Scientific Research Publishing

Energy and Power Engineering, 2023, 15, 353-362
https://www.scirp.org/journal/epe
ISSN Online: 1947-3818
ISSN Print: 1949-243X

Utilizing the Vector Autoregression Model (VAR) for Short-Term Solar Irradiance Forecasting

Farah Z. Najdawi, Ruben Villarreal

Department of Material Science Engineering and Commercialization, Texas State University, San Marcos, USA
Email: fzn1@txstate.edu, rv14908@txstate.edu

How to cite this paper: Najdawi, F.Z. and Villarreal, R. (2023) Utilizing the Vector Autoregression Model (VAR) for Short-Term Solar Irradiance Forecasting. *Energy and Power Engineering*, **15**, 353-362.
https://doi.org/10.4236/epe.2023.1511020

Received: April 14, 2023
Accepted: November 3, 2023
Published: November 6, 2023

Abstract

Forecasting solar irradiance is a critical task in the renewable energy sector, as it provides essential information regarding the potential energy production from solar panels. This study aims to utilize the Vector Autoregression (VAR) model to forecast solar irradiance levels and weather characteristics in the San Francisco Bay Area. The results demonstrate a correlation between predicted and actual solar irradiance, indicating the effectiveness of the VAR model for this task. However, the model may not be sufficient for this region due to the requirement of additional weather features to reduce disparities between predictions and actual observations. Additionally, the current lag order in the model is relatively low, limiting its ability to capture all relevant information from past observations. As a result, the model's forecasting capability is limited to short-term horizons, with a maximum horizon of four hours.

Keywords

Vector Autoregression Model, Hyperparameter Parameters, Augmented Dickey Fuller, Durbin Watson's Statistics

1. Introduction

Solar irradiance is a crucial factor IN the efficient utilization of solar energy resources [1]. Accurate forecasting of solar irradiance is essential for ensuring the reliable and efficient operation of solar energy systems [2]. Solar irradiance represents the amount of solar radiation that reaches the earth's surface [3] and is influenced by various factors such as cloud cover, atmospheric conditions, and geographical location [4]. The availability of solar energy varies across different regions and seasons, and accurate solar irradiance forecasting can help optimize the design and operation of solar energy systems [5].

In recent years, significant progress has been made in developing models and techniques for forecasting solar irradiance, using a variety of data sources such as satellite imagery, ground-based measurements, and numerical weather prediction models [6].

These methods have shown promise in improving the accuracy of solar irradiance forecasting and can be used to provide valuable information for energy management and planning [7]. Alsharif *et al.*, developed a seasonal auto-regressive integrated moving average (SARIMA) model to forecast daily and monthly solar radiation in Seoul, South Korea based on 37 years of hourly data, and found that the auto-regressive integrated moving average (ARIMA) model could represent daily solar radiation while the seasonal ARIMA model could represent monthly solar radiation with expected average monthly solar radiation ranging from 176 to 377 W/m^2 [8]. Shadab *et al.*, use seasonal ARIMA models to forecast monthly solar radiation in the region around Delhi, India using remotely sensed insolation data, and generate monthly average insolation forecasts for the next four years [9]. The accuracy of the forecasts is evaluated, and potential regions for implementing efficient solar power generation projects are identified based on the insolation contours. Nwokolo *et al.*, discuss the importance of predicting and separating beam, diffuse, and global solar radiation for regions in Southern Africa and the Middle East, where access to reliable electricity is limited [10]. The paper proposes a hybrid evolutionary auto-regressive integrated moving average—Gumbel probabilistic (ARIMA-GP) models for predicting and separating these radiometric parameters, which outperforms other models tested. The proposed model can be used to improve solar power generation and support climate mitigation plans. The study also suggests that Gumbel probabilistic (GP) and ARIMA-GP models are more effective for separating beam (Hb) and diffuse (Hd) from global solar radiation (H) than empirical or machine learning models. Brahma & Wadhvani propose a new residual ensemble learning approach that uses advanced base models, Deep Neural Networks (DNNs) and Recurrent Neural Networks (RNNs), for solar irradiance forecasting, which is essential for efficient solar energy systems and sustainable power demand management [11]. The proposed approach consists of three modules that focus on data collection and analysis, feature selection, and the development of an accurate and robust forecast model. The proposed framework is validated using data from four different solar power sites and compared with other models. The results show that the proposed model improves forecast performance by approximately 2.5 percent in prediction error, making it a reliable solar irradiance prediction model. Cargan *et al.*, discuss the importance of accurate solar irradiance forecasting for grid operators, and the challenges associated with traditional time-series methods [12]. The study compares machine learning and deep learning models for forecasting solar irradiance using data from 20 UK locations and commercially available weather data. The authors propose a method that leverages weather measurements from other locations to accurately forecast solar irradiance and suggest that a single global model trained on multiple locations can produce more consistent and accurate results.

The purpose of this study is to employ the Vector Autoregression (VAR) model in forecasting the weather characteristics and solar irradiance levels in the San Francisco Bay Area, California. Accurately predicting solar irradiance is crucial in designing solar photovoltaic (PV) panels, and it requires accounting for the uncertainties in weather modeling. Hence, the VAR model is used to forecast the solar irradiance in this region. The VAR model is a machine-learning method that captures the interdependence of various variables as they evolve over time.

In contrast to the aforementioned models, our approach involves the integration of weather parameters such as atmospheric pressure, temperature, and relative humidity in conjunction with solar irradiance. Utilizing a Vector Autoregressive (VAR) model, we have successfully achieved the capability to predict hourly solar irradiance levels for the San Francisco Bay Area. Furthermore, our model simultaneously provides forecasts for additional meteorological variables, namely, dew temperature, relative humidity, and atmospheric pressure.

2. Methodology

Initially, the researchers obtained the weather dataset for the study area, covering eight years from 2014 to 2021, from the NSRDB portal in its raw form. It has the following weather features: Date in format (YYYY/MM/DD), hourly data recording for each day, actual solar irradiance, dew points, relative humidity, temperature, pressure, and solar zenith angle. The VAR model is solved using Python programming language.

The first step is to solve for VAR model. We clean the dataset to check for any columns containing non-values. It was found that no null entries were present in the dataset. The "minute" column was removed because it contains constant values throughout the data entries and that is not useful to us. Next, the dataset is very large. It contains 70,080 entries with most rows where the solar irradiance (I) value is zero. These are the hours of the day that lie between sunset and sunrise, hence solar power production is null, these rows will be removed to retain only those where solar irradiance is not equal to zero. After doing so, the dataset volume was reduced by more than half. It is now 32,748 entries. In order to observe how the above features (temperature, pressure…) affect solar irradiance, the exploratory data analysis was done, the authors correlate the data weather with solar irradiance and the preliminary results are shown in the next section. In order to use the VAR model, we need to first make sure that the data is stationary over time. We use Augmented Dickey Fuller (ADF) Test to check if the signal is stationary for every variable. The ADF is a unit root test for stationarity and tests for p-value. If the significant level (p-value) is less than 0.05, the null hypothesis is rejected and concludes that the time series is stationary. A transformation is performed by ADF test every time series to obtain stationary data, thus a p-value for each feature is measured and individual time series are differentiated until the p-value becomes smaller than 0.05. Also, hyperparameter optimization is performed to maximize the model performance. The hyperparameters are Akiake Information Criterion (AIC), Final Prediction Error (FPE), Bayesian Information Criterion (BIC) and Hannan-Quinn information criterion (HQIC). These criteria are used in model selection that provides the best tradeoff between goodness-of-fit and parsimony. Here are the mathematical representations for AIC, BIC, FPE, and HQIC respectively Hansen, and Brownlee [13] [14]:

$$\text{AIC} = -2\log L + 2k \tag{1}$$

$$\text{BIC} = -2\log L + k\log n \tag{2}$$

$$\text{FPE} = \frac{(n+k)}{(n-k-2)} \times \sigma^2 \tag{3}$$

$$\text{HQIC} = -2\log L + 2k\log(\log n) \tag{4}$$

where L is the maximum value of the likelihood function of the model; k is the number of parameters in the model; log is the natural logarithm; n is the number of observations; and σ^2 is the estimated variance of the residuals.

AIC penalizes models that contain an excess of parameters, and a lower AIC value signifies a more desirable fit. BIC also penalizes models that possess an excessive number of parameters and is more rigorous than AIC with respect to penalization. A better fit is indicated by the lower FPE values. HQIC is a modified variation of AIC that incorporates the sample size and applies a milder penalty on models with an excessive number of parameters than BIC. Before forecasting our outcomes, we examine if the residuals of the model still exhibit any remaining patterns by using Durbin Watson's statistics (DW).

The Durbin-Watson statistic is used to test whether this correlation is present and to what degree. The Durbin-Watson statistic ranges from 0 to 4, with a value of 2 indicating no autocorrelation. Values between 0 and 2 indicate positive autocorrelation, while values between 2 and 4 indicate negative autocorrelation. The closer the Durbin-Watson statistic is to 0 or 4, the stronger the evidence for autocorrelation Kenton [15]. Finally, we use the trained VAR model to do forecasting on the data. Hourly data from January 1, 2014 (12 midnight) to December 31st, 2021 were taken to train our model. Note that non-daylight hours, when the data of zero solar power production or solar irradiance were removed. Then we tested our model for 13:00, 14:00, 15:00, and 16:00 hours on December 31st, 2021. These hours are the last daylight hours as suggested by our dataset, **Figure 1** flowchart depicts the sequence of steps that are taken to construct the VAR model.

3. Results

This section utilizes Python programming to solve the VAR model and predict

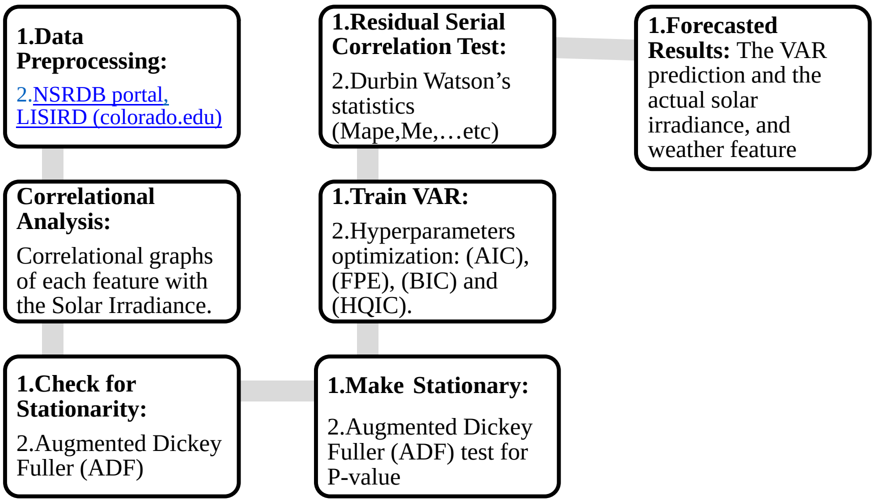

1.Data Preprocessing:
2.NSRDB portal, LISIRD (colorado.edu)
Correlational Analysis:
Correlational graphs of each feature with the Solar Irradiance.
1.Check for Stationarity:
2.Augmented Dickey Fuller (ADF)
1.Make Stationary:
2.Augmented Dickey Fuller (ADF) test for P-value
1.Train VAR:
2.Hyperparameters optimization: (AIC), (FPE), (BIC) and (HQIC).
1.Residual Serial Correlation Test:
2.Durbin Watson's statistics (Mape,Me,…etc)
1.Forecasted Results: The VAR prediction and the actual solar irradiance, and weather feature

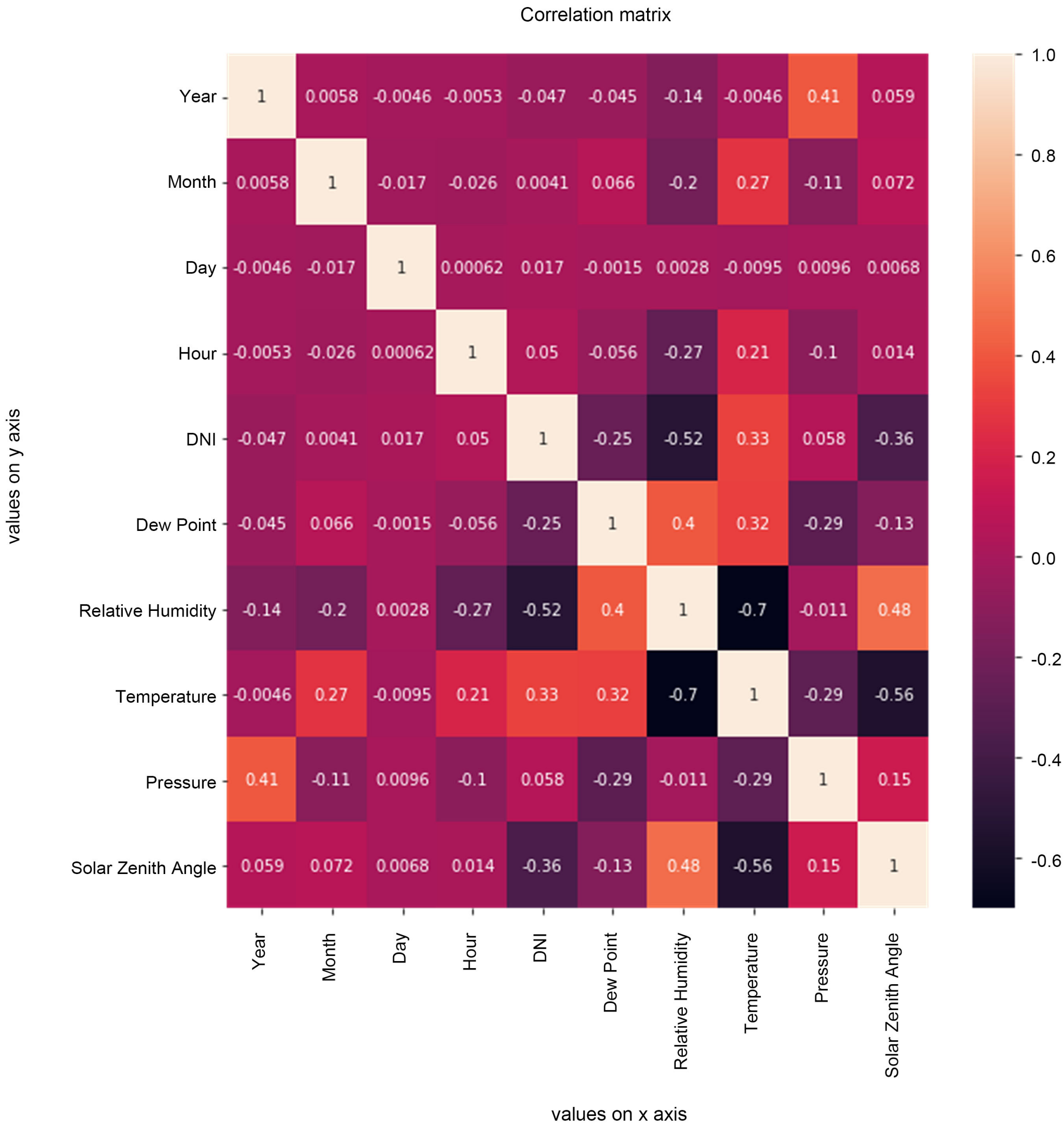
Figure 1. Flowchart depicts the sequence of steps to construct the VAR model.

solar irradiance values in the Bay Area of San Francisco. At the outset, the weather dataset obtained from the NSRDB portal was subjected to data preprocessing in order to clean it. The dataset was quite large, with 70,080 entries, most of which had solar irradiance values of zero. These entries corresponded to the hours between sunset and sunrise when solar power production is non-existent. Therefore, we removed these rows from the dataset, resulting in a reduction of more than half in size. The dataset comprises 32,748 entries that have non-zero solar irradiance values. In the second step of our analysis, we investigated the effect of weather variables on solar irradiance, as can be seen from the heat map in Figure 2 and Figure 3. The Heat Map displayed is utilized to analyze the relationship between each feature and solar irradiance. We found that dew point, humidity, and temperature displayed a significant relationship with solar irradiance, whereas pressure and hour of the day showed little to no correlation. Solar zenith angle exhibited a notable negative correlation. After careful consideration, we have identified the following features to be included in our final feature set with reasoning:

- Temperature: Exhibits a positive correlation coefficient of 0.33 with solar irradiance DNI.
- Pressure: Although it only has a correlation coefficient of 0.05, we have included it in our feature set as our set would become too small without it.
- Dew Point: Displays a negative correlation of −0.25 with solar irradiance.
- Relative humidity: Demonstrates a negative correlation of −0.52 with solar irradiance.
- Hour: Although the correlation between Hour of the Day and solar irradiance is only 0.04, we observed a relationship between the two factors from the heat map, and thus included it in our feature set.

We have chosen to exclude the solar zenith angle, despite its strong correlation with solar irradiance, as our focus is solely on weather conditions at this time.

To utilize the VAR model, a hyperparameter search is necessary to determine the optimal order lag *p*. The forecasting performance of a VAR model can be significantly influenced by its lag order, which determines the number of lags used for each variable. The model is then fitted based on the best values of AIC, BIC, FPE, and HQIC. The fact that the AIC and BIC values are very small suggests that the VAR model is well-suited for forecasting. The FPE values suggest that the selected lag order is relatively small compared to the number of samples available in the data. In this scenario, the 49th lag is chosen due to an inflection point where the FPE decreased from 48 to 49 and then remained constant from 49 to 50. As a consequence of the relatively small lag order, the model may not fully capture all the relevant information from past observations, leading to limited predictive capability for horizons beyond four hours.



	Year	Month	Day	Hour	DNI	Dew Point	Relative Humidity	Temperature	Pressure	Solar Zenith Angle
Year	1	0.0058	-0.0046	-0.0053	-0.047	-0.045	-0.14	-0.0046	0.41	0.059
Month	0.0058	1	-0.017	-0.026	0.0041	0.066	-0.2	0.27	-0.11	0.072
Day	-0.0046	-0.017	1	0.00062	0.017	-0.0015	0.0028	-0.0095	0.0096	0.0068
Hour	-0.0053	-0.026	0.00062	1	0.05	-0.056	-0.27	0.21	-0.1	0.014
DNI	-0.047	0.0041	0.017	0.05	1	-0.25	-0.52	0.33	0.058	-0.36
Dew Point	-0.045	0.066	-0.0015	-0.056	-0.25	1	0.4	0.32	-0.29	-0.13
Relative Humidity	-0.14	-0.2	0.0028	-0.27	-0.52	0.4	1	-0.7	-0.011	0.48
Temperature	-0.0046	0.27	-0.0095	0.21	0.33	0.32	-0.7	1	-0.29	-0.56
Pressure	0.41	-0.11	0.0096	-0.1	0.058	-0.29	-0.011	-0.29	1	0.15
Solar Zenith Angle	0.059	0.072	0.0068	0.014	-0.36	-0.13	0.48	-0.56	0.15	1

Figure 2. The heat map of correlation between the weather data and solar irradiance.

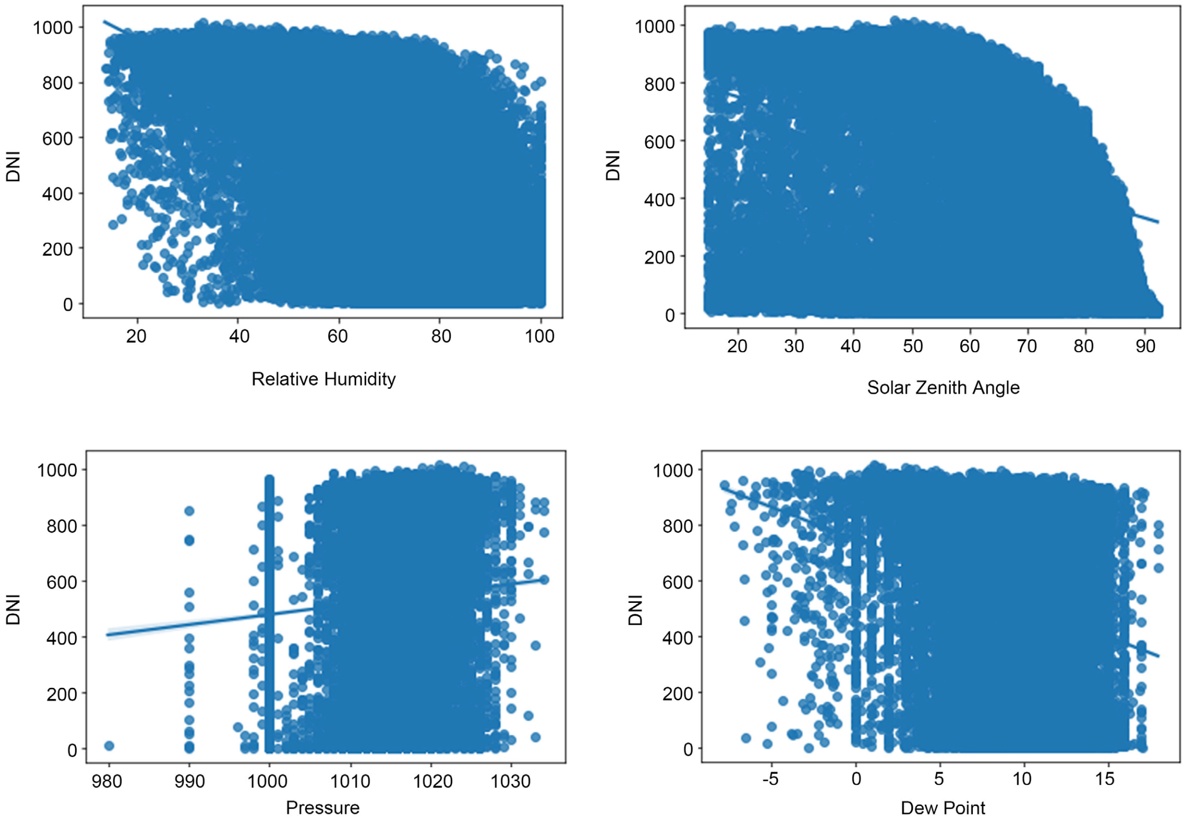


Figure 3. The correlation between the weather data and solar irradiance.

To utilize the Vector Autoregressive Model (VAR), it is imperative to ensure that the data is stationary beforehand. Therefore, Durbin Watson's statistics test is applied to check if there is a correlation left leftover in residual results. Our results indicate that all values are in proximity to 2, suggesting the absence of remaining correlation. Table 1 shows the result of predicting solar irradiance (*I*) and weather conditions (dew point, relative humidity, temperature, and pressure) for four-time instances on December 31st, 2021, at hourly intervals from 1:00 PM to 4:00 PM. For the first time point (1:00 PM), the solar irradiance was high (786 W/m^2), and the dew point, relative humidity, temperature, and pressure were 4.4˚C, 59.57%, 12.3˚C, and 1012 kPa, respectively. For the second time point (2:00 PM), the solar irradiance decreased to 603 W/m^2, while the weather conditions remained similar to the previous time point. For the third time point (3:00 PM), the solar irradiance further decreased to 291 W/m^2, and the dew point and temperature remained similar, but the relative humidity increased to 59.15% and the pressure remained constant at 1011 kPa. For the fourth time point (4:00 PM), the solar irradiance decreased to 154 W/m^2, while the dew point temperature and pressure remained similar to the previous time point. The weather conditions also showed a decreasing trend in temperature and relative humidity, while the dew point and pressure remained constant.

Table 2 shows the performance metrics of the model for predicting the following variables: *I*, dew point, relative humidity, and pressure.

Table 1. The VAR prediction and the actual solar irradiance, and weather feature.

Hour	I_{act} w/m^2	I_{pred} w/m^2	$T_{(D\,act)}$ (˚C)	$T_{(D\,pred)}$ (˚C)	Φ_{act} (%)	Φ_{pred} (%)	T_{act} (˚C)	T_{act} (˚C)	P_{act} (kPa)	P_{pred} (kPa)
13	863	786	4.7	4.4	61.93	59.57	11.8	12.3	1012	1012
14	804	603	4.9	4	63.47	57.74	11.6	12.8	1012	1011
15	676	291	5.6	3.8	67.33	59.15	11.4	12.8	1012	1011
16	344	154	7.4	3.8	76.08	67.02	11.5	11.6	1020	1011

Table 2. The statistical errors result for each weather feature and solar irradiance.

Statistical parameter	I (W/m^2)	Dew Point (˚C)	Relative humidity (%)	Temperature (˚C)	Pressure (Kpa)
Mape	0.589106	0.263855	0.092241	0.069331	0.002700
Me	−290.250000	−1.650000	−6.332500	0.800000	−2.750000
Mae	290.250000	1.650000	6.332500	0.800000	2.750000
Mpe	−0.589106	−0.263855	−0.092241	0.069331	−0.002700
Rmse	332.625540	2.067607	6.844679	0.956556	4.555217
Correlation Coefficient	0.984106	−0.690522	0.918542	−0.077331	−0.333333
Minmax	0.589106	0.263855	0.092241	0.063099	0.002700

Since the MAPE values are relatively low for all features, it shows that the model is good. Experimental MAE values are relatively high for all features, indicating that there is still some error in the predictions. In this case, the RMSE values are relatively high for all features, indicating that there is still some error in the predictions. In this case, the features have a strong positive correlation with DNI, a moderate negative correlation with dew point, a strong positive correlation with Relative Humidity, a weak negative correlation with Pressure, and a weak positive correlation with temperature. Furthermore, Table 2 result clearly shows that the model has good performance, with low MAPE, MAE, and RMSE values, high correlation, and small min-max values. The values for ME and MPE are also generally close to zero, indicating that the model has little bias. The correlation coefficient indicates that there is a strong relationship between the prediction and the actual model. Overall, The VAR model is an effective tool for forecasting solar irradiance, but it may not suffice for the San Francisco area as additional weather features are required to minimize the disparities between prediction and actual observation as shown in Table 1 on December 31st, 2021. The current lag order (FPE) in the model is relatively low, which may limit its ability to capture all the relevant information from past observations. As a result, the model may not be able to predict longer forecasting horizons, which may be evident in its limited ability to forecast beyond a 4-hour horizon. This is because VAR models rely on the autoregressive structure of time series data, meaning that they rely on past values of the same variables to predict future values. As the forecasting horizon extends, the dependency on historical observations becomes progressively more complex. Longer horizons require more accurate predictions of multiple lagged observations, which may not be achievable with a VAR model. Also, the presence of a large 7-year dataset with numerous null values poses a challenge in sourcing adequate historical data for VAR modeling. The lack of past data can make it tough for VAR models to make long horizon predictions.

4. Conclusion

The correlation coefficient and MAPE served as our statistical benchmark, indicating that the VAR model is a reliable tool for forecasting both weather characteristics and solar irradiance. However, our study suggests that incorporating additional weather features or obtaining datasets with fewer null values could improve the model's performance and extend its forecasting horizon beyond four hours. In the future, we plan to compare the VAR model with other available models, especially those capable of predicting both solar irradiance and weather characteristics simultaneously, to further validate the model's accuracy. LIU *et al.*, in 2018 used the VAR model to predict the temperature, solar radiation, and wind speed at 61 locations around the United States [16]. The findings underscore the suitability of their proposed time series approach for very short-term forecasts for six hours of hourly solar radiation, temperature, and wind speed. In the evaluation of model performance, the authors utilized the Mean Absolute Percentage Error (MAPE), which yielded values ranging from 6% to 80% for various meteorological variables, including temperature, wind speed, and solar irradiance. This study introduced a notably higher level of accuracy and improved performance, as indicated by the calculated MAPE results, which were recorded as follows: 0.589106 for solar irradiance, 0.263855 for dew point temperature, 0.0922 for relative humidity, 0.069331 for temperature, and 0.002700 for pressure. However, it is important to note that our model was designed specifically for a 4-hour forecasting horizon for these parameters (temperature, pressure, etc.).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Dhanraj, J.A., *et al.* (2021) An Effective Evaluation on Fault Detection in Solar Panels. *Energies*, **14**, Article 7770. https://doi.org/10.3390/en14227770

[2] Kumar, D.S., *et al.* (2020) Solar Irradiance Resource and Forecasting: A Comprehensive Review. *IET Renewable Power Generation*, **14**, 1641-1656. https://doi.org/10.1049/iet-rpg.2019.1227

[3] Gutiérrez-Trashorras, A.J., *et al.* (2018) Attenuation Processes of Solar Radiation. Application to the Quantification of Direct and Diffuse Solar Irradiances on Horizontal Surfaces in Mexico by Means of an Overall Atmospheric Transmittance. *Renewable and Sustainable Energy Reviews*, **81**, 93-106.

https://doi.org/10.1016/j.rser.2017.07.042

[4] Noia, M., Ratto, C.F. and Festa, R. (1993) Solar Irradiance Estimation from Geostationary Satellite Data: I. Statistical Models. *Solar Energy*, **51**, 449-456. https://doi.org/10.1016/0038-092X(93)90130-G

[5] Kumari, P. and Toshniwal, D. (2021) Deep Learning Models for Solar Irradiance Forecasting: A Comprehensive Review. *Journal of Cleaner Production*, **318**, Article 128566. https://doi.org/10.1016/j.jclepro.2021.128566

[6] Diagne, M., *et al.* (2013) Review of Solar Irradiance Forecasting Methods and a Proposition for Small-Scale Insular Grids. *Renewable and Sustainable Energy Reviews*, **27**, 65-76. https://doi.org/10.1016/j.rser.2013.06.042

[7] Narvaez, G., *et al.* (2021) Machine Learning for Site-Adaptation and Solar Radiation Forecasting. *Renewable Energy*, **167**, 333-342. https://doi.org/10.1016/j.renene.2020.11.089

[8] Alsharif, M.H., Younes, M.K. and Kim, J. (2019) Time Series ARIMA Model for Prediction of Daily and Monthly Average Global Solar Radiation: The Case Study of Seoul, South Korea. *Symmetry*, **11**, Article 2. https://doi.org/10.3390/sym11020240

[9] Shadab, A., Ahmad, S. and Said, S. (2020) Spatial Forecasting of Solar Radiation Using ARIMA Model. *Remote Sensing Applications: Society and Environment*, **20**, Article 100427. https://doi.org/10.1016/j.rsase.2020.100427

[10] Nwokolo, S.C., *et al.* (2022) Hybridization of Statistical Machine Learning and Numerical Models for Improving Beam, Diffuse and Global Solar Radiation Prediction. *Cleaner Engineering and Technology*, **9**, Article 100529. https://doi.org/10.1016/j.clet.2022.100529

[11] Brahma, B. and Wadhvani, R. (2023) A Residual Ensemble Learning Approach for Solar Irradiance Forecasting. *Multimedia Tools and Applications*, **82**, 33087-33109. https://doi.org/10.1007/s11042-023-14616-6

[12] Cargan, T., Landa-Silva, D. and Triguero, I. (2023) Local-Global Methods for Generalised Solar Irradiance Forecasting. https://arxiv.org/abs/2303.06010

[13] Hansen, B. (2017) Vector Autoregressions. The University of Wisconsin-Madison. https://www.ssc.wisc.edu/~bhansen/460/460Lecture25%202017.pdf

[14] Brownlee, J. (2020) Probabilistic Model Selection with AIC, BIC, and MDL. *Machine Learning Mastery*. https://machinelearningmastery.com/probabilistic-model-selection-measures/

[15] Kenton, W. (2023) Durbin Watson Test: What It Is in Statistics, With Examples. Investopedia. https://www.investopedia.com/terms/d/durbin-watson-statistic.asp

[16] Liu, Y., Roberts, M.C. and Sioshansi, R. (2018) A Vector Autoregression Weather Model for Electricity Supply and Demand Modeling. *Journal of Modern Power Systems and Clean Energy*, **6**, 763-776. https://doi.org/10.1007/s40565-017-0365-1