}_- \tag{2}$$

where, $\mathcal{H}_\pm = P_\pm\mathcal{H}$. $P_\pm$ are orthoprojectors in $\{\mathcal{H},\ \langle\cdot|\cdot\rangle\}$. Furthermore the following observations are immediate [19]

$$JP_\pm = \pm P_\pm,\quad J = P_+ - P_-,\ \ P_+P_- = P_-P_+ = 0 \tag{3}$$

The subspaces $\mathcal{H}_\pm$ are mutually orthogonal relative to the inner product $\langle\cdot|\cdot\rangle$ as well as $\langle\cdot|\cdot\rangle_J$. Since $P_+ + P_- = \mathbf{1}$, $P_\pm = \frac{1}{2}(\mathbf{1} \pm J)$

example : Let's consider the Hilbert space $\{\mathbf{R}^2, \langle\cdot|\cdot\rangle\}$ with $\langle x|y\rangle = x_1y_1 + x_2y_2$ and the projectors $P_+ = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ and $P_- = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$ along $y = x$ and $y = -x$ respectively. The fundamental symmetry operator $J = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. This gives $\langle x|y\rangle_J = x_1y_2 + x_2y_1$.

Proposition 1. *Let $|u_\pm\rangle, |v_\pm\rangle \in \mathcal{H}_\pm$ with $|u\rangle = |u_+\rangle + |u_-\rangle$ and $|v\rangle = |v_+\rangle + |v_-\rangle$. Then $\langle u|v\rangle = \langle u_+|v_+\rangle_J - \langle u_-|v_-\rangle_J$.*

Proof :

$$\langle u_+|v_+\rangle_J = \langle Ju_+|v_+\rangle = \langle (P_+ - P_-)u_+|v_+\rangle = \langle (P_+)u_+|v_+\rangle = \langle u_+|v_+\rangle \tag{4}$$

using the fact $P_+|u_-\rangle = 0$. Similarly using $P_-|u_+\rangle = 0$

$$\begin{aligned}\langle u_-|v_-\rangle_J = \langle Ju_-|v_-\rangle = \langle (P_+ - P_-)u_-|v_-\rangle = -\langle (P_-)u_-|v_-\rangle \\ = -\langle u_-|v_-\rangle\end{aligned} \tag{5}$$

Subtracting eqn-5 from eqn-4 we get

$$\langle u_+|v_+\rangle + \langle u_-|v_-\rangle = \langle u_+ + u_-|v_+ + v_-\rangle = \langle u|v\rangle \ \ q.e.d. \tag{6}$$

Note : Existence of fundamental decomposition is in fact often cited as a defining criterion of Krein space where $\langle\cdot|\cdot\rangle_J$ and $-\langle\cdot|\cdot\rangle_J$ are Hilbert spaces. It is to be noted that $\langle u_+|u_+\rangle_J = \langle JP_+u|P_+u\rangle = \langle P_+u|P_+u\rangle = ||u_+||^2$. Similarly, $\langle u_-|u_-\rangle_J = -||u_-||^2$. A J-orthogonal pair of vectors $|u\rangle$ and $|v\rangle$ in the Krein space $\{\mathcal{H}, \langle\cdot|\cdot\rangle_J\}$ satisfies the condition $\langle u|v\rangle_J = 0$ and written as $|u\rangle[\perp]_J|v\rangle$.

Definition 7. *A fundamental decomposition (eqn-2) becomes a J-orthogonal direct sum of subspaces $\mathcal{L}_+$ and $\mathcal{L}_-$ when $\mathcal{L}_\pm = (\mathbf{1} + T)\mathcal{H}_\pm$.*

One can write $\mathcal{H} = \mathcal{L}_+[\dot{+}]\mathcal{L}_-$. The operator T is a hermitian strong contraction i.e.; $||Tu|| < ||u||\ \ \forall\ \ |u\rangle(\neq \mathbf{0}) \in \mathcal{H}$ and $TJ + JT = 0$.

3 $\mathcal{PT}$-Symmetric Formulation in Krein Space

It has been observed that in numerous occasions $\mathcal{PT}$-symmetric Hamiltonians can be analysed in Krein Space framework. Before going into the details of Krein Space behaviour of operators let us consider some general properties of $\mathcal{PT}$-symmetric operator.

Definition 8. *Let us consider a complex Hilbert Space $\{\mathcal{H}, \langle\cdot|\cdot\rangle\}$ and a bounded operator $\mathcal{T}$ defined on it. Assuming the inner product to be linear in the first argument, $\mathcal{T}$ is called a **conjugation** if (i) $\mathcal{T}^2 = \mathbf{1}$ and (ii) $\langle \mathcal{T}u|\mathcal{T}v\rangle = \langle u|v\rangle \ \forall \ |u\rangle, |v\rangle \in \{\mathcal{H}, \langle\cdot|\cdot\rangle\}$. The conjugate operator is antilinear i.e.; $\mathcal{T}(a|u\rangle + b|v\rangle) = \overline{a}\mathcal{T}|u\rangle + \overline{b}|v\rangle \ \forall \ a, b \in \mathbf{C}$.*

We consider a fundamental symmetry on this space $J = \mathcal{P}$ and $\mathcal{PT} = \mathcal{TP}$.

Definition 9. *A linear operator L is called $\mathcal{PT}$-symmetric if $\mathcal{PT}L = L\mathcal{PT}$ in the domain $\mathcal{D}(L)$*

One of the troubles in working with self adjoint operator in Krein Space is that it does not ensure the unitary evolution generated by it. This is natural since an indefinite inner product space can produce negative norms there appears problem of interpretation in relation to quantum probability. This trouble is overcome by introducing a new symmetry operator $\mathcal{C}$ with a definition of inner product $\langle\cdot|\cdot\rangle_{\mathcal{C}}$. It has been shown to be possible to find such an operator specific to a given system. Let's start with the following definition

Definition 10. *Let's consider a Krein Space $\mathcal{K} = \{\mathcal{H}, \langle\cdot|\cdot\rangle_J\}$ with J as the fundamental symmetry. The adjoint of an operator L is given by L^+ and defined by the relation $\langle Lu|v\rangle_J = \langle u|L^+v\rangle_J$ where, $|u\rangle \in \mathcal{D}(L)$ and $|v\rangle \in \mathcal{D}(L^+)$.*

Remark 3. *According to the above definition if $L^\dagger$ is the adjoint relative to $\{\mathcal{H}, \langle\cdot|\cdot\rangle\}$, $\langle Lu|v\rangle_J = \langle u|L^+v\rangle_J \Rightarrow \langle JLu|v\rangle = \langle Ju|L^+v\rangle \Rightarrow \langle u|L^\dagger J|v\rangle = \langle u|JL^+|v\rangle$ or $L^\dagger J = JL^+$. If L is J-selfadjoint $L^+ = L$ or $JL = L^\dagger J$ or $JLJ = L^\dagger$ from the properties of J.*

3.1 Construction of $\mathcal{C}$ Operator

We shall only consider the case when $\mathcal{C}$ is a bounded operator.

Definition 11. *Let there be pre-given fundamental decomposition $\{\mathcal{H}, \langle\cdot|\cdot\rangle_J\} = \mathcal{L}_+[\dot{+}]\mathcal{L}_-$. For an arbitrary element $|w\rangle \in \{\mathcal{H}, \langle\cdot|\cdot\rangle_J\}$ admits a decomposition $|w_{\mathcal{L}_+}\rangle + |w_{\mathcal{L}_-}\rangle$. The operator $\mathcal{C}$ is defined by the following action : $\mathcal{C}|w\rangle = \mathcal{C}(|w_{\mathcal{L}_+}\rangle + |w_{\mathcal{L}_-}\rangle) = |w_{\mathcal{L}_+}\rangle - |w_{\mathcal{L}_-}\rangle$*

Let us define a new inner product $\langle\cdot|\cdot\rangle_{\mathcal{C}} = \langle\mathcal{C}\cdot|\cdot\rangle_J$. From previous section this makes

$$\langle w|w\rangle_{\mathcal{C}} = \langle w_{\mathcal{L}_+}|w_{\mathcal{L}_+}\rangle_J - \langle w_{\mathcal{L}_-}|w_{\mathcal{L}_-}\rangle_J \tag{7}$$

The operator $\mathcal{C}$ is a self adjoint operator in the Hilbert Space $\{\mathcal{H}, \langle\cdot|\cdot\rangle_{\mathcal{C}}\}$.and $\mathcal{C}^2 = \mathbf{1}$. The makes $\mathcal{C}$ a fundamental symmetry of the Krein Space $\{\mathcal{H}, \langle\cdot|\cdot\rangle_J\}$ with an underlying Hilbert Space $\{\mathcal{H}, \langle\cdot|\cdot\rangle_{\mathcal{C}}\}$. Hence, the whole discussion boils down to the relation

$$\{\mathcal{H}, \langle\cdot|\cdot\rangle_{\mathcal{C}}\} = \{\mathcal{H}, \langle\mathcal{C}\cdot|\cdot\rangle_J\} = \{\mathcal{H}, \langle J\mathcal{C}\cdot|\cdot\rangle\} \tag{8}$$

Considering $J\mathcal{C} = e^Q$ with Q as a bounded self-adjoint operator in $\{\mathcal{H}, \langle\cdot|\cdot\rangle\}$. This makes $\mathcal{C} = Je^Q$. Using $\mathcal{C}^2 = \mathbf{1}$ we get $Je^Q = e^{-Q}J$ justifying the anticommutation relation $QJ + JQ = 0$. It has been shown by Kuzhel [20] that the projection operators on $\mathcal{L}_\pm$ is given by

$$P_{\mathcal{L}_\pm} = (\mathbf{1} - T)^{-1}(P_\pm - TP_\mp) \tag{9}$$

Using the definition of $\mathcal{C}$ we get

$$\mathcal{C} = P_{\mathcal{L}_+} - P_{\mathcal{L}_-} = (\mathbf{1} - T)^{-1}(\mathbf{1} + T)J \tag{10}$$

Now choosing $(\mathbf{1} - T)^{-1}(\mathbf{1} + T) = e^Q, T = \tanh\frac{Q}{2}$

Finally the whole idea of $\mathcal{PT}$-symmetric system can be summed up into following steps :

•Given a $\mathcal{PT}$-symmetric operator L in a Hilbert Space $\{\mathcal{H}, \langle\cdot|\cdot\rangle\}$ for a system.

•Interpreting L as a self-adjoint operator in the Krein Space. $\{\mathcal{H}, \langle\cdot|\cdot\rangle_J\}$ with $[J, \mathcal{PT}] = 0$.

•Constructing $\mathcal{PT}$-symmetric operator $\mathcal{C} = Je^Q$.

•Considering L as a self-adjoint operator in $\{\mathcal{H}, \langle\cdot|\cdot\rangle_{\mathcal{C}}\}$.

example : Take the Hilbert Space $\{\mathbf{C}^2, \langle\cdot|\cdot\rangle\}$ where, $\langle u|v\rangle = \overline{u_1}v_1 + \overline{u_2}v_2$ and a linear operator L of the form $L = \sum_{j=0}^{3} c_j\sigma_j$. Here,$c_j \in \mathbf{C} \ \forall j$ and $\{\sigma_j : j = 0, 1, 2, 3\}$ are given by

$$\sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \ \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \ \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \ \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{11}$$

We choose $\mathcal{P} = \sigma_1$ and the action of $\mathcal{T}$ is defined by $\mathcal{T}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} \overline{u_1} \\ \overline{u_2} \end{pmatrix}$ This means the operator $\mathcal{PT}$ is to be represented by $\sigma_1\mathcal{T}$. It is easy to verify that $[L, \sigma_1\mathcal{T}] = 0$ i.e.; L is $\sigma_1\mathcal{T}$-symmetric provided $a_0, a_1, a_2 \in \mathbf{R}$ and $a_3 \in i\mathbf{R}$.

Proposition 2. *The spectrum of L is real iff $a_1^2 + a_2^2 - |a_3|^2 \geq 0$*

Proof : Considering the characteristic equation with $\lambda \in \mathbf{C}$ being the eigenvalue

$$\det(L - \lambda\sigma_0) = 0 \Rightarrow \lambda^2 - L\lambda + \det L = 0 \tag{12}$$

the real roots are possible iff $(L)^2 - 4\det L \geq 0$ giving $a_1^2 + a_2^2 - |a_3|^2 \geq 0$ q. e. d.

Proposition 3. *The general form of symmetry $\mathcal{C}$ in the Krein Space $\{\mathbf{C}^2, \langle\cdot|\cdot\rangle_{\mathcal{C}}\}$ is given by $\mathcal{C} == e^Q J = e^{\rho\Xi} J$, where, $Q = q_2\sigma_2 + q_3\sigma_3 = \rho(\cos\xi\sigma_2 + \sin\xi\sigma_3) = \rho\Xi$*

Proof : Letting $J = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ as a fundamental symmetry in $\{\mathbf{C}^2, \langle\cdot|\cdot\rangle\}$ we consider a hermutian operator Q in $\{\mathbf{C}^2, \langle\cdot|\cdot\rangle\}$ of the form $Q = \sum_{j=0}^{3} q_j\sigma_j : q_j \in \mathbf{C}$. The hermiticity of Q implies $q_j \in \mathbf{R}$ and anticommutation with J leads to $q_0 = q_1 = 0$. This makes $Q = q_2\sigma_2 + q_3\sigma_3$. Choosing $\rho = (q_2^2 + q_3^2)^{\frac{1}{2}}$, $\cos\xi = \frac{q_2}{\rho}$ and $\sin\xi = \frac{q_3}{\rho}$ we get $Q = \rho\Xi$. Therefore

$$\begin{aligned} \mathcal{C} = e^Q J = e^{\rho\Xi} \\ = \cosh\rho\sigma_0 + \sinh\rho\Xi \\ = \begin{pmatrix} -i\sinh\rho\cos\xi & \cosh\rho + \sinh\rho\sin\xi \\ \cosh\rho - \sinh\rho\sin\xi & i\sinh\rho\cos\xi \end{pmatrix} \end{aligned} \tag{13}$$

For arbitrary fundamental symmetry see [10].

4 Comments and Future Scopes

Non-hermutian quantum mechanics both as a method and as a physically meaningful area of investigation is a rich theory that has the potential to transcend the ideological boundary of all that has been so very **QUANTUM** for a span of nearly hundred years. A number of techniques and applications have been discussed in both finite and infinite dimension in [10, 21]. Attempts have also been made in the many particle domain. The existence of real eigenvalues of nonhermitian operator does not necessarily implies $\mathcal{PT}$-symmetry. In a recent article [22], the issue of partial $\mathcal{PT}$-symmetry has been investigated for N-coupled harmonic oscillator Hamiltonian with purely imaginary coupling terms, whereas the reality and partial reality of the spectrum are claimed to have direct correspondences with the classical trajectories. Investigation has also been done by Chakraborty [23, 24] for multiboson Hamiltonian associated to polynomial algebras especially to **Higgs algebra** and the so called partial $\mathcal{PT}$-symmetry has been explored as **Weighted Composition Conjugation** in bosonic **Fock Space** viewed as **Reproducing Kernel Hilbert Space**. Many relevant attributes of non-hermitian system including the existence of so called symmetry breaking and symmetry obeying states are discussed in relation to the reality of eigenvalues. The Krein Space structure of all these cases may be an interesting area of investigation.

References

[1] Cohen-Tannoudji C, Diu B, Laloe F 1991 *Quantum Mechanics Vol-1*(chapter:2, 3) (Michigan : Wiley)

[2] Moretti V 2013 *Spectral Theory and Quantum Mechanics*(chapter:7-13) (Italia : Sringer-Verlag)

[3] Moretti V 2019 *Fundamental Mathematical Structures of Quantum Mechnics*(chapter:2-4) (Switzerland AG: Springer Nature)

[4] Bender C M and Boettcher S 1998 *Phys. Rev. Lett.* **80** 5243

[5] Bender C M, Boettcher S and Meisinger P N 1999 *J. Math. Phys.* **40** 2201

[6] Bender C M, Brody D C and Jones H F 2002 *Phys. Rev. Lett.* **89** 27040

[7] Brody D C 2016 *J. Phys. A: Math. Theor.* **49** 10LT03.

[8] Mostafazadeh A 2002 *J. Math. Phys.* **43** 205

[9] Bender C M 2007 *Rep. Prog. Phys.* **70** 947

[10] Bagarello F, Gazeau J P, Szafraniec F H, Znojil M (Editors) 2015 *Non-Selfadjoint Operators in Quantum Physics* 2015 (chapter-6 specifically pp 323-324 regarding $\mathcal{PT}$ symmetry) (New Jersry : John Wiley and Sons, Inc.)

[11] Moiseyev N 2011 *Non-Hermitian Quantum Mechanics* (chapter-3, 4) (Chambridge(UK) : Chembridge University Press)

[12] Jones H F 2005 *J. Phys. A: Math. Gen.* **38** 1741

[13] Pauli W 1943 *Rev. Mod. Phys.* **15** 175

[14] Sudarshan E C G 1961 *Phys. Rev.* **123** 2183

[15] Lee T D and Wick G C 1969 *Nucl. Phys. B* **9** 209

[16] Mostafazadeh A 2002 *J. Math. Phys.* **43** 2814

[17] Bognár J 1974 *Indefinite Inner Product Spaces* (Berlin, Heidelberg, New York : Springer-Verlag).

[18] Azizov T. Ya. and Iokhvidov I S 1989 *Linear Operators on Spaces with an indefinite Metric* (Chichester, New York, Bisbrane, Toronto, Singapore : John Wiely and Sons).

[19] Bender C M et. al. 2019 *PT Symmetry in Quantum and Classical Physics* (New Jersey, London, Singapore, Beiging, Shanghai, Hong Kong, Taipei, Chennai, Tokyo : World Scientific)

[20] Kuzhel S 2009 *Oper Theory Adv Appl* **190** 375

[21] Bagarello F, Trapani C, Passante R, (Editors) 2016 *Non-hermitian Hamiltonians in Quantum Physics* (Switzerland : Spinger International Publishing.)

[22] Beygi A, Klevansky S P and Bender C M 2015 *Phys. Rev. A.* **91** 062101

[23] Chakraborty A 2020 *J. Phys. A: Math. Theor.* 53 485202

[24] Chakraborty A 2021 *Int. J. Theor. Phys.* **60** 3689

Atmospheric and Space Sciences

Post Monsoon Air Pollution Episodes over Megacity New Delhi

T. Mukherjee[1, 2*], V. Vinoj[1]

[1]School of Earth Ocean and Climate Science, Indian Institute of Technology, Bhubaneswar, India
[2]Department of Atmospheric Science, University of Calcutta, India
*Corresponding Author: tm14@iitbbs.ac.in

Abstract The national capital area of Delhi has experienced several post-monsoonal air pollution episodes in the past few years. The elevated concentration of the pollutants during these episodes created immense discomfort for the resident of Delhi. The present study explores the air quality of Delhi and the pollution sources during the pollution episodes. The review also provides a detailed introspect about the major contributor of these pollution episodes based on previous case studies.
Keywords:Air pollution, Delhi, Crop Residue Burning, Diwali, WRF-Chem.

1. Introduction

The increase in global air pollution levels has become a major concern for scientists. The elevated pollution level can not only impact the health of the inhabitants but also alter the Earth's radiation budget and the climatic systems [1]. It is well established that high air pollution levels can directly create health issues and affect people's livelihood [2]–[5]. Several anthropogenic air pollutants like PM2.5 (particulate matter of size less than 2.5 µm), Black Carbon, Sulphate, Nitrogen Oxides (NO_x), surface ozone have increased significantly over the past few decades. The rapid socio-economic growth of the South Asian region directed the rise of these pollutants over this area [6]–[8].

Among the pollutants, PM2.5 is one of the harmful contaminants which creates several negative impacts on human health and can cause premature deaths[9], [10] PM2.5 particles enter the alveoli, subsequently retained in the lung parenchyma[11]. Thus, it can create several cardiovascular and respiratory diseases and even lung cancer [12], [13]. The global burden of disease study (GBD 2010)[14] reported that PM2.5 is the sixth most significant cause of premature death in the South Asian region. Therefore, the increase of PM2.5 above a certain level can cause severe damage to the inhabitants of a particular area.

The Indo-Gangetic Plain (IGP) is considered one of the world's most populated and polluted areas. The high population and associated industrial activities have caused high pollution levels over this region [15], [16]. In this connection, it can be recalled that as per WHO, 13 out of 20 most polluted cities are located in India [17]. The elevated pollution level impacts the air quality of these cities, including megacities like New Delhi.

New Delhi is considered one of the most polluted cities globally [17]. The drastic degradation of the city's air quality draws the attention of the scientific community. New Delhi has experienced several pollution episodes during the post-monsoon period. During these pollution episodes, the concentration of PM2.5 had reached severe levels [18], which created massive unease for the residents. Therefore, this chapter explains the reason behind such pollution episodes, using the 2016 Diwali pollution episode as a case study.

2. Air Quality of Delhi

The National Capital Territory Region (NCT) Delhi is home to around 46 million people. The rapid anthropogenic emission has become a constant threat to the residents of Delhi [19]–[21]. PM2.5 emerged as a dominant pollutant in more than 75% of the days in a year [22]. Several studies have reported that the PM2.5 concentration over Delhi maintains higher values throughout the year, exceeding the National Air Quality Standards (NAAQS) [22]–[25]. A recent study has shown that the PM2.5 level exceeds ~85% days in a year, increasing to 95% during winter. Even during monsoon, 68% of days are above the NAAQS level [22]. Average PM2.5 and PM10 range from 123±87 µg/m^3 and 208±137 µg/m^3, respectively, over Delhi (2008-2011), which is much higher than the NAAQS slandered value[19]. The continuous increase in PM2.5 concentration [18] indicates the increasing potential for deteriorating air quality in this region. Another study reported that the on-road concentration of PM2.5 exceeded the level in all modes of transport [26]. The wintertime PM2.5 exceeds the NAAQS standard more than eight times in a year [27]. According to a study conducted using the GAINS model, the air quality will continue to degrade even up to 2030 in the present emission control scenario [28]. Considering this alarming situation, the Delhi government had imposed an odd-even rule for the vehicles where only 50% of the total vehicles were permitted on a given day in the NCT area. The policy demonstrated mixed results. However, a moderate improvement in the winter air quality was observed [29]. Several factors like the topography, climate, local and festive emissions, and long-range transport can be attributed to such high loading over the

NCT area[22], [30]. However, recent studies have reported that agricultural crop residue burning has emerged as a significant factor in the degradation of Delhi air quality[18], [31].

3. Diwali and Post-Monsoon Pollution over Delhi

Diwali, or the festival of lights, is one of the most celebrated festivals throughout the Indian subcontinent. A large number of firecrackers are burnt during the festival, emitting pollutants like sulphur dioxide, carbon dioxide, carbon monoxide, suspended particles, aluminium, manganese cadmium, etc. Several studies reported a drastic rise in air pollution throughout the country during Diwali [32]–[38]. A study showed that PM10 concentration could increase more than 35 times during Diwali events compared to the regular days [36]. The Diwali emissions critically impact the air quality level of Delhi, and all pollutants exceed the standard during this period [39]–[42].

Several studies have also reported a high post-monsoonal PM2.5 value over Delhi [19], [43]–[45]. It is important to note that the post-monsoonal aerosol concentration continuously increases over Delhi [18], among which the amount of fine mode particles is much higher (~89%) as compared to the coarse mode particles [46], which reveals that the background concentration of the pollutants is elevated during the post-monsoon period. Therefore, even a minor contribution from the events like Diwali can create severe pollution episodes. Thus, multiple post-monsoonal pollution episodes were observed over Delhi, especially during Diwali. However, recent studies indicated the long-range transport of pollutants due to crop residue burning as a potential cause for these episodes [18], [31].

4. Agricultural Crop Residue Burning in NW India

The 'breadbasket' of India, Punjab and Haryana is located in the northwestern part of the NCT. Two different crop growing periods, summer (harvesting time between October and November) and winter (harvesting time between April and May) can be observed over this region [47]. The farmers of these areas have tended to use mechanized harvesters for the past few decades. Currently, the combined harvester is utilized to harvest more than 75% of the rice crop [47]. However, the process leaves a large amount of crop residue. The farmers burn the residues to prepare the field for the upcoming cropping season [48]. A recent study reported that 62% of biomass burning is contributed by the rise and paddy stocks [49]. The crop residue burning produces a large amount of pollution in the form of particulate matter, multiple greenhouse gases ((CO_2, N_2O, CH_4) and air pollutants (CO, NH_3, NOx, SO_2, NMHC, volatile organic compounds) [50]. It was estimated that 8.57 Mt of CO, 141.15 Mt of CO_2, 0.037 Mt of SOx, 0.23 Mt of NOx, 0.12 Mt of NH_3 and 1.46 Mt NMVOC, 0.65 Mt of NMHC, 1.21 Mt of particulate matter was emitted due to crop burning in the year 2008-09 [51]. These pollutants spread throughout the IGP using weak north-westerly surface wind [52]. 80% of the north-westerly surface flow crosses the crop residue burning areas before incoming in Delhi [53]. A recent study has reported that during the peak residue burning period, the PM2.5 concentration of the Delhi NCT can reach the level of 500 $\mu g/m^3$ [54]. A study based on the dual carbon isotope fingerprint method reveals that crop residue burning can subsidize more than 42% of the post monsoonal pollution of Delhi [31]. A significant rise in the post monsoonal pollution level over Delhi has been observed during the residue burning period from 2012 to 2016 [55]. In both the cropping seasons, PM10 and PM2.5 increases 2-3 times to the NAAQS standards due to residue burning [56].

From the above discussion, t appears that local emissions like Diwali and long-range transport of pollutants due to the crop residue burning both can trigger a post monsoonal pollution episode over Delhi. However, it is crucial to investigate the major contributor to such pollution episodes. Therefore, in the next section, a particular pollution episode, popularly known as the 2016 Diwali pollution episode, is explored based on two different studies.

5. The 2016 Diwali Pollution Episode

The 2016 Diwali pollution episode is an excellent example of the post monsoonal pollution episode over Delhi. The event was initiated after Diwali 2016 (30^{th} October 2016) and lasted more than a week. The PM2.5 concentration over Delhi exceeds 800 $\mu g/m^3$ during that period [18]. Due to such high pollution levels, the judiciary banned selling firecrackers during the Diwali period within the NCT region. However, during that same period, a vast crop residue burning event was taking place in the NW part of the Delhi region (Punjab and Haryana).

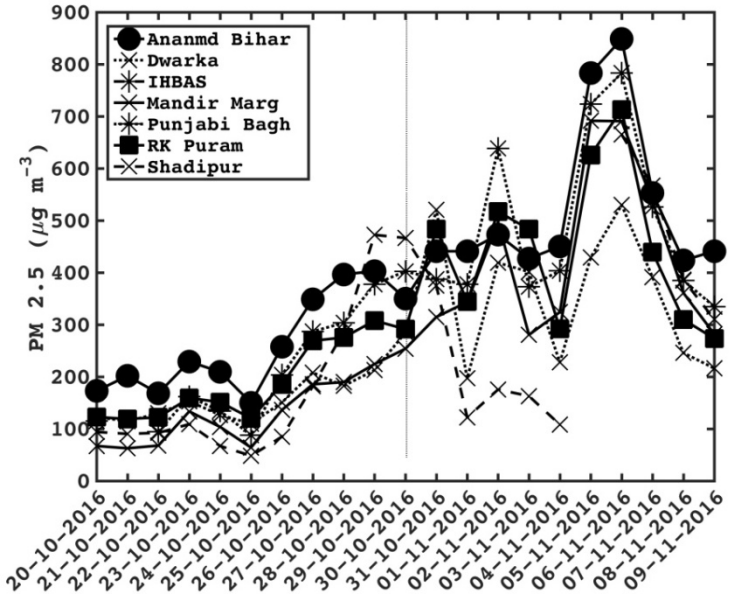


Figure 1.PM 2.5 variability over different stations in Delhi before and after Diwali during 2016. The thin grey line shows the date of Diwali [18]

The background PM2.5 concentration was more than 100 μg/m^3 even before Diwali (Figure 1). However, starting from 26th, an increase is observed to ~ 250 μg/m^3 on the next day of Diwali (30th October 2016). This increasing trend continued unabated until 2nd November, when there was a respite for a couple of days. It has to be noted that the PM2.5 level exceeds the highest value on 5th November, which is 5-6 days after the Diwali events. Therefore, it is challenging to ascertain whether the local emissions due to Diwali is the major contributor to the episode.

5.1. Climatological Features

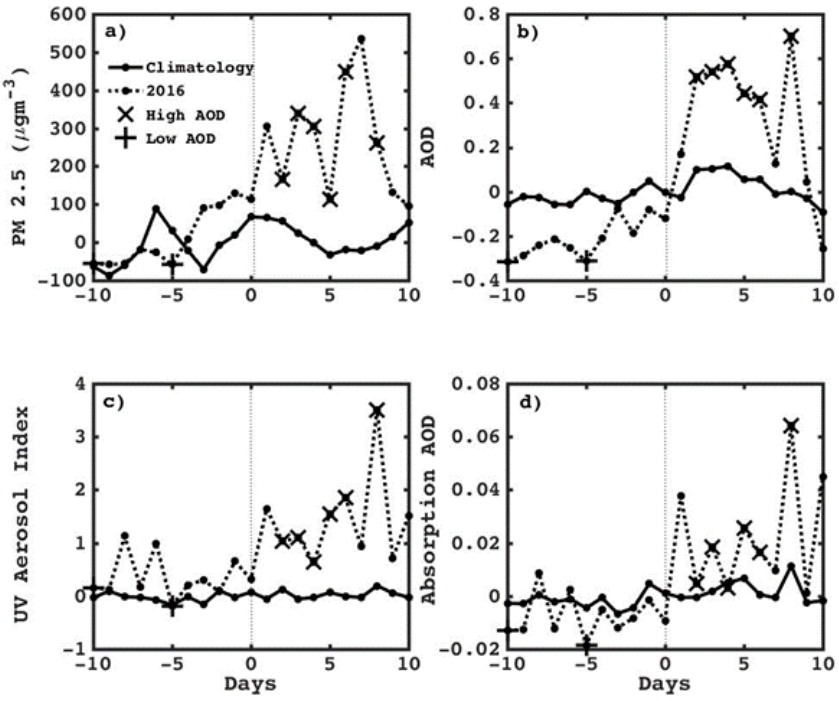


Figure 2. Intercomparison between the climatological anomaly and 2016 anomaly of a) PM2.5 b) Aerosol Optical Depth (AOD), c) UV Aerosol Index and d) Absorption AOD over Delhi [18]

The climatological features of PM2.5 (Figure 2), along with several other aerosol optical parameters like the Aerosol Optical Depth (AOD), UV Aerosol Index (UVAI) and Absorption AOD (AAOD) ascertain the uniqueness of the year 2016. In figure 2, the central vertical line denotes the Diwali days, while the thick and dotted line describes the climatology and year 2016, respectively. All the parameters depict higher values during the 2016 Diwali episode as compared to the climatological values. The before Diwali values are less than the after Diwali values. A past study has

shown that the impact of Diwali emissions subsides in a couple of days after Diwali [57]. However, all the parameters show higher values 5-6 days after the Diwali event. The specific rise of UVAI and AAOD signifies absorbing aerosols that can be generated due to crop residue burning. Therefore, the climatological features indicate the presence of long-range transported aerosols during the specific episode.

5.2. Long-Range Transport of Aerosols

A study based on the Concentration Weighted Trajectory (CWT) method demonstrates the role of long-range transported aerosols in the 2016 Diwali episode (Figure 3). By analyzing the 17 years of satellite-derived AOD and model-generated wind trajectory data, it was found that the NW part of the NCT region is the source of high aerosol loading over Delhi [18].

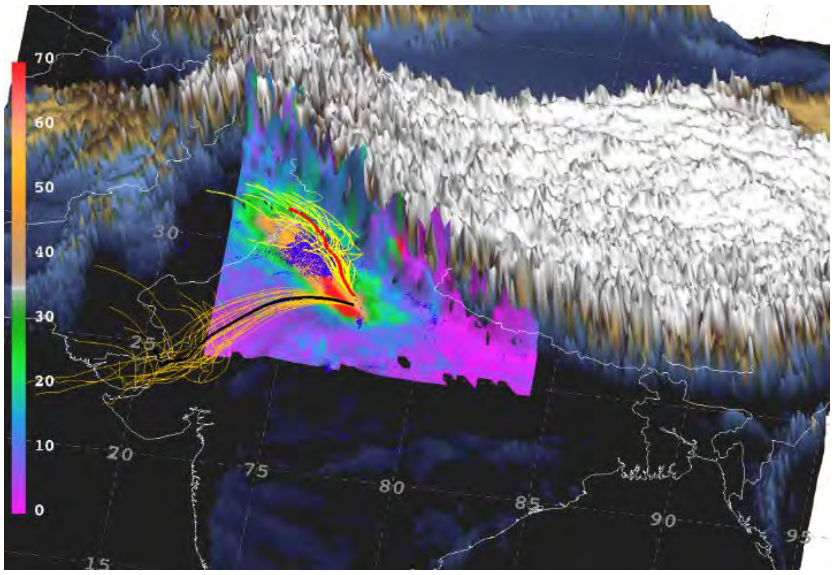


Figure 3. Role of long-range transport on PM2.5 concentration over Delhi. The colour bar represents the PM2.5 levels observed in Delhi. The black trajectory corresponds to the mean trajectory for low loading conditions, whereas the red trajectory corresponds to high loading conditions. The blue dots represent biomass burning during 2016. [18]

It appears that the high loading wind trajectory intercepts the biomass burning areas before reaching Delhi, which transports the fine mode absorbing aerosols to the city, enhancing the PM2.5 concentration during the study period. In that context, it has to be mentioned that several other studies have also reported that the concentration of PM2.5 was found to be higher during October-November, corresponding to the biomass burning over IGP [58]–[62]. A recent study showed that further delay in the burning period could deteriorate the air quality of Delhi by 4.4% [63].

5.3. Chemistry-Climate Modelling Evidence

A study based on a fully coupled online chemistry-climate model WRF-Chem (Weather Research and Forecast model coupled with Chemistry) explored the major contributor to the 2016 pollution episode [64]. Four experiments were designed for the entire study period (Table 1).

Table 1. List of simulations performed [64]

Case	Description
CTRL	All emissions + Anthropogenic and biomass burning
CTRL-BB	All emissions - biomass burning
CTRL-woD	No anthropogenic emissions from Delhi for the entire study period
CTRL+DD	Anthropogenic emissions were doubled for the Diwali day over Delhi

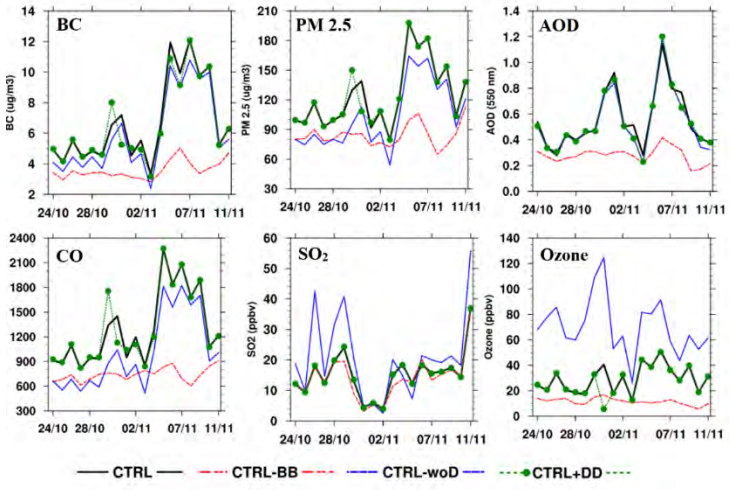


Figure 4. Time series of different pollutants during 2016 Diwali period (24/10/2016-07/11/2016) [64]

It can be observed that even without the Delhi emissions, the PM2.5 concentration matches the observational pattern, while the impact of Diwali emissions fades after 48 hours (Figure 4). However, the PM2.5 values remain low throughout the study period without biomass burning. Another study based on WRF-Chem has quantified that crop residue burning reduces the air quality of Delhi by 58% [65]. The crop residue burning contributes ~20% of the PM2.5 concentration over Delhi. However, during the pollution episodes, the contribution rises to 50-75% [66]. These results indicate that the crop residue burning contributed significantly to the post monsoonal pollution episodes over Delhi.

In that context, it has to be mentioned that local meteorology also plays a vital role to create pollution episodes. Slow wind and low boundary layer height trap the incoming pollutants over Delhi [64], [65]. Therefore, it can be stated that long-range transported aerosols in a favourable meteorological condition can create massive pollution episodes over Delhi. A recent study shows that the increase in the spatiotemporal extent of the pollutants due to residue burning decline the air quality and could pose a serious threat to human health [67], [68].

6. Summary and Conclusion

The megacity of New Delhi has encountered several post monsoonal pollution episodes in the recent past. The elevated pollution level creates massive unease for the local inhabitants. Studies have confirmed that the long-range transported aerosols play a crucial role. These aerosols are generated due to the crop residue burning in the northwest states like Punjab and Haryana. The impact of local emissions from the Diwali event does not last long, for more than 48 hours. The post monsoonal pollution episodes over Delhi provide a unique scenario to introspect where the rural pollution impacts the air quality of a megacity. Therefore, to encounter such pollution episodes, the air quality measurements should be expanded on a large spatial scale along with the local measures.

Acknowledgement

The authors like to acknowledge Dr. Bhupesh Adhikary for providing valuable information and guidance regarding the present article.

References

[1] J. Fenger, "Air pollution in the last 50 years - From local to global," *Atmos. Environ.*, vol. 43, no. 1, pp. 13–22, 2009.
[2] H. Akimoto, "Global Air Quality and Pollution," *Science (80-.).*, vol. 302, no. 5651, pp. 1716–1719, 2003.
[3] A. J. Cohen *et al.*, "The global burden of disease due to outdoor air pollution," *J. Toxicol. Environ. Heal. - Part A*, vol. 68, no. 13–14, pp. 1301–1307, 2005.
[4] B. R. Gurjar, K. Ravindra, and A. S. Nagpure, "Air pollution trends over Indian megacities and their local-to-global implications," *Atmos. Environ.*, vol. 142, pp. 475–495, 2016.
[5] S. K. Pandey, V. Vinoj, K. Landu, and S. S. Babu, "Declining pre-monsoon dust loading over South Asia: Signature of a changing regional climate," *Sci. Rep.*, vol. 7, no. 1, pp. 1–10, 2017.
[6] M. Bollasina, S. Nigam, and K. M. Lau, "Absorbing aerosols and summer monsoon evolution over South Asia: An observational portrayal," *J. Clim.*, vol. 21, no. 13, pp. 3221–3239, 2008.
[7] V. Ramanathan and M. V. Ramana, "Persistent, widespread, and strongly absorbing haze over the Himalayan foothills and the Indo-Gangetic Plains," *Pure Appl. Geophys.*, vol. 162, no. 8–9, pp. 1609–1626, 2005.
[8] S. N. Tripathi, A. Pattnaik, and S. Dey, "Aerosol indirect effect over Indo-Gangetic plain," *Atmos. Environ.*, vol. 41, no. 33, pp. 7037–7047, 2007.
[9] C. A. Pope III, R. T. Burnett, M. J. Thun, E. E. Calle, D. Krewski, and G. D. Thurston, "Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution," *J. Am. Med. Assoc.*, vol. 287, no. 9, pp. 1132–1141, 2002.
[10] M. Brauer *et al.*, "Exposure assessment for estimation of the global burden of disease attributable to outdoor air pollution," *Env. Sci Technol*, vol. 46, no. 2, pp. 652–60, 2012.
[11] D. W. Dockery, "Health Effects of Particulate Air Pollution," *Ann. Epidemiol.*, vol. 19, no. 4, pp. 1–49, 2014.

[12] R. D. Brook *et al.*, "Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the american heart association," *Circulation*, vol. 121, no. 21, pp. 2331–2378, 2010.
[13] G. Hoek *et al.*, "Long-term air pollution exposure and cardio-respiratory mortality: A review," *Environ. Heal. A Glob. Access Sci. Source*, vol. 12, no. 1, 2013.
[14] S. Chowdhury and S. Dey, "Cause-specific premature death from ambient PM2.5exposure in India: Estimate adjusted for baseline mortality," *Environ. Int.*, vol. 91, pp. 283–290, 2016.
[15] S. Dey and S. N. Tripathi, "Aerosol direct radiative effects over Kanpur in the Indo-Gangetic basin, northern India: Long-term (2001-2005) observations and implications to regional climate," *J. Geophys. Res. Atmos.*, vol. 113, no. 4, pp. 1–20, 2008.
[16] D. G. Kaskaoutis, R. P. Singh, R. Gautam, M. Sharma, P. G. Kosmopoulos, and S. N. Tripathi, "Variability and trends of aerosol properties over Kanpur, northern India using AERONET data (200110)," *Environ. Res. Lett.*, vol. 7, no. 2, 2012.
[17] World Health Organization, "Ambient Air Pollution: A global assessment of exposure and burden of disease," *World Heal. Organ.*, pp. 1–131, 2016.
[18] T. Mukherjee *et al.*, "Increasing potential for air pollution over megacity New Delhi: A study based on 2016 diwali episode," *Aerosol Air Qual. Res.*, vol. 18, no. 9, 2018.
[19] S. K. Guttikunda and R. Goel, "Health impacts of particulate pollution in a megacity-Delhi, India," *Environ. Dev.*, vol. 6, no. 1, pp. 8–20, Apr. 2013.
[20] K. J. Maji, A. K. Dikshit, and A. Deshpande, "Disability-adjusted life years and economic cost assessment of the health effects related to PM2.5 and PM10 pollution in Mumbai and Delhi, in India from 1991 to 2015," *Environ. Sci. Pollut. Res.*, vol. 24, no. 5, pp. 4709–4730, 2017.
[21] J. S. Pandey, R. Kumar, and S. Devotta, "Health risks of NO2, SPM and SO2 in Delhi (India)," *Atmos. Environ.*, vol. 39, no. 36, pp. 6868–6874, 2005.
[22] S. K. Sahu and H. Kota, "Significance of PM 2.5 Air Quality at the Indian Capital," *Aerosol Air Qual. Res.*, vol. 17, pp. 588–597, 2017.
[23] A. P. Mitra and C. Sharma, "Indian aerosols: Present status," *Chemosphere*, vol. 49, no. 9, pp. 1175–1190, Dec. 2002.
[24] S. Tiwari, A. K. Srivastava, D. S. Bisht, P. Parmita, M. K. Srivastava, and S. D. Attri, "Diurnal and seasonal variations of black carbon and PM2.5over New Delhi, India: Influence of meteorology," *Atmos. Res.*, vol. 125–126, pp. 50–62, 2013.
[25] A. Srivastava and V. K. Jain, "Seasonal trends in coarse and fine particle sources in Delhi by the chemical mass balance receptor model," *J. Hazard. Mater.*, vol. 144, no. 1–2, pp. 283–291, Jun. 2007.
[26] J. S. Apte *et al.*, "Concentrations of fine, ultrafine, and black carbon particles in auto-rickshaws in New Delhi, India," *Atmos. Environ.*, vol. 45, no. 26, pp. 4470–4480, Aug. 2011.
[27] M. Sharma, "Govt. Of Nct Of Delhi vs Monika Sharma on 26 May, 2016," no. 2, pp. 1–7, 2016.
[28] H. H. Dholakia, P. Purohit, S. Rao, and A. Garg, "Impact of current policies on future air quality and health outcomes in Delhi, India," *Atmos. Environ.*, vol. 75, pp. 241–248, 2013.
[29] S. K. Sharma *et al.*, "Study on Ambient Air Quality of Megacity Delhi, India During Odd–Even Strategy," *Mapan - J. Metrol. Soc. India*, vol. 32, no. 2, pp. 155–165, 2017.
[30] S. Ghosh, J. Biswas, S. Guttikunda, S. Roychowdhury, and M. Nayak, "An investigation of potential regional and local source regions affecting fine particulate matter concentrations in Delhi, India," *J. Air Waste Manage. Assoc.*, vol. 65, no. 2, pp. 218–231, Feb. 2015.
[31] S. Bikkina *et al.*, "Air quality in megacity Delhi affected by countryside biomass burning," *Nat. Sustain.*, vol. 2, no. 3, pp. 200–205, 2019.
[32] K. Ravindra *et al.*, "Short-term variation in air quality associated with firework events: A case study," *J. Environ. Monit.*, vol. 5, no. 2, pp. 260–264, Mar. 2003.
[33] S. Bhatnagar and S. Dadhich, "Assessment of the Impact of Fireworks on Ambient Air Quality," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 3, no. IV, pp. 605–609, 2015.
[34] A. Chatterjee, C. Sarkar, A. Adak, U. Mukherjee, S. K. Ghosh, and S. Raha, "Ambient Air Quality during Diwali Festival over Kolkata – A Mega-City in India," *Aerosol Air Qual. Res.*, vol. 13, no. 13, pp. 1133–1144, 2013.
[35] C. Verma and D. K. Deshmukh, *Recent research in science and technology : an international refereed journal for all aspects of science research.*, vol. 6, no. 1. Recent Research in Science and Technology, 2009.
[36] U. P. Nasir and D. Brahmaiah, "Impact of fireworks on ambient air quality: a case study," *Int. J. Environ. Sci. Technol.*, vol. 12, no. 4, pp. 1379–1386, Apr. 2015.
[37] V. S. Chauhan, B. Singh, S. Ganesh, J. Zaidi, and J. C. Bose, "Status of Air Pollution During Festival of Lights (Diwali) in Jhansi, Bundelkhand Region, India," *Asian J. Sci. Technol.*, vol. 5, no. 3, pp. 187–191, 2014.
[38] S. Nigam, N. Kumar, N. K. Mandal, B. Padma, and S. Rao, "Real Time Ambient Air Quality Status during Diwali Festival in Central, India," *J. Geosci. Environ. Prot.*, vol. 04, no. 01, pp. 162–172, 2016.
[39] N. D. Ganguly, "Surface ozone pollution during the festival of Diwali, New Delhi, India," *J. Earth Sci. India*, vol. 2, pp. 224–229, 2009.
[40] A. Chauhan and R. P. Singh, "POOR AIR QUALITY AND DENSE HAZE / SMOG DURING 2016 IN THE INDO-GANGETIC PLAINS ASSOCIATED WITH THE CROP RESIDUE BURNING AND DIWALI FESTIVAL Vidya College of Engineering , Baghpat Road , Meerut – 250002 , India School of Life and Environmental Sciences , S," pp. 6048–6051, 2017.
[41] N. Parkhi *et al.*, "Large inter annual variation in air quality during the annual festival 'Diwali' in an Indian megacity," *J. Environ. Sci. (China)*, vol. 43, pp. 265–272, 2016.
[42] C. Perrino, S. Tiwari, M. Catrambone, S. D. Torre, E. Rantica, and S. Canepari, "Chemical characterization of atmospheric PM in Delhi, India, during different periods of the year including Diwali festival," *Atmos. Pollut. Res.*, vol. 2, no. 4, pp. 418–427, Oct. 2011.
[43] S. K. Guttikunda and G. Calori, "A GIS based emissions inventory at 1 km × 1 km spatial resolution for air pollution analysis in Delhi, India," *Atmos. Environ.*, vol. 67, pp. 101–111, 2013.
[44] S. Tiwari, A. K. Srivastava, D. S. Bisht, P. Parmita, M. K. Srivastava, and S. D. Attri, "Diurnal and seasonal variations of black carbon and PM2.5 over New Delhi, India: Influence of meteorology," *Atmos. Res.*, vol. 125–126, pp. 50–62, 2013.
[45] S. Tiwari, D. M. Chate, P. Pragya, K. Ali, and D. S. F. Bisht, "Variations in mass of the PM 10,PM 2.5 and PM 1 during the monsoon and the winter at New Delhi," *Aerosol Air Qual. Res.*, vol. 12, no. 1, pp. 20–29, 2012.
[46] S. Tiwari *et al.*, "Aerosol optical properties and their relationship with meteorological parameters during wintertime in Delhi , India," *Atmos. Res.*, vol. 153, pp. 465–479, 2015.
[47] K. P. Vadrevu, E. Ellicott, K. V. S. Badarinath, and E. Vermote, "MODIS derived fire characteristics and aerosol optical depth variations during the agricultural residue burning season, north India," *Environ. Pollut.*, vol. 159, no. 6, pp. 1560–1569, 2011.
[48] D. G. Kaskaoutis *et al.*, "Effects of crop residue burning on aerosol properties, plume characteristics, and long-range transport over northern India," *J. Geophys. Res. Atmos.*, no. July, pp. 5424–5444, 2014.
[49] R. Singh, D. B. Yadav, N. Ravisankar, A. Yadav, and H. Singh, "Crop residue management in rice–wheat cropping system for resource conservation and environmental protection in north-western India," *Environ. Dev. Sustain.*, vol. 22, no. 5, pp. 3871–3896, Jun. 2020.
[50] R. Mathur and V. K. Srivastava, "Crop Residue Burning: Effects on Environment," *Energy, Environ. Sustain.*, pp. 127–140, 2019.
[51] N. Jain, A. Bhatia, and H. Pathak, "Emission of air pollutants from crop residue burning in India," *Aerosol Air Qual. Res.*, vol. 14, no. 1, pp. 422–430, 2014.
[52] R. P. Singh and D. G. Kaskaoutis, "Crop residue burning: A threat to South Asian air quality," *Eos (Washington. DC).*, vol. 95, no. 37, pp. 333–334, 2014.
[53] H. Jethva, D. Chand, O. Torres, P. Gupta, A. Lyapustin, and F. Patadia, "Agricultural Burning and Air Quality over Northern India: A Synergistic Analysis using NASA's A-train Satellite Data and Ground Measurements," *Aerosol Air Qual. Res.*, vol. 18, no. 7, pp. 1756–1773, 2018.
[54] S. Chowdhury, S. Dey, L. Di Girolamo, K. R. Smith, A. Pillarisetti, and A. Lyapustin, "Tracking ambient PM2.5 build-up in Delhi national capital region during the dry season over 15 years using a high-resolution (1 km) satellite aerosol dataset," *Atmos. Environ.*, vol. 204, pp. 142–

150, May 2019.
[55] D. H. Cusworth *et al.*, "Quantifying the influence of agricultural fires in northwest India on urban air pollution in Delhi, India," *Environ. Res. Lett.*, vol. 13, no. 4, p. 044018, Apr. 2018.
[56] P. Saxena *et al.*, "Impact of crop residue burning in Haryana on the air quality of Delhi, India," *Heliyon*, vol. 7, no. 5, p. e06973, 2021.
[57] G. Beig, "System Of Air Quality And Weather Forecasting And Research (SAFAR-India) Indian Institute of Tropical Meteorology, Pune Earth System Science Organization (MoES)," pp. 1–4, 2016.
[58] C. D. Bray, W. H. Battye, and V. P. Aneja, "The role of biomass burning agricultural emissions in the Indo-Gangetic Plains on the air quality in New Delhi, India," *Atmos. Environ.*, vol. 218, p. 116983, Dec. 2019.
[59] S. Kumari, N. Verma, A. Lakhani, and K. M. Kumari, "Severe haze events in the Indo-Gangetic Plain during post-monsoon: Synergetic effect of synoptic meteorology and crop residue burning emission," *Sci. Total Environ.*, vol. 768, p. 145479, May 2021.
[60] K. Ravindra *et al.*, "Real-time monitoring of air pollutants in seven cities of North India during crop residue burning and their relationship with meteorology and transboundary movement of air," *Sci. Total Environ.*, vol. 690, pp. 717–729, Nov. 2019.
[61] K. Kumar, S. Singh, and S. Chandra, "Episodic Measurements of PM2.5 during Crop Residue Burning and Diwali Periods at Delhi," no. October, pp. 40–50, 2020.
[62] P. K. Nagar, M. Sharma, and D. Das, "A new method for trend analyses in PM10 and impact of crop residue burning in Delhi, Kanpur and Jaipur, India," *Urban Clim.*, vol. 27, no. September 2018, pp. 193–203, 2019.
[63] H. Sembhi *et al.*, "Post-monsoon air quality degradation across Northern India: Assessing the impact of policy-related shifts in timing and amount of crop residue burnt," *Environ. Res. Lett.*, vol. 15, no. 10, 2020.
[64] T. Mukherjee, V. Vinoj, S. K. Midya, S. P. Puppala, and B. Adhikary, "Numerical simulations of different sectoral contributions to post monsoon pollution over Delhi," *Heliyon*, vol. 6, no. 3, 2020.
[65] G. Beig *et al.*, "Objective evaluation of stubble emission of North India and quantifying its impact on air quality of Delhi," *Sci. Total Environ.*, vol. 709, p. 136126, Mar. 2020.
[66] S. H. Kulkarni *et al.*, "How Much Does Large-Scale Crop Residue Burning Affect the Air Quality in Delhi?," *Environ. Sci. Technol.*, vol. 54, no. 8, pp. 4790–4799, Apr. 2020.
[67] S. Sarkar, R. P. Singh, and A. Chauhan, "Crop Residue Burning in Northern India: Increasing Threat to Greater India," *J. Geophys. Res. Atmos.*, vol. 123, no. 13, pp. 6920–6934, 2018.
[68] M. Agarwala and A. Chandel, "LETTER Temporal role of crop residue burning (CRB) in Delhi's air pollution," *Environ. Res. Lett.*, vol. 15, no. 11, 2020.

Impact of Environmental Factors on COVID-19 Transmission: An Overview

Souvik Manik[1], Sabyasachi Pal[1,*], Manoj Mandal[1]
[1]Midnapore City College, Bhadutala, Paschim Medinipur, 721129, India
*Corresponding author: sabya.pal@gmail.com

Abstract

The Corona Virus Disease 2019 (COVID-19) broke out in Wuhan province of China in November 2019 and was immediately declared as a worldwide pandemic by the World Health Organization due to its high contagiousness. Multiple research has confirmed that climate conditions have a significant influence on virus transmission. Different studies have found that several air pollutants may have an impact on virus transmission. We have summarised the impact of different air pollutants (PM2.5, PM10, O_3, NO_2, SO_2, and CO) on COVID-19 cases as well as the mortality of the disease. Some of these studies focused solely on environmental variables, while others also looked at non-meteorological or social factors. The review aims at summarising and exploring the impact of environmental factors like temperature, humidity, wind speed, UV index, pressure, different air pollutants, etc. on the COVID-19 spread. There are several epidemiological models (SIR, SIRD, SEIR, SEIRD) widely used to estimate different driving parameters related to an epidemic. We have summarised the impact of different driving parameters on the virus transmission based on several world wide studies.

Keywords:*Coronavirus, COVID-19 transmission, epidemiological models, effective reproduction number, environmental factors*

1 Introduction

The Corona Virus Disease (COVID-19) appeared in Wuhan province of China in November 2019 [1], [2] and affected millions of people globally in a short span of time [3], [4]. The World Health Organization (WHO) immediately declared Corona Virus Disease 2019 (COVID-19) as a global pandemic on March 11, 2020 [5]. The virus evolved significantly with time in the form of several mutant variants of Coronavirus that are dominated throughout the world[1].

The transmission of COVID-19 (SARS-CoV-2) depends on a large number of factors, including environmental factors such as temperature, relative humidity, stability on fomites, precipitation, wind speed, radiation, and also social factors such as over-crowded public places, gatherings, meetings, rallies, festivals, movements of individuals, etc. Similar to the transmission dynamics of various other respiratory viruses, such as influenza [6, 7, 8], COVID-19 can be transmitted through skin contact with virus-contaminated surfaces and also by inhalation of the virus carried in respirable particles.

Meteorological parameters play an important role in influencing infectious diseases such as severe acute respiratory syndrome (SARS) and influenza. A study on Wuhan, China aims to explore the association between Corona Virus Disease 2019 (COVID-19) deaths and weather parameters. A positive association with COVID-19 daily death counts was observed for diurnal temperature range (r = 0.44), but a negative association was observed for relative humidity (r = –0.32) [9].

Several studies were conducted on India to understand the impact of environmental factors on COVID-19 transmission. The daily spread of COVID-19 cases for Indian cities was shown to be fairly associated with temperature, and the higher temperature seems to be disrupting the lipid layer of coronavirus [10, 11]. The outcomes did not clearly reveal the importance of relative humidity in COVID-19 cases on a daily basis.

Earlier, the seasonality of different viruses was observed, like influenza. For COVID-19 and other coronaviruses like SARS, similar seasonal variability may be expected. A correlation was found between the observed spread of COVID-19 (SARS-CoV-2) with temperature and latitude. Along an observed area of 25–55° North latitude and within a climatological band of 4–12 °C, enhanced spread of the disease has been observed between December and February 2019–2020 [12]. In addition, a comparison of the seasonality of other viruses, such as seasonal influenza, reveals that the northern hemisphere's cool season supports disease propagation more than the northern hemisphere's warm season.

There were many possible non-meteorological or social factors, such as governmental interventions, lock-downs, social interactions, overcrowded gatherings, rallies, meetings, herd immunity, migration patterns, population density, personal hygiene, vaccinations, defence mechanisms, festivals, and cultural activity, that could influence the correlation analysis between meteorological factors and COVID-19 transmission. Assembly election took place in India during the pandemic situation. The effect of the election was reflected as a major contributor to the second wave in India. For election-bound states, there was a significant rise in effective contact rate and effective reproduction number during the election-bound period and immediately afterwards, compared to the pre-election period. The impact of pre-election activities, including meetings, political rallies, movements, and over-crowded gatherings, was clearly reflected in the change in effective reproduction number and effective contact rate [13]. States with single-phase elections are comparatively less affected than states

[1]https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/

where the election was conducted in multiple phases. From the first week of April 2021, the election commission imposed additional restrictions on large campaign rallies, meetings, and other political activities, which may help to slow down the effective contact rate and the effective reproduction number in all election-bound states [13].

The present review aims to collect different results on the impact of social and environmental factors on COVID-19 transmission. A summary of different epidemiological models is discussed in Section 2. The impact of different environmental factors on virus transmission is discussed in Section 3. Section 4 concludes the significant views of the study.

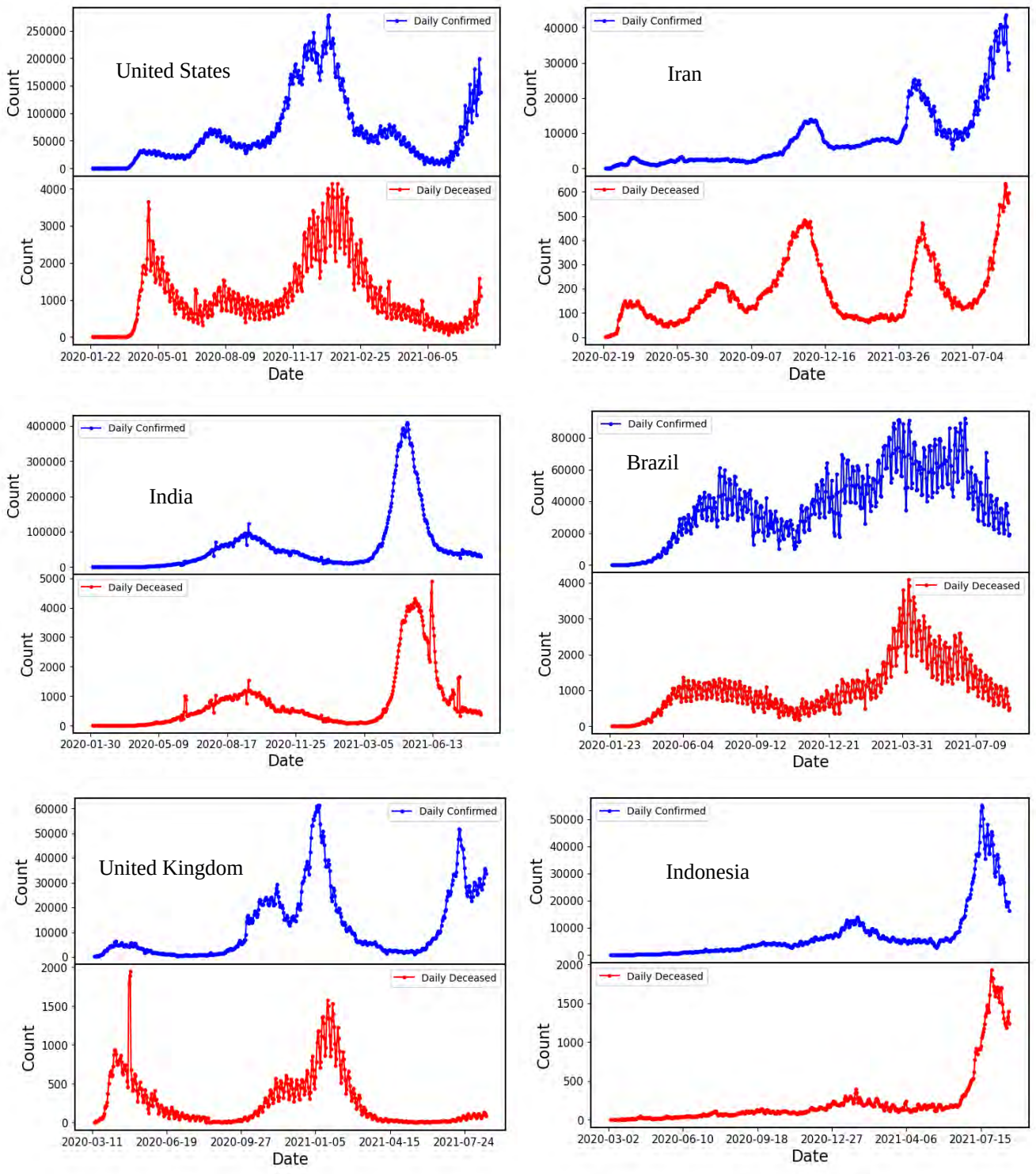


Figure 1: Evolution of daily confirmed and deceased individuals for different countries with 3 days rolling mean.

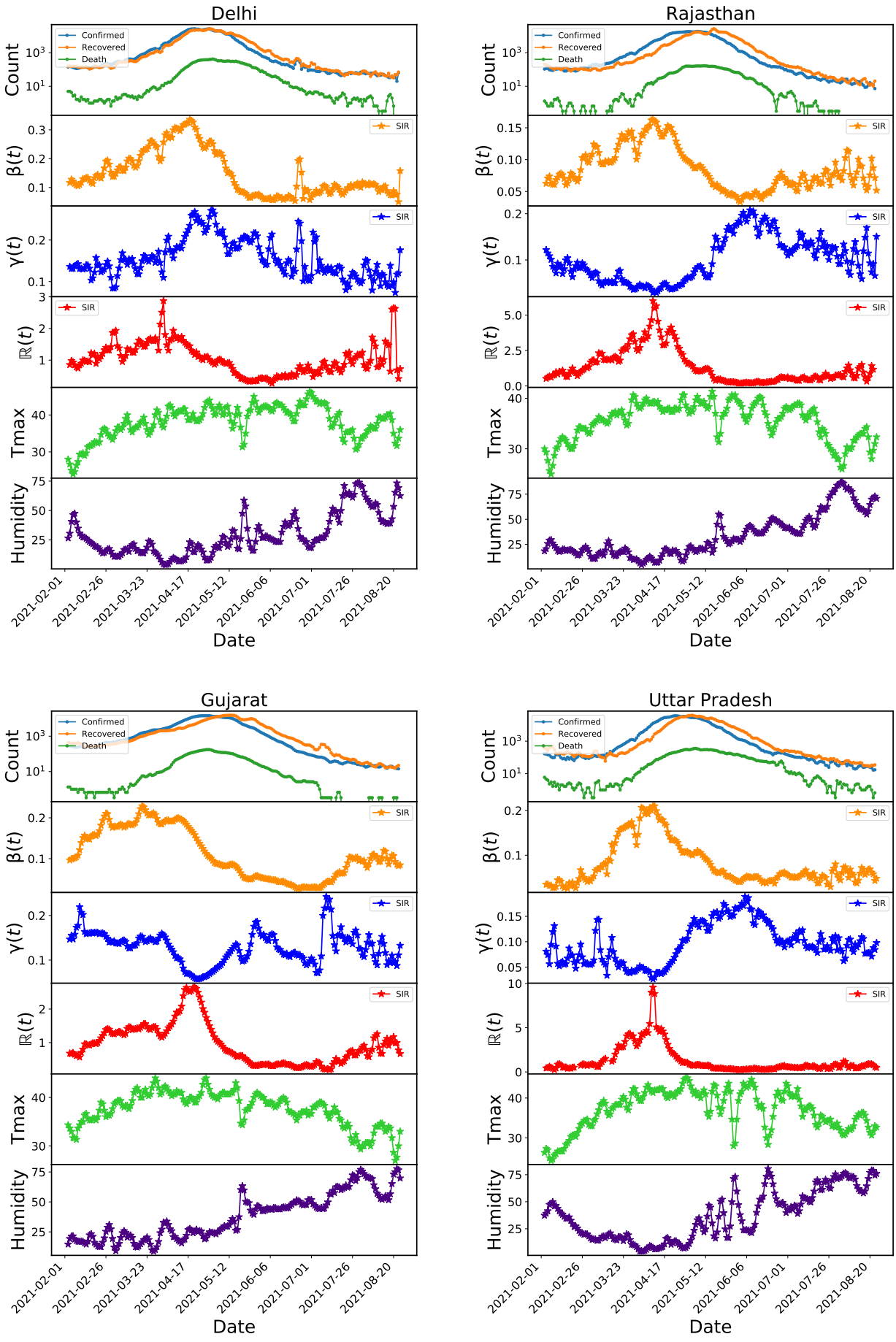


Figure 2: Variation of different driving parameters of the pandemic for different Indian states. Time-dependent parameters are introduced in section 2. $\gamma(t)$ is the recovery rate, $\beta(t)$ is the effective contact rate, and $\mathbb{R}(t)$ is the effective reproduction number.

2 A Brief Review of Different Epidemiological Models

There are several epidemiological models to explain the spread rate of a virus. One of the fundamental models is the SIR model, and two other modified SIR models are known as the SIRD and SEIR models. During the pandemic, several studies were conducted to model the virus spread rate and predict several driving parameters related to a pandemic. The effective contact rate and the effective reproduction number are the most important parameters to be studied to trace the scenario of a pandemic.

The most important part of these models is to estimate the basic reproduction number (R_0) which is the contagiousness of the diseases. R_0 determines the average number of people who can be affected by a single infected person over a course of time. $R_0 = 1$, indicates that the spread is stable, $R_0 > 1$ indicates that the spread is increasing, and $R_0 < 1$ indicates that the spread is expected to stop. We have studied the variation of effective reproduction number $\mathbb{R}(t)$, effective contact rate for India using different models [11, 13, 14].

The Susceptible-Infected-Recovered (SIR) model [15] is the simplest compartmental model which can explain the evolution of the epidemic at the population level. We used this model to study the different time-dependent parameters like contact rate (β), recovery rate (γ), and effective reproduction number ($\mathbb{R}$). The basic reproduction number R_0 can be written by

$$R_0 = \frac{\beta}{\gamma} \tag{1}$$

The effective reproduction number $\mathbb{R}(t)$ can be defined by

$$\mathbb{R}(t) = \frac{\beta_n}{\gamma_n} \tag{2}$$

The SIRD model [16] is the modified version of the SIR model. This model is also used to study the different time-dependent parameters like recovery rate (γ), contact rate (β), and effective reproduction number ($\mathbb{R}$). At any time t, $I(t)$ be the total number of infected individuals, $S(t)$ be the total number of susceptible individuals, $R(t)$ be the total number of recovered, $D(t)$ be the total number of deceased individuals from the epidemic. Here, an additional parameter, α is introduced, which is the mortality rate. The effective reproduction number $\mathbb{R}(t)$ can be given by

$$\mathbb{R}(t) = \frac{\beta_n}{\gamma_n + \alpha_n} \tag{3}$$

The Susceptible-Exposed-Infectious-Removed (SEIR) model [17] is the most studied one. The SEIR model is a variant of the SIR model. The parameter β is the product of the average number of contacts per person and per unit time by the probability of disease transmission in contact between susceptible and infectious individuals. γ is the recovery rate. The additional compartment E of the exposed individuals in the SEIR model makes the model more delicate. At any time t, $S(t)$ be the total number of susceptible individuals, $E(t)$ be the total number of exposed individuals, $I(t)$ be the total number of infected individuals, and $R(t)$ be the total number of removed (recovered or deceased till a given time) individuals from the epidemic.

$$\beta_n = \frac{1}{s_n i_n \sigma}[i_{n+2} + (\gamma_n + \sigma - 2)i_{n+1} + (\sigma - 1)(\gamma_n - 1)i_n] \tag{4}$$

The effective reproduction number $\mathbb{R}$(t) can be written as

$$\mathbb{R}(t) = \frac{\beta_n}{\gamma_n} \tag{5}$$

For detail derivations see [11, 13]. We mostly discuss about the time-dependent reproduction number and the effective contact rate. These are known to be the driving parameters of a pandemic.

3 Impact of Environmental Factors on Virus Transmission

It was widely believed during the early phase of the virus transmission that summer would cause the virus to reduce [18]. Experimental results of other corona viruses also indicated that the virus may be less effective with rising temperature and humidity [19]. There were several studies to investigate the dependency of transmission of COVID-19 on temperature and humidity [19, 20, 21, 22, 23, 24, 25] using early data from the COVID-19 pandemic, which show contradictory results. Using very early data of 100 Chinese cities, [20] found that high temperature and high humidity significantly reduce the transmission of COVID-19 while [21] concluded the opposite using initial data from four of the most affected places from China and five of the most affected places in Italy. A rigorous study with a large amount of data was necessary to effectively study the effect of temperature and humidity on the spread of coronavirus.

Fig. 1 represents the current scenario of virus transmission for different places of the globe. Fig. 1 shows the daily confirmed cases and daily deceased individuals for the United States, Iran, India, Brazil, United Kingdom, and Indonesia. These places are badly affected by COVID-19. The figure shows that most places are faced with multiple waves of diseases.

Fig. 2 shows the variation of different driving parameters of a pandemic for different Indian states from February 2021 to August 2021. The first row shows the variation of confirmed, recovered, and deceased individuals. The second row shows the variation of effective contact rate $\beta(t)$. The third row shows the variation of recovery rate $\gamma(t)$ and the fourth row represent the variation of effective reproduction number. The fifth and sixth rows show the variation of maximum temperature and humidity, respectively.

3.1 Impact of Temperature

There are plenty of studies that focused on the effect of temperature on virus transmission. The studies were conducted in different places of the globe at different times. There was no common conclusion on the correlation between the virus spread rate and temperature. A few studies concluded a positive correlation between the virus spread rate and temperature, which means that with an increase in temperature, the effective contact rate and effective reproduction number also increase. Most studies showed a negative correlation between COVID-19 spread rate and temperature, which implies that with an increase in temperature the value of effective contact rate and reproduction number decreases. A positive correlation between COVID-19 transmission and temperature was shown for several studies in Jakarta [26] and New York [27]. No association was found in countries such as Spain [28], Iran [29, 30], Nigeria[31] and in a worldwide study [32]. A negative correlation was found for multiple worldwide studies [23, 25, 33, 34, 35, 36, 37] including in India [11], California [38], Japan [39], Ghana [40], Spain [41, 42], Italy [43] and in China [44, 45, 46, 47]. A worldwide study of 166 countries (excluding China) found a significant negative correlation between temperature and COVID-19, where a temperature increase of $1^{\circ}C$ was associated with a reduction of 3.08% in cases [48]. The worldwide study including 100 countries up to 18 March 2020 in the temperature range –33.9 to 34.3 $^{\circ}C$, suggested a significant negative correlation between daily temperature and daily cases when temperatures increased above $-15^{\circ}C$, (r = –0.88, $p \leq 0.001$) [49]. A significant negative correlation was found [50] for both average temperature (AT) and diurnal temperature range in cities of China [46]. It is also reported that the decrease in the number of cases per day is 80% (95% confidence interval: 75–85%) and 90% (95% confidence interval: 86-95%), with an increase of $1^{\circ}C$ daily average temperature and 1% increase in the diurnal temperature range. Earlier, a significant negative correlation was reported for provinces in China, with daily COVID-19 transmission in the temperature range $-10^{\circ}C$ to $10^{\circ}C$ [46].

The impact of different environmental factors and conditions such as temperature, humidity, wind speed as well as food, water, sewage, air, insects, inanimate surfaces, and hands on COVID-19 transmission is being studied [51]. The results of studies on the stability of the SARS-CoV-2 at different levels showed that the resistance of this virus on smooth surfaces was higher than others. Increased temperature and sunlight can facilitate the destruction of SARS-COV-2 and its stability on surfaces. When the minimum ambient air temperature increases by 1 °C, the cumulative number of cases decreases by 0.86% [51].

According to a study in Africa, for every $1^{\circ}C$ increase in average daily temperature, the number of cases per day decreased by 13.53%, taking into account the delay in the incubation period [52]. In a study in India, the dependence of temperature varied across 11 states with COVID-19 transmission and four states including Madhya Pradesh (r = 1.43, p $\leq$ 0.05), Maharashtra (r = 2.76, p $\leq$ 0.05), Punjab (r = 1.49, p $\leq$0.05), and Tamil Nadu (r = –15.9, p $\leq$ 0.05) showed a significant correlation with average temperature; maximum temperature showed a significant association with COVID-19 transmission in two states Tamil Nadu (r = 0.43, p $\leq$ 0.05) and Maharashtra (r = –0.32, $p < 0.05$) and minimum temperature reported as significant in association with COVID-19 transmission in two states Gujarat (r = 0.21, $p < 0.05$) and Uttar Pradesh (r = 0.18, $p < 0.05$) [53].

We have looked at the evolution of different driving parameters of a pandemic with temperature for different Indian states. Fig. 3 and Fig. 4 show the variation of the effective reproduction number of COVID-19 with temperature, humidity, pressure, UV index, and precipitation for two different Indian states, Delhi and Uttar Pradesh. For Delhi, a negative correlation was observed with a lower value of the correlation coefficient. A positive correlation was observed for Uttar Pradesh and the correlation coefficient was $\sim 8\%$. There were several detailed studies on India where both positive and negative correlations were observed for different Indian states [11].

There are a few studies in which the result reflects both the positive and negative correlation, which is classified as mixed correlation [11, 54, 55]. It is difficult to conclude a general statement on the dependence of the spread rate of virus and temperature. There are so many other key factors mixed with the effect of temperature, like social distance, personal hygiene, lockdown effect, vaccines, etc., which are not identical for different studies over the globe. These factors may affect a lot on the real scenario of the virus spread.

3.2 Impact of Humidity

There are several studies where the impact of humidity on the spread rate of the virus was investigated, but the results are not conclusive. The studies from Africa, New York, USA, Jakarta, Indonesia, and one global study of 100 countries did not find a significant relationship between humidity and COVID-19 transmission [26, 27, 49, 52]. A mixed results was reported across the Indian region [53], where positive correlations were reported for Punjab (r = 0.584, $p < 0.05$) and Madhya Pradesh (r = 1.211, $p < 0.05$) and a negative correlation was found for Tamil Nadu (r = –6.79, $p < 0.05$). Earlier a negative correlation between daily cases with relative humidity was reported in provinces of China [45] and 1% increase in R_H, daily cases decreased 11–22% when the average temperature was in the range of $5.04^{\circ}C$ to $8.2^{\circ}C$. A significant association with humidity was found in a multivariate analysis in New South Wales, Australia, which assessed relative humidity between 9

a.m. and 3 p.m., where a 1% increase in humidity at 9 a.m. can increase the number of COVID 19 cases by 6.11% [56]. An inverse relationship between relative humidity and daily new cases is reported across the globe, where for every 1% increase in humidity, new daily cases were reduced by 0.85% (95% CI: 0.51–1.19%) [36]. A weak positive association is reported between average relative humidity and new cases in Saudi Arabia (r = 0.194, $p < 0.01$) [57]. In the United States also, a significant positive correlation between relative humidity and new cases is found (RR 0.07 95%CI: 0.05–0.09) [58]. In Spain, a negative correlation between the transmission rate and relative humidity was reported [59], where transmission decreases by 3% for every 1% increase in humidity when adjusting population density, age, and control of public transport.

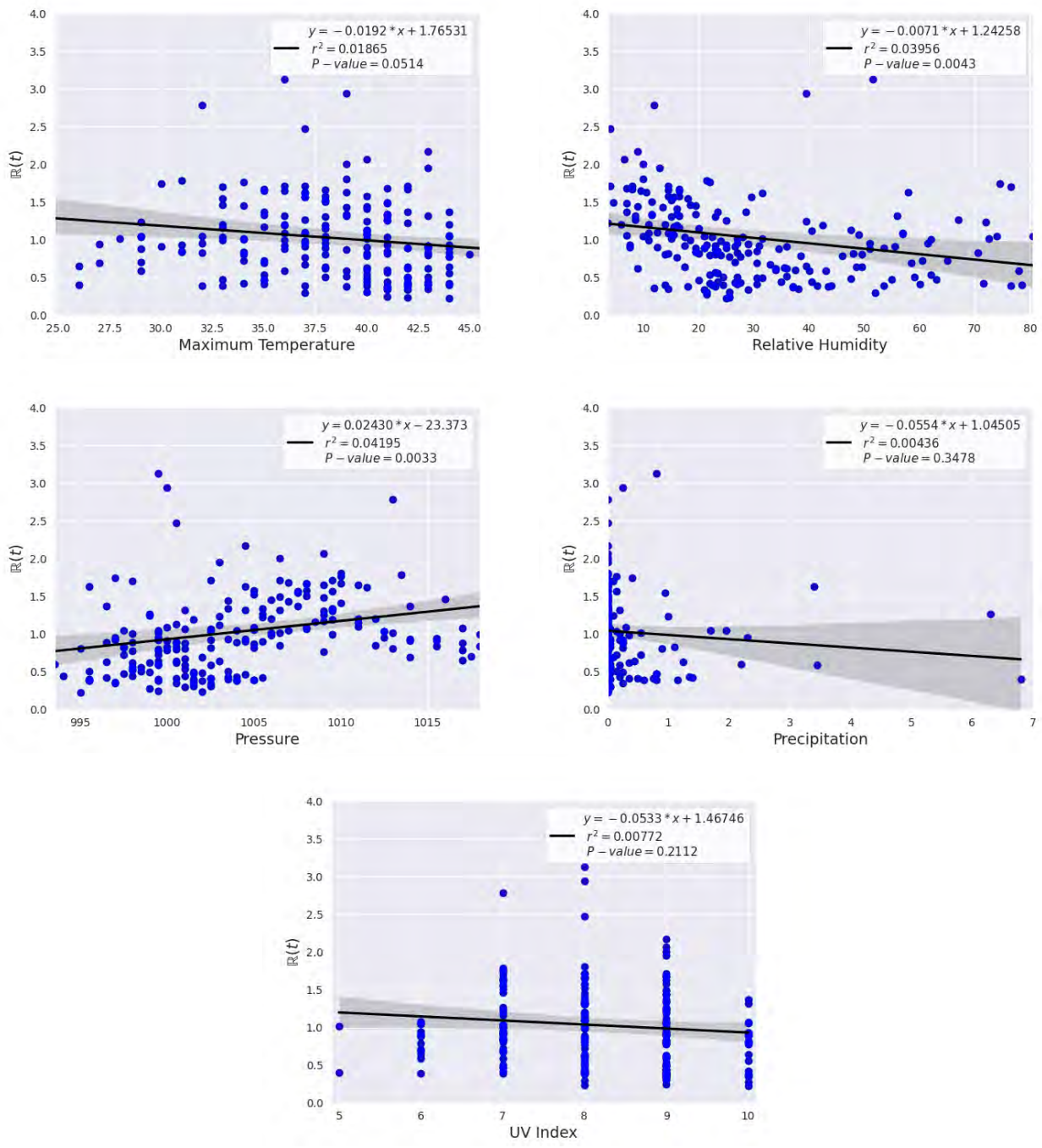


Figure 3: Variation of effective reproduction number ($\mathbb{R}(t)$) with different environmental factors for Delhi during 2nd February, 2021 to 24th August, 2021.

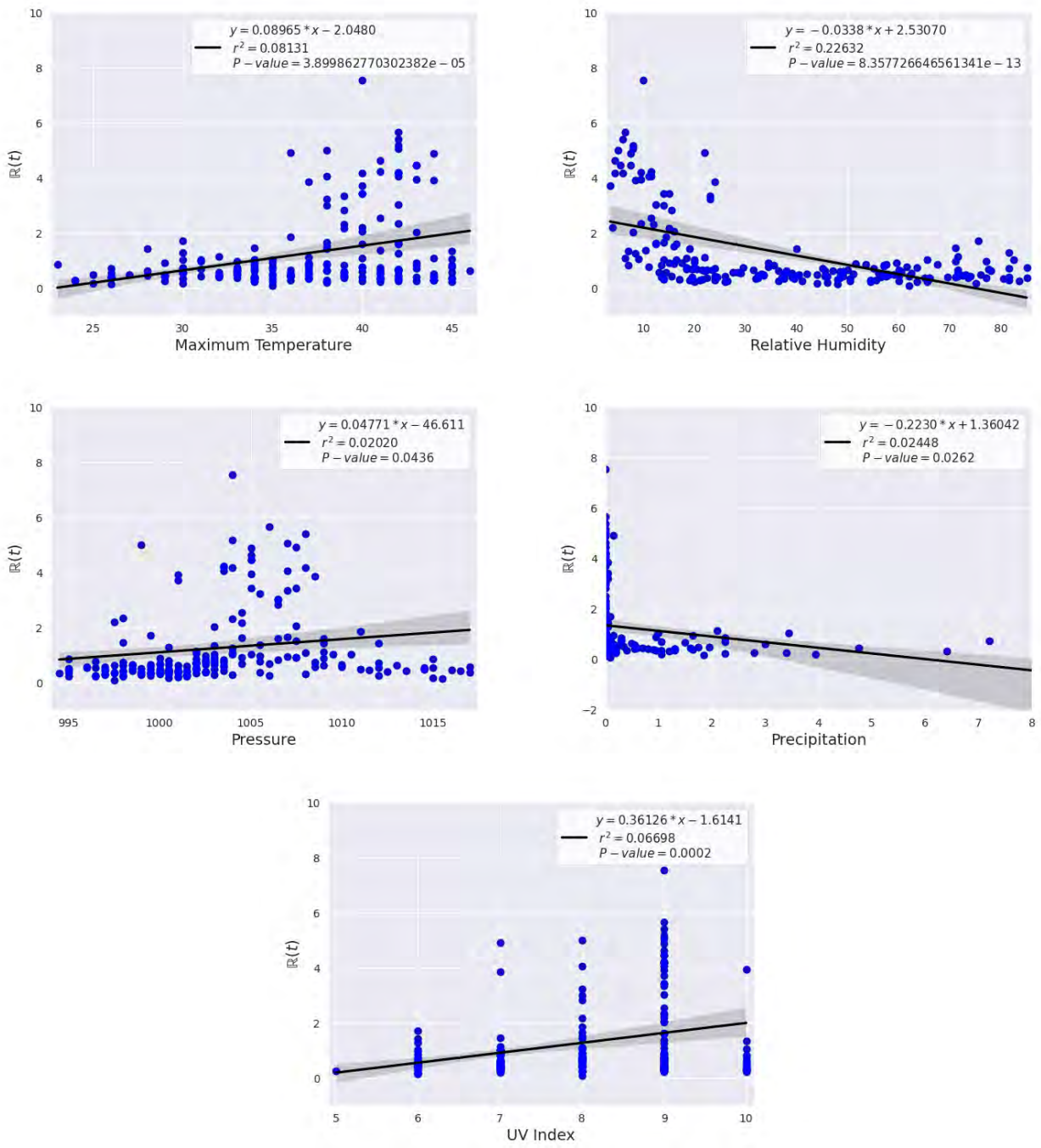


Figure 4: Variation of effective reproduction number ($\mathbb{R}(t)$) with different environmental factors for Uttar Pradesh during 2nd February, 2021 to 24th August, 2021.

All the studies assessing absolute humidity suggested a significant association with the COVID-19 transmission. It was found that in an optimal absolute humidity range 1–9 g m^{-3}, the majority of cases were reported [60]. It was found that 73.8% of confirmed cases are from regions where the absolute humidity was in the range between 3–10 g m^{-3} [61]. A negative correlation was found between absolute humidity and confirmed case across 17 cities in China [62]; when AH increased by 1 g m^{-3}, cases decreased RR of 0.72 (95% CI: 0.59–0.89) and 0.33 (95% CI: 0.21–0.51) respectively. In Singapore, a low-positive correlation between minimum, maximum and average relative humidity (r = 0.19, r = 0.20, and r = 0.21) with COVID-19 cases ($p < 0.05$) is reported, with no significant effect during early outbreak period (from February to March) [63]. This effect increased in intensity as the relative humidity ($80 \pm 4\%$) increased in May. Maximum (r = 0.27) and mean absolute humidity (r = 0.59) were reported to have a strong significant positive correlation association in comparison with relative humidity ($p < 0.01$) [63]. Earlier, a mixed result was reported for absolute humidity in some of the studies, for example, a significant negative association with daily confirmed cases reported for Pichincha ($p < 0.05$) and Rio de Janeiro ($p < 0.01$) and significant positive correlation is reported in Santiago ($p < 0.05$) [64].

Fig. 3 and Fig. 4 show the variation of effective reproduction number of COVID-19 with temperature, humidity, pressure, UV index, precipitation for two different Indian states Delhi and Uttar Pradesh. Both the states Delhi and Uttar Pradesh show a negative correlation between humidity and the reproduction number of COVID-19. There was plenty of

studies on India where both positive and negative correlations were observed for different Indian states [11], which implies that it is difficult to conclude an exact relation as so many social factors are also superposed with environmental factors.

3.3 Effect of Precipitation, Radiation and Wind Speed

There are very few studies analysing the correlation between COVID-19 and other meteorological factors such as precipitation, radiation, and wind speed. In several studies, from New York, Jakarta, Indonesia, Australia [26, 49, 56] no significant correlation was found for rainfall or precipitation. Earlier, a significant negative correlation was found between rainfall and COVID-19 transmission in the US, with daily cases increasing between 1.27–1.74 inches of rainfall and decreasing with rainfall over 1.77 inches of rainfall (<0.0001) [58]. Another significant negative association was reported for Oslo, Norway, with daily precipitation levels recorded at 7 a.m. in ($p < 0.05$) [65].

Wind speed was included in more than 10 studies. In three models, wind speed was included as a confounding factor, and no significant associations were reported [37, 48, 66]. Three of the remaining seven studies included wind speed as a meteorological variable, reported a significant correlation between wind speed and COVID-19 cases. Adekunle et al. reported a significant positive correlation with wind speed, with a 1% increase in average wind speed, there was an increase of 11.21% (95% CI: 0.51–1.19) COVID-19 cases in African countries [52]. Alkhowailed et al. [57] reported a significant negative correlation with maximum and average wind speed ($p < 0.001$ and $p < 0.01$, respectively). Pani et al. [63] also found a significant negative correlation between wind speed and COVID-19, where an increase in wind speed was associated with the decrease in new COVID-19 cases (r = – 0.6, $p < 0.001$). Bashir et al., Bukhari et al., Menebo, and Zhu et al. [27, 60, 64, 65] did not find a significant correlation between wind speed and daily COVID-19 cases.

3.4 Impact of air pollutants

Their several studies worldwide investigate the impact of different air pollutants on COVID-19 transmission. Li et al. (2020) [67] conducted an exploratory analysis looking for the potential association between different environmental factors, including air quality index (AQI), 4 air pollutants (PM2.5, PM10, NO_2, and CO), as well as 5 meteorological variables and COVID-19 incidence/mortality in Wuhan and XiaoGan. It was found that PM2.5 and NO_2 could promote the transmission of COVID-19, while the temperature was also associated with an increased incidence of the disease. Manik et al. (2022) [68] examined the influence of different outdoor air pollutants (CO, NO_2, SO_2, O_3, PM2.5) as well as Air Quality Index (AQI) on COVID-19 transmission over different Indian metropolitan cities (Kolkata, Mumbai, Chennai, Bangaluru, and Delhi). The results showed a significant positive association between daily confirmed cases and particulate matter (both PM2.5 and PM10). The air quality index also shows a positive association with COVID-19 confirmed cases. Gujral and Sinha (2021) [69] investigated the association between PM2.5, PM10, and O_3, and the COVID-19 incidence in Los Angeles and Ventura County, California. The results showed a significant correlation between airborne pollutants and COVID-19 cases. It was seen that short-term exposure to ground-level O3 was positively related to daily confirmed cases, while in contrast, exposure to PM2.5 and PM10 showed a negative association. Meo et al. (2021a) [70] investigated the relationship of wildfire allied pollutants (PM2.5, CO, and O_3) with the new daily cases and deaths due to SARS-CoV 2 infection in 10 counties of California, which were affected by wildfire. These air pollutants were temporally correlated with daily cases and daily deaths due to SARS-COV-2 infection. (Meo et al., 2021b) [71] recently published the findings of another study that looked into the correlation between PM2.5, CO, NO_2, and O_3 levels in the air and SARS-CoV-2-related daily new cases and deaths in five US regions (Los Angeles, New Mexico, New York, Ohio, and Florida). A positive correlation was found between different air pollutants (PM2.5, CO, NO_2, and O_3) and COVID-19 new cases and deceased. Sharma et al. (2021) [72] investigated the association between COVID-19 confirmed cases and deaths with meteorological factors and particulate matter (PM2.5). A significant positive correlation between air pollutants and the COVID-19 confirmed cases and deaths was observed.

3.5 Impact of Sunlight

SARS-CoV-2 loses 90 per cent of its infectivity in simulated saliva on a stainless steel surface after 6.8–12.8 minutes, depending on the intensity of simulated ultraviolet (UV) B radiation levels, when exposed to simulated sunlight representative of the summer solstice at 40° N latitude at sea level on a clear day [73]. These outcomes demonstrate that sunlight may inactivate SARS-CoV-2 rapidly on surfaces.

Experimental studies using SARS-CoV-2 aerosols produced from artificial saliva suggest that simulated sunlight inactivates the virus rapidly [74]. In simulated saliva, the half-life of aerosolized SARS-CoV-2 is approximately 86 minutes in dark conditions. After 6 minutes, the infectious concentration was determined to be 90% lower in the high-intensity sunlight-simulated summer. At 19 minutes, even with low-intensity sunlight simulated in late winter or early fall, a 90% decrease in infectious concentration was seen [74].

3.6 Stability on Surfaces of Fomites

At normal temperature, SARS-CoV-2 can survive for 2 to 4 days on plastic, stainless steel, and glass surfaces [75, 76]. SARS-CoV-2 had a similar persistence on those surfaces to SARS-CoV-1 [75, 76, 77]. The capacity of these viruses to survive on metal surfaces varied depending on the type of material. SARS-CoV-1 and SARS-CoV-2 both persisted less

time on copper (8 and 4 hours, respectively) than they did on stainless steel surfaces [75]. Copper and copper alloys have been shown to have antimicrobial properties against a variety of viruses [78, 79]. Studies have reported that copper alloys ($\geqslant$58% copper) reduce the surface microorganisms when integrated into different hospital furniture and fittings[80, 81]. The use of copper in combination with optimal infection-prevention strategies may further reduce the risk of patients and healthcare workers acquiring COVID-19 infection in healthcare environments. SARS-CoV-2 showed variable persistence on different porous surfaces, such as paper, cardboard, wood, cloth, and masks [75, 76].

4 Conclusion

We summarise the effect and role of different environmental factors on COVID-19 transmission. Based on the results from most of the studies, we can conclude that there is a significant association of different environmental variables (especially temperature, humidity, wind speed, UV index, pressure, rainfall) and various outdoor air pollutants (PM2.5, PM10, O_3, NO_2, SO_2, and CO) the spread of the epidemic. The results are especially evident for PM, and particularly for PM2.5, but the influence of O_3, NO_2, CO, and even SO_2 is not negligible at all.The impact of environmental variables is relatively small compared to other social factors such as governmental interventions, lock-downs, social interaction, herd immunity, migration patterns, population density, personal hygiene, defence mechanisms, etc. Therefore, countries should focus more on public healthcare policies and vaccines while taking into account the influence of weather on outbreaks.

References

[1] F. Wu, S. Zhao, B. Yu, Y. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. Holmes, and Y.-Z. Zhang, "A new coronavirus associated with human respiratory disease in china," *Nature*, vol. 579, pp. 1–8, 03 2020. doi: 10.1038/s41586-020-2008-3

[2] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, P. Niu, F. Zhan, X. Ma, D. Wang, W. Xu, G. Wu, G. Gao, and W. Tan, "A novel coronavirus from patients with pneumonia in china, 2019," *New England Journal of Medicine*, vol. 382, 01 2020. doi: 10.1056/NEJMoa2001017

[3] N. Chen, M. Zhou, X. Dong, J. Qu, F. Gong, Y. Han, Y. Qiu, J. Wang, Y. Liu, Y. Wei, J. Xia, T. Yu, X. Zhang, and L. Zhang, "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study," *The Lancet*, vol. 395, 01 2020. doi: 10.1016/S0140-6736(20)30211-7

[4] Z. Xu, L. Shi, Y. Wang, J. Zhang, L. Huang, C. Zhang, S. Liu, P. Zhao, H. Liu, L. Zhu, Y. Tai, C. Bai, T. Gao, J.-W. Song, P. Xia, J. Dong, J. Zhao, and F.-S. Wang, "Pathological findings of covid-19 associated with acute respiratory distress syndrome," *The Lancet Respiratory Medicine*, vol. 8, 02 2020. doi: 10.1016/S2213-2600(20)30076-X

[5] WHO, "Who director-general's opening remarks at the media briefing on covid-19 - 11 march 2020," *Geneva: World Health Organization*, 2020.

[6] M. Nicas and G. Sun, "An integrated model of infection risk in a health-care environment," *Risk analysis : an official publication of the Society for Risk Analysis*, vol. 26, pp. 1085–96, 09 2006. doi: 10.1111/j.1539-6924.2006.00802.x

[7] M. Nicas and R. Jones, "Relative contributions of four exposure pathways to influenza infection risk," *Risk analysis : an official publication of the Society for Risk Analysis*, vol. 29, pp. 1292–303, 07 2009. doi: 10.1111/j.1539-6924.2009.01253.x

[8] R. Tellier, Y. Li, B. Cowling, and J. Tang, "Recognition of aerosol transmission of infectious agents: A commentary," *BMC Infectious Diseases*, vol. 19, 01 2019. doi: 10.1186/s12879-019-3707-y

[9] Y. Ma, Y. Zhao, J. Liu, X. He, B. Wang, S. Fu, J. Yan, J. Niu, J. Zhou, and B. Luo, "Effects of temperature variation and humidity on the death of covid-19 in wuhan, china," *Science of The Total Environment*, vol. 724, p. 138226, 2020. doi: 10.1016/j.scitotenv.2020.138226

[10] H. Bherwani, A. Gupta, S. Anjum, A. Anshul, and R. Kumar, "Exploring dependence of covid-19 on environmental factors and spread prediction in india," *npj Climate and Atmospheric Science*, vol. 3, p. 38, 09 2020. doi: 10.1038/s41612-020-00142-x

[11] S. Manik, M. Mandal, S. Pal, S. Patra, and S. Acharya, "Impact of climate on covid-19 transmission: A study over indian states," *Environmental Research*, vol. 211, p. 113110, 2022. doi: https://doi.org/10.1016/j.envres.2022.113110

[12] L. Poole, "Seasonal influences on the spread of sars-cov-2 (covid19), causality, and forecastabililty (3-15-2020)," 2020.

[13] S. Manik, S. Pal, M. Mandal, and M. Hazra, "Effect of 2021 assembly election in india on covid-19 transmission," *Nonlinear Dynamics*, vol. 107, p. 1343–1356, 2022. doi: 10.1007/s11071-021-07041-7

[14] S. Manik, S. Pal, and M. Mandal, "Epitools: A python package for modelling and forecasting epidemic," *Submitted*, 2022.

[15] W. O. Kermack, A. G. McKendrick, and G. T. Walker, "A contribution to the mathematical theory of epidemics," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 115, no. 772, pp. 700–721, 1927. doi: 10.1098/rspa.1927.0118

[16] B. Ndiaye, L. Tendeng, and D. Seck, "Comparative prediction of confirmed cases with covid-19 pandemic by machine learning, deterministic and stochastic sir models," *arXiv: Populations and Evolution*, 2020.

[17] M. Li, *An Introduction to Mathematical Modeling of Infectious Diseases*, 01 2018. ISBN 978-3-319-72121-7

[18] Q. Bukhari and Y. Jameel, "Will coronavirus pandemic diminish by summer?" *SSRN Electronic Journal*, 01 2020. doi: 10.2139/ssrn.3556998

[19] K.-H. Chan, J. S. Peiris, S. Lam, L. Poon, Y. ky, and W. H. Seto, "The effects of temperature and relative humidity on the viability of the sars coronavirus," *Advances in virology*, vol. 2011, p. 734690, 10 2011. doi: 10.1155/2011/734690

[20] J. Wang, K. Tang, K. Feng, and W. Lv, "High temperature and high humidity reduce the transmission of covid-19," 03 2020. [Online]. Available: https://arxiv.org/abs/2003.05003

[21] S. Bhattacharjee, "Statistical investigation of relationship between spread of coronavirus disease (covid-19) and environmental factors based on study of four mostly affected places of china and five mostly affected places of italy," 2020.

[22] M. B. Araújo and B. Naimi, "Spread of sars-cov-2 coronavirus likely constrained by climate," *medRxiv*, 2020. doi: 10.1101/2020.03.12.20034728

[23] M. M. Sajadi, P. Habibzadeh, A. Vintzileos, S. Shokouhi, F. Miralles-Wilhelm, and A. Amoroso, "Temperature, Humidity, and Latitude Analysis to Estimate Potential Spread and Seasonality of Coronavirus Disease 2019 (COVID-19)," *JAMA Network Open*, vol. 3, no. 6, pp. e2 011 834–e2 011 834, 06 2020. doi: 10.1001/jamanetworkopen.2020.11834

[24] W. Luo, M. S. Majumder, D. Liu, C. Poirier, K. D. Mandl, M. Lipsitch, and M. Santillana, "The role of absolute humidity on transmission rates of the covid-19 outbreak," *medRxiv*, 2020. doi: 10.1101/2020.02.12.20022467

[25] A. Notari, "Temperature dependence of covid-19 transmission," *Science of The Total Environment*, vol. 763, p. 144390, 2021. doi: 10.1016/j.scitotenv.2020.144390

[26] R. Tosepu, J. Gunawan, D. S. Effendy, L. O. A. I. Ahmad, H. Lestari, H. Bahar, and P. Asfian, "Correlation between weather and covid-19 pandemic in jakarta, indonesia," *Science of The Total Environment*, vol. 725, p. 138436, 2020. doi: 10.1016/j.scitotenv.2020.138436

[27] M. F. Bashir, B. Ma, . Bilal, B. Komal, M. Bashir, D. Tan, and M. Bashir, "Correlation between climate indicators and covid-19 pandemic in new york, usa," *Science of The Total Environment*, vol. 728, p. 138835, 04 2020. doi: 10.1016/j.scitotenv.2020.138835

[28] A. Briz-Redon and A. Serrano-Aroca, "A spatio-temporal analysis for exploring the effect of temperature on covid-19 early evolution in spain," *Science of The Total Environment*, vol. 728, p. 138811, 04 2020. doi: 10.1016/j.scitotenv.2020.138811

[29] M. Ahmadi, A. Sharifi, S. Dorosti, S. Ghoushchi, and N. Ghanbari, "Investigation of effective climatology parameters on covid-19 outbreak in iran," *Science of The Total Environment*, vol. 729, p. 138705, 04 2020. doi: 10.1016/j.scitotenv.2020.138705

[30] M. Jahangiri, M. Jahangiri, and M. Najafgholipour, "The sensitivity and specificity analyses of ambient temperature and population size on the transmission rate of the novel coronavirus (covid-19) in different provinces of iran," *Science of The Total Environment*, vol. 728, p. 138872, 04 2020. doi: 10.1016/j.scitotenv.2020.138872

[31] I. Taiwo and A. Fashola, "Covid-19 spread and average temperature distribution in nigeria," *SSRN Electronic Journal*, 05 2020. doi: https://dx.doi.org/10.2139/ssrn.3585374

[32] T. Jamil, I. Alam, T. Gojobori, and C. Duarte, "No evidence for temperature-dependence of the covid-19 epidemic," *Frontiers in Public Health*, vol. 8, 08 2020. doi: 10.3389/fpubh.2020.00436

[33] M. Arumugam, B. Menon, and S. Narayan, "Ambient temperature and covid-19 incidence rates: An opportunity for intervention?" 04 2020.

[34] K. Chiyomaru and K. Takemoto, "Global covid-19 transmission rate is influenced by precipitation seasonality and the speed of climate temperature warming," 04 2020. doi: 10.1101/2020.04.10.20060459

[35] B. Pirouz, A. Golmohammadi, H. Masouleh, G. Violini, and B. Pirouz, "Relationship between average daily temperature and average cumulative daily rate of confirmed cases of covid-19," 04 2020. doi: 10.1101/2020.04.10.20059337

[36] Y. Wu, W. Jing, J. Liu, Q. Ma, J. Yuan, Y. Wang, M. Du, and M. Liu, "Effects of temperature and humidity on the daily new cases and new deaths of covid-19 in 166 countries," *Science of The Total Environment*, vol. 729, p. 139051, 2020. doi: 10.1016/j.scitotenv.2020.139051

[37] J. Xie and Y. Zhu, "Association between ambient temperature and covid-19 infection in 122 cities from china," *Science of The Total Environment*, vol. 724, p. 138201, 2020. doi: 10.1016/j.scitotenv.2020.138201

[38] D. Gupta and A. Gupta, "Effect of ambient temperature on covid 19 infection rate: Evidence from california," *SSRN Electronic Journal*, 2020. doi: 10.2139/ssrn.3575404

[39] M. Ujiie, S. Tsuzuki, and N. Ohmagari, "Effect of temperature on the infectivity of covid-19," *International Journal of Infectious Diseases*, vol. 95, 04 2020. doi: 10.1016/j.ijid.2020.04.068

[40] W. A. Iddrisu, P. Appiahene, and J. A. Kessie, "Effects of weather and policy intervention on COVID-19 infection in Ghana," *arXiv e-prints*, p. arXiv:2005.00106, 2020.

[41] A. Abdollahi and M. Rahbaralam, "Effect of temperature on the transmission of covid-19: A machine learning case study in spain," *medRxiv*, 2020. doi: 10.1101/2020.05.01.20087759

[42] A. Tobias and T. Molina, "Is temperature reducing the transmission of covid-19 ?" *Environmental Research*, vol. 186, p. 109553, 04 2020. doi: 10.1016/j.envres.2020.109553

[43] G. Livadiotis, "Statistical analysis of the impact of environmental temperature on the exponential growth rate of cases infected by covid-19," *PLOS ONE*, 2020. doi: 10.1371/journal.pone.0233875

[44] B. Oliveiros, L. Caramelo, N. C. Ferreira, and F. Caramelo, "Role of temperature and humidity in the modulation of the doubling time of covid-19 cases," *medRxiv*, 2020. doi: 10.1101/2020.03.05.20031872

[45] H. Qi, S. Xiao, R. Shi, M. Ward, Y. Chen, W. Tu, Q. Su, W. Wang, X. Wang, and Z. Zhang, "Covid-19 transmission in mainland china is associated with temperature and humidity: A time-series analysis," *Science of The Total Environment*, vol. 728, p. 138778, 04 2020. doi: 10.1016/j.scitotenv.2020.138778

[46] P. Shi, Y. Dong, H. Yan, C. Zhao, X. Li, W. Liu, M. He, S. Tang, and S. Xi, "Impact of temperature on the dynamics of the covid-19 outbreak in china," *Science of The Total Environment*, vol. 728, p. 138890, 2020. doi: 10.1016/j.scitotenv.2020.138890

[47] A. Sil and V. N. Kumar, "Does weather affect the growth rate of covid-19, a study to comprehend transmission dynamics on human health," *Journal of Safety Science and Resilience*, 2020. doi: 10.1016/j.jnlssr.2020.06.004

[48] W. Yu, W. Jing, J. Liu, Q. Ma, J. Yuan, Y. Wang, M. Du, and M. Liu, "Effects of temperature and humidity on the daily new cases and new deaths of covid-19 in 166 countries," *Science of The Total Environment*, vol. 729, p. 139051, 04 2020. doi: 10.1016/j.scitotenv.2020.139051

[49] A. Meyer, R. Sadler, C. Faverjon, A. R. Cameron, and M. Bannister-Tyrrell, "Evidence that higher temperatures are associated with a marginally lower incidence of covid-19 cases," *Frontiers in Public Health*, vol. 8, p. 367, 2020. doi: 10.3389/fpubh.2020.00367

[50] H. Liu, Y. Zhang, Y. Tian, Y. Zheng, F. Gou, X. Yang, J. He, X. Liu, L. Meng, and W. Hu, "Epidemic features of seasonal influenza transmission among eight different climate zones in gansu, china," *Environmental Research*, vol. 183, p. 109189, 2020. doi: 10.1016/j.envres.2020.109189

[51] E. Hadi and M. Jalili, "The role of environmental factors to transmission of sars-cov-2 (covid-19)," *AMB Expres*, vol. 10, no. 1, p. 92, 2020. doi: https://doi:10.1186/s13568-020-01028-0

[52] I. A. Adekunle, S. A. Tella, K. O. Oyesiku, and I. O. Oseni, "Spatio-temporal analysis of meteorological factors in abating the spread of covid-19 in africa," *Heliyon*, vol. 6, no. 8, p. e04749, 2020. doi: 10.1016/j.heliyon.2020.e04749

[53] K. Goswami, S. Bharali, and J. Hazarika, "Projections for covid-19 pandemic in india and effect of temperature and humidity," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 5, pp. 801–805, 2020. doi: 10.1016/j.dsx.2020.05.045

[54] S. Hossain, S. Ahmed, and M. J. Uddin, "Impact of weather on covid-19 transmission in south asian countries: an application of the arimax model," *Science of The Total Environment*, 11 2020. doi: 10.1016/j.scitotenv.2020.143315

[55] F. Shahzad, U. Shahzad, Z. Fareed, N. Iqbal, S. Hashmi, and F. Ahmad, "Asymmetric nexus between temperature and covid-19 in the top ten affected provinces of china: A current application of quantile-on-quantile approach," *Science of The Total Environment*, vol. 736, p. 139115, 05 2020. doi: 10.1016/j.scitotenv.2020.139115

[56] M. P. Ward, S. Xiao, and Z. Zhang, "The role of climate during the covid-19 epidemic in new south wales, australia," *Transboundary and Emerging Diseases*, vol. 67, no. 6, pp. 2313–2317, 2020. doi: 10.1111/tbed.13631

[57] M. Alkhowailed, A. Shariq, F. Alqossayir, O. A. Alzahrani, Z. Rasheed, and W. Al Abdulmonem, "Impact of meteorological parameters on covid-19 pandemic: A comprehensive study from saudi arabia," *Informatics in Medicine Unlocked*, vol. 20, p. 100418, 2020. doi: 10.1016/j.imu.2020.100418

[58] L.-C. Chien and L.-W. A. Chen, "Meteorological impacts on the incidence of covid-19 in the u.s," *Stochastic Environmental Research and Risk Assessment*, vol. 34, 07 2020. doi: 10.1007/s00477-020-01835-8

[59] A. Paez, F. A. Lopez, T. Menezes, R. Cavalcanti, and M. G. d. R. Pitta, "A spatio-temporal analysis of the environmental correlates of covid-19 incidence in spain," *Geographical Analysis*, vol. 53, no. 3, pp. 397–421, 2021. doi: 10.1111/gean.12241

[60] Q. Bukhari, J. Massaro, R. D'Agostino, and S. Khan, "Effects of weather on coronavirus pandemic," *International Journal of Environmental Research and Public Health*, vol. 17, p. 5399, 07 2020. doi: 10.3390/ijerph17155399

[61] Z. Huang, J. Huang, Q. Gu, P. Du, H. Liang, and Q. Dong, "Optimal temperature zone for the dispersal of covid-19," *Science of The Total Environment*, vol. 736, p. 139487, 05 2020. doi: 10.1016/j.scitotenv.2020.139487

[62] J. Liu, J. Zhou, J. Yao, X. Zhang, L. Li, X. Xu, X. He, B. Wang, S. Fu, T. Niu, J. Yan, Y. Shi, X. Ren, J. Niu, W. Zhu, S. Li, B. Luo, and K. Zhang, "Impact of meteorological factors on the covid-19 transmission: A multi-city study in china," *Science of The Total Environment*, vol. 726, p. 138513, 2020. doi: 10.1016/j.scitotenv.2020.138513

[63] S. K. Pani, N.-H. Lin, and S. RavindraBabu, "Association of covid-19 pandemic with meteorological parameters over singapore," *Science of The Total Environment*, vol. 740, p. 140112, 2020. doi: 10.1016/j.scitotenv.2020.140112

[64] L. Zhu, X. Liu, H. Huang, R. D. Avellán-Llaguno, M. M. L. Lazo, A. Gaggero, R. Soto-Rifo, L. Patiño, M. Valencia-Avellan, B. Diringer, Q. Huang, and Y.-G. Zhu, "Meteorological impact on the covid-19 pandemic: A study across eight severely affected regions in south america," *Science of The Total Environment*, vol. 744, p. 140881, 2020. doi: 10.1016/j.scitotenv.2020.140881

[65] M. M. Menebo, "Temperature and precipitation associate with covid-19 new daily cases: A correlation study between weather and covid-19 pandemic in oslo, norway," *Science of The Total Environment*, vol. 737, p. 139659, 2020. doi: 10.1016/j.scitotenv.2020.139659

[66] T. To, K. Zhang, B. Maguire, E. Terebessy, I. Fong, S. Parikh, and J. Zhu, "Correlation of ambient temperature and covid-19 incidence in canada," *Science of The Total Environment*, vol. 750, p. 141484, 2021. doi: 10.1016/j.scitotenv.2020.141484

[67] H. Li, X.-L. Xu, D.-W. Dai, Z.-Y. Huang, Z. Ma, and Y.-J. Guan, "Air pollution and temperature are associated with increased covid-19 incidence: a time series study," *International journal of infectious diseases*, vol. 97, pp. 278–282, 2020.

[68] S. Manik, M. Mandal, and S. Pal, "Impact of air pollutants on covid-19 transmission: A study over different metropolitan cities in india," *Submitted*, 2022.

[69] H. Gujral and A. Sinha, "Association between exposure to airborne pollutants and covid-19 in los angeles, united states with ensemble-based dynamic emission model," *Environmental research*, vol. 194, p. 110704, 2021.

[70] S. A. Meo, A. A. Abukhalaf, A. A. Alomar, O. M. Alessa, W. Sami, and D. C. Klonoff, "Effect of environmental pollutants pm-2.5, carbon monoxide, and ozone on the incidence and mortality of sars-cov-2 infection in ten wildfire affected counties in california," *Science of the Total Environment*, vol. 757, p. 143948, 2021.

[71] S. A. Meo, A. A. Abukhalaf, O. M. Alessa, A. S. Alarifi, W. Sami, and D. C. Klonoff, "Effect of environmental pollutants pm2. 5, co, no2, and o3 on the incidence and mortality of sars-cov-2 infection in five regions of the usa," *International Journal of Environmental Research and Public Health*, vol. 18, no. 15, p. 7810, 2021.

[72] G. D. Sharma, S. Bansal, A. Yadav, M. Jain, and I. Garg, "Meteorological factors, covid-19 cases, and deaths in top 10 most affected countries: an econometric investigation," *Environmental Science and Pollution Research*, vol. 28, no. 22, pp. 28 624–28 639, 2021.

[73] S. Ratnesar-Shumate, G. Williams, B. Green, M. Krause, B. Holland, S. Wood, J. Bohannon, J. Boydston, D. Freeburger, I. Hooper, K. Beck, L. Altamura, J. Biryukov, J. Yolitz, M. Schuit, V. Wahl, M. Hevey, and P. Dabisch, "Simulated sunlight rapidly inactivates sars-cov-2 on surfaces," *The Journal of infectious diseases*, vol. 222, 05 2020. doi: 10.1093/infdis/jiaa274

[74] M. Schuit, S. Ratnesar-Shumate, J. Yolitz, G. Williams, W. Weaver, B. Green, D. Miller, M. Krause, K. Beck, S. Wood, B. Holland, J. Bohannon, D. Freeburger, I. Hooper, J. Biryukov, L. Altamura, V. Wahl, M. Hevey, and P. Dabisch, "Airborne sars-cov-2 is rapidly inactivated by simulated sunlight," *The Journal of infectious diseases*, vol. 222, 06 2020. doi: 10.1093/infdis/jiaa334

[75] N. van Doremalen, T. Bushmaker, D. Morris, M. Holbrook, A. Gamble, B. Williamson, A. Tamin, J. Harcourt, N. Thornburg, S. Gerber, J. Lloyd-Smith, E. Wit, and V. Munster, "Aerosol and surface stability of hcov-19 (sars-cov-2) compared to sars-cov-1," *medRxiv : the preprint server for health sciences*, 03 2020. doi: 10.1101/2020.03.09.20033217

[76] S.-M. Duan, X.-S. Zhao, R.-F. Wen, J.-J. Huang, G.-H. Pi, S.-X. Zhang, J. Han, S.-L. Bi, L. Ruan, and X.-P. Dong, "Stability of sars coronavirus in human specimens and environment and its sensitivity to heating and uv irradiation," *Biomedical and environmental sciences : BES*, vol. 16, pp. 246–55, 10 2003.

[77] A. Chin, J. Chu, R. A. Perera, K. Hui, H.-L. Yen, M. Chan, J. S. Peiris, and L. Poon, "Stability of sars-cov-2 in different environmental conditions," 03 2020. doi: 10.1101/2020.03.15.20036673

[78] H. Michels, W. Moran, and J. Michel, "Antimicrobial properties of copper alloy surfaces, with a focus on hospital-acquired infections," *International Journal of Metalcasting*, vol. 2, pp. 47–56, 06 2008. doi: 10.1007/BF03355432

[79] S. Warnes and C. Keevil, "Inactivation of norovirus on dry copper alloy surfaces," *PloS one*, vol. 8, p. e75017, 09 2013. doi: 10.1371/journal.pone.0075017

[80] T. Karpanen, A. Casey, P. Lambert, B. Cookson, P. Nightingale, L. Miruszenko, and T. Elliott, "The antimicrobial efficacy of copper alloy furnishing in the clinical environment: A crossover study," *Infection control and hospital epidemiology : the official journal of the Society of Hospital Epidemiologists of America*, vol. 33, pp. 3–9, 01 2012. doi: 10.1086/663644

[81] D. Montero, C. Arellano, M. Pardo, R. Vera, L. Galvez Arevalo, M. Cifuentes Diaz, M. Berasain, M. Gómez, C. Ramírez, and R. Vidal, "Antimicrobial properties of a novel copper-based composite coating with potential for use in healthcare facilities," *Antimicrobial Resistance & Infection Control*, vol. 8, 01 2019. doi: 10.1186/s13756-018-0456-4

Extreme Weather Events and Its Impact in Agriculture

Javed Akhter[1*], Manish Kumar Naskar[2], Shakil Hassan[2] and Subrata Kumar Midya[1]

[1]Department of Atmospheric Sciences, University of Calcutta
[2]Department of Agricultural Meteorology and Physics, Bidhan Chandra Krishi Viswavidyalaya, Mohanpur, Nadia, 741252
*Corresponding author:*akhterexpressju@gmail.com*

Abstract In the context of climate change several weather aberrations takes place in a frequent manner. Short scale variability of temperature, rainfall, wind speed, relative humidity, evaporation can create havoc imbalance within the atmosphere leading to the weather hazards. The deficiency of moisture is the key element causing drought. Proper crop planning along with efficient water harvesting technologies can mitigate the effect of drought. While flood is the opposite situation of drought where excess inflow of water damages crop production, livestock and human health. Proper land use policies along with preventive river bund Infrastructures are the key factors mitigating flood. Apart from drought and flood some extreme short period weather hazards accounts for tropical cyclone, hail storm and cloudburst which are very difficult to predict. Region specific weather hazards are cold wave, heat wave and Frost. Genetically stress tolerant varieties along with few technological adaptations can improve the mitigation ability of such hazards. Moreover, weather cannot be predicted perfectly at any situation till date. Risk of Crop failure, livestock and human loss are evident. Awareness regarding those extreme weather phenomena can reduce the casualties to some extent.

Keywords:*Extreme weather, Climate Change, Agriculture, Crop planning, Disaster Management*

1. Introduction

Extreme events have been projected to intensify and be more frequent under changing climatic conditions [1]. Extreme events, such as droughts and heat waves, have adverse impact agricultural production and have implications for the livelihoods and food security of communities. The impact is not only limited to the regions immediately experiencing the extreme event, but has far reaching consequences to other parts of the world through reduced exports of agricultural products and higher food prices [2] [3]. The European heat wave and drought in summer 2018, which resulted in widespread crop failures and livestock feed shortages in several areas throughout the continent, is a recent illustration of the consequences of climatic extremes on agricultural production [4]. Understanding the influence of climatic extremes on agricultural yields in the past and current climates is critical for securing and optimising harvests in a changing climate. Keeping this in mind, the article has been organized as following: Firstly, characteristics of various extreme weather events have been mentioned. Next, the impacts of these extreme events on crops have been discussed. Finally, some recommendations for mitigation and adaptation have been provided.

2. Types of Extreme Weather

2.1. Drought

Though drought is created due to prolonged shortage of water it is considered as an extreme weather event. The short-term water deficit leads to longer aridity and dry spell. National Commission on agriculture in 1976 has categorized drought into three types; they are 1) Meteorological drought, 2) Hydrological drought and 3) Agricultural drought.

Meteorological drought is based on precipitation. It is classified into three categories; light, moderate and severe meteorological drought. Light drought happens when the area is less than 75% of its normal value. Moderate drought is the situation when the seasonal rainfall received over that area is in between 50% to 74% of its normal value. While severe drought is the situation less than 50% of its normal value.

Hydrological drought occurs in drying up of surface water resource system and abbreviated through W= G-L (G- inflow of water like rainfall, L-loss of water like evapotranspiration, W- magnitude of water loss or gain)

From the perspective of agriculture and crop production, Agricultural drought is the most important drought to study. It is described as a period during which the soil, precipitation, and rainfall are insufficient to sustain the growth of a healthy crop to maturity, resulting in excessive crop stress and wilt [5]. It is critically divided into five classes such as, early, mid-season, late, terminal, permanent and apparent drought. Mid-season and terminal droughts are often associated with low productivity. Though the apparent drought can be easily escaped through the proper selection of crop variety, planning and irrigation techniques.

The drought assessment has also evolved with time. The in-situ methods like measurement of rainfall, soil moisture status can easily provide an accurate picture of the drought situation. But those are very much location specific and lack spatial

information of a region. While the remote sensing techniques are capable of extracting region-specific drought scenario through image processing. They need ground truthing and are not perfectly accurate in drought accounting. Recently synergistic approaches of both the in-situ and the remote sensing techniques can elaborate the drought scenario more accurately over a diversified area of interest [6].

2.2.Flood

Overflowing or influx of surface water beyond its capacity over land is called flood. The low-lying areas are the worst affected through this kind of weather hazards.

Causes of flood:

1. Insurgence of surplus water during monsoon season along the river basin. The depth of the river basins reduced due to silt deposition in the middle and end section of its tributary. Results in overflowing in the catchment areas.
2. Climate change is also held accountable for the melting of glacial ice results in more influx of water in the river basins.
3. Unplanned catchment area and river basins also results in breaking river bunds causing flood.
4. Decreasing capacity of River dams (like DVC), due to high silt load forces to release water during monsoon season causing devastating floods in the unexpected areas.

2.3.Cloud Burst

Highly concentrated rainfall (10cm/hour) over a small area lasting less than a hour specially in mountainous region can instantly create massive flood along with land slide suddenly [7]. This is known as cloud burst. Highly moisture laden air lifted with strong upward convective currents instantly converts into rainfall with the contact of glacial ice causing flash flood. In India during the monsoon season, such cloud burst is generated by the westward passage of monsoon depressions and mid tropospheric low-pressure systems. It is a highly localised event whose impact is wide spread.

Cloud burst is a natural and common phenomenon in the Himalaya, especially in Garhwal and Kumaon region of Uttarakhand, at least 26 cloud bursts occurred in the Himalayan region from January through July 29, 2021. Cloud burst and associated disaster affects thousands of people every year and cause loss of life, property, livelihood, infrastructure and environment. The 2021 flood in Uttarakhand, that resulted in over 200 dead and missing was the result of an avalanche that dropped about 27 million cubic meters of rock and glacier ice from the nearby Ronti mountain.

2.4.Tropical Cyclone

A tropical cyclone is a strong cyclonic circulation which is categorised by a low-pressure centre along with thunderstorm. The development of a tropical cyclone pre-requisite is high temperature (>27ºC) for at least seven days over ocean and forming a huge low-pressure zone. This strong low-pressure system is also known as deep depression. The cyclone includes a relatively calm region in the centre of the low-pressure area known as “eye”. A fully grown cyclone’s eye can be easily visible from the satellite images as a small, circular cloud free spot in the middle of the big cyclonic circulation. The wind speed can exceed 64 knot or 119km/h. Its speed can substantially reach up to 165knot. There is a wall formed around the eye about a diameter from 16-80km in size. This zone is the root of all destruction as the wind speed in the wall near the eye is maximum.

2.5.Hailstorm

Any thunderstorm which produces hail and reaches the ground is known as a hailstorm. Thunderclouds that are capable of producing hailstones are often seen, obtaining green coloration. Hail is more common along mountain ranges because mountains force horizontal winds upwards (known as orographic lifting), thereby intensifying the updrafts within thunderstorms and making hail more likely.

2.6.Frost

Frost is the frozen condition below freezing point (0ºC), the dew that is formed converts into ice crystals. The water molecule in the air directly converts into solid and form hard ice. It is very slippery. This occurs normally in mountain valleys where cold and heavy air accumulates at the surface level due to temperature inversion.

Kinds of frost:

Radiation frost: During clear nights when there is clear sky the terrestrial radiation easily escapes the earth atmosphere causing the steep fall of the temperature in the atmosphere and convert the water vapour molecule directly into the ice crystal on the earth surface.

Hoar or white frost: It is caused by sublimation of ice crystals on objects such as tree branches and wires when these objects are at temperature below freezing.

Black frost: This type of frost occurs when there is insufficient moisture to form ice crystals thus freezes the vegetation and kill the surface canopy. The blackish appearance of the dried and cold shocked canopy structures yielded its nomenclature "Black frost".

2.7. Heat Waves

In the month of March – July spells of abnormally high temperature along with gusty dry wind is experienced in certain parts of India. The inland when abruptly gets heated a hot and dry air mass is created over the inward portion of the country. This causes the insurgence of very hot and dry air into the neighboring areas in form of wave. This is known as Heat wave. The temperature is increased by several degrees by such insurgence. The daily maximum temperature is the actual indicator for the progress of heat wave [8].

The criteria for describing heat wave are as follows:

1. When the normal maximum temperature (T_{max}) is 40ºC or less

Sl. No	Nomenclature	Departure from normal
1	Normal	-1 to +1ºC
2	Above normal	2ºC
3	Appreciably above normal	3 to 5ºC
4	Moderate heat wave	6 to 7ºC
5	Severe heat wave	8ºC or more

2. When normal T_{max} is more than 40 ºC

Sl. No	Nomenclature	Departure from normal
1	Normal	-1 to +1ºC
2	Above normal	2ºC
3	Heat wave	3 to 4ºC
4	Severe heat wave	5ºC

2.8.Cold Wave

The wind in the higher altitude from the northern latitude of India during November to March insurges a cold atmosphere in the prevailing region. Northern continental landmass gives birth to this kind of cold and relatively dry air masses. The air mass is heavy and relatively very cold with respect to the prevailing temperature of the insurging area. They form a cold front and engulfs the adjoining area. They yield in appreciable fall in minimum temperature (or night temperature). In West Bengal situation cold waves are generally cited during December to March.

The criteria for describing cold wave are as follows:

1. When normal minimum temperature of the area is 10 ºC or more, the cold wave is classified as follows:

Sl no.	Nomenclature	Departure from normal
1	Normal	+1ºC to -1ºC
2	Below normal	-2ºC
3	Appreciably below normal	-3 to -4 ºC
4	Moderate cold wave	-5 to -6 ºC
5	Severe cold wave	-7 ºC or less

2. When normal minimum temperature of the area is less than 10ºC, the cold wave is classified as follows:

Sl no.	Nomenclature	Departure from normal
1	Normal	+1ºC to -1ºC
2	Below normal	-2ºC
3	Cold wave	-3 to -4 ºC
4	Severe cold wave	-5 ºC

3. Types of Impact Extreme Weather Events on Agriculture

3.1.Drought Impact on Agriculture

Drought has a severe effect on Indian agriculture. Especially minute agricultural drought which is not visible through naked eyes impacts the crop production most. Water scarcity during the critical crop growth stages reduces crop yield. In case of cereals drought during grain filling stages causes lower seed yield, chaffy grain associated with economic yield

loss. While in pulses drought during vegetative phases hampers overall growth. Chickpea, lentil, and field pea are susceptible to drought while lathyrus and pigeon pea are relatively more drought tolerant. Vegetables are short lived and have a relatively high economic output. Water scarcity during critical growth stages significantly reduces yield.Economic crop like sugarcane, jute require high water input, shortage of water in any growth stages result in yield loss.

There are several indices to measure the moisture adequacy and its impact on crop growth and behavior.

i.Moisture Adequacy Index

Moisture adequacy index is the ratio of actual evapotranspiration to potential evapotranspiration of any particular point. MAI=AET/PET. It indicates the moisture suitability of any crop in a particular region through the ratio. When the value is 1 there is no water deficit. The value less than 1 defines the stress situation.

ii.Water Requirement Satisfaction Index

WRSI is the ratio of seasonal actual crop evapotranspiration (AETc) to the seasonal crop water requirement, denoted as crop specific potential evapotranspiration by the use of appropriate crop coefficients. Kc value is incorporated in this index to make it crop specific as well as crop growth stage specific. The maximum value can be of 100. The crop when receives moisture stress the index value is decreased and cannot be uplifted by further irrigation after the stress.

iii.IW/CPE ratio

This simply the ratio between irrigation to be given to cumulative pan evaporation. This is also crop specific and stage specific. The ratio falling below certain value in certain growth stages of the crop indicates irrigation requirement to avoid crop stress. If the area receives rainfall, then the rainfall amount is deducted from the cumulative pan evaporation to calculate the value. The ratio values optimum for wheat, mustard and pulses are 0.9, 0.6 and 0.4 respectively.

3.2. Flood Impact on Agriculture:

Flood is one of the greatest natural hazards associated with human life loss and substantial crop failure. In Indian subcontinent, flood during monsoon season is a quite devastating phenomenon. Mainly the low-lying river basins are prone to this type of seasonal floods. Kharif crop especially rice production is highly affected through such weather hazards. Newly transplanted rice plants cannot withstand heavy water logging. Due to puddling condition the situation become worse even more [9].

3.3. Cloud Burst Impact on Agriculture and Livelihood:

i. Destruction of contour bunds due to flash flood.
ii. Life loss of humans as well as livestock due to sudden land slide.
iii. Destruction of properties [10].
iv. Disrupt soil quality and health as stones and pebbles mixes with the upper soil layer.
v. Disruption of roads and bridges heavily damaging trades and economics.

3.4. Tropical Cyclone Impact on Agriculture and Livelihood:

i. Destruction of kachcha houses, electric posts and other small infrastructures.
ii. Excessive rainfall induced floods destroys standing crop.
iii. Danger for the sailing fisherman and coastal adjoining livelihood.
iv. High speed wind during reproductive phase reduces pollination hence decreases crop yield. Lodging of grain crops (specially rice) due to high wind speed which is an irreversible action [11].

3.5. Hailstorm Impact on Agriculture:

i. Apical buds of the emerging crops, especially in the early stages are heavily destroyed.
ii. Ripened grains are also shattered due to the hitting effect of the hails.
iii. Substantial reduction in plant population due to the hail strike and collaterally yield is reduced.

3.6. Frost Impact on Agriculture:

i. Crops are permanently destroyed due to black frost for few days
ii. Radiational frost increases crop duration, becomes infertile and production is reduced [12].
iii. Longer exposure to white frost during reductive phase reduces yield [13].
iv. It preserves the diseases within the frozen environment and provides reproducibility whenever favorable weather situation prevails. Thus, eradication of some diseases becomes impossible in the frost prone regions [14].
v. Cracks are formed on the fruits due to frost (e.g. apple) and economically punishes.

3.7. Heat Wave Impact on Agriculture:

i. Hampers in water supply.
ii. Severe heat waves are responsible for killing working farmers and grazing livestock in the field.
iii. In case of rice, booting to grain development stage is most sensitive phase.
iv. Heat stress is often accompanied by drought increasing the transpiration causing dehydration or stress [15] [16].

v. Carbohydrate depletion is observed as the respiration quotient is high [17].
vi. Hinders nitrogen and lipid metabolism and injure cell membranes [17].
vii. Harmful chemicals like ammonia is released due to heat stress which causes internal injury [17].

3.8. Cold Wave Impact on Agriculture:

i. Cold wave affects both root and shoot.
ii. During blooming stage, it causes poor flowering and pollination and hence low fruit/seed setting.
iii. Higher plants show slow growth rate.
iv. Stunted growth of boro rice seedlings, yellowing of leaves and eventually drying up of young seedlings in nursery bed. This phenomenon is called winter burning of boro rice seedlings.
v. Low soil temperature hinders water absorption and causes water stress in cotton [17][18].

4. Mitigation and Adaptation

In this section, we have mentioned some of recommendations regarding mitigation and adaptation options against the adverse impacts of various extreme events that affect agriculture.

4.1. Drought

i. Crop affected by early drought is advised to go with resowing.
ii. Mid-season drought can be mitigated through spraying of anti-transpirant material like kaolin [1][19].
iii. If the crop is affected by terminal drought, then irrigation is highly recommended to decrease the yield loss.
iv. In permanently drought prone area suitable stress tolerant crops and varieties should be introduced along with adapting good water harvesting technique.
v. Crop selection is also a major adaptation technique to avoid drought as different crop posses different water requirement. Low water requiring crop can be incorporated in the dryer section [20].

4.2. Flood

i. Adaptation methods to prevent flood disasters in agricultural land can be categorized into engineering and non-engineering methods [21].
ii. One of the engineering methods is to improve the drainage and sewer systems. The infiltration volume of paving need to be increased and a reservoir should be built to minimize the damage to agricultural products.
iii. Increasing the height of farmland ridges by 10-60 cm is another option to reduce the flood affected areas.
iv. Deploying pumping station in wasteland would strengthen structural capacity and add more control over floods [21].
v. Among non-engineering methods adjustment of agricultural production period or changing crops can be useful. Flood seasons should be avoided or cropping pattern should be changed to reduce the loss in agricultural outputs.
vi. Fallowing in flood prone areas should be encouraged by providing subsidies.
vii. Restoration of flood-prone areas to its original ecological condition and creation of protected areas is necessary to cope with flood impacts on agriculture [21].

4.3. Tropical Cyclone

i. Wind brakes can be useful to reduce damage caused by tropical cyclones.Deployment of several tiers of wind-brake plants in the coastal areas would be effective to reduce wind speed as well as neutralize adverse effects of high tides by lowering the amount of saline water and sand mass [22].
ii. Plant species like casuarinas, eucalyptus and acacia are suitable for the formation of wind-brake system. These plants also act as bio-drainage system that can help to reduce water-logging caused by cyclone-induced rainfall.
iii. Proper nursery management also required to minimize the damage created by post-cyclonic rainfall. To acquire healthy seedlings early in the season, vegetable seedlings should be grown in low-cost poly houses.
iv. Watermelon, pumpkin, ridge gourd, cucumber, and bitter gourd are creeping vegetables that require a larger spacing and are not suited for transplantation. These crops' seedlings may be grown in poly homes or other protected areas utilising disposable plastic cups and rich soil.

4.4. Hailstorm

i. The impact of the damages caused by hailstorm can be minimized through hail abatement. Physical barriers such as hail nets or similar protective screens can be used to intercept incoming hailstones. This procedure has been successfully implemented in the apple orchards of Himachal Pradesh [23].
ii. Hail abatement can be accomplished through plantation where hail is associated with strong winds. Trees can directly intercept some number of hails and thus protect the crops which are immediately downwind.
iii. The trees also provide a shift in wind, allowing hail to be deflected laterally and partially sheltering the region on the leeward side.
iv. In the lee of the shelter, wind speeds will be lower, resulting in lower total hail kinetic energy, which is the sum of the vertical fall speed of the hail and the wind speeds. If the location's most prevalent wind direction is identified, shelter belts can be planted perpendicular to it to prevent crop damage [24].

v. Another approach to reduce hailstorm damage in high-frequency areas is to plant crops that are less susceptible to hail. Wheat or other crops might be cultivated instead of fruit crops that are more sensitive to hail in some of the most prominent horticultural areas that are frequently prone to severe hail damage [23].

4.5. Frost

i. Frost incidence is more in ploughed soil than in unploughed and compact soil. In frost affected areas, the land should not remain ploughed and open.
ii. Weeds add to radiative cooling and hasten the process of cooling and hence frosting. The field should remain weeded and clean and moist to reduce frost hazard.
iii. Covering materials opaque to long wave radiation from ground may be used to reduce frost hazard to growing trees. These are called hot caps. These should be placed on the plants before evening and taken out next morning. Straw mats, reed (tall grass) screen, hessian net screen, and plastic paper serve this purpose.
iv. Reflective materials such as aluminium foil and protein-based foam materials may also be used to protect the crop from frost.
v. Mulching with straw, dust polythene reduces frost hazard.
vi. Smoking and fogging are used as effective method in reducing frost hazard in vineyard in Europe. Burning wood, straw saw dust, coal tar, etc are usually adopted to reduce frost hazard.
vii. In big orchards in Europe and America, heaters are used to reduce frost hazard.
viii. During frosty nights the process of inversion, i.e. the air layer above the ground remains warmer, occurs. If this inversion is broken, i.e. mixing the upper warmer layer with the lower cooler layer, frost hazard is reduced. For this purpose, wind machines, fans and helicopters are used in temperate countries.
ix. Flooding the crop field is a good method to reduce frost damage. Flooding increases thermal capacity and heat conductivity of the ground and latent heat is released when water freezes.

4.6. Heat Wave

i. Frequent irrigation should be given [25].
ii. No tillage practices can prevent the effect of heat wave to some extent [26].
iii. Heat tolerant crops and varieties should be introduced in heat wave prone zones
iv. Long shade plants planted in a long-term basis can reduce the effect of heat wave to some extent.

4.7. Cold wave

i. Green house cultivation in those areas can be a boon for the farmers
ii. Arrangement for hot air blowers or heater for orchards can prevent further damage the trees.
iii. Irrigation before sunset can prevent cold wave damages to some extent.
iv. Development of stress tolerant varieties [27].

5. Extreme Weather and Crop Modelling

Weather based crop models are usually developed using average quantities of temperature and precipitation. Growing season-averaged weather indices have been utilized by many researchers [28] [29] [30] [31]. However, growing attention to extreme weather indices for crop modelling after observing impact of some major drought events (Russia 2011-12; United States 2013) on regional crop production and global commodity markets yield modelling [32] [33] have investigated the impact of several extreme indices on various crops over United States. Extreme precipitation indices like dry spell, precipitation intensity, and maximum five-day precipitation and temperature indices like hot days, heat waves have been considered in their study. Conditional probabilistic relationships between extreme indices and crop yield have been utilized for the assessment. It has been found that rice is sensitive only to precipitation intensity whereas longer duration dry spell, lower average precipitation intensity and reduced maximum five-day precipitation all resulted in declines of corn and soyabean yields. Spring wheat has been found to be sensitive to extreme temperature indices.Heat waves, the number of hot days, and higher maximum temperatures have shown negative impact on spring wheat yield. Longer duration heatwaves have caused decline in corn and soyabean productivity. The relationship between yields and both precipitation and temperature extreme indices have been reported as non-linear and threshold-type.

A more systemic approach has been adapted by [34] for corn yields over United States. Extreme weather indices, as defined by the CCI/CLIVAR/JCOMM Expert team on Climate Change Detection and Indices (ETCCDI) [35], have been considered as predictors in the regression models. Mutual information has been used to capture non-linear relationships between corn yield and extreme weather indices. A high degree of susceptibility of crop yield to extreme weather have been found. Extreme weather indices like summer days, Heat Wave index and Longest Wet Spell have been more informative than mean weather indices leading to their inclusion for yield modelling. The variability of weather within the growing season not captured by mean weather indices can be represented by extreme indices. For example, a season with little fluctuation in temperature throughout the growing season can have same mean growing season temperature to a season with large variations in temperature. However, the season with more fluctuating temperature may critically impact the yields due to an increased exposure to extreme conditions.

6. Summary

Climate change will continue to prevailits impact over the next few decades and beyond. This implies that extreme events resulting from climate change would remain as a big threat to agriculture and food security. In this article, we have tried to document the possible impacts of climate related extremes on the agriculture sector. However, more intensive and elaborate studies need to be done especially at local to regional level to quantify the effects of extreme events on crops. More emphasis should be given on modelling and simulation projections on the impacts of climate change to be done on various crops cultivation with respect to different locations [36]. More study is needed to distinguish between severe events caused by climate change, natural occurrences, and man-made events such as those caused by inappropriate waste management and other factors. To cope with adverse impacts of extreme events proper strategies are required for adaptation and mitigation. Some of the important recommendations have been shown in this article that can help to minimize the loss and damage caused by extreme events. Besides, there is a need to train the farmers and make them aware about the management practices required to adapt and mitigate the effects of extreme events. Farmer friendly insurance policy needs to deploy to economically compensate their losses due to climate disasters. Finally, climate smart agriculture using modern day technological advances needs to be implemented to make sustainable agriculture system.

Acknowledgement

The first author is highly grateful to University Grants Commission (UGC), India for sponsoring the research work through Dr. DS Kothari Post-Doctoral Fellowship.

References

[1] Abdallah, M. M. S., El-Bassiouny, H. M. S., & AbouSeeda, M. A., "Potential role of kaolin or potassium sulfate as anti-transpirant on improving physiological, biochemical aspects and yield of wheat plants under different watering regimes", *Bulletin of the National Research Centre*, 43(1), 1-12, 2019.

[2] GFSP (2015) Extreme weather and resilience of the global food system. Final project report from the UK–US taskforce on extreme weather and global food system resilience (The Global Food Security programme) (http://foodsecurity.ac. uk/assets/pdfs/extreme-weather-resilience-of-global-foodsystem.pdf)

[3] Puma, M. J., Bose, S., Chon, S.Y., & Cook, B.I., "Assessing the evolving fragility of the global food system", *Environmental Research Letters*, 10, 024007, 2015.

[4] S. Mazumdaru, "Calls for farm support intensify as Europe struggles with heat wave, drought", DW, July 2018. [Online], Available: https://www.dw.com/en/calls-for-farm-support-intensify-as-europe-struggles-with-heat-wave-drought/a-44902321

[5] NCA, V., "National Commission on Agriculture. Report prt V*", Ministry of Agriculture, New Delhi*, 1976.

[6] Hazaymeh, K. and Hassan, Q.K., "Remote sensing of agricultural drought monitoring: A state of art review", *AIMS Environmental Science*, 3(4), 604-630, 2016.

[7] Das S, Ashrit R, Moncrieff MW., "Simulation of a Himalayan cloudburst event", *Journal of earth system science*,115(3), 299-313, Jun.2006.

[8] Smoyer-Tomic, K. E., Kuhn, R., & Hudson, A., "Heat wave hazards: an overview of heat wave impacts in Canada", *Natural hazards*, 28(2), 465-486, 2003.

[9] Kabir, H., & Hossen, N., "Impacts of flood and its possible solution in Bangladesh", *Disaster Advances*, 12(10), 48-57, 2019.

[10] Rana, N., Singh, S., Sundriyal, Y. P., & Juyal, N., "Recent and past floods in the Alaknanda valley: causes and consequences. *Current Science*, 105(9), 1209-1212, 2013.

[11] Hirano, A., "Effects of climate change on spatiotemporal patterns of tropical cyclone tracks and their implications for coastal agriculture in Myanmar", *Paddy and Water Environment*, 1-9, 2021.

[12] Barlow, K. M., Christy, B. P., O'leary, G. J., Riffkin, P. A., & Nuttall, J. G., "Simulating the impact of extreme heat and frost events on wheat crop production: A review", *Field Crops Research*,171, 109-119, 2015.

[13] Zheng, B., Chapman, S. C., Christopher, J. T., Frederiks, T. M., & Chenu, K., "Frost trends and their estimated impact on yield in the Australian wheatbelt", *Journal of experimental botany*, 66(12), 3611-3623, 2015.

[14] Bergot M, Cloppet E, Pérarnaud V, Déqué M, Marçais B, Desprez-Loustau ML., "Simulation of potential range expansion of oak di sease caused by Phytophthora cinnamomi under climate change", *Global Change Biology*., 10(9),1539-52, Sep.2004.

[15] Wreford, A., & Adger, W. N., "Adaptation in agriculture: historic effects of heat waves and droughts on UK agriculture", *International Journal of Agricultural Sustainability*, 8(4), 278-289, 2010.

[16] Souri, A. H., Wang, H., González Abad, G., Liu, X., & Chance, K., "Quantifying the impact of excess moisture from transpiration from crops on an extreme heat wave event in the midwestern US: A top down constraint from Moderate Resolution Imaging Spectroradiometer water vapor retrieval", *Journal of Geophysical Research: Atmospheres*, 125(7), e2019JD031941, 2020.

[17] Turner, N. C., & Kramer, P. J., "Adaptation of plants to water and high temperature stress", *J. Wiley*, 1980.

[18] John, J. B. S., & Christiansen, M. N., "Inhibition of linolenic acid synthesis and modification of chilling resistance in cotton seedlings", *Plant physiology*, 57(2), 257-259, 1976.

[19] Desoky, E. S. M., Tohamy, M. R. A., Eisa, G. S. A., & El-Sarkassy, N. M., "Effect of some antitranspirant substances on growth, yield and flag leaf structure of wheat plant (Triticum aestivum L.) grown under water stress conditions", *Zagazig J. Agric. Res.*, 40, 223-233, 2013.

[20] Glover, J., "The apparent behaviour of maize and sorghum stomata during and after drought", *The Journal of Agricultural Science*, 53(3), 412-416, 1959.

[21] Li, H.C., Hsiao, Y.H., Chang, C.W., Chen, Y.M., & Lin, L.Y., "Agriculture Adaptation Options for Flood Impacts under Climate Change - A Simulation Analysis in the Dajia River Basin", *Sustainability*, 13, 7311, 2021. https://doi.org/ 10.3390/su13137311

[22] Kumar, A., Brahmanand, P.S., & Nayak, A.K., "Management of Cyclone Disaster in Agriculture Sector in Coastal Areas", *Directorate of Water Management, Chandrasekharpur Bhubaneswar*, 108, 2014.

[23] Rao, V.U.M., Bapuji Rao, B., Sikka, A K., Subba Rao, A.V M., Rajbir Singh and Maheswari, M., "Hailstorm threat to Indian agriculture: A historical perspective and future strategies", *Central Research Institute for Dryland Agriculture, Hyderabad -500 059*, 44, 2014.

[24] Vento, Domenico, and Andrea Malossini. *La difesa attiva contro la grandine*. Edagricole, 1982.

[25] van der Velde, M., Wriedt, G., & Bouraoui, F., "Estimating irrigation use and effects on maize yield during the 2003 heatwave in France", *Agriculture, Ecosystems & Environment*, 135(1-2), 90-97, 2010.

[26] Davin, Edouard, Sonia I. Seneviratne, Philippe Ciais, Albert Olioso, and Tao Wang., "Climate benefits of changes in agricultural practices in the context of heat wave mitigation", *In AGU Fall Meeting Abstracts*, vol. 2014, GC14B-04, 2014.

[27] Hazarika, T. K., "Climate change and Indian horticulture: opportunities, challenges and mitigation strategies", *International Journal of Environmental Engineering and Management*, 4(6), 629-630, 2013.

[28] Urban, D., Roberts, M. J., Schlenker, W., and Lobell, D. B., "Projected temperature changes indicate significant increase in interannual variability of US maize yields", *Climatic Change*, 112, 525–533, 2012. doi: 10.1007/s10584-012-0428-2

[29] Osborne TM, Wheeler TR., "Evidence for a climate signal in trends of global crop yield variability over the past 50 years", *Environmental Research Letters,* 4, 8(2), 024001, Apr 2013.

[30] Moore, F. C., and Lobell, D. B., "Adaptation potential of European agriculture in response to climate change", *Nature Climatic Change*, 4, 610, 2014. doi: 10.1038/nclimate2228

[31] Ray, D. K., Gerber, J. S., MacDonald, G. K., and West, P. C., "Climate variation explains a third of global crop yield variability", *Nature Communications,* 6, 5989, 2015. doi: 10.1038/ncomms6989

[32] Otto, F. E., Massey, N., Oldenborgh, G., Jones, R., and Allen, M., "Reconciling two approaches to attribution of the 2010 Russian heat wave", *Geophysical Research Letters*, 39, L04702, 2012. doi: 10.1029/2011GL050422

[33] Troy, T., Kipgen, C., and Pal, I., "The impact of climate extremes and irrigation on US crop yields", *Environmental Research Letters*, 10, 054013. 2015. doi: 10.1088/1748-9326/10/5/054013

[34] Konduri, V.S., Vandal, T.J., Ganguly, S., and Ganguly, A.R., "Data Science for Weather Impacts on Crop Yield", *Frontiers in Sustainable Food Systems*, 4, 52, 2020 doi: 10.3389/fsufs.2020.00052.

[35] Karl, T. R., Nicholls, N., and Ghazi, A., "Clivar/GCOS/WMO workshop on indices and indicators for climate extremes workshop summary", *Weather and Climate Extremes (Dordrecht: Springer)*, 3–7, 1999. doi: 10.1007/978-94-015-9265-9

[36] Durodola, O., "The Impact of Climate Change Induced Extreme Events on Agriculture and Food Security: A Review on Nigeria", *Agricultural Sciences*, 10(4), 487-498, 2019.

Brief Review on Lower Ionosphere and the Effects of Solar Flare Thereon

Sayak Chakraborty[1], Tamal Basak,[2,*]

[1]Indian Centre for Space Physics, 43 Chalantika, Garia Station Road, Kolkata 700084, India
[2]Amity University Kolkata, Major Arterial Road, Action Area II, New Town, Kolkata 700135, India

*Corresponding author: tamalbasak@gmail.com

Abstract

Lower ionospheric response to the impinging solar radiation is a very well studied and numerous scientific groups have prolific achievements out of their investigations in this field. In this article, we perform a brief and systematic review of the lower ionospheric response overall and especially, during solar flares with the help of a good number of important research articles published in last hundred years nearly. Firstly, we discuss about the experimental, theoretical and comparatively new numerical methods of lower ionospheric research. Secondly, we focus on the role of sub-ionospheric very low frequency (VLF) signal propagation effects for lower ionospheric research mainly during x-ray solar flares. Thirdly, we concentrate on the specific research works on the lower ionospheric response time delay (Δt) exclusively during solar flares. We try to look both at the chronological development in this field and the path-breaking outcomes. Finally, we summarize the discussion and add our views to it.

Keywords:*D-region Ionosphere, Solar Flare, Lower Ionospheric Response Time Delay, Very Low Frequency Wave Propagation*

1 Introduction

Solar flare is a sudden high energetic eruption from the surface of the sun. Sometimes tangling, crossing or reorganizing of magnetic field lines near sunspots causes such explosions of energy. Typically, the energy values during solar flares are measured of the order of 10^{20} Joules. A large fraction of this huge amount of energy in the form of x-ray enters into the earth's ionosphere and hits different molecules and ions present there. As a result of photo-electric effect and Compton effect as dominating ionizing mechanisms, this highly energized x-ray ends up generating thousands of free electrons in the ionosphere. Solar flares are often followed by a coronal mass ejection (CME). CME's are huge bubbles of radiation and particles from the sun. They explode into space at very high speed when the sun's magnetic field lines suddenly reorganize or get ruptured. The maximum strength of solar magnetic filed lines can go up to 200-400 Gauss near sunspots. When charged particles from a CME reach areas near earth, they interact with earth's magnetic field and can trigger intense lights in the sky of mainly polar areas. Such intense lighting effects are called auroras. Also, this sudden energy eruption causes perturbation in the ionosphere of the earth, this phenomenon is known as 'solar-ionospheric interaction'. Effect of this solar ionospheric interaction is different in different layers of the ionosphere. Depending upon the amount of energy erupted during a solar flare, it is classified into three main classes, namely, C, M and X-classes. Among them, C-class is the weakest solar flare with energy flux of the order of 10^{-6}W-m^{-2} and X-class is the strongest solar flare with energy flux of the order of 10^{-4}W-m^{-2}. These three classes of solar flares also interacts with these ionospheric layers differently. Solar-ionospheric interaction is a well-studied subject. In this article, we review on the different kind of theoretical, modeling and experimental developments done on this subject, but we focus mainly on the D-region of the ionosphere. D-region is the lowermost region of the ionosphere. It lies within the altitude range of 60-90 km. It is the thinnest in size yet the most chemically dynamic layer among all the other layers of the ionosphere. The major ionization source of D-layer is the solar Lyman-α radiation. This is the reason it gets almost vanished in the night time. The main objective of this article is to summarize different research works done by several scientific groups throughout the world. Starting from the days of early D-region research, for example, launching rocket to the latest developments on the same, for example, studying the sub-ionospheric radio signal propagation effect are presented here. This article aims to help next generation researchers to get an idea of the long spectrum of such works on this field. We divide this review article in three different sections. In Sec.2, we discuss about different methodology adopted by scientists to study D-region dynamics overall and especially during solar flare to have a comparative view among them. In Sec.3, we discuss about one of the most widely accepted and practiced methods among these which is the 'sub-ionospheric radio signal propagation effect'. We talk about its history and present aspects in this field. Sec.4 is about a very interesting phenomenon that is 'lower ionospheric response time delay (Δt)'. Since, the ionosphere does not react instantaneously to the incoming solar irradiation during solar flare, this is an wonderful topic to be studied. Here, we discuss about different research works done on Δt. Finally, we conclude with the evolution of research work on this field and discussing about the possible future work plan.

2 Investigations on Lower Ionosphere: Experimental, Theoretical, and Numerical

The D-layer of ionosphere is formed by absorbing the solar Lyman-α radiation below 85 km altitude. This Lyman-α radiation can ionise molecular nitrogen and oxygen, nitric oxide (NO), and various atoms, such as, sodium and calcium. So, [1] reported that the ionisation and formation of D-layer is explainable by Lyman-α and cosmic rays. Because, the molecular oxygen and nitrogen are also ionized by cosmic rays, but conditions on dominant ionization during to solar flares must be explained by x-rays. At that point of time, the satellite technology to observe the solar x-ray were not so popular among scientists, so they invented some other effective methods to study the x-rays. The United States Naval Research Laboratory reported an interesting method to study solar x-rays and ionospheric disturbance simultaneously by launching a rocket, which is fitted with several sensors to do thorough observation and send those data to the pre-programmed points on the earth ([2], [3], [4],[5], [6], [7], [8], [9], [10], [11]). Possibly, being the only method available to study the ionospheric disturbance, many scientists adopted this methodology for different kind of research purposes on the earth's ionosphere. Reference [12], in a review article, wrote the scientific details about this method in those days. The German V-2 was the vehicle initially used for this purpose. This vehicle was capable of going upto 60 miles (96.56 km) altitude with less play load. Later, the Aerobee sounding rocket came into service. This Aerobee sounding rocket was so updated that with a 150 pound payload, it can attain an altitude between 60 and 80 miles (96.56 and 128.75 km). References [7], [9], [10], [11], [13], [14], [15], [16] and many others mentioned about the alternative aspects of this methodology. Although, the data gathered by such method were very much accurate, but this kind of methodologies were too much costly to be done. So, scientists invented some other effective methods to probe into the ionospheric phenomenon. Reference [17] studied three important parameters of the ionosphere over Calcutta (now Kolkata), namely, (i) the rate of electron production (dN_e/dt), (ii) temperature (T), (iii) effective coefficient of recombination (α_{eff}) during half solar cycle from January 1945 to June 1950 by analysing the some ionospheric measurement apparatus. Reference [18] observed the x-ray spectrum through a high resolution rocket-borne spectograph during an M-class solar flare and recorded several hundreds of emission lines. Although, this work was not directly related to ionospheric disturbance, but the article gave an in-depth idea about the methodology adopted earlier to study solar flare. Later, the sub-ionospheric radio signal propagation effect was introduced, which is discussed in detail in Sec.3.

Along with these observational procedures, theoretical methodologies were performed in this field hand-in-hand since many years. References [19], [20], [21], [22], [23], [24] and many other scientific groups worked on the theoretical aspects of studying the ionosphere in earlier days of development of the subject. Chapman discussed about the formation of different layers of the ionosphere and distinguished between them based on the physical characteristics which is known as "Chapman Layer Formation" ([25]). Appleton-Hartree equation used to be a widely used mathematical tool that was adopted by [21], [23], [26], [27] and many others. Theoretical methods met the realistic scenario with the help of different types of theoretical modelling. These types of abstract models were based on the respective portion of the ionosphere which the scientists wanted to study. COSPAR International Reference Atmosphere (CIRA) 1965 was one of such earlier models which was developed by Committee On SPAce Research (COSPAR) in 1965. Reference [28] used CIRA 1965 model and studied the upper E-layer and F-layer of ionosphere during solar flare. They pointed out a few discrepancies in between the computed recombination coefficient rate constants for ion-atom interchange reaction obtained from ionospheric observations and the same from laboratory experiments. Reference [29] computed ionization rates (q) in the D-region and the associated chemical changes using a coupled atmospheric chemistry and diffusion model with the help of the solar x-ray observations from GOES-2 and ISEEE-3 satellites. They came up with a result of adjustments in equation rate coefficients in order to generate accurate electron density profile.

Some simulation techniques using advanced computational facilities, such as, Long Wave Propagation Capability (LWPC), GEometry ANd Tracking4 (GEANT4), Monte-Carlo were also introduced. References [30], [31], [32], [33], [34] and many others extensively used LWPC code as a tool of their lower ionospheric research. LWPC is a collection of separate and self-complete programs, which is used to simulate the sub-ionospheric VLF signal propagation characteristics. It was developed by Space and Naval Warfare System Centre, San Diego [35]. On the other hand, GEANT4 is a tool based on Monte-Carlo simulation technique. This method was not seen rigorously in this field of study beforehand. Reference [36] opted this particular simulation technique as a tool and did ab-initio calculations to compute the ionospheric ionisation rates during M and X-classes of solar flares. Another global model is the International Reference Ionosphere (IRI), which is an international project sponsored by the COSPAR and the International Union of Radio Science (URSI). This robust system produces an empirical standard model of the ionosphere based on all available data sources across the globe. Generally, the IRI provides a monthly average of the electron density (N_e), electron temperature (T_e), ion temperature (T_i), and ion composition etc over a range of ionospheric altitude (h). Many scientists popularly use this IRI model to obtain the reference/unperturbed values such ionospheric parameters to perform further investigations. Recently, [37] came up with a direct approach of studying the D-region of the ionosphere, where they numerically solved the 'electron continuity equation' ([32], [38], [39], [40], [41], [42]) to generate electron density profile ($N_e(t)$) during solar flares using the solar x-ray profile as observed by GOES-15 satellite and reported that the unperturbed values of the same are close to the standard values provided by IRI model.

3 Evolution of VLF-Lower Ionosphere Interaction in Presence of Solar Flare Effects

Study of lower ionosphere (especially, the D-region) response during a solar flare by the sub-ionospheric radio signal propagation effect is a well-studied subject now. One of the first radio propagation effect attributable to solar flares was the sudden loss of signals on HF-circuits known as the Mogel-Dellinger effect. The effect was discovered by John Howard Dellinger around 1935 and also described by the German physicist Hans Mögel in 1930. But, the HF signal is not capable of providing any scientific information regarding the dynamics of lower ionosphere. Because HF signal reflects back from much higher ionospheric altitude. Later, a notably sensitive method of monitoring the lower ionosphere enhancements in terms of ionization during the solar flare was introduced. It was the observation of the variations in signal amplitude strength (and phase) of sub-ionospheric Very Low Frequency (VLF) radio waves. It gets reflected mostly from the D-region of the ionosphere ([43], [44], [45] and many others). Depending upon the spatio-temporal variation of the D-region electron density (N_e), recombination coefficient (α_{eff}) and few other crucial parameters, the lower ionospheric disturbance was estimated using this method. Reference [43] discussed in details about the observation of the phase (and amplitude) of a continuous wave propagation within the earth-ionosphere cavity and reported that it can be an useful tool to remotely investigate the changes in all relevant parameters of lower ionosphere. He observed phase and amplitude variations of a VLF signal also during a solar flare, which is useful to deduce corresponding changes in D-layer of ionosphere. [44] did a comparative study of solar x-ray emission and VLF sudden phase anomalies. They reported that the enhancement of electron density (N_e) in D-layer during a solar flare lowers the effective VLF signal reflection height (h'). It essentially made the VLF signal profile to get perturbed during a moderate solar flare. They also claimed that, the VLF perturbation had significant solar zenith angle dependency. They reported that, this sub-ionospheric radio signal propagation effect could be a better way to analyse the altitude profile of effective recombination coefficient ($\alpha_{eff}(h)$). Earlier, similar investigation was done by [46] by rocket measurement technique, [47] by electron collision frequency model (σ), [48] by analysis of multiple frequency radio wave absorption and so on. Reference [45] studied the response of D-region during a solar flare by analysing the strength of the VLF signal and its phase variation respectively. They plotted the phase and amplitude of VLF signal for an entire day, which is showing a noticeable spike in the VLF profile during the time when the solar flare had occurred. Reference The method of sub-ionospheric radio signal propagation opened a completely new perspective for the scientists closely working in this field throughout the world. Reference [49] showed that the ionospheric recombination process is comparatively slower at the lower heights of ionosphere. Reference [50] computed the electron density profiles (N_e) from Sudden Cosmic Noise Absorption (SCNA) and VLF phase (and amplitude) propagation method. They made a comparative study between the electron densities obtained by these two methods and finally concluded that the electron density values measured by these two methods are in close agreement. Reference [51] did the opposite type analysis than the earlier case. They measured the electron density profile using full wave methods and predicted the phase/amplitude changes of VLF and some other frequencies as well at the signal receiving stations where the solar flare were observed. Hereafter, we found many scientific groups ([25], [52], [53], [54], [55], [56], [57] and many others) started using this sub-ionosphere radio signal propagation effect to study different ionospheric parameters, such as, (i) effective recombination coefficient, (ii) effective collision frequency, (iii) conductivity profile, (iv) electron density and so on during a solar flare. References [25] & [56] are such most frequently referred books by several scientific communities. Undoubtedly, these book opened so many newer dimensions of research by noting down a number of ionospheric facts from ground level investigations. Reference [54] worked on the effect of solar flare on the field intensity of VLF atmospherics. It was mainly focused on time delay, duration and magnitude during the solar flare. Reference [31] used the numerical programs on the earth-ionosphere wave-guide, such as, LWPC and ModeFinder [35], [58] (developed by Naval Ocean Systems Center (NOSC)) over a significant number of different VLF signal propagation paths. It was done to study the D-region through some traditional parameters, namely, the effective reflection height (h') and the log-linear slope of the altitude profile of electron density (β). They computed the parameters and reported that using these values in LWPC, flare-induced VLF signal phase/amplitude perturbations over a wider range of frequencies can be estimated remotely. Another similar kind of work was done earlier by [30]. They studied the similar parameters through sub-ionospheric VLF signal observations and using LWPC and ModeFinder. They used the x-ray profile as obtained from GOES satellite during solar flare and reported that the effective D-region perturbation has one-to-one correspondence with the x-ray flare intensity. But, that analysis did not take into account the detail analysis on the prediction of VLF phase and amplitude perturbations, which was done successfully by [31]. References [59] & [60] studied these parameters for an extremely intense X45-class solar flare and concluded that the D-region effective reflection height started lowering as the flare was reaching towards its peak phase and at the peak phase, the reflection height was lowest. Reference [60] reported that for a X45-class large solar flare the reflection height was found to have been lowered to 53 km (i.e., 17 km below the normal midday value of 70 km). In connection to this analysis, [61] came up with a new observation. Studying the VLF signal for all possible classes of flares, they reported that the perturbation of VLF amplitude is higher for higher classes of flares i.e., for typical X-class flares, the VLF amplitude is maximum. It indicates that the D-region reflection height (h) goes around its lowest values for X-class flares and highest for C-class flares. Reference [62] studied the effects on VLF signals in presence of the solar flares, solar storms and seismic activity as well. They reported that, studying the VLF profile could be a very important tool to study the ionospheric perturbations due to all these energetic interventions in both day and nighttime.

Reference [63] put forward some interesting set of observations regarding solar-ionosphere interaction during solar flares. They computed the electron density profile ($N_e(t)$) from VLF signal observation during 97 chosen solar flares of varying strengths. They restored the time variation of electron density profile during solar flare by integrating the electron

continuity equation with the help of solar x-ray profile from GOES-12 satellite observation to calculate the lower ionospheric response time delay (Δt). They reported the following set of observations. (i) VLF amplitude profile (ΔA) has an increasing trend with the peak values of increasing solar x-ray flux. (ii) Recombination coefficient (α_{eff}) decreases with increasing peak solar x-ray flux. (iii) D-region electron densities increase with increasing peak solar x-ray flux. Reference [63] gave a very clear idea about the measurement of lower ionospheric response time delay which is discussed in detail in Sec.4. Reference [64] took the help from the theoretical model on the basis of the wave-hop theory and the wave-guide mode theory to compute the phase and amplitude of VLF signal profile. References [32], [34], [40], [42], [65], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], [78] and many others then studied the VLF signals by either direct observations or by any theoretical modelling for different purposes. One important finding was common among almost all of them. It is the consistency in the degree of finite ionospheric perturbation in presence of solar x-rays during flares. References [66] & [67] implemented similar theories to model the VLF signal for solar eclipse also which verifies the robustness of such theories. Reference [69] observed the VLF signal during a solar flare and analyzed them to estimate the reflection height (h) and the sharpness factor (β). They reported that the reflection height (h) decreases, whereas, the sharpness factor (β) increases with the increase of solar flare strength. Moreover, the reflection height was found to be comparatively higher and sharpness factor was smaller at low latitudes than the corresponding values of the same at mid and high latitudes. The perspectives of looking into lower ionosphere remotely through the characterization VLF propagation effects especially during solar flares has evolved with time and got extensively advanced in last two decades. Reference [40] came with another new approach. They reported an inverse method of predicting a solar x-ray spectrum from VLF observations for two different classes of solar flares with satisfactory accuracy.

4 Response Time Delay of Lower Ionosphere During Solar Flares

The term "sluggishness" was coined by E. V. Appleton in the 1950s to describe the time delay between peak irradiance at solar noon and the resulting peak in ionospheric electron density. In our case of discussion, the definition of lower ionospheric response time delay (Δt) is discussed in the introduction part of the article. In this section, a brief review of the significant research works done on the characteristics/behaviour of Δt is performed. Reference [49] was possibly the first work to shape the definition of this response time delay effect. Also, he established an observation that the ionospheric recombination process is comparatively slow at the lower altitudes of ionosphere. This findings eventually motivated the scientific community to further investigate it. Reference [63] is one of the prominent and recent works to directly deal with the concept of lower ionospheric response time delay (Δt). As mentioned in the earlier section, they systematically used the D-region 'electron continuity equation' to derive the mathematical expression for Δt. Reference [79] studied the effects on diurnal VLF signal profile due to a solar flare occurring simultaneously with an annular solar eclipse (ASE) and measured Δt directly as the legitimate time gap between the peaks of solar x-ray flux and VLF amplitude profiles. They reported that the time delay for their observed C1.3-class of solar flare to be nearly 8 minutes. Reference [32] computed the effective recombination coefficient (α_{eff}) and electron density ($N_e(t)$) during a set of solar flares occurred in 2011 using numerical methods and having sub-ionospheric VLF signal observation as an input. They successfully established a profile of effective recombination coefficient (α_{eff}) along with another interesting characteristics of Δt. They computed Δt's for 20 solar flares of different classes and reported that the Δt has a nature of decaying with increasing peak solar x-ray flux (ϕ_{max}). Order of magnitude of Δt values in [32] for an average C-class flare was very close to same reported by [79] although there was no effect of eclipse reported in [32]. Reference [32] reported that for C-class of solar flares the average value of Δt is of the order of 150-300 sec and for M-class, it is 80-120 sec. Reference [71] obtained this Δt by GEANT4-Monte Carlo simulation technique and established a fundamental relation of Δt with ionospheric altitude (h) and flare-energy characteristics. One important point needs to be mentioned that the definition of Δt by [71] was technically different than the same by [32]. Reference [71] defined it as the time gap between peak solar x-ray flux (ϕ_{max}) and the peak electron density ($N_{e,max}$) during a solar flare. They claimed that their computed Δt to be of the order of 32 sec for X-class and the same is 68 sec for the M-class flare. Since, the lower ionospheric electron density and associated VLF perturbations share a cause-effect relationship, they vary hand-in-hand. So, the results associated with both types of Δt's complied with each other as reported by [32]. Reference [71] reported another similar approach, but this time for response time delay associated directly to VLF propagation effects. They developed a numerical model to find the ion densities and resulting VLF signal perturbation (ΔA) during a solar flare. Further, they established the altitude dependency of the Δt. Reference [77] also found similar Δt's to be significant and reported that, this Δt should be taken into account in modeling the ionospheric influence on the Global Navigation Satellite System (GNSS) and Specific Absorption Rate (SAR) signal propagation and in calculations relevant for space geodesy. In continuation of the earlier discussion, [37] followed a direct method of computing electron density by solving the 'electron continuity equation' using numerical method and computed Δt. Their Δt was measured from peak of electron density ($N_{e,max}$) which is similar to the convention adopted by [80]. Reference [37] studied the solar zenith angle (χ) effect on Δt and uniquely reported the seasonal variation of Δt. They readily agreed to the earlier findings that Δt decreases with increasing maximum solar x-ray flux. Reference [37] computed that the values of Δt are 127 sec and 106 sec over 60° S and 30° S latitudes respectively for a C5.2-class flare. The same is $\sim$ 90 sec and $\sim$ 80 sec for a M5.2-class flare, and 26 sec and 16 sec at those same latitudes respectively but for a X4.9-class flare. Since, [37] reported a strong impact of the solar zenith angle (χ) effect on Δt, they further added that these values of Δt gets changed significantly with the change in latitudes and with change in the day of occurrence of a particular solar flare. Reference [78] gave this noble idea a new way by directly measuring ionospheric electron density

from HF observations from 'riometers' and 'SuperDARN' radars, and estimated Δt from this observed electron density profile. They also concluded with the similar results as obtained by [37].

5 Summery and Conclusion

This article reviews a range of lower ionosphere related research works as follows. (i) Various methods of investigation of lower ionosphere overall and also in presence of effects of the solar flares. (ii) Sub-ionospheric VLF signal as a tool to model ionospheric perturbation during solar flares. (iii) Analysis of the response time delay of lower ionosphere during solar flares. We started the discussion with the work done way back in 1931, when mainly theoretical methodologies were available to investigate the solar-ionospheric interaction. We made a survey of the works done till 2021. In the earlier days, scientists started with some fundamental research ideas regarding ionosphere which they had started from the scratch. These fundamental works paved the path of the future scientists to work on different aspects of ionosphere. References [19], [20], [22], [24] gave some fundamental concepts that many scientists have been adopting till date for framing their research problems. Starting from the theoretical methodologies, this article discusses about different other experimental techniques and numerical methodologies that came into picture, such as, study of ionosphere using rocket measurements, sub-ionospheric radio signal propagation effects, different computer simulation techniques and solution of core ionospheric equations using numerical tools.

In section 2, we focused mainly on different methodologies which were adopted for studying the ionosphere during various solar energetic events. We discussed that the high cost of rocket launching experiment method could not stand friendly to the future generation scientists because of its limited measurement accuracy. Alternatively, the method of sub-ionospheric radio signal propagation effect analysis was discovered which became a very effective and important methodology which is extensively used till now. Apart from the study of sub-ionospheric radio signal propagation effect, several computer simulation techniques, such as, CIRA 1965, LWPC, GEANT4 and many more came into the picture. These codes are extensively used by several scientific communities to simulate various ionospheric conditions and successfully verify respective experimental observations. Lastly, the numerical methodology is one such tool that is implemented to solve multi-parametric ionospheric equations using suitable initial conditions depending on the type of perturbations using computer programming.

Section 3 reviews the very popular methodology that is sub-ionospheric radio signal propagation effect. Primarily, the lower ionospheric investigations by this method was mainly focused into the phase and amplitude distortions of radio signals during solar flare. Later [25], [52], [53], [54], [55], [57] and many others started using this sub-ionosphere radio signal propagation effect to study different ionospheric parameters, such as, effective recombination coefficient (α_{eff}), effective collision frequency (ν), conductivity profile (ρ), electron density (N_e) and so on during a solar flare or otherwise. With the further advancement scientific resources, such as, the earth-ionosphere wave-guide (EIWG) programs (LWPC and ModeFinder), this methodology went through a vast development nearly in 1990's. But surprisingly in this survey, we hardly found any other experimental methodology to probe the lower ionosphere during 1980-1990. We further note that [63] opened comparatively many new aspects in this well studied area of lower ionospheric research. Their analysis on lower ionospheric response time delay (Δt) established a new way of research on lower ionospheric response due to solar energetic events.

In section 4, we look at the research works done on lower ionospheric response time delay (Δt). Among the entire range of works on Δt especially during solar flares, we discuss about the contributions by [32], [37], [71], [77], [78], [79], [80], and a few other scientific groups. Almost from all of the outcomes of these articles, we conclude that there is a harsh tendency of Δt to decrease with increasing solar X-ray flux. Some of these articles even tried to give an initial idea about an empirical equation or statistical representation to clarify such nature of Δt along with its seasonal and latitudinal variation.

Finally, the study of lower ionosphere and the impact of solar flare on it is a very enriched field in terms of theoretical, experimental and numerical research, as we explored through the brief review done in this article. Evidently, the methods of research, namely, the theoretical, experimental and numerical are interdependent. We believe that this systematic review will definitely help many scientific groups. Especially, it will give the idea about the history of the subject to the newcomers, so that they can frame their goal, such as, development of some advanced model for D-region using high performance computational facilities etc, and elevate the subject to a newer height in the coming days.

Acknowledgements

Authors thank the editors for giving the opportunity to contribute in the edited book, '*Advances in Modern and Applied Sciences: A Collection of Research Reviews on Contemporary Fields (Volume 1)*'. They also thank all the authors, contributors, journals and publishers of the articles used here and references therein. Sayak Chakraborty acknowledges the support of DST-INSPIRE fellowship, Department of Science and Technology, India (Application Reference No. DST/INSPIRE/03/2021/001103; IF No. IF200266).

References

[1] Nicolet, M., & Aikin, A. C., "The formation of the D region of the ionosphere", *Journal of Geophysical Research*, 65(5), 1469-1483, May.1960.

[2] "Upper Atmosphere Research Report 1", *Naval Research Laboratory*, Rep. R-2955, 1946a

[3] "Upper Atmosphere Research Report 2", *Naval Research Laboratory*, Rep. R-3030, 1946b

[4] "Upper Atmosphere Research Report 3", *Naval Research Laboratory*, Rep. R-3120, 1947a

[5] "Upper Atmosphere Research Report 4", *Naval Research Laboratory*, Rep. R-3171, 1947b

[6] "Upper Atmosphere Research Report 5", *Naval Research Laboratory*, Rep. R-3358, 1948

[7] Durand, E., & Tousey, R., "Analysis of the first rocket ultraviolet spectra", *Astrophysical Journal Letters*, 109, 1-16, Jan.1949.

[8] Durand, E., "Rocket-sonde research at the Naval Research Laboratory, part of a chapter in The atmospheres of the Earth and the planets", *University of Chicago Press*, 1949.

[9] Krause, E. H., "High altitude research with V-2 rockets", *Proceedings of the American Philosophical Society*, 91, 430-446, Apr.1947.

[10] Newell, H. E., JR., "Exploration of the upper atmosphere by means of rockets", *The Scientific Monthly*, 64, 453-463, June.1947.

[11] Newell, H. E., JR., "Upper apmosphere research with V-2 rockets" , *Naval Research Laboratory Report*, R-3294, Feb.1948.

[12] Newell H. E., Jr., "A Review Of Upper Atmosphere Research From Rockets", *Advancing Earth and Space Science*, 31(1), 25. Feb.1950.

[13] Dellinger, J. H., "The Ionosphere", *The Scientific Monthly*, 65(2), 115-126, Aug.1947.

[14] Reifman, A., & Dow, W. G., "Dynamic Probe Measurements in the Ionosphere", *Physical Review Letters*, 76, 987, Oct.1949.

[15] Warwick, J. W., & Zirin, H., "Rocket Observation of X-Ray Emission in a Solar Flare", *Nature*, 180, 500, Sept.1957.

[16] Friedman, H., "Rocket Observations of the Ionosphere", *Proceedings of the IRE*, 47(2), 272-280, Feb.1959.

[17] Baral, S., S., & Mitra, A., P., "Ionosphere over Calcutta: Solar half-cycle January 1945–June 1950", *Journal of Atmospheric and Terrestrial Physics*, 1(2), 95-105, 1950.

[18] Acton, L., Bruner, W., M., Brown, W., A., "Rocket Spectrogram of a Solar Flare in thr 10-100 ÅRegion", *The Astrophysical Journal*, 291, 865-878, Apr.1985.

[19] Chapman, S., "The absorption and dissociative or ionizing effect of monochromatic radiation in an atmosphere on a rotating earth", *Proceedings of the Physical Society*, 43, 26, 1931.

[20] Appleton, E., V., "Fine-Structure of the Ionosphere", *Nature*, 131, 872–873, June.1933.

[21] Taylor, M., "The Appleton-Hartree formula and dispersion curves for the propagation of electromagnetic waves through an ionized medium in the presence of an external magnetic field. Part 1: curves for zero absorption", *Proceedings of the Physical Society (1926-1948)*, IOP Science, 45.

[22] Appleton, E., V., "Radio Exploration of the Ionosphere", *Nature*, 133, 793, May.1934.

[23] Taylor, M., "The Appleton-Hartree formula and dispersion curves for the propagation of electromagnetic waves through an ionized medium in the presence of an external magnetic field. Part 2: curves with collisional friction", *Proceedings of the Physical Society (1926-1948)*, IOP Science, 46.

[24] Appleton, E., V., "British radio observations during the second international polar year 1932-33", *The Royal Society Publishing*, 236, 764, Apr.1937.

[25] Mitra, A., P., *Ionospheric Effects of Solar Flares*, Reidel, Dordrecht, 1974.

[26] Sen, H., K., Wyller, A., A., "On the generalization of the Appleton-Hartree magnetoionic formulas", *Journal of Geophysical Research*, 65(12), 3931-3950, Dec.1960.

[27] Shkarofsky, I, P., "Generalized Appleton-Hartree Equation for Any Degree of Ionization and Application to the Ionosphere", *Proceedings of the IRE*, IEEE, 1857 - 1871, Dec.1931.

[28] Yonezawa, T., "Theory of formation of the ionosphere", *Space Science Reviews*, 5, 3–56, 1966.

[29] Zinn, J., Sutherland, C., D., Ganguly, S., "The solar flare of August 18, 1979: Incoherent scatter radar data and photochemical model comparisons", *Journal of Geophysical Research*, 95(D10), 16705-16718, Sept.1990.

[30] Thomson, N., R., & Clilverd, M., A., "Solar flare induced ionospheric D-region enhancements from VLF amplitude observations", *Journal of Atmospheric and Solar-Terrestrial Physics*, 63(16), 1729-1737, Nov.2001.

[31] McRae, W., M., & Thomson, N., R.,"Solar flare induced ionospheric D-region enhancements from VLF phase and amplitude observations", *Journal of Atmospheric and Solar-Terrestrial Physics* , 66(1), 77-87, Jan.2004.

[32] Basak, T., & Chakrabarti, S. K., "Effective recombination coefficient and solar zenith angle effects on low-latitude D-region ionosphere evaluated from VLF signal amplitude and its time delay during X-ray solar flares", *Astrophysics and Space Science*, 348, 315. Sept.2013.

[33] Schmitter, E., D., "Modeling solar flare induced lower ionosphere changes using VLF/LF transmitter amplitude and phase observations at a midlatitude site", *Annales Geophysicae*, 31(4), 765–773, 2011.

[34] Kumar, A., & Kumar, S., "Solar flare effects on D-region ionosphere using VLF measurements during low- and high-solar activity phases of solar cycle 24", *Earth Planets Space*, 70, 29, Feb.2018.

[35] Ferguson, A., J., "Computer Programs for Assessment of Long-Wavelength Radio Communications", Version 2.0, Technical document 3030, Space and Naval Warfare Systems Center , San Diego, 1998.

[36] Palit, S., Basak, T., Mondal, S., K., Pal, S., Chakrabarti, S., K., "Modeling of very low frequency (VLF) radio wave signal profile due to solar flares using the GEANT4 Monte Carlo simulation coupled with ionospheric chemistry", *Atmospheric Chemistry and Physics*, 13, 9159–9168, 2013.

[37] Chakraborty, S., & Basak, T., "Numerical analysis of electron density and response time delay during solar flares in mid-latitudinal lower ionosphere", *Astrophysics and Space Science*, 365, 184, Dec.2020.

[38] Whitten, R., C., Poppoff, I., G., "A model of solar-flare-induced ionization in the D region", *Journal of Geophysical Research*, 66(9), 2779-2786, Sept. 1961

[39] Rowe, J.N., Ferraro, A.J., Lee, H.S., Kreplin, R.W., Mitra, A.P., "Observations of electron density during a solar flare", *Journal of Atmospheric and Solar-Terrestrial Physics*, 32, 1609-1614, Sept.1970

[40] Palit, S., Ray, S., Chakrabarti, S.K., "Inverse problem in ionospheric science: prediction of solar soft-X-ray spectrum from very low frequency radiosonde results", *Astrophysics and Space Science*, 361, 151, Apr.2016.

[41] Palit, S., Raulin, J.P., Szpigel, S., "Response of Earth's Upper Atmosphere and VLF Propagation to Celestial X-Ray Ionization: Investigation With Monte Carlo Simulation and Long Wave Propagation Capability code", *Journal of Geophysical Research*, 123, 10224, Nov.2018.

[42] Nina, A., Čadež, V., M., Bajčetić, J., Mitrović, S., T., Popović, L., C., "Analysis of the Relationship Between the Solar X-Ray Radiation Intensity and the D-Region Electron Density Using Satellite and Ground-Based Radio Data", *Solar Physics*, 293, 1, Apr.2018.

[43] Crombie, D., D., "On the use of VLF measurements for obtaining information on the lower ionosphere (especially during solar flares)", *Proceedings of the IEEE*, 53(12), 2027-2034, Dec.1965.

[44] Chilton, C., J., & Conner, J., P., Steele, F., K., "A comparison between solar X-ray emission and VLF sudden phase anomalies", *Proceedings of the IEEE*, 53(12), 2018-2016, Dec.1965.

[45] Burgess, B., & Jones, T., B., "Solar flare effects and VLF radio wave observations of the lower ionosphere", *Radio Science*, 2(6), 619-626, June.1967.

[46] Whitten, R.C., Poppoff, I.G., Edmonds, R.S., Berning, W.W., "Effective recombination coefficients in the lower ionosphere", *Journal of Geophysical Research*, 70, 1737, April.1965.

[47] Parthasarathy, R., & Berkey, F., T., "Auroral zone studies of sudden-onset radio-wave absorption events using multiple-station and multiple-frequency data", *Journal of Geophyical Research*, 70(1), 89-98, Jan.1965.

[48] Parthasarathy, R., & Rai, D., B., "Effective Recombination Coefficient in the D Region", *Radio Science*, 1(12), 1397-1400, Dec.1966.

[49] Jones, T., B., "VLF phase anomalies due to a solar X-ray flare", *Journal of Atmospheric and Terrestrial Physics*, 33(6), 963-965, June.1971.

[50] Deshpande, S., D., Mitra, A., P., "Ionospheric effects of solar flares—IV. Electron density profiles deduced from measurements of SCNA's and VLF phase and amplitude" , *Journal of Atmospheric and Terrestrial Physics*, 34(2), 255-266, Feb.1972.

[51] Ananthakrishnan, S., Abdu, M., A., Piazza, L., R., "D-region recombination coefficients and the short wavelength X-ray flux during a solar flare", *Planetary and Space Science*, 21(3), 367-375, March.1973.

[52] Muraoka, Y., Murata, H., Sato, T., "The quantitative relationship between VLF phase deviations and 1–8 Åsolar X-ray fluxes during solar flares", *Journal of Atmospheric and Terrestrial Physics*, 39(7), 787-792, July.1977.

[53] Pant, P., Mahra, H., S., Pande, M., C., "Correlation between Observed VLF Phase Deviation & Solar X-ray Flux during Solar Flares", *NISCAIR Online Periodicals Repository*, 12(2), 40-42, Apr.1983.

[54] Bhattacharya, A., B., Bhattacharya, R., "Effect of Solar Flare on the Field Intensity of VLF Atmospherics", *NISCAIR Online Periodicals Repository*, 12(2), 56-58, Apr.1983.

[55] Wan-tong, L., "Correlative Analysis Between Sudden Phase Anomalies of VLF Signals and Solar X-ray Events", *Chinese Journal of Space Science*, 7, 3, 185, 1987

[56] Mitra, S.,K., *The Upper Atmosphere*, The Asiatic Society, Calcutta, 1992.

[57] Pant, P., "Relation between VLF phase deviations and solar X-ray fluxes during solar flares", *Astrophysics and Space Science*, 209, 297–306, Nov.1993.

[58] Ferguson, J., A., "Ionospheric model validation at VLF and LF", *Radio Science* , 30(3), 775-782, May-June.1995.

[59] Thomson N., R., Waldrom I., McRae W.,M., "D-region electron densities during solar flares from VLF radio measurements", *XXVII-th General Assembly of the International Union of Radio Science*, Aug.2002.

[60] Thomson, N., R., Rodger, C., J., Clilverd, M., A., "Large solar flares and their ionospheric D region enhancements", *Journal of Geophysical Research*, 110(A6), June.2005.

[61] Grubor., D., Šulić, D., Žigman, V., "Influence of solar X-ray flares on the earth-ionosphere waveguide", *Serbian Astronomical Journal*, 171, 29-35, 2005.

[62] Kumar, M., Singh, V., Singh, B., Steinbach, P., Lichtenberger, J., Hamar, D., "Day–night, seismic, and solar flare effect on the propagation of 24 kHz sub-ionospheric VLF transmitter signals", *Physics and Chemistry of the Earth, Parts A/B/C*, 31(4-9), 416-421, 2006.

[63] Zigman, V., Grubour, D., Šulić, D., "D-region electron density evaluated from VLF amplitude time delay during X-ray solar flares", *Journal of Atmospheric and Solar-Terrestrial Physics*, 69(7), 775-792, May.2007.

[64] Pal, S., & Chakrabarti, S., K., "Theoretical models for Computing VLF wave amplitude and phase and their applications", *AIP Conference Proceedings*, 1286, 42, 2010

[65] Kolarski, A., Grubor, D., ˇSulic, D., "Diagnostics of The Solar X-Flare Impact on Lowerionosphere Through the VLF-NAA Signal Recordings", *Open Astronomy*, 20(4), 591–595, Aug.2011.

[66] Pal, S., Chakrabarti, S., K., Mondal, S., K., "Modeling of sub-ionospheric VLF signal perturbations associated with total solar eclipse, 2009 in Indian subcontinent", *Advances in Space Research*, 50(2), 196-204, July.2012.

[67] Pal, S., Maji, S., K., Chakrabarti, S., K., "First ever VLF monitoring of the lunar occultation of a solar flare during the 2010 annular solar eclipse and its effects on the D-region electron density profile", *Planetary and Space Science*, 73(1), 310-317, Dec.2012.

[68] Schmitter, E., D., "Modeling solar flare induced lower ionosphere changes using VLF/LF transmitter amplitude and phase observations at a midlatitude site", *Annales Geophysicae*, 31, 765–773, Apr.2013.

[69] Singh, A., K., Singh, A., K., Singh, R., Singh, R., P., "Solar flare induced D-region ionospheric perturbations evaluated from VLF measurements", *Astrophysics and Space Science volume*, 350, 1-9, Dec.2013.

[70] Sulic, D., M., Sreckovic V., A., "A comparative study of measured amplitude and phase perturbations of VLF and LF radio signals induced by solar flares", *Astrophysics ¿ Earth and Planetary Astrophysics*, Serb.Astron.J, no.188. 2014.

[71] Palit, S., Basak, T., Pal, S., Mondal, S., K., Chakrabarti, S., K., "Effect of solar flares on ionospheric VLF radio wave propagation, modeling with GEANT4 and LWPC and determination of effective reflection height", *2014 XXXIth URSI General Assembly and Scientific Symposium (URSI GASS)*, IEEE, 2014.

[72] Kolarski, A., & Grubor, D., "Comparative Analysis of VLF Signal Variation along Trajectory Induced by X-ray Solar Flares", *Journal of Astrophysics and Astronomy*, 36, 0, Dec.2015

[73] Šulić, D., M., Srećković, V., A., Mihajlov A., A., "A study of VLF signals variations associated with the changes of ionization level in the D-region in consequence of solar conditions", *Advances in Space Research*, Feb.2016.

[74] Macotela, E., L., Raulin, J., P., Manninen, J., Correia, E., Turunen, T., Magalhães A., "Lower Ionosphere Sensitivity to Solar X-ray Flares Over a Complete Solar Cycle Evaluated From VLF Signal Measurements", *Journal of Geophysical Research*, 122(12), 12370-12377, Oct.2017.

[75] Chakrabarti, S., K., Sasmal S., Chakraborty, S., Basak, T., Tucker, R., L., "Modeling D-region ionospheric response of the Great American TSE of August 21, 2017 from VLF signal perturbation", *Advances in Space Research*, 62(3), 651-661, Aug.2018.

[76] George, H., E., Rodger, C., J., Clilverd, M., A., Cresswell-Moorcock, K., Brundell, J., B., Thomson, N., R., "Developing a Nowcasting Capability for X-Class Solar Flares Using VLF Radiowave Propagation Changes", *Advancing Earth and Science*, 17(12), 1783-1799, Dec.2019.

[77] Nina, A., Čadež, V.M., Lakićević, M.D., Radovanović, M.M., Kolarski, A.B., Popović, L.C., "Variations in ionospheric D-region recombination properties during increase of its X-ray heating induced by solar X-ray flare", *Thermal Science: an International Journal, 2019, 23, 6 (Part B)*, Belgrade : Vinča Institute of Nuclear Sciences, , 4043-4053.

[78] Chakraborty, S., Ruohoniemi, JM., Baker, JBH., Fiori, RAD., Bailey, SM., Zawdie, KA., "Ionospheric Sluggishness: A Characteristic Time-Lag of the Ionospheric Response to Solar Flares", *Journal of Geophysical Research*, 126, e2020JA028813, Feb.2021.

[79] Maji, S., K., Chakrabarti, S., K., Mondal, S., K., "Partial Effects on VLF Data due to a Solar Flare During 2010 Annular Solar Eclipse", *AIP Conference Proceedings*, 1286, 214, 2010.

[80] Palit, S., Basak, T., Pal, S., Chakrabarti S., K., "Theoretical study of lower ionospheric response to solar flares: sluggishness of D-region and peak time delay", *Astrophysics and Space Science*, 356, 19–28, Dec.2014.

A Brief Review of ELF/VLF Reception Techniques & Experiments

Bakul Das,[1] Prabir Kr Haldar,[1*]

[1]Department of Physics, Cooch Behar Panchanan Barma University, Cooch Behar, WB, India

[*]Corresponding author: prabirkrhaldar@gmail.com

Abstract

A brief review of reception techniques of Extremely Low Frequency (ELF) and Very Low Frequency (VLF) radio signals generated by natural events or man-made communication transmitters has been presented here. Beginning from the International Geophysical Year (IGY) of 1957-58, we have reviewed the ELF/VLF receivers used in different time periods around the world. We have also reviewed software and hardware that were developed for reception of these radio signals by the time line to date. We have also presented the descriptions of modern ELF/VLF narrow band and broadband receivers like Atmospheric Weather Educational System for Observation and Modeling of Electromagnetics (AWESOME), Automated Geophysical Observatory VLF receiver (AGO-VLF receiver), UK Radio Astronomy Association-UKRAA VLF Receiver, Softpal, INSPIRE project, South Pacific Buoys, etc. Significant outcomes of the experiments and studies over the decades, carried out by received ELF-VLF radio signal are also reviewed in this study.

Keywords:*Very Low Frequency, Earth-Ionosphere Waveguide, ELF/VLF Probe, VLF Transmitter, Pre-Amplifier*

1 Introduction

Radio signals in our natural or artificial environment up to 30 kHz frequency range, have been classified into three categories e.g. Ultra Low Frequency (ULF, <3Hz), Extremely Low Frequency (ELF, 3-3000 Hz) and Very Low Frequency (VLF, 3–30 kHz) [1]. Although some researchers define the ELF band from 300 Hz to 3 kHz, according to the International Telecommunication Union (ITU) ELF radio band is fixed from 3 Hz to 30 Hz with VLF as 3 to 30 kHz and ULF as 30-300 Hz [2]. Considering all types of classifications, we refer to the radio band with frequency range 3 Hz to 30 kHz as ELF/VLF radio signals (Table 1). The radio waves in the ELF-VLF range propagate long to very large distances from their origin, through the Earth-Ionosphere Wave Guide (EIWG) in the same way as light waves propagate within a fiber-optic cable. The present review includes the transmission and reception techniques and the responses of ELF/VLF radio signals due to different types of natural phenomena e.g. earthquakes, geomagnetic storms, tropical cyclones, solar eclipses, solar flares, lightning/thunderstorms, etc. The modification of electron-ion concentration in the boundaries of EIWG during those activities has been investigated in numerous works in terms of characteristic changes measured in ELF/VLF waves. ELF/VLF signals may also be generated in nature during some of the above-said events. Among those, lightning discharges are mostly the primary source of ELF/VLF waves. On the other hand, man-made navigational VLF transmitters use narrowband VLF waves at certain frequencies which can be remotely measured by suitable receivers as these radio waves propagate to long distances with small attenuation rate (<2 dB/1000 km) from the source [3].

Table 1: Bands of radio signals and their call signs

Frequency Band terminology	Frequency Range (kHz)
Ultra Low Frequency (ULF)	< 0.003
Extremely Low Frequency (ELF)	0.003 - 3.0
Very Low Frequency (VLF)	3.0 - 30.0
Low Frequency (LF)	30.0 - 300.0

Measurements of these waves help truly to understand phenomenological spectrum to investigative remote sensing of geophysical-geochemical phenomena and the near-Earth space environment. These radio signals following the earth's magnetic lines of force can travel through the remote ionosphere and interact with the plasma particles present in the ionosphere and after partial reflections, these waves return to the ground [4]. While traveling through EIWG, ELF/VLF waves cover a wide region between Earth's surface to the ionosphere (up to 100 km from the ground), also due to large wavelength they can diffract around huge obstacles. Figure 1 represents the ELF/VLF radio wave propagation through EIWG. Any in-homogeneity or disturbances within the lithosphere-ionosphere cavity may result in abnormal variations

of ELF/VLF signal amplitudes and phases. This allows remote sensing of space weather activities and natural events from an extremely large distance with a suitable ELF/VLF radio receiver [5, 6]. Moreover, ELF/VLF waves have high penetration characteristics (up to a few 100 meters) into the Earth, so these waves can be used for subterranean prospecting and imaging [7]. Though the ELF/VLF radio waves have the characteristics to travel very long distances with low attenuation, the reflected modes face modifications due to ionospheric variations, which is one of the key factors for researchers to investigate the impacts produced in the environment during several natural phenomena. Due to the very long dimension of the wavelength of ELF/VLF radio waves (10-100 km), the use of antennas in the same dimension to receive these signals is impossible. Also due to the high fluctuation level of the received ELF/VLF signals in a few micro-Watts the receiving equipment should be a reliable one, having very stable data acquisition capability for future study. An electric or magnetic field probe equivalent to receiving antenna, connected to a specialized high gain pre-amplifier(receiver) is required to receive these weak signals for scientific investigations [6, 8–10]. The ELF/VLF receiver sometimes also called only 'broadband VLF receiver' defined to record the entire 0.003 to 30 kHz band(ELF-VLF). The narrowband VLF receivers record particular frequencies determined discretely by the user.

As mentioned earlier that radio signals in ELF/VLF range may be generated due to natural phenomena or can be transmitted by man-made transmitters. There are Vast categories of natural phenomena that are responsible for emissions of ELF/VLF electromagnetic signals in the form of radio atmospherics, whistlers, and radio noise emissions e.g. lightning-thunderstorms, tropical cyclones, hurricanes, typhoons, earthquakes, solar phenomena, volcanic eruptions, jet streams, etc. High latitudinal emissions of these signals occur due to the deposition of charged particles during the Aurora events. In this case, the radio signals in the ELF range dominate over all frequency bands. Recent and past research works showed there are presence of ELF and VLF radio signal during the preparation phase of an impending earthquakes [6, 12]. Also, cyclones/typhoons/hurricanes generate natural radio signals in different places of the earth especially in low latitudinal regions [107]. Apart from the above-mentioned sources, unstable layers of Earth's atmosphere such as magnetosphere and plasma-pause may also generate radio noise due to their boundary discontinuities [12]. However from the broad point of view most prominent emission of ELF-VLF radio signal occurs due to the lightning discharges mostly from lightning prone regions around the globe e.g. South Africa, Central and South America, and South-Eastern Asia [11, 12].

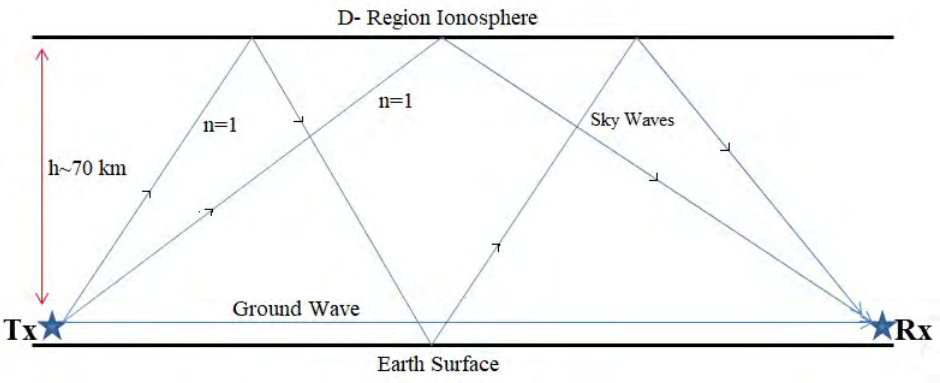


Figure 1: Representation of ELF/VLF radio wave propagation through Earth-Ionosphere Wave Guide in the form of ground and sky waves.

While natural ELF waves are generated as the slow tail of VLF component during lightning discharge event and due to resonance between the fundamental mode of the Earth-ionosphere cavity, known as Schumann resonance (7.83 Hz and its harmonics), there is a very large scale application of ELF waves in submarine communication developed by so many countries around the globe. ELF radio waves can penetrate deep in seawater, so while submerged, communication between the submarines is done by naval ELF transmitters developed in the 19th and 20th centuries in many countries e.g. USA, China, Russia, India, etc [13–16]. Some of them were shut down and some are still working, for example, the Russian Navy ELF transmitter at Murmansk on the Kola Peninsula having call sign ZEVS, operates at 82 Hz. On the other hand for the long-distance propagation of ELF/VLF waves with relatively deep penetration capability of ELF/VLF waves into seawater, a large number of VLF transmitters operate emitting radio signals from 10 to 60 kHz for naval communication with surface ships. Not only in communication purposes, for the first time in early 1970s, a network of VLF transmitters called 'Omega' has been specially designed for navigation at frequency 10-14 kHz, having accurate phase-coherence. This network stopped working in 1997 [17]. In Figure 2 we have depicted the running VLF transmitters around the globe in the year 2017 following a recent research [18]. All the VLF transmitters that are active presently (indicated by their call signs) in the Figure 2 map have frequencies above 15 kHz except the Russian RSDN (11-14 kHz) time transmitters.

A lot of efforts have been given to receive and record the ELF/VLF signals starting from the beginning of the 20th century to the modern days. It is necessary to bring those important pieces of information in a single frame that might be helpful for further development. In this study, at first, we recall some early-stage developments on the reception of ELF/VLF radio signals and then we discuss the modern era of reception techniques. We then describe some experimental results obtained by some recent ELF/VLF radio receivers.

Figure 2: Location of all the active VLF transmitters (red points) on the global map.

2 Short History of ELF-VLF Radio Signal Receiver

The earliest natural ELF/VLF signal observation in history started around early 20th century when some audible noise was observed in long telephone and transmission line and after that first reception of radio waves was made in transatlantic telecommunication during 1901-1904 by G. Marconi. After a pause of approximately 40 years, an enhanced interest in research using ELF/VLF radio signal have been observed during 1950's [19], which was again explored greatly during 1957's International Geophysical Year(IGY), with the theory of propagation of whistlers in the EIWG. Getting encouragements from these early works people focused on developing the software and the hardware necessary for ELF/VLF reception. Use of speech analysis and sound technique to visualize the ELF/VLF incoming signals in the time-frequency domain which later encouraged the use of spectrum analyzer [20]. At the same time people gave efforts to understand the propagation of these long-wave radio signals through the earth-ionosphere cavity by experiments with the radio-atmospherics during lightning events [21–23]. Some on-board satellite receivers also used to record whistler in ELF/VLF range during 1960s.VLF signals from the navigational transmitters were received during 1950-1980, as a function range or azimuth angle to obtain several characteristics like modal interference pattern, D-region ionospheric electron density profile, signal attenuation during propagation over ice, land and seawater etc. [24–26].

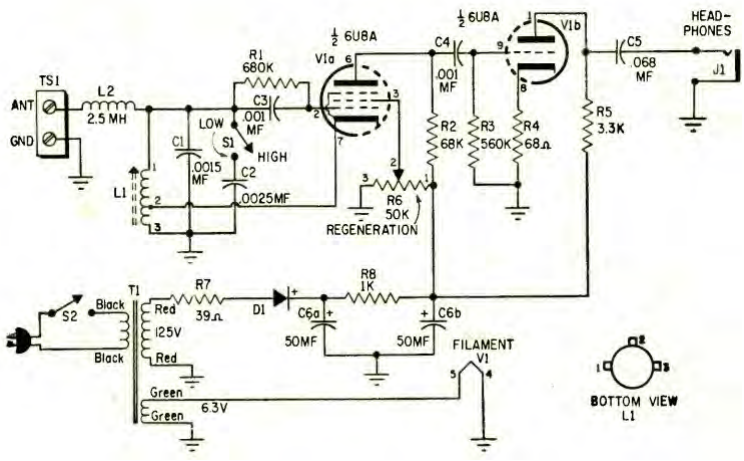


Figure 3: Tube VLF Receiver of 1970: Use of triode-pentode in VLF pre-amplifier circuit [34]

In early days, researchers mainly used large sized (as large as 45 ft length and 22.5 ft breadth) magnetic loops

of conducting wires for ELF/VLF reception at fixed location, as the sensitivity was the prior fact at that time. E. W. Paschal of STAR laboratory of the Stanford University first introduced small to medium sized air-core wire loop antenna and low noise amplifier with impedance matching transformer during the 1980's. The magnetic loops are only sensitive to the magnetic component of ELF/VLF radio waves. But the thermal noise generated due to resistance of the loop was a key limiting factor of these receivers for the sensitivity of received signals. This was solved by R.H. Rorden by introducing sensitivity as inversely proportional to $\sqrt{AM}$, where A is the area of the loop and M indicates mass of the metal wires(M) of loop [27, 28]. On the other hand, ELF/VLF electric field receivers use the capacitive coupling to sense the electric field components of VLF waves. There are very less number of published electric field receiver designs due to difficulty in calibrating and greater dependence of its noise level on front-end circuit design. There are few noticeable electric field receiver designs [29–31] given in recent past will be discussed in later sections. However, in both cases a pre-amplifier is required to amplify very feeble ELF/VLF signals before entering recording unit. Since a storage facility was needed for further scientific analysis of ELF/VLF data, people started development of magnetic-tape or chart paper recordings to archive the data for the first time just after IGY. Later, some upgraded techniques were developed to extract phase information of ELF/VLF signals in Stanford University [19, 28]. Warren K. Grubor produced 'WJ-8940B/MX' and 'WJ-8940B/ELF' receivers (Figure 4a) during 1983's and pointed out some suggestions regarding design of ELF/VLF receiver for 20Hz-10 kHz band in terms of problems faced with WJ-8940B ELF tuner [32]. He suggested to give prior concentration in power line noise suppression and reduction of local oscillator noise by improving shape factor by implementing Intermediate Frequency(IF) filters, phase noise of local oscillator and mixer balance.

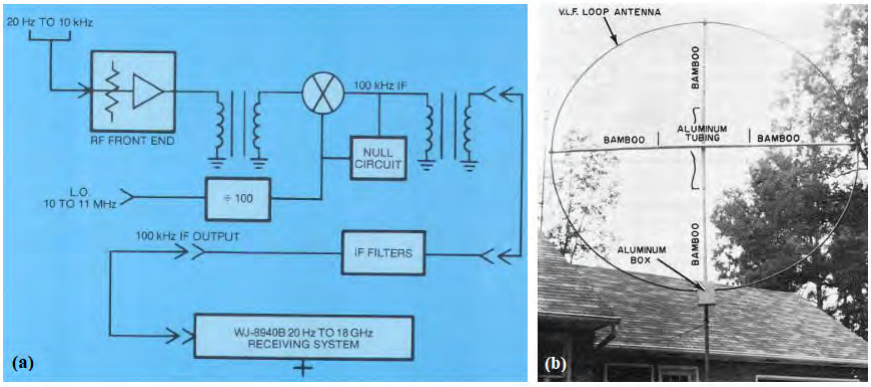


Figure 4: (a) Warren K Grubor's ELF tuner during 1983 [32] (b) Active VLF loop antenna of 1963's [33]

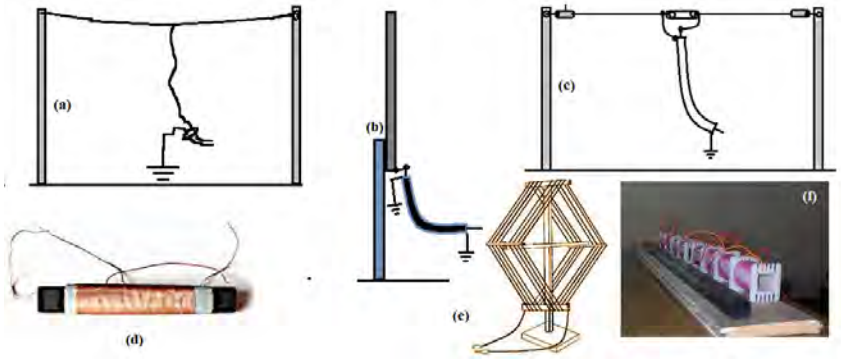


Figure 5: Various types of ELF/VLF probes have been used from begging of this field of research: E-field (a) Marconi T random wire (b) Whip antenna (c) B-field Dipole (d) Ferrite core(http://www.vlf.it) (e) Multi-turn loop(https://www.eeweb.com) (f) long induction coil of ELF detection (http://www.vlf.it)

For pre-amplifier, circuit designs from beginning includes triode-pentode, Field Effect Transistor (FET), Junction Field Effect transistors, BJT transistor amplification, operational amplifier (OP-Amp) or both combined. Circuits have been developed to attain low noise amplification which is essential for magnetic field sensors. Narrow band pre-amplifiers required filter circuit to avoid unwanted frequencies but for broadband receiver amplification with low noise is important than filtering. So many people from telecommunication, radio amateur and scientific community have built pre-amplifier designs and received ELF/ VLF radio signals for long. Some of the oldest are 1963's active VLF loop antenna (Figure 4b) given by Richard A. Genaille, 1970's One Tube VLF Receiver with a tuning range of 13-28 kHz which used 6U8A dual triode-pentode given by Hartland Smith (Figure 3). The receiver contains a tuned circuit, composed by two fixed capacitors in parallel for switching in or out signal and adjustment to the tuning frequency band. A variable inductor was also there to accomplish the tuning which was originally from a TV horizontal oscillator. The author reported excellent day and night reception of undecodable frequency shift Keying signals from Michigan

location, of NAA transmitter at 24kHz and other transmitters from different location of US, along with continuous wave decodable signals [33, 34]. After this so many low cost VLF receiver designs have evolved during last two decades (1990-2010), so many of them are from amateur-enthusiast community or from scientific community [35–38].

Few numbers of these ELF/VLF receiver designs were based on system approach and some others are only an amplifier circuit with so much limitation to use in scientific experiment. There were an ease of building prototype of circuit designs upgrading over the years although a few were available as built in kits. Some ready kit of VLF receivers like SuperSID or SolarSID were being started to distributing worldwide(mostly free of cost) specifically for educational purposes [39]. Each ELF/VLF pre-amplifiers uses a receiving antenna/probe in the form of single or multiple turn air core loop or an electric field areal (for reception of B-field ELF/VLF component). Induction coil may be of large number of turns ($\sim$70000) on a cylindrical metal core of significant length. Different types of E-field sensor has been used so far e.g. Marconi's long-wire antenna, whip antenna, dipole antenna, Active differential antenna, An Earth dipole (to listen to the interior of the earth), Ferrite antenna etc. Use of OP-Amp based pre-amplifier circuit also increased gradually depending on the amount of thermal noise created by amplifier circuit. Best op-amps in terms of VLF receptions are OP27, AD777, OP07 etc (http://www.vlf.it/).

3 Topology of Some Existing ELF/VLF Receivers

ELF/VLF receiver systems includes real-time digitization of data and storage of the data for post analysis. With The launch of the global positioning system (GPS/GNSS), real time accurate data acquisition has become possible during the start of last decade [29]. Modern broadband VLF receiver (Figure 6) consists of a B-field or E-field probe/antenna, Low noise amplifier (LNA), Analog to Digital Converter (ADC), an accurate time reference, a good quality more than one-channel sound interface and a recording/processing unit (mostly a computer). Some modern ELF/VLF receiver uses Anti-Aliasing Filter (AAF) between LNA and ADC. As already discussed this LNA is most important for ELF/VLF signal amplification and historically it is placed very closed to the detecting probe/antenna to avoid signal attenuation due to cable resistance, rest of the parts shown in Figure 5 can be placed remotely from antenna i.e. in indoor area [40]. Characteristics of some present day ELF/VLF receivers will be discussed here. Receivers like Atmospheric Weather Educational System for Observation and Modeling of Electromagnetics (AWESOME), Automated Geophysical Observatory VLF receiver (AGO-VLF receiver), UK Radio Astronomy Association-UKRAA VLF Receiver, Softpal, INSPIRE project, South Pacific Buoys (SPB-ELF/VLF Receiver) and many more are being designed and deployed for scientific data acquisition during recent years.

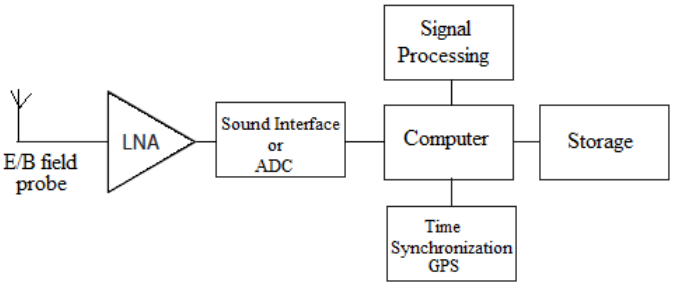


Figure 6: Block diagram of a typical present day ELF-VLF receiver

3.1 PENGUIn AGO-VLF Receiver

In around 1990's Stanford University launched an ELF/VLF receiver for combined recording of ELF/VLF broadband and narrow band signals. The receiver was a part of Stanford University's Polar Experiment Network for Geospace Upper-atmosphere Investigations (PENGUIn) project which included eight Automated Geophysical Observatories (AGOs) powered with wind generators and solar panels. The receiver used orthogonal loops(facing towards magnetic N-S and E-W directions) of two 1.7 x 1.7 m^2 square loop antennas(Figure 7) connected to dual-channel low-noise pre-amplifier unit. AGO ELF/VLF receiver sensitivity was set to 1. 89 x 10^{-4} μVolts m^{-1} $Hz^{-1/2}$ having limiting boundary due to relatively small loop antenna deployed in each observatory. Two same frequency range (1-2 kHz) of narrow-bands with two different antennas (N-S and E-W) were capable to record 'hiss' signals from two orthogonal directions for post-processing. Also each receiver consist of a digital broadband snapshot system which captured broadband data of 2s with bandwidth from 30 Hz to 10kHz in every 15 minute interval. Recording of narrow-band data done by five channels referred to as 'hiss' filters (30Hz-1kHz, 1-2kHz E-W, 1-2kHz N-S, 2-4kHz). Two additional narrow-band channels (30-40 kHz) were also tuned to record the signals of high power navigational VLF transmitters [41, 42]. Sampling rate of AGO-VLF receiver was extremely low which gives reduced storage requirements so that the system can be operated long time keeping unattended and the system was communicated by Iridium modem facility. The power consumption of AGO-VLF receiver was only $\sim$30 Watts, which allowed the system to be run by on-board power package as mentioned above. Although having a requirement of low storage, but slow sam-

pling rate gives less data quality to detect natural events. The natural events like lightning discharge, whistlers were not recorded which takes place in a time scale of $<$1s.

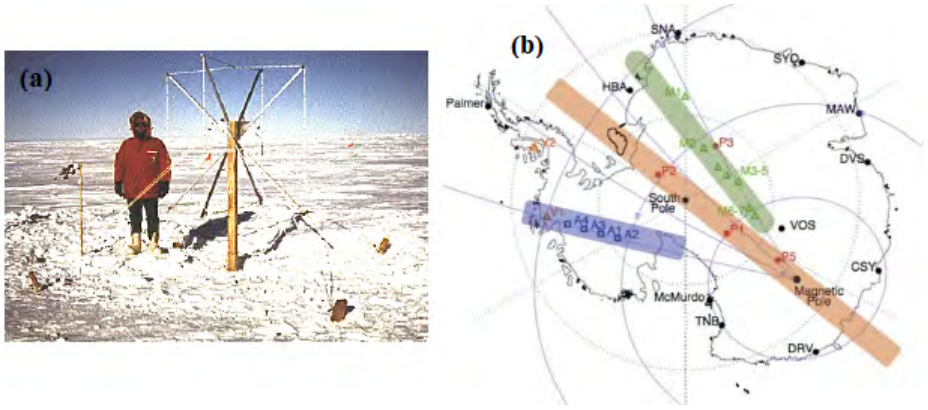


Figure 7: Snapshot showing AGO-VLF antenna (left) and the map showing some of the observatory locations in Antarctica (right) as described in Lessard M. R. et al., 2009 [42]

3.2 UKRAA VLF Receiver

The United Kingdom Radio Astronomy Association(UKRAA) is the commercial arm of the British Astronomy Association Radio Astronomy Group (BAA-RAG)(www.ukraa.com). The components of an UKRAA receiver includes a loop sensor/antenna, antenna tuning unit, receiver, and measurement device. The receiver also has optional devices like controller and signal generator for testing. For real time visualization this receiver has an ease to connect it with external voltmeter, a data-logger software for charting and storage of data. The loop antenna designed for this system is an multiturn loop of number of turns $>$100. The standard loop parameters like loop-dimension 0.4m, number of turns 125 of 24 AWG wire, Inductance 125 mH, 17.1Ω resistance, Q-factor 50 and resonant frequency $\sim$ 50 kHz are used for this system [43]. Like other B-field receivers, following Faraday's law of electromagnetic induction, this loop responses to the changing magnetic flux of ELF/VLF signals and for which an induced current appears across each turn and total current is the addition of all the turns. This loop is attached to an antenna tuning unit where an external capacitor is connected in parallel with the loop wire ends which is adjusted for resonance of the antenna. The capacitance for tuning the loop consists of several fixed capacitors connected in parallel by switches and also there is a variable capacitor to fine tune the loop to receive narrow band frequencies. The pre-amplifier(receiver) part is mainly enabled for VLF reception (10-35 kHz) which requires power supply of 15-18 volt DC.

The pre-amplifier is a simple tunable audio-amplifier which includes some functional blocks like first radio frequency(RF) amplifier, band-pass filter, second RF amplifier, detector and associated low-pass filter, and output buffers. The RF amplifier at the first stage of the receiver where the tuned signal is injected first is consists of a junction field effect transistor (J-FET) as an untuned cascade amplifier with $\sim$1.5 voltage gain and the output from this stage is taken out by a bipolar junction transistor (BJT). The cascode amplifier in this application has low voltage gain (approximately 1.5) but gain is not its only purpose. This arrangement in this stage of the receiver prevents antenna loading by the bandpass filter capacitance multiplication effects. The 1 MΩ biasing resistor at the J-FET input adjusts the input impedance of the receiver. Next stage is an 4-OPAmp band pass filter wich allows the desired band of VLF waves into the 2nd RF amplifier. The detector consists of two biased diodes in an attempt to eliminate voltage drop in the output circuit. An R-C filter is attached in the out of detector which helps to smooth the output when the signal is noisy or when sudden ionospheric disturbances (SID) occurr. Two buffers at the final output delivers the signal to computer data logger with constant voltage gain = 2 (for 0 to 5 V analog output). There is temperature sensor in the UKRAA VLF receiver and provision to install a Maxim MAX186 low resolution 12-bit analog to digital converter. These provisions allow the UKRAA receiver to be connected to a computer printer port to use computer as data logging device [40,44].

3.3 SuperSID VLF Receiver

The Stanford Solar center in collaboration with other institutes like American Association of Variable Star Observers (AAVSO) has designed a basic level and robust, easy to use Sudden Ionospheric Disturbance (SID) VLF pre-amplifier to especially monitor the space weather activity like Solar flares. It was nicknamed as SuperSID (previously SolarSID). This was a project under the United Nation's Basic space Science Initiative (UNBASSI)of International Heliophysical Year (IHY-2007) programs. This is an op-amp based pre-amplifier circuit distributed worlwide freely by the Society of Amateur Radio Astronomers (SARA) to obtain global VLF-SID data. Anyone having scientific background may ask to obtain this built in pre-amplifier kit. To record VLF radio signals, this pre-amplifier is suggested to connect with a multiple turn loop antenna (http://solar-center.stanford.edu/SID/sidmonitor/). A sound card with maximum 96 kHz sampling rate and a free automated data logger record the data with suitable resolution for which an user need a dedicated computer with minimum configuration.

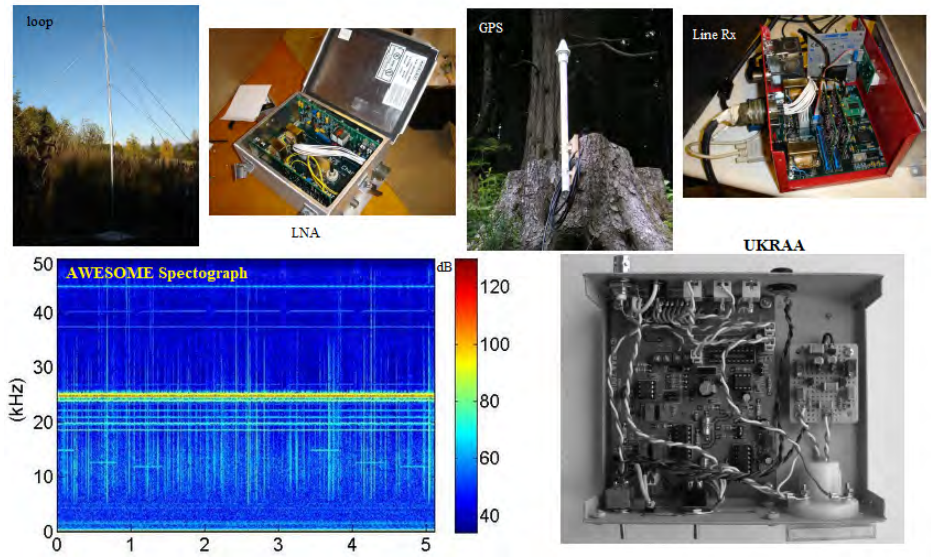


Figure 8: Snapshots showing different parts of AWESOME ELF/VLF receiver. Upper photos are loops, LNA, GPS Antenna and line receiver and lower panel shows typical ELF-VLF spectograph(obtained from Stanford Solar centre documentation [47]). Lower right black & white photo showing assembled UKRAA VLF Receiver [40]

3.4 AWESOME

The Atmospheric Weather Educational System for Observation and Modeling of Electromagnetics (AWESOME) is a ground based B-field broad band (300 Hz-50 kHz) ELF/VLF receiver consists of two orthogonal magnetic loops (for East-West and North-South radio flux) of several turns with a very high sensitivity to weak signals constructed and distributed by Stanford University [45]. It is a part of UNBASSI/International Heliophysical Year (IHY-2007) programs. ELF/VLF radio flux induces current in the loops which is then amplified by a line receiver (LNA) whose impedance is matched accurately with the loops. Both impedance and size of the antenna determines how much sensitive will be the receiver. The LNA is separated by a long multi-wire cable, which carries the induced amplified signal to the indoor line receiver (see Figure 8). This cable separates the antenna-LNA part from indoor part of the system to avoid ELF/VLF interference from high power lines, generators, and any other sources. The signal processing (includes digitization and filtration) is done by the line receiver and delivers the signal into a computer and a custom software controls the sampling rate and clock synchronization [46].

The direct Analog to digital conversion topology is followed by this receiver. In this case they used National Instruments Data Acquisition, or NI-DAQ, Card in the computer with sampling rate of 16 bit-100 kHz. The impedance matching between LNA and antennas are attained at nominal 1Ω, 1mH impedance, which is a standard used in most ELF/VLF B-field receivers [47]. The software part that the AWESOME uses is called VLF_DAQ, which have facility of both broadband and narrow-band recordings. This software was written for windows operating system. It can accumulate huge data like 1.5 GB in one hour and readily can transfer them to external memory or in some online storage [46, 47]. With so many facilities this receiver has a basic limitation regarding power consumption. Overall after deployment it required power 60 Watt to 200 Watt for which AC main signal is necessary for uninterrupted data acquisition. That means deploying this receiver is much difficult in very remote location like islands or antarctic region or in the hill areas where stable and high required power supply is not possible. Another problem is that if AC main is essential then there will be so much local hum noise around the receiver.

3.5 HAARP and SPB-ELF/VLF Receiver

HAARP stands for ”The High Frequency Active Auroral Research Program”, was a high frequency heating facility in Gakona, Alaska for studying the ionosphere. Te U.S. Air Force and U.S. Navy proposed the HAARP project in early 1990s, and the Air Force began construction in 1993. There are 180 crossed 10kW radiating elements which gave effective radiation power 300MW-3GW in the HF frequency band 2.75-10 MHz. It was developed for generation of ELF/VLF radio signal through ionospheric heating process. The ELF/VLF signal generated by these process was detected at 700km to 4400km was used to diagnose the auroral electrojet direction [48]. Received ELF/VLF signals at geomagnetic conjugate region which was injected in space were also useful to probe magnetospheric events. The HAARP generated ELF/VLF signal was used to investigate ionospheric D-region properties [49–51]. A parallel research platform called the South Pacific Buoys were deployed in March, 2007 specifically to receive ELF/VLF radio signal at the geomagnetic conjugate point, generated due to HF-heating of ionosphere during the HARRP program. SPB receivers on the buoys were mainly deployed to study the one hop, two hop propagation characteristics of ELF/VLF radio signals. The buoy receivers were capable to record both the ELF/VLF broadband and narrow band signals, having receiving probes like 6ft square vertical and 5.5 foot circular horizontal antennas providing measurement of magnetic field along three directions. The block diagram of the system is shown in Figure 9b. Like the AWESOME

receiver, the LNA device here acts in a similar way. Due to less area on the buoys the electronic parts of the receivers were attached with the antennas into a magnetic shielding (nickel-iron alloy) to avoid system generated noise. Digitization procedure of the received VLF signals was similar to the AWESOME i.e. 16-bit resolution with 100 kHz sampling rate. To make uninterrupted signal recordings, this receiving system was powered by solar panels to supply atleast 15 watt power continuously. Although the system was constructed to operate in severe weather but failed to do so in Antarctic region during half of the year without sunlight. To ensure operation the system was supplied heavy lead acid batteries but still there were problem, the below freezing point and also the battery banks were problematically heavy [52, 53].

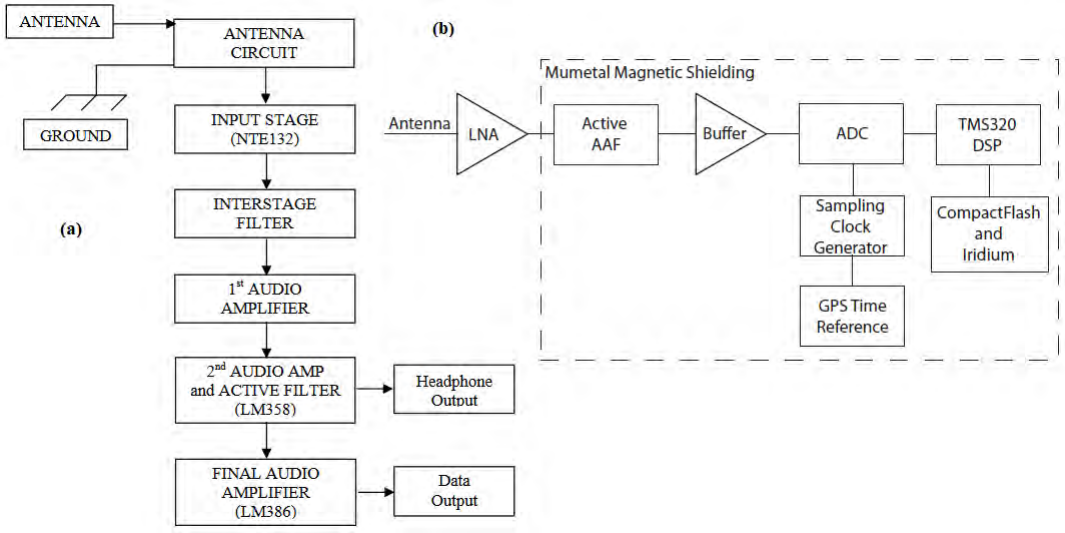


Figure 9: Block diagram showing different sections of (a) INSPIRE VLF-3 (https://theinspireproject.org/default.asp?contentID=3) and (b) HAARP-SPB-ELF/VLF receiver as described in Klein, 2010 [53]

3.6 IN-SPIRE VLF-3 Radio Receivers

The Interactive NASA Space Physics Ionosphere Radio Experiments (INSPIRE) Project was launched in 1989 with some primary stage VLF Receivers to distribute among students internationally to do ionospheric radio observations. The most upgraded VLF receivers of this project is the VLF-3 after its predecessors, RS-4 and VLF-2 receivers. VLF-2 and RS-4 were the standard receivers used by this group over a decade and in late 2006 the upgraded version was launched to provide a low cost value for distribution among students, simple and readily useful design and a short size E-field vertical probe to pick up natural radio waves. This ELF/VLF receiver was designed mainly for reception of whistler signals having electric field strength between 5μV/M to 4 mV/M, generated in mid-latitude regions. To record those signals using 1-3 meter whip good quality amplifier was required. Block diagram of INSPIRE VLF-3 receiver is shown in Figure 9a where we can see that the signals first injected in an antenna circuit, which consists of a passive circuit of inductor, capacitors, and resistances which is followed by the Input stage. In this stage the signal is amplified by field effect transistor (FET, converting very high antenna-impedance to a lower value. For the frequency under consideration (300 Hz to 20 kHz), E-Field probe has impedance as high as 30-800 MΩ. FET circuit in the input stage converts the high impedance to as low as 100-Ohm with 3dB signal gain. In the 3rd stage this receiver uses a low-pass filter which is then coupled with the first audio amplifier consists of 2N2222A transistor which supplies a 10 dB signal gain. The 4th stage is another audio amplifier comprise of two LM358 Op-Amps, first one gives an output signal with another 15 dB amplification but the second LM358 is basically a unity gain low-pass filter with a flat response in the above mentioned frequency band. Last part is another audio amplifier where a LM386 IC, coupled with preceding sections by resistors and capacitors, have been used for acquiring variable amplification facility to drive a recorder or headphone/speaker. Power supply part is maintained by 9V battery foe portable mobile use [54, 55].

3.7 Software Defined VLF Receivers

3.7.1 Softpal ELF/VLF Receiver

The Software Phase and Amplitude Logger (SoftPAL) is a software based radio receiver dedicated for the absolute phase and amplitude measurement of the VLF navigational transmitter signal up to 45 kHz. Very tiny phase and amplitude variations can be recorded by SoftPAL receiver on time scales from tens of milliseconds. A simple whip antenna of about 1.5 m long gives high gain in the MSK (Minimum Shift Keying) band (20-25 kHz) when isolated from AC power line noises. Antenna connection to this receiver is of almost pure capacitance of 10-20 pF with respect to ground. Incident static charges on the capacitive antenna is drained off readily by through 10MΩ resistor of pre-amplifier in $<$ 1s. Very high voltage $\sim \pm$1000V forms the lightning pulses around few hunderd meters away from antenna can be clipped to $\sim \pm$10V (http://www.lfsoftpal.com/) [56]. SoftPAL receiver with 2-bit demodulation

algorithm is the modern version of AbsPAL and OmniPAL systems where custom built DSP cards are used during the time when PCs had very slow speed. SoftPAL receiver gives a real time (GPS locked) display of broadband VLF spectrum and analysis of narrow-band data with the Labchart software data logger in Windows OS. The SoftPAL ELF/VLF receiver uses sigma-delta ADCs and a high quality digital signal processing device, and measures the time of a GPS 1 PPS signal with respect to the ADC clock at an accuracy of a few nano-seconds once in every second which is being used to phase-lock a software frequency synthesizer to the GPS PPS. GM-44UB is the GPS receiver with very less timing error ($\sim$ 25ns) is considered for SoftPAL with thermally insulated sound card crystal. Input signal from several antennas can be recorded in this receiver depending on the number of channels in the sound card, for example, a 4 input-channel sound card records signals from 3 antennas where one input is dedicated for GPS PPS signal [57].

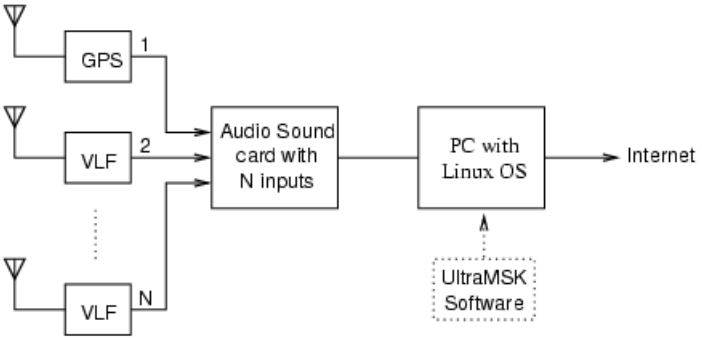


Figure 10: Block diagram showing parts of Ultramsk receiver, where N is the number of input channels. Similar figure will be for SoftPAL receiver but runs in windows OS (https://www.ultramsk.com/).

3.7.2 Ultramsk VLF Radio Receiver

Very similar kind to SoftPAL, the UltraMSK is also a software defined MSK VLF receiver but runs on linux operating system. This software based receiver has been developed by Dr. James Brundell, University of Otago, New Zealand. Here user can simultaneously measure and record the phase and amplitude of the MSK (Minimum Shift Key) modulated signals between 10-50 kHz frequency from several VLF and LF transmitters with the narrow band of 200 Hz around the main signal. VLF antenna of both E- field and B-field can be used to catch VLF navigational signals. A sophisticated N-channel sound card is a must essential to inject all the VLF signal and one GPS 1PPS time signal. The block diagram of Ultrmsk VLF receiver is shown in Figure 10. The software packages used in the Ultramsk receiver to record and plot the VLF amplitude and phase follows the VLF Software Receiver Toolkit (http://abelian.org/vlfrx-tools/notes.html). Mini computer board like Raspberry Pi is also suitable for this receiver. UltraMSK makes use of standard audio sound cards for data acquisition which requires a sound card with enough input channels to connect each of VLF antenna signals plus one additional channel to input the GPS 1 PPS signal. Thus for a single vertical electric field antenna, a standard 2 input channel sound card would be enough. For a setup with orthogonal VLF loop antennas, a multichannel card would be required. The sound card must be capable of sampling at either 48 kHz or 96 kHz and must use sigma-delta analog to digital converters(https://www.ultramsk.com/requirements/).

3.8 Other ELF/VLF Radio Receivers

Apart from the above discussed ELF/VLF receivers, there are many designs that have been published in recent times. An Indonesia based group of researchers have developed a high sensitive monolithic Op-Amp driven VLF amplifier in 2015. This receiver was designed only to receive VLF signals within 10-30 kHz with very high sensitivity which is required for long distance weak signals. It was designed to deliver its output to computer external/internal sound card to record the VLF data for post analysis. The measured sensitivity was linear upto -12 dBm (decible per miliwatt) and quadratic above -12 dBm signal level [58]. Tan and Ghanbari, 2016 designed a VLF receiver for reception of 1-20 kHz band which was a successor of their previous design. They have used two orthogonal loop as a VLF probe for N-S and E-W reception. For signal digitization, a two channel sound card was attached with this receiver and for logging phase and amplitude they have used Ultramsk software tools. The time-stamp was calibrated and synchronized with GPS 1PPS time signal [59].

Similarly, there is an increased interest in recent decades to study the techniques employed in Schumann Resonance(SR) experimental detection. Specialized sensors are necessary to improve signal to noise ratio in ELF-SR frequencies to have a good reception. A 10-30 Hz SR reception was successfully done for earthquake study by a Japan based research team during Chi-Chi earthquake in Taiwan in 1999. They have used perm alloy of 1.2 m long with copper wire of 100,000 turns in a highly sensitive ELF-preamplifier, low-pass filter of 10 and 30 Hz and an output amplifier. Group of researchers from different countries like China, Italy, Poland, Mexico have followed the similar kind of method to receive SR and other ELF signals upto 300 Hz during the last decade [60–64]. Votis et al, 2018, a Greek research group, have designed a portable ELF amplifier to receive and monitor SR (Figure 11). Six filtering circuit and amplification blocks were followed after a long induction coil. The self resonance frequency was 480 Hz,

with a low noise input signal stage and a good gain ∼95dB given a satisfactory result of reception of sixth harmonic of SRs [65]. Researchers from Wuhan University have also developed a ground based ELF/VLF digital receiver in recent time including magnetic loop antenna design, low-noise analog front-end and digital receiver for data sampling and transmission. The structure adopted in this receiver includes analog front end which gives good common-mode rejection to remove unwanted interference. Further, a field programmable gate array (FPGA) device and Universal Serial Bus (USB) architecture with real time synchronization facility have been added to this receiver [66].

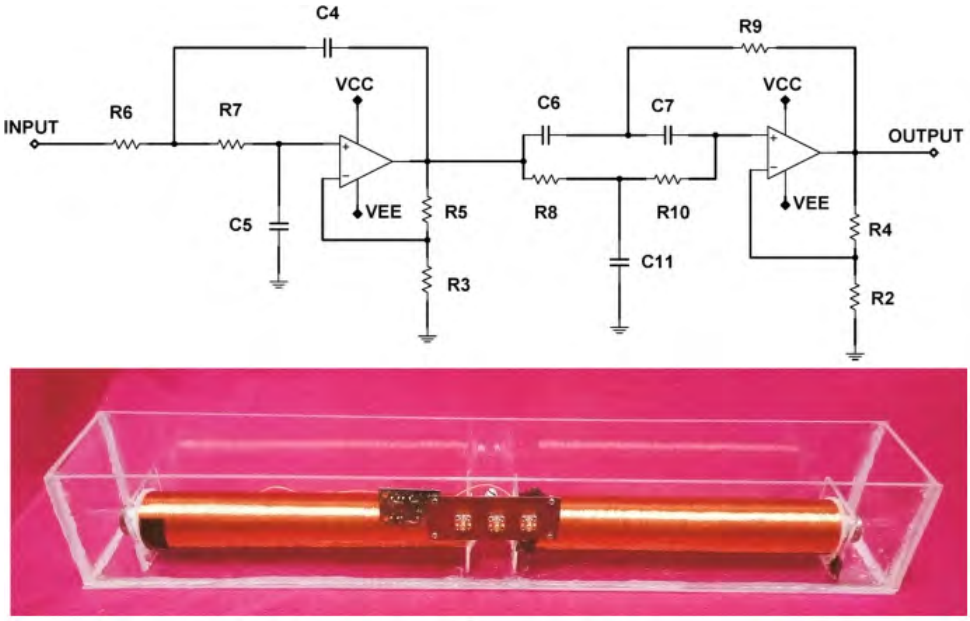


Figure 11: ELF-SR receiver. Circuit diagram (Upper) and induction coil (lower) are designed by Votis et al., 2018 [65]

4 Some Research with Ground-Based Observation through ELF/VLF Signals

One of the oldest ELF/VLF radio propagation experiment was done on transient radiation of electromagnetic signal from lightning discharges, presently known as radio-atmospherics or sferics. The vertical electric field of these signals can be as large as 1 V/m at ∼1000 km [67]. During 1950-1970s, several studies have been carried out for investigation of different types of sferics and their propagation characteristics through the earth-ionosphere wave-guide. For better determination of propagation pattern, research have been carried out in different ways to obtain photos of amplitude spectrum of sferics. By obtaining FFT of those digital photos and using narrow-band filtering process, efforts were delivered to obtain amplitude spectrum and as well as the propagation pattern of sferics pulses. The attenuation constant of the earth-ionosphere waveguide for ELF/VLF sferics propagation for the frequency band 100 Hz-12.5 kHz was calculated by Chapman and Macario, 1956 [68]. Mainly two types of sferics were detected using ELF/VLF reception technique one is slow tail and another is tweek. Effective height change of the ionosphere during solar eclipse and wave guide theory were also implemented to model tweeks-propagation through EIWG [69–73]. With the development of computer technology many studies have been done to diagnose the D-region ionosphere from sferics propagation originated from single flash or multiple flash lightning discharge events. A robust computer code called the Long Wave Propagation Capability(LWPC) was programmed to solve the sub-ionospheric VLF navigational signal propagation problems for the Naval Oceans Systems Center(NOSC) [74]. Using the ELF/VLF receiver data, Inan et al. 1993, concluded about the ionospheric heating due to lightning sferics. Cummer 1997 investigated the sprite discharges from lightning and measured change in vertical charge moment using quasi-electrostatic heating model [75, 76]. Holographic Array for Ionospheric Lightning (HAIL) was an VLF remote sensing setup to monitor ionospheric changes due to lightning discharges, covering mostly the North America. Experiments in estimation of lightning-induced electron precipitation (LEP) were done using ELF/VLF experiments in early 21st century. Clilverd et al. 2002 estimated a (600 x 1500)km area of precipitation region using multiple VLF receivers on the Antarctic peninsula [77]. Extended studies on the LEP event statistics were carried out in Stanford University using HAIL data regarding the onset delays, pole ward propagation tendency, and duration of the events similar to non-ducted whistlers [78]. In another study, Clilverd et al. were able to quantify the relation between sub-ionospheric VLF signal perturbation with the electron precipitation flux due to lightning event [79]. Early VLF events which coincides lightning stroke time were also studied using ELF/VLF technique and the experimental results showed that the VLF signal get perturbed upto 6dB [75]. Observation of lightning-sprites were made in 2003 when many sprites were observed above thunderstorms in central France. A sensitive camera and VLF radio signal of HWU received from Crete were used in this study [80]. The day time early VLF events were also studied from Suva, Fiji for the first time

with the received signals of NWC(19.8 kHz) and NPM (21.4 kHz) by Kumar et al., 2008 and from the same receiving station Kumar and Kumar, 2013 have studied forward and backward scattering of day time early VLF events with the help of NWC, NPM, VTX and NLK [81, 82].

Singh et al, 2010 installed a network of modern AWESOME ELF/VLF receivers under International Heliophysical Year 2007/United Nations Basic space Science Initiative (UNBASSI) program to study space weather and geophysical phenomena like Solar flares, lightning induced whistlers, LEPs, cosmic gamma ray flares, geomagnetic storm and their effects on D-region ionosphere [83]. Study of the effects induced in the ionosphere by several natural disturbances like solar flares, Gamma ray bursts, earthquakes, geomagnetic storms, cyclonic activity etc. have been carried out in recent decades using ELF/VLF (mostly VLF) with high-end or low cost receiving systems [84–90]. VLF navigational signal evolves as a very authentic tool to probe the ionospheric D-region during solar flares. McRae and Thomson, 2004 stated that the solar flare influence can make changes in D-region as a change in electron density height profile and these variations in the D-region can also be studied by VLF signal phase changes during a solar flare [91–93]. The electron-ion recombination effects in ionosphere and time delay in VLF signal response during solar flares were also calculated in terms of sub-ionospheric VLF navigational signal perturbation [94, 95].

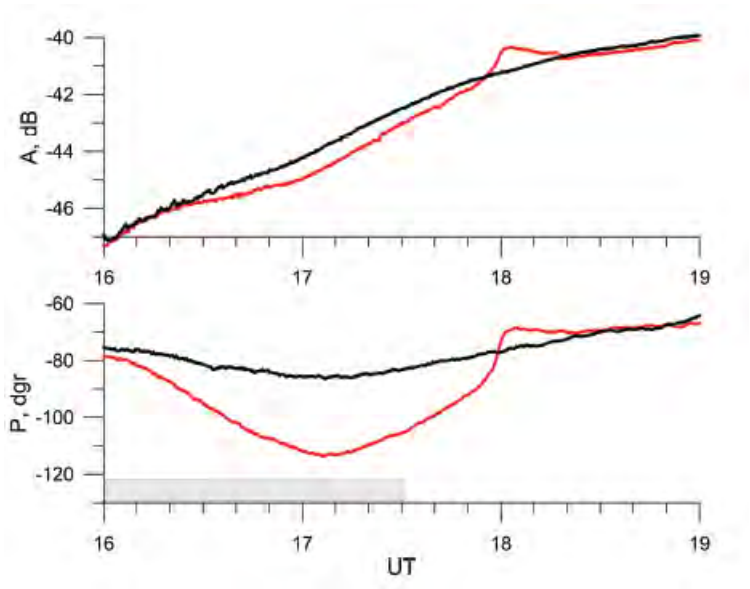


Figure 12: An example of VLF amplitude and phase disturbances during solar eclipse that detected by ultrmask receiver (Rozhnoi et. al., 2020) [96]

.

The solar eclipse, a Sun-Moon-Earth phenomena can produce a night-time situation during course of an eclipse-day and VLF signal propagation can be effected as well. The ionospheric D-region electron-ion modifications can be diagnosed by VLF signal in terms of their amplitude and phase variations during different types of solar eclipses. Starting from Bacewell 1952, many researcher have used this technique to study solar eclipse circumstances at sub-ionospheric height which is not possible by satellites [96–103]. A significant observation of amplitude and phase disturbance is obtained from Rozhnoi et. al., 2020 and shown in Figure 12. Apart from solar phenomena extra terrestrial radio emissions were also detected as evident from Mondal et al., 2012, for which they used Gyrator-II type VLF receiver to detect and study the Soft Gamma Repeater (SGR-J1550-5418) in 2009, using the VTX (18.2 kHz) communication signal from, Vijaynarayanam, India [89].

Hurricane/Typhoons or tropical cyclones are another class of natural activities that have been studied using ELF/VLF propagation techniques in recent times. Most recently Pal et. al., 2020 observed significant VLF signal amplitude disturbance during Severe cyclone 'Fani' during May, 2019 [107] is presented in Figure 13. Especially, the Atmospheric Gravity waves (AGW) generated in the troposphere during this events are believed to travel up to ionospheric height were detected and analyzed by fixed location VLF recordings. Research group from Suva, Fiji, studied 4 VLF transmitter signals having call signs NPM (21.4 kHz), NLK (24.8 kHz), NAA (24.0 kHz) and JJI (22.2 kHz) recorded at Suva, Fiji to study the AGWs generated by Tropical cyclone Evan in the December, 2012, using SoftPal VLF receiver [104]. With 41 Tropical cyclones and 27 Tropical depressions, a statistical correlation study was conducted using NAA VLF military transmitter signal amplitude received received from Belgrade (Serbia) [105]. Wavelet analysis was implemented to extract wave period of AGWs generated by the above mentioned atmospheric events and correlations between cyclone parameters and VLF amplitude fluctuation were also evident as well [106, 107]. Using VLF amplitude data of two fixed location VLF receiver(Coocbehar and Kolkata) spatial dimension of ionospheric disturbance was calculated by Das et al., 2021, during Extremely severe cyclonic storm Fani (May, 2019) [108]. Later they also showed the first time, the shifting of VLF terminator time minima of VTX and NWC signals received from CoochBehar during the Super Cyclonic Storm Amphan during May 2020 [109]. Analysis of responses of VLF sferics at discrete frequencies (4.1 kHz, 7.1 kHz and 9.1 kHz) generated by lightning discharges during the cyclones Fani and

Amphan were also done recently using two VLF receiver [110].

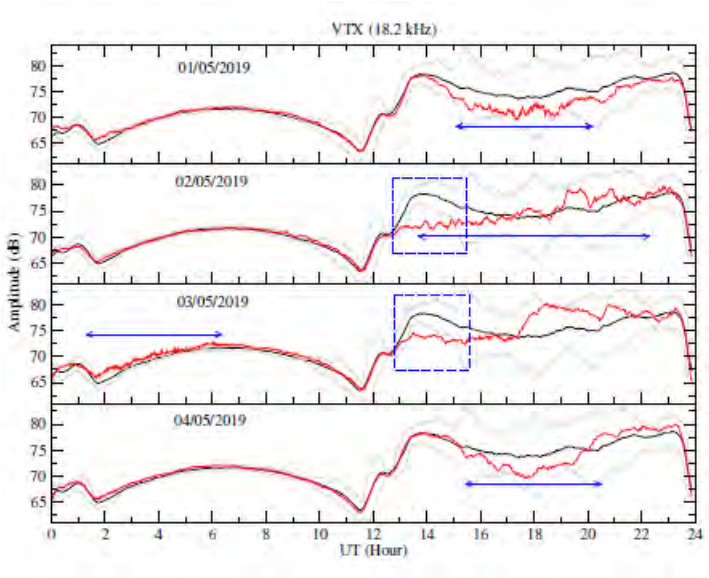


Figure 13: Another example of VTX(18.2 kHz) amplitude disturbances during Extremely Severe Cyclonic Storm Fani, over Bay of Bengal during May, 2019, received from Kolkata, India by Near Earth Space and atmospheric observatory-NESAO-ELF/VLF receiver. (Pal et. al., 2020) [107]

.

Probing geomagnetic storms by VLF technique evolved as useful starting from the study of Kikuchi, 1981 when Omega VLF transmitter signals were analyzed to estimate precipitating electrons due to mid-latitude geomagnetic storms. VLF signal amplitude and phase perturbations also helped to understand traveling ionospheric disturbances from high latitude to mid-latitude. In few cases, ELF signals were also used to understand geomagnetic storm characteristics. Low latitude ionospheric effects induced by geomagnetic storm is expected to be feeble but prominent effects were reported by a study where NWC signal received by inverted L shaped VLF antenna with suitable pre-amplifier from Agartala, India used as key diagnostic tool [111–114]. ELF/VLF signals were significantly used to study pre-seismic or simply seismic effects on the lower ionospheric boundary. Although complex but ELF/VLF method have given some promising results. During last two decades, many studies have been done to show correlation between ELF/VLF signal perturbation and occurrence of the earthquakes. ELF receivers were also used to monitor seismic radio emissions with the parallel VLF observations in some experiments [115–117]. Hayakawa and Molchanove, 2000 studied Earthquake effect on ionosphere in terms of deviation of sunrise and sunset minima-times of VLF diurnal pattern. In a very recent study, researchers have used VLF signal amplitude of VTX signal to estimate $\sim$20% electron density variation for an earthquake very near to receiving station [118, 119].

5 Summary

In this short review, we have presented ELF/VLF reception techniques from the past and present. Since IHY-1957 or before, naturally or technically generated electromagnetic radiations in 3 Hz - 30 kHz frequency band are being recorded by huge number of ground based stations. We have given our focus mainly on the ground based observation techniques to receive the ELF/VLF signal from different sources. There were large number of ELF/VLF observatories in middle and high latitude region at the starting of this science. Later, low and subtropical observations were also carried out with significant results. Reception of ELF/VLF signal through a proper receiver indicates a system with good and stable Signal to noise ratio. Propagation to large distances and strong interactions with D-region ionosphere makes the ELF/VLF radio signals a very strong and unique tool for diagnosing the variability of ionosphere influenced by different natural activities like space weather phenomena, solar eclipse, extra-terrestrial emissions, earthquakes, geomagnetic storms, and lithosphere-upper atmosphere coupling during cyclonic activities, lightning thunderstorms etc. Construction of ELF/VLF pre-amplifier from the very early time to modern days have been discussed. With time, the pre-amplifier/receiver becomes as modern as available nowadays. Use of tube diode/triode etc. for amplification of the radio signal, DAQ card for recording/storage, and loop/Marconi wire antenna (probe) have been there for long time. From 1980's, the designs of receiving system adopted the us of Operational amplifier for amplification and filtering to avoid unwanted noise. Use of sophisticated sound card for the computer and automated data acquisition were also included later. Stanford Solar center played a great role in developing and deploying modern type ELF/VLF receivers like AWESOME, SuperSID. The Ultramsk and Softpal software based receivers triggered the automated reception technique a lot. A huge number of experiments have been done with the received signals. Detection of ELF/VLF signals requires a probing element equivalent to antenna. Several types of antenna for electric or magnetic field reception have been used. Portable systems generally give preferences of the use of Ferrite rod or small loop antenna to trap those signals. For fixed location observatories, large loop or long whip is more suitable. Experimental results

in probing the ionosphere are becoming more reliable with the developments of these reception techniques. Low cost and scientifically correct receiving systems are now more achievable. In recent times, the study of ELF/VLF signal propagation both natural lightning generated radio-atmospherics and transmitter signals, led to the greater understanding of various geophysical events. Correlations between impacts of geophysical events and disturbances in ELF/VLF signals helped to interpret the mechanisms of generation of AGWs, traveling ionospheric disturbances, interaction of lithosphere/troposphere to ionosphere, lightning discharges etc.

Acknowledgements

Authors thankful to the Science & Technology and Biotechnology Department of Govt. of West Bengal, India, for their financial support(Sanction Memo no. 917(Sanc.)/STBT-11012(20)/42/2019-ST SEC). Authors also thank the editors for reviewing the article.

References

[1] Dea, J.Y., Hansen, P.M., Boerner, W. M., "Low-frequency (ULF/ELF/VLF) radio polarimetry and some applications," In Radar Polarimetry *International Society for Optics and Photonics: Bellingham*, Proc. SPIE 1748, Feb.1993. 23–30. https://doi.org/10.1117/12.140621

[2] International Telecommunication Union. ITU-R Recommendation V.431-7: Nomenclature of the frequency and wavelength bands used in telecommunications. Geneva, May 2000.

[3] Taylor, W.L., "VLF attenuation for east-west and west-east daytime propagation using atmospherics," *Journal of Geophysical Research*, 65, 7, 1933-1938, 1960.

[4] Carpenter, D.L., and Miller, T.R., "Ducted magnetospheric propagation of signals from the Siple, Antarctica, VLF transmitter," *Journal of Geophysical Research*, 81(A16) 2692–2700 1976.

[5] Wolf, T. G. and Inan, U. S., "Path-dependent properties of subionospheric VLF amplitude and phase perturbations associated with lightning," *Journal of Geophysical Research*, 95(A12), 1990.

[6] Hayakawa, M., Kasahara, Y., Nakamura, Y., Hobara, Y., Rozhnoi, A., Solovieva, M., Molchanov, O. A., "On The Correlation Between Ionospheric Perturbations as Detected by Subionospheric VLF/LF Signals and Earthquakes as Characterized by Seismic Intensity," *Journal of Atmospheric and Solar-Terrestrial Physics* 70 982-987 2010.

[7] Mcneil, J. D. and Labson, V. F., "Geological mapping using VLF radio field," *Electromagnetic Methods in Applied Geophysics* 2 7 Tulsa, OK Soc. Explor. Geophys., 1991.

[8] Brijraj, S., Collier, A. B., "Investigation of Anomalous Perturbations in VLF Signals and Correlation with Seismic Activity," *XXXth URSI General Assembly and Scientific Symposium (URSI GASS)* Istanbul 1-4 2010.

[9] Biagi, P. F., Righetti, F., Maggipinto, T., Schiavulli, L., Ligonzo, T., Ermini, A., Moldovan, I. A., Moldovan, A. S., Silva, H. G., Bezzeghoud, M., Contadakis, M. E., Arabelos, D. N., Xenos, T. D., Buyuksarac, A., "Anomalies Observed in VLF and LF Radio Signals on the Occasion of the Western Turkey Earthquake (Mw = 5.7) on May 19, 2011," *International Journal of Geosciences* 3(4A) 856-865 2012.

[10] Mullayarov, V., Druzhin, G., Argunov, V., Abzaletdinova, L., Mel'nikov, A., "Variations of VLF Radio Signals and Atmospherics During The Deep Earthquake With M = 8.2 Occurred on 24 May 2013 Near Kamchatka Peninsula," *Natural Science* 6(3): 144-149 2014.

[11] Kikuchi, H., *Environmental and Space Electromagnetics*, Springer, Tokyo, 1991, 8-10.

[12] Zeren, Z., Hu Yunpeng, H., Mirko, P., Shen Xuhui, S., Angelo, D. S., Rui, Y., YanYan, Y., Shufan, Z., Zhenxia, Z., Qiao, W., Jianping, H., Feng G., "The Seismic Electromagnetic Emissions During the 2010 Mw 7.8 Northern Sumatra Earthquake Revealed by DEMETER Satellite/", *Frontiers in Earth Science* 8 459 2020 doi:10.3389/feart.2020.572393

[13] China's NYC-Sized Earthquake Warning System' Array Sounds More Like A Way To Talk To Submarines; https://www.thedrive.com/the-war-zone/25728/chinas-new-york-city-sized-earthquake-warning-system-sounds-more-like-way-to-talk-to-subs

[14] U.S. Navy: Vision...Presence...Power, SENSORS - Subsurface Sensors. US Navy. Accessed 7 February 2010.

[15] ZEVS, The Russian 82 Hz ELF Transmitter: An Extreme Low Frequency transmission-system, using the real longwaves By Trond Jacobsen at ALFLAB, Halden in Norway, http://www.vlf.it/zevs/zevs.htm

[16] Latest defence news https://www.janes.com/defence-news/

[17] Swanson, E. R., "Omega," *Proc. IEEE*, 71, 10, 1140–1155, Oct. 1983.

[18] Maria, G., Dejan, V., Gottfried, S., Aleksandra, N., Detlef, K., Esko, L.,. "Meteor Observations as Big Data Citizen Science,". 2017.

[19] Helliwell, R. A., "Whistlers and Related Ionospheric Phenomena," New York: Dover, 11-17 1965.

[20] Potter, R. K. "Analysis of audio-frequency atmospherics," *Proc. IRE*, 39, 9, 1067–1069, Sep. 1951.

[21] Chapman, F. W., Jones, D. D., Todd, J. D. W., and Challinor, R. A., "Observations on the propagation constant of the Earth-ionosphere wave-guide in the frequency band 8 c/s to 16 kc/s," *Radio Sci.*, 11, 1273–1282, Nov. 1966.

[22] Budden, K., "The Wave-Guide Mode Theory of Wave Propagation," Moscow, ID: Logos Press, 1961.

[23] Wait, J. R., "Electromagnetic Waves in Stratified Media," New York: Pergamon, 1962.

[24] Weeks, K., "The ground interference pattern of very-low-frequency radio waves". *Proceedings of the IEEE* 97 (III), 100-107 1950.

[25] Bickel, J.E., Heritage, J.L., Weisbrod, S., "An experimental measurement of VLF

eld strength as a function of distance using an aircraft," *Naval Electronics Laboratory Report No. 767* 1957.

[26] Burgess, B., Jones, T. B., "The propagation of LF and VLF radio waves with reference to some systems applications," *The Radio and Electronic Engineer* 45, 47-61 1975.

[27] Paschal, E. W., "The design of broad-band VLF receivers with air-core loop antennas," *Tech. Rep STARLab, Stanford Univ.*, Stanford, CA, 1980.

[28] Cohen, M. B., Inan, U. S., and Paschal, E. W., "Sensitive Broadband ELF/VLF Radio Reception With the AWESOME Instrument," *IEEE Transactions on Geoscience and Remote Sensing*, 48, 1, 3-17, Jan. 2010, doi: 10.1109/TGRS.2009.2028334.

[29] M Fullekrug, M., "Measurement Science and Technology Wideband digital low-frequency radio receiver", *Meas. Sci. Technol.* 21, 015901, 2009.

[30] Fraser-Smith, A. C., and Helliwell, R. A., "The Stanford University ELF/VLF radiometer project: Measurement of the global distribution of ELF/VLF electromagnetic noise," *IEEE Internat. Symp. on Electromag. Compatability*, IEEE Catalog No. 85CH2116–2, 305–311, August 1985.

[31] Volkan Gurses, B., Kevin T. Whitmore, and Morris B. Cohen, "Ultra-sensitive broadband "AWESOME" electric field receiver for nanovolt low-frequency signals", *Review of Scientific Instruments* 92, 024704 2021. https://doi.org/10.1063/5.0031491

[32] Gruber, W. K., "Design strategies aid ELF/VLF receivers", *Microwaves*, 22, 75, 1983.

[33] Richard, A. Genaille, "V.L.F. Loop Antenna" *Electronics World* 69 01 p 49 1963.

[34] Hartland, S., "One-tube bottom scrapper," *Science and Electronics* 28 6 29 1971.

[35] Coyle, L., "A Modular Receiver for Exploring the LF/VLF Bands," *QST, Part 1* Nov 2008.

[36] Stokes, A., "A Gyrator Tuned VLF Receiver" *Communications Quarterly*, Spring 1994.

[37] Stokes, A., "Gyrator II - An Improved Gyrator Tuned VLF Receiver", *American Association of Variable Star Observers - Solar Division*, 10, 1, Jul. 1999.

[38] Gentges, F and Ratzlaff, S., "AMRAD Low Frequency Upconverter," *QST*, Apr 2002.

[39] Stanford Solar Center, "SID Monitors," http://solar- center.stanford.edu/SID/

[40] Reeve, W., "Application of the UKRAA Very Low Frequency Receiver System," 2010 https://www.semanticscholar.org

[41] Shafer, D. C, Brown, A. D., Trabucco, W. J. and Inan U. S., "A programmable and low power ELF/VLF receiver for automatic geophysical observatories," *Antarctic Journal of The United States*, 29, 361–362, 1994.

[42] Lessard, M. R., Weatherwax, A., Spasojevic, M., Inan, U. S., Gerrard, A. J., Lanzerotti, L. J., Ridley, A. J., Engebretson, M. J., Petit, N. J., Clauer, R., LaBelle, J., Mende, S. B., Frey, H. U., Pilipenko, V. A., Rosenberg, T. J., & Detrick, D. L., "PENGUIn multi-instrument observations of dayside high-latitude injections during the 23 March 2007 substorm," *Journal of Geophysical Research*, 114 2009.

[43] Radio Instruments and Measurements, US Department of Commerce, *National Bureau of Standards*, 1937

[44] Freeman, R. A., "Continuous Tracking of Lava Effusion Rate in a Lava Tube at Kilauea Volcano Using Very Low Frequency (VLF) Monitoring" *Electronic Theses and Dissertations* 2364. https://dc.etsu.edu/etd/2364

[45] Inan, U. S., Cohen, M., Scherrer, P. and Scherrer, D., "VLF Remote-Sensing of the Lower Ionosphere with AWESOME Receivers: Solar Flares, Lightning-induced Electron Precipitation, Sudden Ionospheric Disturbances, Sprites, Gravity Waves and Gamma-ray Flares," *2nd UN/NASA Workshop on International Heliophysical Year and Basic Space Science*, 27 November-1 December, 2006 at Indian Institute of Astrophysics, Bangalore, Abstract Book., p. 66, 2006.

[46] Scherrer, D., Cohen, M., Hoeksema, T., Inan, U., Mitchell, R. and Scherrer, P., "Distributing space weather monitoring instruments and educational materials worldwide for IHY 2007: The AWESOME and SID project," *Advances in Space Research*, 42(11) 1777–1785, 2008.

[47] Cohen, M., "Stanford university ELF/VLF receiver- Atmospheric Weather Educational System for Observation and Modeling of Electromagnetics(AWESOME)" 2006 http://solar-center.stanford.edu/SID/AWESOME/

[48] Cohen, M. B., Gołkowski, M. and Inan, U. S., "Orientation of the HAARP ELF ionospheric dipole and the auroral electrojet," *Geophys. Res. Lett.*, 35, L02806, 2008a, doi:10.1029/2007GL032424.

[49] Cohen, M. B., Inan, U. S. and Gołkowski, M., "Geometric modulation: A more effective method of steerable ELF/VLF wave generation with continuous HF heating of the lower ionosphere," *Geophys. Res. Lett.*, 35, L12101, 2008b doi:10.1029/2008GL034061.

[50] Gołkowski, M., Inan, U. S., Cohen, M. B. and Gibby, A. R.,""Amplitude and phase of nonlinear magnetospheric wave growth excited by the HAARP HF heater," *J. Geophys. Res.*, 115, A00F04, 2010, doi:10.1029/2009JA014610. Gołkowski, M., M. B. Cohen, D.

[51] Jin, G., Spasojevic, M., Cohen, M. B., Inan, U. S. and Lehtinen N. G., "The relationship between geophysical conditions and ELF amplitude in modulated heating experiments at HAARP: Modeling and experimental results," *J. Geophys. Res.*, 116, A07310, 2011 doi:10.1029/2011JA016664.

[52] Gol-kowski, M., Inan, U. S., Gibby, A. R. and Cohen, M. B., "Magnetospheric amplification and emission triggering by ELF/VLF waves injected by the 3.6 MW HAARP ionospheric heater," *Journal of Geophysical Research*, 113(A10), 2008.

[53] Klein, M. E., "Autonomous ultra-low power ELF/VLF receiver systems," 2010 https://vlf.stanford.edu/sites/default/files/publications/klein_thesis.pdf

[54] Robert Bennett "The INSPIRE VLF-3 Receiver: Theory of Operation" Las Cruces, NM https://theinspireproject.org/

[55] Taylor, B., "Data processing Techniques I use on INSPIRE data" *The INSPIRE Journal* 12-17 04 01 1995.

[56] Singh, A. K. and Singh, A. K., "Phase and amplitude perturbations observed on subionospheric VLF signal recorded at Varanasi (L = 1.07) using SoftPAL Receiver," *XXXth URSI General Assembly and Scientific Symposium 2011*, pp. 1-4, 2011. doi: 10.1109/URSIGASS.2011.6051034.

[57] User Manual, SoftPAL VLF Receiver, 2009. http://www.lfsoftpal.com/2009/

[58] Kusnandar, Kusmadi, A. Najmurrokhman, Sunubroto, Chairunnisa and A. Munir, "Development of high sensitivity amplifier for VLF receiver application," *International Conference on Electrical Engineering and Informatics (ICEEI)*, 328-331, 2015 doi: 10.1109/ICEEI.2015.7352520.

[59] Tan, L. M. and Keyvan, G., "Development of the new ELF/VLF receiver for detecting the Sudden Ionospheric Disturbances. *Acta Technica CSAV (Ceskoslovensk Akademie Ved)* 57 2016.

[60] Ohta, K., Umeda, K., Watanabe, N., Hayakawa, M., "ULF/ELF emissions observed in Japan, possibly associated with the Chi-Chi earthquake in Taiwan," *Natural Hazards and Earth System Science* 1(1/2), 37–42 2001.

[61] Sierra, F. P., Vazquez, H. S., Andrade, M. E., Mendoza, B., Rodriguez-Osorio, D., "Development of a Schumann-resonance station in Mexico: preliminary measurements," *IEEE Antennas and Propagation Magazine* 56(3), 112–119 2014.

[62] Ouyang, X., Zhang, X., Nickolaenko, A. P., Hayakawa, M. Shen, X., Miao, Y., "Schumann resonance observation in China and anomalous disturbance possibly associated with Tohoku M9.0 earthquake," *Earthq. Sci.* 26(2), 137–145 2013 https://doi.org/10.1007/s11589-013-0009-0

[63] Rossi, C., Palangio, P., Rispoli, F., "Investigations on diurnal and seasonal variations of Schumann resonance intensities in the auroral region," *Ann. Geophys.* 50(3), 301–311 2007.

[64] Fornieles-Callejón, J., Salinas, A., Toledo-Redondo, S., Portí, J., Méndez, A., Navarro, E. A., Morente-Molinera, J. A., Soto-Aranaz, C., Ortega-Cayuela, J. S., "Extremely low frequency band station for natural electromagnetic noise measurement," *Radio Sci.* 50(3), 191–201 2015.

[65] Votis, C. I., Tatsis, G., Christofilakis, V., Chronopoulos, S. K., Kostarakis, P., Tritakis V., & Repapis, C., "A new portable ELF Schumann resonance receiver: design and detailed analysis of the antenna and the analog front-end," *J Wireless Com Network 2018*, 155 2018. https://doi.org/10.1186/s13638-018-1157-7

[66] Chen, Y., Yang, G., Ni, B., Zhao, Z., Gu, X., Zhou, C., Wang, F., "Development of ground-based ELF/VLF receiver system in Wuhan and its first results," *Advances in Space Research*, 57 9 1871-1880 2016.

[67] Taylor, W. L., "Daytime attenuation rates in the very low frequency band using atmospherics," *Journal of Research of the National Bureau of Standards* 64D, 349-355 1960.

[68] Chapman, F. W., Macario, R. C. V., "Propagation of audio frequency radio waves to great distances," *Nature* 177 930-933 1956.

[69] Hepburn, F., "Waveguide interpretation of atmospheric waveforms," *Journal of Atmospheric and Terrestrial Physics*, 10, 121-135 1957b.

[70] Hepburn, F., "Classification of atmospheric waveforms" *Journal of Atmospheric and Terrestrial Physics* 12, 1-7 1958.

[71] Al'pert, Ya. L., Fligel, D. S., Michailova, G. A., "The propagation of atmospherics in the Earth-ionosphere waveguide," *Journal of Atmospheric and Terrestrial Physics* 29, 29-42 1967.

[72] Taylor, W. L., "VLF transmission loss calculated from spectral analyses of atmospherics," *Radio Science* 2, 139-145 1967.

[73] Barr, R., "The ELF and VLF amplitude spectrum of atmospherics with particular reference to the attenuation band near 3 kHz," *Journal of Atmospheric and Terrestrial Physics* 32, 977-990 1970.

[74] Ferguson, J. A., Snyder, F. P., "The segmented waveguide program for long wavelength propagation calculations," *Technical Document No. 1071, Naval Ocean Systems Center*, San Diego, California, USA 1987.

[75] Inan, U. S., Rodriguez, J. V., Idone, V. P., "VLF signatures of lightning-induced heating and ionization of the nighttime D-region," *Geophysical Research Letters* 20, 2355-2358 1993.

[76] Cummer, S. A., "Lightning and ionospheric remote sensing using VLF/ELF radio atmospherics," *PhD Thesis*, 1997. https://vlf.stanford.edu/wp-content/uploads/2010/06/cummerthesis.pdf

[77] Clilverd, M. A., Nunn, D., Lev-Tov, S. J., Inan, U. S., Dowden, R. L., Rodger, C. J., and Smith, A. J., "Determining the size of lightning-induced electron precipitation pathes," *J. Geophys. Res.*, 107(A8), 1168, 2002. doi:10.1029/2001JA000301.

[78] Peter, W. B., and Inan U. S., "On the occurrence and spatial extent of electron precipitation induced by oblique nonducted whistler waves," *J. Geophys. Res.*, 109, A12215, 2004, doi:10.1029/2004JA010412.

[79] Clilverd, M. A., Rodger, C. J. and Nunn, D., "Radiation belt electron precipitation fluxes associated with lightning," *J. Geophys. Res.*, 109, A12208, 2004, doi:10.1029/2004JA010644.

[80] Mika, A. C., Haldoupis, C., Marshall, R. A., Neubert, T. and Inan, U. S., "Subionospheric VLF signatures and their association with sprites observed during EuroSprite 2003," *J. Atmos. Sol.-Terr. Phys.*, 67, 1580–1597, 2005.

[81] Kumar, S., Kumar, A., and Rodger, C. J., "Subionospheric early VLF perturbations observed at Suva: VLF detection of red sprites in the day", *J. Geophys. Res.*, 113, A03311, 2008 doi:10.1029/2007JA012734.

[82] Kumar, S., Kumar, A., "Lightning-associated VLF perturbations observed at low latitude: Occurrence and scattering characteristics," *Earth Planet Sp* 65, 25–37 2013 https://doi.org/10.5047/eps.2012.05.019

[83] Singh, R., Veenadhari, B., Cohen, M. B., Pant, P., Singh, A. K., Maurya, A. K., Vohat, P., & Inan, U. S., "Initial results from AWESOME VLF receivers: set up in low latitude Indian regions under IHY2007/UNBSSI program," *Current Science*, 98(3), 398–405 2010. http://www.jstor.org/stable/24111589

[84] Thomson, N. R., & Clilverd, M. A., "Solar flare induced ionospheric D-region enhancements from VLF amplitude observations," *Journal of Atmospheric and Solar-Terrestrial Physics*, 63(16), 1729–1737 2001. https://doi.org/10.1016/S1364-6826(01)00048-7

[85] Mitra, A., P., "Ionospheric Efects of Solar Flares," Springer, New York 1974.

[86] Tatsuta, K., Hobara, Y., Pal, S., and Balikhin, M., "Sub-ionospheric VLF signal anomaly due to geomagnetic storms: a statistical study," *Ann. Geophys.* 33 1457- 1467 2015 https://doi.org/10.5194/angeo-33-1457-2015.

[87] Molchanov, O. A., and Hayakawa, M., "Subionospheric VLF signal perturbations possibly related to earthquakes" *Journal of Geophysical Research Atmospheres* 103(A8) 17489-17504 1998 doi: 10.1029/98JA00999

[88] Fishman, G., & Inan, U. S., "Observation of an ionospheric disturbance caused by a gamma-ray burst," *Nature*, 331, 418 1988.

[89] Mondal, S. K., Chakrabarti, S. K., Sasmal, S., "Detection of ionospheric perturbation due to a soft gamma ray repeater SGR J1550-5418 by very low frequency radio waves," *Astrophysics and Space Science*, 330, No. 2 2002.

[90] Rozhnoi, A., Solovieva, M., Levin, B., Hayakawa, M., & Fedun, V., "Meteorological effects in the lower ionosphere as based on VLF/LF signal observations," *Natural Hazards and Earth System Sciences*, 14, 2671 2014. https://doi.org/10.5194/nhessd-2-2789-2014

[91] McRae, W. M., Thomson, N. R., "Solar flare induced ionospheric D-region enhancements from VLF phase and amplitude observations," *Journal of Atmospheric and Solar-Terrestrial Physics*, 66, 77 2004a.

[92] Thomson, N., and Clilverd, M., "Solar cycle changes in daytime VLF subionospheric attenuation," *Journal of Atmospheric and Solar-Terrestrial Physics*, 62 601-608 2000 https://doi.org/10.1016/S1364-6826(00)00026-2.

[93] Zigman, V., Grubor, D., Sulic, D., "D-region electron density evaluated from VLF amplitude time delay during X-ray solar ares," *Journal of Atmo- spheric and Solar-terrestrial Physics* 69 775-792 2007. https://doi.org/10.1016/j.jastp.2007.01.012.

[94] Pal, S., and Chakrabarti, S., K., "Theoretical models for Computing VLF wave amplitude and phase and their applications," *AIP Conference Proceedings* 1286 42 2010 https://doi.org/10.1063/1.3512894

[95] Basak, T., & Chakrabarti, S., "Effective recombination coefficient and solar zenith angle effects on low-latitude D-region ionosphere evaluated from VLF signal amplitude and its time delay during X-ray solar flares," *Astrophysics and Space Science*, 348, 2013, https://doi.org/10.1007/s10509-013-1597-9.

[96] Rozhnoi, A., Solovieva, M., Shalimov, S., Ouzounov, D., Gallagher, P., Verth, G., McCauley, J., Shelyag, S., Fedun, V., "The Effect of the 21 August 2017 Total Solar Eclipse on the Phase of VLF/LF Signals", *AGU, Earth and Space*, 7, 2, e2019EA000839, Feb. 2020, doi:10.1029/2019EA000839

[97] Bracewell, R. N., "Theory of formation of an ionospheric layer below E layer based on eclipse and solar flare effects at 16 kc/sec", *Journal of Atmospheric and Terrestrial Physics*, 2, 226-235. 1952, doi 10.1016/0021-9169(52)90033-0

[98] Clilverd, M. A., Rodger, C. J., Thomson, N. R., Lichtenberger, J., Steinbach, P., Cannon, P. and Angling, M. J., "Total solar eclipse effects on VLF signals: Observation and modeling", *Radio Sci.*, 2001 36(4), 773–788 2001 doi:10.1029/2000RS002395

[99] Gupta, A. S., Goel, G. K., and Mathur, B. S., "Effect of the 16 February 1980 solar eclipse on VLF propagation", *Journal of Atmospheric and Terrestrial Physics*, 42, 907–909 1980.

[100] Lynn, K. J. W., "The total solar eclipse of 23 October 1976 observed at VLF", *Journal of Atmospheric and Terrestrial Physics*, 43, 1309–1316 1981 doi:10.1016/0021-9169(81)90156-2

[101] Guha, A., De, B. K., Roy R., and Choudhury, A., "Response of the equatorial lower ionosphere to the total solar eclipse of 22 July 2009 during sunrise transition period studied using VLF signal," *Journal of Geophysical Research* 115, A11302 2010. doi:10.1029/2009JA015101

[102] Pal, S., Chakrabarti, S. K., Mondal, S. K., "Modeling of sub-ionospheric VLF signal perturbations associated with total solar eclipse, 2009 in Indian subcontinent", *Advances in Space Research*, 50, 2, 196-204 2012, https://doi.org/10.1016/j.asr.2012.04.007.

[103] Inui and Hobara, Y., "Spatio-temporal characteristics of sub-ionospheric perturbations associated with annular solar eclipse over Japan: Network observations and modeling," *XXXIth URSI General Assembly and Scientific Symposium (URSI GASS)*, Beijing, 1-3, 2014, doi:10.1109/URSIGASS.2014.6929555

[104] Kumar, S., Amor, S. N., Chanrion, O., Neubert, T., "Perturbations to the Lower Ionosphere by Tropical Cyclone Evan in the South Pacific Region," *J. of Geophys. Res: Space Physics* 122 (8), 8720–8732 2017 https://doi.org/10.1002/2017JA024023

[105] Nina, A., Radovanovic, M., Milovanovic, B., Kovacevic, A., Bajcetic, J. P., Luka, C., "Low Ionospheric Reactions on Tropical Depressions prior Hurricanes" *Adv. Space Res* 2017 https://doi.org/10.1016/j. asr.2017.05.024.

[106] Correia, E., Tiago, L., Raunheitte, M., Valentin, J., Dino, B., DAmico, E., "Characterization of gravity waves in the lower ionosphere using VLF observations at Comand ante Ferraz Brazilian Antarctic Station," *Ann. Geophys.* 1–15 2019 https://doi.org/10.5194/angeo-2019-123

[107] Pal, S., Sarkar, S., Midya, S. K., Mondal, S. K., Hobara, Y., "Low-Latitude VLF Radio Signal Disturbances Due to the Extremely Severe Cyclone Fani of May 2019 and Associated Mesospheric Response," *J. Geo. Res. Space Phys.* 125, 5 2020 https://doi.org/10.1029/2019JA027288.

[108] Das, B., Sarkar, S., Haldar, P. K., Midya, S. K., Pal, S., "D-region ionospheric disturbances associated with the Extremely Severe Cyclone Fani over North Indian Ocean as observed from two tropical VLF stations," *Adv. Space Res.* 67 (1), 75–86 2021.

[109] Das, B., Sen, A., Pal, S., Haldar, P. K., "Response of the Sub-Ionospheric VLF Signals to the Super Cyclonic Storm Amphan: First Observation from Indian Subcontinent," *Journal of Atmospheric and Solar-Terrestrial Physics*, 220, 105668, 2021, https://doi.org/10.1016/j.jastp.2021.105668.

[110] Pal, S., Das, B., Sen, A., Barman, K., Haldar, P. K., Mondal, S. K., Midya S. K., "Impact of two tropical cyclones on the Radio Atmospherics observed using VLF receivers," *Earth and Space Science Open Archive* 2021 doi:10.1002/essoar.10507567.1

[111] Kikuchi, T., "VLF phase anomalies associated with substorm," *Mem. Natl Inst. Polar Res.*, Spec. Issue, 18, 3-22 1981.

[112] Ringlee, R. J., and Stewart, J. R., "IEEE Power Engineering Society," *IEEE Power Engineering Review*, 9, 7, c2-c2, 1989, doi: 10.1109/39.39055.

[113] Victor, U. J., Nwankwo, Chakrabarti, S. K., Ogunmodimu, O., "Probing geomagnetic storm-driven magnetosphere–ionosphere dynamics in D-region via propagation characteristics of very low frequency radio signals," *Journal of Atmospheric and Solar-Terrestrial Physics*, 145, 154-169, 2016 https://doi.org/10.1016/j.jastp.2016.04.014.

[114] Choudhury, A., De, B. K., Guha, A., & Roy, R., "Long-duration geomagnetic storm effects on the D region of the ionosphere: Some case studies using VLF signal" *Journal of Geophysical Research: Space Physics*, 120, 778–787 2015. https://doi.org/10.1002/2014JA020738

[115] Maurya, A., Venkatesham, K., Tiwari, P., Kathamana, V., Singh, R., Singh, A., Ramesh, D., 2016, "25 April 2015 Nepal Earthquake: Investigation of precursor in VLF sub-ionospheric signal: 25 April 2015 Nepal EQ and VLF Precursor," *Journal of Geophysical Research: Space Physics* 121 2016 doi:10.1002/2016JA022721.

[116] Parrot, M., Mogilevsky, M. M., "VLF emissions associated with earthquakes and observed in the ionosphere and the magnetosphere," *Physics of the Earth and Planetary Interiors*, 57, 1–2, 86-99, 1989 https://doi.org/10.1016/0031-9201(89)90218-5.

[117] Singh, R. P., Kumar, M., Singh, O. P., Singh, B., "Subsurface VLF electric field emissions associated with regional earthquakes," *Indian Journal of Radio and Space Physics*, 38. 220-226 2009.

[118] Hayakawa, M., Molchanov, O. A., "Effect of earthquakes on lower ionosphere as found by subionospheric VLF propagation," *Advances in Space Research*, 26, 8, 1273-1276, 2000 https://doi.org/10.1016/S0273-1177(99)01217-X.

[119] Das, B., Sen, A., Haldar, P. K. and Pal, S., "VLF radio signal anomaly associated with geomagnetic storm followed by an earthquake at a subtropical low latitude station in northeastern part of India," *Indian J Phys* 2021. https://doi.org/10.1007/s12648-020-01966-2

Ionospheric Effects of Cyclonic Storms: A Brief Review

Kheyali Barman[1,2], Bakul Das[1,2], Sujay Pal[3,2,*], Prabir Kumar Haldar[1,2]

[1]Department of Physics, Cooch Behar Panchanan Barma University, Cooch Behar, India
[2]Near-Earth Space and Atmospheric Observatory, Kolkata, India
[3]Department of Physics, Srikrishna College, Nadia, India
*Corresponding author: myselfsujay@gmail.com

Abstract

Mesoscale convective systems in the troposphere such as tropical cyclones that form over warm Ocean in tropical regions consisting of large-scale rotating cloud mass, have an exemplary impact on the upper atmosphere including the ionosphere. Interactions of the troposphere with the ionosphere during tropical cyclones using numerous multidisciplinary methods have now become an emerging subject of interest to the scientific community in recent years. In this article, we present a short review on the ionospheric effects of cyclonic storms as observed by various experimental techniques. We also review the proposed theoretical mechanisms responsible for linking the tropospheric cyclones to the upper atmosphere. Further, in this context, we highlight the possibilities of forecasting the cyclonic storms using ionospheric observations.

Keywords: *Cyclonic Storms; Ionosphere; Atmospheric Gravity Waves; VLF/LF waves; Atmosphere-Ionosphere Coupling;*

1. Introduction

Tropical Cyclones (TCs) are one of the most devastating natural disasters with respect to loss of lives and economic damages. Over the past few years, there has been a strong increase not only in the frequency but also in the intensity of TCs over the North Indian Ocean. All TCs have a low pressure center, known as eye, with rapid rotating wind-storm and clouds spiraling towards the eye-wall. The wind is normally calm and without clouds. A typical TC has a diameter of 200-500 km and can extend up to ~1000 km (World Meteorological Organization - WMO). TCs are mainly tropical or subtropical phenomena over sea-water with convection and wind circulation in counter-clockwise (Northern hemisphere) or clockwise (southern hemisphere) direction (WMO). This weather phenomenon has been termed in several ways by the WMO, depending on its locationand strength. TCs taking place in the Caribbean Sea, the Gulf of Mexico, the North Atlantic Ocean, and the eastern and central North Pacific Ocean are called Hurricane. In the western North Pacific and south-eastern Asia this event is termed as Typhoon, and in the Bay of Bengal and Arabian Sea it is known as cyclone. 'Severe tropical cyclone' and 'tropical cyclone' are the terms used in western South Pacific, southeast Indian Ocean and southwest Indian Ocean. Different stages of cyclonic storm are categorized in several classes. In Arabian Sea and Bay of Bengal the cyclonic storm is classified by the Indian Meteorological Department (IMD) from lowest phase depression (D) to maximum phase Super Cyclonic Storm (SuCS). This classification in Indian subcontinent has been done on the basis of maximum sustained wind speed (v in km/h) and the pressure drop (p in hPa) in the center of low pressure or in the eye of the cyclonic storm. The classification is as follows: low-pressure area (v~32 km/h, p~1.0 hPa), depression (v~32–50 km/h, p~1.0–3.0 hPa), deep depression (v~51–59 km/h, p~3.0–4.5 hPa), cyclonic storm (v~60–90 km/h, p~4.5–8.5 hPa), severe cyclonic storm (v~90–119 km/h, p~8.5–15.5 hPa), very severe cyclonic storm (v~119–165 km/h, p~15.5–39.5 hPa), extremely severe cyclonic storm (v~166–220 km/h, p~40–65.5 hPa) and super cyclonic storm (v >220 km/h, p > 65.5 hPa) [1].

Forecasting the formation, intensity change, and track of the TCs as well as understanding the physical processes affecting the intensity of TCs are of great concern in the field of meteorology and atmospheric science. Generally, it is believed that TCs are likely to form over warm Ocean when sea-surface temperature is above 26.5°C in association with low level tropical disturbances such as easterly waves or tropical cloud cluster [2]. Although various external and internal factors such as monsoon circulations, intraseasonal oscillations, the intertropical convergence zone, organization of convection, tropospheric moisture, vertical wind shear, sea surface temperature play important role in the formation of TCs [3]. Life span of TCs can be of few days to weeks and generally dissipate as it proceeds over the land orcooler sea water surface. The effects of TCs are not limited to the surface but also reached out to the stratosphere, mesosphere to ionosphere. In this report, we concentrate on the effects of TCs in the ionosphere which is the partially ionized layer of the upper atmosphere extending from 60km to ~1000 km. The ionpsphere is mainly divided into three regions, namely the D-region (60-90 km), the E-region (90-150 km) and the F-region (from 150 km to more than 500 km). All of these regions show variability in the time scale ranging from seconds, hour, and day, seasonal, annual to solar cycle response. Electron-ion distributions of the ionospheric regions are not only depends on the ionizing sources from above such as Cosmic Rays, solar UV and X-ray, high energy gamma rays but also on the various effects connected to underlying atmospheric layers such as TCs.

The first report about weather-ionosphere relationship was summarized by Mitra in 1952 [4] but was not enough conclusive. Bauer 1957, for the first time, investigated a possible relationship between virtual height and critical frequency of the ionosphere F2-layer with a frontal passage of air mass in the lower atmospheric layer through statistical analysis which is well supported by the hypothesis given by Martyn, 1950 and also consistent with the troposphere-ionosphere dynamic coupling as per the theory of atmospheric oscillation [5-7]. Further, in the next work, he analyzed the ionospheric perturbation due to four hurricanes namely Hazel (1954), Connie, Diane and Ione (1955) having tracks with closest approach to monitoring station at Washington D.C. using critical frequency and virtual height data in 1958. There were significant deviations in critical frequency and virtual height value in the post hurricane phase or during the peak hurricane phase. In the absence of any otherionospheric or extraterrestrial events the study was well suggestive of the changes due to divergent field of tropospheric pressure system [8]. These early studies opened a new path in the field of atmosphere-ionosphere connection and is now emerged as a new topic in research community during the past few decades. Recently, there has been much interest on quantifying the ionospheric effects of cyclonic storms using high resolution data and finding relationship with various atmospheric parameters so as to use the ionospheric observation techniques for now-casting of cyclonic storms. This article briefly reviews recent findings on ionospheric effects of cyclonic storms and the various hypotheses linking the tropospheric phenomenon with the ionospheric effects.

2. Effects on the F-Region Ionosphere

Detecting the upper ionospheric response due to the cyclonic effect is one of the most intriguing investigations in the ionospheric study. Several investigations showed very weak effects of TCs on the upper ionosphere. Space weather events such as solar flares, geomagnetic storms dominate over the meteorological event generated disturbances in the ionospheric heights. The F-region ionospheric variability mainly arises from the TIDs generated during cyclone from the sources at tropospheric level at the cyclone formation region. These TIDs are wave-like oscillations with periods few minutes to hour. Baker and Davies, 1969 obtained the time period of TC generated TIDs about 2-5 minutes. In other studies like Hung and Kuo, 1978 obtained time period of TIDs as about 20-90 minutes; Huang et al., 1985 obtained it as 13-14 minutes, and Xiao et al., 2007 calculated the time period of TIDs as about 20 minutes [9-12]. Although the results of Hung and Kuo, 1978 were not statistically significant i.e., they detected F-region variability only for two Typhoons out of twelve during 1982-1983 [9]. On the other hand, Xiao et al. 2007, showed the F-region irregularities due to TIDs for 22 typhoons out of 24 samples [11] for the time period of 1987-1992.

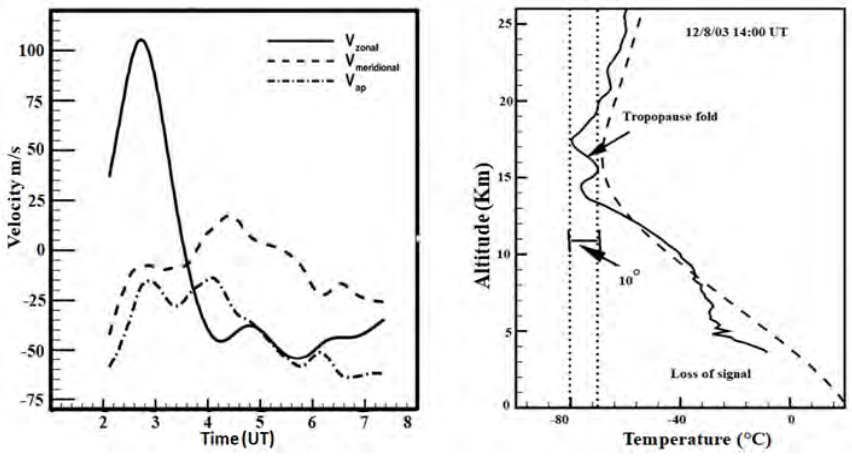


Figure 1. Left panel represents the three ion drift velocity components; zonal (magnetic east), meridional (magnetic north), and anti-parallel direction at 367 km height. Large fluctuations have been observed in all three components during the 'Odette'. Temperature profile taken within the storm path by CHAMP satellite (solid line) and MSIS-90 (dashed line) are presented, as described in Bishop et al., 2006 [13].

Bishop et. al., 2006 studied the ionospheric irregularities associated with the tropical storm 'Odette' by analyzing data of incoherent scatter radar, ionosonde and a satellite based GPS receiver [13]. They found existence of wave-like plasma drift-component variation with higher velocity (period ~90 min) at a height of ~367 km in the F-region (shown in Figure 1). Another interesting result was obtained with temperature profile from the GPS radio occultation observation by the CHAMP satellite near the storm. The temperature profile when compared with the MSIS-90 temperature profile revealed much colder tropopause at higher height in addition to tropopause fold as shown in Figure 1 [13]. The double tropopause or tropopause folds indicate dynamical coupling between the troposphere and stratosphere associated with cyclones which was reported by the works of Uccellini et al., 1985 [14].

Ground-based GPS receivers able to monitor total electron content (TEC) of the ionosphere have also been used to study the characteristic of ionospheric response due to cyclones. Using GPS TEC data, Bondur et al., 2008 reported a sharp enhancement of electron density in the F-layer over the hurricane Katrina during its maximum phase and they hypothesized penetration of electric field, generated by the hurricane, into the ionosphere as the main cause of F-region disturbances [15]. At the same time, Afraimovich et al., 2008 reported that geomagnetic disturbances can generate

ionospheric variation of higher amplitude and consequently may suppress the detection of ionospheric response due to TCs if occur in the same period. Thus a quiet geophysical condition is favorable for detection of F-region ionospheric response using GPS TEC [16]. Effects of TCs on the F-region ionosphere are not instantaneous like solar flares or solar eclipses rather the effects are very specific, like the ionospheric f_0F_2 value increases with the increase in TC intensity and maximum is occurred when TC makes landfall. Shen, 1982 and Liu et al., 2006 indicated a decrease of ionospheric parameters like electron concentration, TEC, and f_0F_2 immediately after the post-landfall days [17]. In the post-landfall period of TC, there may be a protracted decrease in f_0F_2 value. Using 50 GPS TEC stations, an increase of 5 units of TEC value with respect to the mean was observed in the pre-landfall day and 1 unit of TEC increment in the post landfall day for the typhoon Matsa [18]. Thus the tendency of the F-region ionosphere can be generalized as an enhancement of ionospheric parameters with the TC life cycle, maximum values around landfall day, and decrease in post-landfall period [8]. Sharp decrease in ionospheric vertical TEC value obtained along the TC path during post-landfall period of TCs can be explained in several ways. Presence of AGWs in ionospheric boundary is taken as a common phenomenon during the landfall of TC/hurricane/typhoons [19]. During the TC, injection of up stream of neutral particles from TC redistributes the neutral gas particles of troposphere in horizontal direction as well as there are also upward forcing of those particles up to the ionosphere. Belyaev et al., 2015 proposed that the vertical submerged jet in the lower ionosphere injects particles into the upper ionosphere with different directional speed from the injection height responsible for unusual electron density variation over the TC [20].

For the first time over Indian sector, Guha et al., 2016, studied the F-region ionospheric response using GPS TEC data during the TCs Mahasen (2013) and Hudhud (2014). Sharp reduction of 3.8 TEC and 2.1 TEC units with respect to the monthly mean background TEC were observed for TCs Mahasen and Hudhud respectively on the day of landfall. They have also found vertical TEC reduction of 1.5, 1.9, and 2.1 units for three different GPS receivers associated with the TC Vongfong over Japan. Authors proposed combined effects of TC induced AGWs, ejection of neutral gas particles from TC, and TC associated lightning induced electric field responsible for anomalous changes in the ionospheric TEC [21]. Dube et al. 2020, also reported the F-region ionospheric disturbances over Indian sector due to Very Severe Cyclonic Storm 'Phailin' of October, 2013. GPS-TEC data obtained from 7 GPS receiver located at the adjacent areas of cyclone track and lightning data from Global Lightning detection360 network were used in this study [22]. Instead of decrease in TEC values like Guha et al. 2016, they showed increase in TEC values resulting enhanced variation in the differential TEC values on the cyclone days. This increase or decrease in ionospheric TEC values associated with different TCs actually supports the mechanism of AGW propagation in the ionospheric heights from their origin in the troposphere. In general, the low-pressure convective zone associated with a TC in the troposphere generates non-stationary AGWs in a broad-spectrum range. From classical point of view, the coriolis force, pressure gradient force or the gravity force often influence the atmosphere during TCs in such a way that atmospheric particles lose their equilibrium. The gravity force arises as the restoring force to bring back the particles to equilibrium, which results as the evolution of non-stationary AGWs with wavelength 1-6 km in vertical direction, and 10-1000 km along horizontal direction [23]. Some of these waves may dissipate in the upper mesosphere and lower ionosphere regions affecting the atmospheric circulation at those altitudes. Breaking of AGWs in the ionospheric heights also contributes to the modulation of atmospheric densities [24]. Under favorable conditions, the AGWs into the ionospheric heights may also convert into the TIDs [25-27].

3. Effects on the Lower Ionosphere

Effects of TCs are not limited to the ionospheric F-region, but a considerable amount of research also shows the effects in the lower ionosphere, namely the E- and D-regions. Large numbers of AGWs that propagate upward break into the lower ionosphere and create turbulence in the region by depositing energy and momentum. Observable TIDs can be produced from the the turbulence in the ionosphere during passage of TCs as reported in many studies. Existence of meteorological storm induced ionospheric disturbances in the E-layer was reported by Olga et al. 2020 in which the observations were focused on the sporadic E_S-layer. Significant variation of E_S-layer critical frequency was observed in the vicinity of the storm tracks. Propagation of acoustic gravity waves generated due to convective vortex structure of storms was considered as the cause for the reduction of E-layer critical frequency well below the threshold sensitivity of the ionosonde [28]. Few case studies described that creation of E_S- layer and F-region disturbance coincides with same periodicity [29-31]. Apart from the Es-layer irregularity, in-homogeneity of horizontal winds, anomalous variation of E-layer critical frequency and recombination process in the upper mesosphere-lower thermosphere region are also described in connection to TCs [32-35].

The D-region (60-90 km) just below the E-region is also expected to response during the TCs. For example, Perevolva et. al., 2009 showed perturbation of electron concentration of both the D and E-region of ionosphere due to underlying activity of TCs [36]. Among few direct and indirect methods, Very low frequency (3-30 kHz-VLF)/Low frequency (30-300 kHz-LF) electromagnetic radio wave remote sensing technique has been evolved as the most convincing and suitable one to study the D-region ionosphere. Using ground based VLF/LF radio receiver one can easily and uninterruptedly monitor amplitude and phase of VLF/LF signals transmitted by man-made communication transmitters or lightning strikes. VLF/LF radio signals propagate long distances with minimum attenuation [37] through multiple reflections between the conducting medium formed between the lower boundary of ionosphere and the surface of earth and can be received by suitable receiver [38]. This technique has long been used scientifically since 1950's

Heliophysical Year and revealed so many significant facts about lower ionospheric changes during several geophysical processes like space weather activities, solar eclipses, lightning perturbations, earthquakes, Gamma ray bursts, Tropical Cyclones, etc [39-42].

Existence of AGWs in troposphere-stratosphere level was found during TCs and it was also detected at large distance from cyclone track. Correlation between cyclone intensity and amplitude of AGWs was also established in some recent studies [43-44]. Rozhnoi et al., 2014, presented convincing results of VLF/LF radio signals perturbations due to the passage of TCs in the troposphere. They reported prominent gravity waves in the range of 7-16 min and 15-55 min periods in the D-region ionosphere due to the TCs [45]. Nina et al., 2017 studied NAA-VLF amplitude data for 69 tropical depressions, and showed that the tropical depressions which grew into Hurricanes in later stage affected the lower ionosphere [46]. In another study Fiji based researchers monitored the D-region disturbances during a TC [47] using the VLF signals from the NPM (21.4 kHz), NLK (24.8 kHz), NAA, and JJI (22.2 kHz) transmitters. They found reduction in VLF amplitudes during the depression phase of the cyclone. Further, they simulated the reference D-layer ionospheric height using the Long wave capability code (LWPC) and found a sharp 5.2 km lowering of nighttime D-layer height along the JJI-Fiji path and ~6.0-7.5 km increase in the reference height along the NPM-Fiji path. To confirm the existence of wave-like signatures of AGWs, researchers applied wavelet analysis on the residual VLF signals. For example, Kumar et. al. 2017, NaitAmor et al. 2018, and Correia et. al. 2019 have used wavelet analysis to characterize the AGWs associated with the VLF signal perturbations in connection to TCs. While Kumar et al., 2017 reported AGW signatures with periodicity 7 min to 5.5 hr, NaitAmor et al. 2018 found AGW signatures with periodicity 2 to 3 hr in the D-region ionosphere [42,47-48]. Thus a broad spectrum of AGWs ranges from several minutes to few hours can be found in the D-region ionosphere due to TCs. Further, the VLF/LF signal disturbances can be observed even if the cyclonic storms remain 1,000 to 2,000 km far from the radio propagation path or from receiver [42]. Though VLF/LF signals capture the signatures of AGWs in the lower ionosphere, the relationship of AGW activities with TC intensity is not well established yet. Even, the changes in temperature or chemical composition in lower ionosphere due to TCs are not well understood.

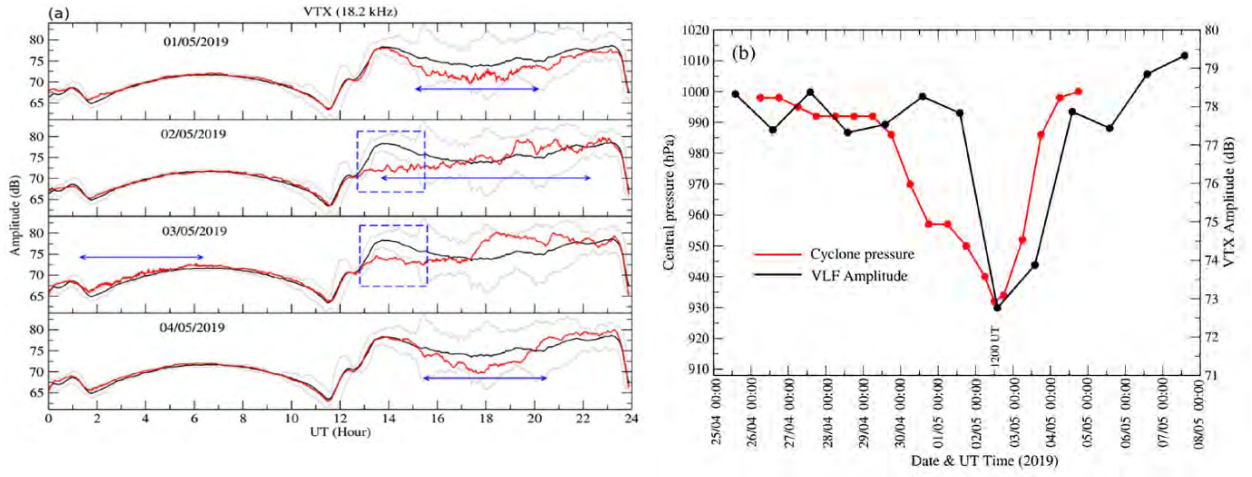


Figure 2. (a) Anomalous nighttime signal variation of the VTX amplitude above 3σ levels during the TC Fani. (b) Correlation between the cyclone pressure (red) is compared with the corresponding variation of VTX amplitude (black) as obtained by Pal et al. 2020 [49].

In an attempt to find the correlation of AGW activities in the D-region ionosphere with TC intensity, Pal et. al., 2020 investigated VLF signals from the VTX (18.2 kHz, India) and NWC (19.8 kHz, Australia) transmitters during the extremely severe cyclone Fani over the Bay of Bengal. They showed anomalous amplitude variations and wave-like oscillations of the nighttime VLF signals during the cyclone period as shown in Figure 2a. VLF signal deviation was correlated with the cyclone pressure change (shown in Figure 2b) and also the percentage change in both VLF amplitude and cyclone pressure was similar. Using wavelet analysis they found significant wave bands of period 10 min to 2 hr in the D-region ionosphere. High frequency component with periodicity 10-30 min exhibited a strong anti-correlation between AGW amplitudes and cyclone pressure which further adds a possibility of monitoring cyclone intensity from VLF signal measurements. They also showed temperature and Ozone anomalies in the D-region ionosphere (shown in Figure 3) around the VLF reflection heights during the cyclone indicating changes in electron-neutral collision frequency and chemical composition in the D-region. While the maximum ozone anomaly and maximum VLF anomaly occurred on the day of maximum cyclone intensity, the maximum temperature anomaly was found on the landfall day [49].

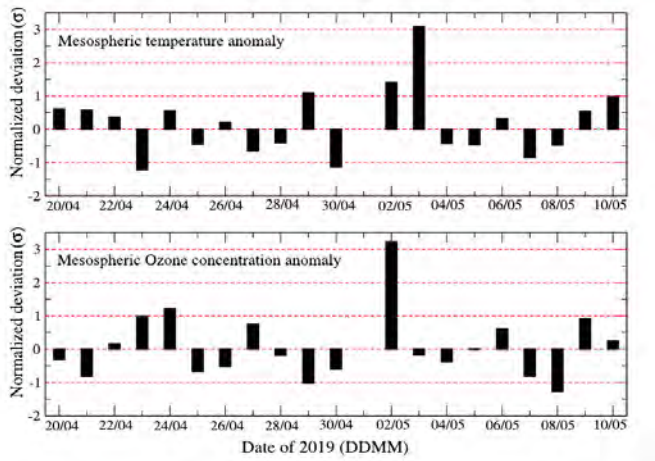


Figure 3. Temperature and Ozone anomalies in the D-region ionosphere during the extremely severe cyclone Fani, as described in Pal et al., 2020[49].

Das et al., 2021 generalizes the study of TC induced ionospheric perturbations over Indian sector using multi-path VLF signals from two places and they attempted to find the disturbances in all atmospheric layers starting from tropopause to D-region ionosphere via the stratosphere. They showed VLF signal disturbances both during day and night. The VLF propagation path closest to the TC track was exhibited strong perturbation. Both pressure and wind speed was correlated with VLF signal amplitude deviation, though the correlation was found better with central pressure. They showed that atmospheric temperature anomalies at the tropopause, stratopause, and ionosphere were well connected to the TC Fani. Further, the LWPC simulation exposed the perturbation characteristics of the D-region ionosphere above the cyclone as shown in Figure 4. Variation of the VLF reflection heights in the D-region due to the cyclone along the propagation path followed a Gaussian curve from which the authors estimated a ~1650 km spatial size of the ionospheric disturbances bigger than the cloud image of the cyclone [50].

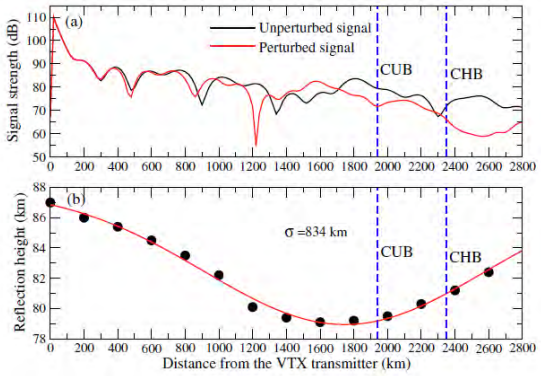


Figure 4. Comparison of perturbed and unperturbed VTX signal from transmitting point to receivers (upper panel). Gausian behavior of the D-region reflection heights during Fani. This estimation can be used to calculate spatial size of the disturbance in the ionosphere generated by TC [50]

Another significant recent study showed that morning and evening terminator times along with both day and nighttime amplitude perturbations on the VTX and NWC signals over the Indian sector associated with the super cyclonic storm Amphan in 2020. Significant shift of morning and evening terminator times and ~10 minute increase of VLF day-time were observed due to the TC generated disturbances other than the Sun. In case of the TC Fani, a decrease in the VTX signal level and increase in the NWC signal were reported, but in case of the TC Amphan, intensity of which was higher, the variations were opposite for the VTX and NWC signals as observed from the same place. The reason of this opposite behaviour could be resulted from the difference in the orientation of the cyclone tracks with respect to the recording site or change in modal interference pattern associated with disturbed D-region ionosphere during two cyclones. The authors also estimated the size of the D-region perturbed region using the multipath VLF observations associated with the cyclone during its maximum intensity [51].

4. Conclusion

Impact of tropospheric cyclone activities on the upper and lower ionosphere have been reviewed here in detail. Investigations revealed that the evolution of AGWs during cyclone can be the most probable cause of troposphere-ionosphere coupling. Except the AGW channel, the electric field produced by lightning discharge and the modification of atmospheric dc electric field have the ability to perturb the ionosphere during the TCs. The ionospheric perturbation is

commonly observed within cyclonic period, sometimes before the pre-cyclone formation stage i.e., in the depression or tropical disturbance stage and mostly during the peak cyclone stage or during landfall or post-landfall period. The tropical disturbances or deep depressions transforming into cyclones/tropical cyclones/hurricanes or typhoons are the biggest sources of ionospheric perturbations. GPS TEC, ionosonde, and satellite based GPS radio occultation observations are very much efficient technique to monitor the F-region ionosphere during the passage of TCs. Monitoring sporadic E-layer can be possible only by ionosonde or incoherent radar experiments. The D-region ionosphere can be continuously monitored using VLF/LF remote sensing methods. VLF/LF signals provide a good opportunity to monitor the lower ionosphere disturbances due to the tropical cyclones. A dense VLF/LF network with sufficient number of receivers has the capability to monitor the exact dimension of ionospheric disturbances during the tropical cyclones. This will also allow monitoring the gravity waves in the ionosphere arising from the cyclones in the troposphere. Indeed, a more detail investigation is needed to know further about the formation and evolution of ionospheric disturbances during the tropical cyclones and whether the strength of the ionospheric disturbances can be taken back to calculate the intensity of the storm.

Acknowledgements

P.K. Haldar thanks the financial support provided by Science & Technology and Biotechnology Department of Govt. of West Bengal, India (Sanction Memo no. 917(Sanc.)/STBT-11012(20)/42/2019-ST SEC). S. Pal acknowledges the support from the SERB research grant SRG/2020/001104.

References

[1] Guha, A., Paul, B., Chakraborty, M. and De, B.K, "Tropical cyclone effects on the equatorial ionosphere: First result from the Indian sector," *J Geophys Res Space Physics*, 121, 5764–5777, 2016. doi:10.1002/2016JA022363.

[2] Montgomery, M. T., and Farrell, B. F., "Tropical Cyclone Formation," *Journal of the Atmospheric Sciences*, 50, 285-310, 1993. https://doi.org/10.1175/1520-0469(1993)050<0285:TCF>2.0.CO;2

[3] Brian, H., Tang, Juan, F., Alicia, B., Gerard, K., Masuo, Na., Myung-Sook, P., Rajasree, V. P. M., Zhuo, W., Allison, A.., Wing, L. , "Recent advances in research on tropical cyclogenesis," *Tropical Cyclone Research and Review*, 9, 2, 87-105, 2020, https://doi.org/10.1016/j.tcrr.2020.04.004

[4] Mitra, S. K, *The Upper Atmosphere*, Asiatic Society, Calcutta, 08-11, 1952.

[5] Bauer, S. J, "A possible troposphere-ionosphere relationship," J Geophys Res, 62. 425-430, 1957.

[6] Martyn, D. F, Procedings of Conference on Ionospheric Physics, July 1950, Part 1 (edited by L. Katz and N. C. Gerson), Geophysical Res. Division, Air Force, Cambridge Res. Centre, Geophy. Res. Papers, 12, 31-33, 62-64, 1952

[7] Wilkes, M. V, Oscillations of the earth's atmosphere, Cambridge University Press, England, 1949.

[8] Bauer S. J, "An apparent ionospheric response to the passage of hurricanes," J Geophys Res, 63. 265–269, 1958.

[9] Hung, R. J., Kuo, J. P, "Ionospheric observation of gravity waves associated with Hurricane Eloise," J Geophys, 45. 67–80, 1978.

[10] Huang, Y. N., Cheng, K. and Chen, S. W, "On the detection of acousticgravity waves generated by typhoon by use of real time HF Doppler frequency shift sounding system," Radio Sci, 20. 897–906, 1985.

[11] Xiao, Z., Xiao, S., Hao, Y. and Zhang, D, "Morphological features of ionospheric response to typhoon" J Geophys Res, 112. A04304, 2007. doi:10.1029/2006JA011671.

[12] Baker, D. M., Davies, K, "F2-region acoustic waves from severe weather," *J Atmos Terr Phys*, 31. 1345–1352, 1969.

[13] Bishop, R. L., Aponte N., Earle G. D., Sulzer M., Larsen M. F. and Peng G. S, "Arecibo observations of ionospheric perturbations associated with the passage of Tropical Storm Odette," *J Geophys Res*, 111, 2006. doi:10.1029/2006JA011668

[14] Uccellini, L. W., Keyser, D., Brill, K. F., Wash, C. H., "The President's day cyclone of 18–19 February 1979: influence of upstream trough amplification and associated tropopause folding on rapid cyclogenesis," *Monthly Weather Review* 113: 962–988, 1985.

[15] Bondur, V. G., Pulinets, S. A., Uzunov, D, "Ionospheric effect of large-scale atmospheric vortex by the example of hurricane Katrina," *EarthRes from Space*, 6. 3–11, 2008.

[16] Afraimovich, E. L., Voyeikov, S.V., Ishin, A. B., Perevalova, N. P. and Ruzhin, Y. A, "Variation of Total Electron Content during powerful typhoonon 5–11 August 2006 near the southeastern coast of China," Geomagn Aeron, 48 (5). 674–679, 2008.

[17] Shen, C. S, "The correlations between the typhoon and the foF2 of ionosphere," *Chin J Space Sci*, 2. 335–340, 1982.

[18] Mao, T., Wang, J. S., Yang, G. L., Yu, T., Ping, J. S. and Suo, Y. C, "Effects of typhoon Matsa on ionospheric TEC," *Chin Sci Bull*, 55(8). 712–717, 2010. doi:10.1007/s11434-009-0452-4.

[19] Ming, C. F., Chen, Z. and Roux, F, "Analysis of gravity-waves produced by intense tropical cyclones," *Ann Geophys*, 28. 531–547, 2010. doi:10.5194/angeo-28-531-2010.

[20] Belyaev, G., Boychev, B., Kostin, V., Trushkina, E. and Ovcharenko, O, "Modification of the ionosphere near the terminator due to the passage of a strong tropical cyclone through the large island," *Sun Geosphere*, 10, 31–38, 2015.

[21] Guha, A., Paul, B., Chakraborty, M. and De, B. K, "Tropical cyclone effects on the equatorial ionosphere: First result from the Indian sector," J Geophys *Res Space Physics*, 121, 5764–5777, 2016. doi:10.1002/2016JA022363.

[22] Dube, A., Singh, R., Maurya, A. K., Kumar, S., Sunil, P. S. and Singh, A. K, "Ionospheric perturbationsinduced by a very severe cyclonic storm(VSCS): A case study of Phailin VSCS," *Journal of Geophysical Research SpacePhysics*, 125, e2019JA027197, 2020. https://doi.org/10.1029/2019JA027197.

[23] Tsuda, T., "Characteristics of atmospheric gravity waves observed using the MU (Middle and Upper atmosphere) radar and GPS (Global 8Positioning System) radio occultation," *Proc. Japan Acad. Ser. B, Phys. Biol. Sci*, 90 (1), 12–27, 2014.

[24] Singh, R., Pallamraju, D., "Effect of cyclone Nilofar on mesospheric wave dynamics as inferred from optical nightglow observations from Mt. Abu, India: Effect of cyclone Nilofar in mesosphere," *Journal of Geophysical Research: Space Physics*, 121, 2016. https://doi.org/10.1002/2016JA022412.

[25] Hocke, K., Schlegel, K, "A review of atmospheric gravity waves and traveling ionospheric disturbances: 1982–1995," *Ann. Geophys*, 14, 917–940, 1996.

[26] Kazimirovsky, E., Herraiz, M., De La Morena, B. A, "Effects of the ionosphere due to phenomena occurring below it," *Surv Geophys*, 24. 139–184, 2003.

[27] Lastovicka, J, "Forcing of the ionosphere by waves from below," *J Atmos Sol-Terr Phys*, 68, 479–497, 2006.

[28] Olga, B., Karpov, I. V., Karpov, M., Korenkova, N., Vlasov, V. and Leshchenko, V, "Impact of meteorological storms on the E-region of the ionosphere in 2017–2018," *Solar-Terrestrial Physics*, 6, 74-79, 2020. https://doi.org/10.12737/stp-64202011.

[29] Van Eyken A. P., Williams P. J. S., Maude A. D., and Morgani G, "Atmospheric gravity waves and sporadic-E," *J AtmosSolar-Terr Phys*, 44 (1), 25–29, 1982. https://doi.org/10.1016/0021-9169(82)90089-7.

[30] Mathews, J. D, "Sporadic E: current views and recent pro-gress," *J Atmos Solar-Terr Phys*, 60 (4), 413–435, 1998. https://doi.org/10.1016/S1364-6826(97)00043-6.

[31] Parkinson, M. L., Dyson, P. L, "Measurements of mid-latitude E-region, sporadic-E, and TID-related drifts using HF Doppler-sorted interferometry," *J Atmos Solar-Terr Phys*, 60 (5), 509–522, 1998. https://doi.org/10.1016/S1364-6826(97)00058-8.

[32] Barta, V., Haldoupis, C., Sátori, G., Buresova, D., Chum, J., Pozoga, M., Kitti, A. B., József, B., Martin, P., Árpád, K., Pál, B., "Searching for effects caused by thunder-storms in midlatitude sporadic E-layers," *J Atmos Solar-Terr Phys*, 161, 150–159, 2017. https://doi.org/10.1016/j.jastp.2017.06.006.

[33] Haldoupis, C, "Midlatitude sporadic E. A typical paradigm of atmosphere-ionosphere coupling," *Space Sci Rev*, 168, 441–461, 2012. https://doi.org/10.1007/s11214-011-9786-8.

[34] Šauli, P., Bourdillon, A, "Height and critical frequency varia-tions of the sporadic-E layer at midlatitudes," *J Atmos Solar-Terr Phys*, 70 (15), 1904–1910, 2008. https://doi.org/10.1016/ j.jastp.2008.03.016

[35] Didebulidze, G. G., Dalakishvili, G., Lomidze, L. and Mati-ashvili, G, "Formation of sporadic-E (Es) layers under the influence of AGWs evolving in a horizontal shear flow," *J Atmos Solar-Terr Phys*, 136(B), 163–173, 2015. https://doi.org/10.1016/ j.jastp.2015.09.012

[36] Perevalova, N. P., Ishin, A. B, "Effects of Tropical Cyclones in the Ionosphere from Data of Sounding by GPS Signals," *AtmosOcean Phys*, 47, 1072–1083. 2009. https://doi.org/10.1134/S000143381109012X.

[37] Davies, K, *Ionospheric radio*, Peregrinus, London, 1990.

[38] Das, B., Sen, A., Haldar, P. and Pal, S, "VLF radio signal anomaly associated with geomagnetic storm followed by an earthquake at a subtropical low latitude station in northeastern part of India," *Indian Journal of Physics*, 2021. https://doi.org/10.1007/s12648-020-01966-2.

[39] Inan, U. S., Cummer, S. A. and Marshall, R. A, "A survey of ELF and VLF research on lightning-ionosphere interactions and causative discharges," *J Geophys Res Space Phys*, 115, A6, 2010. https://doi.org/ 10.1029/2009JA014775.

[40] Pal, S, *Numerical modelling of VLF radio wave propagation through earth-ionosphere waveguide and its application to sudden ionospheric disturbances*, PhD thesis to University of Calcutta, 2015. arXiv:1503.05789 [astro-ph.EP].

[41] Das, B., Pal, S. and Haldar, P.K, "VLF radio signal perturbations during two recent solar eclipses observed from a VLF receiving station, Cooch Behar, India," in *2021 XXXIVth General Assembly and Scientific Symposium of the International Union of Radio Science (URSI GASS)*, 1-4. https://doi.org/10.23919/URSIGASS51995.2021.9560244.

[42] NaitAmor, S., Cohen, M.B., Kumar, S., Chanrion, O. and Neubert, T, "VLF Signal Anomalies During Cyclone Activity in the Atlantic Ocean," *Geophys Res Lett,* 45 (10), 185–210, 2018. https://doi.org/10.1029/2018GL078988.

[43] Nolan, D. S., Zhang, J. A, "Spiral gravity waves radiating from tropical cyclones," *Geophysical Research Letters*, 44, 3924, 2017. https://doi.org/10.1002/2017GL073572

[44] Hoffmann, L., Wu, X. and Alexander, M.J, "Satellite observations of stratospheric gravity waves associated with the intensification of tropical cyclones," *Geophysical Research Letters*, 45, 1692, 2018. https://doi.org/10.1002/2017GL076123

[45] Rozhnoi, A., Solovieva, M., Levin, B., Hayakawa, M. and Fedun, V, "Meteorological effects in the lower ionosphere as based on VLF/LF signal observations," *Nat Hazards Earth Syst Sci,* 14 (10), 2671–2679, 2014. https://doi.org/10.5194/nhessd-2-2789-2014.

[46] Nina, A., Radovanovic, M., Milovanovic, B., Kovacevic, A., Bajcetic, J. P. and Luka, C, "Low Ionospheric Reactions on Tropical Depressions prior Hurricanes," *Adv Space Res*, 2017. https://doi.org/10.1016/j.asr.2017.05.024.

[47] Kumar, S., NaitAmor, S., Chanrion, O. and Neubert, T, "Perturbations to the lower ionosphere by tropical cyclone Evan in the South Pacific Region," *J Geophys Res Space Physics*, 122, 8720–8732, 2017. https://doi.org/10.1002/2017JA024023.

[48] Correia, E., Tiago, L., Raunheitte, M., Valentin, J., Dino, B. and DAmico, E, "Characterization of gravity waves in the lower ionosphere using VLF observations at Comand ante Ferraz Brazilian Antarctic Station," Ann Geophys, 1–15, 2019. https://doi.org/10.5194/angeo-2019-123.

[49] Pal, S., Sarkar, S., Midya, S.K., Mondal, S.K. and Hobara, Y, "Low-Latitude VLF Radio Signal Disturbances Due to the Extremely Severe Cyclone Fani of May 2019 and Associated Mesospheric Response," *J Geo Res Space Phys*, 125, 5, 2020. https://doi.org/10.1029/2019JA027288.

[50] Das, B., Sarkar, S., Haldar, P.K., Midya, S.K. and Pal, S, "D-region ionospheric disturbances associated with the Extremely Severe Cyclone Fani over North Indian Ocean as observed from two tropical VLF stations," Advances in Space Research, 67(01), 75-86, 2021. https://doi.org/10.1016/j.asr.2020.09.018

[51] Das, B., Sen, A., Pal, S. and Haldar, P.K, "Response of the Sub-Ionospheric VLF Signals to the Super Cyclonic Storm Amphan: First Observation from Indian Subcontinent," *Journal of Atmospheric and Solar-Terrestrial Physics*, 220, 105668, 2021. https://doi.org/10.1016/j.jastp.2021.105668.

Possible Precursory Effects of Seismic Events in VLF Radio Signals

Suman Ray[1*]

[1]Department of Physics, Gobardanga Hindu College, Gobardanga, North 24 Parganas, West Bengal, India.
*Corresponding author: sumanray07@email

Abstract 'Very Low Frequency' (VLF) radio waves (3-30 KHz) propagate through the Earth-ionosphere wave-guide which is formed by lower part of the ionosphere and upper part of Earth's surface. Normally, patterns of VLF signal depend on regular solar flux variations. However, an extra source of ionization (e.g., solar flares, gamma ray bursts, possible seismic events) can change height of ionospheric layers and/or ion densities and these changes can perturb VLF signal amplitude. By measuring amplitude and phase of radio signals reflected from the ionosphere, it is possible to detect various kinds of energetic phenomena. Here we shall mainly discuss about the possibilities of predicting seismic events by using these VLF signals.

Keywords: *VLF Radio Signal, Earth-ionosphere Waveguide, Seismic events.*

1. Introduction

Every year so many natural disasters occur which cause huge human and economic losses. Among them seismic events is one of the greatest natural disasters. The causes of seismic events are very complex and it is mainly due to the sudden abnormal movements in tectonic plates which we unable to fix. The best way to deal with it is to predict it so that we can save so many lives. But the predictions of the seismic events are not an easy task for scientific community. Already so many attempts have been taken by scientists and yet a fruit full result is still awaiting.

There are mainly two types of predictions. One is long-term predictions and another one is short term predictions. Long-term predictions of seismic events are mainly based on the analysis of Earth's geological structures which is the job of the geologists and they are trying to do that. On the other hand short term predictions of seismic events can be done by identifying the ionospheric disturbances. It is believed that before any seismic events it starts to release huge energies which may create disturbances in the Earth's ionosphere. Very Low Frequency (VLF) radio signals may be used to identify these ionospheric disturbances and hence it may be used to predict the seismic events.

'Very Low Frequency' (VLF) is one of the bands of Radio waves having frequencies lying between 3-30 KHz, with wavelengths 100-10 Km. It propagates through the Earth-ionosphere wave-guide which is formed by lower part of the ionosphere and upper part of Earth's surface. Thus it may be perturbed due to the ionospheric disturbances, associated with pre-seismic activities. Scientific works regarding this started in 1960's. The first paper about the seismo-ionospheric correlation was published by Bolt et al. in 1964 [1] after Alaska Good Friday earthquake which occurred on Friday, March 27 (local time), 1964. They used ionosonde method to find out this correlation. After that several papers have been published where different workers had used different methods to find out relation between seismic events and ionospheric anomalies by using VLF signals [2] [3] [4] [5]. Then another important work was reported by [6]. They observed the signals received by Omega navigation transmitter (~10 kHz) during 1983-1986, and they found that 250 out of 350 earthquakes with magnitude (M) greater than 4

were associated with phase and/or amplitude variations. The most convincing evidence regarding earthquake precursor effects in VLF signals, was obtained by Hayakawa et al. in 1996 [7] [8]. They analyzed VLF data during an eight month period centered on great Hyogo-Ken Nambu (Kobe) earthquake, which occurred on 17th January, 1995. The magnitude of this earthquake was 7.2. They found a significant amount of shift in sunrise and sunset terminator times few days before earthquake. This method of prediction is known as "Terminator Time method".

In Indian context, we have also reported several papers [9] - [15] where we found the evidences of observing VLF signal anomalies associated with pre-seismic activities. In the next sections, we shall discuss these results in detail.

2. Possible Pre-Seismic Effects on "VLF Day Length"

As we have mentioned, after the Kobe earthquake occurred in 1996, several workers have been reported that the terminator times are shifted towards night few days before an earthquake. To verify this, in the context of Indian sub-continent we have introduced a new parameter, namely "VLF day length" which is defined as the difference between sunset and sunrise terminator times. Our theory is very simple. If the terminator shifts really happens then it will increase the value of "VLF day length" and we shall get an anomalous "VLF day length". To investigate this, we have analyzed the VLF signals for VTX-Malda propagation path. The latitude and longitude of the transmitting station VTX is 08.43^{o} and 77.73^{o} respectively. VTX transmits the VLF signals at 18.2 KHz. Our receiving station is situated at Malda branch of Indian Centre for Space Physics (Latitude 22.56^{o} and Longitude 88.04^{o}). Here we have used a whole year data of 2008 to find out the pre-seismic effects, if any, on the anomalous "VLF day length". For this at first we find out both the terminator times from each day of the VLF signals and then we calculate the "VLF day length". Finally we calculate the correlation coefficient of the deviation of this "VLF day length" with the effective magnitude of the seismic events. We have presented this plot in Figure 1. We found that the correlation coefficients become higher one day before the earthquake. This result indicates that anomalous "VLF day length" may be used as a precursory effect of the earthquakes.

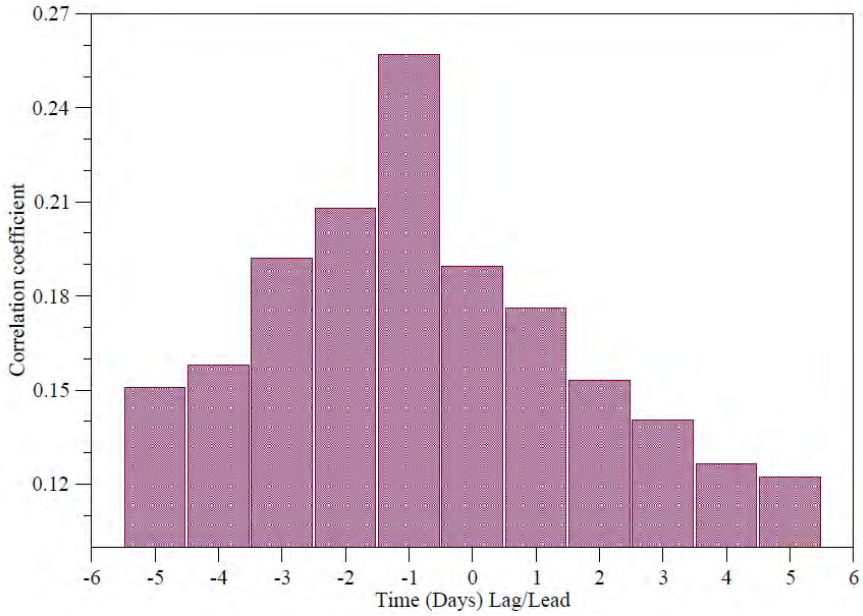


Fig. 1: Correlation between effective magnitude of earthquake at mid-point between transmitter and receiver and variation of anomalous 'VLF day length' [12] [15].

We have also reported similar type of results for NWC-Salt Lake propagation path. NWC (latitude 22.56^{o} and longitude 22.56^{o}) transmits VLF signals at 19.8 KHz which is received at Salt Lake,

Kolkata (latitude 22.56^o and longitude 22.56^o). In June, 2010, two major earthquakes (depth ~ 10 Km and magnitude greater than 5) occurred at Nicobar (latitude 22.56^o and longitude 22.56^o) and Andaman (latitude 22.56^o and longitude 22.56^o) Islands, India. The first one occurred on 13th June, 2010 and another one occurred on 19th June, 2010. The epicentres of both the earthquakes are very close to NWC-Salt Lake propagation path. This is shown in Figure 2. We have analyzed the VLF signals for this propagation path during 5-25th June, 2010, centred on these earthquakes days. For both the cases, we have observed a significant amount of shift in sunset terminator time one day before the earthquake. This is shown in Figure 3. We have also calculated the "VLF day length" and found that it also becomes anomalously high just one day before both of the earthquakes. This is shown in Figure 4.

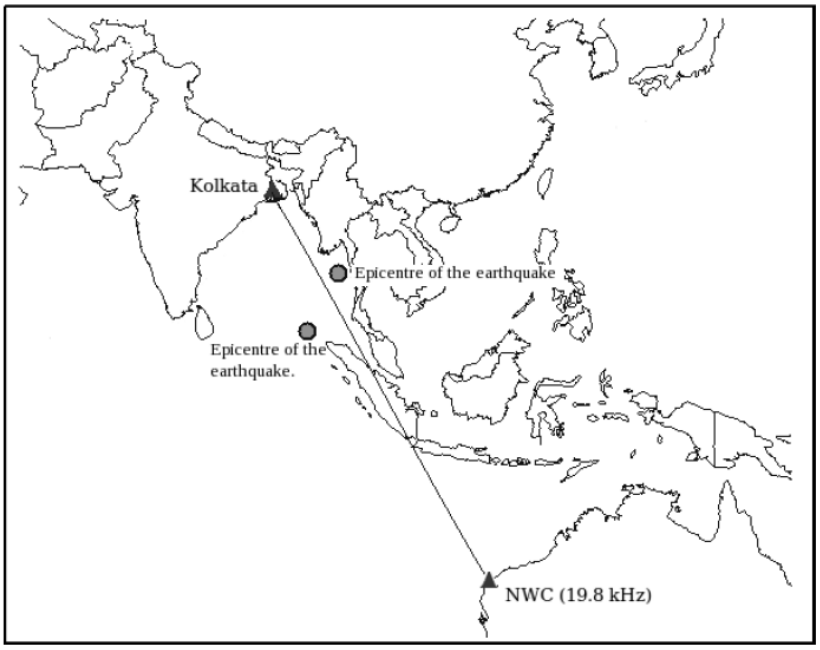


Fig. 2: NWC-Kolkata VLF propagation path and the locations epicenters of the earthquakes [15].

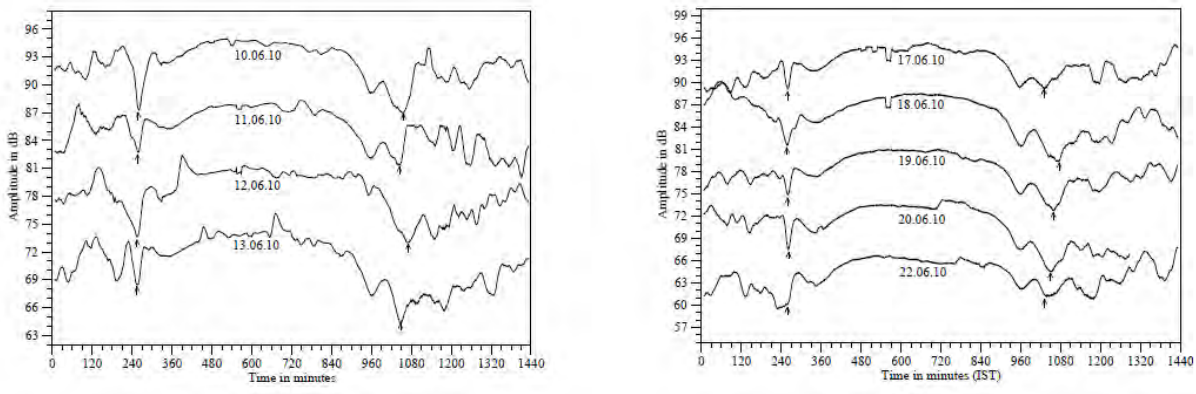


Fig. 3: Variation of amplitude of VLF signal is plotted as a function of time in both plots. Signals are plotted in both left and right panel, around the earthquake day which occurred on 13th June, 2010 and also on 19th June, 2010, respectively. SRTs and SSTs of VLF signals are marked by 'arrow' symbols [15].

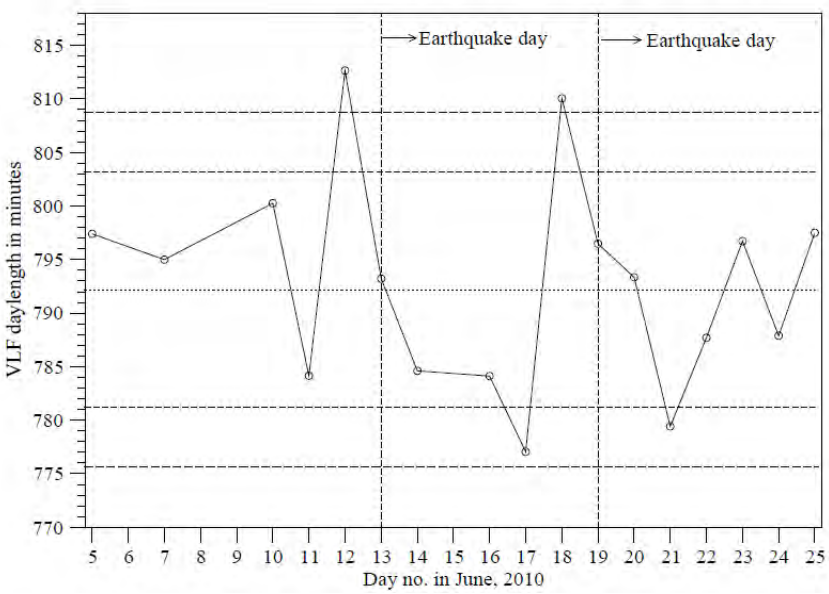


Fig. 4: VLF day-lengths are plotted as a function of day number during 5-25 June, 2010. 'Solid' curve represents variation of 'VLF day-length'. 'Dotted' curve represents mean of these 'VLF day-length' while 'small-dashed' and 'big-dashed' curves represent mean±2σ and mean±3σ curves. Two vertical lines indicate two different earthquake days. Note that 'VLF day-lengths' of 12th June, 2010 and 'VLF day-length' of 18th June, 2010 crossed 3σ line. These could be precursor effects of earthquakes which occurred on 13th and 19th June, 2010, respectively [15].

3. Correlation of Anomalous "Nighttime VLF Amplitude Fluctuations" with Seismicity

We have analyzed the night time amplitude variation of the VLF signals for VTX-Kolkata propagation paths to find out its correlations, if any, with seismicity. The transmitting station, VTX (18.2 KHz) is located at Vijayanarayanam (latitude 8.43^o N, longitude 77.73^o E) and our receiving station is situated at Indian Centre for Space Physics, Kolkata (latitude 22.56^o N, longitude 88.56^o E). For this present work, we have used a whole year night time data of 2007, received by an AWESOME receiver. Our night time starts at 19:30 h (14:00 UT) and ends at 04:30 h (23:00 UT) of the (local) next day with a one hour gap just prior to midnight for data analysis. We stayed away from sunrise and sunset terminators by at least an hour to avoid contamination from the D-layer formation or disappearance effects. To calculate the night time amplitude fluctuations, we first calculate the standard deviation of the night time signal for each day. Then we subtract this value from its mean value. By this way we calculate the deviation of the night time amplitude for each day. Then we calculate the effective magnitude of the earthquake occurred near the VTX-Kolkata propagation path during the year 2007. Finally we calculate the correlation co-efficient between the deviation of the night time amplitude of the VLF signals and the effective magnitude of the earthquake. This result is shown in Figure 5. It indicates that the night time fluctuations of the VLF signals became anomalously high three days before the earthquake.

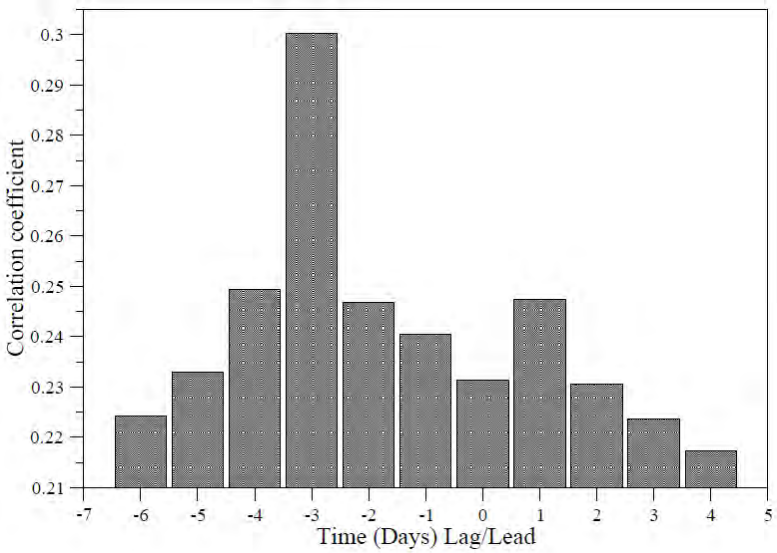


Fig. 5: Correlation between effective magnitude of earthquake at mid-point between transmitter and receiver and variation of anomalous night time amplitude fluctuations of the.VLF signals. Note that correlation coefficient is the highest three days before an earthquake [13] [15].

We have also carried out few case by case studies to ensure our results. For example we may present our analysis for VTX-Kolkata propagation path during 11th to 25th January, 2011 centred on an earthquake day which occurred on 18th January, 2011 at Southwestern Pakistan (latitude 28.09^{o} N, longitude 64^{o} E). The magnitude of the earthquake was 7.4. Here also we have observed similar types of result which we found in case of year-long study. In Figure 6, we have presented the night time amplitude variations of the VLF signals for 13th and 14th January, 2011. Here we have observed that the variation of the signal of 13th January, 2011 is quite but a huge fluctuation is present in the data of 14th January, 2011. This anomalous fluctuation in the night time VLF signal of 14th January 2011 could be the precursor effect for the earthquake (M = 7.4) occurred on 18th January 2011 in Pakistan.

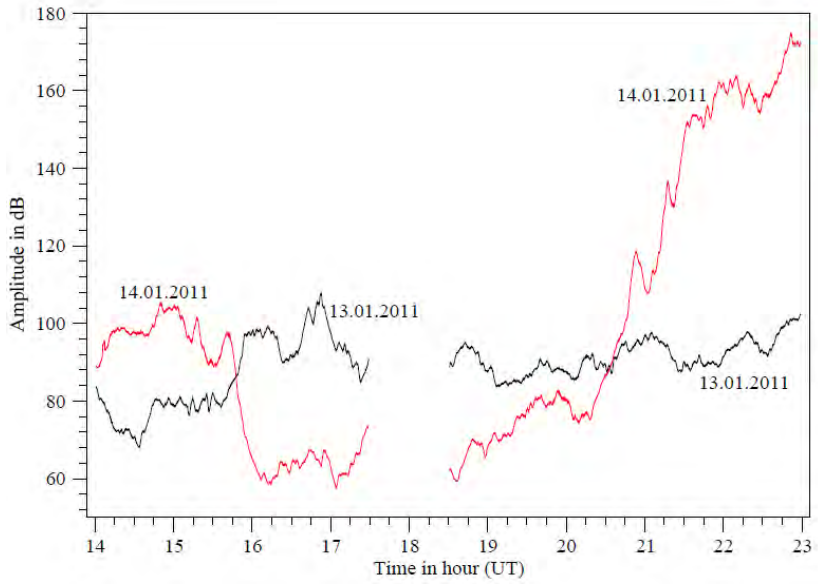


Fig. 6: Amplitude of night time VLF signals are plotted as a function of time. Black curve is for 13th January, 2011 and red curve is for 14th January, 2011. Note that amplitude of signal of 13th January, 2011 is quiet but high fluctuation is present in signal of 14th January, 2011. This anomalous fluctuation in night time VLF signal of 14th January, 2011 could be precursor effect for 7.4 earthquake occurred on 18th January, 2011 at Pakistan [14] [15].

In Figure 7, we have plotted the night time VLF amplitude fluctuations as a function of day number for two weeks centred on the earthquake day (i.e. 18th Jan, 2011). Here we have observed that the amplitude fluctuations of the VLF signals crossed 3σ line four days before earthquake. This indicates that the anomalous night time amplitude fluctuations of the VLF signals may be considered as a precursory effect of the earthquakes.

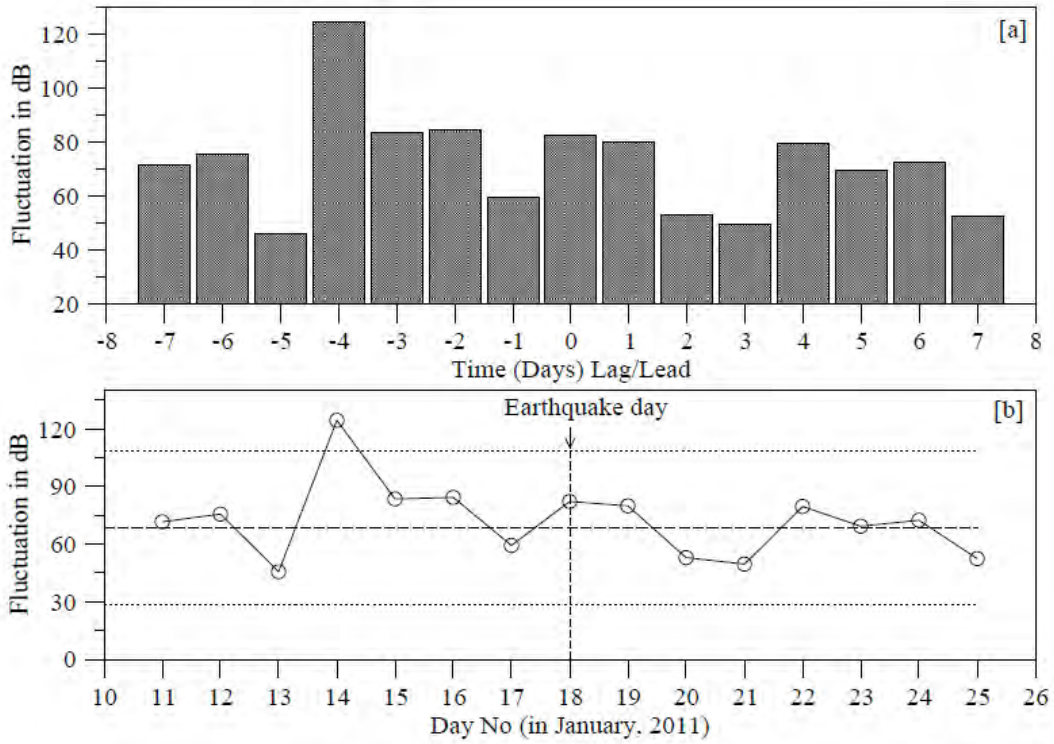


Fig. 7: Night time VLF amplitude fluctuations for two weeks centered on 18th January, 2011. Note that peak appears four days before the event. (b) Amplitude of night time VLF fluctuations (open circles) signals are plotted as a function of day number. Dashed curve represents average night time fluctuations while dotted curves represent the ±3σ (σ is the standard deviation) lines. Vertical line is earthquake day. Fluctuation four days before the event crossed 3σ line [14] [15].

4. Anomalies in "DLPT" and "DLDT"

It is also reported that the "D-layer preparation time" (DLPT) and "D-layer disappearance time" (DLDT) become anomalously high few days before an earthquake [9] [10] [15]. This method of prediction is known as "DLPT and DLDT method".

We now present the analysis of the whole year data of 2008 for VTX-Malda propagation path. We find out the value of 'DLPT' and 'DLDT' for each day. Then we calculate the deviation of 'DLPT' and 'DLDT' by subtracting its value from its mean value. In Figure 8, we have plotted the Correlation coefficient between effective magnitude of earthquake at mid-point of propagation path and anomalous variations of the DLPT (left panel) and DLDT (right panel). We note that in case of DLPT, a peak in correlation coefficient appears just two days before the earthquake. But for DLDT we find the peak to form on the day of the earthquake. Since we have chosen a 1 day bin-size, more accurate analysis is required to judge whether the peak forms at least a few hours before an earthquake.

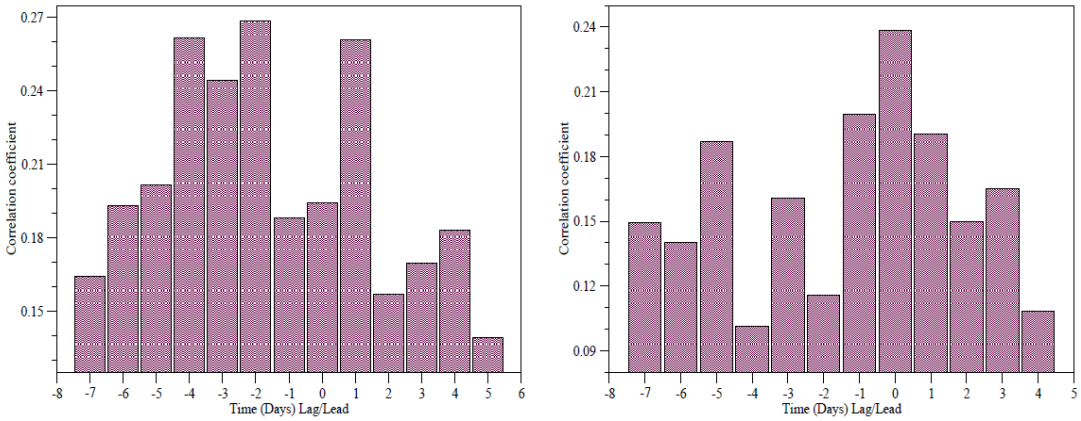


Fig. 8: Correlation coefficient between effective magnitude of earthquake at mid-point of propagation path and anomalous variations of the DLPT (left panel) and DLDT (right panel) [13] [15].

5. Conclusion

We have discussed about the possibilities of prediction of seismic events by using VLF signals. Search for correlation between seismic events and ionospheric signal anomaly is on for about half a century, and yet, a definite answer to this tantalizing problem is not in hand. Reason is that the problem is multi-parametric and highly non-local. The effect is also non-linear. In these circumstances, the best strategy would be to analyze as many cases as possible and get a good amount of statistics of what we observe and what we do not observe. Here we mainly discussed about three methods namely "VLF day length method", "night time fluctuation method" and "DLPT and DLDT method". We found that in all these three methods anomalies are present in the signals one to four days before an earthquake. Thus VLF signals may be used for predicting the seismic events. In future for the progress of this subject we have to make a theoretical model to explain these observed anomalies.

References

[1] Bolt, B. A., 'Seismic air waves from the Great 1964 Alaskan Earthquake', *Nature*, 202, 1095-1096, (1964).

[2] Davis, K., Baker, D. M., 'Ionospheric effects Observed around the time of the Alaskan Earthquake of March 28,1964', *J. Geophysical Res.*, 70, 2251-2253, (1965).

[3] Row, R. V., 'Acoustic-Gravity Waves in the Upper Atmosphere due to a Nuclear Detonation and an Earthquake', *J. Geophys. Res.*, 72, 1599-1610, (1967).

[4] Yuen, P. C., Weaver, P. F., Suzuki, R. K., Furumoto, A. S., 'Continuous traveling coupling between Seismic Waves and the Ionosphere Evident in May 1967 Japan Earthquake Data', *J. Geophys. Res.*, 74(9), 2256-2264, (1969).

[5] Tanaka, T., Ichinose, T., Okuzawa. T., Shibata, T., Sato, Y., Nagasawa., C. and Ogawa, T., 'HF-Doppler observations of acoustic waves excited by the Urakawa-oki earthquake on 21 March 1982', *J. Atmos Terr. Phys.*, 46, 233-245, (1984).

[6] Gokhberg,M. B., Gufeld, I. L., Rozhnoy, A. A.,Marenko, V. F., Yampolsky, V. S., Ponomarev, E. A., 'Study of seismic influence on the ionosphere by super long-wave probing of the Earthionosphere waveguide', *Phys. Earth Planet. Inter.*, 57, 64-67, (1989).

[7] Hayakawa,M.,Molchanov, O. A., Ondoh, T., and Kawai, E., 'The precursory signature effect of the Kobe earthquake on VLF subionospheric signals', *J. Comm. Res. Lab., Tokyo*, 43, 169-180, (1996).

[8] Molchanov, O. A., and Hayakawa, M., 'Subionospheric VLF signal perturbations possibly related to earthquakes', *J. Geophys. Res.*, 103, 17489-17510, (1998).

[9] Chakrabarti, S., Sasmal, S., Saha, M., Khan, M., Bhowmik, D. and Chakrabarti, S. K., 'Unusual behaviour of D-region ionization time at 18.2kHz during seismically active days', *Indian J. Phys.*, 81, 531-538, (2007).

[10] Chakrabarti, S. K., Sasmal, S., and S.K. Chakrabarti., 'Ionospheric anomaly due to seismic activities - II: Evidence from D-Layer preparation and disappearance times', *Nat. Haz. Earth. Sys. Sc.*, 10, 1751-1757, (2010).

[11] Sasmal, S., and Chakrabarti, S.K., 'Ionospheric Anomaly due to Seismic Activities -I: Calibration of the VLF signal of VTX 18.2KHz Station From Kolkata and Deviation During Seismic events', *Nat. Hazards Earth Syst. Sci.*, 9, 1403-1408, (2009).

[12] Ray, S., Chakrabarti, S. K., Sasmal, S. and Choudhury, A. K., 'Correlations between the Anomalous Behavior of the Ionosphere and the Seismic Events for VTX-MALDA VLF Propagation', *AIP Conference Proceeding*, 1286, 298-308, (2010).

[13] Ray, S.; Chakrabarti, S. K., Mondal, S. K., Sasmal, S., 'Ionospheric anomaly due to seismic activities-III: correlation between night time VLF amplitude uctuations and effective magnitudes of earthquakes in Indian sub-continent', *Nat. Hazards Earth Syst. Sci.*, 11, 2699-2704, (2011).

[14] Ray, S., Chakrabarti, S. K., Sasmal, S., 'Precursory effects in the nighttime VLF signal amplitude for the 18th January, 2011 Pakistan earthquake', *Indian J. Phys.*, 86(2), 85-88, (2012).

[15] Ray, S. and Chakrabarti, S. K., 'A study on the behavior of the terminator shifts using multiple VLF propagation paths during Pakistan earthquakes (M=7.4), occurred on 19th Jan., 2011', *Nat. Hazards Earth Syst. Sci.*, 13, 1501-1506, (2013).

Effects of Sudden Stratospheric Warming (SSW) on the Upper Atmosphere

Arnab Sen[1,2,*], Sushanta K. Mondal[1], Sujay Pal[3,4]

[1]Department of Physics, Sidho Kanho Birsha University, Purulia-723104
[2]North East Regional Institute of Education, NCERT, Umiam, Meghalaya-793103
[3]Department of Physics, Srikrishna College, Nadia-702512
[4]Near-Earth Space and Atmospheric Observatory, Kolkata, West Bengal
*Corresponding author: arsenphy@gmail.com

Abstract

Sudden Stratospheric Warming (SSW) is a large-scale meteorological phenomenon that causes rapid warming of the upper and middle stratosphere during winter at an altitude of about 25-45 km primarily over high latitude regions in the northern hemisphere. A rapid rise in stratospheric pressure takes place over the extreme northern latitude during the SSW that lasts for about a few days or weeks. During a SSW, the polar vortex may break or weaken which allows extreme cold air to spill out of the polar region to middle latitudes resulting in adverse weather in the northern US and Europe during winter or early spring. The effects of SSWs are not only limited to the surface but can extend to mid-latitude to equatorial mesosphere-ionosphere regions. This work briefly reviews the SSW mechanism, and its effects on the upper atmosphere including the lower ionosphere.

Keywords: *Polar Vortex, Stratospheric Warming, Planetary Waves, Atmosphere-Ionosphere Coupling, VLF Techniques*

1. Introduction

Sudden Stratospheric Warming (SSW) is considered as one of the most prominent meteorological phenomena in the middle atmosphere. During a SSW, a sudden rise of polar stratospheric temperature (up to 50 K within a few days) is observed while the zonal-mean flow weakens. Major stratospheric warmings are said to be occurred when the zonal-mean winds become easterly at 60° N and 10 hPa level during winter and the zonal-mean temperature gradient at 10 hPa between 60° N and 90° N becomes positive [1,2]. This has a large effect on the weather of northern Europe and US when cold winds come from the polar region, because of the shift from westerly to easterly winds, reducing the temperature of the surface air rapidly in those regions. When there is no reversal of the westerlies observed but still temperature rises rapidly in the upper stratosphere during winter, then a minor SSW is said to have occurred. Thus the minor SSWs are less intense and only slow down the westerlies but do not reverse the direction. The effect on surface weather is much less in case of minor SSW [3]. Typically, five to six major SSWs occur in a decade. Also, significant variability is observed in different decades. According to the observational record, there was a long period with no major warmings (1992–98) and there were periods with major warmings almost every year (the 2000s). Minor warmings do occur in almost every winter [4]. Left panel of Figure 1 shows the zonal temperature anomaly during 2013 SSW. Sudden rise of stratospheric temperature in mid-January is clearly observed at very high latitude at around 10 hPa (~31 km) level. While there is a sudden rise of stratospheric temperature near the pole, there is little decrease in temperature near the equator at the same time as could be seen in the right panel of Figure 1. (Source: https://www.cpc.ncep.noaa.gov/).

SSW is the result of interaction of planetary scale Rossby waves and atmospheric gravity waves with the zonal-mean flow in the stratosphere. These waves are formed in the troposphere primarily due to topography and land-sea contrast and then propagate from the troposphere toward the extra-tropical stratosphere [5]. There are also fluctuations in the temperature gradients either vertically or horizontally or both ways, which also induce planetary waves and gravity waves. Since these waves are mainly forced by topography and land-sea contrasts, these are stronger in the northern hemisphere (more land) than in the southern hemisphere. The forcing of planetary waves is usually not enough in the southern hemisphere to initiate a major SSW. Since the amplitude of planetary waves determines the overall likelihood of SSWs, these events mostly occur during the winter in the northern hemisphere [6]. Although the major SSWs occur mostly in the northern hemisphere (approximately six events in a decade), the first major SSW was observed in the southern hemisphere in late September, 2002 [7].

The Stratosphere over the polar region is characterized by a strong west-to-east (westerly) cold polar vortex. The polar vortex is a large-scale low pressure region at very high latitude, near the north pole and south pole. The altitude may extend

from tropopause (~10 km) to stratopause (~55 km). Each low pressure polar vortex has a diameter of around 1000 km and rotates counter-clockwise at the north pole and clockwise at the south pole and their rotation is driven by Coriolis effect similar to cyclonic storms. The polar vortices strengthen in the winter due to large temperature difference between the equator and pole and weaken in the summer. When the polar vortex is strong, the polar jet stream stays near the pole and exhibits a zonal flow with less meandering. A weakened polar vortex in winter often disturbs the winter weather because the cold air is pushed towards south reducing the temperature of surface air rapidly during winter. When planetary waves propagate from troposphere to stratosphere in winter and interact with the polar vortex, they deposit their westward angular momentum into it. Therefore the angular momentum of the polar vortex decreases. Since the adiabatic thermodynamic process is involved and the potential vorticity (PV) is a conserved quantity in adiabatic process, the vortex has to deform to achieve this either by getting displaced from the pole ("wave-1" warming) or by getting split into two ("wave-2" warming) [5].

The effects of SSWs are not only confined to the polar stratosphere but extend to earth's surface as well as to mesosphere and beyond. The ionosphere also responds to the changing dynamics and energetics of the lower-lying atmosphere during the SSWs. The amplified planetary waves and atmospheric gravity waves propagate upward through the stratosphere to the mesosphere that lie in the lower ionosphere and dissipate its energy during SSWs. This plays a significant role in altering the E- and F-region dynamos [8]. The ionosphere experiences changes in the distribution of ionization not only at high latitude due to the processes taking place locally during SSWs but also at mid and lower latitude regions. These changes in the ionosphere due to the SSW events can be detected prominently if there is prolonged solar minimum and quiet geomagnetic activities during this period. SSW couples the atmospheric layers of all latitudes through the interaction of planetary waves with the tidal and small-scale gravity waves components [9,10,11,12,13].

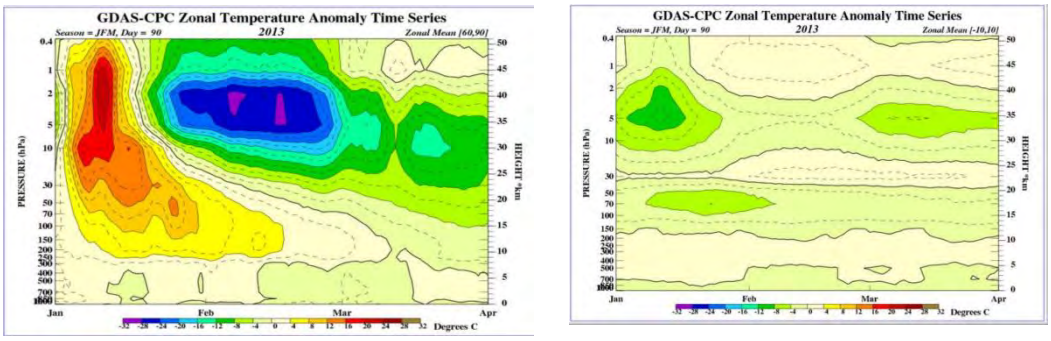


Figure 1. (Left) SSW event during 2013. Sudden rise of stratospheric temperature in mid-January was observed at very high latitude at around 10 hPa level. (Right) Sudden increase of stratospheric temperature near the pole but little decrease in temperature near the equator at the same time. (Source: https://www.cpc.ncep.noaa.gov/)

2. Brief Historical Background

A SSW event was first observed by Prof. Richard Scherhag in 1952 using radiosonde measurement above Berlin, Germany. In the early 1950s, when the knowledge of the stratosphere was inadequate, Prof. Scherhag started exploring the stratosphere using radiosondes at the Free University of Berlin. In 1951, he started using an improved version of radiosonde which would allow reliable measurements of temperature up to a height of 40 km or more. He reported a sudden increase of stratospheric temperature as explosive warming in the winter of January 1952. The warming was too strong to be explained by advection and Scherhag reported this as a Berlin phenomenon since it was observed above Berlin [14]. This was followed by another stronger warming at 10 hPa in February 1952, almost a month later. The reversal of circulation (westerlies to easterlies) in the middle atmosphere was observed [15]. Figure 2 shows the temperature variation at different heights (hPa) during the second Berlin phenomenon [14]. The sudden stratospheric temperature rise was a surprising observation at that time since the temperatures could not rise to such high values during winter [4, 14]. Sudden warming of the stratosphere during winter was initially thought to be the effect of severe solar eruption [16, 17]. But, it is now well known that SSWs are not triggered by solar activities alone. Another stratospheric warming was reported by the British Meteorological Office in February 1951 from the measurements using radiosonde and radar over England and Scotland. The reversal of the lower stratospheric winds (westerlies to easterlies), which again turned into westerlies before the summer, had been observed [6,18]. Scherhag continued to study the SSW events at the Free University of Berlin where he formed a group of meteorologists. This group mapped the stratospheric temperature in the northern hemisphere up to 10 hPa using radiosondes and rocketsondes. The work of this group revealed more details about the nature and evolution of SSWs. After the initial reporting in 1952, the next strong SSWs were reported only in the winters 1956-57 and 1957-58 [19] since the strong SSW events do not happen every year (typically 5-6 events in a decade).

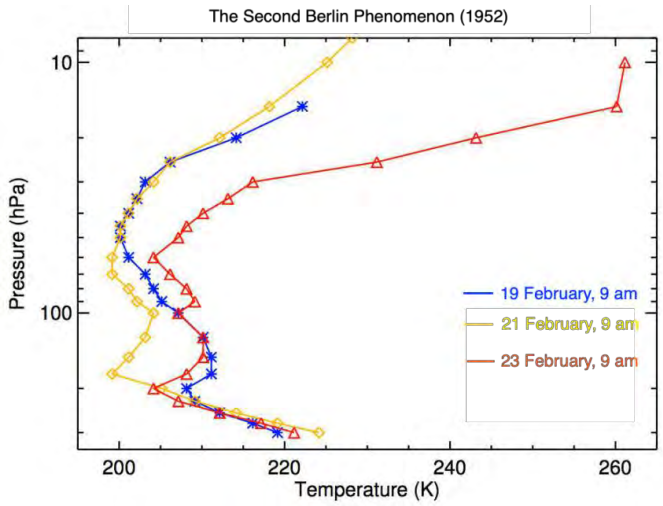


Figure 2. Temperature variation at different heights (hPa) during the second Berlin phenomenon with a temperature maximum of -260.6 K (a warming of ~37K within 2 days) at 10 hPa on February 23 [adapted from 14, 15].

During the International Geophysical Year (IGY) (July 1957 to December 1958), the number of radiosonde balloons reaching over the 10 hPa level increased significantly. Stratospheric maps of the northern hemisphere started to be published on a daily or weekly basis for 100 hPa, 50 hPa, 30 hPa, and 10 hPa by several groups. This provided great scope for knowing the nature and evolution of SSWs [6]. The International Years of the Quiet Sun (IQSY) was observed during 1964–65. The World Meteorological Organization (WMO), Commission for Atmospheric Sciences (CAS) initiated an international SSW monitoring program across the meteorological centers at Melbourne, Tokyo, Berlin, and Washington D. C. using radiosonde and rocketsonde to obtain an increased number of high-altitude soundings during sudden stratospheric warmings. The program reported information on the intensity and movement of the warmings over those centers. According to the WMO/IQSY [20], SSWs were classified based on their time of occurrence and intensity. These are named as major warmings, minor warmings, mid-winter warmings and final warmings. During major warmings, the zonal-mean wind or zonal-westerlies near 10 hPa level reverses its direction, which causes the breakdown of the polar vortex. The minor SSWs are less intense and only slow down the westerlies but do not reverse their direction. In minor warmings also, the polar temperature gradient reverses similarly to the major warmings. A final warming stated to have occurred on this transition of winter and summer, when the westerlies becomes easterlies and would remain so until the next winter. It is called final warming because another warming is not possible in the next summer, and it is the final warming of that winter.

The coupling of troposphere and stratosphere during SSWs was pointed out in some early studies. The sudden warmings are caused by the stationary planetary waves generated in the troposphere. These waves amplify in the stratosphere preceding the onset of sudden warming and also create blockings in the troposphere simultaneously [21,22]. Figure 3 illustrates an example of stratosphere-troposphere coupling at 10 hPa level maps during 1963 SSW. The left panel shows the 10 hPa height map at the beginning (January 18) of 1963 SSW. The middle panel indicates the 10 hPa map at the peak (January 27) of the SSW and the right panel is the surface pressure map on January 31, 1963 [23]. According to Labitzke [24], the tropospheric blockings happened about ten days after a stratospheric warming. There are significant thermodynamic disturbances during winter stratospheric warmings which cover a height ranging from the troposphere to the upper mesosphere and the lower thermosphere. Quiroz [25] found tropospheric temperature variations after stratospheric warming.

Stratospheric temperatures were started to be measured using the Stratospheric Sounding Units (SSU) on board the NOAA operational satellites in 1979. It provided global stratospheric temperature data at a higher altitude (above the lower stratosphere). The data from these measurements gave insights into the stratospheric and tropospheric dynamics with its implications on weather forecasting. McIntyre and Palmer [26] gave maps of large-scale distribution of potential vorticity in the middle atmosphere. These maps demonstrate the breaking of planetary-scale Rossby waves from the troposphere into the stratosphere causing the sudden warmings. Since then the satellite data have become the key for finding more about SSW events.

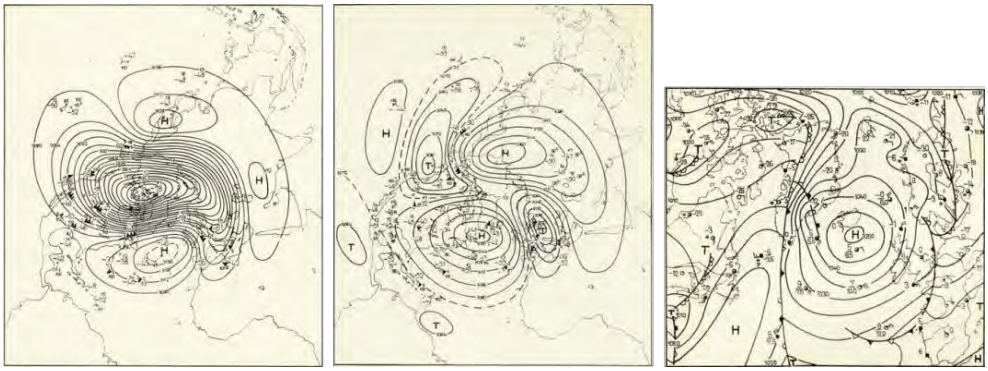


Figure 3. 10 hPa height map during January 1963 SSW: Left- January 18, 1963 (beginning of SSW); Middle- January 27, 1963 (peak of SSW); Right- Sea level pressure on January 31, 1963 (adapted from [23], [6] © Springer)

Some studies in the 1960s and 1970s mentioned the concept of stratosphere-troposphere coupling. Many compelling theories evolved during this period had stated that the sudden warmings are caused by the upward propagation of planetary-scale waves originating from the troposphere. The direction and strength of the winds in the stratosphere control the ability of upward propagation of these waves. During summer, when stratospheric winds are easterly, these disturbances cannot propagate upward and interact with the stratospheric winds. But during winter, when stratospheric winds are westerly, the planetary-scale wave disturbances can propagate from the troposphere to the stratosphere, and interact with the zonal-mean wind. This is the reason that SSWs occur in winter but not in summer.
According to linear planetary-wave theory [27], planetary-scale waves can propagate from troposphere to stratosphere when the stratospheric winds are moderately westerlies. These waves cannot propagate in easterlies and also when the velocity of the zonal mean wind does not reach the critical level. Matsuno [28] gave the critical-layer theory for the tropospheric origin of SSWs, which is now widely accepted today. In his paper, Matsuno presented a model which has two aspects, the vertical propagation of the planetary waves forced from below (troposphere) and their interaction with zonal winds in the stratosphere. The rise in temperature near the pole and the deceleration of the zonal-mean wind are attributed to the vertically propagating planetary waves forced from the troposphere. Matsuno's model suggests that SSWs require a pulse of anomalously intense wave forcing from the troposphere to initiate. Lately, Matsuno's theory was confirmed by many other studies using satellite data. But Matsuno's simplified model did not consider the wave-wave interactions which certainly is an important factor for stratospheric variability. Another aspect in his theory that the stratosphere is initially zonally symmetric, also needed to be amended since the initial asymmetry of the stratosphere is also relevant.

Many differences have been noted in the frequency and nature of SSWs observed since the 1950s. These variabilities are random in nature and the likelihood of occurrence of SSW is influenced by many external factors including the Quasi-Biennial Oscillation (QBO), El Niño Southern Oscillation phenomenon (ENSO), the 11-year solar cycle, the Madden-Julian Oscillation (MJO). Studies with general circulation models have shown that the occurrence of a major SSW is enhanced during both the warm and cold phase of ENSO processes. Quasi-biennial oscillation (QBO) influences the zonal-wind structure in the stratosphere which affects the propagation of planetary waves from the troposphere [4].

3. Effects of SSW Events

Though the SSWs are stratospheric phenomena and cause rapid warming of the upper and middle stratosphere over primarily high latitude region in the northern hemisphere, there are significant thermodynamic disturbances during SSWs over a height ranging from the troposphere to the upper mesosphere and lower thermosphere.

3.1. Effects on the Surface Weather

The impact of SSWs on the troposphere was mentioned by Quiroz [25]. He reported tropospheric warming in the polar region and cooling in middle latitudes after the SSW in December-January 1976-77. The anticyclonic circulation anomalies at the high latitudes associated with the SSW descended down to the earth's surface. Baldwin and Dunkerton [29,30] found that the trend of downward propagation of extratropical zonal wind anomalies in the northern hemisphere used to appear within 1-2 weeks after an SSW. These tropospheric anomalies tend to persist for a long time (about 60 days) which provides a source of memory for seasonal weather forecasts. The tropospheric circulation shifts the Northern Annular Mode (NAM, also known as Arctic Oscillation (AO)) toward its negative phase, which implies that circulating winds around the Arctic get weakened and more distorted allowing southward migration of colder, arctic air masses. This results in the decrease in surface temperature over North America and Western Europe and warm anomalies over Newfoundland, Greenland, and Southern Europe [4, 31]. Figure 4 shows the phase of Northern Annular Mode (NAM)/ Arctic Oscillation (AO).

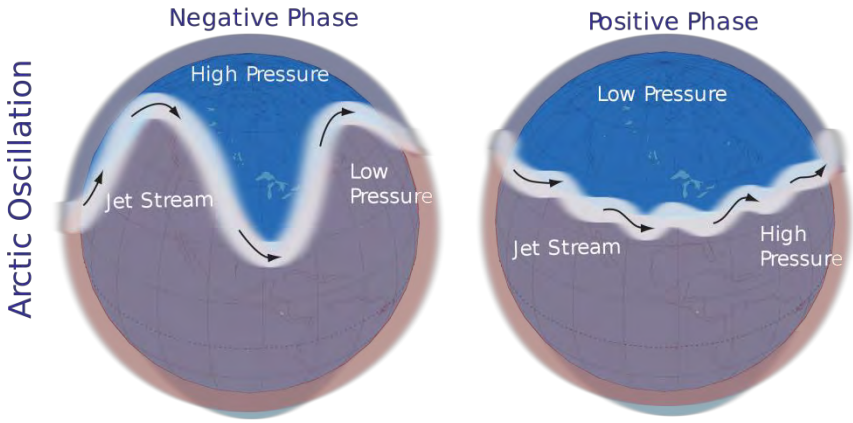


Figure 4. Positive and Negative phases of Northern Annular Mode (NAM)/ Arctic Oscillation (AO). The positive phase indicates a strong polar vortex and the negative phase indicates a weakened polar vortex, more susceptible of breaking during SSW. (Source: National Climatic Data Center, National Oceanic and Atmospheric Administration.)

3.2. Effects on the Upper Atmosphere

Atmospheric waves, mainly in the form of planetary waves, tides, and small-scale gravity waves, couple the stratosphere-mesosphere to ionosphere. These waves travel vertically and convey the momentum from the troposphere to high altitude (stratosphere, mesosphere, and thermosphere). During SSW events, the tropospheric transient planetary waves propagate vertically upward into the stratosphere and interact with the zonal-mean flow. This induces a downward circulation in the stratosphere and an upward circulation in the mesosphere [32]. The stratosphere becomes warmer by several tens of Kelvin (K) at very high latitude (near north-polar region) due to adiabatic heating caused by the downward circulation. On the other hand, the mesopause becomes cooler due to adiabatic cooling caused by the upward circulation in the mesosphere. Thus, the SSW events are usually accompanied by mesospheric coolings [33,34,35]. SSWs lead to the propagation of gravity waves in the middle atmosphere. According to Holton [36], the reversal of polar stratospheric winds reduces the transmission of gravity waves into the mesosphere. The changes in the mesosphere-lower thermosphere (MLT) during SSWs are primarily due to the changes in gravity wave drag. The changes in the MLT region during SSWs are only weakly correlated with the changes in the stratosphere [37]. Therefore the coupling between the stratosphere and MLT region remains a complex process due to the lack of any direct linear correspondence.

The ionosphere which is embedded within the mesosphere and thermosphere, also responds to the changing dynamics and energetics of the lower-lying atmosphere. The ionospheric absorption of radio waves in the middle latitudes shows a distinct winter anomaly and evidence of stratospheric warmings [38]. The high and medium frequency radio waves get absorbed in the lower ionosphere during winter anomaly. These anomalies were weakly correlated to the solar and magnetic activities but occurred largely due to increase of upper stratospheric temperature. Ionospheric effects during SSW events have been extensively studied by ionosonde and GPS TEC method ([13], [39] and references therein). Funke et al. [51] found observational evidence of dynamic coupling between the lower atmosphere and upper atmosphere (upto 170 km in thermosphere) during the January, 2009 major sudden stratospheric warming using temperature data from Michelson interferometer for passive atmospheric sounding (MIPAS) on board ESA's Envisat satellite. Using the TEC data from the global network of GPS receivers, a large scale distribution of electron density in the daytime ionosphere due to vertical ion drifts during SSW events has been observed. The analysis of TEC data during SSW becomes important because of its large scale variation compared to its pre-SSW behaviour. The semidiurnal variability of TEC and the signature of its perturbation before and after SSW give a scope for studying predictability of SSW [52]. Pedatella and Forbes [39] used GPS TEC data and found that the non-migrating semi-diurnal tides generated by the interaction of planetary waves and the migrating diurnal tides are related to the coupling of SSWs and the ionosphere. Significant variability has been observed in GPS TEC semidiurnal tide during the SSW period. But it should also be noted that other phenomena e.g., geomagnetic variations or solar activities may also cause the observed perturbations. The vertical total electron content (VTEC) during the major SSW event of January 2009 was studied using GPS receivers over 17 GPS locations covering from equatorial to mid latitude regions, which showed depression in TEC variations during few days following the SSW peak [13]. This observed TEC depletion was the characteristic of the SSW event since the period was geomagnetically quiet.

Ionospheric disturbances due to SSW may occur simultaneously or precede the maximum stratospheric disturbances or may occur after the maximum stratospheric warming. This provides potentially improved forecasting of ionosphere variability during SSW. Since the ionosphere is forced externally during the SSW events, it is less sensitive to the initial conditions compared to the troposphere-stratosphere [40]. This provides predictability of the SSWs in the absence of external factors like solar activity or other effects from the lower-lying atmosphere. Wang et al. [41] and Pedatella et al. [42] showed that ionospheric variability could be forecast ~10 days before the SSW. Therefore SSWs may be forecast from the variability of the ionosphere or vice versa.

4. VLF Remote Sensing and Ionospheric Response to SSWs

Very Low Frequency (VLF) radio waves with frequency range between 3 – 30 kHz and wavelength range 100 – 10 km respectively may originate from transmitters used for naval communication or from natural sources like lightning discharges, meteor echoes, Aurora Borealis, etc. The VLF band is used for radio navigation and communication systems, especially in naval communication. Many countries around the world operate VLF transmitters for navigation and military communication. The transmitted signal travels through the earth-ionosphere waveguide formed by the lower ionosphere (60–100 km) and earth's surface. The lower most region of the ionosphere reflects the VLF signals back to earth. During day-time, the ionospheric D-layer is formed. Thus reflection of VLF waves occurs from the E- layer in the nighttime and from the D-layer in the daytime. Any change in the ionization of the lower ionosphere can be detected on the ground by receiving the VLF signals. The VLF signals exhibit regular diurnal and seasonal variation because of the ionospheric variation due to solar ionization. Any perturbation in the ionosphere can easily be detected by studying the recorded signals on the ground.

Since there is a change in the distribution of ionization over low to high latitude due to the processes that take place locally during SSWs, the VLF signals are expected to detect the disturbances caused by SSW events when the planetary waves couple the neutral atmosphere and ionosphere. Very few reports exist about the D-region and VLF signal disturbances caused by SSW events. Belrose [43] first reported a possible correlation of VLF phase advancement with medium frequency (300 kHz to 3 MHz) radio wave absorption associated with the SSW events of 1952. A significant increase of D-region electron density by a factor of 10 at 80 km was noted associated with the event. Larsen [44] investigated short path VLF amplitude and phase along with ionosonde data at high latitude during the SSW event of 1969 and noted that there were no amplitude change except small phase retardation which was explained by a 3 km increase in VLF reflection height. Using the full wave computer model he concluded that there were no large changes in D-region electron density as observed during the 1952 event since SSW events do not always exhibit the same development pattern. Thus the effects in the upper mesosphere and ionosphere due to SSW events are not always the same. Cavalier and Deland [45] showed positive VLF phase fluctuations with temperature at 60 km altitude during the SSW event of 1970/1971. Subsequent studies [46,47] further showed D-region electron density enhancement during SSW events and revealed the existence of planetary wave influence in the VLF signals. Although, VLF anomaly due to the SSW event is not completely understood yet both by observation and theoretical studies.

Pal et al. [48,49] studied the VLF/LF signal disturbances caused by the lower ionospheric variability associated with the major SSW event in January 2009 during quiet solar and geomagnetic conditions. The VLF/LF disturbances near the polar vortex and also far from it were studied using the VLF/LF network data from Germany and Japan. A significant increase or decrease (or sometimes both) of VLF/LF signals amplitude during the peak of an SSW event, depending on the VLF/LF signals propagation path, was found. Figure 5 shows the daytime and nighttime amplitude fluctuations in the VLF signal from the NAA and NRK transmitters received in Kiel, Germany during the entire SSW period. Further, analysis of TIMED/SABER temperature and pressure data for NRK-Kiel (high latitude to middle latitude) propagation path shows a decrease in mesospheric temperature and increase in mesospheric pressure during the SSW period. Although, there was a delay of 2 days in mesospheric cooling from the peak of stratospheric warming. The authors also calculated the VLF signal amplitude using the Long Wave Propagation Capability (LWPC) code during the SSW period and showed that there was an enhancement of D-region electron density along with increased electron-neutral collision frequency. Sen et al. [50] also studied the effects of the SSW event of 2016 using several VLF propagation paths and showed significant perturbations due to the 2016 event. It is believed that the interaction of upward propagating tropospheric transient planetary Rossby waves with the tidal waves and small-scale gravity waves to be the main cause of electron density fluctuations in the ionosphere which in-turn affected the VLF/LF propagation in the earth-ionosphere waveguide.

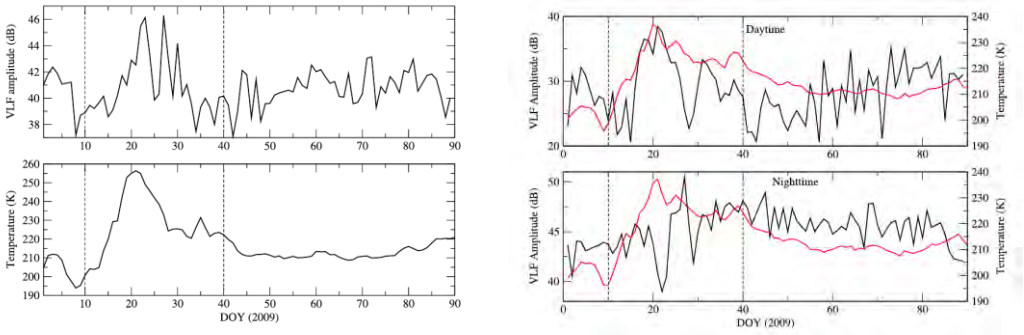


Figure 5. (Left) Nighttime VLF amplitude of the NAA signal received at (top) Kiel, Germany was compared with the (bottom) stratospheric temperature variation over the middle of propagation path. (Right) Daytime and nighttime amplitude of the NRK signal (black) received at the same place was compared with stratospheric temperature variation (red) at 10 hPa level [47].

5. Summary

SSW remains one of the most dramatic and prominent meteorological phenomena in the middle atmosphere. In this paper, we have reviewed some important scientific aspects of SSW events including early results. First, we have attempted to present a brief historical overview of SSW observations in the stratosphere and its characteristics. The physical mechanism and hypothesis of SSW formation are discussed. An SSW requires a pulse of anomalously intense wave forcing from the troposphere to initiate. The simple model of Matsuno about vertical propagation of planetary-scale Rossby waves forced from the troposphere which interact with zonal winds in the stratosphere can describe SSW formation. This model is now widely accepted and experimentally verified to some extent. Although the vertical propagation of planetary Rossby waves is allowed only under moderately westerlies during northern winter. The impact of an SSW on the troposphere can persist for several weeks after the SSW and in general warming in the polar region and cooling in the mid-latitude regions are observed. SSW being a global-scale phenomenon, can couple the high latitude upper atmosphere to low latitude upper atmosphere. Many observations found mesospheric cooling before or during or after the maximum warming in the mid-latitude region. The correlation of the connection between stratospheric warming and changes in the mesosphere-ionosphere can be weak or can depend on various factors like solar and geomagnetic conditions, mesospheric wind circulation, QBO phase, ENSO, etc. High-latitude and mid-latitude D-region ionospheric disturbances during an SSW in the northern hemisphere can be measured using the ground-based VLF radio remote sensing method. Several SSW events had been detected and analyzed by VLF radio signals in different solar conditions, but prominent and distinct VLF perturbations were observed during quiet solar and geomagnetic conditions. Though SSW affects the equatorial ionosphere, there is no evidence of low-latitude VLF signal disturbances due to an SSW to date. In order to understand the complete effects of SSW in the whole atmosphere from the troposphere to the ionosphere on a global scale, numerous ground and space based measurements are necessary.

Acknowledgements

A. Sen acknowledges the support received from Sidho Kanho Birsha University and NERIE-NCERT in completing this study. S. Pal acknowledges the research support from the SERB grant SRG/2020/001104.

References

1. Charlton, A. J., & L. M. Polvani, (2007). A New Look at Stratospheric Sudden Warmings. Part I: Climatology and Modeling Benchmarks, Journal of Climate, 20(3), 449-469.
2. Medvedeva, Irina & Semenov, Anatoly & Chernigovskaya, M. & Perminov, Vladimir. (2012). Studying Manifestations of 2008–2011 Sudden Stratospheric Warmings in East-Siberia and European Russia. Geophysica. 48. 91–103.
3. Labitzke, K., (1977), Interannual Variability of the Winter Stratosphere in the Northern Hemisphere, Monthly Weather Review, 762-770, https://doi.org/10.1175/1520-0493(1977).
4. O'Neill, A. J., Charlton-Perez, L.M., Polvani, L., Middle Atmosphere, Stratospheric Sudden Warmings, Editor(s): Gerald R. North, John Pyle, Fuqing Zhang, Encyclopedia of Atmospheric Sciences (Second Edition), Academic Press, (2015), Pages 30-40, ISBN 9780123822253, https://doi.org/10.1016/B978-0-12-382225-3.00230-9.
5. Shepherd, T. G., (2007), Transport in the Middle Atmosphere, Journal of the Meteorological Society of Japan. Ser. II, 85B, P165-191, https://doi.org/10.2151/jmsj.85B.165.
6. Baldwin, M. P., B., Ayarzagüena, T., Birner, N., Butchart, A. H., Butler, A. J., Charlton-Perez, et al., (2021). Sudden stratospheric warmings. Reviews of Geophysics, 59. https://doi.org/10.1029/2020RG000708.
7. Charlton, A. J., A. O'Neill, W. A. Lahoz, and P. Berrisford, (2005), The Splitting of the Stratospheric Polar Vortex in the Southern Hemisphere, September 2002: Dynamical Evolution, Journal of the Atmospheric Sciences 62, 3: 590-602, https://doi.org/10.1175/JAS-3318.1
8. Yadav, S., T. K., Pant, R. K., Choudhary, C., Vineeth, S., Sunda, K. K., Kumar, Mukherjee, S., (2017), Impact of sudden stratospheric warming of 2009 on the equatorial and low-latitude ionosphereof the Indian longitudes: A case study, Journal of Geophysical Research: Space Physics, 122, 10, 486–10,501. https://doi.org/10.1002/2017JA024392
9. Liu, H. L., and R. G. Roble (2005), Dynamical coupling of the stratosphere and mesosphere in the 2002 Southern Hemisphere major stratospheric sudden warming, Geophys. Res. Lett., 32, L13804, doi:10.1029/2005GL022939.
10. Goncharenko, L., and S. R. Zhang, (2008), Ionospheric signatures of sudden stratospheric warming: Ion temperature at middle latitude, Geophys. Res. Lett., 35, L21103, doi:10.1029/2008GL035684.
11. Fuller-Rowell, T., R. Akmaev, F. Wu, M. Fedrizzi, R. A. Viereck, and H. Wang (2011), Did the January 2009 sudden stratospheric warming cool or warm the thermosphere, Geophys. Res. Lett., 38, L18104, doi:10.1029/2011GL048985.
12. Laskar, F. I., D. Pallamraju, and B. Veenadhari, (2014), Vertical coupling of atmospheres: Dependence on strength of sudden stratospheric warming and solar activity, Earth Planets Space, 66, 94.
13. Fagundes, P. R., L. P. Goncharenko, A. J. de Abreu, K. Venkatesh, M. Pezzopane, R. de Jesus, M. Gende, A. J. Coster, and V. G. Pillat (2015), ionospheric response to the 2009 sudden stratospheric warming over the equatorial, low, and middle latitudes in the South American sector, J. Geophys. Res. Space Physics, 120, 7889–7902, doi:10.1002/2014JA020649.
14. Scherhag, R. (1952b). Einu von Sonneneruptionen auf Stratosphrenwetternachgewiesen. Wetterkarte des Deutschen Wetterdienstes in der US-Zone, 14. Mrz 1952
15. Wiehler, J., (1955), Die ergebnisse der berliner radiosonden-hochaufstiege der jahre 1951-1953, Met. Abh. FU-Berlin, Band III.
16. Scherhag, R., (1952a): Die explosionsartigen Stratosphärenerwärmungen des Spätwinter 1951/1952 (The explosive warmings in the stratosphere of the late winter 1951/1952). Ber. Dtsch. Wetterdienstes U.S. Zone, 38, 51–63.
17. Willett, H. C. (1952), Atmospheric reactions to solar corpuscular emissions, Bull. Amer. Meteor. Soc, 33 (6), 255-258.
18. Scrase, F. J., (1953), Relatively high stratospheric temperatures of February 1951, Meteorol. Mag., 82, 19–27.
19. Teweles, S., & F. G., Finger, (1958). An abrupt change in stratospheric circulation beginning in mid-January 1958. Mon. Wea. Rev, 86, 2328.

20. WMO/IQSY. (1964). International Years of the Quiet Sun (IQSY) 1964-65. Alert messages with special references to stratwarms. WMO/IQSY Report No 6, Secretariat of the World Meteorological Organization, Geneva, Switzerland. World Meteorological Organization
21. Julian, P. R., & K. G., Labitzke, (1965). A study of atmospheric energetics during the January February 1963 stratospheric warming. J. Atmos. Sci., 22 , 597-610. doi: 10.1175/1520-0469
22. Tung, K. K., & R. S., Lindzen, (1979), A Theory of Stationary Long Waves. Part I: A Simple Theory of Blocking, Monthly Weather Review, 107(6), 714-734.
23. Scherhag, R. (1965). Neuere Ergebnisse der Meteorologie der Hochatmosphre. Die Naturwissenschaften, 11 , 279-286.
24. Labitzke, K. (1965). On the mutual relation between stratosphere and troposphere during periods of stratospheric warmings in winter. Journal of Applied Meteorology.
25. Quiroz, R. S. (1977). The tropospheric-stratospheric polar vortex breakdown of january, Geophys. Res. Lett., 4, 151-154. doi: 10.1029/GL004i004p00151.
26. McIntyre, M., T., Palmer, (1983), Breaking planetary waves in the stratosphere. Nature 305, 593–600. https://doi.org/10.1038/305593a0
27. Charney, J. G., and P. G. Drazin, (1961): Propagation of planetary-scale disturbances from the lower into the upper atmosphere. J. Geophys. Res., 66, 83–109, doi:10.1029/JZ066i001p00083.
28. Matsuno, T. (1971), A dynamical model of the stratospheric sudden warming, J. Atmos. Sci., 28(8), 1479–1494.
29. Baldwin, M. P., and T. J. Dunkerton, (1999): Propagation of the Arctic Oscillation from the stratosphere to the troposphere. J. Geophys. Res., 104, 30 937–30 946.
30. Baldwin, M. P., and T. J. Dunkerton, (2001): Stratospheric harbingers of anomalous weather regimes. Science, 294, 581–584.
31. Thompson, D. W. J., M. P., Baldwin, and J. M., Wallace, (2002), Stratospheric Connection to Northern Hemisphere Wintertime Weather: Implications for Prediction. Journal of Climate 15, 12, 1421-1428, doi: https://doi.org/10.1175/1520-0442(2002)015.
32. Liu, H. L., and R. G. Roble (2002), A study of a self-generated stratospheric sudden warming and its mesospheric-lower thermospheric impacts using the coupled TIME-GCM/CCM3, J. Geophys. Res., 107(D23), 4695, doi:10.1029/2001JD001533.
33. Mukhtarov, P., et al., (2007), Large-scale thermodynamics of the stratosphereand mesosphere during the major stratospheric warming in 2003/2004, J. Atmos. Sol. Terr. Phys.,69, doi:10.1016/j.jastp.2007.07.012.
34. Whiteway, J. A. and A. I. Carswell, (1994), Rayleigh Lidar Observations of Thermal Structure and Gravity Wave Activity in the High Arctic during a Stratospheric Warming, J. Atmos. Sci., 51, 3122–3136, https://doi.org/10.1175/1520-0469(1994)051.
35. Cho, Y. -M., Shepherd, G. G., Won, Y. -I., Sargoytchev, S., Brown, S., Solheim, S., (2004). MLT cooling during stratospheric warming events. Geophysical Research Letters. 31. 10.1029/2004GL019552.
36. Holton, J.R., (1983), The influence of gravity wave breaking on the general circulation of the middle atmosphere, J. Atmos. Sci., 40, 2497–2507.
37. Smith, A. K., N. M., Pedatella, & Z. K. Mullen, (2020), Inter-hemispheric Coupling Mechanisms in the Middle Atmosphere of WACCM6, J. Atmos. Sci., 77 (3), 1101-1118. doi: 10.1175/JAS-D-19-0253.1
38. Shapley, A., W., Beynon, (1965), 'Winter Anomaly' in Ionospheric Absorption and Stratospheric Warmings, Nature 206, 1242–1243, doi:10.1038/2061242a0
39. Pedatella, N.M. and Forbes, J.M, 2010, Evidence for stratosphere sudden warming-ionosphere coupling due to vertically propagating tides, Geophys. Res. Lett., vol. 37, L11104.
40. Siscoe, G., & S. C., Solomon, (2006). Aspects of data assimilation peculiar to space weather forecasting, Sp. Weather, 4(4).
41. Wang, H., R. A., Akmaev, R. A., Fang, T. J., Fuller-Rowell, F., Wu, Maruyama, N., & Iredell, M. D., (2014), First forecast of a sudden stratospheric warming with a coupled whole-atmosphere/ionosphere model IDEA, J. Geophys. Res. Sp. Phys., 119 (3), 2079-2089. doi: 10.1002/2013JA019481.
42. Pedatella, N. M., Liu, H. L., Marsh, D. R., Raeder, K., Anderson, J. L., Chau, J. L., Siddiqui, T. A., (2018). Analysis and Hindcast Experiments of the 2009 Sudden Stratospheric Warming in WACCMX+DART, J. Geophys. Res. Sp. Phys., 123 (4), 3131-3153. doi: 10.1002/2017JA025107
43. Belrose, J., The "Berlin" Warming. Nature 214, 660–664 (1967) doi:10.1038/214660a0
44. Larsen, T. R., (1971), Short path VLF phase and amplitude measurements during a stratospheric warming in February 1969, J. Atmos. Sol. Terr. Phys., 33, 1251–125.
45. Cavalieri, D. J., R. J. Deland, T. A. Potemra, and R. F. Gavin (1974), The correlation of VLF propagation variations with atmospheric planetary-scale waves, J. Atmos. Terr. Phys., 36(4), 12 561–574.
46. Muraoka, Y. (1983), Winter anomalous effects of mode conversion observed in mid-latitude VLF transmission, J. Geophys. Res. Sp. Phys., 88(A1), 311–317.
47. Muraoka, Y. (1985), The D-region winter anomaly and dynamical effects of atmospheric planetary4 scale waves., J. Geomagn. Geoelectr., 37(5), 509–530.
48. Pal, S., Y. Hobara, S. K. Chakrabarti, and P. W. Schnoor, 2017a, Effects of the major sudden stratospheric warming event of 2009 on the subionospheric very low frequency/low frequency radio signals, J. Geophys. Res. Space Physics, 122, 7555–7566, doi:10.1002/2016JA023813
49. S. Pal, Y. Hobara, S. K. Chakrabarti, and P. W. Schnoor, Response of the sub-ionospheric VLF/LF signals to the major SSW event of 2009, URSI GASS, 2017, DOI: 10.23919/URSIGASS.2017.8105404, Publisher: IEEE.
50. A. Sen, S. Pal, S. K. Mondal and Y. Hobara, "Mid-latitude and high latitude ionospheric disturbances during Sudden Stratospheric Warming events observed by VLF/LF signals," 2019 URSI Asia-Pacific Radio Science Conference (AP-RASC), 2019, pp. 1-2, doi: 10.23919/URSIAP-RASC.2019.8738682.
51. Funke, B. & López-Puertas, M. & Bermejo Pantaleón, Diego & Garcia-Comas, M. & Stiller, G. & Clarmann, T. & Kiefer, M. & Linden, A.. (2010). Evidence for dynamical coupling from the lower atmosphere to the thermosphere during a major stratospheric warming. Geophysical Research Letters - GEOPHYS RES LETT. 37. 10.1029/2010GL043619.
52. Goncharenko, L. P., J. Chau, H.-L. Liu, and A. J. Coster (2010), Unexpected connections between the stratosphere and ionosphere, Geophys. Res. Lett. , 37 , L10101, doi:10.1029/2010GL043125.

Computer Science & Mathematics

Crosstalk-Avoided Resource Allocation in Spectrally-Spatially Elastic Optical Networks: An Overview

Imran Ahmed[1], Eiji Oki[2], and Bijoy Chand Chatterjee[1,*]

[1]Department of Computer Science, South Asian University, New Delhi, India
[2]Graduate School of Informatics, Kyoto University, Kyoto, Japan

*Corresponding author: bijoycc@ieee.org

Abstract

Spectrally-spatially elastic optical networks (SS-EONs) emerge as a promising solution to satisfy the continuously increasing bandwidth requirement. Inter-core and inter-mode crosstalks are a significant obstacle in SS-EONs, which reduces spectral and spatial resource utilization. This work presents an overview of SS-EONs technologies considering different types of fibers. We present a crosstalk model considering both inter-core and inter-mode crosstalks simultaneously. Furthermore, we discuss routing, spectrum, core, and mode allocation (RSCMA) in SS-EONs. Finally, few challenging issues related to SS-EONs are higlighted.

Keywords: Space division multiplexing, Elastic optical networks, Crosstalks.

1 Introduction

An optical network is a conventional data communication network that uses the fiber optics technology. In an optical network, optical fiber cables serve as the primary communication medium where data is converted and transmitted as light pulses between source and destination nodes [1, 2]. Figures 1(a), (b), and (c) explain the concept of refraction by considering the behavior of light rays associated with plane waves in two different mediums [3]. We consider that refractive index of medium 1 (n_1) is greater than medium 2 (n_2), i.e., $n_2 < n_1$. In Fig. 1(a), if the angle of incidence (θ_1) is less than the critical angle (θ_c), the light ray refracts away from the normal. In Fig. 1(b), if the angle of incidence (θ_1) is equal to the critical angle (θ_c), the light ray refracts at 90° to the normal. In Fig. 1(c), if the angle of incidence (θ_1) is greater than the critical angle (θ_c), the total internal reflection occurs. Critical angle is defined as $\theta_c = \sin^{-1}(\frac{n_2}{n_1})$, where n_1 and n_2 are the refractive indices of the two medium.

A physical structure of an optical fiber is indicated in Fig. 1(d) that consists of core, cladding, and buffer coating, which are explained the following. Core is built from highly purified glass, and the maximum light energy is confined to the core. Cladding protects optical fields from the interference of fiber's outer layers. Buffer layers surround the cladding. These layers play no role in light propagation. They primarily serve to give mechanical support to the glass fiber as well as to guard the fiber from exterior damage. Figure 1(e) expresses the propagation of light inside the core that has two scenarios: (i) the rays which always traverse through the axis of fiber that leads to high optical intensity at the center of the fiber's core; these rays are known as meridional rays, (ii) the rays which never cross the axis of the fiber that leads to high optical intensity towards the rim and low intensity at the center of the fiber; these rays are known as skew rays.

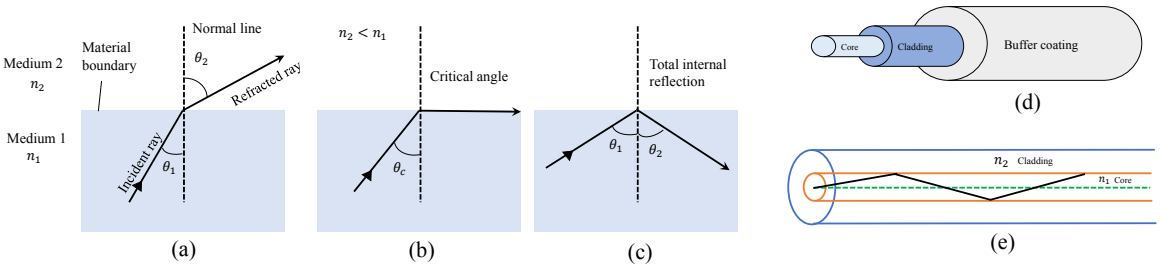


Figure 1: Reflection of light in optical fiber.

The wavelength division multiplexing (WDM) based optical network utilizes the spectrum, which is split into different channels [4, 5]. The international telecommunication union (ITU)-T standards specify the spacing between adjacent channels either 50 GHz or 100 GHz, as indicated in Fig. 2(a). Figure 2(b) shows that a WDM-based optical network has comparatively large frequency spacing between two adjacent channels. If a client needs low bandwidth and no traffic can be transmitted in the sizeable unused frequency gap, most of the spectrum will be wasted. Jinno et al. [6] presented a spectrum

efficient elastic optical network (EON) which utilizes orthogonal frequency-division multiplexing (OFDM) technology to overcome the barriers of conventional optical networks [7,8]. As indicated in Figs. 2(c) and (d), the EON utilizes the flexible spectrum grid, which enhances transmission spectral efficiency than fixed-grid that is used by WDM optical networks.

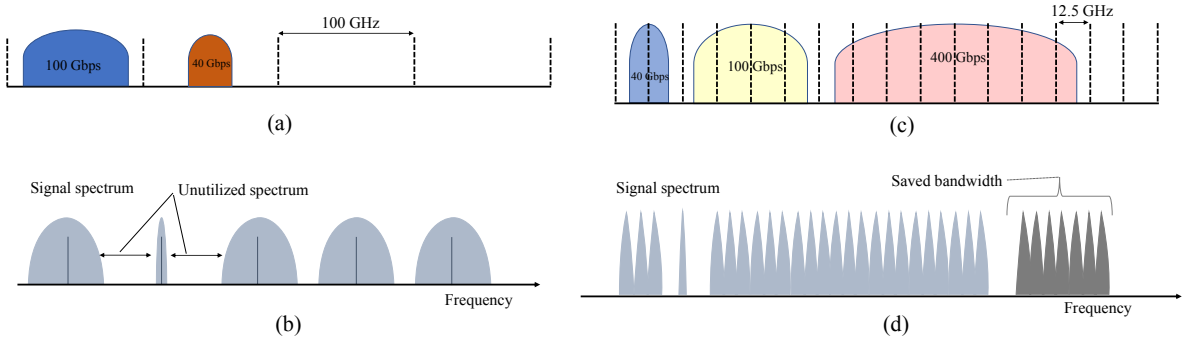


Figure 2: (a) ITU-T fixed grid, (b) spectrum allocation in WDM-based optical networks, (c) flexible grid, and (d) spectrum allocation in OFDM-based optical networks.

In recent times, due to the ever-increasing popularity of different network services, such as ultra-high-definition audio/video, smart homes, and Internet of Things, the overall Internet traffic has been rapidly increasing. However, EONs have reached the transmission capacity limit [9]. According to the Cisco Annual Internet Report, 2018-2023, internationally, the total number of Internet users is expected to increase from 3.9 billion in 2018 to 5.3 billion by 2023 at a compound annual growth rate (CAGR) of 6%, which will enhance transport traffic. As a result, it is required an optical network that supports high-capacity transport traffic [10]. It seems that the conventional optical transport technology will not fulfill the high bandwidth requirement due to physical impairment and electrical bandwidth bottleneck limitation [11]. Therefore, we need to develop a high-capacity transport network. One of the possible solutions is the integration of space division multiplexing (SDM) technology in EONs to improve bandwidth capacity. In the SDM technology, multi-core fiber is used, in which each core can have multiple modes [12–14]. When the SDM technology is incorporated in EONs, emerging technology is formed, which is named spectrally-spatially elastic optical networks (SS-EONs) [15].

In EONs, when spectrum allocation is performed, it must follow the routing and spectrum allocation (RSA) constraints, which are spectrum contiguity and continuity constraints, to satisfy the requested bandwidth requirement [16]. In spectrum contiguity constraint, the required spectrum slots must be contiguous to each other in the spectrum domain. The spectrum continuity constraint ensures that the same spectrum slots are utilized in all links of a lightpath. When spatial (core and mode) dimensions are included in EONs, the RSA problem has been becoming more complex, which is known as the routing, spectrum, core, and mode allocation (RSCMA) problem in SS-EONs. The RSCMA problem has two different phases, which are (i) routing and (ii) spectrum, core, and mode allocation (SCMA) in SS-EONs [17, 18]. (i) Routing is used to determine the appropriate route between a source-destination pair. (ii) SCMA is an approach, which is used to find the suitable spectrum slot in each mode of each core for the allocation of each request. In SS-EONs, the resource allocation is performed by satisfying the following constraints: spectrum contiguity, spectrum continuity, core continuity, mode continuity, inter-core crosstalk, and inter-mode crosstalk. As previously stated, the spectrum contiguity and continuity constraints remain unchanged. The same core on each hop and the same mode in each core must be utilized in the end-to-end path of a lightpath under the core continuity and mode continuity constraints, respectively [18]. During lightpath allocation in SS-EONs, inter-core and inter-mode crosstalks are considered a significant issue in multi-core multi-mode fibers (MCMMFs). Inter-core crosstalk happens due to evanescent waves when the same spectrum slots are used for lightpath establishment between neighboring cores [19]. Inter-mode crosstalk is produced when the same spectrum slots of neighboring modes are utilized for lightpath establishment in the same core [20].

2 Types of Optical Fibers for SS-EONs

The performance of SS-EONs depends on the position of the spatial channel in a given fiber structure. SS-EONs uses the channel in different ways, such as distinguish cores, multiplexed linearly polarized (LP) modes, and multiple cores in which each core supports a few multiplexed LP modes. In SS-EONs, there are various kinds of optical fibers used, which can be classified as follows [21–25]. (i) Single-mode fiber bundle (SMFB) — The most widely used form of optical fiber in SS-EONs is the single-mode fiber bundle, which consists of several traditional single-mode fibers, as shown in Fig. 3(a) [26]. SMFB has been a commercially accessible technology for many years, and it is used in current optical network technology. (ii) Multi-core fiber (MCF) — MCF consists of multiple cores embedded in a fiber cladding, each of them containing a single mode. MCF can be divided into two main layouts based on the arrangement of cores, uncoupled MCF and coupled MCF. For MCFs, there are various core arrangement structures have been developed as indicated in Fig. 3(b), (c), (d), (e), and (f) [27]. In MCFs, the cores can have even or uneven spacing and loose or tight packing. (iii) Multi mode fiber (MMF) (few mode fiber (FMF)) — MMF is the most well-known concept of a fiber for mode division multiplexing. In MMF, the transmission is performed using a higher number of transverse LP modes (as shown in Fig. 4), which propagates in a high index core. Since MMFs have been used a higher number of modes, there are some impairments and interference, such

as model dispersion and model interference produced in MMFs; in order to reduce these obstacles, FMFs are proposed. FMFs are utilized a fewer number of LP modes than MMF for propagating. FMFs are worked for long transmission distances compared to MMFs. (iv) Few-mode multi-core fiber (FM-MCF) — FM-MCF comprises multiple higher index cores in which each core contains numerous modes. FM-MCF combines the benefits of both FMF and MCF while avoiding their disadvantages. When FMF is combined with MCF, it enhances transport capacity; the authors in [28] showed that the integration of FM-MCF and Dense WDM is achieved a transmission capacity of 255 Tb/s. Recently, in [29], a fully integrated FM-MCF (with six modes and seven cores) amplifier has been realized in a cladding-pumped configuration. The key issue in MCFs is inter-core crosstalk and in FM-MCFs is inter-core and inter-mode crosstalks that can be suppressed by using the crosstalk-aware and crosstalk-avoided approach; we discuss these approaches in the subsequent sections.

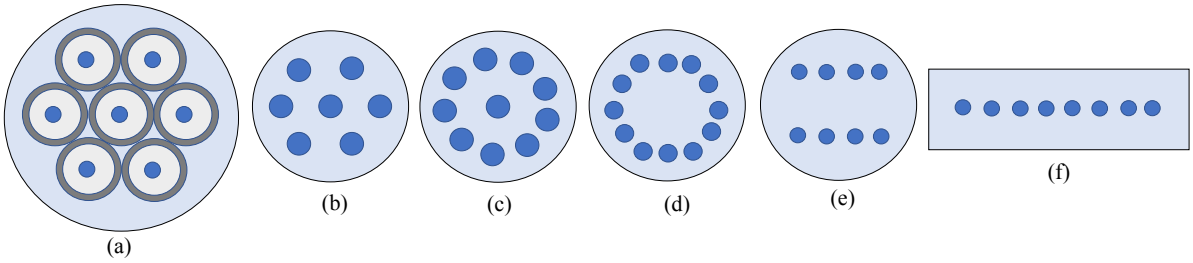


Figure 3: Different structure of fiber (a) single-mode fiber bundle, (b) hexagonal close packed MCF, (c) two pitch MCF, (d) one ring MCF, (e) linear array MCF, and (f) linear array rectangular shape MCF.

Figure 4: Different fundamental LP modes utilized in MMF and MCMMF (a) LP_{01}, (b) LP_{11}, and (c) LP_{02}.

3 Inter-core and Inter-mode Crosstalks in SS-EONs

In multi-mode multi-core fibers (MCMMFs), inter-core crosstalk and inter-mode crosstalk need to be suppressed, which reduces the resource utilization. To deal with the crosstalks issue, there are two main approaches, which are crosstalk-avoided and crosstalk-aware approaches. Crosstalk-avoided and crosstalk-aware approaches are the two major techniques that have been considered for RSCMA in SS-EONs [30–33]. In the former approach, identical spectrum slot must not be allocated in the adjacent modes and cores for different connections. Whereas, in the latter approach, the same spectrum slots in the adjacent modes and cores can be used if the generated crosstalk is below the predetermined crosstalk threshold limit. Although the crosstalk-aware approach provides better spectrum utilization than the crosstalk-avoided approach, it requires higher computational complexity. On the other hand, the crosstalk-avoided approach results in a more simplified network management policy with moderate spectrum utilization. Hence, the XT-avoided approach is favourable to deal with crosstalk in SS-EONs.

We model inter-core and inter-mode crosstalks as follows. The mean inter-core crosstalk among cores i and j is computed by using an analytical model that is based on the coupled-power theory [34,35]. The crosstalk ($XT_{i,j}$) in a multi-core fiber is calculated by:

$$XT_{i,j} = \frac{2\xi_{i,j}^2 R_b}{\Lambda \delta_{i,j}} L, \tag{1}$$

where $\xi_{i,j}$, R_b, Λ, $\delta_{i,j}$, and L denote the coupling coefficient between cores i and j, bending radius, mode propagation constant, core pitch between cores i and j, and length of fiber, respectively.
The expression of the coupling coefficient ($\xi_{i,j}$) in (1) is calculated by [35,36]:

$$\xi_{i,j} = \frac{\sqrt{\kappa}}{C_\mathrm{r}} \frac{\Upsilon^2}{\Delta^3 K_1^2(D)} \sqrt{\frac{\pi C_\mathrm{r}}{D\delta_{i,j}}} \exp\left(-\frac{\delta_{i,j}}{C_\mathrm{r}} D\right), \tag{2}$$

where $\Upsilon = C_\mathrm{r}\sqrt{(\eta^2 n_1^2 - \gamma^2)}$, $D = C_\mathrm{r}\sqrt{(\gamma^2 - \eta^2 n_0^2)}$, $\Delta = 2\pi C_\mathrm{r} n_1 \sqrt{(2|\kappa|)}/\lambda$, $\gamma = \eta \times n_\mathrm{eff}$, and $\eta = 2\pi/\lambda$; η and λ is a wave number and wavelength of light in vacuum, respectively. κ and C_r are the relative refractive index difference of core and cladding and core radius, respectively. n_0 and n_1 represent the refractive index of the cladding and core, respectively. $K_1(D)$ is the Bessel function of the second type with first order.

In order to calculate the crosstalk in an end-to-end path of a lightpath request, we consider the summation of crosstalk that is produced in each link for lightpath r [37, 38]. L_r represents the set of links that are considered for lightpath r. $XT(r,b) = \sum_{m \in L_r} XT(r,m,b) = \sum_{m \in L_r} K(r,m,b)h(m)L(m)$ calculate the crosstalk in an end-to-end path of lightpath r, where $XT(r,b)$, $XT(r,m,b)$, $K(r,m,b)$, $h(m)$, and $L(m)$ denote the average crosstalk for lightpath r on spectrum slot b, the mean crosstalk for lightpath r on slot b of link m, the number of assigned cores that are adjacent to the core captured by lightpath r on slot b of link m, the coefficient of power coupling for link m, and the length of link m, respectively. In [38], the crosstalk, $XT(r)$, for lightpath r is estimated by taking into account the most impacted spectrum slot in Ψ_r. Ψ_r is a set of spectrum slots utilized by lightpath r. $XT(r)$ is given by $XT(r) = \max_{b \in \Psi_r}\{XT(r,b)\}$.

In SS-EONs, when spectral and spatial resource allocation is performed, inter-core crosstalk and inter-mode crosstalk are generated. Inter-core crosstalk happens when at least two lightpaths are established in the same spectrum slots in neighboring cores in a fiber [19]. Similarly, when two or more lightpaths are established in the same spectrum slots in neighboring modes, inter-mode crosstalk is produced [20].

The authors in [13] presented a crosstalk-avoided model for both inter-core and inter-mode crosstalks simultaneously in SS-EONs; we adopt the same crosstalk-avoided model in the following. We consider a seven cores fiber in which each core has six modes. Figure 5(a) depicts the hexagonal structure with seven cores, which are based on [29]. The seven cores are arranged as follows: the central core represents core 7, and cores 1, 2, 3, 4, 5, and 6 are the outer cores of the fiber. Each core supports six spatial LP modes. Figure 5(b) shows an auxiliary graph that denotes these six LP modes, which are based on [39].

For avoiding inter-core and inter-mode crosstalks, there are two crosstalks constraints that must be followed. The inter-core crosstalk constraint ensures that adjacent cores do not use the same spectrum slots of the same mode for allocation. The inter-mode crosstalk constraint prevents adjacent modes of the same core from assigning the same spectrum slots. Figure 5(c) shows the spectrum condition considering MCMMFs with seven cores and six modes, and 26 lightpath requests (LR) are allocated with satisfying inter-core and inter-mode crosstalks simultaneously. In Fig. 5(c), if we consider slots 1, 2, 3, and 4 of mode 1 of core 1 for allocation of LR 1, we cannot use slots 1, 2, 3, and 4 of mode 1 of core 2, 6, and 7 for allocation of lightpaths due to inter-core crosstalk constraints. If we consider slots 1, 2, 3, and 4 of mode 1 of core 1 for allocation of LR 1, we cannot utilize slots 1, 2, 3, and 4 of mode 2 and 3 of core 1 for allocation of lightpaths due to inter-mode crosstalk constraints. Note that those slots used to avoid inter-core and inter-mode crosstalks and are not utilized for lightpath establishment are called crosstalks-avoided unutilized slots, as indicated in Fig. 5(c).

Since the impact of inter-core crosstalk of different slots between adjacent cores is considered negligible, different spectrum slots of the same mode can be utilized among adjacent cores. Since the impact of inter-mode crosstalk of different slots between adjacent modes is considered negligible, different spectrum slots of the same core can be utilized among adjacent modes. For example, in Fig. 5(c), if we use slots 1, 2, 3, and 4 of mode 1 of core 1, slots 5 and 6 of mode 1 are also used in adjacent cores 2 and 6 due to negligible effect of inter-core crosstalk. If we use slots 1, 2, 3, and 4 of mode 1 of core 1, slots 5 and 6 of adjacent mode 2 of core 1 are also used due to negligible effect of inter-mode crosstalk.

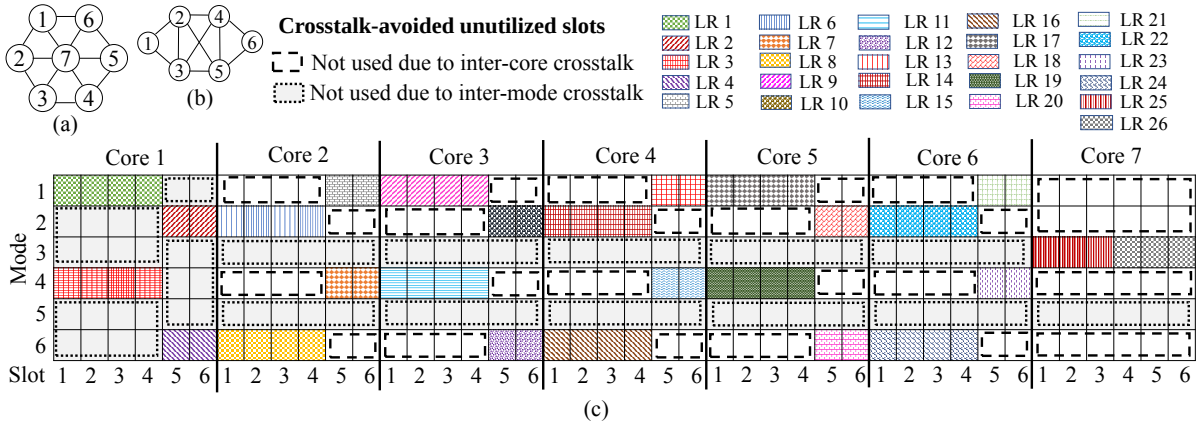


Figure 5: Demonstration for avoiding crosstalks: (a) structure of 7-core, (b) arrangement of 6-mode, and (c) spectrum condition.

4 Basic Concept of RSCMA

In SS-EONs, spectral and spatial resource allocation and network design problems can be categorized into two types static and dynamic [26]. In a static scenario, the traffic matrix for all lightpath requests is previously known, and allocation is performed simultaneously in the network. Routing, spectrum, core, and mode are determined for lightpath allocation in an offline manner. In the dynamic scenario, it is considered that the lightpaths are unknown in advance, but they are allocated in real-time and torn down as per requirement. Routing, spectrum, core, and mode are selected dynamically for lightpath allocation based on the current state of the network.

In order to manage routing problems in SS-EONs, we can consider similar approaches that are applied in EONs. Routing approaches can be mainly divided into two types: on-time (on-demand) and pre-computational approaches. The on-time approach is preferable for efficient resource utilization, but it is not scalable. The pre-computation approach can be subdivided as follows: single-route, multiple-route, anycast, and multicast. Single-route and multiple-route approaches are

used for general purposes; for example, minimum hop routing and shortest path routing are used for single-routes, and K-shortest path routing is used for multiple-routes. Anycast and multicast approaches are considered for specific applications, including content-oriented networks and cloud computing environments.

In SS-EONs, we consider different spectrum, core, and mode allocation policies such as core-mode-spectrum first fit (CMS-FF) and CMS random fit (CMS-RF). The work in [13, 18, 40] presented the following spectrum, core, and mode allocation policies: (i) core-mode-spectrum first fit (CMS-FF), (ii) core-spectrum-mode first fit (CSM-FF), (iii) spectrum-core-mode first fit (SCM-FF), (iv) spectrum-mode-core first fit (SMC-FF), (v) mode-core-spectrum first fit (MCS-FF), and (vi) mode-spectrum-core first fit (MSC-FF). The working of the CMS-FF policy is explained the following [13]. The CMS-FF policy performs the establishment of lightpath requests by considering the first fit searching order of core index (id), mode id, and spectrum slot id. In the CMS-FF policy, the highest priority is given to the core, and the lowest priority is given to the spectrum. This implies that resource allocation is performed from core id =1, mode id = 1, and slot id =1. All slots are explored while keeping the core id and mode id unchanged. When all slots have been explored, the mode id is incremented, and all spectrum slots are explored. In this manner, when all modes are explored, the core id is incremented, and all modes are explored. These processes are repeated until all cores have been explored in order to meet the bandwidth requirement of a lightpath. When the lightpath request fails to determine the required slots, it is assumed to be a blocked request.

In the CSM-FF, SCM-FF, SMC-FF, MCS-FF, and MSC-FF policies, the procedure of searching resources is performed in the same manner as described above, which is based on the priority of core, mode, and spectrum. In the CSM-FF policy, the highest priority is given to core, and the lowest priority is given to mode. In the SCM-FF policy, the highest priority is given to spectrum, and the lowest priority is given to mode. In the SMC-FF policy, the highest priority is given to spectrum, and the lowest priority is given to core. In the MCS-FF policy, the highest priority is given to mode, and the lowest priority is given to spectrum. In the MSC-FF policy, the highest priority is given to mode, and the lowest priority is given to core.

The flowchart of the RSCMA problem is shown in Fig. 6. First, the route is determined by considering any routing policy, such as shortest path routing. The number of required slots is calculated to satisfy the requested bandwidth demand considering distance-adaptive modulation [41]. Thereafter, the available spectrum slots, core, and mode are identified by considering inter-core and inter-mode crosstalks and spectrum contiguity constraints. When the spectrum slots, core, and mode are determined, the lightpath request is allocated by satisfying spectrum, core, and mode continuity constraints. If all the SS-EONs constraints are satisfied, the request is allocated. Otherwise, it is rejected.

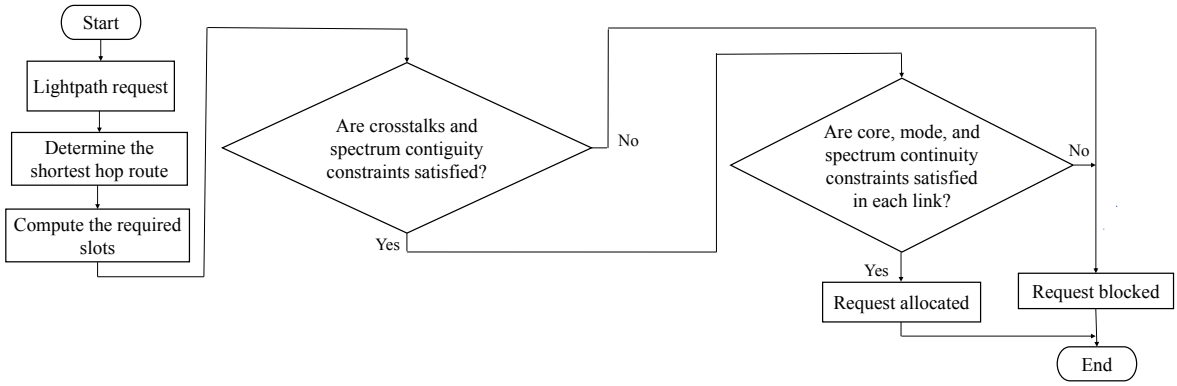


Figure 6: Routing, spectrum, core, and mode allocation considering inter-core and inter-mode crosstalks in SS-EONs.

5 Challenges of SS-EONs and Future Directions of Research

In SS-EONs, inter-core and inter-mode crosstalks are challenging issues. In order to suppress the crosstalks, several approaches [13, 42–45] have been addressed in the literature. Due to the crosstalks issue, spectral and spatial resource utilization is reduced in SS-EONs. It is required to investigate and introduce the different resource allocation policies to suppress crosstalk effects. In addition, different types of fiber need to be developed to diminish the crosstalk effects.

The fragmentation problem in SS-EONs is more complex compared to EONs [30]. The fragmentation problem reduces spectral and spatial resource utilization and increases the blocking in SS-EONs. Therefore, it is necessary to diminish spectrum fragmentation by introducing efficient algorithms.

There are several SDM technologies to enhance transport capacity. There are some discussions on pros and cons of a bundle of single-mode fibers (SMFs) transmission versus multi-core multi-mode fiber transmission [12, 26, 46]. A bundle of SMFs, named fiber-bundles, is considered an alternate option for early SDM implementation. Multi-core fibers are potential candidates for SDM technology, and significant research efforts have shown impressive results. However, the major practical limitation for multi-core fibers is the additional cost of replacing the existing fiber infrastructure with the new SDM fiber types. Therefore, It is essential to study further which technology provides an effective solution.

6 Conclusion

In this work, we provided an overview related to spectrally-spatially elastic optical networks. We presented a crosstalk model considering both inter-core and inter-mode crosstalks simultaneously. We described a crosstalk-avoided model for avoiding crosstalks in MCMMFs fibers for SS-EONs. Furthermore, we discussed routing, spectrum, core, and mode allocation problems and challenging issues in SS-EONs.

Acknowledgements

This work is supported in part by the Core Research grant (Grant Number: CRG/2020/002663), Govt. of India, Inspire Faculty Scheme (Grant Number: DST/INSPIRE/04/2016/001316), Govt. of India, and Japan-India Science Cooperative Program between JSPS and DST (Grant Number: JPJSBP120207712 and DST/INT/JSPS/P-320/2020).

References

[1] B. C. Chatterjee, N. Sarma, and E. Oki, "Routing and spectrum allocation in elastic optical networks: A tutorial," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 3, pp. 1776–1800, 2015.

[2] B. C. Chatterjee and E. Oki, *Elastic Optical Networks: Fundamentals, Design, Control, and Management: Fundamentals, Design, Control, and Management*. CRC Press, 2020.

[3] G. Keiser, *Optical fiber communications*. McGraw-Hill Science, Engineering & Mathematics, 1983.

[4] R. M. C. Siva and G. Mohan, *WDM Optical Networks: Concepts, Design and Algorithms*, PHI, 2003.

[5] B. Mukherjee, *Optical WDM Networks (Optical Networks)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[6] M. Jinno, H. Takara, B. Kozicki, Y. Tsukishima, Y. Sone, and S. Matsuoka, "Spectrum-efficient and scalable elastic optical path network: architecture, benefits, and enabling technologies," *IEEE communications magazine*, vol. 47, no. 11, pp. 66–73, 2009.

[7] G. Zhang, M. De Leenheer, A. Morea, and B. Mukherjee, "A survey on OFDM-based elastic core optical networking," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 65–87, 2012.

[8] G. Keiser, *Optical fiber communications*. McGraw-Hill New York, 2000, vol. 2.

[9] P. J. Winzer, "Spatial multiplexing in fiber optics: The 10x scaling of metro/core capacities," *Bell Labs Technical Journal*, vol. 19, pp. 22–30, 2014.

[10] G. Li, N. Bai, N. Zhao, and C. Xia, "Space-division multiplexing: the next frontier in optical communication," *Advances in Optics and Photonics*, vol. 6, no. 4, pp. 413–487, 2014.

[11] C. V. Saradhi and S. Subramaniam, "Physical layer impairment aware routing (PLIAR) in WDM optical networks: Issues and challenges," *IEEE Communications Surveys & Tutorials*, vol. 11, no. 4, pp. 109–130, 2009.

[12] D. J. Richardson, J. M. Fini, and L. E. Nelson, "Space-division multiplexing in optical fibres," *Nature photonics*, vol. 7, no. 5, pp. 354–362, 2013.

[13] B. C. Chatterjee, A. Wadud, I. Ahmed, and E. Oki, "Priority-based inter-core and inter-mode crosstalk-avoided resource allocation for spectrally-spatially elastic optical networks," *IEEE/ACM Transactions on Networking*, vol. 29, no. 4, pp. 1634–1647, 2021.

[14] B. C. Chatterjee, I. Ahmed, A. Wadud, M. Maity, and E. Oki, "Bpria: Crosstalk-avoided bi-partitioning-based counter-propagating resource identification and allocation for spectrally-spatially elastic optical networks," *IEEE Transactions on Network and Service Management*, pp. 1-15, 2022 [to appear].

[15] Z. Zhu, X. Liu, and M.-S. Alouini, "Advances in optical communications and network technologies," *IEEE Communications Magazine*, vol. 57, no. 10, pp. 12–12, 2019.

[16] Y. Wang, X. Cao, and Y. Pan, "A study of the routing and spectrum allocation in spectrum-sliced elastic optical path networks," in *2011 Proceedings IEEE Infocom*. IEEE, 2011, pp. 1503–1511.

[17] H. Tode and Y. Hirota, "Routing, spectrum and core assignment on SDM optical networks," in *Optical Fiber Communication Conference*. Optical Society of America, 2016, pp. Tu2H–1.

[18] ——, "Routing, spectrum, and core and/or mode assignment on space-division multiplexing optical networks," *Journal of Optical Communications and Networking*, vol. 9, no. 1, pp. A99–A113, 2017.

[19] K. Morita and K. Hirata, "Dynamic spectrum allocation method for reducing crosstalk in multi-core fiber networks," in *2017 International Conference on Information Networking (ICOIN)*. IEEE, 2017, pp. 686–688.

[20] S. Guo, W. Ju, S. Yin, B. Wang, Y. Yuan, and S. Huang, "Crosstalk-aware routing, spectrum and mode assignment in few mode fiber with MIMO equalization," in *Asia Communications and Photonics Conference*. Optical Society of America, 2018, pp. M2E–2.

[21] H. Takara, A. Sano, T. Kobayashi, H. Kubota, H. Kawakami, A. Matsuura, Y. Miyamoto, Y. Abe, H. Ono, K. Shikama *et al.*, "1.01-pb/s (12 SDM/222 WDM/456 gb/s) crosstalk-managed transmission with 91.4-b/s/hz aggregate spectral efficiency," in *European Conference and Exhibition on Optical Communication*. Optical Society of America, 2012, pp. Th–3.

[22] M.-J. Li, B. Hoover, V. N. Nazarov, and D. L. Butler, "Multicore fiber for optical interconnect applications," in *2012 17th Opto-Electronics and Communications Conference*. IEEE, 2012, pp. 564–565.

[23] O. Egorova, S. Semjonov, A. Senatorov, M. Salganskii, A. Koklyushkin, V. Nazarov, A. Korolev, D. Kuksenkov, M.-J. Li, and E. Dianov, "Multicore fiber with rectangular cross-section," *Optics letters*, vol. 39, no. 7, pp. 2168–2170, 2014.

[24] S. Matsuo, K. Takenaga, Y. Arakawa, Y. Sasaki, S. Taniagwa, K. Saitoh, and M. Koshiba, "Large-effective-area ten-core fiber with cladding diameter of about 200 μm," *Optics letters*, vol. 36, no. 23, pp. 4626–4628, 2011.

[25] K. Takenaga, Y. Arakawa, S. Tanigawa, N. Guan, S. Matsuo, K. Saitoh, and M. Koshiba, "Reduction of crosstalk by trench-assisted multi-core fiber," in *Optical Fiber Communication Conference*. Optical Society of America, 2011, p. OWJ4.

[26] M. Klinkowski, P. Lechowicz, and K. Walkowiak, "Survey of resource allocation schemes and algorithms in spectrally-spatially flexible optical networking," *Optical Switching and Networking*, vol. 27, pp. 58–78, 2018.

[27] G. M. Saridis, D. Alexandropoulos, G. Zervas, and D. Simeonidou, "Survey and evaluation of space division multiplexing: From technologies to optical networks," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2136–2156, 2015.

[28] R. G. Van Uden, R. A. Correa, E. A. Lopez, F. Huijskens, C. Xia, G. Li, A. Schülzgen, H. De Waardt, A. Koonen, and C. M. Okonkwo, "Ultra-high-density spatial division multiplexing with a few-mode multicore fibre," *Nature Photonics*, vol. 8, no. 11, pp. 865–870, 2014.

[29] Y. Jung, M. Wada, K. Shibahara, S. Jain, I. A. Davidson, P. Barua, J. R. Hayes, T. Sakamoto, T. Mizuno, Y. Miyamoto *et al.*, "High spatial density 6-mode 7-core fiber amplifier for l-band operation," *Journal of Lightwave Technology*, vol. 38, no. 11, pp. 2938–2943, 2020.

[30] B. C. Chatterjee, A. Wadud, and E. Oki, "Proactive fragmentation management scheme based on crosstalk-avoided batch processing for spectrally-spatially elastic optical networks," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 9, pp. 2719–2733, 2021.

[31] J. Halder, M. Maity, E. Oki, and B. C. Chatterjee, "Shared backup path protection-based resource allocation considering inter-core and inter-mode crosstalk for spectrally-spatially elastic optical networks," *IEEE Communications Letters*, vol. 26, no. 3, pp. 637–641, 2022.

[32] K. Takeda, T. Sato, B. C. Chatterjee, and E. Oki, "Joint inter-core crosstalk- and intra-core impairment-aware lightpath provisioning model in space-division multiplexing elastic optical networks," *IEEE Transactions on Network and Service Management*, 2022 [to appear].

[33] ——, "Jointly inter-core xt and impairment aware lightpath provisioning in elastic optical networks," in *ICC 2021-IEEE International Conference on Communications*. IEEE, 2021, pp. 1–6.

[34] T. Hayashi, T. Taru, O. Shimakawa, T. Sasaki, and E. Sasaoka, "Design and fabrication of ultra-low crosstalk and low-loss multi-core fiber," *Optics express*, vol. 19, no. 17, pp. 16 576–16 592, 2011.

[35] D. Kumar and R. Ranjan, "Optimal design for crosstalk analysis in 12-core 5-LP mode homogeneous multicore fiber for different lattice structure," *Optical Fiber Technology*, vol. 41, pp. 95–103, 2018.

[36] K. Okamoto, *Fundamentals of optical waveguides*. Academic press, 2006.

[37] J. Strand and A. Chiu, "Impairments and other constraints on optical layer routing", rfc 4054," 2005.

[38] M. Klinkowski and K. Walkowiak, "Impact of crosstalk estimation methods on the performance of spectrally and spatially flexible optical networks," in *2018 20th International Conference on Transparent Optical Networks (ICTON)*. IEEE, 2018, pp. 1–4.

[39] D. Gloge, "Weakly guiding fibers," *Applied optics*, vol. 10, no. 10, pp. 2252–2258, 1971.

[40] F. Arpanaei, N. Ardalani, H. Beyranvand, and S. A. Alavian, "Three-dimensional resource allocation in space division multiplexing elastic optical networks," *Journal of Optical Communications and Networking*, vol. 10, no. 12, pp. 959–974, 2018.

[41] G. Bosco, V. Curri, A. Carena, P. Poggiolini, and F. Forghieri, "On the performance of nyquist-wdm terabit superchannels based on pm-bpsk, pm-qpsk, pm-8qam or pm-16qam subcarriers," *Journal of Lightwave Technology*, vol. 29, no. 1, pp. 53–61, 2011.

[42] A. Muhammad, G. Zervas, D. Simeonidou, and R. Forchheimer, "Routing, spectrum and core allocation in flexgrid sdm networks with multi-core fibers," in *2014 International Conference on Optical Network Design and Modeling*. IEEE, 2014, pp. 192–197.

[43] M. Yang, Y. Zhang, and Q. Wu, "Routing, spectrum, and core assignment in SDM-EONs with MCF: node-arc ILP/MILP methods and an efficient XT-aware heuristic algorithm," *Journal of Optical Communications and Networking*, vol. 10, no. 3, pp. 195–208, 2018.

[44] M. Klinkowski and G. Zalewski, "Dynamic crosstalk-aware lightpath provisioning in spectrally–spatially flexible optical networks," *Journal of Optical Communications and Networking*, vol. 11, no. 5, pp. 213–225, 2019.

[45] C. Rottondi, P. Martelli, P. Boffi, L. Barletta, and M. Tornatore, "Crosstalk-aware core and spectrum assignment in a multicore optical link with flexible grid," *IEEE Transactions on Communications*, vol. 67, no. 3, pp. 2144–2156, 2018.

[46] B. C. Chatterjee, F. He, E. Oki, A. Fumagalli, and N. Yamanaka, "A span power management scheme for rapid lightpath provisioning and releasing in multi-core fiber networks," *IEEE/ACM Transactions on Networking*, vol. 27, no. 2, pp. 734–747, 2019.

Performance of Non-Defragmentation and Batch Processing Based Proactive Fragmentation Management Scheme in Elastic Optical Networks

Abdul Wadud[1,2], Eiji Oki[3], and Bijoy Chand Chatterjee[1*]

[1]Department of Computer Science, South Asian University, New Delhi, India
[2]Bangladesh Institute of Governance and Management, Bangladesh
[3]Graduate School of Informatics, Kyoto University, Kyoto, Japan

[*]Corresponding author: bijoycc@ieee.org

Abstract

Fragmentation occurs in elastic optical networks (EONs) due to dynamic lightpath assignment and tearing down in the network, which is considered as one of the major challenging issues for resource allocation. It is difficult to utilize fragmented slots in the network during the allocation process. Therefore, it is necessary to manage and suppress fragmented slots in the network. This work presents a non-defragmentation and batch processing-based proactive fragmentation management routing and spectrum allocation (BPBP-RSA) scheme in EONs. The presented scheme creates batches of incoming lightpath requests before establishing them to the network and allocates requests belonging to each batch sequentially. Since we do not take any action or rearrange any slot after request allocation and complete all the processes before allocation, we consider the presented scheme as non-defragmentation and proactive. We consider a traditional first-fit routing and spectrum allocation (FF-RSA) as a benchmark and compare BPBP-RSA with FF-RSA. Numerical results ensure that the presented BPBP-RSA scheme outperforms the traditional FF-RSA scheme in terms of blocking probability and traffic admissibility. The presented BPBP-RSA scheme allocates 28.5% more traffic than that of the traditional FF-RSA scheme when the blocking probability is considered 1% in the network.

Keywords: *Fragmentation, Routing and Spectrum allocation, Elastic optical networks.*

1 Introduction

In the era of fifth-generation (5G) and sixth-generation (6G) networks, bandwidth-hungry applications, such as the Internet of things (IoT), cloud computing, and live streaming, require high-speed networks [1]. A highly compatible optical backbone network is required to address the requirements of 5G and 6G technologies [2]. The elastic optical network (EON) is a prospective candidate to facilitate high-speed bandwidth support to 5G and 6G networks. EONs adopt flexible resource allocation to establish lightpath requests by using the orthogonal frequency division multiplexing (OFDM) technology [3]. When a lightpath request arrives at the network, typically, the route of the request is determined, and thereafter the request is allocated to spectrum slots of the specific links in its routes by satisfying the spectrum contiguity and continuity constraints; this process is known as routing and spectrum allocation (RSA) [4–6].

There are two distinct types of resource allocations in EONs, i.e., static and dynamic allocation. Fragmentation arises inevitably in dynamic resource allocation in EONs due to dynamic assignment and tearing down of lightpath requests, which is a challenging issue that needs to be addressed [6,7]. Different approaches have been presented to handle the fragmentation problem in EONs [7–11]. A taxonomy of fragmentation management is presented in [7, 12] that shows the fragmentation can be handled using the non-defragmentation and defragmentation approaches. In the non-defragmentation approaches [13–16], necessary precautions are taken to avoid fragmentation before the establishment of a lightpath. However, no action is taken for in-service lightpaths. Whereas in the case of defragmentation approaches [17–20], a necessary action is taken for in-service lightpaths in order to suppress the fragmentation effect. The authors in [7, 12] further stated that the non-defragmentation and defragmentation-based approaches are not mutually exclusive. Network operators generally favor non-defragmentation approaches due to lower capital expenditures (CAPEX) and operational expenditures (OPEX).

This work presents a non-defragmentation and batch processing-based proactive fragmentation management routing and spectrum allocation (BPBP-RSA) scheme in EONs. In the presented scheme, a static batch processing for the lightpath requests method is discussed, which classifies lightpath requests into batches. The batch with the highest priority is allocated first, while the last batch possesses the least priority. If any batch request is blocked, a fairness policy is triggered to provide equal opportunities for all requests to be allocated. The presented scheme creates the set of batches, finds routes of each request, allocates lightpath requests by maintaining resource allocation constraints using first-fit, and triggers fairness policy if any request gets blocked.

The rest of the paper is organized as follows. The presented BPBP-RSA scheme is discussed in section 2. The performance of BPBP-RSA is evaluated in section 3, and we conclude the paper in section 4.

Note that the key contribution of this work is to investigate the impact of batch processing in the fragmentation management approach in the context of EONs. We adopt batch processing for fragmentation management in EONs from the study done by us in our previous work [21].

2 Presented Non-defragmentation and Batch Processing Based Proactive Fragmentation Management Scheme

This section presents the non-defragmentation and batch processing based proactive fragmentation management scheme. The presented scheme has two steps, which are batch processing and resource allocation.

2.1 Model and Assumptions

We model the optical network as a directed graph $G(V, E)$, where V is the set of nodes and E is the set of edges. We assume that each link $e \in E$ has $|S|$ number of spectrum slots, where S is the set of spectrum slices in a link. R represents the set of all requests, and a request is denoted by $r \in R$. We assume that the capacity of each lightpath request $r \in R$ is given. We consider the shortest path routing to estimate the end-to-end routing of each lightpath request [22]. Furthermore, we consider polarized-multiplexed optical signals for transmission. Modulation formats, which are polarized multiplexed-binary phase shift keying (PM-BPSK), polarized multiplexed-quadrature phase shift keying (PM-QPSK), polarized multiplexed-8quadrature amplitude modulation (PM-8QAM), polarized multiplexed-16quadrature amplitude modulation (PM-16QAM), polarized multiplexed-32quadrature amplitude modulation (PM-32QAM), and polarized multiplexed-64quadrature amplitude modulation (PM-64QAM), are used for lightpath allocation based on requested capacity and transmission distance of each lightpath request according to Table 1 [23, 24]. We assume that the optical bandwidth of each elastic transceiver is 37.5GHz, and the size of each spectrum slice is 12.5 GHz. The number of required slots δ_r for a lightpath requests $r \in R$ is estimated according to the transmission reach model presented in [21]. The number of links in the end-to-end route of a lightpath request is denoted by l_r. B denotes the set of generated batches, and an element of set B is also a set, and it is denoted by $\beta_k \in B$, where $k \in [1, |B|]$.

Table 1: Transmission reach (KM) with respect to different traffic volume and modulation format

Traffic volume (Gbps)	PM-BPSK	PM-QPSK	PM-8QAM	PM-16QAM	PM-32QAM	PM-64QAM
50	3400	3300	1300	1000	500	300
100	1700	1700	700	500	200	100
150	1200	1100	400	300	100	100
200	900	900	300	200	100	100
250	700	700	300	200	100	0
300	600	600	200	200	100	0
350	500	500	200	100	0	0
400	400	400	100	100	0	0

The lightpath requests in R are allocated by satisfying spectrum contiguity and continuity constraints. The contiguity constraint ensures contiguous slot allocation of a lightpath request. The continuity constraint ensures using the same slots over other links if more than one link exists in the route of a request.

2.2 Batch Processing

The batch processing method has two steps, namely sequence generation and batch generation. The batch processing method is explained in the following.

Step 1: Estimate number of links l_r in the end-to-end routes of all $r \in R$ and classify requests into different classes based on their l_r. For instance- requests having 1 link in their end-to-end route belongs to class 1, requests having 2 link in their end-to-end route belongs to class 2, and so on. We have eight requests $r_1,\ r_2,\ r_3,\ r_4,\ r_5,\ r_6,\ r_7,$ and r_8 having 1 and 2 hops in the end-to-end route in Fig. 1(b), where $c_1 = \{r_1, r_3, r_5, r_7\}$ and $c_2 = \{r_2, r_4, r_6, r_8\}$.

Step 2: Assign weights w to each class based on its index. For example- c_1 has the weight $w = 1$, and c_2 has the weight $w = 2$.

Step 3: Sort classes in descending order based on assigned weights. For example- we obtain $c_2 = \{r_2, r_4, r_6, r_8\}$ and $c_1 = \{r_1, r_3, r_5, r_7\}$ after sorting by class in Fig. 1(b).

Step 4: Form a super-class combining all the classes maintaining the sorted class order. For example- super class U for Fig. 1(b) is $\{r_2, r_4, r_6, r_8, r_1, r_3, r_5, r_7\}$.

Step 5: Generate batches of requests $\beta \in B$ maintaining the batch capacity constraint and minimizing the cost factor. The batch capacity constraint limits the size of a batch, where the total number of required slots $\sum \delta_r$ for all requests r in a batch $\beta \in B$ must not exceed $|S|$, i.e.- $\sum_{r \in R} \delta_r x_r^n \leq |S|, \forall n \in [1, |B|]$. The cost-factor $\mathcal{C}_f$ is the total cost for generating all batches, and it is defined by- $\mathcal{C}_f = \sum_{n \in [1,|B|]} \left(n \sum_{r \in R} l_r x_r^n\right)$, where n is the index of batch, l_r is the number of links in the end-to-end route of r, and x_r^n is the binary variable representing attachment of request r to nth batch. If request r belongs to nth batch, $x_r^n = 1$, otherwise $x_r^n = 0$. A request $r \in R$ can be a member of only one batch, i.e.- $\sum_{n \in [1,|B|]} x_r^n = 1, \forall r \in R$. Therefore, we obtain batches $\beta_1 = \{r_2, r_4, r_8\}$, $\beta_2 = \{r_6, r_1\}$, $\beta_3 = \{r_3, r_5\}$, and $\beta_4 = \{r_7\}$ by generating batches satisfying all batch generation constrains in Fig. 1.

Steps 1 to 4 work for generating sequence, and step 5 works for batch processing. After generating batches, the BPBP-RSA scheme allocates each batch using the first-fit resource allocation policy.

2.3 Resource Allocation

A lightpath request $r \in R$ is allocated using spectrum slices in EONs. The presented BPBP-RSA scheme processes batches of lightpath requests using a batch processing method mentioned in Section 2.2 after estimating the end-to-end route and requested slots. Afterward, requests are allocated by satisfying spectrum continuity and contiguity constraints. Suppose any lightpath request of a batch is rejected during resource allocation. In that case, the presented scheme triggers the fairness policy that gives an equal opportunity to all requests, irrespective of their number of hops, requested capacities, and arriving sequences.

In the fairness policy, if any request $r \in \beta_k$ considering all batches fails to satisfy the resource allocation constraints, the presented scheme discards one request from R randomly and updates R. The updated set of requests R is now reconsidered for batch generation. The same procedure is iterated until R becomes empty. In each iteration, exactly one request is rejected. Therefore, the number of iterations equals the number of blocked requests. During this resource allocation process, expired requests are automatically torn down from the network.

2.4 Benefit of BPBP-RSA over FF-RSA

This subsection presents the benefit of the presented BPBP-RSA scheme over the traditional FF-RSA scheme. The presented scheme follows the non-defragmentation process by partitioning the incoming lightpath requests into batches before establishment and taking no measures after allocation. Figure 1 demonstrates the resource allocation in EON using BPBP-RSA. A sample 4-node network and a set of requests $R = \{r_1, r_2, r_3, r_4, r_5, r_6, r_7, r_8\}$ with the same arrival time, predefined routes and capacity, and estimated required slots δ_r for each requests $r \in R$ are considered for demonstration purpose [see Fig. 1(a) and (b)]. The number of required slots δ_r for each request is calculated based on the model described in [24]. Modulation format used for request r_1, r_2, r_3, r_4, r_5, r_6, r_7, and r_8 are PM-8QAM, PM-QPSK, PM-BPSK, PM-QPSK, PM-QPSK, PM-BPSK, PM-QPSK, and PM-32QAM, respectively, using Tab. 1. In the sample network, $V = \{A, B, C, D\}$ and $E = \{AB, BC, BD, DC, AD\}$.

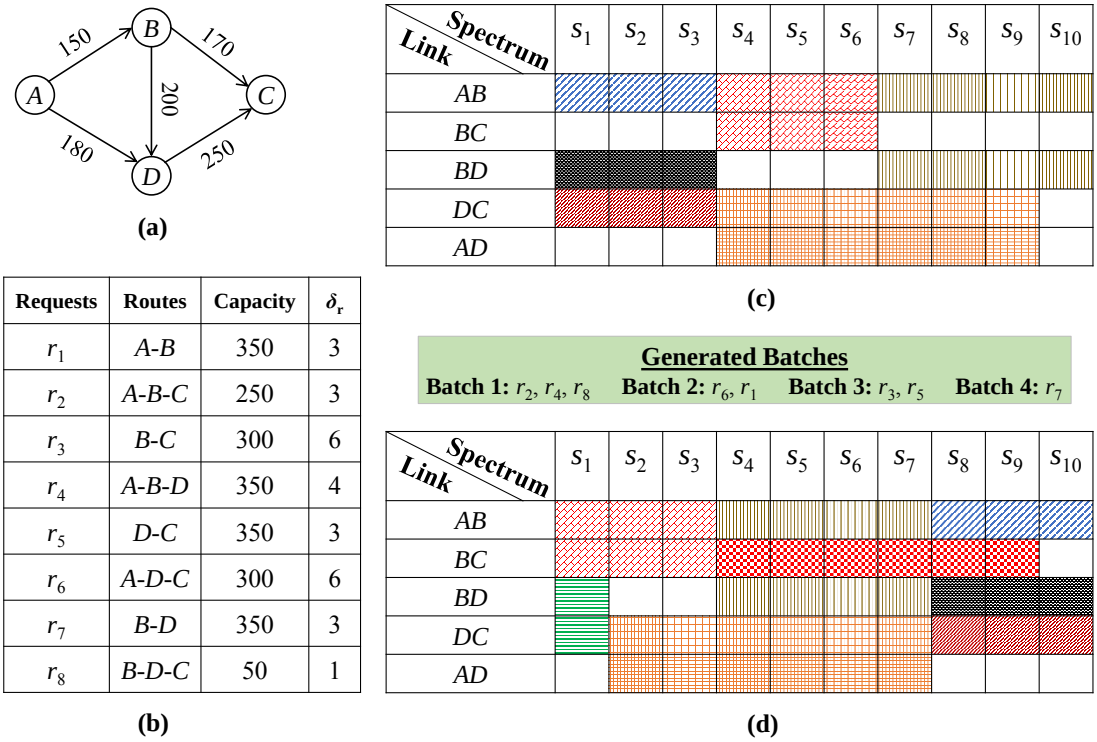
Figure 1(c) depicts the network status after allocation the requested lightpaths $r \in R$ using FF-RSA. Two requests r_3 and r_8 are rejected in Fig. 1(c) and hence the blocking ratio is $\frac{2}{8} = 0.25$. For allocating the same set of requests using BPBP-RSA, we need to generate batches of lightpath requests using the batch processing method. After applying batch processing method, we obtain batches $\beta_1 = \{r_2, r_4, r_8\}$, $\beta_2 = \{r_6, r_1\}$, $\beta_3 = \{r_3, r_5\}$, and $\beta_4 = \{r_7\}$ by satisfying all batch generation constrains (see in 2.2). Each batch is sequentially allocated in the network using the first-fit allocation process.

Figure 1(d) shows the network status after allocating all eight requests in the sample network. No request is blocked using the BPBP-RSA approach. Hence we claim that the BPBP-RSA approach works better than traditional FF-RSA approach by reducing blocking probability and enhancing resource utilization. The external fragmentation metric is $\Gamma_l = 1 - \frac{P}{Q}$, where Γ_l, P, and Q represents fragmentation in link l, maximum number of contiguous available slots, and total number of available slots, respectively [25]. Link fragmentations for all the links in Fig. 1(c) using traditional first-fit RSA scheme are $\Gamma_{AB} = 0$, $\Gamma_{BC} = 0.43$, $\Gamma_{BD} = 0$, $\Gamma_{DC} = 0$, and $\Gamma_{AD} = 0.25$. Similarly, link fragmentations using BPBP-RSA in Fig. 1(c) are $\Gamma_{AB} = 0$, $\Gamma_{BC} = 0$, $\Gamma_{BD} = 0$, $\Gamma_{DC} = 0$, and $\Gamma_{AD} = 0.25$. Hence, we can conclude that link fragmentation after allocating all eight requests are using BPBP-RSA is less than link fragmentation using traditional first-fit after allocating only six lightpath requests. Note that, as no request is rejected using BPBP-RSA, the fairness policy is not considered in this example.

3 Simulation and Results

We consider the Indian network [26] for analyzing the performance of the presented scheme, where the average node degree is 3.14. We consider Erlang traffic load for generating the lightpath requests. The load is defined by $\rho = \lambda \times H$, where λ and H represent inter-arrival rate and average holding time, respectively. The exponential distribution is maintained for generating both inter-arrival rate and holding times of lightpath requests. All the simulations are performed on a Linux-based HPE ProLiant ML350 server with an Intel Xeon-Bronze 3106 processor (1.7GHz) and 64GB of memory.

We use blocking probability and traffic admissibility as performance metrics. The blocking probability is defined as a ratio of the number of rejected requests to the number of requests served in the network. The obtained results for the



Requests	Routes	Capacity	δ_r
r_1	A-B	350	3
r_2	A-B-C	250	3
r_3	B-C	300	6
r_4	A-B-D	350	4
r_5	D-C	350	3
r_6	A-D-C	300	6
r_7	B-D	350	3
r_8	B-D-C	50	1

(b)

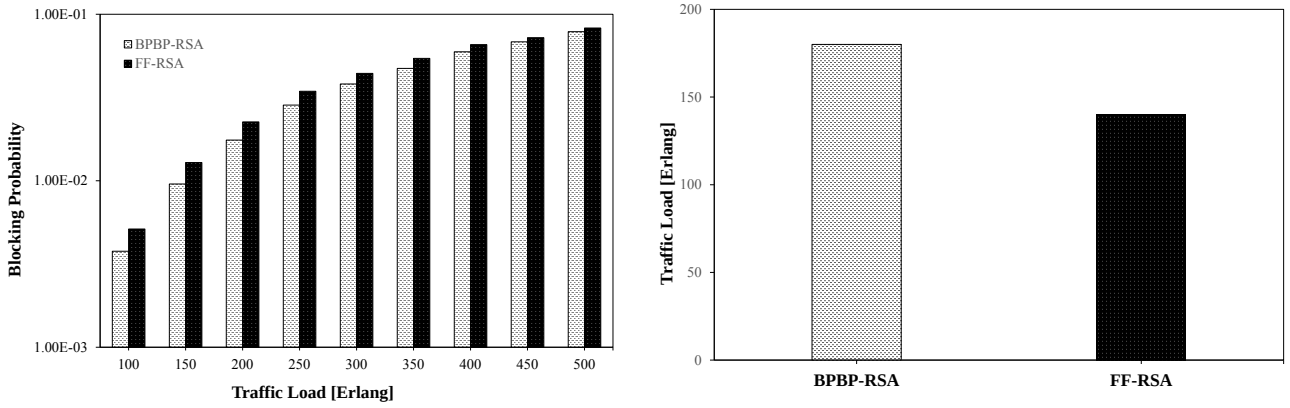
Figure 1: Resource allocation in EONs: (a) sample network, (b) lightpath requests with predefined routes and their capacity, (c) resource allocation using FF-RSA, and (d) resource allocation using BPBP-RSA.

presented BPBP-RSA scheme and the traditional FF-RSA scheme for traffic admissibility are considered under the 1% blocking of requests.

Figure 2a depicts the blocking probability comparison between the presented BPBP-RSA scheme and the traditional first-fit RSA scheme; the comparison shows that the presented scheme suppresses the blocking probability compared to the traditional FF-RSA scheme. Figure 2b plots the traffic admissibility comparison between the two schemes. The performance comparisons in Fig. 2 illustrate that the presented scheme works better for resource allocation in EONs. It suppresses fragmentation and reduces the wastage of the existing spectral resources. As a result, the resource utilization in the network is enhanced, and blocking probability is reduced. Finally, the traffic admissibility ensures that the presented BPBP-RSA approach is capable of assigning more lightpath requests in the network compared to the traditional FF-RSA scheme.

4 Conclusion

Fragmentation is a severe problem for resource allocation in elastic optical networks (EONs). It suppresses resource utilization and traffic admissibility in elastic optical networks. This work introduced a distance-adaptive resource allocation scheme based on batch processing in EONs, which deals with the fragmentation problem during resource allocation and



(a) Blocking probability comparison.

(b) Traffic admissibility comparison.

Figure 2: Performance comparison between traditional RSA using first-fit and presented BPBP-RSA scheme.

enhances resource utilization. The performance evaluation section ensured that the presented BPBP-RSA scheme outperformed the traditional first-fit resource allocation scheme in terms of blocking probability and traffic admissibility.

Acknowledgements

This work is supported in part by the Core Research grant (Grant Number: CRG/2020/002663), Govt. of India, Inspire Faculty Scheme (Grant Number: DST/INSPIRE/04/2016/001316), Govt. of India, and Japan-India Science Cooperative Program between JSPS and DST (Grant Number: JPJSBP120207712 and DST/INT/JSPS/P-320/2020)

References

[1] P. J. Winzer, "Spatial multiplexing: The next frontier in network capacity scaling," in *IET Conference Proceedings*. The Institution of Engineering & Technology, 2013.

[2] X. Liu, N. Deng, M. Zhou, Y. Wang, M. Tao, L. Zhou, S. Li, H. Zeng, S. Megeed, A. Shen *et al.*, "Enabling technologies for 5g-oriented optical networks," in *Optical Fiber Communication Conference*. Optical Society of America, 2019, pp. Tu2B–4.

[3] F. Yousefi, A. G. Rahbar, and M. Yaghubi-Namaad, "Fragmentation-aware algorithms for multipath routing and spectrum assignment in elastic optical networks," *Optical Fiber Technology*, vol. 53, p. 102019, 2019.

[4] O. Gerstel, M. Jinno, A. Lord, and S. B. Yoo, "Elastic optical networking: A new dawn for the optical layer?" *IEEE communications Magazine*, vol. 50, no. 2, pp. s12–s20, 2012.

[5] B. C. Chatterjee and E. Oki, *Elastic Optical Networks: Fundamentals, Design, Control, and Management: Fundamentals, Design, Control, and Management*. CRC Press, 2020.

[6] B. C. Chatterjee, N. Sarma, and E. Oki, "Routing and spectrum allocation in elastic optical networks: A tutorial," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 3, pp. 1776–1800, 2015.

[7] B. C. Chatterjee, S. Ba, and E. Oki, "Fragmentation problems and management approaches in elastic optical networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 1, pp. 183–210, 2017.

[8] B. Bao, H. Yang, Q. Yao, A. Yu, B. C. Chatterjee, E. Oki, and J. Zhang, "Sdfa: A service-driven fragmentation-aware resource allocation in elastic optical networks," *IEEE Transactions on Network and Service Management*, 2021.

[9] X. Wang, R. Gu, and Y. Ji, "Multipath routing and spectrum allocation for network coding enabled elastic optical networks," *Current Optics and Photonics*, vol. 1, no. 5, pp. 456–467, 2017.

[10] L. Al-Tarawneh and S. Taebi, "Minimizing blocking probability in elastic optical networks by varying the bandwidth granularity based on optical path fragmentation," in *Photonics*, vol. 4, no. 2. Multidisciplinary Digital Publishing Institute, 2017, p. 20.

[11] R. V. Fávero, L. H. Bonani, and M. L. F. Abbade, "Spectral reallocation in lightpaths encompassing the most fragmented link of elastic optical networks," in *2016 18th International Conference on Transparent Optical Networks (ICTON)*. Ieee, 2016, pp. 1–4.

[12] E. Oki, T. Sato, and B. C. Chatterjee, "Spectrum fragmentation management in elastic optical networks," in *2019 21st International Conference on Transparent Optical Networks (ICTON)*. IEEE, 2019, pp. 1–4.

[13] B. C. Chatterjee, N. Stol, and E. Oki, "Impairment-aware spectrum allocation in elastic optical networks: A dispersion-sensitive approach," *Optical Fiber Technology*, vol. 61, p. 102431, 2021.

[14] B. C. Chatterjee and E. Oki, "Dispersion-adaptive first–last fit spectrum allocation scheme for elastic optical networks," *IEEE Communications Letters*, vol. 20, no. 4, pp. 696–699, 2016.

[15] B. C. Chatterjee, W. Fadini, and E. Oki, "A spectrum allocation scheme based on first–last-exact fit policy for elastic optical networks," *Journal of Network and Computer Applications*, vol. 68, pp. 164–172, 2016.

[16] W. Fadini, B. C. Chatterjee, and E. Oki, "A subcarrier-slot partition scheme with first-last fit spectrum allocation for elastic optical networks," *Computer Networks*, vol. 91, pp. 700–711, 2015.

[17] B. C. Chatterjee and E. Oki, "Defragmentation based on route partitioning in 1+ 1 protected elastic optical networks," *Computer Networks*, vol. 177, p. 107317, 2020.

[18] T. Sawa, F. He, T. Sato, B. C. Chatterjee, and E. Oki, "Defragmentation using reroutable backup paths in toggled 1+ 1 path protected elastic optical networks," in *2018 24th Asia-Pacific Conference on Communications (APCC)*. IEEE, 2018, pp. 422–427.

[19] S. Ba, B. C. Chatterjee, and E. Oki, "Defragmentation scheme based on exchanging primary and backup paths in 1+ 1 path protected elastic optical networks," *IEEE/ACM Transactions on Networking*, vol. 25, no. 3, pp. 1717–1731, 2017.

[20] S. Ba, B. C. Chatterjee, S. Okamoto, N. Yamanaka, A. Fumagalli, and E. Oki, "Route partitioning scheme for elastic optical networks with hitless defragmentation," *Journal of Optical Communications and Networking*, vol. 8, no. 6, pp. 356–370, 2016.

[21] B. C. Chatterjee, A. Wadud, and E. Oki, "Proactive fragmentation management scheme based on crosstalk-avoided batch processing for spectrally-spatially elastic optical networks," *IEEE JSAC Special Issue on Latest Advances in Optical Networks for 5G Communications and Beyond*, vol. 39, no. 9, pp. 2719-2733, 2021.

[22] J.-C. Chen, "Dijkstra's shortest path algorithm," *Journal of formalized mathematics*, vol. 15, no. 9, pp. 237–247, 2003.

[23] M. Salani, C. Rottondi, and M. Tornatore, "Routing and spectrum assignment integrating machine-learning-based qot estimation in elastic optical networks," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019, pp. 1738–1746.

[24] G. Bosco, V. Curri, A. Carena, P. Poggiolini, and F. Forghieri, "On the performance of nyquist-wdm terabit superchannels based on pm-bpsk, pm-qpsk, pm-8qam or pm-16qam subcarriers," *Journal of Lightwave Technology*, vol. 29, no. 1, pp. 53–61, 2011.

[25] M. S. Johnstone and P. R. Wilson, "The memory fragmentation problem: Solved?" *ACM Sigplan Notices*, vol. 34, no. 3, pp. 26–36, 1998.

[26] B. C. Chatterjee, N. Sarma, and P. P. Sahu, "Priority based routing and wavelength assignment with traffic grooming for optical networks," *Journal of Optical Communications and Networking*, vol. 4, no. 6, pp. 480–489, 2012.

Applications of Computational Geometry in Clustering: A Review

Tuhin Kumar Biswas[1], Kinsuk Giri[2,*]

[1]Department of CSE, National Institute of Technology, Durgapur - 713209, India
[2]Department of CSE, National Institute of Technical Teachers' Training & Research,
Block - FC, Sector - III, Salt Lake, Kolkata - 700106, India

[*]Corresponding Author: **kinsuk@nitttrkol.ac.in**

Abstract

Data analytic has extended its foundation in every sphere of our lives. In connection with the same, from the last two decades, data clustering is considered to be an important part of different research areas in order to identify the common contact relationship and inner characteristics of various objects. There are plenty of research works are established for data clustering in a efficient way. On the other hand computational geometry, which is widely used at various technical domain deals with different geometrical shapes and figures for different computing algorithms. The application of computational geometry in data clustering is becoming an emerging field of data science research. In this article, we have reported few important research works using computational geometry in the context of centroid hinged k-means clustering and density based DBSACN clustering.

Keywords: *Computational Geometry, Clustering, Empty Circles, Voronoi Diagram, Convex Hull*

1 Introduction

The field of data science and data mining are the most valuable research areas in today's era and for this regards the analysis of data objects distributed in a given space has always given the importance in area of extensive research. Computation geometry which is widely used at many technical domain, deals with various geometrical shapes and figures is a field of computer science that effectively represents various geometrical models through computing algorithms. While discussing on computational geometry, the convex hull, the voronoi diagram, and the empty circles are widely known in this field. Most of the modern technologies make uses the concept of computational geometry effectively. During spatio-temporal streaming of data, it becomes very essential to give correct results and present information in the best organized way possible.

In the other hand, one of the important area in unsupervised learning that deals with organization of data set of points is clustering. Computational geometry and clustering technology are two fields that have been successfully applied in numerous occasions to solve practical problems in various applications. The concepts of different geometrical figures of computational geometry, can be effectively used to in various data clustering techniques to achieve an improvement on clustering quality.

This article mainly focus on the detail analysis of combination between computational geometry and the clustering algorithms. It also focuses on the application of various concepts of computational geometry within various data clustering models. There are several types of clustering approaches available, such as hierarchical clustering, centroid based clustering, density based clustering and so on. Among them the k-means clustering and the DBSCAN clustering algorithms are the main two types of clustering that falls under the centroid based clustering and spatial clustering respectively. This report also includes a short analysis of improvement on traditional k-means & DBSCAN algorithm using computational geometry.

2 Computational Geometry

Computational geometry was first proposed by M. Shamos in his PhD. Thesis in the year of 1970. Since the inception of this area, plenty of research works have been done using computational geometry and the modification of these concept is also carried out till now. To demonstrate various real world objects and its representation the computational geometry is widely used. The difference between mathematical geometry and computational geometry is that computational geometry is vastly powerful discipline and very interesting and also it represents the direction of various points, lines, surfaces towards a particular planes. In the field of data mining the computational geometry plays a vital role to represents structural layers. It gives data its shape and tells us how to manipulate it efficiently as per our search requirement. One of the important basis computational geometry is the concept of convex hull, voronoi diagram and voronoi/empty circles. A brief introductions for these terms are given in the next subsections.

2.1 Convex Hull

The convex hull is a boundary among a set of data points such that all the points remain inside the boundary, with some points lies on the boundary line, considering as hull points. It means that for a given data set distributed in a space a convex hull is a boundary which is created with some points so that all the other points remain inside that boundary. So the convex hull is the smallest convex polygon that contains maximum data points in its interior and the rest points are the vertices of the polygon itself. Few popular algorithms are there to find the CH of a given set of data points, those are Graham Scan [1], Jarvis March [2] etc.

2.2 Voronoi Diagram

Voronoi diagram is used to generate the partitions in a space having a set of data points. This partitioning is performed in such a way that every region contains only one data point/object. The Russian mathematician Georgy Voronoi has introduced the concept of voronoi diagram [3]. It is defined as the partitioning of given set of points in a plane into some convex polygons or cells such that every point inside that convex polygon is closer to its generating point than any other point on the plane. The cross points of any two voronoi edges is known as voronoi vertices.

2.3 Empty Circles

From the name it is obvious that empty circle does not contain any point of dataset inside it. The voronoi vertices which can be extracted from voronoi diagram can be used as a center of empty circles and hence the empty circle can also be termed as voronoi circle. The importance of empty circle in clustering approaches is that the data points that remain on the boundary line or in the circumference of circles are closer to its center point than any other data point. So, for centroid based clustering approach it is easier to detect the closest points from empty circles

In figure 1, for the given data points(black large dots), the convex hull(Boundary with black lines), voronoi diagram(saffron lines) and empty circles(green circles) are shown [4].

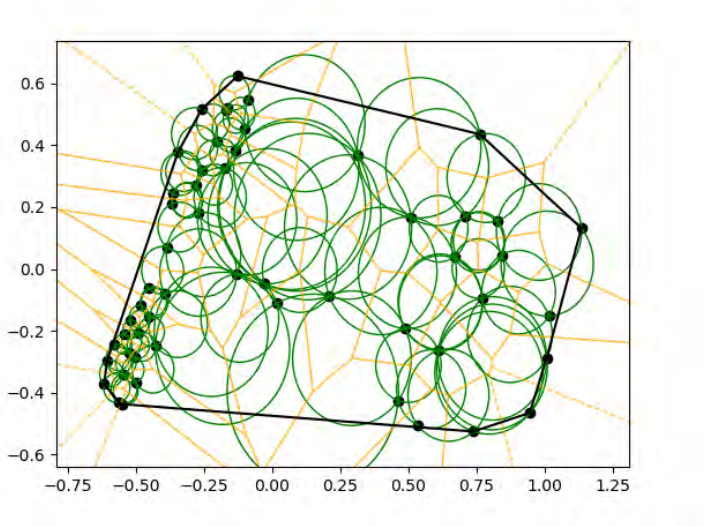


Figure 1: Convex Hull, Voronoi Diagram and Empty Circles for an artificial data set [4]

3 Data Clustering

Clustering is a technique to group data objects into various classes with a certain defined characteristics. In modern era of technology it is very much important to analyze the data into an organized manner. Clustering is such types of technique to make groups of similar data objects having a defined characteristics. Various types of clustering algorithm are available in literature. Few of those are: hierarchical clustering, centroid based clustering, spatial clustering etc. Among these. here we describe twy types,viz., centroid based k-Means clustering and spatial clustering DBMS.

3.1 k-Means Clustering

k-means clustering (KM) falls under the category of centroid based clustering. The main idea behind the KM is first of all the number of cluster is to be defined beforehand and the algorithm choose the initial center points for each clusters randomly from the given data set. Then it compute the Euclidean distance for every data point and initial centres. The data points are grouped into a specific cluster from which the Euclidean distance is less than the any other cluster. In the next when all the data points are grouped into various clusters the average values of data points in every cluster is computed and

this values becomes the new initial centroid for corresponding cluster for next successive steps. When the centroid values of two successive steps remain unchanged then we stop the clustering process. The random initialization method in traditional KM makes the clustering process ineffective and may leads the clustering up to a long processing time.

3.2 DBSCAN Clustering

DBSCAN [5] is one of the most applied clustering algorithm, which is density-based spatial clustering of applications with noise. It is one of the foremost common clustering algorithm works based on the density of objects. It depends on two important parameters viz., epsilon, the radius of neighbourhood around a data point and $minPts$, the minimum number of data points during a neighborhood to define a cluster. The DBSCAN algorithm find a core point if any point having the $minPts$ number of points within the radius of epsilon as well as it defines a point as border point if any point having lesser than the $minPts$ number of point within epsilon radius but one point itself is a core point. The DBSCAN algorithm is capable to detect noise point in such a way, when one point is neither a core point nor a border point.

4 Clustering with Computational Geometry

The concept of computational geometry was successfully used in a research work [6], by Reddy et al. k-means clustering can not ensure about the global optimum results always, as the algorithm select the initial clusters' centers in a random manner. This clustering is one of the most used clustering method in the field of data mining / data science due to it's simple operations and powerful output. But, this clustering technique uses a distance based approach to select the suitable data points of any cluster. On the other hand the first selection of initial centers of the clusters is done by a random selection approach. This can lead the clustering process stuck to a local optima. The authors in [6] addressed this issues into their proposed algorithm and tried to improve the clustering process using the circumference points of largest voronoi circles, which are constructed from the voronoi diagram of a given data set. Basically, the main idea of the proposed algorithm is to supply the initial cluster centers from voronoi circles instead of selecting it in a random manner. The authors have shown that their proposed algorithm outperforms against the traditional k-means and modified traditional k-means, in terms of miss classified patterns and error rate. This algorithm is capable enough to produce a better clustering in a qualitative nature as it is tested on various artificial and real world data. The experimented outcomes are shown in figure 2 for the algorithm proposed in [6].

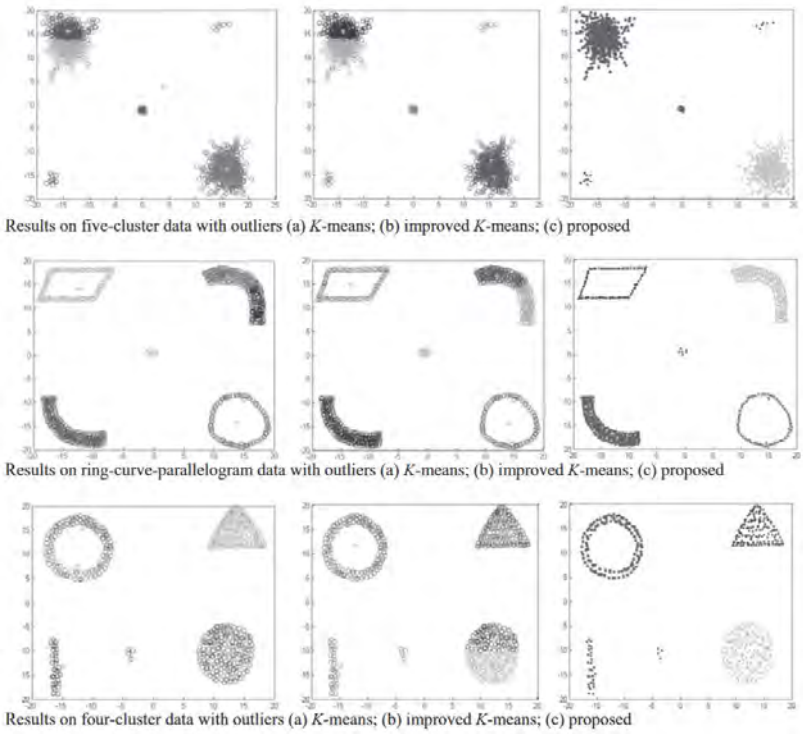


Figure 2: Results of proposed algorithm described in [6]

Another research work on data clustering from the same group was successfully presented in [7] with the integration of computational geometry. The works mainly focused on defining a new clustering method based on voronoi diagram. The voronoi diagram is efficient to make a complete partition of data objects into some specific regions having only one data object within a particular region. This partition ensure that the object inside the region is closer to its' generating points. In the voronoi diagram the cross point of any two edges is considered as a voronoi vertex. The proposed algorithm in [7] compute the largest voronoi circles, which can be drawn from voronoi vertices as a center of circles. This circles

have been used to locate the closer points represented by the voronoi vertices. This closer points are referred as a cluster prototypes. i.e. using several prototypes in every iterations the algorithm can detect the closer points of every clusters as well as using those points a new voronoi diagram is created to detect closer points of new voronoi vertices. In this way the authors successfully shown that the final generated clusters are much more better in qualitative nature than other existing algorithms. At a later stage the authors have established the validity of the proposed algorithm using a defined index termed as DVI(Dynamic Validity Index), whwere the proposed clustering method outperforms over traditional $K-means$, FCM & $CTVN$ clustering algorithms. The experimented outcomes are shown in figure 3 for the algorithm proposed in [7].

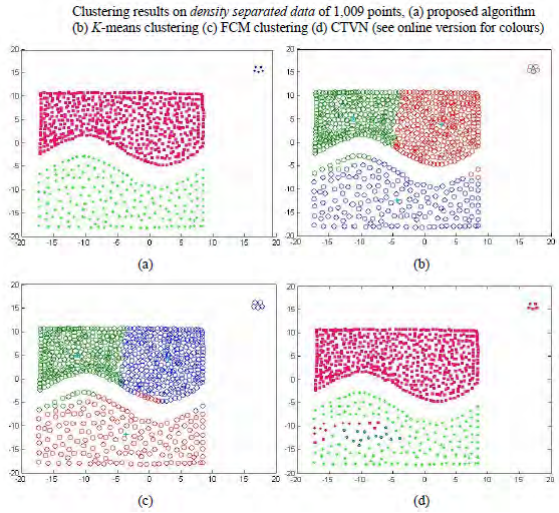


Figure 3: Results of proposed algorithm described in [7]

One more algorithm to detect the hardness of k-means clustering problem using the concept of computational geometry was introduced in [8]. k-means is a well known clustering approach problem which use the euclidean distance metric for the distance calculation among the data objects and cluster centers'. This euclidean k-means problem is extensively studied in the field of computational geometry. Instead of choosing random initial centers the authors proposed a method such that sum of squared distances of any data point to its nearest center is minimized. Using this concept the authors have reduced the vertex cover problem into euclidean k-means method to established the hardness of approximation of euclidean k-means algorithm.

The concept of border peeling clustering is successfully established [9], which is a non parametric clustering technique. Here the authors used the concept of layered architecture instead of finding a particular parameter for DBSCAN algorithm. Traditionally, we make the group/cluster of data objects first and the objects which can not satisfy the criteria for any generated clusters are simply treated as noise points but in this mentioned work the authors considered the entire clustering process as a layered architecture and defined that the border points can be considered as an external layer. In contrast of this, an external layer can explicitly identify the actual clusters. The authors has shown the proposed algorithm and the outcome of experimental analysis which is shown in figure 4.

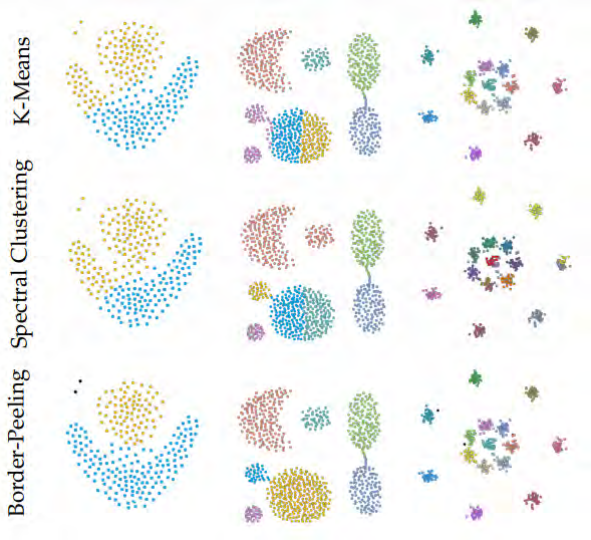


Figure 4: Border Peeling Clustering using the Algorithm Described in [9]

In the same context, a different approach for the optimization of time complexity for $DBSCAN$ was proposed by Jang

et al.[10]. They termed their method as $DBSCAN++$, a step towards a fast and scale able $DBSCAN$. They provided two simple strategies: uniform and greedy k-center-based sampling in which it is shown that for simulated and real data sets, $DBSCAN++$ runs in a fraction of the time compared to the normal $DBSCAN$.

Apart from all the above mentioned research works, we also have defined two new concepts for modification on centroid and spatial clustering using computational geometry approaches. In this context we have used the idea of convex hull, voronoi diagram and empty circles in our proposed work. It is well known that k-means clustering is very much sensitive to its initial cluster centers. We have focused on this part to achieve improvements on cluster quality. i.e. instead of random selection of initial cluster centers, our initial centers are obtained from the circumference points of the larger and largest empty circles [11]. Also, in order to catch the values from a feasible region, we have selected those empty circles which are interior to the convex hull of any given data set. For our proposed method an extensive experiments are done on various real world and artificial data. It is to be noted that our proposed algorithm outperforms over traditional k-means in both qualitative and quantitative nature. The performance of our proposed algorithm is shown in figure 5[11] for a synthetic data where the results of original k-means for the same data is also shown for meaningful comparison.

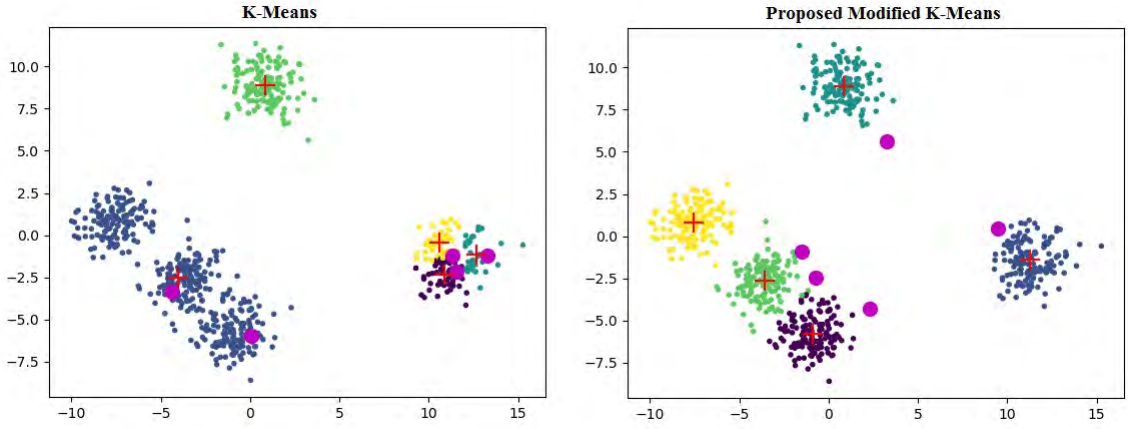


Figure 5: Modified K-Means using the Algorithm Described in [11]

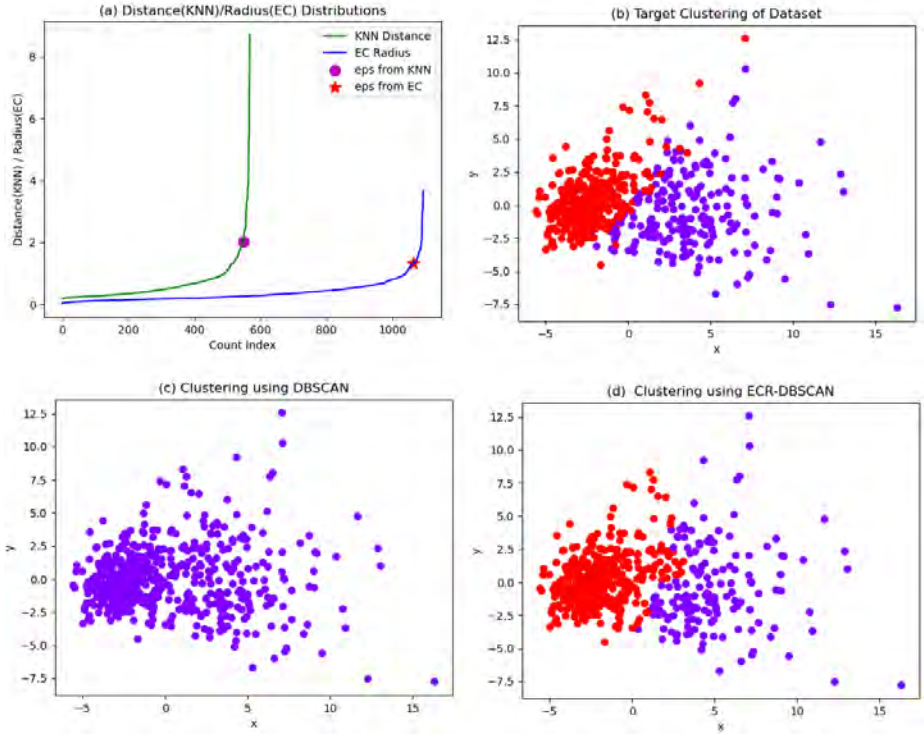


Figure 6: Spatial Clustering using the Algorithm Described in [12]

We have also applied the concept of empty circles into spatial clustering approaches in [12]. The DBSCAN is a well known density based spatial clustering method which is sensitive to noise points. Traditionally the epsilon parameter of DBSCAN is selected using the KNN distance based approach and the other parameter minimum points within the radius of epsilon can be treated as a free parameter. In our modified DBSCAN (termed as ECR-DBSCAN [12]), we have used the radii of empty circles [13] to select the epsilon parameter. All the sorted radii (in increasing order) of the circles are

used to find the elbow value of those radii. Next, we have intelligently used this elbow value as the epsilon parameter of the DBSCAN. This coupling between computational geometry and spatial clustering in ECR-DBSCAN enhance the performance of traditional DBSCAN while we redistribute the noise points into existing clusters using a distance metric. For the improvement on clustering quality through this proposed method is represented at figure 6[12].

5 Conclusions

The goal of writing this article was to make a very brief analysis between the possible relationship of computational geometry and data clustering methods. Clustering data into various group is a essential part in the field of data science. It is obvious that some defined concept of computational geometry can be effectively used in the operation of data clustering approaches. Empty circles can play a crucial role to in order to optimize the traditional clustering methods. However, we have discussed here only two types of clustering.The same can be to the other clustering to enhance the clustering quality. In future, we have a plan to carry out this analysis in detail.

References

[1] R. Graham, "An efficient algorithm for determining the convex hull of a finite planar set," *Inf. Process. Lett.*, vol. 1, pp. 132–133, 1972.

[2] R. Jarvis, "On the identification of the convex hull of a finite set of points in the plane," *Information Processing Letters*, vol. 2, no. 1, pp. 18–21, 1973.

[3] G. Voronoi, "Nouvelles applications des paramètres continus à la théorie des formes quadratiques. premier mémoire. sur quelques propriétés des formes quadratiques positives parfaites." *Journal für die reine und angewandte Mathematik (Crelles Journal)*, vol. 1908, pp. 97 – 102.

[4] T. K. Biswas, "A novel approach on finding initial centers for k-means clustering based on empty circles," Master's thesis, National Institute of Technical Teachers' Training & Research, Kolkata, 2020.

[5] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96. AAAI Press, 1996, p. 226–231.

[6] D. Reddy and P. K. Jana, "Initialization for k-means clustering using voronoi diagram," *Procedia Technology*, vol. 4, pp. 395–400, 2012, 2nd International Conference on Computer, Communication, Control and Information Technology(C3IT-2012) on February 25 - 26, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212017312003404

[7] D. Reddy and P. Jana, "A new clustering algorithm based on voronoi diagram," *Int. J. of Data Mining*, vol. 6, pp. 49 – 64, 01 2014.

[8] P. Awasthi, M. Charikar, R. Krishnaswamy, and A. K. Sinop, "The hardness of approximation of euclidean k-means," 2015.

[9] H. Averbuch-Elor, N. Bar, and D. Cohen-Or, "Border-peeling clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 7, pp. 1791–1797, 2020.

[10] J. Jang and H. Jiang, "Dbscan++: Towards fast and scalable density clustering," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3019–3029.

[11] T. K. Biswas and K. Giri, "A novel approach for initializing centroid at k-means clustering in paradigm of computational geometry," in *Accepted in Proceedings of 3rd International Conference on Recent Trends in Advanced Computing - Artificial Intelligence and Technology*. Springer Proceedings: Lecture Notes in Electrical Engineering, 2020.

[12] K. Giri, T. K. Biswas, and P. Sarkar, "Ecr-dbscan:an improved dbscan based on computational geometry," *Accepted in Machine Learning with Applications (MLWA), Elsevier*, 2021.

[13] K. Giri and T. K. Biswas, "Determining optimal epsilon(eps)on dbscan using empty circles," in *Accepted in Proceedings of nternational conference on Artificial Intelligence and Sustainable Engineering (AISE 2020)*. Springer Proceedings: Lecture Notes in Electrical Engineering, 2021.

Verifiable Visual Cryptography for Gray Scale Images with Meaningful Shares

Ujjal Kumar Das

Dept. of Computer Science, Srikrishna College, Bagula, Nadia, India
email: ujjal@srikrishnacollegebagula.ac.in

Abstract

In this paper a Verifiable Visual Cryptography for gray scale images with meaningful shares is proposed. In this approach we generate two gray scale meaningful shares which are verifiable in nature it means they can authenticate themselves at the receiver side. Proposed approach is a good combination of pixel based blind fragile watermarking approach, visual cryptography approach and scrambling approach together. Hence this approach ensures all security requirements like authentication, integrity verification, confidentiality etc. Visual cryptography approach is used to securely transmit the secret image over the internet. Due to meaningful shares, it is less vulnerable to cryptanalysis. Watermarking is used to verify the integrity of the shares at the time of any dispute. Scrambling is used to provide confidentiality to the secret image. A symmetric key is used by sender and receiver in order to authenticate them and for scrambling/unscrambling. Proposed approach also fulfils the contrast and security requirements of the VC approach. Experimental results demonstrate that proposed approach is good enough to localise the tampering with more than 90% accuracy at the same time it is able to recover the secret image with 100% accuracy.

Key words:*Visual cryptography, Verifiability, Meaningful shares,Gray scale secret, Share authentication.*

1.Introduction

There are three major security approaches are used in order to protect an image that are Visual Cryptography (VC), Digital Image Watermarking and Encryption or Scrambling approach. Every approach has its own significance and used to provide different security requirements like Authentication, Integrity verification, Confidentiality etc. For example, VC is used to securely transmit the secret image and it provides the security to the image at the time of transmission. Similarly fragile watermarking is used to provide integrity verification when an image is stored or during transmission if tampered. Scrambling is used to encrypt an image for providing confidentiality to the image. There are various state of the art approaches are proposed in all three fields.

VISUAL cryptography (VC) is a kind of secret sharing scheme, which is first proposed by Naor et al. [11], which allows the decryption of secretly shared images without any cryptographic computation. k-out-of-n is a special case where, a secret image is encoded into n number of shares. Every share is represented by random binary pattern. The shares are then printed onto transparencies, respectively, and distributed among *n* participants. Single share does not have any visual information about the shares. As per k-out-of-n Visual cryptography approach, if we superimpose less than k shares then no information regarding secret will be revealed but k or more then k shares can decode the secret without any computation just by stacking them. VC may also be used in various other applications like access control, watermarking, copyright protection [12], identification [13] and visual authentication. In order to understand the concept of visual cryptography, we can take an example of 2-outof- 2 VC scheme where k = 2 and n = 2 as shown in Fig. 1. Each pixel p of secret binary image is encoded into a pair of black and white subpixels for both shares. If p is white/black, one of the first/last two columns tabulated under the white/black pixel in Fig. 1 is chosen randomly so that selection probability will be 50%. Then, the first two subpixels in that column are allotted to share 1 and the following other two subpixels are allotted to share 2. Both the pixels either it is black or white will be encoded by two subpixels of black and white. Due to similar pattern used to encode both the intensities, an individual share has no information about the intensity of the secret image pixel whether p is black or white. The last row of Fig. 1. shows the stacking of both the shares, If the pixel p is black, the output of superimposition will be two black subpixels corresponding to a gray intensity 1. If p is white, then result of superimposition will be one white and one black subpixel, corresponding to a gray intensity as 1/2. By this method, one can visually recover all the

Here are the drawbacks of the existing VC approaches:

1. One can only secure binary images.
2. Secret image cannot be recovered with full accuracy and contrast.
3. All the shares are expanded in size as well random in nature which are the major problems.

So, in this paper we have considered all these issue in our account and developed a novel approach which will be able to secretly share a gray scale image with meaningful shares without pixel expansion.

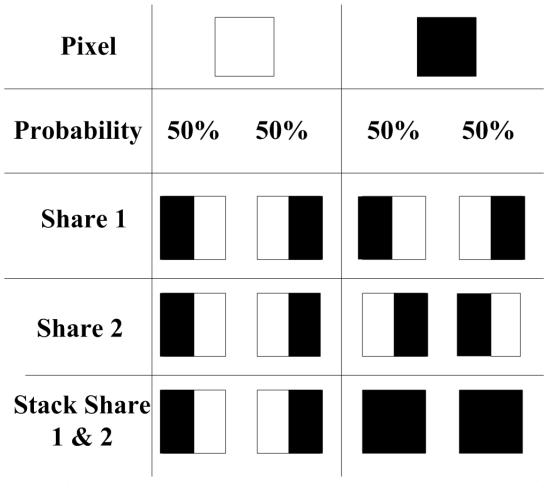


Fig. 1. 2-out of 2 VSS, where a secret pixel is encoded into two subpixels in each of the two shares

With increasing rate of the use of multimedia data, their alteration is also increased. There are various occasions when we need an authentic image without any tampering. Authentication and integration must be ensured at the time of storage as well as transmission of the secret [3]. These real demands lead to an emerging multimedia technology, known as image hashing. We study a new yet robust image hashing in this paper. Image hashing not only allows us to quickly find imagecopies in large databases, but also can ensure content security of digital images. Image hashing maps an input image to a short string, called image hash, and has been widely used in image retrieval [1], image authentication [4], digital watermarking [5], image copy detection [6], tamper detection [7], image indexing [8], multimedia forensics [9], and reduced-reference image quality assessment [10]. In the proposed approach, a self embedding method of fragile watermarking is used in order to ensure the integrity of the shares at the time of storage and transmission. Ateniese et al. [14] proposed the method of extended visual cryptography (EVC). In EVC, the shares contain both, shares are meaningful in nature. In case of EVC scheme, we recover very low quality of secret image and it is restricted for only binary images. Nakajima et al. [15] enhances the power of the EVC approach for gray scale images to improve the visual quality of images. Shyongjian [16] proposed a visual cryptography approach for multi tone color images but meaningless shares are the major problem of this technique. To generate meaningful shares, Zhou and Arce [2] proposed Halftone Visual Cryptography (HVC) which can be applied on error diffused half toned version of the secret images. This approach is more improved version of extended visual cryptography. The main drawback of HVC approach is the pixel expansion value. Zhonmin et al. [17] has proposed more refined version of extended HVC in which AuxiliaryBlack Pixel (ABP) is used in place of complimentary shares which can avoid the use of complementary shares. This approach again suffers with the problem of pixel expansion.

Rest of the paper is organized as: Section 2 provides the proposed approach. Section 3 demonstrates the experimental results and analysis. Paper is concluded in section 4 followed by references.

2.Proposed Approach

Figure 2 shows the flow diagram of the proposed approach. Here approach is divided into two phases: one is dedicated for sender end whereas another one is dedicated to receiver end. The block of sender end consists of three steps that are pre-processing, verifiable share generation and meaningful share generation. Similarly block of receiver end consists of two steps namely tamper detection and secret image recovery. We will see every step-in details in next subsections.

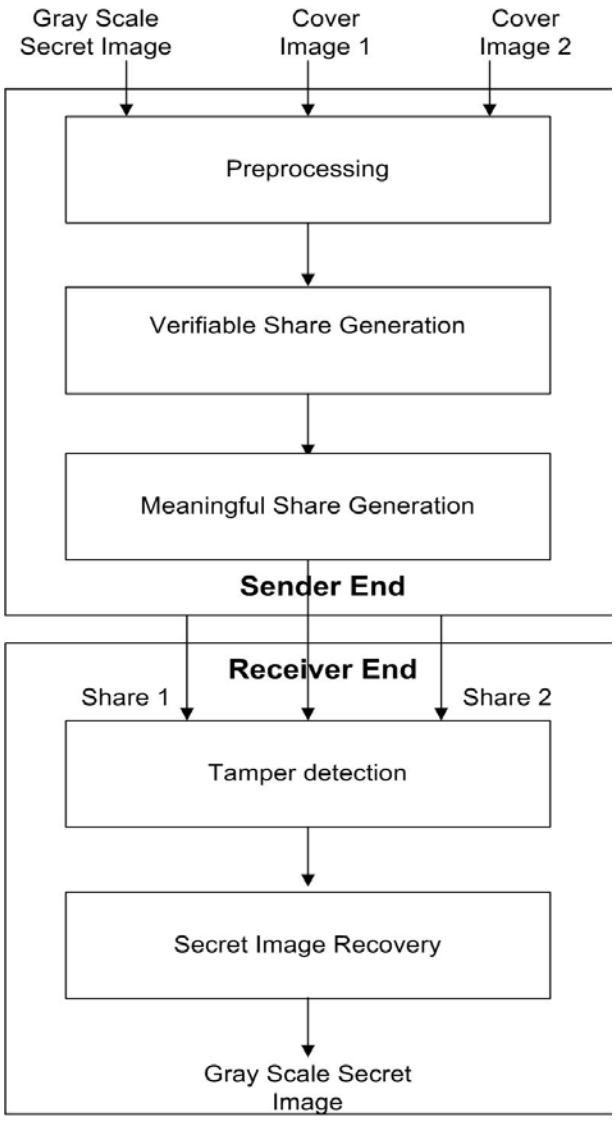


Figure 2- Flow Diagram of Proposed approach

2.1.Pre-processing

Proposed approach requires three gray scale images as input. One image is for secret image and other two images are for cover images. Cover images are those images which will be displayed on the meaningful shares. Before using these three images in the algorithm of share generation, they need to be pre-processed. Following steps are used to pre-process the given images.

Step 1-Suppose input gray scale secret image is denoted by I and cover images are denoted by C_1 and C_2 respectively.

Step 2- First of all convert$I, C_1\ and\ C_2$ into gray scale image if they are in colour image format.

Step 3-Resize all images $I, C_1\ and\ C_2$in same dimension let's say $\times N$.

Step 4- Now apply bit scrambling method like Arnold on all pixels of secret image I using a secret key k_1.

$$Is_{m,n} = Scramble((I_{m,n}), k_1)$$

Here $(I_{m,n})$ denotes the mnthpixel of input secret image I and $Is_{m,n}$ denotes the mnthpixel of scrambled secret image Since this image is of gray intensities. Hence pixels will be represented into eight bits. For example, 11000100 is binary bit stream of a pixel intensity of image I then after applying the scrambling it may look like 00110001. This scrambling approach is reversible in nature. It means at the receiver end, by using the same secret key k_1, one can get the original bit stream.

2.2.Verifiable Share Generation

Verifiable shares mean, a share will have the capability to authenticate itself. It is quite possible that during creation of the shares or during transmission of the shares, they might get tampered because of any intentional or unintentional attacks. Since shares are very sensible objects which carries the information of secret images, hence they need to be protected by any means. In this paper we have protected them by using self-embedding approach of fragile watermark. Here pixel wise blind authentication approach is proposed. It means at the receiver end, we do not need any information related to the original secret image and alteration can be detected at pixel level. Following steps are required in order to produce two authentication bits $Au_1$and Au_2 for every pixelof $Is, C_1\ and\ C_2$. Au_1will be used to protect pixels of share 1 S_1whereas Au_2 is used to protect share 2S_2. Following steps are required to get verifiable share generation.

Step 1- Take the gray level scrambled secret image Is, and cover images C_1 and C_2 as input. All are having the same dimension as $M \times N$.

Step 2- Following steps are used to generate the first authentication bit Au_1 for every pixel of *Is* and corresponding pixel of C_1.

Step 3-Let *p* be the pixel of *Is* which is having the location index *(i,j).*Similarly *q* denotes the pixel of cover image C_1 of the same index *(i,j).*

Step 4-*b(p,k)* and *b(q,k)* denotes the k^{th} binary bit of pixel *p* and *q* respectively.

Step 5-Compute the following

$$B[] = \sum_{k=1}^{3} b(p,k) \oplus b(p,k+1)$$

Where $B[]$ is an array which will have three bits.

Step 6- Now perform the following bitwise operation on *B* and *q*:

$$Au_1 = \sum_{k=1}^{3} (B(k) \quad \oplus q(9-k)) mod\ 2$$

Step 7 – Embed the bit Au_1 into the first LSB of pixel $S_1(i,j)$. Where S_1 is matrix of dimension $M \times N$ where all pixels are initialized with zeros of eight bits.

Step 8- Following steps are used to generate the first authentication bit Au_2 for every pixel of *Is* and corresponding pixel of C_2.

Step 9- Let *p* be the pixel of *Is* which is having the location index *(i,j).* Similarly *q* denotes the pixel of cover image C_2 of the same index *(i,j).*

Step 10-*b(p,k)* and *b(q,k)* denotes the k^{th} binary bit of pixel *p* and *q* respectively.

Step 11- Compute the following

$$B[] = \sum_{k=1}^{3} b(p,k) \oplus b(p,k+1)$$

Where $B[]$ is an array which *will* have three bits.

Step 12- Now perform the following bitwise operation on *B* and *q*:

$$Au_2 = \sum_{k=1}^{3} (B(k) \quad \oplus q(9-k)) mod\ 2$$

Step 13– Embed the bit Au_2 into the first LSB of pixel $S_2(i,j)$. Where S_2 is matrix of dimension $M \times N$ where all pixels are initialized with zeros of eight bits.

2.3.Meaningful Share Generation

Once we get two shares S_1 and S_2 which are preloaded with the authentication bits at first LSBs. Now our next objective is to generate meaningful shares from that. For creating the meaningful shares we use first three MSBs of the cover images. Here we use the concept of human visual system that our eyes cannot distinguish the visual intensity values of two gray values having minor difference. Following algorithm is used to create the meaningful shares.

Step 1- Take scrambled secret gray scale image *Is* of size $M \times N$and two cover images $C_1$and C_2 of same size $M \times N$as input. Here our output will be meaningful shares S_1 and S_2 of same size $M \times N$ which are already created and initialized.

Step 2-Repeat for*k=6 to 8*

$$P_{s1}(k) = P_{C1}(k)$$

$$P_{s2}(k) = P_{s1}(k)$$

Where $P_{s1}(k)$ and $P_{s2}(k)$ is k^{th} bit of pixel of S_1 and S_2 respectively. Whereas $P_{C1}(k)$ and $P_{C2}(k)$ is k^{th} bit of pixel of C_1 and C_2 respectively.

Step 3-*End for*

Step 4- Repeat for *j=2 to 5*

$$P_{s1}(j) = P_{ls}(j-1)$$

Where $P_{ls}(j)$ is j^{th} bit of pixel of scrambled secret image *ls* and $P_{s1}(j)$ is j^{th} bit of pixel of S_1.

Step 5-*End for*

Step 6- Repeat for j=2 to 5

$$P_{s2}(j) = P_{ls}(j+3)$$

Where $P_{ls}(j)$ is j^{th} bit of pixel of scrambled secret image *ls* and $P_{s1}(j)$ is j^{th} bit of pixel of S_2.

Step 7-*End for*

Aforesaid algorithm is used for embedding of all the cover image pixel's bits and all the scrambled image pixel's bits into the corresponding pixel's bits of shares. After applying this the eight bits of every pixel of S1 and S2 will hold three information that is content of cover image, content of scrambled secret image and authentication bit respectively as shown in figure 3.

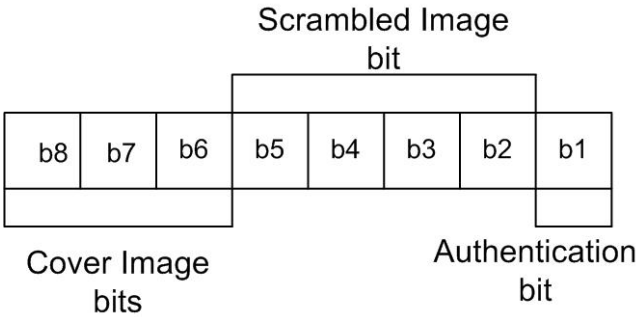


Figure 3: Bit assignment sequence for a pixel of the share

So these three steps are used to create two verifiable meaningful shares for a single secret image. Here shares as well as secret image both will be of gray scale in nature. Now we assume that these two shares are transmitted to the receiver for recovery. So it may be possible that during the transmission, because of any intentional or unintentional attacks, shares may get tampered. In that case receiver has to perform tamper detection test first before starting the recovery process. So in next subsection, we will see the remaining two steps of the proposed approach that is tamper detection and secret recovery.

2.4.Tamper Detection

This step is the responsibility of receiver. When he receives the shares, in order to ensure its authenticity, he will perform following operations. If alteration is detected in any of the given shares then receiver will ask to the sender to resend it.

By using proposed approach, sender not only will know about the unauthentic shares but also, he will know about the exact location of the tampering. Following are the steps for tamper detection.

Step 1- First of all extract the authentication bit Au_1 and Au_2 from last LSBs of every pixel of S_1 and S_2 respectively.

Step 2-Create two binary matrix namely M_1 and M_2 of dimension $M \times N$ by using extracted Au_1 and Au_2 respectively.

Step 3-Recalculate the Au_1 and Au_2 by using same algorithm which was used to create them at sender end by using seven MSBs of every pixel of S_1 and S_2 respectively.

Step 4- Create two binary matrix namely Mc_1 and Mc_2 of dimension $M \times N$ by using recalculated Au_1 and Au_2 respectively.

Step 5-Do pixel wise comparison between M_1 and Mc_1 as well as M_2 and Mc_2. If corresponding pixels of both the pair images are same then treat that pixel as tampered one else untampered one.

2.5. Secret Image Recovery

Once receiver finds the unaltered version of the shares, he needs to recover the secret image. In the proposed approach, we are getting the recovered secret with 100% accuracy, hence contrast condition is achieved by the proposed approach. Due to scrambled nature of the binary bit stream, security condition is also achieved. Following steps are used to recover the secret image:

Step 1- Takeverifiable meaningful sharesS_1 and $S_2$and secret key $k_1$as input.

Step 2-Perform following operation

$$B_1[\ \ ] = S_1(p_{i=5\ to\ 2})$$

Where $B_1[\ \ ]$ is an array which holds the fifth to second bits of a pixel of share.$S_1(p_{i=5\ to\ 2})$denotes the pixel p of the share S_1.

Step 3- Perform following operation

$$B_2[\ \ ] = S_2(p_{i=5\ to\ 2})$$

Where $B_2[\ \ ]$ is an array which holds the fifth to second bits of a pixel of share.$S_2(p_{i=5\ to\ 2})$denotes the pixel p of the share S_2.

Step 4- Append both the arrays in a single array with following order.

$$B[\ \ ] = B_1|B_2$$

Step 5-Convert the eight bit binary stream B into the decimal format which will indicate the i^{th} pixel intensity of the scrambled secret image Is.

Step 6-Since obtained secret is scrambled in nature hence if someone extract the bits to get the secret, due to scrambling secret will not be revealed. So by this way we can achieve confidentiality. Unscrambling is done by the same symmetric key k_1.

$$I_{m,n} = Unscramble((Is_{m,n}), k_1)$$

So finally, receiver will get the recovered secret with 100% accuracy and contrast. In next section we will see the efficiency of the proposed approach by the experimental results.

3. Experimental Result and Analysis

Experiments have been performed on various set of gray scale images for secret as well as for cover images. For all the sets satisfactory results are obtained. Experiments have been done into two phases. In first phase tamper detection capabilities is checked. For those two meaningful shares are generated for a gray scale secret image. Both the shares are then embedded with the verifiable bits. Now one or both the shares are intentionally tampered during transmission or storage. Now at the receiver end both the shares are checked for alteration detection. Some results are demonstrated to

show the efficacy of the proposed approach for getting the good tamper detection results. Figure 4 shows the result of tamper detection and recovery with tampered shares.

Figure 4- Example of intentional attack on the share, alteration detection and recovery with the altered shares.

In figure 4, image (a) shows the gray scale secret image which is going to be shared secretly. Image (b) and (c) are the meaningful shares which are also gray scale in nature. Both the shares are verifiable. It means if during the transmission they get tampered then both can self authenticate themselves. For example, image (d) shows the tampered version of the share 1 where an objectionable text is intentionally written over the image and share 2 is left unaltered. When they received by the receiver, he will apply tamper detection algorithm on both the shares. Images (f) and (g) are the results of tamper detection where we can see that black pixels show the unaltered pixels with their positions whereas white pixels show the altered portion. As we can see the image (f) corresponds to share 1 hence there are some white pixels on altered portions which shows the altered pixel localization and image (g) is completely black as there is no changes are done in share 2. Image (h) shows the recovery with the altered shares which is not same as the original secret image. It means when any share will be altered then we cannot get the secret image with 100% accuracy. Table 1 shows the quantitative results for the tamper detection capabilities of the proposed approach. These results justify our tamper detection capabilities because in most of the cases, we are achieving more 90% accuracy which is quite good. At the same time proposed approach minimizes the time complexity as shown in the table. As authentication is done with only a single bit per pixel hence time taken during the authentication is quite less.

Secret Image	**Altered pixel in Share 1**	**Altered pixel in Share 2**	**Detected Pixels in Share 1**	**Detected Pixels in Share 2**	**Accuracy %**	**Time in Second**
Barbara	339	93	309	79	88	10
Lena	519	578	489	469	92	12
Cameraman	119	629	110	658	90	17

Girl	432	859	403	823	95	15
Boat	331	531	319	499	93	13

Table 1- Tamper detection results for multiple secret images.

Figure 5 shows the plot between the number of altered pixels and number of detected pixels. We are getting the straight line which shows the effectiveness of the proposed approach. As it means that the number of detected pixels are almost same as the tampered pixels.

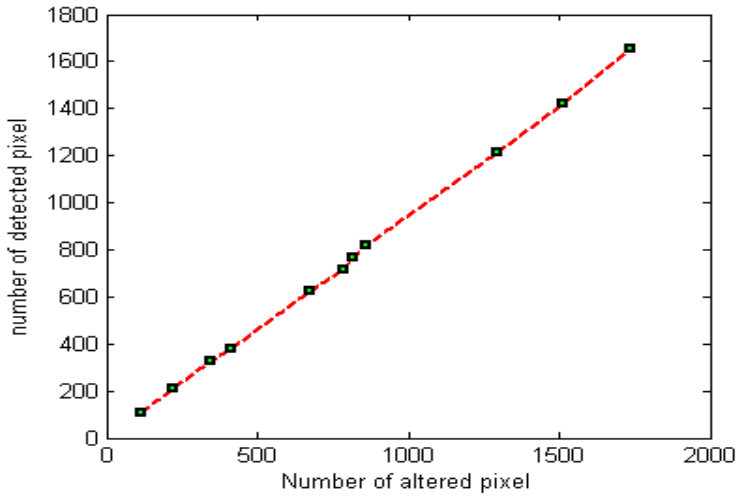


Figure5: Plot between the altered and detected pixels

Figure 6 shows the experimental results when no attack is done on any of the shares. In this result image (a),(b) and (c) are the gray scale secret image, meaningful share 1 and meaningful share 2. All three images are gray in nature. Images (d) and (e) show the shares after embedding the watermark or authentication bits at first LSB of every pixel of the shares. Here we can see that image (b) and (c) and corresponding images (d) and (e) are visually similar. There is no much difference which can be perceived by our human visual system. Though there may be maximum 1 intensity difference between corresponding pixels. Because, we are embedding the authentication bits only in a single first LSB. This difference can only be measured with various similarity metrics like PSNR, SNR, RMS error etc. Image (f) shows the recovered image by using both the shares. Image (a) that is original secret and image (f) that is recovered secret are completely identical. There is no contrast loss between these images during recovery.

Figure 6- Example of Meaningful and Verifiable Shares and Recovery of secret without any attack.

4.Conclusion

This paper proposes Verifiable Visual Cryptography for gray scale images with meaningful shares. Proposed approach generates two gray scale meaningful shares which can self authenticatethemselves. This approach provides all security requirements like authentication, integrity verification, confidentiality etc by combining pixel based blind fragile watermarking approach, visual cryptography approach and scrambling approach together. Visual cryptography approach is used to securely transmit the secret image over the internet. Watermarking is used to authenticate and to verify the integrity of the shares at the time of recovery. Scrambling is used to provide confidentiality to the secret image. A symmetric key is used by sender and receiver in order to authenticate them. Proposed approach fulfils the contrast and security requirements of the VC approach. 100% recovery is ensured by the proposed approach when no attack is done. Experimental results show that proposed approach is good enough to localise the tampering with more than 90% accuracy at the same time it is able to recover the secret image with 100% accuracy.

References

[1] M. Slaney and M. Casey, "Locality-Sensitive Hashing for Finding Nearest Neighbors," IEEE Signal Processing Magazine, vol. 25, no. 2, pp.128-131, Mar. 2008.

[2] Zhou Z, Arce GR, Di Crescenzo G (2006) Halftone visual cryptography. IEEE Trans Image Process 15(8):2441–2453.

[3] S. Wang and X. Zhang, "Recent Development of Perceptual Image Hashing," J. Shanghai Univ. (English ed.), vol. 11, no. 4, pp. 323-331, 2007.

[4] F. Ahmed, M.Y. Siyal, and V.U. Abbas, "A Secure and Robust Hash-Based Scheme for Image Authentication," Signal Processing, vol. 90, no. 5, pp. 1456-1470, 2010.

[5] C. Qin, C.C. Chang, and P.Y. Chen, "Self-Embedding Fragile Watermarking with Restoration Capability Based on Adaptive Bit Allocation Mechanism," Signal Processing, vol. 92, no. 4, pp. 1137- 1150, 2012.

[6] C.S. Lu, C.Y. Hsu, S.W. Sun, and P.C. Chang, "Robust Mesh-Based Hashing for Copy Detection and Tracing of Images," Proc. IEEE Int'l Conf. Multimedia and Expo, vol. 1, pp. 731-734, 2004.

[7] Z. Tang, S. Wang, X. Zhang, W. Wei, and S. Su, "Robust Image Hashing for Tamper Detection Using Non-Negative Matrix Factorization," J. Ubiquitous Convergence and Technology, vol. 2, no. 1, pp. 18-26, 2008.

[8] E. Hassan, S. Chaudhury, and M. Gopal, "Feature Combination in Kernel Space for Distance Based Image Hashing," IEEE Trans. Multimedia, vol. 14, no. 4, pp. 1179-1195, Aug. 2012.

[9] W. Lu and M. Wu, "Multimedia Forensic Hash Based on Visual Words," Proc. IEEE Int'l Conf. Image Processing, pp. 989-992, 2010.

[10] X. Lv and Z.J. Wang, "Reduced-Reference Image Quality Assessment Based on Perceptual Image Hashing," Proc. IEEE Int'l Conf. Image Processing, pp. 4361-4364, 2009.

[11] M. Naor and A. Shamir, "Visual cryptography," Advances in Cryptograhy: EUROCRYPT94, LNCS, vol. 950, pp. 112, 1995.

[12] M. S. Fu and O. C. Au, "Joint visual cryptography and watermarking," in Proc. IEEE Int. Conf. Multimedia and Expo, Taipei, Taiwan, Jun. 2004.

[13] M. Naor and B. Pinkas, "Visual authentication and identification," Crypto97, LNCS, vol. 1294, pp. 322340, 1997.

[14]Ateniese G, Blundo C, De Santis A, Stinson DR (2001) Extended capabilities for visual cryptography. TheorComputSci 250:143–161

[15] Nakajima M, Yamaguchi Y (2002) Extended visual cryptography for natural images. In: J. WSCG, vol 10, pp 303–310

[16] Shyu SJ (2007) Image encryption by random grids. PattRecog 40(3):1014–1031

[17] Wang Z, Arce GR, Crescenzo GD (2009) Halftone visual cryptography via error diffusion. IEEE Trans Inf Forensics Secur 4(3):383–396

Alexander-Spanier Cohomology Theory on Topological Spaces: A Review

Tushar Kanti Biswas[1]

[1]Department of Mathematics, Srikrishna College, Bagula, Nadia, India.
[1]Corresponding author: tusharsxc18@gmail.com

Abstract

In this article we review and analyse Alexander-Spanier cohomology theory on topological spaces and topological G spaces respectively, where G is a finite group. We will also review the Eilenberg-Steenrod axioms, tautness with respect to Alexander-Spanier cohomology theory in both topological spaces and topological G spaces context.

Keywords: *Alexender-Spanier cohomology, Eilenberg-Steenrod axioms, Bredon-Illman cohomology, Tautness.*

1 Introduction

In 1935 on compact metric spaces a cohomology module was introduced by James W. Alexander [1]. In 1948 Edwin H. Spanier generalized that cohomology theory for all topological spaces [14], which is known as the Alexander-Spanier cohomology theory. On the other hand in 1988 Hannu Honkasalo constructed equivariant version of Alexander-Spanier cohomology theory on a topological G space [18], where G is a finite topological group [17]. We will briefly review Alexander-Spanier Cohomology theory and equivariant Alexander-Spanier Cohomology theory in Section 2 and 3 respectively. In Section 4 we will explore the possibilities of some generalizations of equivariant Alexander-Spanier cohomology theory on topological G spaces with some suitable coefficient system.

Let X be a topological space [12] and A be a subspace of X. By an open neighbourhood [2] of A we mean a superset $N \subset X$ of A such that N is open in X. If N contains an open set containing A, then N is called a neighbourhood of A. A collection of subsets $\mathcal{U} = \{U_\alpha\}_{\alpha\in\Lambda}$ of X where Λ is an indexed set is said to be an covering of X if $X \subset \bigcup_{\alpha\in\Lambda} U_\alpha$. $\mathcal{U}$ is said to be an open covering if each U_α is open subset of X. A covering $\mathcal{V} = \{V_\beta\}_{\beta\in\Gamma}$ of X is a refinement of a covering $\mathcal{U} = \{U_\alpha\}_{\alpha\in\Lambda}$ if each V_β is contained in some U_α. A covering $\mathcal{U}$ of X is said to be locally finite if for any $x \in X$ there exists a neighbourhood N_x of x such that $\{\alpha \in \Lambda | U_\alpha \cap N_x \neq \phi\}$ is finite. A Housdorff topological space X is said to be a paracompact space [16] if every open covering has a locally finite open refinement. By $\bar{A}$ we denote the closure of a subset A of X. Let X and Y be two topological spaces. Then two maps $f, f' \to Y$ are said to be homotopic [4] if there exists a continuous map $F : X \times [0, 1] \to Y$ such that $F(x, 0) = f(x)$ and $F(x, 1) = f'(x)$ for all $x \in X$.

Now let X be a topological G space [18] and A be a G subspace (that is, A is invariant under G action), where G is a topological group (not necessarily finite). A covering $\mathcal{U} = \{U_\alpha\}_{\alpha\in\Lambda}$ is said to be a G covering of X if for all $g \in G$ and $U_\alpha \in \mathcal{U}$, $gU_\alpha \in \mathcal{U}$. Let X and Y be two topological G spaces. Then $X \times Y$ is also a G space with diagonal action. A map $f : X \to Y$ is said to be a G map if for all $g \in G$ and $x \in X$, $f(gx) = gf(x)$. Two G maps $f, f' : X \to Y$ are said to be G homotopic if there exists a G map $F : X \times [0, 1] \to Y$ such that $F(x, 0) = f(x)$ and $F(x, 1) = f'(x)$ for all $x \in X$. Note that G has trivial G action on $[0, 1]$.

A contravariant coefficient system is a contravariant functor [10] from the category of G spaces $G/H(H$ is a subgroup of $G)$ and maps between them to the category of Abelian groups (or R modules [3], where R is a ring with unity). We have used the same notation m to denote a R module (in Section 2) as well as a contravariant coefficient system (in Section 3).

2 Alexander Spanier Cohomology Theory

Let X be a topological space and m be a R module [3], where R is a ring with unity. For $q \geq 0$ let $C^q(X; m)$ be the R module consisting of all function $c : X^{q+1} \to m$. On $C^q(X; m)$ addition and scalar multiplication of functions are defined pointwise. Define a homomorphism $\delta : C^q(X; m) \to C^{q+1}(X; m)$ by setting

$$\delta c(x_0, x_1, \ldots x_{q+1}) = \sum_{i=0}^{q+1} (-1)^i c(x_0, x_1, \ldots \hat{x_i} \ldots x_{q+1})$$

Then it is easy to check that $\delta^2 = 0$. That is δ is a coboundary homomorphism and $C^*(X;m) = \{C^q(X;m), \delta\}_{q\geq 0}$ is a cochain complex.

An element $c \in C^q(X;m)$ is said to be locally zero cochain if there exists an open covering $\mathcal{U}$ of X such that $c(x_0, x_1, \ldots x_q) = 0$ whenever $x_0, x_1, \ldots x_q \in U$ for some $U \in \mathcal{U}$. This simply implies that c vanishes on $U^{q+1}, \forall U \in \mathcal{U}$. Let $C_0^q(X;m) = \{c \in C^q(X;m) | c \text{ is locally zero on } X\}$. It follows that if $c \in C_0^q(X;m)$, then $\delta c \in C_0^{q+1}(X;m)$. That is, $C_0^*(X;m) = \{C_0^q(X;m), \delta\}_{q\geq 0}$ is a subcomplex of $C^*(X;m)$. Let $\bar{C}^*(X;m)$ be the quotient cochain complex defined by $\bar{C}^*(X;m) = \{C^q(X;m)/C_0^q(X;m), \delta\}_{q\geq 0}$.

Definition 2.1. *The q^{th} dimensional cohomology module of the cochain complex $\bar{C}^*(X;m)$ is denoted by $\bar{h}^q(X;m)$ which will be called the q^{th} dimensional Alexander-Spanier cohomology module of the topological space X with coefficient m.*

For any two topological spaces X and Y and any map $f : X \to Y$, the induced cochain map $f^\# : C^*(Y;m) \to C^*(X;m)$ defined by $(f^\# c)(x_0, x_1, \ldots x_q) = c(f(x_0), f(x_1) \ldots f(x_q))$, where $c \in C^q(Y;m)$ and $x_0, x_1, \ldots x_q \in X$ induces a cochain map $f^\# : \bar{C}^*(Y;m) \to \bar{C}^*(X;m)$.

Now if A be a subspace of X , then the inclusion map $i : A \to X$ induces a cochain map $i^\# : \bar{C}^*(X;m) \to \bar{C}^*(A;m)$ which is an epimorphism. Let us denote the kernel of the map $i^\#$ by $\bar{C}^*(X, A;m)$. Then $\bar{C}^*(X, A;m)$ is a cochain subcomplex of $\bar{C}^*(X;m)$.

Definition 2.2. *The q^{th} dimensional cohomology module of the cochain complex $\bar{C}^*(X, A;m)$ is denoted by $\bar{h}^q(X, A;m)$ which will be called the q^{th} dimensional relative Alexander-Spanier cohomology module of the pair of topological space (X, A) with coefficient m.*

If $f : (X, A) \to (Y, B)$ be a map between pair of topological spaces (X, A) and (Y, B) then from naturality it follows that there exists a induced cochain map $f^\# : \bar{C}^*(Y, B;m) \to \bar{C}^*(X, A;m)$ which interns induces a map $f^* : \bar{h}^*(Y, B;m) \to \bar{h}^*(X, A;m)$.

2.1 Eilenberg-Steenrod Axioms

It can be shown that Alexander-Spanier cohomology modules satisfies the following axioms [15].

Theorem 2.1. *(Exactness Axiom) For each pair of topological spaces (X, A) there exists a long exact sequence* $\ldots \to \bar{h}^{q-1}(A;m) \to \bar{h}^q(X, A;m) \to \bar{h}^q(X;m) \to \bar{h}^q(A;m) \to \ldots$

Theorem 2.2. *(Dimension Axiom) For a one-point space X*

$$\bar{h}^q(X, m) = m \text{ if } q = 0$$
$$= 0 \text{ if } q \geq 0$$

Theorem 2.3. *(Excision Axiom) For a topological space X and two subspaces A and B with the property that $A \subset B$ and B is contained in an open neighbourhood U such that $\bar{U} \subset intA$, the inclusion map $i : (X - B, A - B) \to (X, A)$ induces isomorphism $i^* : \bar{h}^*(X, A;m) \to \bar{h}^*(X - B, A - B;m)$.*

Theorem 2.4. *(Homotopy Axiom) For any two topological spaces X, Y and two homotopic maps $f, f' : X \to Y$, $f^* = f'^*$.*

That is, these cohomology groups satisfy all the Eilenberg-Steenrod axioms for a cohomology theory.

2.2 Tautness

Let X be any topological space and A be a subspace of X. If N is any neighbourhood of A then the inclusion map $i : A \to N$ induces a restriction homomorphism $i^* : \bar{h}^*(N;m) \to \bar{h}^*(A;m)$. Now the collection of all neighbourhoods of A forms a downward directed set by inclusion. Hence as N varies over all neighbourhoods, these homomorphisms i^* determines a morphism $\eta : \lim_{\to} \bar{h}^*(N;m) \to \bar{h}^*(A;m)$.

Definition 2.3. *A is said to be a taut subspace* [15] *of X with respect to Alexander-Spanier cohomology if η is an isomorphism.*

E. Spanier [13] has shown that

Theorem 2.5. *If X is a paracompact Housdorff space and A be any closed subspace of X , then A is a taut subspace with respect to Alexander-Spanier cohomology.*

2.3 Relation with Equivariant Singular Cohomology

For $q \geq 0$ let $H^q(X;m)$ denote the q^{th} dimensional singular cohomology [5] of the topological space X. It can be shown that

Theorem 2.6. *If X is a paracompact Housdorff space, then there exists a isomorphism between $\bar{h}^*(X;m)$ and $H^*(X;m)$ provided each $x \in X$ is taut with respect to singular cohomology* [15].

3 Equivariant Alexander Spanier Cohomology Theory

Throughout this section G is a finite group. Let X be a G space and m be a contravariant coefficient system.

For $q \geq 0$, let $V_q(X) = \{\phi : G/H \times \{0,1,\dots q\} \to X$ is a G map, where H varies over the subgroups of $G\}$. H will be called the $\tilde{t}$-type of ϕ and will be denoted by $\tilde{t}(\phi) = H$. Note that we can rewrite the set $V_q(X)$ as $V_q(X) = \{\phi = (\phi_0, \phi_1 \dots \phi_i \dots \phi_q)|$ where $i \in \{0,1,\dots q\}$ and $\phi_i|_{G/H\times\{i\}} : G/H \to X$ is a G map$\}$.

Let $\hat{m} = \bigoplus_{H\leq G} m(G/H)$ and $C^q(X;\hat{m})$ be the cochain module of all maps $c : V_q(X) \to \hat{m}$ along with the coboundary homomorphism $\delta : C^q(X;\hat{m}) \to C^{q+1}(X;\hat{m})$, where for $c \in C^q(X;\hat{m}), \delta c \in C^{q+1}(X;\hat{m})$ and

$$\delta c(\phi_0, \phi_1 \dots \phi_i \dots \phi_{q+1}) = \sum_{i=0}^{q+1} (-1)^i c(\phi_0, \phi_1 \dots \hat{\phi_i} \dots \phi_{q+1})$$

A $\tilde{t}$-type preserving cochain [9] $c \in C^q(X;\hat{m})$ has the property that $\forall \phi \in V_q(X), c(\phi) \in m(G/\tilde{t}(\phi)$. Also a $\tilde{t}$-type preserving cochain $c \in C^q(X;\hat{m})$ is said to be equivariant if $c((\phi_0 o\gamma, \phi_1 o\gamma \dots \phi_i o\gamma \dots \phi_q o\gamma) = m(\gamma)(c(\phi_0, \phi_1 \dots \phi_i \dots \phi_q))$ whenever $\phi = (\phi_0, \phi_1 \dots \phi_i \dots \phi_q)$ and $\gamma : G/K \to G/\tilde{t}(\phi)$ is a G map.

Let

$$C^q_{\tilde{t}}(X;\hat{m}) = \{c \in C^q(X;\hat{m}) | c \text{ is } \tilde{t}\text{-type preserving}\}$$
$$C^q_G(X;\hat{m}) = \{c \in C^q_{\tilde{t}}(X;\hat{m}) | c \text{ is equivariant}\}$$

Note that if $c \in C^q_{\tilde{t}}(X;\hat{m}),$ then $\delta c \in C^{q+1}_{\tilde{t}}(X;\hat{m})$. Hence $C^*_{\tilde{t}}(X;\hat{m}) = \{C^q_{\tilde{t}}(X;\hat{m}), \delta\}_{q\geq 0}$ is a cochain subcomplex of $C^*(X;\hat{m}) = \{C^q(X,\hat{m}), \delta\}_{q\geq 0}$. A cochain $c \in C^q_{\tilde{t}}(X;\hat{m})$ is defined to be locally zero cochain [6] on X if there exists an open G covering $\mathcal{U}$ of X such that $c(\phi_0, \phi_1 \dots \phi_i \dots \phi_q) = 0,$ whenever $\phi_i(e\tilde{t}(\phi)) \in U,$ for some $U \in \mathcal{U}$ and for all $i \in \{0,1,\dots q\}$, where $e \in G$ is the identity element. By $C^q_0(X,\hat{m})$ let us denote all the elements $c \in C^q_{\tilde{t}}(X;\hat{m})$ which are locally zero on X.

It is easy to check that if $c \in C^q_G(X;\hat{m}), \delta c \in C^{q+1}_G(X;\hat{m})$ also if $c \in C^q_0(X;\hat{m}), \delta c \in C^{q+1}_0(X;\hat{m})$. This implies that $C^*_G(X;\hat{m}) = \{C^q_G(X;\hat{m}), \delta\}_{q\geq 0}$ and $C^*_0(X;\hat{m}) = \{C^q_0(X;\hat{m}), \delta\}_{q\geq 0}$ are cochain subcomplexes of $C^*_{\tilde{t}}(X;\hat{m})$. Now let us consider the following quotient complex $\bar{C}^*_G(X;\hat{m}) = \{\bar{C}^q_G(X;\hat{m}), \delta\}_{q\geq 0}$, where

$$\bar{C}^q_G(X;\hat{m}) = \frac{C^q_G(X;\hat{m})}{C^q_G(X;\hat{m}) \cap C^q_0(X;\hat{m})}$$

Definition 3.1. *The q^{th} dimensional cohomology module of the cochain complex $\bar{C}^*_G(X;\hat{m})$ is denoted by $\bar{h}^q_G(X;\hat{m})$ which will be called the q^{th} dimensional Equivariant Alexander-Spanier cohomology module of the topological G space X with coefficient system m.*

For any two topological G spaces X and Y and any G map $f : X \to Y$, the induced cochain map $f^\# : C^*(Y;\hat{m}) \to C^*(X;\hat{m})$ defined by $(f^\# c)(\phi_0, \phi_1, \dots \phi_q) = c(fo\phi_0), fo\phi_1) \dots fo\phi_q))$, where $c \in C^q(Y;\hat{m})$ and $(\phi_0, \phi_1, \dots \phi_q) \in V_q(X)$ induces a cochain map $f^\# : \bar{C}^*_G(Y;\hat{m}) \to \bar{C}^*_G(X;\hat{m})$.

Now if A be a G subspace of X , then the inclusion map $i : A \to X$ induces a cochain map $i^\# : \bar{C}^*_G(X;\hat{m}) \to \bar{C}^*_G(A;\hat{m})$ which is an epimorphism. Let us denote the kernel of the map $i^\#$ by $\bar{C}^*_G(X,A;\hat{m})$. Then $\bar{C}^*(X,A;\hat{m})$ is a cochain subcomplex of $\bar{C}^*_G(X;\hat{m})$.

Definition 3.2. *The q^{th} dimensional cohomology module of the cochain complex $\bar{C}^*_G(X,A;\hat{m})$ is denoted by $\bar{h}^q_G(X,A;\hat{m})$ which will be called the q^{th} dimensional relative Equivariant Alexander-Spanier cohomology module of the pair of topological space (X,A) with coefficient system m.*

If $f : (X,A) \to (Y,B)$ be a G map between pair of topological G spaces (X,A) and (Y,B) then from naturality it follows that there exists a induced cochain map $f^\# : \bar{C}^*_G(Y,B;\hat{m}) \to \bar{C}^*_G(X,A;\hat{m})$ which interns induces a map $f^* : \bar{h}^*_G(Y,B;\hat{m}) \to \bar{h}^*_G(X,A;\hat{m})$.

3.1 Eilenberg-Steenrod Axioms for an Equivariant Cohomology

Hannu Honkasalo [6] has shown that Equivariant Alexander-Spanier cohomology modules satisfies the following axioms

Theorem 3.1. *(Exactness Axiom) For each pair of topological G spaces (X, A) there exists a long exact sequence* $\ldots \to \bar{h}_G^{q-1}(A;\hat{m}) \to \bar{h}_G^q(X, A;\hat{m}) \to \bar{h}_G^q(X;\hat{m}) \to \bar{h}_G^q(A;\hat{m}) \to \ldots$

Theorem 3.2. *(Dimension Axiom) For any subgroup H of G,*

$$\begin{aligned}\bar{h}_G^q(G/H;\hat{m}) &= 0 \quad if\ q > 0,\\ &= m(G/H)\ \ if\ q = 0.\end{aligned}$$

For $q = 0$ there is an isomorphism $\bar{h}_G^q(G/H;\hat{m}) \to m(G/H)$, natural with respect to G maps $\hat{g} : G/H \to G/K, g^{-1}Hg \subset K$.

Theorem 3.3. *(Excision Axiom) For a topological G space X and two G subspaces A and B with the property that $A \subset B$ and B is contained in an open G neighbourhood U such that $\bar{U} \subset intA$, the inclusion map $i : (X - B, A - B) \to (X, A)$ induces isomorphism $i^* : \bar{h}_G^*(X, A;\hat{m}) \to \bar{h}_G^*(X - B, A - B;\hat{m})$.*

Theorem 3.4. *(Homotopy Axiom) For any two topological G spaces X, Y and two G homotopic maps $f, f' : X \to Y$, $f^* = f'^*$.*

That is, these cohomology groups satisfy all the Eilenberg-Steenrod axioms for an equivariant cohomology theory.

3.2 Tautness

Let X be any topological G space and A be a G subspace of X. If N is any G neighbourhood of A then the inclusion map $i : A \to N$ induces a restriction homomorphism $i^* : \bar{h}_G^*(N;\hat{m}) \to \bar{h}_G^*(A;\hat{m})$. Now the collection of all G neighbourhoods of A forms a downward directed set by inclusion. Hence as N varies over all G neighbourhoods, these homomorphisms i^* determines a morphism $\eta : \lim_{\to} \bar{h}_G^*(N;\hat{m}) \to \bar{h}_G^*(A;\hat{m})$.

Definition 3.3. *A is said to be a taut subspace* [6] *of X with respect to Equivariant Alexander-Spanier cohomology if η is an isomorphism.*

H. Honkasalo [6] has shown that

Theorem 3.5. *If X is a paracompact G space and A be any closed G subspace of X , then A is a taut subspace with respect to Equivariant Alexander-Spanier cohomology.*

3.3 Relation with Equivariant Singular Cohomology

For $q \geq 0$ let $H_G^q(X; m)$ denote the q^{th} dimensional equivariant singular cohomology [8] of the topological G space X. It can be shown that [6]

Theorem 3.6. *If X is a paracompact G space and every orbit $Gx \subset X$ is taut with respect to equivariant singular cohomology, then there exists a isomorphism between $\bar{h}_G^*(X; m)$ and $H_G^*(X; m)$.*

4 Conclusion

H. Honkasalo constructed equivariant Alexander-Spanier cohomology for a topological G space X, where G is a compact lie group [7]. Equivariant Alexander-Spanier cohomology may be constructed and studied when G is an arbitrary topological group.

A. Mukherjee and G. Mukherjee in [11] constructed Bredon-Illman cohomology $H_G^*(X; M)$ with a suitable local coefficient [11] M. This cohomology satisfies all the Eilenberg-Steenrod axioms with respect to the coefficient system M. If M is simplie in some sense [11], then Bredon-Illman cohomology reduces to the equivariant singular cohomology with contravariant coefficient system [8]. Hence it is indeed a generalization of the equivariant singular cohomology [8]. So there is a scope of generalization of equivariant Alexander-Spanier cohomology with respect to a local coefficient system [11]. It may also be verified that the generalized equivariant Alexander-Spanier cohomology with respect to a local coefficient system satisfies all the Eilenberg-Steenrod axioms, some tautness property in the local coefficient system context. It may also share some isomorphic relation with Bredon-Illman cohomology [11] under some suitable condition.

References

[1] JW Alexander, *On the chains of a complex and their duals*, Proceedings of the National Academy of Sciences of the United States of America **21** (1935), no. 8, 509.

[2] Jacques Dixmier, *General topology*, Springer Science & Business Media, 2013.

[3] David S Dummit and Richard M Foote, *Abstract algebra*, Vol. 1999, Prentice Hall Englewood Cliffs, NJ, 1991.

[4] Marvin J Greenberg and John R Harper, *Algebraic topology: a first course*, CRC Press, 2018.

[5] Allen Hatcher, *Algebraic topology*, Cambridge University Press, 2005.

[6] Hannu Honkasalo, *Equivariant alexander-spanier cohomology*, Mathematica Scandinavica (1988), 179–195.

[7] ———, *Equivariant alexander-spanier cohomology for actions of compact lie groups*, Mathematica Scandinavica (1990), 23–34.

[8] Sören Illman, *Equivariant singular homology and cohomology i*, American Mathematical Soc., 1975.

[9] ———, *Equivariant alexander–spanier cohomology and pa smith theory*, Topology and its Applications **210** (2016), 269–291.

[10] Nathan Jacobson, *Basic algebra i*, Courier Corporation, 2012.

[11] Amiya Mukherjee and Goutam Mukherjee, *Bredon-illman cohomology with local coefficients*, The Quarterly Journal of Mathematics **47** (1996), no. 2, 199–219.

[12] James R Munkres, *Elements of algebraic topology*, CRC press, 2018.

[13] Edwin Spanier, *Tautness for alexander-spanier cohomology*, Pacific Journal of Mathematics **75** (1978), no. 2, 561–563.

[14] Edwin H Spanier, *Cohomology theory for general spaces*, Annals of Mathematics (1948), 407–427.

[15] ———, *Algebraic topology*, Springer Science & Business Media, 1989.

[16] Arthur H Stone, *Paracompactness and product spaces*, Bulletin of the American Mathematical Society **54** (1948), no. 10, 977–982.

[17] Mikhail Tkachenko and AV Arkhangelskii, *Topological groups and related structures* (2008).

[18] Tammo tom Dieck, *Transformation groups*, Vol. 8, Walter de Gruyter, 2011.

Material Science

Machine Learning inElectrochemical Energy Storage: Practice and Discussion

Sourav Ghosh[a,b,*],G. Ranga Rao[b], Tiju Thomas[a,*]

[a]Department of Metallurgical and Materials Engineering, Indian Institute of Technology Madras, Chennai 600036, India
[b]Department of Chemistry, Indian Institute of Technology Madras, Chennai 600036, India
*Corresponding author: Tiju Thomas (tijuthomas@iitm.ac.in), Sourav Ghosh (srvg742@gmail.com)

AbstractData-based materials designing and selection is an integral part of modern materials research. Even in the case of electrochemical energy storage, it is an emerging field. A discussion on the fundamentals of ML in materials research is provided here. Further, a brief review of fingerprinting and learning strategies is offered, followed by an introductory guide to the fair practices of ML in materials science. Cautions are crucial in every stage of the work: (a) the database curation, (b) the modeling part, and (c) fitting, testing, and benchmarking the data. A careful trade-off between the predictiveaccuracy and understanding of the physicochemical insights needs to be carried out while practicing ML. Furthermore, the application of ML is discussed, especially in energy storage materials and relevant applications. Finally, the importance of experimental validation of ML work is discussed and justified for better reproducibility and reliability.

Keywords:*Machine Learning; Fingerprinting; Overfitting; Energy Storage Materials; Validation.*

1. Introduction

Data-to-knowledge is an integral part of the modern information revolution, and machine learning (ML) is one of the crucial aspects. Data transformation to knowledge is the fundamental criteria if 'data-intensive discovery' [1]. The transformation ensures data transition from an active role to a passive one. The passive role denotes the use case of data, where it is limited to confirming or refuting a particular hypothesis. On the contrary, in the modern information-centric era, data have a pivotal role in raising new hypotheses at an unprecedented scale. In order to succeed in the passive-to-active conversion, the fundamental idea is to translate the data into a machine-readable format [2]. However, further indulging in ML in materials science, it is essential to recognize the root of this technique.

A vital part of ML is the process of decision-making. After encountering a new situation, cognitive systems (including even humans) tend to decide based on similar experiences from the past. Consequently, a judgemental error may occur depending on the new situation from the past or experienced encounters. The error leads to new lessons, which become part of the experience. Hence, ideally, the intrinsic capability of the cognitive system and the decision-making ability are supposed to improve based on the increasing encounters with novel scenarios. In ML, the experience is popularly denoted as the 'training data,' new encounters are known as the 'test data,' and the cognitive system can be called the 'model.' Although the data-driven science is comparatively novel compared to its theoretical and experimental counterparts, application of data-science has its roots in history.

Ramprasad *et al.* coined a timeline containing classic instances based on data-driven scientific and engineering efforts [**Fig. 1.**][3]. The timeline shows that data-driven discovery dates back to the 6th century BC. Ancient populations from India and Sri Lanka used alloying elements to inhibit the rusting tendency of iron-based on their creativity and experience. Data science has developed a long way since the early 21st century, when the terms like 'data-driven', 'materials-informatics' etc., became a critical part of modern-day materials research. Despite a worldwide acceptance and active research on ML in materials science, there are many unresolved questions on the appropriateness of ML in several problems. In this article, an effort has been made to provide a brief overview of ML and informatics in materials science. The basis and introductory best practices are discussed with the help of the literature. Finally, recent data-driven advances in materials science are provided with a holistic view of the field.

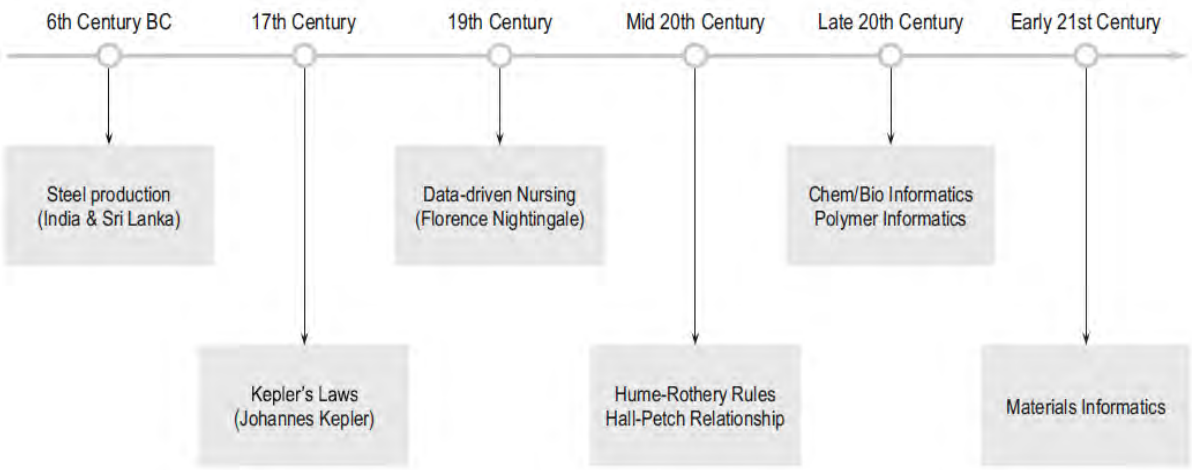


Fig.1. Timeline of data-driven efforts in science and engineering[3].

2. Machine Learning in Materials Science: The Basis

A significant issue of materials science is the problem of dimensionality. The development of novel materials and the study of their functional properties combined with the already well-studied materials raise the complexity of the structure-property relationship. For the characterization of the material, microstructure, composition, and even processing conditions are considered crucial parameters. The complexity is even enhanced when applications are considered. Thus, material modeling becomes challenging due to high-parameter spaces. Hence, relying on just experimental studies can often be a matter of concern from a time and cost perspective. The use of data can aid in reducing complications. Interestingly, irrespective of the materials science problem, the prerequisite of ML is the same – the available data. ML-based problem solving can have a particular structure regardless of the problem to simplify the concept.

A clean and reliable dataset is ML studies' first and foremost criteria. Once the dataset is ready, the next step is to find suitable descriptors to identify fingerprints. Fingerprinting is a numerical representation of the input data, where the input case is reduced to a set of numbers. In short, this is the step where the dataset is made readable to machines. Finally, fingerprinting goes through a learning mechanism so that fingerprints can be mapped to property values. Thus, the material-property correlation is established, and the model can be used for further prediction. The most significant step in the process is to reduce the input data to fingerprints as it requires knowledge of the relevant materials class and the application. However, the learning part is a typical statistical process, where cross-validation and training are carried out on the data. The training process ensures model development at a level where the original dataset gains the ability to accommodate new data.

Overall, machine learning in material science is comprised of four steps – (i) dataset creation, (ii) fingerprinting, (iii) mapping and learning stage, and (iv) new data infusion to test the model for prediction purposes. The main objective behind the whole process is to deliver a recommendation system capable of improving through continuous adaptability.

2.1. Further Insight on the Fingerprints and Learning

As mentioned above, fingerprints are arguably the most significant component of the ML paradigm. The fundamental need of ML is to convert a large dataset into a quantitative scheme, and the selection of suitable fingerprints is the pivotal part of that. This step requires domain knowledge, experience, and a clear idea of the endgoal - hence warrants a deeper discussion. Fingerprints are the authentic materials' proxies, also known as descriptors and features. The process of fingerprinting is dependent on the desired level of accuracy of the model. For instance, if the accuracy of the prediction is deemed to be less critical, then the fingerprints can be at a grosser level. The case would be precisely the opposite in a case where a higher degree of accuracy is demanded – in the case of prediction of specific material property.

A basic understanding of more acceptable fingerprinting is often associated with higher prediction accuracy and more complexity of the ML model. Material representation should adhere to a certain level of invariance against stiff rotation or translation of the material. A comparison can be drawn with the problem of biometric or facial recognition schemes. The representation cannot be dependent on the rotated or enlarged facial images. Similarly, if a fingerprint considers information on the atomic position, that should not be altered because of the permutation of the atoms. Another important criterion is that the fingerprints combinedly should capture all yet only the relevant aspects of the dataset. Type of learning (supervise, unsupervised) is also vital here and discussed later in this section.

A classic example of utilization of gross level fingerprints is finding out if a mixture of a couple of metals can result in a solid solution, i.e., Hume-Rothary rules. Crystal structure, atomic sizes, oxidation state, and electronegativities of each

metal are considered features or fingerprints [3]. Experiments were the critical way to develop such rules in the past and can be possible to reciprocate by current ML and data mining tactics. For efficient identification of probable nonlinear multivariate correlation, the initial fingerprints can be selected in terms of relevant primary parameters. Then multiple algebraic combinations can be performed to create a high number of nonlinear mathematical functions. Next, the ample space of functions needs to be tested to identify a highly correlated subset with the targeted property. The associated approaches have been recently developed in compressed sensing, genetic programming, and information technology [4,5].

The process mentioned above of fingerprinting consists of three distinct steps – (i) identification of fundamental and relevant components, (ii) creation of a more extensive set of nonlinear mathematical functions, and (iii) finding suitable and strongly property correlated subsets. The second and third stages are related to scalability problems associated with feature selection and dimensionality reduction. Dimension reduction helps to make the dataset thinner and more comprehendible [6,7]. Dimensionality reduction technique means converting a data matrix 'A' of dimensions M × K into another new matrix 'B' with a dimension of M × k where k << K, after affirming that the original information is preserved. However, the accurate completion of the task is nearly impossible, and often approximate solutions are considered for the job.

Among many approaches, the least absolute shrinkage and selection operator (LASSO) and principal component analysis (PCA) has gathered massive attention in this field [8,9]. Both of them are capable of reducing the dimension but follow different mechanisms. LASSO works by discarding redundant and lesser relevant (or even irrelevant) features. The relevance is determined by how strongly the features are correlated with the property to be predicted. On the other hand, PCA creates a set of new features (smaller in dimension) by aggregating the original features. Each principal component is selected in a way so that they remain uncorrelated to other principal components and stay along the direction of the largest variance. On a larger scale, these techniques are classified into two different types (a) linear projection methods and (b) nonlinear projection methods. PCA, multidimensional scaling, and time-lagged independent component analysis are a few linear projection methods used in materials science. Isomap, diffusion map, sketchmap etc., are nonlinear methods.

Linear projection methods work on linear manifolds and often depend on a specific parameter. The techniques are generally fast and robust. Nonlinear methods are required to work on curved-twisted manifolds and highly nonlinear and complex manifolds. These techniques are dependent on a few parameters and work fast on the manifolds that can be made linear easily. However, highly complex manifolds need increasing sophistication in the dimensionality reduction techniques and are even challenging to fine-tune to make problem-specific [10].More efficient and complex techniques require higher computational power and significantly high trial and error before final deployment. Detailed discussion on each technique can be found elsewhere in the literature.

After establishing successful fingerprints, the next focus of the ML becomes the learning process. There are three types of learning methods: (a) supervised learning, (b) unsupervised learning, and (c) reinforcement learning. Supervised learning connects a set of known input variables to the known output variable. Depending on the continuity of the space, supervised ML algorithms perform either a regression task (continuous space, e.g., polarizability, energy, etc.) or a classification task (involves categorical values, e.g., direct or indirect bandgap, acid, or base, etc.). A supervised learning method is trained to correlate input patterns or structures to known output values to deduce an input-output relationship. Thus, the ML model can predict the outcome beyond the training data. On the contrary unsupervised learning handles the problems with known input with no associated labels. The target here is to identify and interpret any underlying structure of the dataset. Dimensionality reduction techniques become handy in unsupervised learnings. Reinforcement learning is involved defining a method and receiving feedback. Feedbacks are generally rewarding (for the desired result) or punishment (for undesired result). The goal is to seek maximum overall reward to accomplish an optimized solution [11].

3. Best Practices and an Introductory Guide

A powerful tool like ML should be applied with utmost care. A careful evaluation is required to justify the appropriateness of utilizing ML for a material science problem. Ideally, machine learning is helpful where the data and especially the inter-data interactions are too complex in terms of interpretability and conceptualization from a 'human-learning' point of view. Keeping this clarity is vital because in cases especially with smaller data sizes, the human mind can often understand the relationship with data better than the machine. Another vital aspect of ML is the interpretability and predictability trade-off [12]. A complex and less interpretable model often provides higher efficiency. However, finding interpretable physical and chemical insights are pretty rare in a complex and robust model with high prediction efficiency like neural networks. It is hard to understand and extract significant insights from the high-performance 'black-box' models.

On the contrary, a simpler model can potentially offer an easier route to significant physical and chemical insights but lack prediction efficiency [13]. A rough schematic of ML work in materials science is provided in **Fig.2.** Keeping this discussion as a backdrop, a guide for better practice of using ML in materials science is discussed in this section.

3.1. Dataset Creation, Processing, and Splitting

First, during dataset creation, the size and balance of the dataset are crucial. The dataset should be large with enough data points containing the material space one desire to study. On the other hand, a larger dataset is computation-heavy and time-consuming during the prototype phase. However, the balance is essential to avoid unwanted bias towards a particular type of condition in the database. Dataset imbalance remedies can be classified into four techniques discussed by Garcia et al.[14]: (a) Resampling methods, (b) modifications of learning algorithms, (c) measurement of classification performance problems in the imbalance domain, and (d) understanding of the correlation between class imbalance and other data-complexity. A brief discussion is provided below.

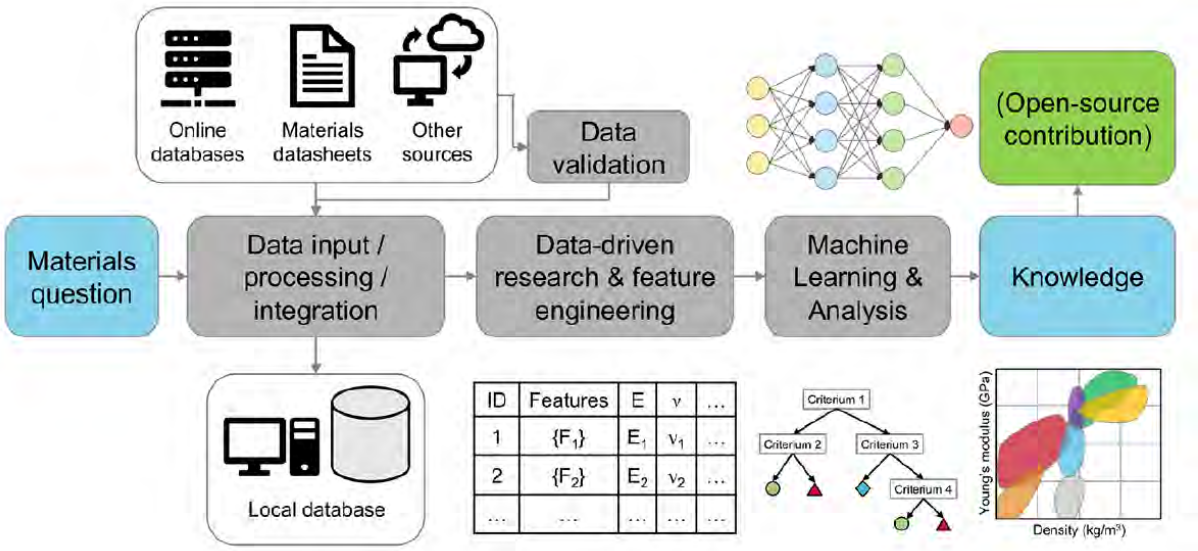


Fig. 2. Schematic of machine learning-based work in materials science [12].

Resampling is a usual method to avoid or overcome the class imbalance. It can be done in two-way, (i) oversampling the smaller represented class and (ii) under-sampling the more represented class. The resampling measures are used until the classes are balanced enough. The resampling technique demands adequate caution as the oversampling method can (a) enlarge the dataset size to increase the learning computationally expensive and (b) potentially cause overfitting. On the other hand, under-sampling can cause data reduction, including potentially important information [12–14].

Overfitting is an issue where a model perfectly fits the current dataset but fails to describe the pattern. Hence, overfitting results in poor performance while giving a prediction for novel data. Data scientists often split the dataset into training and testing parts to avoid the issue. It is done is usually two ways, (a) make test and train splitting of the whole dataset or (b) using n-fold cross-validation[15]. In the former method, a fraction of the dataset is used to train the model, and the rest is used as a test set. In the later one, the partitions are iterated over the whole dataset until every partition is used as a test set, at least once. Then, the errors are averaged from all the iterations. The iterations affect the accuracy and adaptability of the model for new inputs [16].

An alternate method of resampling is modifications of existing algorithms. Cost-sensitive learning and ensemble classifiers (multiclass classifier systems) are worth noting in this aspect. Another method used by the researchers is to assign internal bias to the learning process to avoid class imbalance. Pazzani et al. used unique weights for the different classes [17]. Barandela et al. suggested a weighted version of the distance functions for the k-nearest neighbour classification [18]. The fundamental idea is to assign a modified weight to compensate for the present class imbalance. Alternatively, one-class classifier and classifiers ensembles can be found in literature as solutions to this problem [19,20].Finally, to assess the overall classifier performance factional area under the ROC curve can be considered. The value of the area is between 0 to +1, where a value closer to +1 is more desired.

After the dataset curation process, the next step is to clean and process the dataset. The clean-up process includes removing the missing and unrealistic values, outliers, and poorly formatted values. In short, removing the irregularities properly and without introducing any bias. After the dataset processing is over, the next job is to split the data for training and test the model. One reproducible way of splitting the dataset is to assign a random seed. It is crucial to ensure that the same data should not stay in the test set if they are in the validation set or train set. The models should only be introduced to the training set during the 'training' procedure, followed by a validation process to tune hyperparameters. The test data should be used in the final evaluation step, i.e., to assess the model's performance. Furthermore, the dataset can be separated for training and testing by using cross-validation (discussed above).

3.2.Modeling

The modeling part of the ML work starts with selecting appropriate models and features. The data size can play a crucial role during the model selection process. For instance, a smaller dataset often favours statistical and classical approaches like regression, decision trees, k-nearest neighbours, etc. [21,22]. On the contrary, neural networks are more suitable for greater data (data points in thousands or more) [12]. These algorithms can be improved further by boosting, stocking, or bagging approaches. Python libraries such as sci-kit-learn (for general models), PyTorch, and TensorFlow (for neural networks) can be used here. Alternate solutions are WEKA, R, MATLAB etc [23,24]. Another important aspect of modeling is feature engineering. A well-engineered feature selection can help enhance the model's performance. A composition-based feature vector selection and Onehot-encoding are a few popular feature engineering choices. Similarly, in WEKA, information gain evaluator, ranking, etc., can be used to find the worth of each feature in a dataset.

After that, scaling is often found beneficial before proceeding further. The input data is often scaled to have unit variance, and zero mean. It helps to obtain a stable gradient, and model convergence is achieved faster. The stable and faster performance is because the dimensions of the features become similar in scale. It must be noted that train, test, and validation sets must be scaled only using the standard deviation and mean value obtained from the training dataset. Then the hyperparameters of the model should be tuned. Hyperparameters are vital for the models' performance, speed, and complexity.

The next step of modeling is evaluation and comparisons between (a) several models and (b) several combinations of hyperparameters in a single model. For evaluation, train models havecompared their performance against the test dataset with the help of various test metrics such as recall, precision, receiver operating characteristics curve, the area under the curve, etc., for a classification problem. For regression problems, mean absolute error, root mean squared error, Pearson correlation coefficient, etc., are used. Finally. To report a new model algorithm or architecture, all necessary information is to be described thoroughly. The instructions to reproduce the model and results must be shared with the prospective authors.

3.3.Fit-Test-Benchmark

Caution is required during the fitting of the models. Every ML problem is usually expected to perform two different tasks – (a) minimize the error of prediction on the 'training' dataset and (b) maximize the ability to generalization of a novel dataset (prediction-accuracy). Based on the modeling-related steps as discussed above, the outcome of ML can be either of two – (i) wanted outcome: adequate representation of the dataset patterns or (ii) unwanted outcome: memorization of training dataset or overfitting. The malice of overfitting must be avoided and discussed above. In general, overfitting occurs more frequently in complex models with high initial performance accuracy. Finally, one thing is necessary to perform a fair practice during fitting and testing – data evaluation should not be tested on the test data set until the model is fine-tuned and finalized to its optimum form. After the dataset is finalized, models are selected, optimized, fitted, and tested (prediction of newly infused data) is essential. After that, it is indispensable to make sure of the reproducibility of the result and set a benchmark to judge the result. Prediction results become much more reliable when compared to the experimental result.

4. Recent Advances in Machine Learning in Electrochemical Energy Storage

The fourth generation of research in materials science is based heavily on the 'big-data' informatics. A typical trend of this era shows (i) designing of the synthesis procedures, (ii) designing of the materials using a reverse engineering mechanism, and (iii) virtual testing and validation followed by a lifecycle assessment[25]. Overall, the focus is on building a time-efficient and economic culture of materials research ranging from fundamental conceptualization to the manufacturing industry.

Based on the above background, Yang et al. proposed a novel deep learning method to understand the correlation between process, structure, and property of materials [26]. Several traditional ML models are used, followed by a feature engineering process to establish the required correlations. As discussed above, deep learning is operable in a feature-engineering-free environment as a more advanced technique. Behera et al. investigated and observed several vital features like (i) ionic radius, (ii) valence, (iii) bond lengths, and electronegativity for the prediction of ABO_3 perovskites' crystal structure by the Light Gradient Boosting Machine (Light GBM) process [27]. In another work, Wang and co-workers employed deep learning to rapidly screen energy storage materials [28]. The models are trained using a nominal percentage (>5%) of the whole dataset of 64 different structures and followed by a prediction of the bandgaps of the rest of the dataset containing 1,439 structures. Hence, (i) a total of 75materials are found to be carrier-transport materials, (ii) around 33 materials are deemed suitable for electrode/electrocatalytic purposes, (iii) about 299 materials are screened in favour of power switching application, and finally (iv) 114 sensing materials are explored. This work demonstrates how deep learning-based models are time-efficient as the process is approximately 104 times quicker in terms of the computational speed of the *ab initio* process.

The idea of data-based accelerated screening is used in the case of novel battery materials selection [29]. In these cases, active materials, electrolytes (liquid and solid), are identified using ML. For instance, in **Fig. 3.**an unsupervised learning method is exercised. An unsupervised clustering method is utilized on several Li-based compounds. Further, the model

predicted 16 novel Li-based conducting materials with high conductivity (10^{-4} – 10^{-1} S cm^{-1}) at room temperature with the help of simulations based on ab initio-molecular dynamics [30]. A smaller amount of conductivity data was used in the unsupervised scheme to find out Li_6KBiO_6, Li_8N_2Se, and $Li_5P_2N_5$ as potential electrode materials. Besides that, ML is used to predict battery parameters such as state of charge, cyclic stability, state of health, etc. [31–33]. Meredig and co-authors studied a combinatorial screening method for predicting the thermodynamic stability of several ternary materials [34]. The authors predicted a high number of stable materials (4500) out of 1.6 million candidates. This prediction-based approach demands extensive research before it becomes a norm for material discovery and selection. In an article, Volker L Deringer mentioned that an ML-based study could potentially achieve a quantum-mechanical standard accuracy much more time-efficiently [35].

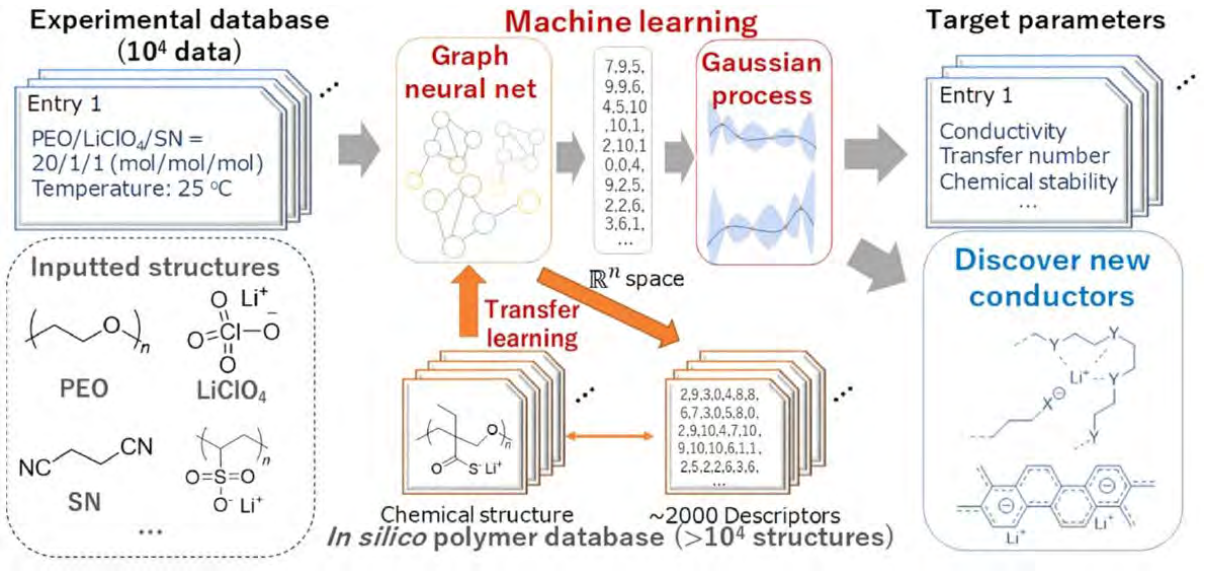


Fig. 3. Workflow of ML-based study to predict Li-based conducting materials. Reproduced with permission. Copyright (2020) American Chemical Society [30].

Apart from batteries, data-based methods have gained massive attention in supercapacitors (SCs) in recent years. Zhu et al. predicted carbon-based SCs' specific capacitance using neural networks. The authors used pore volume, specific surface area, pore size, amount of nitrogen-doping, defect of materials, and voltage window for this study [36]. Similar to the discussion above, the neural network model provided a highly accurate prediction result but could not offer a deep understanding of the features' impact on performance. Hence, it emphasizes the prediction performance and model complexity trade-off even more. Because of that, Su et al. provided an alternate solution [37]. The specific capacitance of carbon and related materials is predicted with a better understanding of the features used. The critical result of the work is to determine the specific capacitance using random tree model-based analytics. In another work, Dubey et al. evaluated the electrolytic effect on the performance of carbon-based electrodes for energy storage [38]. This is one of the instances where electrolytic performance is considered a valuable feature to determine the device's performance.

Simulation of experimental data is another aspect where ML or deep learning can be used. For example, Dongale et al. studied the cyclic voltammetry of MnO_2-based electrodes using a neural network approach [39]. The prediction result shows a low value of error (<2%) compared to the experimental specific capacitance values. In another work, the performance prediction of a novel material was carried out using ML by Ghosh et al. [40]. ML methods such as random forest and multilayer perceptron are used for performance prediction of cerium oxynitride for SC electrode. Experimental data further validate the prediction result. The experimental value of charge storage (26 mAh g^{-1}) is reasonably close to the ML-based prediction of~26.6 mAh g^{-1}.The schematic of the work is provided in **Fig.4.** Hence, the literature observes an emphasis on experimental validation for benchmarking purposes. Another similar work is executed by Ding et al. for hydrogen storage. The authors studied $LiBH_4$-based materials to predict their hydrogen release ability using ensemble ML methods [41].

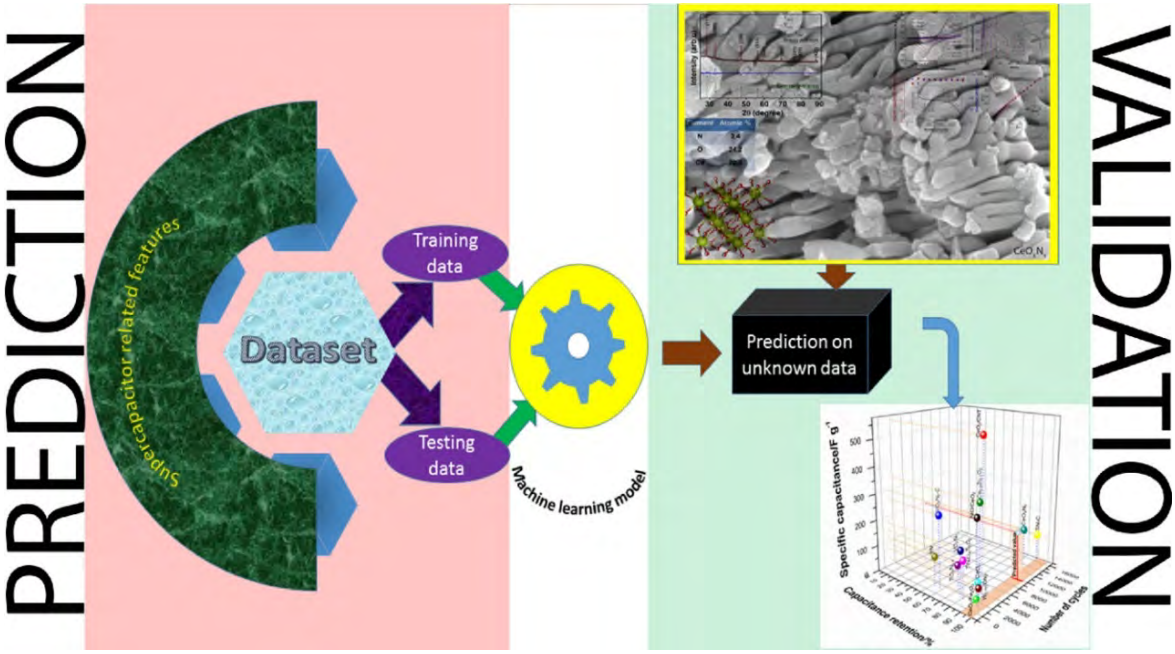


Fig.4. Scheme of performance prediction of novel materials. Copyright (2021) Elsevier publication [40].

Overall, material science – a truly interdisciplinary field, went through a considerable change in recent times from a heavy experimental-oriented research methodology to computation-centered predictive designing-based research. Nonetheless, the data-driven methods are still in their infancy and require careful assessment and a robust experimental background with profound physical and chemical insights.

5. Conclusion

ML-based predictions are inherently statistical, and uncertainty is a crucial aspect. The nature of the predictions is often interpolative and deduced from the previously observed data. In literature, several models, from classical models to modern complex neural network-based models, are explored for material science application. However, various challenges related to their proper materials informatics utilization still exist. A careful methodology has to be followed from the stage of dataset creation to the evaluation of the models to ensure a robust prediction model. Drawbacks like class imbalance, overfitting, etc., need to be checked and reduced (if possible, removed) with great attention. Data-driven studies are a statistical approach for understanding of several physicochemical insights. Applications of ML in electrochemical energy storage (Li-ion batteries and supercapacitors) are discussed briefly. It is understood that in the case of electrochemical energy storage devices, ML is used chiefly in case of prediction of (i) device health, (ii) performance, and (iii) novel material discovery. Finally, experimental validation and benchmarking are advised to increase the reliability of the study. Overall, recent attention is welcome in this field, but at the same time, cautions need to be undertaken to ensure more reproducible science in this field.

References

[1] C. Chen, Y. Zuo, W. Ye, X. Li, S.P. Ong, Learning properties of ordered and disordered materials from multi-fidelity data, Nature Computational Science. 1 (2021) 46–53. https://doi.org/10.1038/s43588-020-00002-x.

[2] J.F. Rodrigues, L. Florea, M.C.F. de Oliveira, D. Diamond, O.N. Oliveira, Big data and machine learning for materials science, Discover Materials. 1 (2021) 12. https://doi.org/10.1007/s43939-021-00012-0.

[3] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, Machine learning in materials informatics: recent applications and prospects, Npj Computational Materials. 3 (2017) 54. https://doi.org/10.1038/s41524-017-0056-5.

[4] H. Wang, C. Ma, L. Zhou, A Brief Review of Machine Learning and Its Application, in: 2009 International Conference on Information Engineering and Computer Science, 2009: pp. 1–4. https://doi.org/10.1109/ICIECS.2009.5362936.

[5] K. de Jong, Learning with genetic algorithms: An overview, Machine Learning. 3 (1988) 121–138. https://doi.org/10.1007/BF00113894.

[6] S. Sajid, A. Haleem, S. Bahl, M. Javaid, T. Goyal, M. Mittal, Data science applications for predictive maintenance and materials science in context to Industry 4.0, Materials Today: Proceedings. 45 (2021) 4898–4905. https://doi.org/https://doi.org/10.1016/j.matpr.2021.01.357.

[7] G.T. Reddy, M.P.K. Reddy, K. Lakshmanna, R. Kaluri, D.S. Rajput, G. Srivastava, T. Baker, Analysis of Dimensionality Reduction Techniques on Big Data, IEEE Access. 8 (2020) 54776–54788. https://doi.org/10.1109/ACCESS.2020.2980942.

[8] R. Tibshirani, Regression Shrinkage and Selection Via the Lasso, Journal of the Royal Statistical Society: Series B (Methodological). 58 (1996) 267–288. https://doi.org/https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[9] K. Pearson, LIII. On lines and planes of closest fit to systems of points in space, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 2 (1901) 559–572. https://doi.org/10.1080/14786440109462720.

[10] A. Glielmo, B.E. Husic, A. Rodriguez, C. Clementi, F. Noé, A. Laio, Unsupervised Learning Methods for Molecular Simulation Data, Chemical Reviews. 121 (2021) 9722–9758. https://doi.org/10.1021/acs.chemrev.0c01195.

[11] J.A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller, A. Tkatchenko, Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems, Chemical Reviews. 121 (2021) 9816–9872. https://doi.org/10.1021/acs.chemrev.1c00107.

[12] A.Y.-T. Wang, R.J. Murdock, S.K. Kauwe, A.O. Oliynyk, A. Gurlo, J. Brgoch, K.A. Persson, T.D. Sparks, Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices, Chemistry of Materials. 32 (2020) 4954–4965. https://doi.org/10.1021/acs.chemmater.0c01907.

[13] N. Wagner, J.M. Rondinelli, Theory-guided machine learning in materials science, Frontiers in Materials. 3 (2016). https://doi.org/10.3389/fmats.2016.00028.

[14] J.S. Sánchez, R. Alejo, V. García, J.S. Sánchez, R.A. Mollineda, R. Alejo, J.M. Sotoca, The class imbalance problem in pattern classification and learning, 2007. https://www.researchgate.net/publication/230582747.

[15] M. Stone, Cross-Validatory Choice and Assessment of Statistical Predictions, Journal of the Royal Statistical Society: Series B (Methodological). 36 (1974) 111–133. https://doi.org/https://doi.org/10.1111/j.2517-6161.1974.tb00994.x.

[16] H.-J. Lu, N. Zou, R. Jacobs, B. Afflerbach, X.-G. Lu, D. Morgan, Error assessment and optimal cross-validation approaches in machine learning applied to impurity diffusion, Computational Materials Science. 169 (2019) 109075. https://doi.org/https://doi.org/10.1016/j.commatsci.2019.06.010.

[17] M.J. Pazzani, C.J. Merz, P.M. Murphy, K.M. Ali, T. Hume, C. Brunk, Reducing Misclassification Costs, in: Proceedings of the Eleventh International Conference on International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1994: pp. 217–225.

[18] R. Barandela, J.S. Sánchez, V. García, E. Rangel, Strategies for learning in class imbalance problems, Pattern Recognition. 36 (2003) 849–851. https://doi.org/https://doi.org/10.1016/S0031-3203(02)00257-1.

[19] P.K.-F. Chan, S. Stolfo, Toward Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection, in: KDD, 1998.

[20] B. Raskutti, A. Kowalczyk, Extreme Re-Balancing for SVMs: A Case Study, SIGKDD Explor. Newsl. 6 (2004) 60–69. https://doi.org/10.1145/1007730.1007739.

[21] A. Agrawal, A. Choudhary, Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science, APL Materials. 4 (2016) 053208. https://doi.org/10.1063/1.4946894.

[22] M.-L. Zhang, Z.-H. Zhou, ML-KNN: A lazy learning approach to multi-label learning, Pattern Recognition. 40 (2007) 2038–2048. https://doi.org/https://doi.org/10.1016/j.patcog.2006.12.019.

[23] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA Data Mining Software: An Update, n.d.

[24] G. van Rossum, Python reference manual, (1995).

[25] R. Jose, S. Ramakrishna, Materials 4.0: Materials big data enabled materials discovery, Applied Materials Today. 10 (2018) 127–132. https://doi.org/https://doi.org/10.1016/j.apmt.2017.12.015.

[26] Z. Yang, Y.C. Yabansu, R. Al-Bahrani, W. Liao, A.N. Choudhary, S.R. Kalidindi, A. Agrawal, Deep learning approaches for mining structure-property linkages in high contrast composites from simulation datasets, Computational Materials Science. 151 (2018) 278–287. https://doi.org/https://doi.org/10.1016/j.commatsci.2018.05.014.

[27] S. Behara, T. Poonawala, T. Thomas, Crystal structure classification in ABO3 perovskites via machine learning, Computational Materials Science. 188 (2021) 110191. https://doi.org/https://doi.org/10.1016/j.commatsci.2020.110191.

[28] Z. Wang, Q. Wang, Y. Han, Y. Ma, H. Zhao, A. Nowak, J. Li, Deep learning for ultra-fast and high precision screening of energy materials, Energy Storage Materials. 39 (2021) 45–53. https://doi.org/https://doi.org/10.1016/j.ensm.2021.04.006.

[29] Z.-H. Shen, H.-X. Liu, Y. Shen, J.-M. Hu, L.-Q. Chen, C.-W. Nan, Machine learning in energy storage materials, Interdisciplinary Materials. n/a (2022). https://doi.org/https://doi.org/10.1002/idm2.12020.

[30] K. Hatakeyama-Sato, T. Tezuka, M. Umeki, K. Oyaizu, AI-Assisted Exploration of Superionic Glass-Type Li+ Conductors with Aromatic Structures, J Am Chem Soc. 142 (2020) 3301–3305. https://doi.org/10.1021/jacs.9b11442.

[31] M.-F. Ng, J. Zhao, Q. Yan, G.J. Conduit, Z.W. Seh, Predicting the state of charge and health of batteries using data-driven machine learning, Nature Machine Intelligence. 2 (2020) 161–170. https://doi.org/10.1038/s42256-020-0156-7.

[32] D. Roman, S. Saxena, V. Robu, M. Pecht, D. Flynn, Machine learning pipeline for battery state-of-health estimation, Nature Machine Intelligence. 3 (2021) 447–456. https://doi.org/10.1038/s42256-021-00312-3.

[33] S. Stock, S. Pohlmann, F.J. Günter, L. Hille, J. Hagemeister, G. Reinhart, Early Quality Classification and Prediction of Battery Cycle Life in Production Using Machine Learning, Journal of Energy Storage. 50 (2022) 104144. https://doi.org/https://doi.org/10.1016/j.est.2022.104144.

[34] B. Meredig, A. Agrawal, S. Kirklin, J.E. Saal, J.W. Doak, A. Thompson, K. Zhang, A. Choudhary, C. Wolverton, Combinatorial screening for new materials in unconstrained composition space with machine learning, Physical Review B. 89 (2014) 94104. https://doi.org/10.1103/PhysRevB.89.094104.

[35] V.L. Deringer, Modelling and understanding battery materials with machine-learning-driven atomistic simulations, Journal of Physics: Energy. 2 (2020) 041003. https://doi.org/10.1088/2515-7655/abb011.

[36] S. Zhu, J. Li, L. Ma, C. He, E. Liu, F. He, C. Shi, N. Zhao, Artificial neural network enabled capacitance prediction for carbon-based supercapacitors, Materials Letters. 233 (2018) 294–297. https://doi.org/https://doi.org/10.1016/j.matlet.2018.09.028.

[37] H. Su, S. Lin, S. Deng, C. Lian, Y. Shang, H. Liu, Predicting the capacitance of carbon-based electric double layer capacitors by machine learning, Nanoscale Advances. 1 (2019) 2162–2166. https://doi.org/10.1039/C9NA00105K.

[38] R. Dubey, V. Guruviah, A data-driven approach for evaluation of electrolyte informatics on electrochemical performance of carbon-based electrode materials, Ionics (Kiel). 28 (2022) 2169–2183. https://doi.org/10.1007/s11581-022-04480-z.

[39] T.D. Dongale, P.R. Jadhav, G.J. Navathe, J.H. Kim, M.M. Karanjkar, P.S. Patil, Development of nano fiber MnO2 thin film electrode and cyclic voltammetry behavior modeling using artificial neural network for supercapacitor application, Materials Science in Semiconductor Processing. 36 (2015) 43–48. https://doi.org/https://doi.org/10.1016/j.mssp.2015.02.084.

[40] S. Ghosh, G.Ranga Rao, T. Thomas, Machine learning-based prediction of supercapacitor performance for a novel electrode material: Cerium oxynitride, Energy Storage Materials. 40 (2021) 426–438. https://doi.org/https://doi.org/10.1016/j.ensm.2021.05.024.

[41] Z. Ding, Z. Chen, T. Ma, C.-T. Lu, W. Ma, L. Shaw, Predicting the hydrogen release ability of LiBH4-based mixtures by ensemble machine learning, Energy Storage Materials. 27 (2020) 466–477. https://doi.org/https://doi.org/10.1016/j.ensm.2019.12.010.

Thermoelectric Device: Recent Status and Applications in Biomedical Instrument

S. Sau[1], A. Jana[2], S. Mahakal[2], Diptasikha Das[3] and K. Malik[2,*]

[1]Department of Physics, Bangabasi College, Kolkata, India
[2]Department of Physics, Vidyasagar Metropolitan College, Kolkata, India
[3]Department of Physics, ADAMAS University, Kolkata, India

[*]Corresponding author: kartickmalik@vec.ac.in

Abstract

Thermoelectric device (TED) converts thermal to electrical energy or vice versa. Worldwide research in thermoelectric (TE) are become attractive for conversion of waste heat into electricity. Efficiency of TED are directly related with the physical properties of material viz., seebeck coefficient (S), resistivity (ρ) and thermal conductivity (κ). Optimization of these interrelated material properties may only give highest TED efficiency. TE materials are special type of semiconductor with narrow band gap. Efficiency of TED depends on material properties as well as external parameters. TED is prepared by interconnection of n- and p-type TE material. Devising of TED required dedicated techniques to reduce mismatch of Thermal expansion coefficient (CTE),TE Enhance the carrier conductivity through electrode to reduce Joule heating etc. Choice of inter layer material is another important task to prepare TED. Inter layer material hinders diffusion of TE material and electrode material particle. Diffusion or formation composite between electrode and TE material may reduce the efficiency of TED. Theoretical modelling of device considering external and internal parameters is very important to obtain high efficiency and viability of application in definite place. Use of TED has been started in various field viz., space mission where utility is more important than cost. TED may play crucial role in Biomedical application also. Thermal energy harvesting from the body may extend the use of biomedical wearable and implantable devices beyond lifetime of batteries. Embedding a TE device rather than an electronic battery in a biological body is a promising way to supply power in the long term to a medical device. It may resolve service life mismatch between the implantable medical device and its battery.

Keywords: *Efficiency of Thermoelectric Device, Figure of Merit, Thermoelectric Material and Device, Biomedical Device, Wearable Medical Device, Implantable Medical Device.*

1 Introduction

Transformation of energy is strongly associated with the human civilization. Since the dawn of the industrial age, the ability to harness and use different forms of energy has transformed living conditions for billions of people. It enables the human race to enjoy a level of comfort as well as perform productive tasks. However, humanity now finds itself confronting an enormous green and sustainable energy challenge. The overwhelming reliance on fossil fuels become threats to Earth's climate to an extent may have grave consequences for integrity of biological ecosystem. Consumption of fossil fuel should be reduced for sustainable development and grow human society in harmony with nature. However, it may only be achieved through technological development and use of diversified renewable energy viz., wind, solar and hydropower etc. Thermoelectricity is currently emerging as a promising alternative energy source amid other alternative energy sources. Characteristic of TE technology are no moving parts, environment friendly, require almost no maintenance and directly converts heat into electricity. Hence, TE energy conversion may become promising method to solve environmental pollution and energy problems effiiently [1]. TE technology is usually applied in the form of TED. However, performance of TED is limited by internal and external parameters. The continuous efforts to improve the efficiency of TED may be categorized in the following way: (i) optimization of interrelated material properties; (ii) Modification of the structure/geometry of the thermo elements and (iii) the improvement of the thermal and electrical energy management [2, 3, 4].

Biomedical devices are currently receiving considerable attention with advancement of microelectronics. These are potential to use in real-time health monitoring during ongoing assessments of personal health [5]. However, long time power harvesting or generation is still a main challenge in such devices. Limited battery life time and the risk of battery leaking toxic substances are crucial problem in recent mode of power generation for biomedical devices. A big drawback of life-saving medical implants viz., pacemakers and defibrillators are run out of batteries after certain time. It require continuous monitoring and evolution. However, patients require frequent surgery to replace these batteries [6]. This kind of immediate surgeries may be avoided by employing body energy harvesting techniques to supply power to the life saving

devices. Power sources within the body are blood flow inside the vessels, patient's heartbeat, movement of the body parts and the body temperature (heat). TED is one of such environmental friendly device with excellent stability and efficient for energy harvesting using body heat [7, 8].

This review is dedicated to discuss the configuration, fabrication and medical application of TEDs. Further, effective factors to improve performance of device are also discussed. A brief background and the basic criteria of TE material and device are covered. It provides a theoretical basis for the enhancement of TE performance. Application of TEDs on the wearable and implantable devices has been discussed in details. It is envisioned that the study will provide profound knowledge on advancement of TEDs and specific medical applications, which will be helpful for future endeavors. At the end of the review, the technical obstacles to improve efficiency of TEDs are summarized. Further, future research works are also prospected.

2 Thermoelectric Material and Device: Basic Criteria

In general, electrical current and heat flux are coupled through the phenomena of electron and phonon transport within conductor and semiconductors. Electrical current and heat flux are coupled through the equation [9].

$$J = \sigma E - \sigma S \nabla T (\frac{A}{m^2}) \tag{1}$$

$$q = \pi J - \kappa \nabla T (\frac{W}{m^2}) \tag{2}$$

Where E and T are the electric field and temperature in the material respectively. And σ, κ, S, π are electrical conductivity, thermal conductivity, Seebeck coefficient and Peltier coefficient respectively. Following conclusion may be drawn from equations (1) and (2) for electrical and heat transport through material: (i) A electric field in absence of electrical current may be generated owing to temperature gradient in the material and (ii) an electric field causes a thermal gradient in absence of thermal current.

There are three well- known major effects involved in the TE phenomena: the Seebeck, Peltier and Thomson effects. In 1821 Thomas Seebeck, a German physicist discovered that voltage may be generated in open circuit of a thermocouple if two junctions are kept at different temperatures. The effect may be considered as conversion of thermal to electrical energy. Peltier effect was discovered by a French watch maker Jean Peltier in 1834. It may be considered as reverse Seebeck Effect. Absorption or liberation of heat is observed during current flow through a thermocouple and the effect is known as Peltier effect.

However, heat is given out or absorbed depend on the pairs of metals and the direction of the current. William Thomson discovered a third TE effect. It provides a inter-relation between Seebeck effect and Peltier effect. Thomson found that when a current is passed through a wire of single homogeneous material in presence of temperature gradient, evolution or absorption of heat occurs within material other than the Joule heat.

TEDs are created by connecting a p-type and n-type TE legs electrically in series and thermally in parallel. TED may be built for power generation (Figure 1(a)) or cooling system (Figure 1(b)) based on the TE effects. Maximum heat-to-power conversion efficiency (η_{max}) for an ideal TED with temperature independent TE properties is:

$$\eta_{max} = \frac{T_H - T_C}{T_H} \frac{\sqrt{1 + Z_d T_m} - 1}{\sqrt{1 + Z_d T_m} + \frac{T_C}{T_H}} \tag{3}$$

and the cooling efficiency of a TE cooling device is characterized by the coefficient of performance (COP_{max}):

$$COP_{max} = \frac{T_C}{T_H - T_C} \frac{\sqrt{1 + Z_d T_m} + \frac{T_H}{T_C}}{\sqrt{1 + Z_d T_m} + 1} \tag{4}$$

Where, T_H,T_C and $T_m = \frac{T_H + T_C}{2}$ are hot side, cold side and average respectively. And Z_d is the device figure of merit. Here device signifies that TE device consisting of p-type and n-type TE material leg. Zd is defined as [10].

$$Z_d = \frac{(S_p + |S_n|)^2}{(\sqrt{\kappa_n \rho_n} + \sqrt{\kappa_p \rho_p})^2} \tag{5}$$

Where S_p,ρ_p,κ_p,S_n,ρ_n and κ_n are the Seebeck coefficient, electrical resistivity and thermal conductivity of the p-type and n-type legs of the TE module respectively. It is evident that a higher Z_d corresponds to a better performance in power generation and cooling. However, in general research focuses on improving a TE properties of a single material at a time. For a single material, TE figure of merit is defined as:

$$ZT = \frac{\sigma S^2 T}{\kappa_e + \kappa_l} \tag{6}$$

Where the subscripts e and l in k refer to electronic and lattice contributions of thermal conductivity. ZT is known as Figure of Merit. Hence, efficiency of TED will be maximum for highest ZT. Worldwide there is resurgence to enhance ZT of suitable TE materials in industrial and biomedical applications.

Commercially viable TEDs are limited by the efficiency. Effciency of a TED strongly depends on intrinsic factors, inter-related materials properties and extrinsic factors viz., heat loss or contact resistance, selection of electrodes, geometrical size and shape of TEDs etc. It is of equal importance to consider both intrinsic and extrinsic factors to enhance the efficiency of a TED module. Hence all the factors should be included during theoretical modelling and preparation of TED module.

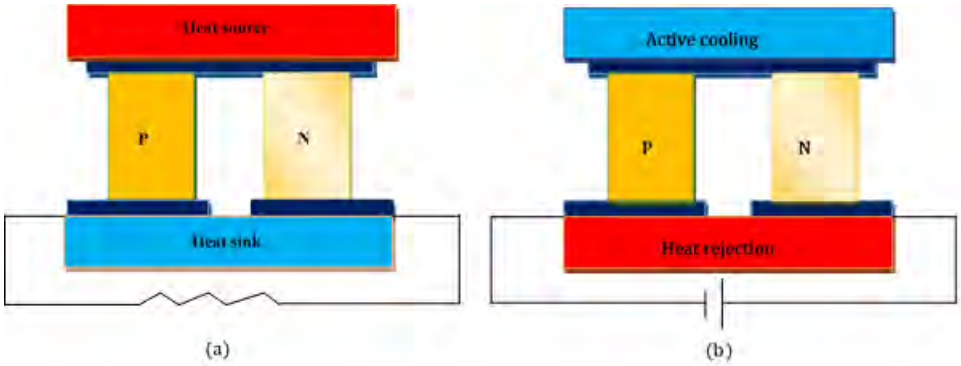


Figure 1: Illustration of TE Devices for (a) Power Generation and (b) Cooling

Highest ZT of TE material is obtained only through optimization of interrelated material properties; S, ρ and κ. It is evident from Eq. (6) that increasing σ, S and decreasing κ simultaneously may only enhance ZT. It is noteworthy to mention that these are interrelated material property and optimization may only give highest ZT. S in a degenerate semiconductor with parabolic band may be written as [11],

$$S = \frac{8\pi^2 {k_b}^2 m^* T}{2eh^2} \left(\frac{\pi}{3n}\right)^{\frac{2}{3}} \tag{7}$$

Where, $m^*$is the effective mass and n is the carrier concentration. According to this equation, S is increased with decreasing carrier concentration and increasing effective mass. Further, σ is related with m^* through Drude expression[12]:

$$\sigma = \frac{ne^2\tau}{m^*} \tag{8}$$

The other tunable parameter except S and σ to enhance ZT is κ. It is the sum of lattice and electronic thermal conductivities. Further, σ and κ_e are related through the Wiedmann-Franz law [13]:

$$\kappa_e = L_0 \sigma T \tag{9}$$

κ_l (due to phonons) is larger than κ_e for a typical semiconductor. Hence, by minimizing κ_l overall ZT may be increased with out hindering electron transport too much.

According to Figure 2, semiconductors have highest ZT due to optimization of S, ρ and κ. Hence, good TE materials should have low κ (property of glass), high σ (property of metal) and large S (property of semiconductor). However, summary of temperature dependent ZT for various n-type and p-type state of the art TE materials are plotted in Figure 3.

Choice of TE materials is one of primary steps to obtain maximum efficiency. However, there are also various parameters which control efficiency viz., selection of electrodes, inter-diffusion material, contact resistance and geometrical size and shape of TE materials etc. It is noteworthy to mention that electrode and inter layer materials not only influence TE power out put but also related with reliability of the TED.

In order to obtain best performance for long time electrode and interlayer material(IM) of TED module should have excellent matching between coefficient of thermal expansion (CTE) with TE materials. Mismatch of CTE may generate cracks and stress in inter-layered region. It strongly affects electrical and thermal transport of TED. IM should posses high σ and κ and low contact resistance (electrical and thermal) at the interfaces to achieve optimized efficiency along

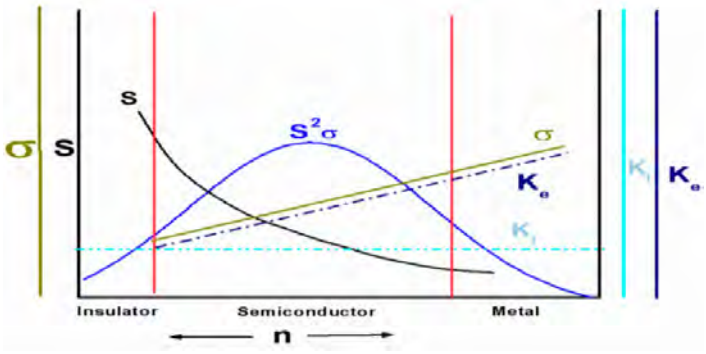


Figure 2: S, σ, $S^2\sigma$, and κ_e and κ_l as a function of carrier concentration (n) [14]

.

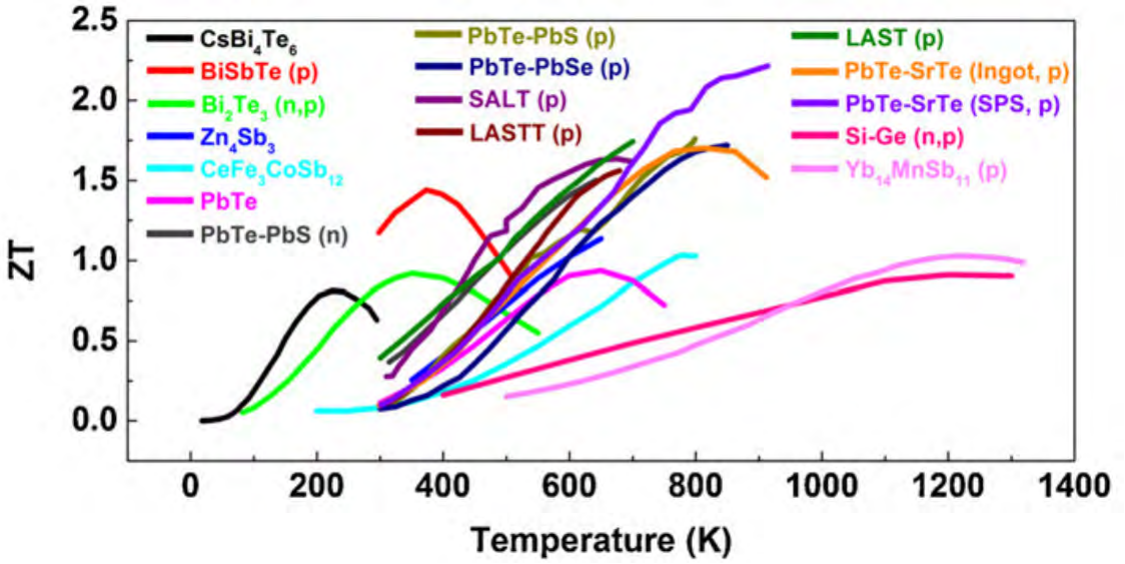


Figure 3: TE figure of merit (ZT) verses temperature in some recent bulk TE materials [15].

with reliability. Another important criteria for best is lack of inter-diffusion and avoid chemical reaction in the operating temperature range of the TED. Cu, Al, Ag, Mo, Ni and their alloys are well known electrode for TED module [16].

There are two contact regions for interlayer material, interlayer material-electrode and interlayer material-TE materials. Contact resistance may be categorized broadly in two ways, thermal contact resistance (TCR) and electrical contact resistance (ECR). In general, TE performance is better for low TCR [17]. However, better performance is observed for thermoelectrically nontrivial contact layer with high TCR [18]. TRC causes temperature difference at contact region which further led to evolve Seebeck effect and reduces overall S of TE leg of TED [17]. However, there is no experimental evidence of this concept. It may be concluded TED require low TCR in general. ECR evolves owing to contact between electrode (metal)/contact layer and TE materials (semiconductor)/contact layer. It is noteworthy to mention that the junctions may be rectifying (Schottky) or non-rectifying (Ohmic) behavior. It solely related with the work function of the interacting materials[19]. Schottky junction will be developed if work function of metal (ϕ_M) is greater than work function of semiconductor(ϕ_S). Hence, junction may be considered as diode. And the junction is Ohmic if $\phi_M < \phi_S$ and and junction act as a resistor i.e. non-rectifying.

Size and shape of TED are important for convenient use in various purpose. Further, geometrical shape and size of TE leg also influence the performance of TED. Conversion efficiency of TED module may be enhanced by operating at the larger temperature difference and increasing length of TE leg of TED . But shorter thermoelements are required for better power output [2]. Hence, length of TE material is usually optimised for maximum power output and conversion efficiency [1]. However, it is also observed that power is also corroborated with cross-sectional area of the TE legs of TED [2]. There are several studies on shape and size dependent conversion efficiency and power output of thermoelements. TED with triangular leg is superior for power out-put compared to square /cylindrical legs. Size and shape of base also strongly affect out-put power[4]. Further, hollow geometries result in higher output electrical power than the filled geometries. It is also reported that layered geometries show the highest benefit in terms of output electrical power [4].

3 Configuration of Thermoelectric Device

TE technology is usually applied in the form of TED. The standard structure of a conventional TED follows the π-shaped configuration. It is also known as 'flat bulk TE device', where the electrical current and thermal current are parallel to each other [Figure 4(a)] [20, 21]. It consists of a bunch of the column-like or cube like p- and n-type TE materials. TE materials are connected in electrically series and thermally parallel through electrode. TED module is sandwiched between two polymer or ceramic plates which serve as both electrical insulators and thermal conductors. The ceramics are commonly made from alumina (Al_2O_3). however, high κ like beryllia or aluminum nitride (AIN) is required for large lateral heat transfer[17]. However, TED module converts heat energy to electrical energy. Hence, heat energy should be captured efficiently. And basic target is tight-fitting contact between heat source and heat sink with TED module. But, it is very difficult to properly attached of π-shaped TED with heat source and sink like oil pipelines, cooling channels for power station transformers etc. In these case, the tube-shaped module becomes a better option specially, when the diameter of the cylindrical heat source is less than 1 cm. The tube-shaped TE module has been designed by the coaxial arrangements of ring-shaped TE materials which is shown in Figure 4(b) [23]. These ring-shaped TE materials are connected in electrically series and joined alternatively at their inner and outer perimeters through interleaved ring-shaped electrodes. Both inner surface and outer surface of the modules are covered with ceramics. It acts as electrically insulating and thermally conducting material. Temperature gradient is established during heat flow through outer surface or inner surface and concomitantly electrical power is generated.

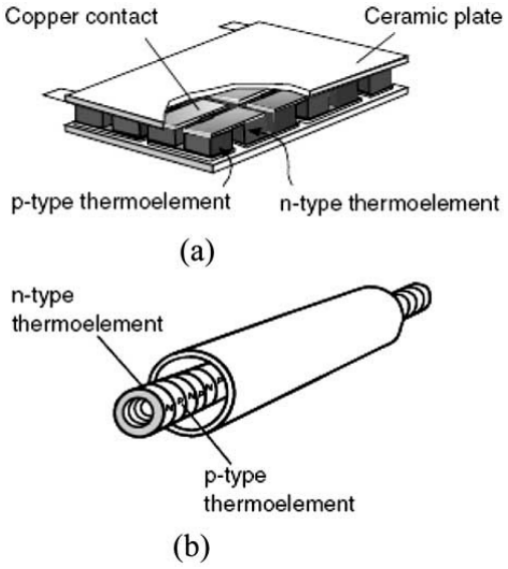


Figure 4: Schematic diagram of TE modules, (a) π-shaped TE module, (b) Tube shaped TE module [22].

Micro-thermoelectric generators (μTEGs) and micro-thermoelectric coolers (μTECs) are fabricated for applications in spot cooling or cooling in power electronics [24, 25, 26, 27]. Thickness of these type of TE modules are in the micrometer to sub-millimeter range. It is designed for high densities of heat flux. Cooling power is attainable by proper design of heat sink and substrate. μTECs can be operated in both steady-state and transient mode. In transient mode, it is provided higher cooling power [28, 29].

4 Fabrication of Thermoelectric Device

In general a typical π-shape TE device module is shown in Figure 5. It consists by one p-type and one n-type semiconductor legs and electrodes. After that it is inserted between two ceramic plates. The uni-couple TE legs are covered with a protective coating to minimize the degradation of TE materials at the high operation temperature. An inter layer or a barrier layer material is introduced to prevent inter-diffusion between electrode and TE materials. It also relaxes the stress at the joint. There are several methods are available to connect TE materials with electrode. Also, various methods have been investigated for barrier layers between electrode and TE legs.

4.1 Soldering Method

Due to the cost-effective, mature and industry-scalable manufacturing process, the soldering method has been widely used to prepare various type of TED modules. In this process selection of soldering material is very important for reliable joining of TE materials with electrodes. The soldering material should have excellent matching of CTE with the TE material and

electrode. Further, soldering material should have high electrical and κ and also have low electrical and thermal contact resistance at the interface. At the operating temperature, soldering material can react with TE materials at the hot side and form a very thick intermetallic compounds (IMCs) at junction. It may degrade the electrical and mechanical properties of TE module. Therefore, IMis needed to prevent reaction between soldering materials and TE materials and also inter diffusion of the soldering material into the TE materials. Several groups have been studied on IMbetween electrode and TE leg. Additionally, High thermal stability and low thermal stresses at the interface are very important for the long-term use of a TED [16].

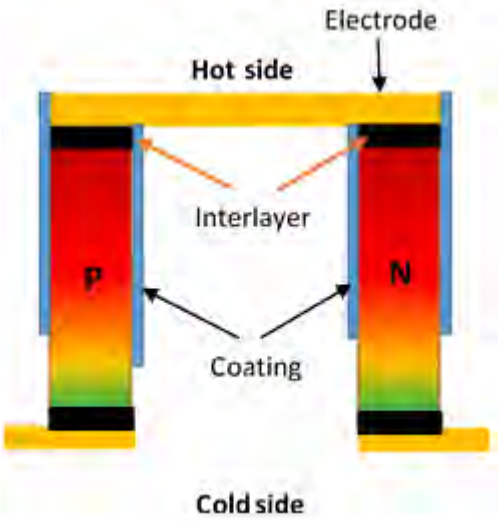


Figure 5: Schematic diagram of a typical π-shaped TE uni-couple.

There are several studies concerning the preparation and reliability of TE joints using soldering method. Sn, Sn-Bi, Sn-Pb, Sn-Ag and Sn-Ag-Cu etc. are used as a soldering material [30]. Due to the high temperature, soldering material can react with TE materials at the hot side and form a very thick intermetallic compounds (IMCs) at junction. It may degrade the electrical and mechanical properties of TE module [31, 32]. To prevent inter-diffusion of soldering material into the TE leg and also reaction between these, Mengali et al. placed a nickel layer between the for bismuth telluride based TE material and Sn soldering material [33]. However, under high temperature and high electric current operation conditions, Sn-based solder will consume Ni to form Ni3Sn4 and penetrate through the barrier layer to react with telluride of BT-based TE material. Therefore, nickel is not suitable for a diffusion barrier layer at the hot side. Lin et al. investigated palladium (Pd), nickel/gold (Ni/ Au), silver (Ag), and titanium/gold (Ti/Au) as diffusion barrier layers, finding that titanium/gold formed the best diffusion barrier for BT-based TE elements among the four candidates [34]. Also there have difficulty, after aging at 250°C for 50 hours a crack occurs between the titanium and gold layer. Apparently, there is still no appropriate diffusion barrier layer at the junction between BT-based TE elements and soldering material for power generation of a TED. The electrical, thermal, and mechanical properties of the IMshould be known for the reliability of TE modules. Moreover, for a high-quality soldering surface, the specimens were polished to remove surface oxide, if any, and then cleaned consecutively with acetone, isopropyl alcohol, and deionized water in an ultrasonic bath. Apart from the above mentioned, Mo, Ti, Co-Fe-Ni etc. also can be used as an interlayer material. Which depends on the types of TE materials and electrode.

4.2 Thermal Spraying Method

Thermal spraying represent a potential efficient technology to fabricate high temperature TEDs.

R.Puschmann et al. have successfully presented various thermal spray processes like Atmospheric Plasma Spray (APS), High Velocity Oxy-Fuel (HVOF) spray, High Velocity Air-Fuel (HVAF) spray, to develop a manufacturing technology for TED [35]. Zhang et al. in Figure 6, showing fabrication of $Bi_2Te_3/Mo/Al$ TED by arc spraying method. In this process, at first a desire frame has fabricated by using polymer or ceramic and then TE materials are inserted into this frame [16]. After that, Mo and Al layered will be created on the material by the Thermal gun (i.e. Arc spray). Here Mo use for interface material and Al use for electrode. Compared with soldering method, thermal spraying process is more simple, efficient and also has good scalability.

4.3 Diffusion Welding Method

Diffusion welding method is the another possible method to fabricate TE device, which is also known as “one step sintering” method [16]. In this method powder of the electrode material, the barrier layer material and TE materials are put into a graphite die, and additional sheet is used to separate p- and n-legs for uni-couple fabrication.The die assembly is then heated by employing a hot press or by using a spark plasma sintering method shown in Figure 7.

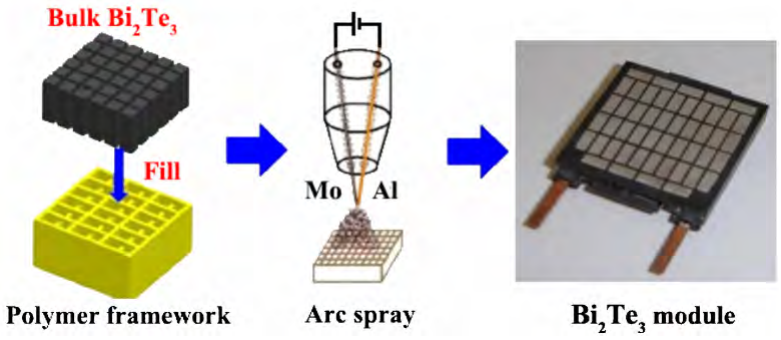


Figure 6: Schematic diagram showing fabrication of $Bi_2Te_3/Mo/Al$ TE device by arc spraying [16].

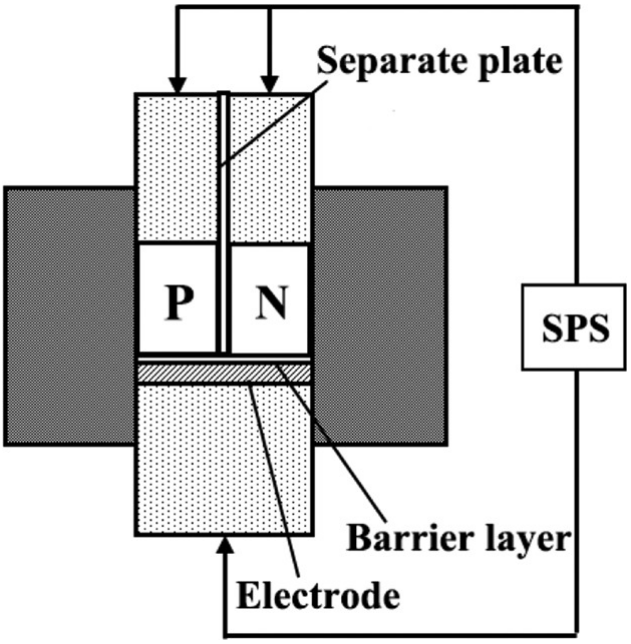


Figure 7: Schematic diagram of the one-step sintering of uni-couples TE material [16].

5 Advancement on Thermoelectric Device

All over the world many researchers are involving to developed a highly efficient TED and also used in commercial purpose. Already many TEDs have been developed in laboratory and many of them also used in different places. Kaibe et al. 2005 have been developed a Cascade type TED by using p- type Sb_2Te_3/higher manganese silicide And n-type $Bi_2Te_3 - Bi_2Se_3/Mg_2Si$ TE material. He reported that this device gives 12.1 % of efficiency at 520°C temperature difference [36]. Zhang et al. 2017 reported a TED which is achieved a recorded high efficiency of up to 12% under a temperature difference of 541°C. The device is configured Segment type by joining of skutterudite with Bi_2Te_3 for n-type material and Sb_2Te_3 for p-type material [37]. There are few recently developed TEDs are listed in table 1, with their maximum efficiency (η_{max}) and power output (P_{output}). There have lots of theoretical model to improve the efficiency and output power of a TED.

Kim at al. 2015 derived maximum efficiency formulae based on a cumulative temperature dependent model including Thomson effect [38]. The effect of cumulative Joule and Thomson heat enables the formulae to predict the efficiency and output power more reliably than the conventional model. They defined an engineering figure of merit (ZT)eng and an engineering power factor (PF)eng as a direct indicators of practical efficiency and output power.

Bjork 2016 designed an analytical model, capable of calculating the efficiency of a TE generator including both electrical and thermal contact resistances [39]. In order to determine the validity of the developed model, they have computed the efficiency of 16 individual TE legs of different materials. The model presented here was shown to accurately calculate the efficiency for all system and all contact resistance considered, with a global average difference between the analytical and the numerical model of -0.07 ± 0.35 pp. This makes the model more accurate than previously published model.

Lamba and Kaushik, 2016 developed a thermodynamic model based on first and second law of thermodynamics for an exoreversible trapezoidal TE generator including influence of Thomson effect as well as influence of leg geometry on the performance of the device [40]. This study will be helpful in the designing of the practical trapezoidal TE generator system with improved energy and energy efficiency. The result of this study shows that when the shape parameter is increased from flat plate to trapezoidal (Figure 8) TE generator, then the energy and energy efficiency improve by 2.3 % and 2.31 %

Table 1: Recently Developed TEDs with Their Maximum Efficiency and Power Output.

Thermoelectric Modules	**Temperature difference(∇T)**	P_{output} / η_{max}	**Referance**
Low, Medium and High temperature TE modules	603 K	225 W	[41]
Bi_2Te_3	445 K	608.85W	[42]
Bi_2Te_3	Th = 573 K	13.08 W	[43]
p-type material is Ni doped with Mo, and the n-type one is SrTiO3 substituted by La. Tubular module $Bi_{0.5}Sb_{1.5}Te_3/Ni$	10K (91K)	100 μW/cm^2 (8.2 W/pipe)	[44]
A full Heusler materials are Fe-V-Al for both p- and n-types	280 K	0.25 W/cm^2	[44]
n-Mg-Si and π-shape module of n-Mg-Si and p-Mn-Si	530 K	0.75 W/cm^2 and 0.99 W/cm^2	[44]
Flexible Bi-Te	70 K	0.15 W/cm^2 (targeted)	[44]
Three type of Bi-Te	493 K (523 K)	2.5 $\sim$4 W (240 W)	[44]
Sukutterudite module	550 K	1.2 W/cm^2	[44]
Bi-Te	100 K	3.6 %	[44]
p-Sb_2Te_3/Zn_4Sb_3/skutterudite n-Bi_2Te_3/skutterudite	500 K	10 %	[45]
p- Bi_2Te_3/nanostructured PbTe n- Bi_2Te_3/PbTe	590 K	11 %	[46]
Half-Heusler	656 k	11.4 %	[47]
p- Bi_2Te_3/nanostructured PbTe n- Bi_2Te_3/PbTe	590 k	12 %	[48]

respectively.

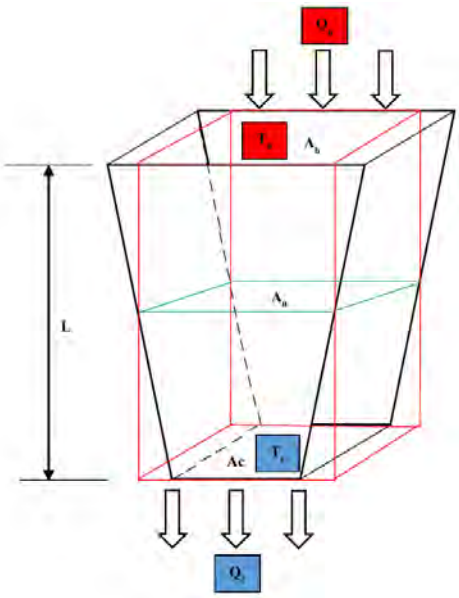


Figure 8: Geometric configuration of TE leg, rectangular area is shown in red colour and trapezoidal cross- section is shown in black colour.[40]

6 Medical Applications of Thermoelectric Devices

The human body is subject to the same laws of physics as other objects. Heat absorption and liberation by biological body also follow similar equations of conduction, convection and radiation. Conduction of heat occurs if body skin comes in contact with a cold or warm object. convection in human body is achieved through blood, gases and other fluids within it. And heat transfer through radiation is carried out due to thermal exchange between human body and the surface surrounding environment. However, these three mechanism occur simultaneously. Metabolism in living body generates heat as a heat engine. Heat transport in human body may be considered stable state. Hence, it absorbs and emits energy in equilibrium and related with metabolism [49]. TEDs are potential for both wearable home healthcare solutions and implantable medical devices owing to their solid-state nature. Further, stability and energy harvesting efficiency from low-grade heat made them attractive [5]. There are currently more implantable medical devices being used than ever before. One of the best example

is implantable cardiac pacemaker. Thermal energy harvesting from the body may extend the use of biomedical devices and overcome the restriction of battery lifetime.

6.1 Wearable Medical Devices

TEDs may be potential source of energy for low power electronics by converting body heat to electrical energy. Small and lightweight TEDs can be integrated into wearable devices to provide battery-less devices with reliability. These wireless medical devices or sensors can easily control safety and physiological conditions, health, and emergent issues and overall analysis of the patient in the hospital or at home. Body heat may act as a thermal engine to provide power to these smart devices. Furthermore, wearable medical devices are not only specified for patients but also may be used to examine during sports.

6.1.1 Wireless Autonomous Pulse Oximeter

Torfs et al. have successfully presented a wireless pulse oximeter (Figure 9) to measure the pulse and blood oxygen saturation[7]. The device is fully powered by a TE generator in the form of a watch using the person's body heat. The overall system consists of two independent blocks, a thermoelectric power supply and the wireless pulse oximeter sensor.

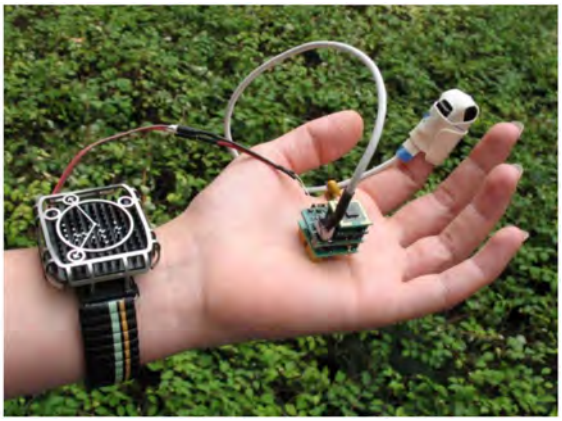


Figure 9: Wearable pulse oximeter, using 3D-stack with connectors and commercial finger sensor [7]

The whole system consumes 62 μW power in average. The measurement is completed in every 15 s. However, it requires 89μW input power and it is provided by thermoelectric generator [8]. It is noteworthy to mention that commercial BiTe-based TE generates produce 100 μW at 22°C temperature. Here all signal processing is done locally using sensor.

6.1.2 Biomedical Hearing Aids

Our ears are exquisitely sensitive. We can detect sounds when tympanic membrane or eardrum is vibrated. It is sensetive to few picometers. The initial stages of sound perception involve purely mechanical energy. Sound waves displace the eardrum and its vibration is transmitted to the inner ear or cochlea. Transportation of vibration occurs by three small bones in the middle ear: malleus, incus, and stapes. Hearing aids are primarily useful to improve the hearing and speech comprehension of people with hearing loss. It mainly results due to damage of small sensory cells in the inner ear, hear cells. This type of hearing loss is called sensorineural hearing loss. TED may play a crusial to provide electricity to hearing aids employing human body heat. Figure 10, presents the use of TEDs to supply a medical hearing prosthesis. It is a thin-film-based TED named as MPG-D602 make by Micropelt. [8].

Figure 10: Biomedical hearing aids for hearing and speech comprehension [8].

6.1.3 Electroencephalography (EEG) and Electrocardiography (ECG) Devices

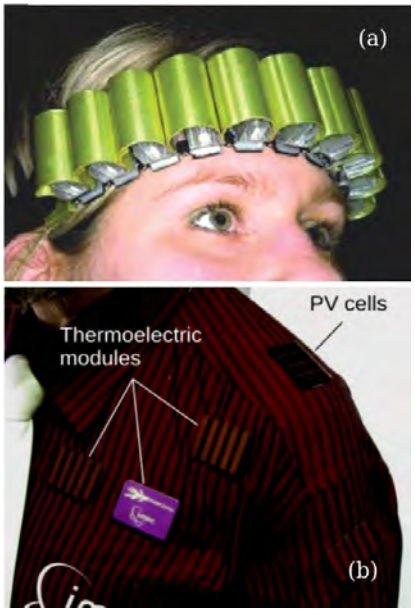


Figure 11: Wearable TEDs used in medical applications: (a) electroencephalography (EEG) headband [51], (b) Electrocardiography (ECG) shirt [52]

In order to improve the efficiency and applicability implanted/wearable TED should be thin and flexible. Leonov et.al. [50] published an interesting review on wearable TEDs and focused on rigid/flexible type TEDs. A body-powered Electroencephalogram acquisition system which produces 2–2.5 mW of power and worn as a headband is shown in Figure 11(a). Further, Leonov et.al. have been designed a TED with sixteen one-stage thermopiles from Thermix (with about two thousand BiTe thermocouples in total)[51]. The TEDs are integrated into an office-style shirt[51].

The bulk-micromachined TED produces >1 mW at the temperature of $11 - 13°$C, and generates an electrical voltage of 2V. Leonov had also presented 8 TE generators for powering implanted electrocardiography (ECG) systems in wearable textiles [52](Figure 11(b)). The used TE generator presents a figure of merit Z = 0.0025 K^{-1} and generates 0.5–5 mW at a room temperature of $15 - 27°$C. The device is comfortable because of cotton layer on the skin and a radiator made of carbon fabric and cotton.

6.2 Implantable Medical Devices

Implantable medical devices are drawing attention in the field of diagnosis, treatment and monitoring of various diseases due to advancement in microelectronics and nanofabrication. A variety of subcutaneous devices such as pacemakers, drug pumps, gastric stimulators as well as muscle, retinal and neurological stimulators are now being used for clinical applications. Some common implantable devices are shown in Figure 12.

However, power harvesting or generation is still a main challenge in such conventional implantable devices. Most of the implantable devices are powered solely by employing batteries. But, the replacement of batteries require surgery. Further, limitations of these conventional devices include their size, lifespan, and the risk of batteries leaking toxic substances. TED is one such environmental friendly device which has excellent stability and energy harvesting efficiency from low grade temperature. Implantable medical TED designing should be fabricated on the basis of power and usage requirements. Typical power requirements for implantable medical devices are in the range between 30 to 100 μW. A list of common implantable medical devices along with their typical power requirements are shown in Table 2 [5].

The implantable devices may be powered from 'inside the body' viz., such as biofuel cell, blood glucose or 'outside of the body' viz., body motion, skin thermal gradient etc. However, application and power constraints require careful device design to maximize power output from TEDs. Yang et al. performed some experimental studies in order to evaluate the feasibility of TEDs to power medical devices [54]. He found that there is available temperature differences ranging from 1 to 5K in the fat body to produce power for implanted device [54].

6.2.1 Cardiovascular Devices

Implantable Cardiovascular devices, including pacemaker, cardiac defibrillator, blood pump and cardiac loop recorders are designed to help control or monitor irregular heartbeats in people with certain heart rhythm disorders and heart failure [5].

Humans body parts may fail to work due to some internal failures such as in case of heart, sino-atrial node (SA node) may be malfunctioned which may leads to abnormal heartbeats. These arrhythmias can be very serious and may cause heart attacks and even death. In this situation pacemaker play crucial role to sustain life. However, battery provides the

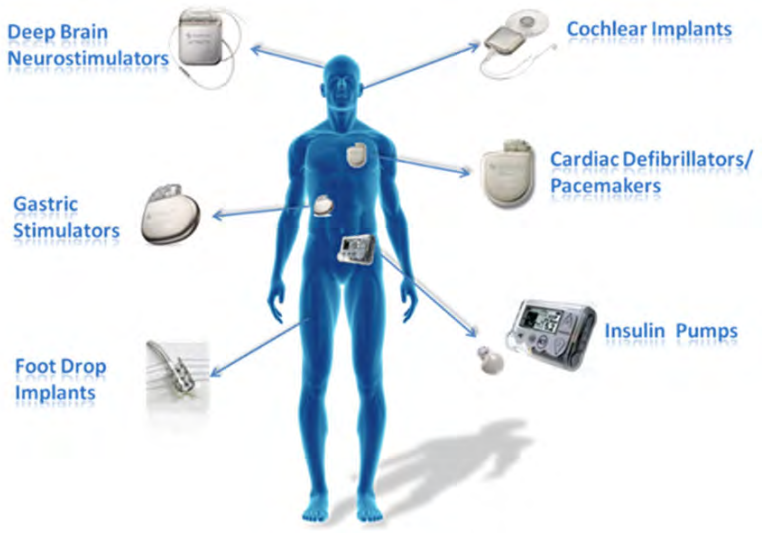


Figure 12: Common implantable devices [53]

Table 2: Power requirements of common implantable devices.

Implanted Device	**Applications**	**Typical Power Requirement**
Cardiac pacemakers	Conduction disorders	30- 100 μW
Cardiac defibrillator	Ventricular tachycardia	30- 100 μW (Idle)
Neurological stimulator	Essential tremor	30 μW to several mW
Drug pump	Spasticity	100 μW – 2 mW
Cochlear implant	Auditory assistance	Up to 10 mW
Glucose monitor	Diabetes care	> 10 µW

energy or power to the whole circuit of the pacemaker to operate. Currently, the life of these battery of pacemakers are maximum of 10 years. The depleted battery has to be surgically replaced with a new battery. This repetitive surgery causes the discomfort and life risk too to the patient.

worldwide research has been going on to convert body heat to electrical energy by using TEDs (Figure 13(a)). A study has been conducted by employing 4000 thermocouples in series of a size of about 6.0 cm^2 which generate 4V for a 10°C temperature difference[6]. The basic charging circuit along with the PN junction array for the pacemaker is shown in Figure 13(b). The primary function of the charging circuit is to monitor the voltage level using TED. IC NE555 timer is used for cut off above the threshold voltage. When the charger is disconnected from the power line, it automatically switches off within 1 min. The study estimates that the pacemaker life can be extended by more than 30 years by continuously charging with a temperature difference of 2°C.

6.2.2 Deep Brain Stimulators

Deep brain stimulation (DSB) uses a surgically implanted medical device, similar to a pacemaker, to deliver mild electrical pulses to precisely targeted areas of the brain (Figure 14).

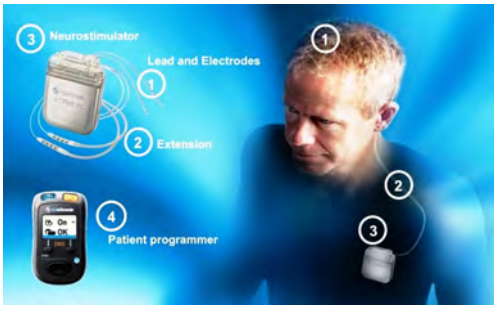


Figure 14: :Deep Brain Stimulation hardware. Source: https://iranmedtour.com/treatments/medical/deep-brain-stimulator/

In deep brain stimulation, electrodes are placed in the targeted areas of the brain. The electrodes are connected by

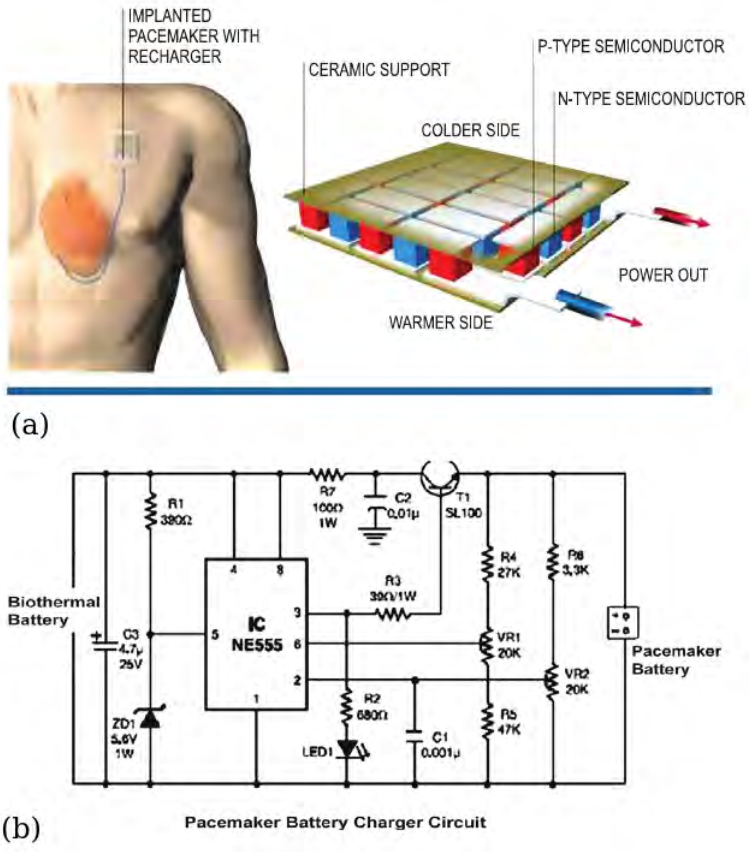


Figure 13: P-N junction array for cardiac pacemaker, (b) battery charging circuit [6]

wires (extensions) to a pacemaker type device also called neurostimulator. It is generally implanted under the clavicle or alternatively in the abdomen. The extensions are platinum (Pt)/ iridium (Ir) insulated wires that transmit the electric impulses to the electrodes. Once activated, the pulse generator sends continuous electrical pulses to the target areas in the brain. It modify the abnormal activity in that area of the brain that causes symptoms [55]. The signals are typically constant-voltage or constant-current amplitude pulses. In fact, deep brain stimulation is referred to as "the pacemaker for the brain".

7 Conclusions

Energy challenge and environmental crisis made TEDs significant as TE power generation and TE cooling device. However, utilization of TED are limited due to the low TE conversion rate. Commercially viable TE generator or cooling device require improvement in efficiency of TED . In this review we have summarized the basic criteria of a good TE material and device, fabrication process, configurations and other external factors which affect performance of the device. Choice of proper TE materials as well as proper modelling and structure are important to improve the performance of a TED. Theoretical modelling of device considering internal parameters viz., Thomson heat and the external parameters viz., contact resistance, geometric configuration of TE legs are very important to obtain high efficiency and viability of application in definite place. The feasibility of TEDs for wearable and implantable devices are discussed in detail in the review. Further, remarkable advantage of implantable TEDs in the place batteries have been highlighted. The power requirements of common implantable devices are very important and compared with generated power by TED. Further study is required on the biocompatibility of TE materials and any toxic effects on patients. Interdisciplinary nature of TE technology demand collaboration of scientists among different areas.

Acknowledgement

This work is supported by the Science and Engineering Research Board (SERB), Govt. of India in the form of sanctioning research project, File Number: EEQ/2018/001224. Author SM is thankful to CSIR, India for providing Research Fellowships.

References

[1] Rowe, D. M., *CRC Handbook of Thermoelectric.*, CRC Press,Florida,1995.

[2] Rowe, D. M. and Min, G., "Evaluation of Thermoelectric Modules for Power Generation", *J. power sources*, 73, 193-198, 1998.

[3] Min, G. and Rowe, D. M., "Optimisation of thermoelectric module geometry for waste heat electric power generation", *J. power sources*, 38, 253-259, 1992.

[4] Thimont, Y. and LeBalance. S, "The impact of Thermoelectric leg geometries on thermal resistance and power output", *J. Appl. Phys.*, 126, 095101, 2019.

[5] Chen, A. and Wright, P. K., *Medical Applications of Thermoelectrics*, CRC Press, 26-1-26-22, April, 2012.

[6] Bhatia,D., Bairagi,S., Goel,S. and Jangra, M.," Pacemakers charging using body energy," *J. Pharm. Bioallied Sci.* 51. Feb.2010.

[7] Torfs, T., Leonov, V. and Vulliers, R.J.M., "Pulse Oximeter Fully Powered by Human Body Heat," *IFSA*,80.1230.June.2007

[8] Lay-Ekuakille, A., Vendramin, G., Trotta, A. and Mazzotta, G., "Thermoelectric generator design based on power from body heat for biomedical autonomous devices," *In: International Workshop on Medical Measurements and Applications*,IEEE,1-4.

[9] Chen, G., *Nanoscale Energy Transport and Conversion."Oxford University Press*, Oxford, 2005.

[10] Ioffe, A. F., *Physics of Semiconductors*, publishing house of the U.S.S.R., Academy of Science, Moscoiw, 315-316, 1957.

[11] Snyder, G. J., Toberer, E. S., "Complex thermoelectric materials," *Nat Mater*,7,105-114, Feb, 2008.

[12] Ashcroft, N. W. and Mermin, N. D., *Solid State Physics*, Harcourt college Publishers,California,1976.

[13] Tritt, T. M., "Thermoelectric phenomena, materials, and applications.", *Annual Review of Materials Research*, 41,433-448, April,2011.

[14] Zoui, M. A., Bentouba, S., Stocholm, J. G. and Bourouis, M., "A Review of Thermoelectric Generators: Progress and Applications.", *energies*, 13,3606, June.2020.

[15] He, J., Kanatzidis, M. G. and Dravid, v. p., "High performance bulk thermoelectric via a panoscopic approach.",*Materials Today* ,16, 166-176,may.2013.

[16] Zhang, Q. H., Huang, X. Y., Bai, S. Q., Shi, X., Uher, C., and Chen, L. D., "Thermoelectric Devices Power Generation: Recent Progress and Future Challenge", *Adv. Eng. Mater*, 18, 194, 2016.

[17] He, R., Schierning, G. and Nielsch, K., "Thermoelectric Devices: A Review of Devices, Architectures, and Contact Optimization", *Adv. Mater. Technol.* , 3, 1700256,2018.

[18] Ju, Y. S. and Ghoshal, U., "Study of interface effects in thermoelectric microrefrigerators", *J. Appl. Phys.*, 88, 4135, 2000.

[19] Streetman, Ben G., Banerjee, S., *Solid State Electronic Devices*, Pearson Publishers, London, 1972

[20] Chen,G., Dresselhaus, M. S., Dresselhaus, G., Fleurial, J.P., and Caillat, T., " Recent developments in thermoelectric materials", *Int. Mater. Rev.*, 48, 45-66, 2003.

[21] El-Genk, M. S. and Saber, H. H., *In CRC Handbook of Thermoelectrics: Micro to Nano*, CRC Press, Boca Raton, FL, USA 2006, Ch. 43.

[22] Min, G., and Rowe, D. M., "Ring-structured thermoelectric module", *Semicond. Sci. Technol.*, 22, 880-883, 2007.

[23] Min, G, *In CRC Handbook of Thermoelectrics: Micro to Nano, CRC Press*, Boca Raton, FL, USA 2006, Ch.11.

[24] Fleurial, J. P., Snyder, G. J., Herman, J. A., Giauque, P. H., Phillips, W. M., Ryan, M. A., Shakkottai, P., Kolawa, E. A. and Nicolet, M. A., "THICK FILM THERMOELECTRIC MICRODEVICES", *In Eighteenth Int. Conf. on Thermoelectrics*, IEEE, Piscataway, New Jersey, 294–300.

[25] Fleurial, J. P., Borshchevsky, A., Ryan, M. A., Phillips, W., Kolawa, E., Kacisch, T. and Ewell, R., "Thermoelectric microcoolers for thermal management applications", *In XVI Int. Conf. on Thermoelectrics*, IEEE, Piscataway, New Jersey, 641–645.

[26] Enright, R., Lei, S., Nolan, K., Mathews, I., Shen, A., Levaufre, G., Frizzell, R., Duan, G. H. and Hernon, D., "A Vision for Thermally Integrated Photonics Systems", *Bell Labs Tech. J.*, 19, 31, 2014.

[27] Shakouri, A. and Yan, Z., "On-chip solid-state cooling for integrated circuits using thin-film microrefrigerators", *IEEE Trans. Compon. Packag. Technol.*, 28(1), 65, 2005.

[28] Yang, R., Chen, G., Ravi Kumar, A., Snyder, G. J. and Fleurial, J. P, "Transient cooling of thermoelectric coolers and its applications for microdevices", *Energy Convers. Manage.*, 46, 1407, 2005.

[29] Ezzahri, Y., Dilhaire, S., Patiño-Lopez, L. D., Grauby, S., Claeys, W., Bian, Z., Zhang, Y. and Shakouri, A., "Dynamical behavior and cut-off frequency of Si/SiGe microcoolers", *Superlattices Microstr.*, 41, 7, 2007.

[30] Liao, C. N., Lee, C. H., and Chen, W. J., "Effect of Interfacial Compound Formation on Contact Resistivity of Soldered Junctions Between Bismuth Telluride-Based Thermoelements and Copper", *Electrochem. Solid-State Lett.*,10 (9), 23-25, 2007.

[31] Buist, R. J. and Roman, S. J., "Development of a Burst Voltage Measurement System for High-Resolution Contact Resistance Tests of Thermoelectric Heterojunctions", *In Eighteenth Int. Conf. on Thermoelectrics*, IEEE, Piscataway, New Jersey, 249– 251.

[32] LI, F., HUANG, X., JIANG, W., and CHEN, L., "Interface Microstructure and Performance of Sb Contacts in Bismuth Telluride-Based Thermoelectric Elements", *J. Electron. Mater.*, 42, 1219, 2013.

[33] Lan, Y. C., Wang, D. Z., Chen, G. and Ren, Z. F., "Diffusion of nickel and tin in P-type (Bi, Sb)2Te3 and n type Bi2(Te, Se)3 thermoelectric materials", *Appl. Phys. Lett*., 92, 101910, 2008.

[34] Lin, W. P, Wesolowski, D. E, and Lee, C. C, "Barrier/bonding layers on bismuth telluride (Bi2Te3) for high temperature thermoelectric modules", *J. Mater. Sci.: Mater. Electron*., 22, 1313, 2011.

[35] Puschmann, R., Barbosa, M. M., Scheitz, S., Berger, L. M., Toma, F.L., Leyens, C., Beyer,E. and Dresden/D, "Technological approach for a full thermally sprayed thermoelectric generator.", *International Thermal Spray Conference*, DVS, 6.

[36] Kaibe, H. et al. “Development of thermoelectric generating stacked modules aiming for 15% of conversion efficiency”, *24th International Conference on Thermoelectrics*, 242–247, 2005.

[37] Zhang, Q. et al. “Realizing a thermoelectric conversion efficiency of 12% in bismuth telluride/skutterudite segmentedmodules through full-parameter optimization and energyloss minimized integration”, *Energy Environ. Sci.*, 10, 956–963, 2017.

[38] Kim, H. S., Liu, W., and Ren, Z., “ Efficiency and output power of thermo electric module by taking into account corrected Joule and Thomson heat.”, *J. Appl. Phys.*, 118,115103, Aug, 2015.

[39] Bjork, R., “An analytical model for the influence of contact resistance on thermoelectric efficiency.” *J. Electron. Mater*, 45(3), 1301-1308, May, 2016.

[40] Lamba, R. and Kaushik, S.C., “Thermodynamic analysis of thermoelectric generator including influence of Thomson effect and leg geometry configuration.”, *ELSEVIER*, 144, 388-398, Aug, 2016

[41] Mori, M., Yamagami, T., Sorazawa, M., Miyabe, S.T.T. and Haraguchi, T., “Simulation of fuel economy effectiveness of exhaust heat recovery system using thermoelectric generator in a series hybrid.”, *SAE Int. J. Mater. Manuf.*, 4 (1), 1268–1276, 2011.

[42] Quan, R., Liu, G., Wang, C., Zhou, W., Huang, L. and Deng, Y., “Performance investigation of an exhaust thermoelectric generator for military SUV application.”, Coatings, 8 (1), 45, 2018.

[43] Cao, Q., Luan, W. and Wang, T., “Performance enhancement of heat pipes assisted thermoelectric generator for automobile exhaust heat recovery.”, *Appl. Therm. Eng.*, 130, 1472–1479, 2018.

[44] Shinohara, Y., “Recent progress of thermoelectric devices or modules in Japan”, *Materials Today: Proceedings*, 4, 12333-12342, 2017.

[45] Caillat, T., Fleurial, J., Snyder, G.J. and Borshchevsky, A., “Development of high efficiency segmented thermoelectric unicouples”, *20th Int. Conf. on Thermoelectrics*, 282–285, 2001.

[46] Hu, X., Jood, P., Ohta, M., Kunii, M., Nagase, K., Nishiate, H., Kanatzidis M.G. and Yamamoto, A., “Power generation from nanostructured PbTe-based thermoelectrics: comprehensive development from materials to modules”, Energy Environ. Sci., 9, 517–529, 2016.

[47] Zhu, H. et al. “Discovery of TaFeSb-based half-Heuslers with high thermoelectric performance”, *Nat. Commun.*, 10, 270, 2019.

[48] Jood, P., Ohta, M., Yamamoto, A. and Kanatzidis, M.G., “Excessively doped PbTe with m Geinduced nanostructures enables high-efficiency thermoelectric modules”, Joule 2, 1339– 1355, 2018.

[49] Shang, Z. G. and Jiang, G. T., "dynamic analysis of inner metabolizing status based on the surface temp erature distribution of body", *23th IEEE.EMBS*, Oct.2001.

[50] Leonov, V., Vullers, R. J. M. and Hoff, C. V., “Thermoelectric Generator Hidden in a Shirt with a Fabric Radiator,”*9th European Conference on Thermoelectric, AIP Conf. Proc.*,556.

[51] Leonov, V., “Thermoelectric Energy Harvesting of Human Body Heat foe Wearable Sensors,” *IEEE Sensor Journal,13,2284,June.*, 2013.

[52] Leonov, V., Torfs, T., Hoof, C. V. and Vullers, R. J. M., “ Smart Wireless Sensor Integrated in Clothing: an Electrocardiography System in a Shirt Powered Using Human Body Heat,” *IFSA*,107.165.Aug.2009.

[53] Kumar, P. M., Babu, V. J., Subramanian, A., Bandle, A., Thakor, N., Ramakrishna, S. and Wei, H., “ The Design of a Thermoelectric Generator and Its Medical Applications”, *Designs*, 3-22.Apr,2019.

[54] Yang, Y., Xu, G. D. and Liu, j. “ A Prototype of an Implantable Thermoelectric Generator for Permanent Power Supply to Body Inside a Medical Device.”, *Journal of Medical Devices*, Vol. 8,014507, March, 2014

[55] Cagnan, H., Denison, T., McIntyre, C., and Brown, P., “Emerging technologies for improved deep brain stimulation”, *Nat. Biotechnol*, 37, 1024-1033, Sep,2019.

Advanced Electronic and Energy Applications of Chitin and Chitosan based Composites

Chinta Haran Majumder[1, 2], Krishanu Chatterjee[2], Arpan Kool[3,*], Somtirtha Kool Banerjee[4,#]

[1]Department of Physics, Santipur College, Santipur, Nadia, West Bengal, India
[2]Department of Physics, Techno India University, Salt Lake, Kolkata, India
[3]Department of Physics, Vidyanagar College, Vidyanagar, South 24 Parganas, West Bengal, India
[4]Department of Microbiology, Lady Brabourne College, Kolkata, West Bengal, India
*Corresponding author: akphysics89@gmail.com

Abstract Sustainable green energy resources chitin and its derivative chitosan have been considered as promising materials to reach the global energy demand in an environment-friendly way. The inherent properties such as antimicrobial, non-toxic, biocompatibility, biodegradability in addition to ease of fabrication and low-cost availability have made chitin and chitosan a suitable material for green energy harvesting, biological and industrial fields. This review provides a detailed study of chemical structures, extraction routes along with physical and chemical properties of chitin, chitosan and their composites followed by their numerous electronic and energy storage device applications including different sensors, energy harvesting devices, solar cells, fuel cells, super capacitors and Li-ion batteries.

Keywords: *Chitin, Chitosan, Electronics, Energy, Composites*

1. Introduction

Rapid advancements of modern civilization with increasing industrial and technological fields along with growing population, lead to high global demand of energy in daily life. Conventional energy resources like fossil fuels such as coal, natural gas, petroleum and its derivatives are trying to fulfil this huge energy demand. However, these fossil fuels are not unlimited in nature [1] and being exhausted rapidly [2, 3]. On the other hand, its uncontrolled exploitation and combustion also emit various pollutants in the air which are harmful for our environment [4, 5]. Hence, to overcome this energy crisis problems and environmental pollution, many research works are being carried out to find biodegradable and biocompatible sustainable green energy resources [6]. Recently, much interest have been taken on the polysaccharides specifically on chitin and its derivative chitosan by the researchers while exploring new materials for green energy applications [7]. Chitin and chitosan have been considered as attractive material for green energy resources due to their ease of fabrication, low cost, availability, biocompatibility and biodegradability properties [7, 8, 9]. Chitosan is derived from deacetylation of chitin which is the most ubiquitous natural biopolymer after cellulose. Chitosan has better solubility in water and organic solvents which makes it is more suitable than chitin for its applicability in biological fields [10, 11]. Due to its non- toxicity and biodegradability nature, chitosan has been used widely in various fields such as agriculture [12], water treatment [13, 14], food packaging [15-17] and biomedical applications [18, 19]. In recent years, chitin and chitosan based composites have gained utmost importance in electronic and electrical energy storage applications like sensors, energy harvesting devices, solar cells, organic light emitting diodes (OLEDs), supercapacitors, fuel cells, diodes and photoelectrical applications not only because of their high stretchability and better electrical conductivity but also for their non-toxic and biocompatible nature. [20-23]

2. Chitin and Chitosan Structures and Extraction

2.1. Structure of Chitin

 # Somtirtha Kool Banerjee was previously known as Somtirtha Banerjee

Chitin [$(C_8H_{13}O_5N)_n$] is a long chain natural biopolymer with two monomer units (N-acetyl-D-glucosamine and D-glucosamine) linked with β-(1-4) glycosidic bonds as shown in Figure 1 [24].

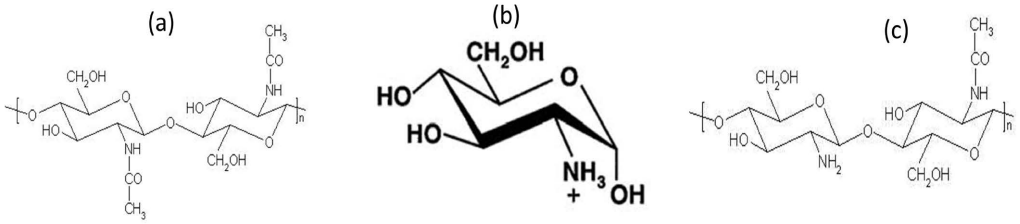


Figure 1. (a) Structure of chitin, (b) Glucosamine and (c) Chitosan. [24]

Natural chitin has three crystalline polymorphic forms namely α, β and γ (Figure 2) chitin having different orientations of microfibrils. The most abundant α-chitin has highest crystallinity with antiparallel alignment of microfibrils and found in crabs, shrimps, insect cuticles, yeast cells marine sponges and other species [25, 26]. The β- crystalline form has parallel orientation and γ-structure has a mixed alignment with two parallel microfibrils followed by one antiparallel one. β-chitin is found in chaetae of certain annelids, squids chrysalides, crustaceous, and fungi whereas γ-Chitin is rare and found in cocoons of moth and stomach of Loligo [27, 28]. These different polymorphs of chitin has different physicochemical properties depending on microstructures.

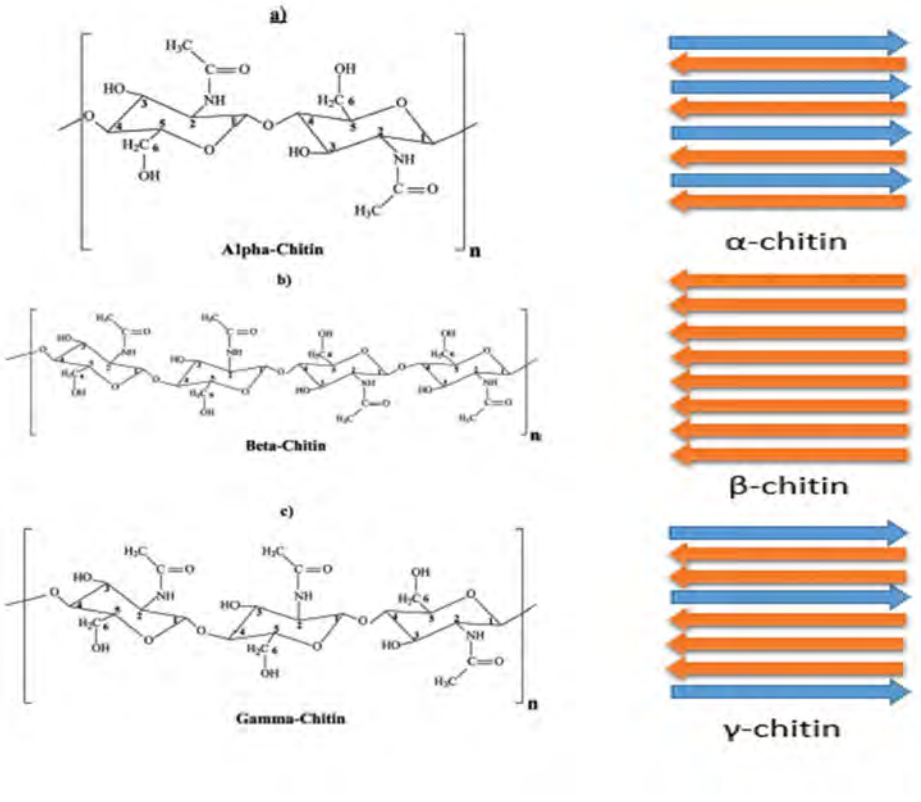


Figure 2. Schematic representation of the three polymorphic forms of chitin. [25]

2.2. Structure of Chitosan

Chitosan has a little compositional difference between the two monomers N-acetyl-D-glucosamine and D-glucosamine compared to the chitin structure. (Figure 1) Chitosan has higher number of D-glucosamine (2-amino-2-deoxy-D-glucose) monomer unit whereas chitin has more N-acetyl-D-glucosamine (N-acetyl 2-amino-2-deoxy-D-glucose) units. Thus, chitosan has a heteropolysaccharide structure with linear β-(1-4) linkage

between monomer units and it can be easily obtained from chitin by the process of deacetylation. Chitosan and has better solubility in water and acidic-aqueous solutions compared to chitin due to presence of positive charges by its amino group. Degree of deacetylation highly influence the solubility, conductivity, crystallinity, biocompatibility, biodegradability, flexibility, antioxidant, antimicrobial and other properties of cationic chitosan polymer.

2.3. Extraction of Chitin

The second most abundant natural bio-polymer chitin is mainly extracted from the shells of crabs, mussel shrimps, insect cuticles or squid gladius. Chitin can be extracted from the exoskeletons of these species either by chemical or by biological extraction techniques. The expensive, less efficient and non-ecofriendly chemical technique is primarily used for industrial applications whereas biological method of extraction is used for laboratory purpose with longer processing time though more eco-friendly nature.

Chemical extraction of chitin can be performed by three steps namely demineralization, deproteinization and decolouration. In the demineralization step minerals like calcium carbonate, calcium phosphate are removed using dilute HCl or dilute sodium hypochlorite solution whereas deproteinization or removal of protein can be performed using NaOH solution followed by washing with deionized water for removal of alkaline. Finally for the purpose of obtaining colourless product organic solvents like acetone are used. A schematic diagram for chitin extraction via chemical route is shown in Figure 3.

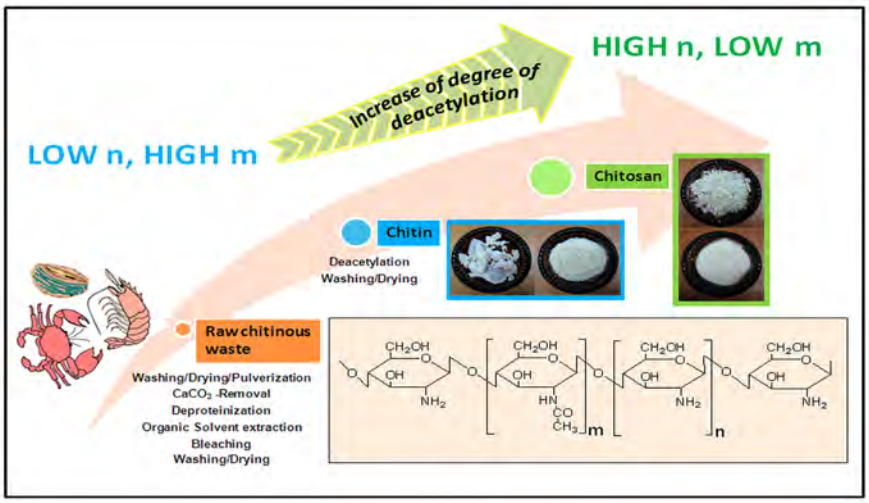


Figure 3. Chemical structures and extraction of chitin and chitosan. [Reprinted (adapted) with permission from [29]. Copyright (2013) American Chemical Society]

On the other hand, in biological extraction process of chitin is less hazardous and energy-consuming compared to the chemical method. In biological method, demineralization step is done with the help of lactic acid forming bacteria because bio-generated lactic acid removes minerals like calcium carbonate by producing calcium lactate which can be easily eliminated by washing. The deproteinization step in biological extraction method is performed by fermentation with the help of bacteria like *Pseudomonas aeruginosa K-187, Bacillus s subtilis, Bacillus cereus etc.* [30].

2.4. Extraction of Chitosan

Chitosan is obtained by the deacetylation of chitin and thus extraction of chitosan comprises of two steps: extraction of chitosan and deacetylation of it. Deacetylation is done by chemical hydrolysis or enzymatic treatment of chitin. Chemical hydrolysis method uses alkaline or acids with alkaline deacetylation having higher efficiency [31]. However, chemical hydrolysis releases many soluble and insoluble products which are harmful for environment. This is why instead of lower production rate, enzymatic treatment is used conventionally for the deacetylation of chitin. Additionally, variety of methods namely photochemical, electrochemical,

sonochemical or microwave irradiation techniques have gained attention among the scientists. Deacetylation of chitin by microwave irradiation is especially important as it not only increases the yield but also reduces the reaction time and unnecessary side reactions [32]. However, a suitable combination of chemicals with power controlled microwave irradiation technique can provide controlled deacetylation with better efficiency [33].

3. Physical and Chemical Properties

Chitin is a biodegradable, biocompatible and non-toxic polymer that is a very useful in biological and industrial field due to its unique physicochemical properties [34, 35]. The extraction methods, protein content and sources significantly affect its properties. Due to the presence of two hydroxyl and an acetamide group, chitin shows more crystallinity than chitosan with strong hydrogen bonding [36]. Chitin shows the first step of thermal decomposition due to water evaporation in the temperature range of 50°C-110°C and the second step of thermal decomposition due to degradation and dehydration of saccharide rings in the temperature range of 300°C-400°C [37]. Chitin is insoluble in water and hydrophobic in nature due to presence of larger number of N-acetyl-D-glucosamine monomer units [38]. On the other hand, chitosan is a basic natural polysaccharide with respective to other polysaccharides such as cellulose, dextran, pectin etc. which are neutral or acidic in nature. Chitosan shows hydrophilicity due to presence of large number of amino and hydroxyl groups within its structure as shown in Figure 1 though being insoluble to alkaline aqueous solution in its crystalline form [39]. Presence of amino group makes chitosan to pH sensitive and governs its cationic and solubility [40]. Chitosan shows the first step of thermal decomposition due to dehydration and attains the peak value at a temperature of 168°C. Chitosan shows the main thermal degradation in the temperature range 230°C-400°C and attains the peak value at 273°C [41]. Chitosan has received much attention to the researchers due to its availability, low-cost [42] and hydrophilic nature [43].

4. Chitin and Chitosan Based Composites for Advanced Electronics Applications

Since the invention of electronic devices, they have played an indispensable role for comfortable human life. However, in present situation, care should be taken about the electronic waste, which have now become the biggest threat to the environment. Therefore, to develop a sustainable future, natural environment-friendly materials should be used for sustainable green electronics growth [44, 45].

4.1. Sensor Applications

Recently, polymer and polymer nanocomposite based materials have gained interest among researchers to develop long-lived, low-cost multi-purpose sensors. Sensor is an electronic device that responds to a signal and convert it into electrical or magnetic form Chitosan has been confirmed as a specific polymer for sensing applications due to its chemical versatility, high adsorption capacity, mechanical robustness, flexibility, biocompatibility, biodegradability, hydrophilicity and gel forming ability along with antimicrobial and anti-oxidative properties. Chitin and chitosan based sensors can be broadly categorized in 3 types namely biosensors which senses biological reactions and convert it into detectable electrical signals, chemical sensors which detects chemical ions or gases and physical sensors which can sense physical movement or mechanical strains. [41,46,47]. The presence of 2 hydroxyls (–CH_3OH) and one acetyl (for chitin) or amino group (for both chitin and chitosan) makes chitin and chitosan suitable for chemical and biological sensing applications as the lone pair in the amino groups show affinity towards metal ions and better compatibility in some aqueous system.

4.1.1. Biosensors

Biosensors act as analytical tools in medical science for clinical detection of bio-chemical moieties [48]. In a typical electrochemical biosensor the biological element which has to be sensed is associated and interfaced with a transducer. Chitosan being biocompatible and having functional groups with possibility of chemical modification can be deposited easily on the surface of transducer forming adhesive films for the immobilization process of the sensing elements. The elements like alcohol, lactate, glutamate, glucose etc. can be detected either directly by means of their oxidation and followed by immobilization of their oxidases on chitosan composites or by the immobilizations of some dehydrogenase enzymes [49] (Figure 4) on chitosan composite films with NAD

or FAD as cofactors. Different types of biosensors for detecting variety of biochemical compounds are enlisted in Table 1 with type of chitosan composites used along with immobilized compounds on the film.

TABLE 1: ELECTROCHEMICAL BIOSENSORS BASED ON CHITIN AND CHITOSAN COMPOSITES

Types of Biosensors	Purpose of Sensing	Immobilized Agent	Chitosan composite used	Referencs
Glucose biosensors	Detection of glucose in blood or any other biological system	Glucose oxidase	Chitosan carbon nanotubes	[50]
		Glucose oxidase	Multi layered chitosan biofilms- gold nanoparticles	[51]
		Glucose oxidase	Fe_3O_4 Chitosan nafion	[52]
Lactate biosensor	Food and important medical compounds monitoring	Lactate oxidase	Chitosan-polyvinylimidazole-Os-carbon nanotubes	[53]
Glutamate sensor	Sensing of Glutamate	Glutamate oxidase	Chitosan/graphene oxide-polymerized riboflavin	[54]
Xanthine biosensor	Xanthine detection in biological systems	Xanthine oxidase	Chitosan-polypyrrole-gold nanoparticles	[55]
Galactose biosensor	Galactose detection	Galactose oxidase	Chitosan single walled carbon nanotubes	[56]
Cholesterol biosensor	Detection of cholesterol	Cholesterol oxidase	Multiwalled chitosan carbon nanotubes	[57]
Choline sensor	Detection of choline	Choline oxidase	Chitosan/titanate nanotubes	[58]
Immuno sensor	Monitoring organophosphorus(OP) pesticides chlorolpyriphos	Anti chlorpyriphos monoclonal antibody	Multiwalled carbon nanotube-chitosan-thionine	[59]
	Detection and determinationof organophosphorus(OP) pesticides	OP hydrolase	Chitosan-carbon-nanoparticles-hydroxy-apatite nanocomposite.	[60]
	Detection and monitoring of fungal hepatocarcinogen, aflatoxin B1	Polyclonal anti aflatoxin B1	Chitosan-gold nanoparticles	[61]
	To detect alpha fetoprotein in human serum	Alpha-fetoprotein antigen	Gold nanoparticles/ carbon nanotubes/chitosan nano complex	[62]
	To detect HIV1- related capsid protein P24 in human serum	P24 antigen	Gold free-single walled carbon nanotube chitosan complex	[63]

	To detect carcinoembryonic antigen	Carcinoembryonic antibodies	Chitosan gold nanoparticles	[64]
	To detect hepatitis B	Hepatitis B antibodies	Chitosan/ferrocene/ gold nanoparticles biofilm	[65]
DNA biosensor	To detect typhoid	*Salmonella typhi* single-stranded(ss) DNA	Chitosan/graphene oxide/ITO nanocomposites	[66]
	Detection of Escherichia coli	*Escherichia coli* stranded(ss) DNA	Chitosan/graphene oxide hybrid nanocomposites	[67]

In addition to the enlisted electrochemical biosensors, some biosensors were fabricated by a group of researchers for the detection and quantification of drugs and neurotransmitters like acetaminophen and mefenamic acid [68], dopamine and morphine [69], paracetamol, 5-hydroxytryptamine and dopamine [70] etc. using chitosan-multi walled carbon nanotube composite films

4.1.2. Chemical Sensors

Chemical sensors have broad range of applications in the food industry along with environmental monitoring due to their capability of sensing toxic elements, ions or gases present in food, water or air. Some chemical compounds like ethanol, nitrite, hydrogen peroxide etc. can be detected in a mechanism similar to that of electrochemical biosensors as discussed in the previous section. A sensor for sensing ethanol was fabricated by Wen et al. 2007 [71] using chitosan-eggshell film via immobilization of alcohol oxidase. This sensor is effective to study the reduction in oxygen level with respect to the ethanol concentration. Quan and Shin in 2010 [72] prepared nitrite sensor via the immobilization of Cu-containing nitrite reductase on the vilogen-chitosan film which catalyzes the nitrite reduction. For the purpose of detection of phenolic compounds Liu et al.[73] developed a sensor with horseradish pentoxide being immobilized on alumina-chitosan nanocomposite. Yang et al. in 2012 [74] devised a sensor for detecting catechol and other phenolic compounds by immobilizing tyrosinase on chitosan-nickel nanocomposite film. On the other hand laccase immobilized on ZnO-chitosan nanocomposite for sensing chlorophenol was fabricated by Mendes et al. [75] Sensors for detection of hydrogen peroxide via electrocatalytic reduction of it were developed by Akhter et al. [76] using graphene oxide-polypyrrole-chitosan film with screen-printed carbon electrodes and by Dong et al. [77] using immobilization of catalase on chitosan-β-cyclodextrin via electrocatalytic reduction of hydrogen peroxide. Teepoo et al. in 2017 devised hydrogen peroxide sensor and detector by utilizing horseradish peroxide immobilized on chitin-gelatin nanofiber composite. On the other hand Abu-Hani et al. [78] developed a high-sensitive low temperature H_2S gas sensor using glycerol ionic liquid blended conductive, transparent and flexible chitosan film. The mechanism of detection is based on the proton transfer between H_2S gas and basic amino groups from chitosan chains. This type of device is very sensitive due to large extent of H-bonding for the presence of excess OH groups coming from glycerol. This sensor can operate at a temperature of 20 ^oC at as low as 15 ppm level of gas with around 15 second response time. In addition to these sensors, many chemical sensors were developed by the researchers based on chitosan nanocomposites for the detection of trace amounts of toxic and carcinogenic metal ions. Sugunan et al in 2005 [79] and Borgohain et al. [80] developed Cu(II) and Zn(II) ion sensors using chitosan-gold nanocomposites and chitosan capped ZnS quantum dot composites respectively. In other research works, Cd(II) and Hg(II) were sensed using chitosan-carbon nanotube composite by Janegitz et al.[81] whereas Ahmed and Fekry [82] devised a Ni(II), As(II) and Pb(II) sensor using chitosan-α-Fe_2O_3 nanocomposites.

4.1.3. Physical Sensors

The field of advanced electronics of lightweight and wearable devices especially pressure sensing devices are growing fast for their various applications such as electronic skin [83,84,85], flexible touch displays [86,87], soft robotics and energy harvesting devices [88,89,90]. A piezoresistive sensor based on conducting flexible aerogel comprised of chitotsan, polyaniline and bacterial cellulose has been developed by Huang et al. [91]. In addition to piezoresistive sensor, some strain sensors have also been developed by scientists. Liu et al. [92] fabricated a high sensitive strain sensor using chitosan-carbon black conducting aerogels for sensing human activities like breathing or joint bending and another strain sensor based on spiral natural rubber, latex with carbon black and chitin nanocrystal [93]. The second sensor has high strain sensitivity and can be used efficiently to monitor human activities like movement of fingers (Figure 4) or pronunciation.

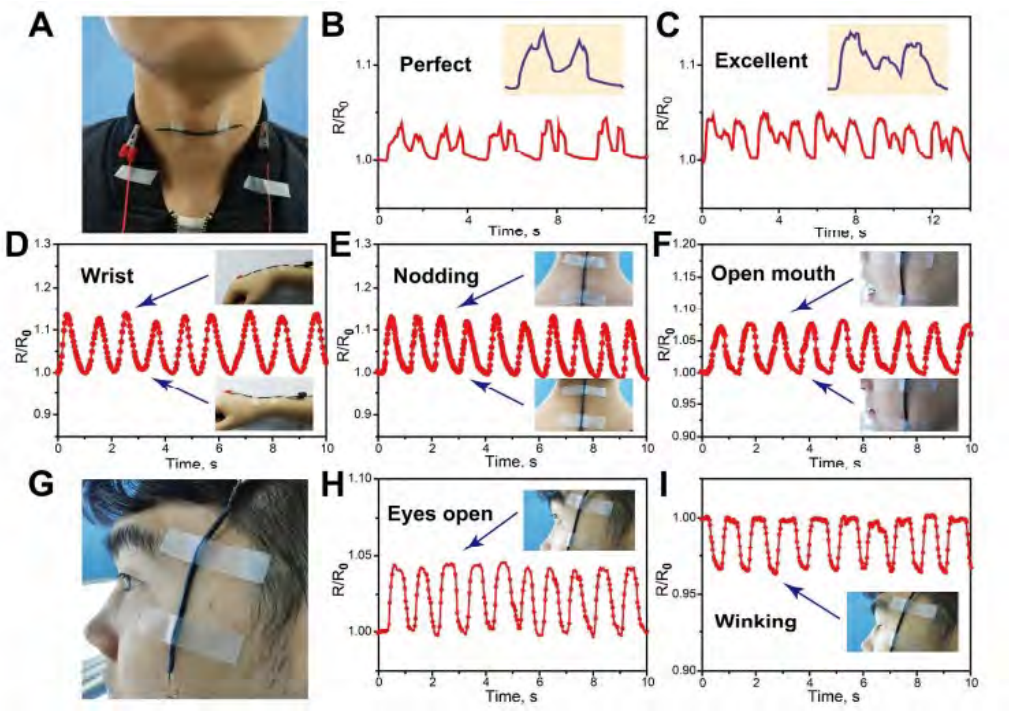


Figure 4. High-sensitive strain sensor attached to throat and cheek (A, G) along with the current signals from the sensor attached to the throat while speaking different words (B, C) and making different movements (D-F) and recorded current patterns from the sensor attached to the cheek while conducting different eye movements (H,I) [Reprinted (adapted) with permission from [93]. Copyright (2018) American Chemical Society].

4.2. Energy Harvesting Electronic Device Applications

Harvesting ambient waste energies into electricity has become very popular not only due to the huge energy crisis of modern society but also due to the recyclability and ease of access. However, some energy harvesting devices release harmful materials to the environment during the fabrication or decomposition. For the purpose of overcoming this limitation, biocompatible energy harvesting devices have gained utmost importance in the energy research field. [94,95,96] An innovative green electrical energy generation device using water vapour cell and chitosan film has been developed by Balyan et al. [97] The amine groups of chitosan acts as the active sites for the conversion of water vapour into electrical energy. The generation starts at 78% relative humidity with highest power generation of 120.13 μW at 4% chitosan concentration and this power is maintained at 90% relative humidity level. Li et al. [98] also generated electricity from chitin nanofibrils. In addition to harvesting water vapour into electricity, many researches are focused on the generation of electrical energy, harvesting ambient mechanical energy based on chitin and chitosan composites which is realised by means of the piezo- and tribo-electric properties of those composites. Hänninen et al [99] compared the piezoelectric response of pure chitosan film, pure cellulose nanofiber films and their blends. They interestingly found the best

piezoelectric sensitivity (4 pC/N) for the plain chitosan film which also has the highest elongation during its break making it most flexible among others. Hoque et al. [100] extracted chitin from the waste crab shells and fabricated pure chitin based along with chitin doped poly-vinylidene fluoride (PVDF) based piezoelectric nanogenerators. Only chitin based generator showed an open circuit voltage (V_{oc}) of ~22 V and short circuit current (I_{sc}) of ~0.12 µA whereas PVDF-chitin composite film showed ~49 V of V_{oc} and 1.9 µA of I_{sc}. On the other hand, sustainable power sources for harvesting mechanical energy by means of triboelectric power generation using pulse laser processed surface modified chitosan films have been developed by Wang et al. in 2018. [101] In a recent work, Eom et al. [102] got promising output current from the triboelectric nanogenerator using epitaxially grown PVDF-TrFE (polyvinylidene fluoride tetrafluoroethylene) on chitosan with perpendicular orientation of PVDF-TrFE having best performance (Figure 5).

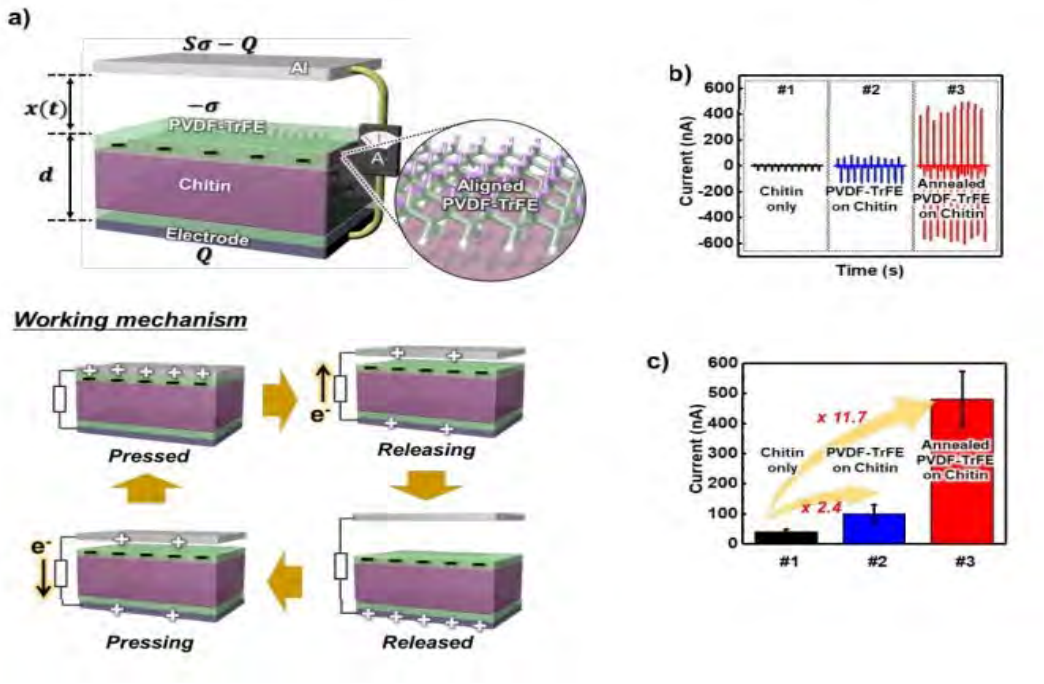


Figure 5. (a) Structure and Working output performance of PVDF-TrFE/chitin triboelectric nanogenerator, (b, c) Triboelectric output current as a function of time with different film-based sensors [Reprinted (adapted) with permission from {102}. Copyright (2020) American Chemical Society].

Kim et al. [103] developed an unconventional diatom frustule embedded chitosan based triboelectric nanogenerator (TENG) with output voltage reached upto 150 V for 0.1% diatom frustule embedded chitosan. In a latest work, Pongampai et al. [104] fabricated triboelectric-piezoelectric hybrid nanogenerator using chitosan-barium titanate ($BaTiO_3$) nanocomposites with enhanced performance by means of self-powered charge pumping mechanism.

4.3. Other Electronic Applications

There are other miscellaneous electronic applications of chitin and chitosan composites including organic light emitting diodes, Schottky diodes, solar cells etc. Lian et al. [105] fabricated organic light emitting diode (OLED) using Cu nanowire/chitosan composite as the anode and obtained the current density and luminance to be higher than the ITO device. Uzun et al. [106] prepared diodes using Al as metal and p-Si as semiconductor with 5-(2,4-dichlorophenyl)-2-furoic acid (C524D2FA) and anthraquinone-2-carboxylic (CA2CA) blended chitosan as interface layers. Al/CA2CA/p-Si diode is found to be more ideal than Al/C524D2FA/p-Si diode and most of other diodes that uses Al and p-Si as metal and semiconductor respectively both in dark and illumination of 100 mW/cm^2. Du et al. [107] prepared flexible organic thin film transistor (OTFT) using Y_2O_3/chitosan thin film as the dielectric gate of flexible OTFT where a P-type semiconductor, poly (3-hexylthiophene) (P3HT) was used as

the semiconductor layer on polyimide substrate. In this thin film transistor on/off current ratio was found to be 100 times increased along with the improvement of dielectric properties and low leakage current.

5. Energy Storage Applications

Chitin and chitosan nanocomposites based electrochemical energy storage devices such as solar cells, fuel cells, Lithium-ion batteries, and super capacitors have revealed the applications of chitin and chitosan as a potential materials for sustainable green energy storage devices [108-113].

5.1. Solar Cells

To overcome global energy crisis and huge environmental issues of conventional fossil energy resources, solar energy may be the alternative of fossil energy resources due to its no cost and renewability. As long as sun, there is no problem of harvesting energy from solar energy. To replace silicon based solar cell for its high manufacturing cost and environmental issues, a low cost, stable dye-sensitized based solar cell (DSSC) was developed by researchers.

Buraidah et al. [114] fabricated polymer electrolyte based on chitosan blended with polyethylene oxide powder (PEO) for dye-sensitized solar cells applications. A fixed amount of ammonium iodide (NH_4I) was mixed with chitosan blend. For 16.5 wt% of chitosan, 38.5 wt% of PEO and 45wt% of NH_4I showed highest ionic conductivity of 3.66×10^{-6} S/cm with current density Jsc of 2.71mA/cm^2, open circuit voltage Voc of 0.58V and efficiency 0.78%. Lojpur et al. [115] fabricated a novel electrolyte blend based on Sb_2S_3 with chitosan and polyethylene glycol (PEG).The fabricated electrolyte based solar cell offered efficiencies of 23.1%, 2.9%, 0.75% respectively at intensities of 5%, 35% and 100% Sunlight. Zhang et al. [116] developed efficient cathode interlayer film instead of substrate materials in organic solar cells (OSCs) using chitosan derivatives obtained by electrostatic layer-by-layer self-assemble technique. The film as a cathode interlayer exhibited a power conversion efficiency of 9.34%. Zulkifli et al. prepared phthaloyl chitosan (PhCh) based gel polymer electrolytes (GPE) using dimethyl formamide (DMF), ethylcarbonate (EC) and a mixed composition of potassium iodide (KI) with iodine (I_2) [117]. The maximum ionic conductivity 4.94×10^{-2} S/cm of PhCh based GPE was achieved for the 0.0012 mol of KI: I_2. When the gel polymer electrolyte sample I_2 applied to dye-sensitized solar cells (DSSCs) it showed conversion efficiency of 3.57% with ionic conductivity of 2.08×10^{-2} S/cm, a short circuit current density (J_{sc}) of 20.33 mA/cm^2, open circuit voltage Voc of 0.37 V and fill factor (FF) of 0.65. In another work, Ratan et al. [118] prepared bio-polymer based electrolyte by incorporating succinonitrile (SN) in N-Phthaloyl chitosan. The electrolyte was obtained by using dimethyl formamide (DMF), chitosan, phthalic anhydride, polyethylene oxide (PEO), ethylene carbonate (EC), tetra propyl ammonium iodide (TPAI), iodine and succinonitrile. The incorporation of succinonitrile enhanced conductivity value of 1.30×10^{-2} S/cm at 2 wt% of SN as compared to 7.84×0^{-3} S/cm at 0 wt% of SN at 25°C. The formed GPE showed an overall efficiency of 4.82% with open circuit voltage (Voc) of 0.63 V in dye-sensitized solar cells (DSSCs) applications.

5.2. Fuel Cells

Fuel cells are electrochemical devices which convert electrochemical energy into electrical energy. Proton conducting membrane based fuel cell is a promising alternative to conventional power sources. Chitin and chitosan has been extensively investigated and found as a novel material for primarily microbial fuel cell applications. Dashtimogadam et al. [119] developed a low cost, biodegradable polyelectrolyte membrane by modified chitosan structure by various amount of sulfo succinic acid/glutaraldehyde as a crosslinking agent. The formed membrane showed proton conductivity of 0.04525 S/cm and methanol permeability of 9.6×10^{-7} cm^2/sec showing a favourable power density of 17 mW/ cm^2 at 30° C and 41 mW/ cm^2 at 60° C in 2M methanol feed. Xiang et al. [120] fabricated polymer electrolyte membrane by crosslinking sulfonic groups of chitosan sulfate with amido groups of pure chitosan monomers. The obtained membrane exhibited the conductivity of 0.03 S/cm at 80°C and observed much lower methanol permeability than Nafion 112. A proton exchange membrane by modifying chitosan were fabricated by Binsu et al. [121] by introducing phosphonic acid group with it and its

composite membranes were also formed with variable compositions of polyvinyl alcohol. The obtained membrane showed good proton transport number, conductivity and higher selectivity, lower methanal permeability than Nafion 117. In another research work Hasani-Sadrabadi et al. [122] developed a low-cost, triple layer proton exchange membranes for direct methanol fuel cells applications (DMFCs). The membranes were formed by modifying structure and showed output power density of 68.10 mW/cm^2 at 5M methanol with improved proton transport conductivity and methanol permeability. He et al. [123] architectured a bioanode material with hierarchically porous chitosan and vacuum stripped graphene (CHI/VSG) for high performance microbial fuel cell applications. The formed material showed a maximum power density of 1530 mW/cm^2. Liu et al. [124] on the other hand, prepared a promising proton exchange membrane by chitosan (CS) with silica coated carbon nanotubes (SCNTs). The composite membrane showed higher mechanical properties, proton exchange conductivity than pure chitosan membrane. In another work a chitosan/polyvinyl alcohol based composite membranes with intercalated glycine betaine layered double hydroxides (LDHs) for direct methanol fuel cell applications has been developed by Hu et al. [125].The membrane with 5 wt% LDHs showed ionic conductivity of ~35.7 mS/cm at 80°C and power density of 97.8 mW/cm^2. Gong et al. [126] developed an anion exchange membrane using modified chitosan/polyvinyl alcohol with layer double hydroxides (LDHs) carbon nanotube which has been considered as a promising membrane material for direct methanol fuel cells. This membranes showed a conductivity of 47 mS/cm with 1 wt% of LDH@CNTs and maximum power density of 107.2 mW/cm^2 with 2M methanol and 5M KOH at 80°C. Li et al. [127] degradated chitin anaerobically by electroactive Aeromonas hydrophila bacteria for energy recovery and used effectively in microbial fuel cells (MFCs) for its faster degradation than fermentation system whereas Vijayalekshmi et al. [128] developed chitosan based electrolyte membrane by doping methanosulphonic acid (MSA) and sodium salt of dodecylbenzene sulfonic acid (SDBS) with crosslinked chitosan. The membranes with 15 wt% MSA showed proton conductivity of 2.86×10^{-4} S/cm at 100°C and conductivity of 4.67×10^{-4} S/cm with 10 wt% of SDBS at 100°C.

5.3. Supercapacitors

Supercapacitors are considered as a dominant components of energy storage devices due to their high charging and discharging rates, high power density and long life [129]. For safety aspects, supercapacitors require biocompatibility along with high power density and high energy density. Chitin and chitosan based supercapacitors have been widely reported recently [130-139]. Zhang et al. [130] prepared hierarchically porous carbon microspheres (HCM) with chitin/chitosan used as a forming agent. The HCM displayed specific surface area of 1450 m^2/g and polyaniline (PANI) were deposited on HCM nanocluster to use it as an electrode material for supercapacitors. In a different work, Zhang et al. [131] prepared nitrogen enriched (N-enriched) carbon nanofiber aerogels (NCNAs) by using Chitin nanofiber aerogels as the precusor. The NCNAs showed high surface area of (490-1597) m^2/g, specific capacitance 221 F/g at a current density 1A/g and high cycling stability with 92% capacitance retentivity after 8000 cycles. In another research paper, Zhang et al. [132] reported the development of 3D nitrogen doped grapheme aerogels (NGAs) by using graphene oxide (GO)

and chitosan via a self-assembly process. Furthermore, the NGAs carbonized at different temperature and NGA-900 showed excellent electrochemical performance with a high specific capacitance 244.4 F/g at a current density of 0.2 A/g and excellent cycling stability with 96.2% capacitance retentivity after 5000 cycles. Sunnetha et al. [133] developed Zn doped chitosan nanocomposites modified electrode which exhibited good capacitance and has been considered as a potential candidate for supercapacitor applications. Ba et al. [134] developed nitrogen doped hierarchical porous carbon (NHPC) materials by using chitosan and polyethelene glycol (PEG). The sample obtained (3:2 chitosan, PEG) exhibits high surface area of 2269 m^2/g and optimized pore structure. It exhibits high capacitance of 356 F/g at a current density 1A/g in 1M H_2SO_4 and 271 F/g at a current density 1 A/g in 2M KOH electrolytes. The cycling stability with 94% in 1M H_2SO_4 and 97% in 2M KOH retention after 10000 cycles 1 A/g.

5.4. Li-Ion Battery

Presence of nitrogen within chitin and chitosan structure results in an increase in their conductivity. Between the two, chitosan has been studied extensively for the use as a membrane material for lithium ion battery [140]. Some studies admitted that chitin and chitosan with other materials can be used as a potential binder and also can be used as electrodes, separators and electrolytes for Li/Na ion batteries [141-148]. Zhang et al. [141] prepared advanced sustainable separators for Li/Na ion batteries from chitin nanofiber. However, this separator shows limited performance and applications. To overcome their complicated pore forming process, low ionic conductivity and low mechanical strength, cyanoethyl groups is grafted on the surface of chitin nanofibers [149]. Wu et al. in a different work [144] developed alginate-carboxymethyl chitosan composite as a water soluble binder for Li-ion batteries. This binder exhibits excellent cycling stability with a capacity of 750 mAh/g remaining after 100 cycles. Sustainable 3D crosslinked chitosan-poly (ethylene glycol) diglycidyl ether (PEGGE) based electrolyte gel were developed by Wen et al. [142]. The obtained gel shows excellent Li-ion transportation for Li-ion batteries with mechanical strength 5.5MPa, lithium ion conductivity of 2.74×10^{-4} S/cm and an initial discharge capacity of 146.8 mAh/g with capacitance retentivity 88.49% after 360 cycles. Lee et al. [143] prepared chitosan binder with $LiFePO_4$ electrode and obtained high electrical conductivity with higher discharge capacity of 159.4 mAh/g compared to 127.9 mAh/g (for PVDF binder). The prepared binder also shows higher capacity retention ratio of 98.38% compared to 85.13% (for PVDF binder). Tang et al. [145] developed water based chitosan-oligosaccharide (COS) binder for lithium ion batteries. This binder shows initial discharge capacity of 225.6 mAh/g and 66.1 mAh/g can be obtained after 1000 cycles. Some others reported about cross-linked chitosan with silicon/graphite and cross-linked chitosan with glutaraldehyde for Li-ion batteries [146]. Recently, N-rich biochars via pyrolysis of chitosan in the temperature range 284°C-540°C has been prepared by Nistico et al. [150] The obtained biochars showed a good homogeneity, good capacity retention and improved coulombic efficiency.

6. Conclusions

Chitin and chitosan being 2nd most abundant polysaccharide and having unique combination of properties like biocompatibility, biodegradability, presence of amine and hydroxyl groups, aqueous solubility etc. have become very popular among researchers not only regarding their bio-medical and biochemical applications, but also for advanced electronics and energy storage device applications. Due to the acetyl deficiency, chitosan is a better functional material as compared to chitin. Chitin can be extracted from the exoskeletons of different natural species via demineralization and deproteinization steps which can be performed either by chemical or biological routes. Numerous biosensors for the detection of biomolecules have been developed by the scientists using composites containing chitin or chitosan. In addition to the biosensors various chemical sensors for tracing the toxic chemicals and physical sensors for sensing body movements have been developed by group of researchers using chitin or chitosan composites. On the other hand, energy harvesting devices using chitin and chitosan composites have utter significance in green energy research. In addition to these advanced electronic device applications, energy storage devices like solar cells, fuel cells, supercapacitors or Li-ion batteries have revealed the suitability and better efficiency of composites of chitin and chitosan. Thus, in summary electronic and energy storage devices based on chitin and chitosan composites are ubiquitous, easy to fabricate, cost-effective and environmentally benign in nature for which chitinous composites have become best suited material for green energy and electronic applications.

Acknowledgements

Authors acknowledge Techno India Group for their support and inspiration to complete the work.

References

[1] Abas, N., Kalair, A., Khan, N., "Review of fossil fuels and future energy technologies," *Futures,* 69. 31–49.March. 2015.

[2] Mülhaupt, R., "Green polymer chemistry and bio-based plastics: dreams and reality," *Macromol. Chem. Phys.,* 214(2). 159–174. Nov. 2012

[3] Thakur, V. K., Singha, A. S., Mehta, I. K., "Renewable resource-based green polymer composites: analysis and characterization," *Int J Polym Anal Charact*, 15(3). 137–146. 2010.

[4] Withagen, C., "Pollution and exhaustibility of fossil fuels," *Resource and Energy Economics,* 16. 235-242. 1994.

[5] Clauss, R., Mayes, J., Hilton, P., Lawrenson, R., "The influence of weather and environment on pulmonary embolism: pollutants and fossil fuels," *Medical Hypotheses*, 64.1198–1201. Nov.2004.

[6] Midilli, A., Dincer, I. Ay M., "Green energy strategies for sustainable development," *Energy Policy,* 34. 3623–3633, Sep. 2005.

[7] Thakur, V.K., Thakur, M. K., Raghavan, P., Kessler, M.R., "Progress in green polymer composites from lignin for multifunctional applications: a review," *ACS Sustain Chem Eng,* 2(5). 1072–1092. March.2014.

[8] Thakur,V.K., Singha, A.S., Kaur, I., Nagarajarao, R.P., Liping, Y., "Studies on analysis and characterization of phenolic composites fabricated from lignocellulosic fibres," *Polym Compos,* 19.505–511 . Jul. 2011.

[9] Thomas, S., Visakh, P.M., Mathew, A.P., "Natural Polymers: Their Blends, Composites and Nanocomposites: State of Art, New Challenges and Opportunities," *Advances in Natural Polymers*.,18. 1-20. DEC. 2012.

[10] Cho, Il-H., Kim, D.H., Park, S., "Electrochemical biosensors: perspective on functional nanomaterials for on-site analysis," Biomaterials Research, 24. 1, August 2021.

[11] Banerjee, S., Bagchi, B., Bhandary, S., Kool, A., Hoque N.A., Das, K., Karmakar, P., Das, S., "Antimicrobial and biocompatible fluorescent hydroxyapatite-chitosan nanocomposite films for biomedical applications," *Colloids and Surfaces B: Biointerfaces, 171. 300-307. Nov. 2018.*

[12] Chung, Y.C., Wang, H.L., Chen, Y.M., Li, S.L., "Effect of abiotic factors on the antibacterial activity of chitosan against waterborne pathogens" *Bioresource Technology,* 88(3). 179 –184. Jul. 2003.

[13] Northcott, K.A., Snape, I., Scales, P.J., Stevens, G,W., "Dewatering behaviour of water treatment sludges associated with contaminated site remediation in Antarctica,"*Chemical Engineering Science,* 60 . 6835–6843. Jul. 2005.

[14] Crin, G., "Recent developments in polysaccharide-based materials used as adsorbents in wastewater treatment," *Progress in Polymer Science,* 30(1).) 38–70. Jan. 2005.

[15] Suntornsuk, W., Pochanavanich, P., Suntornsuk,, L., "Fungal chitosan production on food processing by-products," *Process Biochemistry,* 37(7). 727–729. Feb. 2002.

[16] Pichavant, F.H., Sebe, G., Pardon, P., Coma, V., "Fat resistance properties of chitosan-based paper packaging for food applications," *Carbohydrate Polymers,* 61(3). 259-265. Aug. 2005.

[17] Devlieghere, F., Vermeulen, A., Debevere, J., "Chitosan: antimicrobial activity, interactions with food components and applicability as a coating on fruit and vegetables," *Food Microbiology,* 21(6). 703–714. Dec. 2004.

[18] Berger, J., Reist, M., Mayer, J.M., Felt, O., Gurny, R., "Structure and interactions in chitosan hydrogels formed by complexation or aggregation for biomedical application," *European Journal of Pharma-ceutics and Biopharmaceutics,* 57(1). 35-52.Jan. 2004.

[19] Ng, L.T.,Swami, S., " IPNs based on chitosan with NVP and NVP/HEMA synthesised through photoinitiator-free photopolymerisation technique for biomedical applications," *Carbohydrate Polymers*, 60(4). 523–528. Jun. 2005

[20] Amdursky, N., Glowacki, E.D., Meredith, P., "Macroscale Biomolecular Electronics and Ionics," *Advanced Materials,* 31(3). 1802221. 1-28. Oct. 2018.

[21] Jin, J., Lee, D., Im, H.G., Han, Y.C., Jeong, E.G.,Rolandi, M.,Choi, K.C.,Bae, B.S., "Chitin Nanofiber Transparent Paper for Flexible Green Electronics,"*Adv. Mater.* 28(26). 5169-75. Jul.2016.

[22] Aksoy, Ö., Uzun, I., Topal, G., Ocak, Y.S., Çelik, Ö., Batibay, D., "Synthesis, characterization, and Schottky diode applications of low-cost new chitin derivatives," *Polymer Bulletin,* 75. 2265–2283. Aug.2017.

[23] Street, R.M., Huseynova, T., Xu, X., Chandrasekaran, P., Han, L., Shih, W.Y., Shih, W.H., Schaueret, C.L.,"Variable piezoelectricity of electrospun chitin," *Carbohydr. Polym.,* 195. 218-224. Sept. 2018.

[24] Akakuru, O.U.,Louis ,H., Amos, P.I., Akakuru, O.C.,Nosike, E.I.,Ogulewe, E.F., "The Chemistry of Chitin and Chitosan Justifying their Nanomedical Utilities," *Biochem. Pharmacology*, 7(1). 55042909. 2018.

[25] Rufato, K.B., Galdino, J.P.,Ody,K.S., Pereira, A.G.B., Corradini, E., Martins, A.F., Paulino, A.T.,Fajardo, A.R., Aouada, F.A.,La Porta, F.A.,Rubira, A.F., *Muniz, E.C.,* "Hydrogels Based on Chitosan and Chitosan Derivatives for Biomedical Applications," April. 2019.

[26] Peter, S., Lyczko, N.,Gopakumar, D., Maria, H., Nzihou, A., Thomas, S., "Chitin and Chitosan Based Composites for Energy andEnvironmental Applications: A Review" *Waste and Biomass Valorization*, 12. 4777-804. Sept. 2020.

[27] Kaya, M., Mujtabaa, M., Ehrlichb, H., Salaberriac, A. M., Barand, T., Amemiyae, C.T., Gallig, R., Akyuzh, L., Sargina, I., Labidic, J., "On Chemistry of γ-chitin," *Carbohydr. Polym.,* 176. 177-186, Aug. 2017

[28] Jang, M-K., Kong, B-G., Jeong, Y-I., Lee, C-H., Nah, J-W., "Physicochemical Characterization of α-Chitin, β-Chitin, and γ-Chitin Separated from Natural Resources," J. Polym. Sci: Part A: Polym. Chem., 42, 3423–3432, March 2004

[29] Suginta, W., Khunkaewla, P., Schulte, A., "Electrochemical Biosensor Applications of Polysaccharides Chitin and Chitosan," *Chem. Rev.* 113(7). 5458-79. Jul. 2013.

[30] Aljawish, A., Chevalot, I., Jasniewski, J., Scher, J., Muniglia, L., "Enzymatic synthesis of chitosan derivatives and their potential applications," *J. Mol. Catal. B,* 112. 25–39.Feb.2015

[31] El Knidri, H., Belaabed, R., Addaou, A., Laajeb, A., Lahsini, A., "Extraction, chemical modification and characterization of chitin and chitosan," *Int. J. Biol. Marcomol.* 120.1181–1189. Dec.2018

[32] Safavy, A., Raisch, K.P., Mantena, S., Sanford, L.L., Sham, S.W., Krishna, N.R., Bonner, J.A., "Design and development of watersoluble curcumin conjugates as potential anticancer agents," *J. Med. Chem.* 50(24). 6284–6288. Nov.2007.

[33] El Knidri, H., El Khalfaouy, R., Laajeb, A., Addaou, A., Lahsini, A., "Eco-friendly extraction and characterization of chitin and chitosan from the shrimp shell waste via microwave irradiation," *Process Saf. Environ. Prot.*, 104. 395-405. November 2016.

[34] Dash, M.,Chiellini, F., Ottenbrite, R., Chiellini, E., "Chitosan—a versatile semi-synthetic polymer in biomedical applications," *Prog Polym Sci,* 36(8). 981–1014. Aug.2011.

[35] Kim, I.Y., Seo, S.J., Moon, H.S., Yoo, M.K., Park, I.Y., Kim, B.C., Cho, C.S., "Chitosan and its derivatives for tissue engineering applications," *Biotechnol Adv,* 26(1). 1–21. Jan–Feb. 2008.

[36] Su, Z., Zhang, M., Lu, Z., Song, S., Zhao, Y., Hao, Y., "Functionalization of cellulose fiber by in situ growth of zeoliticimidazolate framework-8 (ZIF-8) nanocrystals for preparing a cellulose-based air filter with gas adsorption ability,"*Cellulose,* 25.1997–2008. Feb. 2018.

[37] Paulino, A.T., Simionato, J.I., Garcia, J.C., Nozaki, J., "Characterization of chitosan and chitin produced from silkworm crysalides," Carbohydr. Polym. (2006). 64(1). 98-103. April 2006.

[38] M. Rinaudo, "Chitin and chitosan:Properties and applications**,"** *Progress in Polymer Science***,** 31(7). 603-632.Jul.2006.

[39] Sencadas,V., Correia,D.M., Areias,A.,Botelho,G., Fonseca, A.M., Neves,I.C., Ribelles,J.L.G., Mendez, S.L., "Determination of the parameters affecting electrospun chitosan fiber size distribution and morphology,"*Carbohydr.Polym,* 87.(2). 1295-1301.Jan. 2012.

[40] Shukla, S.K., Mishra, A.K., Arotiba,O.A., Mamba, B.B., "Chitosan-based nanomaterials: a state-of-the-art review," *International Journal of Biological Macromolecules,* 59. 46-58. Aug. 2013.

[41] Gopi, S., Pius, A., Kargl, R., Kleinschek, K.S., Thomas, S., "Fabrication of cellulose acetate/chitosan blend films as efficient adsorbent for anionic water pollutants, *Polym. Bull,* 76. 1557–1571. Jul. 2018.

[42] Pishbin, F., Simchi, A., Ryan, M.P., Boccaccini, A.R., "Electrophoretic deposition of chitosan / 45S5 Bioglass composite coatings for orthopaedic applications," *Surface and Coatings Technology,* 205(23–24). 5260-5268. Sept. 2011.

[43] Muzzarelli,R.A.A., "Chitins and chitosans for the repair of wounded skin, nerve, cartilage and bone," *Carbohydr. Polym.* 76 (2). 167-182.Mar. 2009.

[44] Bhutta, M. K. S., Omar, A., Yang, X., "Electronic waste: a growing concern in today's environment" *Econ. Res. Int. 2011*, No.474230. 2011

[45] Feig, V.R., Tran, H., Bao, Z., "Biodegradable polymeric materials in degradable electronic devices,"*ACS Cent. Sci.* 4. 337−348..Feb. 2018.

[46] Teles, F., Fonseca, L., "Applications of polymers for biomolecule immobilization in electrochemical biosensors," *Materials Science and Engineering: C*, 28(8).1530–1543. Dec. 2008.

[47] Zou, Y., Xiang, C., Sun, L.X., Xu, F., "Glucose biosensor based on electrodeposition of Platinum nanoparticles onto carbon nanotubes and immobilizing enzyme with chitosan-SiO2 sol-gel," *Biosens Bioelectron, 23(7). 1010-6. Feb. 2008.*

[48] Merzendorfer, H., Cohen, E., "Chitin/Chitosan: Versatile Ecological, Industrial, and Biomedical Applications," *Extracellular Sugar-Based Biopolymers Matrices*, 541-624. July.2019

[49] Zhang, M.G., Smith, A., Gorski, W., "Carbon nanotube-chitosan system for electrochemical sensing based on dehydrogenase enzymes," *Anal Chem* 76(17):5045–5050. Sept. 2004.

[50] Liu, Y., Wang, M.K., Zhao, F., Xu, Z.A., Dong, S.J. "The direct electron transfer of glucose oxidase and glucose biosensor based on carbon nanotubes/chitosan matrix," *Biosens Bioelectron,* 21(6).984–988.Dec.2005.

[51] Wu, B.Y., Hou, S.H., Yin, F., Li, J., Zhao, Z.X., Huang, J.D., Chen, Q., "Amperometric glucose biosensor based on layer-by-layer assembly of multilayer films composed of chitosan, gold nanoparticles and glucose oxidase modified Pt electrode," *Biosens Bioelectron,* 22(6).838–844. Jan.2007

[52] Yang, L.Q., Ren, X.L., Tang, F.Q., Zhang, L., "A practical glucose biosensor based on Fe3O4 nanoparticles and chitosan/nafion composite film," *Biosens Bioelectron,* 25(4). 889–895. Dec.2009.

[53] Cui, X.Q.,Li, C.M.,Zang, J.F.,Yu, S.C., "Highly sensitive lactate biosensor by engineering chitosan/ PVI-Os/CNT/LOD network nanocomposite," *Biosens Bioelectron,* 2212).3288–3292.Jun.2007.

[54] Celiesiute, R., Radzevic, A., Zukauskas, A., Vaitekonis, S., Pauliukaite, R., "A strategy to employ polymerised riboflavin in the development of electrochemical biosensors," *Electroanal* , 29(9). 2071–2082 . Jun.2017.

[55] Dervisevic, M., Dervisevic, E., Cevik, E., Senel, M., "Novel electrochemical xanthine biosensor based on chitosan-polypyrrole-gold nanoparticles hybrid bio-nanocomposite platform," *J Food Drug Anal.,* 25(3). 510–519. July.2017.

[56] Tkac, J., Whittaker, J.W., Ruzgas, T., "The use of single walled carbon nanotubes dispersed in a chitosan matrix for preparation of a galactose biosensor," *Biosens Bioelectron* ,22(8).1820–1824.March.2007.

[57] Tsai, Y.C., Chen, S.Y., Lee, C.A., "Amperometric cholesterol biosensors based on carbon nanotube-chitosan-platinum-cholesterol oxidase nanobiocomposite," *Sensor Actuators B-Chem,* 135(1). 96–101.Dec.2008.

[58] Dai, H., Chi, Y.W., Wu, X.P., Wang, Y.M., Wei, M.D., Chen, G.N , "Biocompatible electrochemi luminescent biosensor for choline based on enzyme/titanate nanotubes/chitosan composite modified electrode," *Biosens Bioelectron*, 25(6).1414–1419.Feb.2010.

[59] Sun, X., Cao, Y.Y., Gong, Z.L., Wang, X.Y., Zhang, Y.,Gao, J.M., "An amperometric immunosensor based on multi-walled carbon nanotubes-thionine-chitosan nanocomposite film for chlorpyrifos detection," *Sensors,* 12(12).17247–17261.Dec.2012.

[60] Stoytcheva, M., Zlatev, R., Montero, G., Velkova, Z., Gochev, V., "A nanotechnological approach to biosensors sensitivity improvement: application to organophosphorus pesticides determination," *Biotechnol Biotechnol Equip,* 32(1). 213–220.Oct.2017.

[61] Masoomi, L., Sadeghi, O., Banitaba, M.H., Shahrjerdi, A., Davarani, S.S.H., " A non-enzymatic nanomagnetic electro-immunosensor for determination of Aflatoxin B-1 as a model antigen," *Sensor Actuat B-Chem,* 177. 1122–1127.Feb.2013.

[62] Lin, J.H., He, C.Y., Zhang, L.J., Zhang, S.S., "Sensitive amperometric immunosensor for alphafetoprotein based on carbon nanotube/gold nanoparticle doped chitosan film," *Anal Biochem,* 384(1).130–135.Jan.2009.

[63] Giannetto, M., Costantini, M., Mattarozzi, M., Careri, M., "Innovative gold-free carbon nanotube/ chitosan-based competitive immunosensor for determination of HIV-related p24 capsid protein in serum," *RSC Adv.,* 7(63). 39970–39976. Aug. 2017.

[64] Liu, Y.X., Yuan, R., Chai, Y.Q., Hong, C.L., Liu, K.G., Guan, S , "Ultrasensitive amperometric immunosensor for the determination of carcinoembryonic antigen based on a porous chitosan and gold nanoparticles functionalized interface," *Microchim Acta* , 167(3). 217–224. Oct. 2009.

[65] Qiu, J. D., Liang, R.P., Wang, R., Fan, L.X., Chen, Y.W., Xia, X.H., "A label-free amperometric immunosensor based on biocompatible conductive redox chitosan-ferrocene/gold nanoparticles matrix," *Biosens Bioelectron,* 25(4). 852–857. Dec.2009.

[66] Singh, A., Sinsinbar, G., Choudhary, M., Kumar, V., Pasricha, R., Verma, H.N., Singh, S.P., Arora, K., "Graphene oxide-chitosan nanocomposite based electrochemical DNA biosensor for detection of typhoid," *Sensor Actuat B-Chem,* 185. 675–684. Aug.2013.

[67] Xu, S.C., Zhang, Y.Y., Dong, K., Wen, J.N., Zheng, C.M., Zhao, S.H., "Electrochemical DNA biosensor based on graphene oxide-chitosan hybrid nanocomposites for detection of Escherichia coli O157:H7," *Int J Electrochem Sc* 12. 3443–3458.March.2017.

[68] Babaei, A., Afrasiabi, M., Babazadeh, M., "A glassy carbon electrode modified with multiwalled carbon nanotube/chitosan composite as a new sensor for simultaneous determination of acetaminophen and mefenamic acid in pharmaceutical preparations and biological samples," *Electroanal,* 22(15).1743–1749.July. 2010.

[69] Babaei, A., Babazadeh, M., "Multi-walled carbon nanotubes/chitosan polymer composite modified glassy carbon electrode for sensitive simultaneous determination of levodopa and morphine," *Anal Methods.* 3(10)..2400–2405. Sept. 2011.

[70] Xu, H.R., Wang, L., Luo, J.P., Song, Y.L., Liu, J.T., Zhang, S., Cai, X.X., "Selective recognition of 5-hydroxytryptamine and dopamine on a multi-walled carbon nanotube-chitosan hybrid film modified micro electrode array," *Sensors,* 15(1).1008–1021.Jan.2015.

[71] Wen, G.M., Zhang, Y., Shuang, S.M., Dong, C., Choi, M.M.F., "Application of a biosensor for monitoring of ethanol," *Biosens Bioelectron,* 23(1).121–129.April 2007.

[72] Quan, D., Shin, W., "A nitrite biosensor based on co-immobilization of nitrite reductase and viologen-modified chitosan on a glassy carbon electrode," *Sensors,* 10. 6241–6256.Jun.2010.

[73] Liu, X.J., Luo, L.Q., Ding, Y.P., Xu, Y.H., "Amperometric biosensors based on alumina nanoparticles-chitosan-horseradish peroxidase nanobiocomposites for the determination of phenolic compounds," *Analyst* ,136(4). 696–701.Feb.2011

[74] Yang, L., Xiong, H., Zhang, X., Wang, S., "A novel tyrosinase biosensor based on chitosan carbon-coated nickel nanocomposite film," *Bioelectrochemistry,* 84. 44–48.April.2012.

[75] Mendes, R.K.,Arruda, B.S.,deSouza, E.F.,Nogueira, A.B.,Teschke, O.,Bonugli, L.O.,Etchegaray, A., "Determination of chlorophenol in environmental samples using a voltammetric biosensor based on hybrid nanocomposite," *J Brazil Chem Soc,* 28(7).1212–1219.July.2017.

[76] Akhtar, M.A., Hayat, A., Iqbal, N., Marty, J.L., Nawaz, M.H., "Functionalized graphene oxidepolypyrrole-chitosan (fGO-PPy-CS) modified screen-printed electrodes for non-enzymatic hydrogen peroxide detection," *J Nanopart Res,* 19. 334. Oct.2017.

[77] Dong, W.B.,Wang, K.Y., Chen, Y., Li, W.P., Ye, Y.C., Jin, S.H., "Construction and characterization of a chitosan-immobilized-enzyme and beta-cyclodextrin-included-ferrocene-based electrochemical biosensor for H2O2 detection," *Materials,* 10(8). 868. July.2017.

[78] Abu-Hani, A.F.S., Greish, Y.E., Mahmoud, S.T., Awwad, F., Ayesh, A.I., "Low-temperature and fast response H2S gas sensor using semiconducting chitosan film," *Sens. Actuators B,* 253. 677-684. Dec. 2017.

[79] Sugunan, A., Thanachayanont, C., Dutta, J., Hilborn, J.G., "Heavy-metal ion sensors using chitosan-capped gold nanoparticles," *Sci Technol Adv Mater,* 6. 335–340. March.2005.

[80] Borgohain, R., Boruah, P.K.,, Baruah, S., "Heavy-metal ion sensor using chitosan capped ZnS quantum dots," *Sens. Actuators B,* 226. 534-539. April 2016.

[81] Janegitz, B.C., Figueiredo, L.C.S., Marcolino, L.H., Souza, S.P.N., Pereira, E.R., Fatibello, O., "Development of a carbon nanotubes paste electrode modified with crosslinked chitosan for cadmium (II) and mercury(II) determination," *J Electroanal Chem* 660(1).209–216.Sept.2011.

[82] Ahmed, R.A., Fekry, A.M., "Preparation and characterization of a nanoparticles modified chitosan sensor and its application for the determination of heavy metals from different aqueous media," *Int J Electrochem Sci,* 8(5). 6692–6708.May.2013.

[83] Kim,D.H., Lu, N., Ma, R., Kim, Y.S., Kim, R.H., Wang, S., Wu, J., Won, S.M., Tao, H., Islam, A., Yu,K.J., Kim, T.,Chowdhury, R., Ying,M., Xu, L., Li, M.,Chung,H.J., Keum,H., McCormick, M., Liu, P., Zhang, Y.W., Omenetto, F.G., Huang, Y., Coleman, T.,Rogers, J.A., "Epidermal electronics," *Science*, 333. 838–843. Aug.2011.

[84] Mannsfeld, S.C.B., Tee, B.C.K., Stoltenberg, R.M., Chen, C.V.H.H., Barman, S., Muir,B.V.O., Sokolov,A.N., Reese, C., Bao,Z., "Highly sensitive flexible pressure sensors with microstructured rubber dielectric layers," *Nat. Mater.*, 9. 859 –864. Sept.2010.

[85] Schwartz, G., Tee,B.C.K., Mei,J., Appleton,A.L., Kim,D.H., Wang,H., Bao, Z., "Flexible polymer transistors with high pressure sensitivity for application in electronic skin and health monitoring," *Nat.Commun.*,4.1859.May.2013.

[86] Lipomi, D.J., Vosgueritchian, M., Tee,B.C.K., Hellstrom, S.L., Lee,J.A., Fox,C.H.,Bao, Z.,"Skin-like pressure and strain sensors based on transparent elastic films of carbon nanotubes," *Nat.Nano*,6. 788–792. Oct.2011.

[87] Fan, F.R., Lin, L., Zhu, G., Wu, W., Zhang, R., Wang, Z.L.,"Transparent triboelectric nanogenerators and self-powered pressure sensors based on micropatterned plastic films," *Nano Lett.,* 12. 3109 – 3114.May.2012.

[88] Wang, Z. L., Wu, W., "Nanotechnology-enabled energy harvesting for self-powered micro-/nanosystems," *Angewandte Chemie International Edition,* 51(47). 11700 –11721. Nov. 2012.

[89] Yang,Y., Zhang,H., Lin,Z.H., Zhou,Y.S., Jing, Q., Su, Y., Yang, J., Chen,J., Hu, C., Wang, Z.L., "Human skin based triboelectric nanogenerators for harvesting biomechanical energy and as self-powered active tactile sensor system," *ACS Nano,* 7. 9213–9222. Sept. 2013.

[90] Hu, Y., Yang, J., Jing, Q., Niu, S., Wu, W., Wang, Z.L., "Triboelectric nanogenerator built on suspended 3D spiral structure as vibration and positioning sensor and wave energy harvester," *ACS Nano.,* 7. 10424 –10432. Oct.2013.

[91] Huang, J., Li, D., Zhao, M., Ke, H , Mensah, A., Lv, P., Tian, X., Wei, Q., "Flexible electrically conductive biomass-based aerogels for piezoresistive pressure/strain sensors," *Chemical Engineering* Journal, 373. 1357-1366. Oct. 2019.

[92] Liu, Y., Zheng, H., Liu, M., "High performance strain sensors based on chitosan/carbon black composite sponges," *Mater. Design,* 141. December 2017.

[93] Liu, Y., Wu, F., Zhao, X., Liu, M, "High-Performance Strain Sensors Based on Spirally Structured Composites with Carbon Black, Chitin Nanocrystals, and Natural Rubber," *ACS Sustainable Chem. Eng.* 2018, 6, 8, 10595–10605, June 11, 2018.

[94] Torres, F. G., De-la-Torre, G.E.,"Polysaccharide-based triboelectric nanogenerators:A review, "*Carbohydrate Polymers*,, 251. 117055.Jan. 2021.

[95] Slabov, V., Kopyl, S., Santos, M.P.S., Kholkin, A.L., "Natural and Eco-Friendly Materials for Triboelectric Energy Harvesting," *Nano-Micro Letters,* 12. 42. Jan.2020.

[96] Wang, Y.M.,Zeng, Q.,He, L., Yin, P., Sun, Y., Hu, W., Yang, R., "Fabrication and application of biocompatible nanogenerators,"iScience, 24(4).102274. April 2021.

[97] Balyan, M., Nasution, T.I., Nainggolan, I., Mohamad, H., Ahmad, Z.A., "Energy harvesting properties of chitosan film in harvesting water vapour into electrical energy," *Journal of Materials Science: Materials in Electronics,* 30. 16275–16286. Aug.2019.

[98] Li, M., Zong, L., Yang, W., Li, X., You, J., Wu, X., Li, Z., Li, C., "Biological Nanofibrous Generator for Electricity Harvest from Moist Air Flow," *Advanced Functional Materials, 29(*32).1901798. Jun. 2019.

[99] Hänninen, A., Sarlin, E., Lyyra, I., Salpavaara, T., Tuukkanen, S., "Nanocellulose and chitosan based films as low cost, green piezoelectric materials," *Carbohydrate Polymers,* 202. 418-424. Dec. 2018.

[100] Hoque, N.A., Thakur, P., Biswas, P., Saikh, M., Roy, S., Bagchi, B.,Das, S., Ray, P.P., "Biowaste crab shell-extracted chitin nanofiber-based superior piezoelectric nanogenerator,"*J. Mater. Chem. A,* 6. 13848-13858. Jun. 2018.

[101] Wang, R., Gao, S., Yang, Z., Li, Y., Chen, W., Wu, B., Wu, W., "Engineered and Laser-Processed Chitosan Biopolymers for Sustainable and Biodegradable Triboelectric Power Generation," *Advanced Materials,* 30(11). 1706267.Jan. 2018.

[102] Eom, K., Shin, Y.E., Kim, J.K., Joo, S.H., Kim, K., Kwak, S.K., Ko, H., Jin, J., Kang, S.J., "Tailored Poly(vinylidene fluoride-co-trifluoroethylene) Crystal Orientation for a Triboelectric Nanogenerator through Epitaxial Growth on a Chitin Nanofiber Film," *Nano Lett.* 20.9.6651–6659.Aug. 2020.

[103] Kim, J. N., Lee, J., Go, T.W., Abhari, A.R., Mahato,M., Park, J.Y., Lee, H., Oh , I.K., "Skin-attachable and biofriendly chitosan-diatom triboelectric nanogenerator," *Nano Energy,* 75.104904. Sept.2020.

[104] Pongampai, S., Charoonsuk, T., Pinpru,N., Pulphol, P., Vittayakorn, W., Pakawanit, P., Vittayakorn, N., "Triboelectric-piezoelectric hybrid nanogenerator based on BaTiO3-Nanorods/Chitosan enhanced output performance with self-charge-pumping system," *Composites Part B: Engineering,* 208. 108602. March 2021.

[105] Lian, L., Wang, H, Dong, D., He, G., "Highly robust and ultrasmooth copper nanowire electrode by one-step coating for organic light-emitting diodes," *J. Mater. Chem. C, 6(34). 9158-9165.* Aug.2018.

[106] Uzun, I., Orak, I., Yağmur, H.K., Karakaplan, M., Yalman, M , "Determination of Electrical and Photoelectrical Properties of Schottky Diodes Made Using New Chitin Derivatives Synthesized as Interface Layer," *Silicon,* Oct.2020.

[107] Du, B.W., Hu, S.H., Singh, R., Tsai, T.T., Lin, C.C., Ko,F.H., "Eco-Friendly and Biodegradable Biopolymer Chitosan/Y2O3 Composite Materials in Flexible Organic Thin-Film Transistors," *Materials* ,10(9).Sept. 2017.

[108] Conway, B.E., *Electrochemical Supercapacitors: Scientific Fundamentals and Technological Applications,* Plenum Press, New York , 1999.

[109] Zheng, J.P., Jow, T.R., "A New Charge Storage Mechanism for Electrochemical Capacitors," *Journal of The Electrochemical Society,* 142. L6 - L8. 1995.

[110] Soudan, P., Gaudet, J., Guay, D., Bélanger,D., Schulz, R., "Electrochemical properties of ruthenium-based nanocrystalline materials as electrodes for supercapacitors," *Chem. Mater.,* 14(3). 1210–1215. Feb. 2002.

[111] Conway, B.E., Birss, V., Wojtowicz, J., "The role and utilization of pseudocapacitance for energy storage by supercapacitors," *Journal of Power Sources,* 66.(1-2). 1-14. May-Jun. 1997.

[112] Lin, C., Ritter, J A., Popov, B.N., "Characterization of Sol-Gel-Derived Cobalt Oxide Aerogels as Electrochemical Capacitors," *Journal of the Electrochemical Society,* 145. 4097-4103. Jul. 1998.

[113] Wu, N. L., "Nanocrystalline oxide supercapacitors," *Materials Chemistry and Physics*, 75 (1–3). 6-11. April. 2002.

[114] Buraidah, M. H., Teo, L. P., Yong, C. M. A., Shah, S., Arof. A. K., "Performance of polymer electrolyte based on chitosan blended with poly (ethyleneoxide) for plasmonic dye-sensitized solar cell," *Optical Materials*, 57, 202-211. July.2016.

[115] Lojpur, V., Kristi, J., KacarevicPopovic, Z., Mitric, M., Rakcevic, Z., Validzic, I. L. J., "Efficient and novel Sb2S3 based solar cells with chitosan/poly (ethylene glycol)/electrolyteblend," *Int. J. Energy Res.* 42. 2. 843-852. Sept. 2017

[116] Zhang,K.,Xu,R.,Ge,W.,Qi,M.,Zhang,G.,Xu,Q.H.,Huang,F.,Cao,Y.,Wang,X., "Electrostatically self-assembled chitosan derivatives working as efficient cathode interlayers for organic solarcells,"*NanoEnergy,*34(C).164-171.2017

[117] Zulkifli,A.M.,Said,N.I.A.M.,Aziz,S.B.,Dannoun,E.M.A.,Hisham,S.,Shah,S.,Bakar,A.A.,Zainal,Z.H.,Tajuddin,H.A.,Hadi,J.M.,Brza,M.A.,Saeed, S.R.,Amin,P.O., "Characteristics of Dye-Sensitized Solar Cell Assembled from Modified Chitosan-Based Gel Polymer Electrolytes Incorporated with Potassium Iodide," *Molecules*,25(18).4115.Sept.2020

[118] Ratan,A.,Buraidah,M.H.,Teo,L.P.,Singh,P.K.,Arof,A.K., "Enhanced photo-current conversion efficiency by incorporation of succinonitrile in N-Phthaloyl chitosan based bio-polymer electrolyte for dye-sensitized solar cell," *Optik*,222.165467.Nov.2020.

[119] Dashtimoghadam, E., Hasani-Sadrabadi, M.M., Moaddel, H., "Structural modification of chitosan biopolymer as a novel polyelectrolyte membrane for green power generation," *Polym. Adv*. Technol. 21(10). 726-734. Jun.2009.

[120] Xiang, Y., Yang, M., Guo, Z., Cui, Z. "Alternatively chitosan sulfate blending membrane as methanol-blocking polymer electrolyte membrane for direct methanol fuel cell," *J. Membr. Sci*.,337 (1-2). 318-323. July.2009.

[121] Binsu, V.V., Nagarale, R.K., Shahi, V.K., Ghosh, P.K., "Studies on N-methylene phosphonic chitosan/poly (vinyl alcohol) composite proton-exchange membrane," *Reactive and Functional Polymers,* 66(12).1619-1629. Dec.2006.

[122] Hasani-Sadrabadi, M.M., Dashtimoghadam, E., Mokarram, N., Majedi, F.S., Jacob, K.I., "Triple-layer proton exchange membranes based on chitosan biopolymer with reduced methanol crossover for high-performance direct methanol fuel cells application," *Polymer (Guildf)* (2012).Polymer 53 (13), 2643-2651,Jun. 2012

[123] He, Z., Liu, J., Qiao, Y., Li, C.M., Tan, T.T.Y., "Architecture engineering of hierarchically porous chitosan/vacuum-stripped graphene scaffold as bioanode for high performance microbial fuel cell," *Nano Lett.* (2012).Nano letters 12 (9). 4738-4741. Aug.2012.

[124] Liu, H., Gong, C., Wang, J., Liu, X., Liu, H., Cheng, F., Wang, G., Zheng, G., Qin, C., Wen, S., "Chitosan/silica coated carbon nanotubes composite proton exchange membranes for fuel cell applications," *Carbohydr. Polym.* (2016)Carbohydrate polymers 136. 1379-1385. Jan. 2016

[125] Hu, Y., Tsen, W.C., Chuang, F.S., Jang, S.C., Zhang, B., Zheng, G., Wen, S., Liu, H., Qin, C., Gong, C., "Glycine betaine intercalated layered double hydroxide modified quaternized chitosan/polyvinyl alcohol composite membranes for alkaline direct methanol fuel cells ," *Carbohydrate Polymers*, 213.320-328.June 2019.

[126] Gong, C., Zhao, S., Tsen, W.C., Hu, F., Zhong, F., Zhang, B., Liu, H., Zheng, G., Qin, C., Wen, S., "Hierarchical layered double hydroxide coated carbon nanotube modified quaternized chitosan/polyvinyl alcohol for alkaline direct methanol fuel cells," *Journal of Power Sources, 441. 227176. Nov. 2019*

[127] Li, S.W., He, H., Zeng, R.J., Sheng, G.P., "Chitin degradation and electricity generation by Aeromonas hydrophila in microbial fuel cells," *Chemosphere* (2017). Chemosphere 168. 293-299. Feb.2017.

[128] Vijayalekshmi, V., Khastgir, D., "Eco-friendly methanesulfonic acid and sodium salt of dodecylbenzene sulfonic acid doped cross-linked chitosan based green polymer electrolyte membranes for fuel cell applications," *Journal of Membrane Science*, 523. 45-59, Feb.2017.

[129] Salanne, M., Rotenberg, B., Naoi, K., Kaneko, K., Taberna, P. L., Grey, C. P., Dunn, B., Simon, P., "Efficient storage mechanisms for building better supercapacitors," *Nature Energy*, 1. 16070. May. 2016.

[130] Gao, L., Xiong, L., Xu, D., Cai, J., Huang, L., Zhou, J., Zhang, L., "Distinctive Construction of Chitin Derived Hierarchically Porous Carbon Microspheres / Polyaniline for High Rate Supercapacitors," *ACS Appl. Mater. Interfaces, Aug*2018

[131] Ding, B., Huang, S., Pang, K., Duan, Y., Zhang, J., "Nitrogen-Enriched Carbon Nanofiber Aerogels Derived from Marine Chitin for Energy Storage and Environmental Remediation," *ACS Sustainable Chem. Eng*., 6(1). 177–185.Oct.2017.

[132] Zhang, Y., Zhu, J.Y., Ren, H.B., Bib, Y.T., Zhang, L., "Facile synthesis of nitrogen-doped grapheme aerogels functionalized with chitosan for supercapacitors with excellent electrochemical performance," *Chinese Chemical Letters,* 28(5). 935-942. May. 2017.

[133] Suneetha, R.B., Selvi, P., Vedhi, C., "Synthesis,structural and electrochemical characterization of Zn doped ironoxide/ grapheneoxide/ chitosan nanocomposite for supercapacitor application, "*Vacuum,* 164. 396-404. Jun. 2019.

[134] Ba , Y., Pan, W., Pi, S., Zhao,Y., Mi, L., "Nitrogen-doped hierarchical porous carbon derived from a chitosan / polyethyleneglycol blend for high performance supercapacitors," *RSC Adv.*, 8.7072-7079. Feb. 2018.

[135] Lin, Z., Xiang, X., Peng, S., Jiang, X., Hou, L., "Facile synthesis of chitosan-based carbon with rich porous structure for supercapacitor with enhanced electrochemical performance," *Journal of Electroanalytical Chemistry,* 823. 563-572.2018

[136] Gopalakrishnan, A., Vishnu, N., Badhulika, S., "Cuprous oxide nanocubes decorated reduced grapheme oxide nanosheets embedded in chitosan matrix : A versatile electrode material for stable supercapacitor and sensing applications," *Journal of Electroanalytical Chemistry,* 834.187-195. Feb.2019.

[137] Raj, C.J., Rajesh, M., Manikandan, R., Yu,K.H., Anusha, J.R., Ahn, J.H., Kim, D.W., Park, S.Y., ChulKim, B., "High electrochemical capacitor performance of oxygen and nitrogen enriched activated carbon derived from the pyrolysis and activation of squid gladius chitin," *Journal of Power Sources,* 386. 66 -76. May. 2018.

[138] Genovese, M., Wu, H., Virya, A., Li, J., Shen, P., Lian, K., "Ultrathinall-solid-state supercapacitor devices based on chitosan activated carbon electrodes and polymer electrolytes," *ElectrochimicaActa,* 273. 392-401. May2018.

[139] Ramkumara, R., Minakshi, M., "Fabrication of ultrathin CoMoO4 nanosheets modified with chitosan and their improved performance in energy storagedevice," *DaltonTrans.* ,44(13). 6158-6168. Feb. 2015.

[140] Jafarizadeh, H.M., Jahanian, M.A., Berenjian, A., "Potential applications of chitosan nanoparticles as novel support in enzyme immobilization,"*Am J BiochemBiotechnol*., 8(4). 203–219. Sept. 2012

[141] Zhang, T.,Shen, B., Yao, H.B., Ma, T., Lu, L.L., Zhou, F., Yu, S.H.,"Prawn Shell Derived Chitin Nanofiber Membranes as Advanced Sustainable Separators for Li/Na-Ion Batteries,"*NanoLett.*, 17(8). 4894-4901. Aug. 2017

[142] Xu, D., Jin, J., Chen, C., Wen, Z., "From Nature to Energy Storage: A Novel Sustainable 3D CrossLinkedChitosan−PEGGE-Based Gel Polymer Electrolyte with Excellent Lithium-Ion Transport Properties for Lithium Batteries," *ACS Appl. Mater. Interfaces,* 10. 38526−38537. Oct. 2018.

[143] Prasanna, K., Subburaj, T., Jo, Y.N., Lee, W.J., Lee, C.W., "Environment-Friendly Cathodes Using Biopolymer Chitosan with Enhanced Electrochemical Behavior for Use in Lithium Ion Batteries,"*ACS Appl. Mater. Interfaces,*7. 7884−7890. Mar.2015.

[144] Wu, Z.H., Yang, J.Y., Yu, B., Shi, B.M., Zhao, C.R.,Yu, Z.L.,"Self-healing alginate–carboxymethylchitosan porous scaffold as an effective binder for silicon anodes in lithium-ion batteries,"*Rare Metals, 99067078. Jun. 2016.*

[145] Tang, H., Weng, Q., Tang, Z., "Chitosan oligosaccharides: a novel and efficient water soluble binder for lithium zinc titanate anode in lithium-ion batteries,"*ElectrochimicalActa*, 151. 27-34. Jan. 2015.

[146] Zhao, X., Yim, C.H., Du, N., Abu-Lebdeh, Y., "Crosslinked Chitosan Networks as Binders for Silicon/Graphite Composite Electrodes in Li-Ion Batteries," *Journal of The Electrochemical Society,* 165 (5). A1110-A1121. April 2018.

[147] Chen, C., Lee, S.H., Cho, M., Kim, J., Lee, Y., "Cross-Linked Chitosan as an Efficient Binder for Si Anode of Li-ion Batteries,"*ACS Appl. Mater. Interfaces,* 8. 2658−2665. Jan. 2016,

[148] Gao, H., Zhou, W., Jang, J.H., Goodenough, J.B., "Cross-Linked Chitosan as a Polymer Network Binder for an Antimony Anode in Sodium-Ion Batteries," *Adv. Energy Mater*., 6. 1502130. Jan. 2016.

[149] Zhang, T.W., Chen, J.L., Tian, T., Shen, B., Peng, Y.D.,Song, Y.H., Jiang, B., Lu, L.L., Yao, H.B.,Yu, S.H., "Sustainable Separators for High-Performance Lithium Ion Batteries Enabled by Chemical Modifications," *Adv. Funct. Mater.,* 2019, 29, 1902023

[150] Nisticò, R.,Guerretta, F.,Benzi, P.,Magnacca, G.,"Chitosan-derived biochars obtained at low pyrolysis temperatures for potential application in electrochemical energy storage devices," *International Journal of Biological Macromolecules,*164 . 1825–1831. Aug 2020.

Advances in Modern and Applied Sciences

A Collection of Research Reviews on Contemporary Research (Volume 1)
Sujay Pal, Tushar Kanti Biswas. Editors

This book materializes our long-cherished dream of publishing a series of volumes consisting of review papers on contemporary research fields from a broad spectrum of basic sciences. The present volume, which is our first baby-step towards that fulfillment, includes a collection of twenty-five review articles contributed by about fifty researchers and scientists whose vocations are in diverse fields of science including astrophysics, astronomy, space and atmospheric sciences, computer sciences to material sciences.

The main objective of this book is to provide an insight into the advances that modern day science has made and bring forth a better understanding of this vast and exhilarating discipline called science. We are certain that new graduates, PhD scholars, teachers, and researchers from diverse fields will benefit from this volume, which can be considered as a stock-taking of the new developments in recent day science.

The editors have compiled and edited the articles duly to suit the purpose of the book and at the same time to keep a balance between diverse topics. The book is organized into four specific chapters. Chapter 1 consists of nine articles from Astrophysics, Astronomy. Chapter 2 is devoted to Atmospheric and Space Sciences comprising eight articles. In Chapter 3, researchers have explored advances in modern Computer Science and Mathematics through five articles focusing on recent topics. Finally, Chapter 4 contains three articles from Material Sciences giving an overview of the current state of the art on topics like data-based material designing using Machine Learning, thermo-electric devices, and green energy resources.