Journal of Data Analysis and Information Processing, 2020, 8, 282-294
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

# Identifying Extreme Rainfall Events Using Functional Outliers Detection Methods

**Mohanned Abduljabbar Hael[1,2], Yongsheng Yuan[2]**

[1]Department of Statistics and Informatics, Taiz University, Taiz, Yemen
[2]College of Science, Hohai University, Nanjing, China
Email: Mohanned2020@taiz.edu.ye

**How to cite this paper:** Hael, M.A. and Yuan, Y. (2020) Identifying Extreme Rainfall Events Using Functional Outliers Detection Methods. *Journal of Data Analysis and Information Processing*, **8**, 282-294.
https://doi.org/10.4236/jdaip.2020.84016

**Received:** September 19, 2020
**Accepted:** November 2, 2020
**Published:** November 5, 2020



## Abstract

Outlier detection techniques play a vital role in exploring unusual data of extreme events that have a critical effect considerably in the modeling and forecasting of functional data. The functional methods have an effective way of identifying outliers graphically, which might not be visible through the original data plot in classical analysis. This study's main objective is to detect the extreme rainfall events using functional outliers detection methods depending on the depth and density functions. In order to identify the unusual events of rainfall variation over long time intervals, this work conducts based on the average monthly rainfall of the Taiz region from 1998 to 2019. Data were extracted from the Tropical Rainfall Measuring Mission and the analysis has been processed by R software. The approaches applied in this study involve rainbow plots, functional highest density region box-plot as well as functional bag-plot. According to the current results, the functional density box-plot method has proven effective in detecting outlier compared to the functional depth bag-plot method. In conclusion, the results of the current study showed that the rainfall over the Taiz region during the last two decades was influenced by the extreme events of years 1999, 2004, 2005, and 2009.

## Keywords

Rainfall Data, Outlier Detection, Rainbow Plot, Functional Bag-Plot, Functional Box-Plot

## 1. Introduction

Outlier detection approaches help in identifying characteristics that might have been neglected when using classical statistics and mathematical models. This area of study has much attention when analyzing data in a functional context.

[1] introduced the theoretical foundations and methodologies about functional data analysis with many applications besides [1] explained the characteristics of a functional form of a continuous variable over age or time. Also, many studies of several non-/parametric methods for the functional data analysis were extended by [2]. In the state-of-the-art literature on outlier detection in high dimensional data and functional data, various methods were presented by [3] [4] [5].

Mostly, in the functional setting of outlier detection methods, [6] introduced useful graphical tools for visualizing univariate cases of functional data and detecting functional outliers. These proposed graphical methods, including functional bagplot and functional highest density region (HDR) plots, are presented to detect outlier functional data graphically. The enhanced functional boxplot was proposed by [7]. This tool can identify outliers graphically, which might not be visible through the original data plot and visualizing functional data. Moreover, [8] demonstrated some methods for visualizing large amounts of functional time series data and detecting outliers for functional data.

Therefore, several functional methods of outlier detection have been successfully carried out in many different applications. For instance, in the depth-based approach, [9] [10] identified abnormal nitrogen oxides levels of the set of curves based on functional depth measures. Besides [11] applied the functional depth method to the detection of outliers in gas emissions from urban areas. In depth and density approaches, [12] used functional outlier detection techniques such as rainbow plots and the functional bag-plot and box-plot for examining the outliers of the daily flood-flow series. In recent, [13] applied different functional methods such as rainbow plots to visualize large amounts of hydrological data and the functional/bivariate bag-plot and box-plot for the detection of outliers graphically. In the density-based approach, [14] employed the functional highest density region box-plot to identify functional outliers in streamflow hydrograph of flood data.

Above all, our interest in the current work will focus on applying the functional outlier detection methods, which have been proposed by [6] [8]. These approaches have high-speed computing in revealing functional outliers as well as their ability to visualize the behavior of unusual observations.

In the analysis of rainfall data, the outlier detection techniques play a critical role in examining data, and detecting the extreme points that may affect the accuracy of the results. The functional framework could be seen as an appropriate method to obtain additional insight compared to classical analyses. Moreover, it enables the comprehensive analysis of rainfall data by performing one analysis of the entire data graphics rather than the several analyses. Rainfall is a critical climatological scale and needs to be precisely analyzed. The rainfall rate change is one of the critical weather matters that effect on Taiz region in Yemen. Therefore, it is crucial to conduct this study due to the urgent need to introduce a suitable statistical approach by functional techniques to study the entire rainfall graphics. The present paper may help researchers in conducting future studies.

The results could also make a significant contribution to the effective planning of water resources management.

The primary objective of this study is to detect the extreme events of rainfall data over the Taiz region from 1998 to 2018 using the functional (HDR) box-plot and functional bag-plot. Besides, this study adapts rainbow plots for visualizing the temporal rainfall patterns based on the depth and density orderings.

The current study is organized into four main sections. In the first section, the introduction and literature review are stated. In the second section, methods and materials are described. Results and discussion are presented in the third section. Conclusions are reported in the last section.

## 2. Methods and Materials

This section gives brief information about the rainbow approach used to visualizing data graphically. This section mainly provides the methodology used to detect outliers, particularly functional bagplot, and functional boxplot that will be conducted in the current study.

### 2.1. Rainbow Approach

The rainbow approach is a visual representation of all the functional data. The only additional features are a rainbow color palette depending on a time-default order of the data.

#### 2.1.1. Depth Functions

In the procedure of depth functions, data can be ordered according to Tokay's half-space location depth, and the bivariate principal component score depth is defined by:

$$OT_i = D(s_i, S), \quad S = \left\{ s_j \in \mathbb{R}^2, j = 1, \cdots, N \right\} \tag{1}$$

where $D(.,.)$ is the half-space depth function; this depth function is given as the smallest number of data points involved in a closed half-space containing on its border. Then, the curves are ordered decreasingly based on their depth values, and the first ordered curve represents the median curve. In contrast, the last curve in a sample of curves can be considered the external curve.

#### 2.1.2. Density Functions

In the procedure of density functions, the observations can also be ordered by highest density regions, and the kernel density estimate at the bivariate principal component scores is given by:

$$OD_i = \hat{f}(s_i) = \frac{1}{n} \sum_{j=1}^{N} \frac{1}{b_j} K\left( \frac{s_i - s_j}{b_j} \right), \ i \neq j, \ i = 1, \cdots, N \tag{2}$$

where $b_j$ is the bandwidth for the *jth* bivariate score points $\{s_j\}$ and $K$ is the kernel function. The functional data $y_i(t)$ are ordered in decreasing order with respect to $OD_i$. Therefore, the modal curve is represented by the first or-

dered curve with the highest OD. In contrast, the last ordered curves can be considered the most unusual curves with the lowest OD. The curves are drawn with colors based on the values of $OT_i$ and $OD_i$, so that the most outlying curves are violet while the curves that close to the center are red.

## 2.2. Functional Outlier Detection Methods

Examining extreme values is an essential step in analyzing data before modeling and forecasting. The outliers may affect the accuracy of results. Functional outlier detection techniques facilitate the investigation and identification of specific outliers' specific properties that may not be apparent with summary statistics or classical approaches. [6] proposed two graphical approaches, which firstly visualizing functional data by the rainbow plots, after that identifying functional outliers through the functional bag-plot and the functional highest-density region (HDR) box-plot. The robust functional principal component analysis is used to decompose functional data into the first two functional principal components and their scores $S_i = \left(s_{i,1}, s_{i,2}\right)$; the first functional principal component scores are given as:

$$s_{i1} = \int \xi_1(t)\left[x_i(t) - \bar{x}(t)\right]\mathrm{d}t \tag{3}$$

And the second functional principal component scores are given as:

$$s_{i2} = \int \xi_2(t)\left[x_i(t) - \bar{x}(t)\right]\mathrm{d}t \tag{4}$$

where $\bar{x}(t)$ is the mean function of a sample of curves $x_i(t)$. The $\xi_1(t)$ and $\xi_2(t)$ are the first two functional principal components, respectively. The functional bag-plot and bag-plot approaches are computationally faster and more able to detect outliers, as explained in the next sections.

### 2.2.1. Functional Bag-Plot Method

According to Tukey's half-space location depth, the bivariate principal component scores are ordered and plotted in a simple two-dimensional graph as given in formula (1). Therefore, the functional bag-plot is considered as a map of the bivariate bag-plot of the first and second principal component scores to the functional curves. It displays the inner and outer regions and the median curve of Tukey's. The inner region is assigned as the region bordered by all curves corresponding to the points in the bivariate bag and includes 50% of the observed curves. The outer region is achieved by distention the inner region by a factor $\alpha$ that could take the values 1.96 or 2.58 in order to cover either 95% or 99% of the curves, respectively. The colors of detected bivariate outliers are matched to the same colors of the functional outliers, where all the points outside the outer region are identified as outliers.

### 2.2.2. Functional Highest Density Region (HDR) Box-Plot Method

The highest density regions order the bivariate principal component scores. These scores are the construct of a two-dimensional kernel density estimate,

where the bandwidths are chosen by a technique based on [15]. The bivariate HDR boxplot is constructed using the bivariate kernel density estimates $\hat{f}(s_i)$ which are given in formula (2). The highest density region (HDR) is defined as

$$R_\alpha = \left\{ s : \hat{f}(s_i) \geq f_\alpha \right\} \tag{5}$$

where $f_\alpha$ is such, $\int_{R_\alpha} \hat{f}(s_i) \mathrm{d}s = 1-\alpha$, that is the region with coverage probability $(1-\alpha)$; all points within the covered region have a higher density estimate than any of the points outside the highest density region. In consideration of the bivariate density with an expanding coverage as $\alpha$ decreases, the highest density regions can be represented as contours. The bivariate HDR boxplot displays the 50% inner and 99% or 95% outer highest density regions, and the mode highest density point defined as

$$\max \hat{f}(s_i) \tag{6}$$

Then all points excluded from the outer HDR are identified as outliers. The functional Highest Density Region (HDR) boxplot is a map of the bivariate HDR boxplot of the first and second robust principal component scores to the functional curves. It displays the inner and outer regions, and the highest density modal curve. The inner region is determined as the region bordered by all curves corresponding to points inside the 50% bivariate HDR; therefore, 50% of curves are in the inner region. The outer region is bounded by entire curves corresponding to the points within the outer bivariate HDR, almost 95% or 99% of the curves are in the outer region. The colors of bivariate outliers match the same colors of functional outliers, where curves outside the outer functional HDR region are considered outliers.

### 2.3. An Application to Rainfall Data

The study area is the Taiz region, which is considered one of the largest cities located in the southwest of Yemen at the geographical coordinates of 13˚34'46"N 44˚01'15"E [16]. Its average height from sea level is about 1311 m. The mean annual rainfall is approximately between 800 - 1200 mm, while the mean monthly evaporation is almost 140 mm. The primary data source of average monthly rainfall over the Taiz region, Yemen, during the period from January 1998 to December 2018 were obtained as satellite data from Tropical Rainfall Measuring Mission, https://giovanni.gsfc.nasa.gov. In the present dataset, we have $N = 21$ discrete measurements $y_i(t_j)$, $t_j \in [1,12]$, $i = 1,\cdots,21$, the discrete observation $y_i(t_j), j = 1,\cdots,12$ denotes rainfall records for the *ith* year and the *jth* month. In this study, rainfall data will be processed by using the R software rainbow package for visualizing functional time series data and detecting outliers [8] besides the rainbow package developed by [17].

## 3. Results and Discussion

This section is divided into three main sections. Rainbow plots for visualizing

rainfall data will be displayed in the first section. Detecting outliers using functional and bivariate bag-plots will be demonstrated in the second section. The third section will discuss the results of detecting outliers using the functional and bivariate HDR box-plots.

### 3.1. Visualizing Rainfall Data by Rainbow Plots

The rainbow plot represents the points of all the data in a single plot having a distinct feature of color palette following the order of rainbow colors. By default, time order is followed by the lines of rainbow plot in a way which line of the recent past appears in violet. In contrast, the line belongs to the remote past is appear in red. Figure 1 displays the rainbow plot of rainfall data for all years from 1998 to 2018, which are ordered by time. The curves from the distant past years are shown in red (1998); the curves from the middle years are shown in green (2008); while the most recent years are shown in violet (2018).

In Figure 2, rainbow plots of rainfall data are drawn based on the depth and density order indexes in addition to the colors following the order of the rainbow that is reflecting the ordering. In the depth order, as shown in Figure 2 (top panel), the red lines show the curves with a higher depth closer to the center of the data. In contrast, the outlying curves with the lower depth are represented by the violet lines. The black line describes the median curve. Regarding the density ordering, the isolated and outlying curves have a lower density. Then these curves are distinguished in the violet lines as given in Figure 2 (bottom panel).

On the contrary, the higher density curves are distinguished in the red lines, and the black line describes the modal curve. The red curves are mostly ambiguous, while the violet curves are clearly visualized. However, the majority of the data are superimposed over curves. Results show that both methods lead to a different ordering, particularly for the years associated with high or low ordering such as 1999, 2004, 2005, and 2014, as drawn in Figure 2.

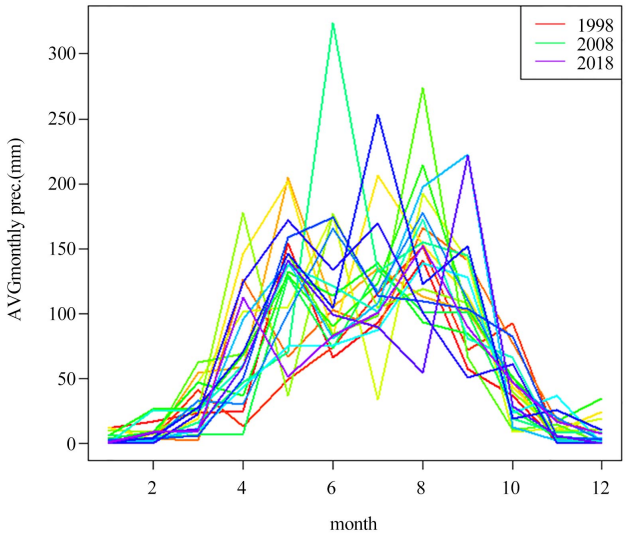


Figure 1. Rainbow plot with time ordering for years 1998-2018.

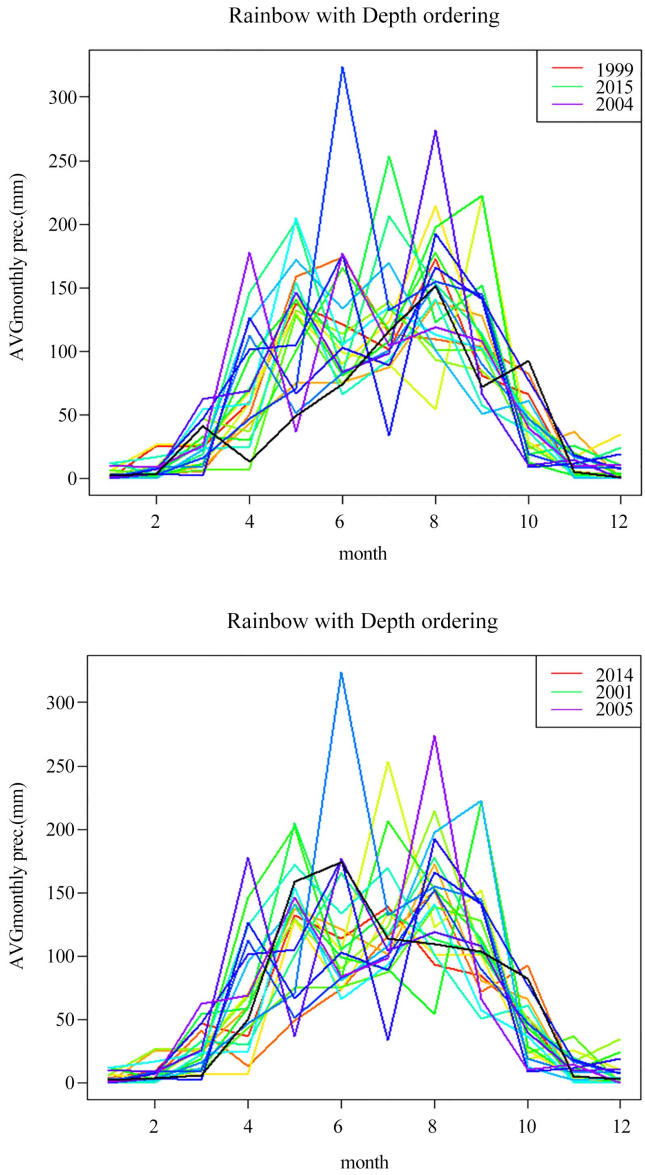


**Figure 2.** Rainbow plots with depth ordering (top panel) and density ordering (bottom panel).

### 3.2. Detecting Outliers by Functional (Bivariate) Bag-Plot

In **Figure 3**, the principal components scores of bivariate bag-plot and the functional bag-plot with 95% of probability coverage are drawn graphically. **Figure 3(a)** displays the bivariate bag-plot scores of the rainfall data, with 95% probability coverage. We can see the dark and light gray regions that exhibit the bag and fence regions of the bivariate bag-plot. The bag is specified as the smallest depth region, including at least 50% of the observations. The bag-plot's outer region is the convex hull of the region obtained by inflating the bag using a factor $\alpha = 1.96$ that cover 95% of probability of estimated curves. Distinctly, the red asterisk is the Tukey median. All the individual points outside the fence regions are defined as outliers.

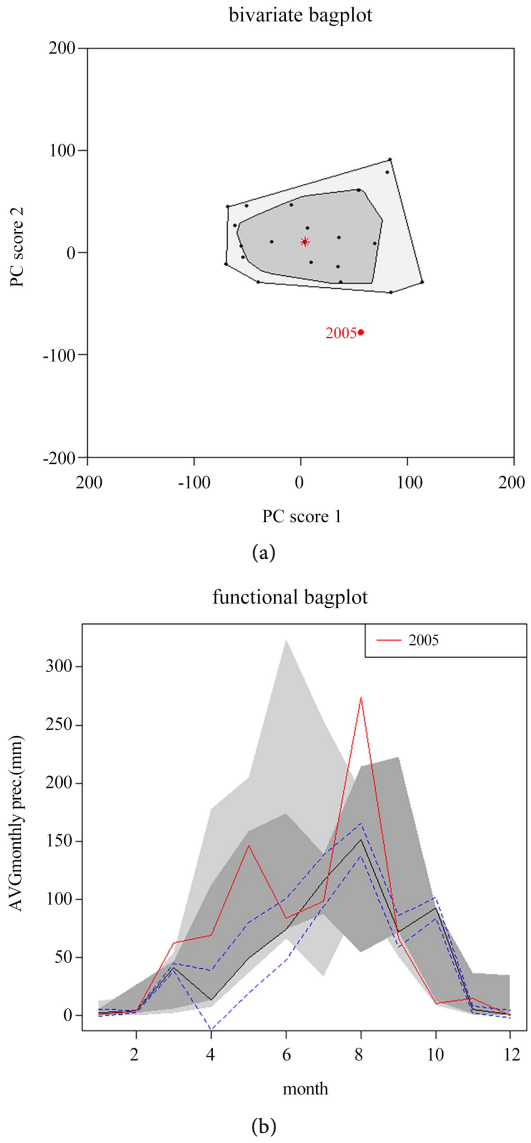


**Figure 3.** The bivariate bag-plot (a), and the functional bag-plot (b), with 95% of probability coverage.

Figure 3(b) displays the corresponding functional bag-plot with 95% probability coverage. The dark and light gray regions show the bag and fence regions of the functional bag-plot, respectively. The black line is the median curve, which is surrounded by 95% point-wise confidence intervals, as shown in blue, and the curves outside the fence regions are defined as outliers. In Figure 3(a), the point corresponding to 2005 is located outside the outer bivariate bag-plot region. It corresponds to the red curve in the associated functional bag-plot in Figure 3(b). Then this year is detected as an outlier according to Tukey's depth.

The rainfall rates were the greatest during the end of the summer season in the Taiz region, especially in August. The outlier curve corresponding to 2005 is discriminated by a very high peak that seems to correspond to a heavy rainfall event.

### 3.3. Detecting Outliers by Functional (Bivariate) HDR Box-Plot

The depth method has failed to recognize outliers of the rainfall data; it means that insufficiently distant from the median curve. However, the HDR box-plot method correctly identifies outliers of the curves with the highest density from the modal curve. The functional bag-plot may be a convenient technique to detect outliers when outliers are at a considerable distance from the median curve. Therefore, the depth method is unable to identify outliers accurately when outliers are near the median. In this situation, the functional HDR box plot is more suitable and gives better results. In Figure 4 and Figure 5, the bivariate and functional HDR box-plots for rainfall data with 95% probability coverage according to two stages are presented.

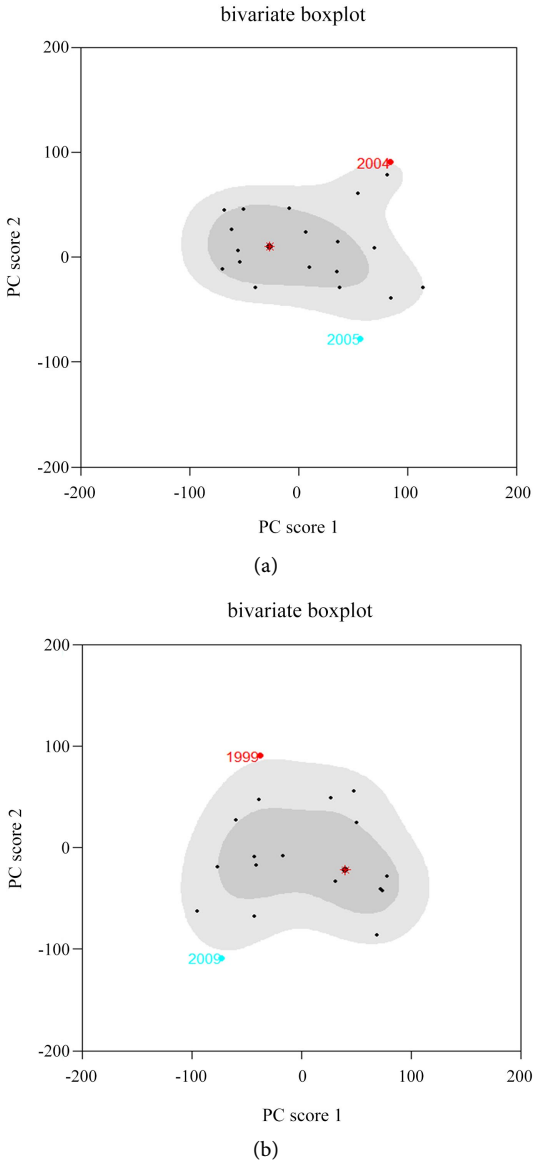


Figure 4. The bivariate HDR box-plots in first stage (a) and in second stage (b), with 95% of probability coverage.

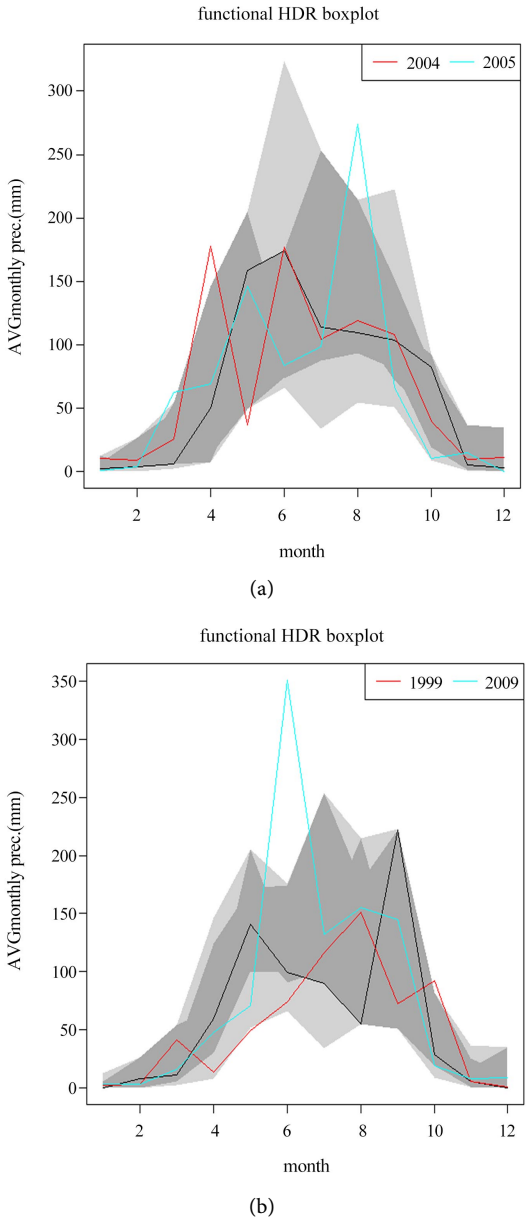


**Figure 5.** The functional HDR box-plots in the first stage (a) and in the second stage (b), with 95% of probability coverage.

**Figure 4** displays the bivariate box-plot scores with 95% probability coverage; the light and dark gray parts present the fence and box regions, respectively. The red asterisk is the mode, and the points outside the fence regions are identified as outliers. **Figure 5** displays the corresponding functional HDR box-plot with 95% probability coverage; the light and dark gray parts present the fence and box regions, respectively. The modal curve is drawn in the black line. The curves outside the fence regions are detected as functional outliers.

In **Figure 4(a)**, the point corresponding to the years 2005 and 2004 are outside the outer bivariate HDR bag-plot regions that correspond to the red and green curves in the associated functional HDR bag-plot as drawn in **Figure 5(a)**.

Hence, these years are considered outliers according to the highest density regions. Besides, in Figure 4(b), the points corresponding to the years 2009 and 1999 are outside the outer bivariate HDR bag-plot region. They match the red and green curves in the associated functional HDR bag-plot, respectively, as drawn in Figure 5(b). Hence, these years are considered outliers according to the highest density regions.

Apparently, in the first stage, the detected outliers are 2005 and 2004, outside the HDR box-plot outer region. In the second stage, after excluding the outliers detected in the first stage, the detected outliers are 2009 and 1999 outside the HDR box-plot outer region. From Figure 5, it is concluded that the notable curves of 2005 and 2009 are significantly different from the general shape of curves and the location curves. Actually, on the basis of the rainfall rate, the curves of 2005 and 2009 are distinctive by a very high peak and different sizes that seem to correspond to rainy years since the rainfall amount was the greatest over the Taiz region in June and August during the summer season.

In comparison with previous studies specifically with similar works, this study differs slightly from previous studies in terms of using the functional highest density region (HDR) box-plot method based on a two-stage procedure, as shown in Figure 4 and Figure 5. Unlike the previous studies, the presented approaches in the current study have been adapted to detect outliers for rainfall data.

Generally, the results of the present study are thoroughly consistent with previous studies that applied the same methodology in different types of data. Table 1 summarizes the procedures and methods used in the current study and compare it with previous related studies.

Table 1. Comparison of the procedures used in the current study with the previous studies.

| Applications | Visualization tool | Methods | | Ref. |
|---|---|---|---|---|
| | | Depth-based | Density-based | |
| Rainfall data | Rainbow plots with depth and density ordering | Functional bag-plot method | Functional (HDR) box-plot method with a two-stage | current |
| Flood data | - | - | Functional (HDR) box-plot method | [14] |
| Hydrological data | Rainbow plots with depth and density ordering | Functional bag-plot method | Functional (HDR) box-plot method | [13] |
| Hydrograph streamflow | Rainbow plots by depth and density estimates | Functional bag-plot method | Functional (HDR) box-plot method | [12] |
| Fertility rates & Sea surface temperature | Rainbow plots with depth and density ordering | Functional bag-plot method | Functional (HDR) box-plot method | [8] |
| Mortality rates & Sea surface temperature | Rainbow plots with depth and density ordering | Functional bag-plot method | Functional (HDR) box-plot method | [6] |
| Gas emissions | - | Functional depth | - | [11] |
| Nitrogen oxides emissions | - | Functional depth | - | [10] |

In short, the functional outlier detection methods performed in this study, particularly functional (HDR) box-plot, made it easier to identify the extreme events of rainfall over the Taiz region and effectively detect unusual curves. Moreover, these approaches have demonstrated the temporal rainfall patterns in the Taiz region during the past two decades, influenced by many different outlier events due to the monsoonal effects.

## 4. Conclusions

In this study, rainfall data analyzed using functional outliers detection approach through a case study based on 21 years recorded during the period from January 1989 to December 2018 over the Taiz region in Yemen. The practical methods included rainbow plots according to time order, depth and density functions, outlier detection by functional highest density region box-plot, and functional bag-plot. The study demonstrated significant results for extreme rainfall events over the region during the last two decades. The functional procedures of outlier detection can be more effective and preferable than classical ones; moreover, these methods are sophisticated and adaptable in climate studies such as rainfall change. Besides, the approaches employed in this study have a fast computation in detecting outliers and graphical representing the actual phenomena that performed a great variety of structure of rainfall data.

Based on the current results, the functional density box-plot method is more reliable and more effective in detecting outlier than the functional depth bag-plot method. The functional detected outliers, such as years in 1999, 2004, 2005, and 2009 were identified as real observations and had different shapes and magnitudes compared to the other observed curves. Therefore, the detected outliers events were the most significant rainfall over Taiz region due to the convective storms of high intensity, and the monsoonal effects. Visibly the years 2005 and 2009, were the most outliers above the mean curve; the years 2004 and 1999 were the most oscillated under the mean curve.

In addition, the functional results gave more information concerning the rainfall regime over the Taiz region by adding continuous temporal aspects. The entirely functional results showed that there are significant variations around rainfall rates. Overall, the variability was very high during the summer season, high in spring, moderate in autumn, and less or none in winter. May and September were the most rainfall rates due to the monsoonal affections. In future perspectives, it is recommended to apply the most advanced functional approaches to outlier detection. Furthermore, the appropriate functional methods should be used in modeling and forecasting rainfall with other climatic variables.

## Conflicts of Interest

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

## References


[1] Ramsay, J.O. (2005) Functional Data Analysis (Springer Series in Statistics). Springer, New York. https://doi.org/10.1007/b98888

[2] Ferraty, F. and Vieu, P. (2006) Nonparametric Functional Data Analysis: Theory and Practice (Springer Series in Statistics). Springer, New York.

[3] Filzmosera, P., Maronna, R. and Werner, M. (2008) Outlier Identification in High Dimensions. *Computational Statistics & Data Analysis*, **52**, 1694-1711. https://doi.org/10.1016/j.csda.2007.05.018

[4] López-Pintado, S. and Romo, J. (2009) On the Concept of Depth for Functional Data. *Journal of the American Statistical Association*, **104**, 718-734. https://doi.org/10.1198/jasa.2009.0108

[5] Narisetty, N.N. and Nair, V.N. (2016) Extremal Depth for Functional Data and Applications. *Journal of the American Statistical Association*, **111**, 1705-1714. https://doi.org/10.1080/01621459.2015.1110033

[6] Hyndman, R.J. and Shang, H.L. (2010) Rainbow Plots, Bagplots, and Boxplots for Functional Data. *Journal of Computational and Graphical Statistics*, **19**, 29-45. https://doi.org/10.1198/jcgs.2009.08158

[7] Sun, Y. and Genton, M.G. (2011) Functional Boxplots. *Journal of Computational and Graphical Statistics*, **20**, 316-334.

[8] Shang, H.L. (2011) Rainbow: An R Package for Visualizing Functional Time Series. *The R Journal*, **3**, 54-59. https://doi.org/10.32614/RJ-2011-019

[9] Febrero, M., Galeano, P. and González-Manteiga, W. (2007) A Functional Analysis of $NO_x$ Levels: Location and Scale Estimation and Outlier Detection. *Computational Statistics*, **22**, 411-427. https://doi.org/10.1007/s00180-007-0048-x

[10] Febrero, M., Galeano, P. and González-Manteiga, W. (2008) Outlier Detection in Functional Data by Depth Measures, with Application to Identify Abnormal $NO_x$ Levels. *Environmetrics*, **19**, 331-345. https://doi.org/10.1002/env.878

[11] Martínez Torres, J., *et al.* (2010) Detection of Outliers in Gas Emissions from Urban Areas Using Functional Data Analysis. *Journal of Hazardous Materials*, **186**, 144-149. https://doi.org/10.1016/j.jhazmat.2010.10.091

[12] Chebana, F., Dabo-Niang, S. and Ouarda, T.B.M.J. (2012) Exploratory Functional Flood Frequency Analysis and Outlier Detection. *Water Resources Research*, **48**, W04514. https://doi.org/10.1029/2011WR011040

[13] Hussain, I. (2019) Outlier Detection Using Graphical and Nongraphical Functional Methods in Hydrology. *International Journal of Advanced Computer Science and Applications*, **10**, 438. https://doi.org/10.14569/IJACSA.2019.0101259

[14] Suhaila, J. (2019) Application of Functional Data Analysis in Stream Flow Hydrograph. *Proceedings of the Third International Conference on Computing, Mathematics and Statistics* (*iCMS*2017), Springer, Berlin, 245-251. https://doi.org/10.1007/978-981-13-7279-7_30

[15] Hyndman, R.J. (1996) Computing and Graphing Highest Density Regions. *The American Statistician*, **50**, 120-126. https://doi.org/10.1080/00031305.1996.10474359

[16] Hael, M.A. (2020) Modeling of Rainfall Variability Using Functional Principal Component Method: A Case Study of Taiz Region, Yemen. *Modeling Earth Systems and Environment*. https://doi.org/10.1007/s40808-020-00876-w

[17] Shang, H.L. and Hyndman, R. (2019) Rainbow: Rainbow Plots, Bagplots and Boxplots for Functional Data. R Package Version 3.6. https://CRAN.R-project.org/package=rainbow