Scientific Research Publishing

Atmospheric and Climate Sciences, 2025, 15(3), 645-667
https://www.scirp.org/journal/acs
ISSN Online: 2160-0422
ISSN Print: 2160-0414

Machine Learning-Based Outlier Detection in Long-Term Climate Data: Evidence from Burkina Faso's Synoptic Network

Zamantakonè Guillaume Ki, Wenceslas Somda, Marcel Bawindsom Kébré, Soumaila Gandema, François Dabilgou

Laboratoire de Matériaux et Environnement (LAME), Université Joseph Ki-Zerbo, Ouagadougou, Burkina Faso
Email: zamantakoneky@gmail.com

How to cite this paper: Ki, Z.G., Somda, W., Kébré, M.B., Gandema, S. and Dabilgou, F. (2025) Machine Learning-Based Outlier Detection in Long-Term Climate Data: Evidence from Burkina Faso's Synoptic Network. *Atmospheric and Climate Sciences*, **15**, 645-667.
https://doi.org/10.4236/acs.2025.153032

Received: May 25, 2025
Accepted: July 5, 2025
Published: July 8, 2025

Abstract

In recent decades, the impact of climate change on natural resources has increased. However, the main challenges associated with the collection of meteorological data include the presence of missing, outlier, or erroneous data. This work focuses on outliers detection in long-term climate data by using machine learning models. The study uses meteorological data collected over 40 years (1981-2021) from ten synoptic stations operated by Burkina Faso's National Meteorological Agency (ANAM). The methodology is based on the use of 18 machine learning algorithms from the PyOD library, including probabilistic, linear, proximity-based, and ensemble models. Univariate and multivariate analyses are performed. For the multivariate analysis, this paper focuses on two key variables, maximum temperature and minimum relative humidity which consistently exhibit strong correlations across all stations. A robust approach is adopted to optimize the detection of outliers, using thresholds based on extreme percentiles. The results show that models such as KPCA, LSCP, LOF, and Feature Bagging are best suited to capturing anomalies in complex time series. These results will contribute to more reliable climate analyses and improved modeling of extreme climate events in data-scarce regions.

Keywords

Machine Learning, Climate Data, Anomaly Detection, Burkina Faso, PyOD

1. Introduction

Outlier detection, also known as anomaly detection, refers to the task of identifying data points that significantly diverge from the majority and may signal mean-

ingful anomalies in a given context, otherwise, focuses on identifying abnormal data points that do not follow the general distribution of a dataset [1]. It is a key research challenge in various fields due to the diversity of data structures and the complexity of real-world phenomena. The objective is to determine whether certain observations deviate from expected patterns, by being abnormally high or low, and are therefore considered outliers. Data point that conform to the general trend are not typically the focus of such analyses, which concentrate instead on unusual or exceptional values.

In the context of time series, these anomalies can appear in various forms, such as point anomalies, contextual anomalies, or collective anomalies, depending on whether the deviation is observed in isolation, in specific temporal contexts, or in groups [2]. Proper identification and classification of anomalies in time-dependent data are essential for ensuring data integrity and improving the robustness of downstream analyses [3]-[5].

In climate studies, anomaly detection has gained increasing importance due to its potential to reveal critical deviations from typical weather or environmental patterns. Climate outliers are often linked to extreme events such as droughts, floods, heatwaves, or unexpected temperature drops, which can lead to significant human and environmental consequences [6]. As climate variability increases and new extremes emerge, it becomes crucial to detect deviations from long-term climate norms to support reliable climate modeling and disaster risk management. While climate prediction models aim to capture general cyclical behavior, outlier detection helps identify and isolate unusual patterns or inconsistencies, including sensor faults and extreme meteorological events [4]-[6].

Over the years, various techniques have been developed to detect climate-related anomalies, including statistical approaches (e.g., z-score, percentiles, Density-Based Spatial Clustering of Applications with Noise-DBSCAN), time series models (e.g., moving averages, ARIMA), machine learning techniques (e.g., Isolation Forest, One-Class SVM), and hybrid methods that consider multivariate relationships, seasonal patterns, or historical data comparisons [3] [5]-[7]. These methods have been applied with some success, although they often suffer from limitations in scalability, long-term accuracy, and sensitivity to data sparsity, which may reduce their reliability when dealing with large-scale or incomplete climatic datasets. Consequently, the more robust approaches, such as machine learning techniques, are increasingly being explored to overcome these constraints and improve anomaly detection performance by capturing the complex temporal dependencies and seasonal patterns inherent in climate data [6].

As [8] and [9] emphasized, the choice of anomaly detection method should depend on the nature of the dataset and the problem at hand. While statistical models offer transparency and simplicity, machine learning approaches are often more flexible and better suited to capturing nonlinear and high-dimensional patterns. These methods are increasingly applied in combination to optimize the detection process [4] [5] [10] [11].

This study aims to explore and apply machine learning techniques for outlier detection in long-term meteorological data. The dataset consists of daily observations collected over 40 years (1981-2021) from synoptic weather stations in Burkina Faso. Given the extended time frame and potential for both seasonal variability and data inconsistencies, this study seeks to distinguish between erroneous measurements and true climatic anomalies, comparing different machine learning methods for time series anomaly detection. The results are expected to contribute to more reliable climatic data analysis and improved modeling of extreme weather events in West African contexts.

2. Methodology

In this study, we use 18 machine learning methods from the PyOD library to detect anomalies in our dataset among various methods in the literature for time series data in general [3] [11] or climate data [4] [12]. PyOD is a library designed for outlier detection in datasets. It provides access to more than 50 different algorithms for outlier detection and is compatible with both Python 2 and 3 [13] [14]. PyOD is a widely used Python library for anomaly detection in multivariate data. It has been successfully applied in various domains, including astrophysics, where [15] used it to detect multimessenger transient events. Additionally, [16] integrated PyOD into PyODDS, an automated anomaly detection system, demonstrating its adaptability across different fields.

We first conducted a univariate analysis using the Isolation Forest (IF) method to identify an overall trend for each climatic parameter regarding the size of outliers or erroneous data in the dataset. Next, considering the interrelationship between meteorological parameters, a multivariate analysis was conducted. As a reminder, a univariate outlier is a data point that consists of extreme values in a single variable, whereas a multivariate outlier is an unusual combined score across at least two variables. Multivariate analysis takes multiple weather parameters into account simultaneously (e.g., temperature, humidity, wind speed) and detects anomalies using multivariate statistical methods or machine learning models [7].

Isolation Forest detects anomalies by isolating observations. It builds binary trees (called iTrees) by recursively partitioning the data using random split thresholds. Since anomalies are easier to isolate than normal instances, they tend to have a lower average path length.

The anomaly score used in the Isolation Forest algorithm is defined as [17]:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}} \tag{1}$$

where $E(h(x))$ is the average path length to isolate observation x, and $c(n)$ is the average path length in a binary search tree of size n. In scikit-learn, the decision_function used returns:

$$\text{decision\_function}(x) = E(h(x)) - c(n) \tag{2}$$

A negative value indicates a likely outlier, while a positive value suggests normal

behavior.

One of the limitations of the PyOD library lies in its reliance on domain knowledge for effective use. For instance, models like Isolation Forest require manual specification of the contamination or the fraction of outliers—the expected proportion of anomalies in the dataset. By default, this parameter is set to 0.1 (or 10%) [17], which may not be suitable for all applications without prior data insight. This value may lead to either an overestimation or an underestimation for our dataset. In the absence of deeper meteorological insight, determining an appropriate outlier fraction remains challenging. We then demonstrate the impact of this parameter when using the Isolation Forest implementation from the scikit-learn package. The method allows setting the proportion of outliers within a range of 0 to 50%, offering flexibility depending on the characteristics of the dataset.

Table 1 below presents the various methods used for anomaly detection in this study. [11], [18] and [19] provide further details on each of these methods. These algorithms are categorized into different methodological approaches: probabilistic models, linear models, proximity-based methods, and ensemble models. Probabilistic models, such as ECOD, COPOD, KDE, and Sampling estimate the data distribution and identify anomalies as observations with extremely low probabilities, while the linear models like PCA, KPCA, MCD, and OCSVM leverage linear transformations or hyperspherical separations to detect outliers. Proximity-based methods, where we count LOF, HBOS, kNN, AvgKNN, MedKNN, and ROD assess local density or distance to neighboring points to assign an anomaly score. Ensemble models (e.g., IForest, INNE, FB, LSCP) combine multiple approaches to enhance robustness in anomaly detection.

Table 1. Summary of anomaly detection methods used.

	Abbr	Algorithm	Type	Supervision Type
1	ECOD	Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions	Probabilistic	Unsupervised
2	COPOD	COPOD: Copula-Based Outlier Detection	Probabilistic	Unsupervised
3	KDE	Outlier Detection with Kernel Density Functions	Probabilistic	Unsupervised
4	Sampling	Rapid distance-based outlier detection via sampling	Probabilistic	Unsupervised
5	PCA	Principal Component Analysis (sum of weighted projected distances to eigenvector hyperplanes)	Linear Model	Unsupervised
6	KPCA	Kernel Principal Component Analysis	Linear Model	Unsupervised
7	MCD	Minimum Covariance Determinant (Mahalanobis distance as outlier score)	Linear Model	Unsupervised
8	OCSVM	One-Class Support Vector Machines	Linear Model	Unsupervised
9	LOF	Local Outlier Factor	Proximity-Based	Unsupervised
10	HBOS	Histogram-based Outlier Score	Proximity-Based	Unsupervised

Continued

11	kNN	k Nearest Neighbors (distance to the kth nearest neighbor as outlier score)	Proximity-Based	Supervised
12	AvgKNN	Average kNN (average distance to k nearest neighbors as outlier score)	Proximity-Based	Supervised
13	MedKNN	Median kNN (median distance to k nearest neighbors as outlier score)	Proximity-Based	Supervised
14	ROD	Rotation-based Outlier Detection	Proximity-Based	Unsupervised
15	IForest	Isolation Forest	Outlier Ensembles	Unsupervised
16	INNE	Isolation-based Anomaly Detection Using Nearest-Neighbor Ensembles	Outlier Ensembles	Unsupervised
17	FB	Feature Bagging	Outlier Ensembles	Unsupervised
18	LSCP	LSCP: Locally Selective Combination of Parallel Outlier Ensembles	Outlier Ensembles	Unsupervised

Additionally, these algorithms differ in their supervision type: most operate in an unsupervised manner, while some kNN-based variants require labeled data (supervised). This methodological diversity allows for a comprehensive evaluation of anomalies within our dataset. Although most studies in the climate domain face the problem of data set unavailability, the lack of labeled data [6], as well as expensive and scarce data.

3. Study Area and Dataset

The study is conducted in Burkina Faso, a West African country with predominantly flat terrain and an average elevation of around 300 meters (**Figure 1**). The country has a tropical climate, alternating between a dry season (November to May) and a wet season (June to October), with rainfall increasing from north (≈600 mm/year) to south (≈1200 mm/year) [20].

Meteorological data were collected from ten synoptic stations operated by the Burkina Faso National Meteorological Agency (ANAM). These stations are distributed across the country to capture regional climatic variability. The dataset spans 41 years (1981-2021) and includes daily measurements of precipitation (mm), evapotranspiration (mm), minimum, average, and maximum air temperatures (˚C), minimum, average, and maximum relative humidities (%), wind speed at 2 meters above ground (m/s), wind direction at 2 meters above ground (degrees), sunshine duration (hours), and daily global solar radiation on a horizontal plane (J/cm^2/day).

4. Assessment of Missing Data and Preprocessing Approach

Figure 2 provides a visual overview of missing values across the dataset, highlighting significant gaps prior to 1996, likely linked to technical limitations and non-automated data collection during that period.

The visualization reveals frequent data gaps, especially before 1996, with notable missingness in wind-related variables (WS-MEAN and WD-MEAN), global radiation (GRAD), and some inconsistencies in precipitation (RAIN). These patterns guided the selection of variables for subsequent analysis. The heatmap, shown in Figure 3, reveals significant missing data heterogeneity across stations and variables. Notably, GRAD and ETP exhibit high missing rates in several locations, with Bogandé showing the most extensive data gaps. In contrast, core temperature and humidity variables are well preserved across most stations. As we can observe in Figure 4, most stations exhibit over 90% data availability, with Ouagadougou reaching nearly complete coverage. In contrast, Bogandé stands out with substantial missing data, retaining only 63.3% of observations.

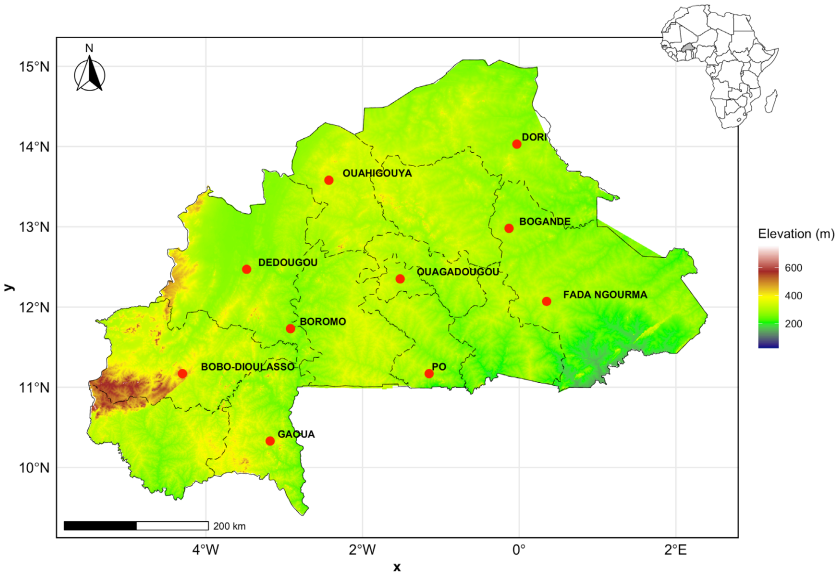


Figure 1. Location of the study area within Africa and spatial distribution of the ten synoptic stations used in this study. Elevation is derived from a Digital Elevation Model (DEM) and is represented in meters. The map highlights the topographic gradient and the geographic spread of stations used for data collection (1981-2021).

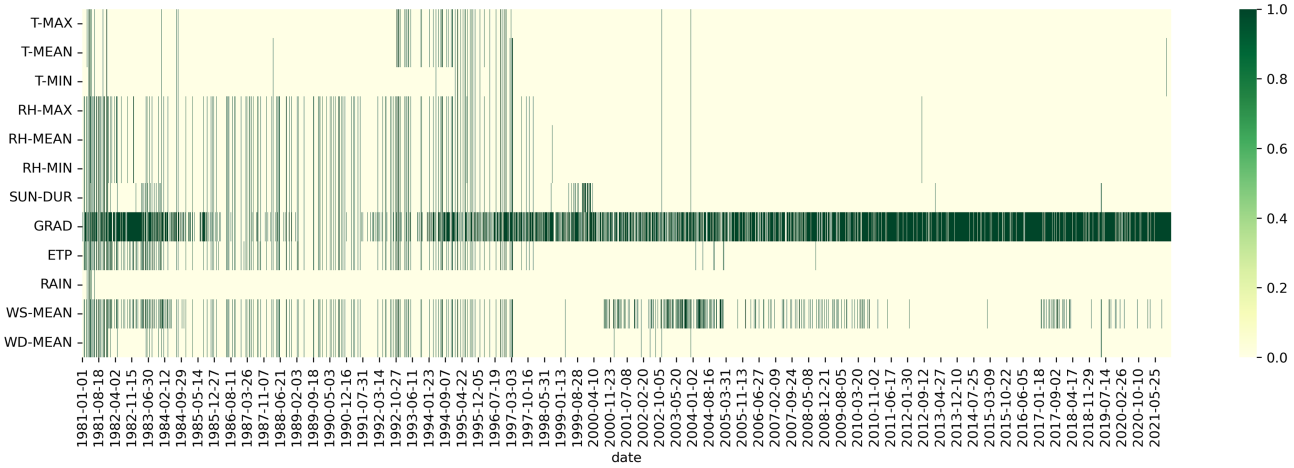


Figure 2. Heatmap of missing data across meteorological variables (1981-2021).

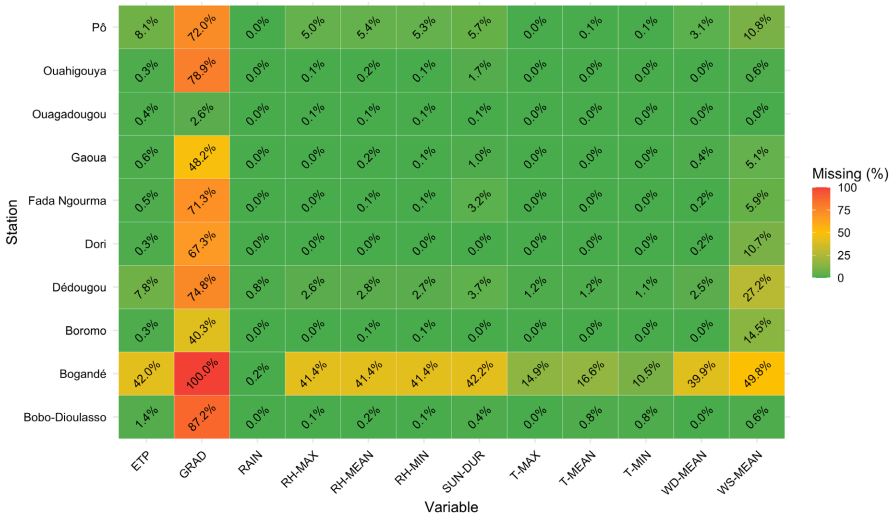


Figure 3. Heatmap of missing data proportions by station and climate variable (1981-2021).

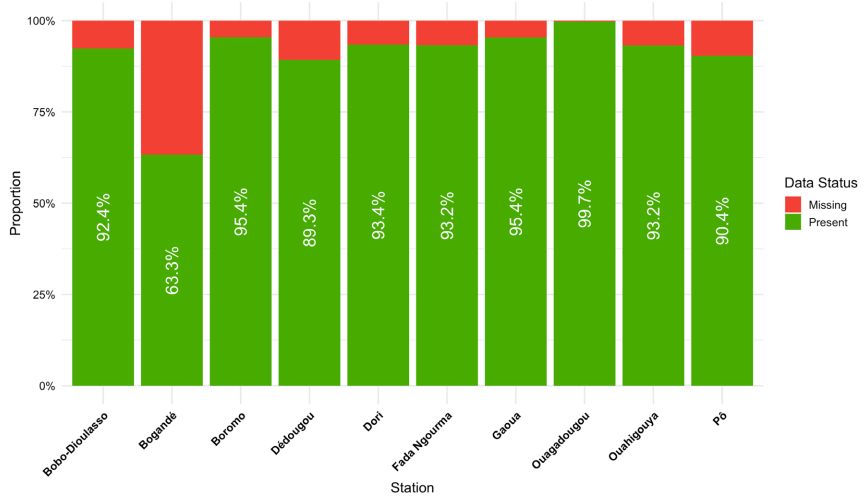


Figure 4. Proportion of present and missing data per meteorological station (1981-2021).

Following this missing data analysis and given the sensitivity of machine learning algorithms to missing values, the Bogandé station with consistently incomplete records was excluded from the initial analysis with the Isolation Forest method. The Bogandé station will be the subject of a separate study, as it serves as the gateway between the Sahelian and Sudanian zones, and high-quality data is crucial for related research. After this selection, an imputation method based on interpolation, which previously demonstrated satisfactory results in similar contexts [21], was applied to the remaining variables. This preprocessing step ensured a completer and more reliable dataset for subsequent machine learning analyses.

5. Results and Discussions

5.1. Visual Assessment of Climate Data Distribution and Outlier Structure

To initiate the outlier detection process, violin plots are employed to explore the

distribution of meteorological variables across stations over the study period. This format not only emphasizes the central tendency of the data but also reveals the kurtosis of the distributions, making it possible to visually assess the degree of data flattening and the potential presence of outliers [17]. These graphical insights will serve as a foundation for a more detailed investigation once outliers have been formally detected. In the following, violin plots are presented for selected groups of meteorological variables, including temperature, humidity, precipitation, and wind characteristics. In these figures, red dots indicate the mean values, while the shape of each violin reflects the density and spread of the data. Asymmetries in the violin plots and the elongation of lower or upper tails highlight the presence of extreme values, reflecting either climatic variability across stations, including potential extreme weather events, or erroneous values that should be treated as missing data.

So, **Figure 5** presents the distribution of daily temperature data across the study period, allowing for a visual comparison of the mean values for T-MIN, T-MEAN, and T-MAX across the meteorological stations. Overall, minimum temperatures (T-MIN) display a relatively narrow range across stations, with average values generally between 21˚C and 23˚C. The distributions are fairly symmetrical and compact, suggesting limited variability and few extreme low-temperature events. Mean temperatures (T-MEAN), by contrast, range from approximately 27.8˚C to 29.9˚C, with the highest averages observed in Dori, Fada N'Gourma, and Ouagadougou, which are located in warmer climatic zones. The violin shapes reveal moderate dispersion, especially in the lower tails, hinting at occasional deviations. Maximum temperatures (T-MAX) show the greatest variability, with averages spanning from around 33.5˚C in Bobo-Dioulasso to above 37˚C in Dori. The broader upper tails in several stations, notably Dédougou and Dori, indicate the presence of extreme heat events and highlight regions more exposed to temperature extremes. This comparative view across the three temperature indicators reinforces the need to account for station-specific variability in subsequent anomaly detection procedures.

The distribution of relative humidity values (RH-MIN, RH-MEAN, RH-MAX) across stations shows distinct patterns compared to temperature, as shown in **Figure 6**. These violin plots reveal a strongly asymmetric distribution, particularly for RH-MIN and RH-MEAN, with pronounced elongation in the lower tails. This suggests a high frequency of very low humidity values, especially in stations like Dori and Ouahigouya, which are located in drier climatic zones. In contrast, RH-MAX values are more compact and concentrated near the upper bounds (above 60% - 80%), indicating that maximum humidity tends to cluster at high values across most stations. These shapes also point to a greater presence of outliers on the lower end, especially for minimum humidity, which is consistent with the nature of extreme dry episodes in semi-arid environments.

The violin plots of annual rainfall (RAIN) and annual potential evapotranspi-

ration (ETP), see Figure 7, reveal distinct patterns across stations. While RAIN shows a wide range of annual totals, with higher values observed in Gaoua and Bobo-Dioulasso, and much lower amounts in Dori, the overall distributions remain continuous and relatively well-centered. In contrast, the distributions for ETP in the stations of Pô and Dédougou appear highly compressed and distorted. This is consistent with earlier missing data diagnostics (see Figure 1 and Figure 3), which showed that these two stations exhibit more than 10% of missing ETP values during the period 1981-1996. Such a rate significantly biases the shape of the distribution and underrepresents interannual variability.

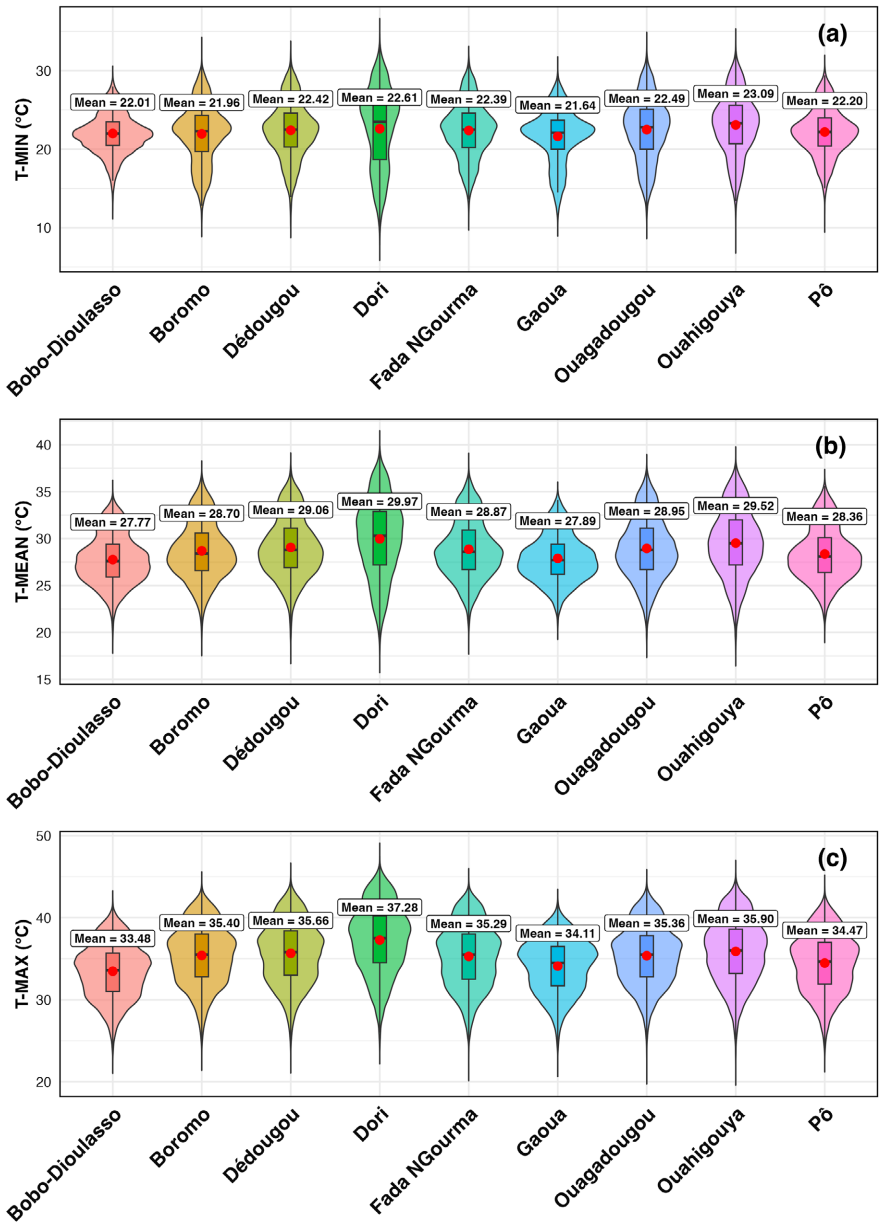


Figure 5. Distribution of daily temperatures data across the study period: (a) Minimum temperature (T-MIN); (b) Mean temperature (T-MEAN); (c) Maximum temperature (T-MAX).

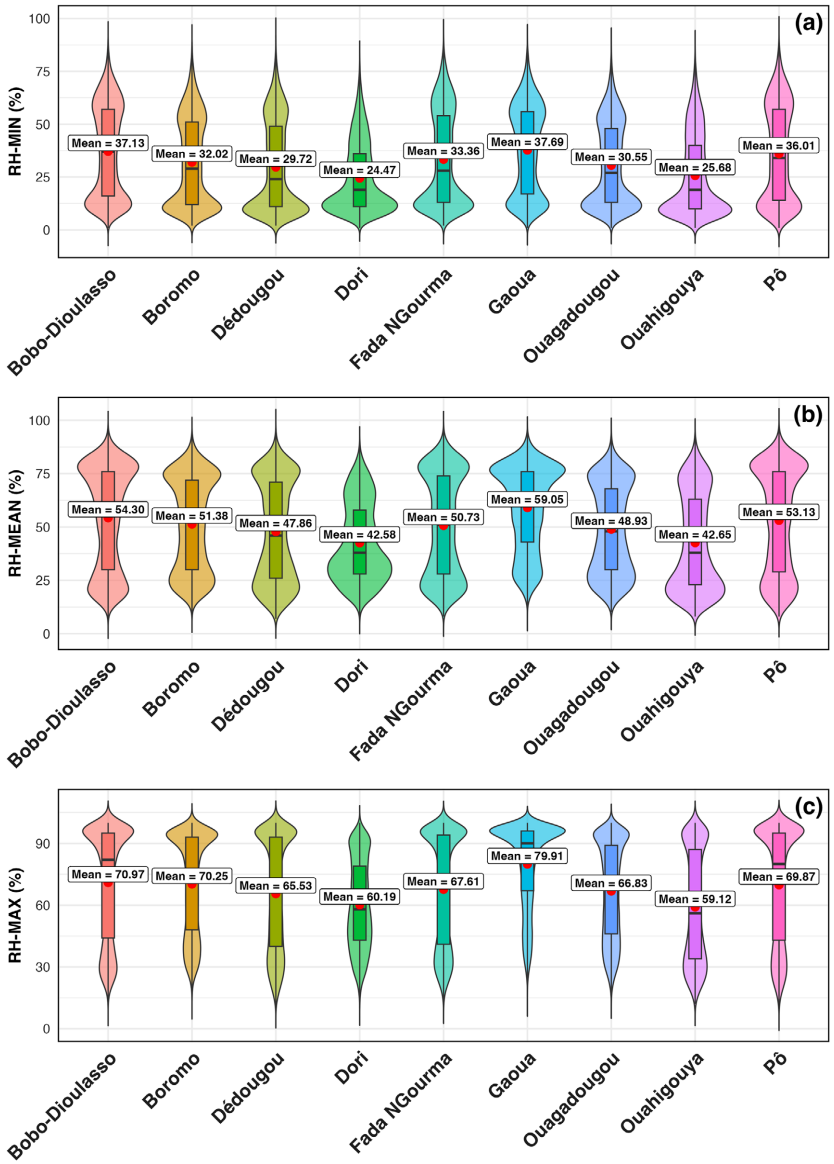
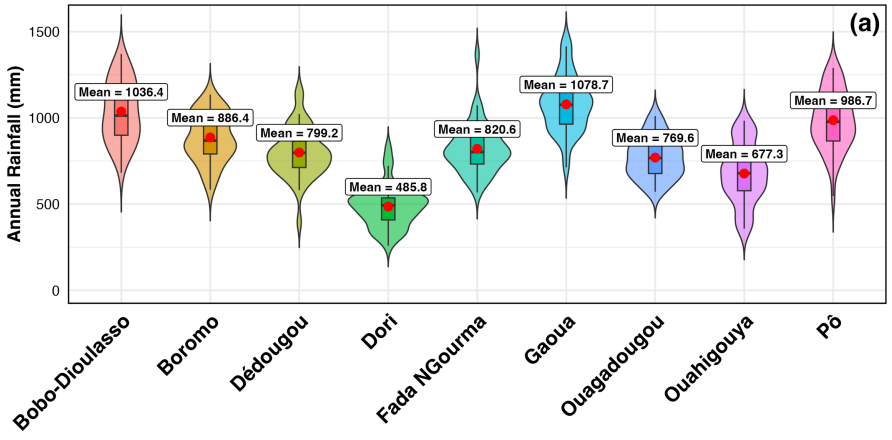


Figure 6. Distribution of daily relative humidity (RH) levels across stations over the study period: (a) Minimal RH (RH-MIN); (b) Mean RH (RH-MEAN); (c) Maximum RH (RH-MAX).



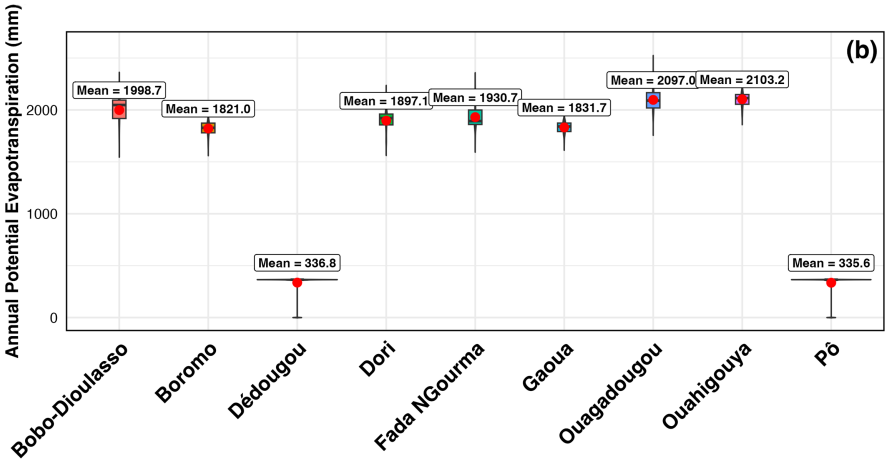


Figure 7. Distribution of annual rainfall (a) and Annual potential evapotranspiration (b) across stations over the study period.

A similar issue is observed for wind speed (WS-MEAN), where Pô and Dédougou again show highly compressed and skewed distributions with unrealistically low mean values (Figure 8). These distortions further confirm the impact of missing data on the interpretation of the climatic behavior at these stations and justify their cautious treatment in subsequent analyses.

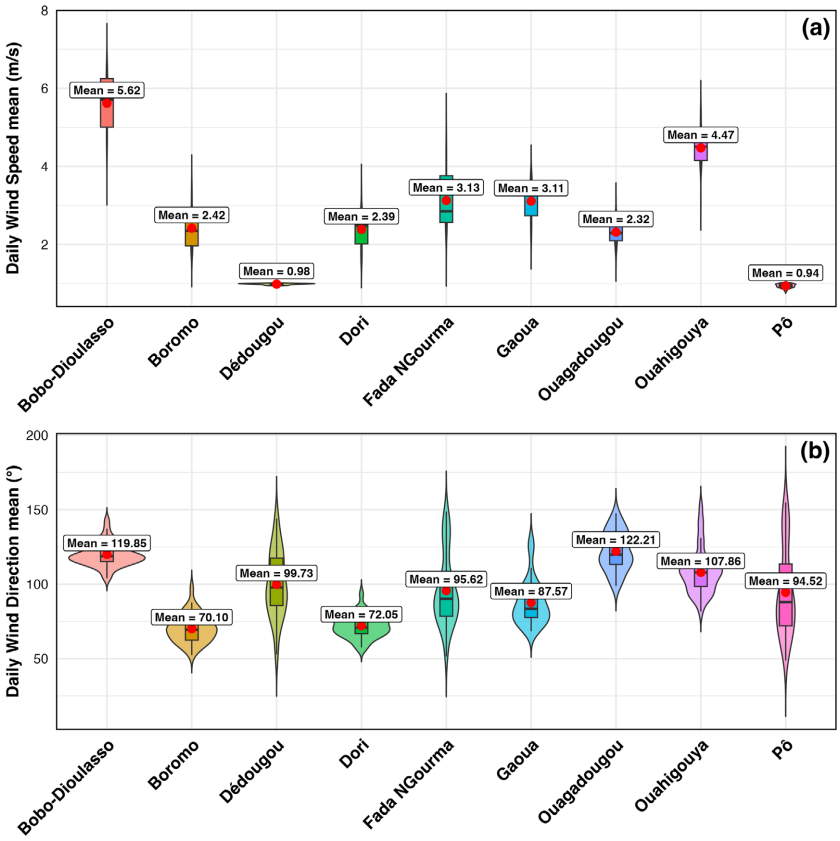


Figure 8. Distribution of daily wind speed (a) and Direction (b) levels across stations over the study period.

For wind direction (WD-MEAN), the distribution shown in Figure 8 appears more consistent across stations, with most values ranging between 70˚ and 130˚. The violin plots are generally symmetrical, but a few stations (e.g., Dédougou, Gaoua) display extended tails or slight bimodality, which may indicate directional shifts or transitional wind regimes. Outliers remain limited but noticeable in certain stations, especially where longer tails appear, reflecting occasional deviations in wind orientation likely tied to seasonal variability or isolated weather events.

We now turn to correlation analysis to explore the interrelationships among meteorological variables. Understanding these dependencies is key for guiding multivariate approaches. This step supports subsequent tasks such as anomaly detection and pattern analysis.

5.2. Analysis of the Correlation Structure among Meteorological Variables

Figure 9 displays correlation heatmaps across the nine climatic stations for the study period. Strong and consistent positive correlations are observed among temperature variables (T-MAX, T-MEAN, T-MIN), as well as within the humidity group (RH-MAX, RH-MEAN, RH-MIN), reflecting internal climatic consistency. Sunshine duration and global radiation are also moderately to strongly correlated in most stations, highlighting their shared dependence on solar input. In contrast, rainfall tends to be negatively correlated with both temperature and solar-related variables. Wind direction, however, exhibits weak and inconsistent correlations, confirming its relative independence from the other meteorological variables.

Following this correlation analysis, we propose to focus the subsequent multivariate outlier detection using machine learning methods from the PyOD library on variables exhibiting relatively strong correlations, particularly maximum and minimum temperature (T-MAX, T-MIN) and relative humidity (RH-MAX, RH-MIN). These variables are known to be the most sensitive to extreme values, as confirmed by violin plot analyses (see Figure 5 & Figure 6). Solar-related parameters (GRAD, SUN-DUR) and precipitation (RAIN) will be reserved for more specific analyses that consider additional influencing factors. It is important to note that the ultimate goal is to isolate erroneous data points from actual climatic extremes, treat them as missing values, and apply imputation techniques to build a high-quality dataset. We also exclude ETP from this stage, as it is derived by ANAM based on other meteorological variables. Wind-related parameters will be explored further through univariate analysis.

5.3. Univariate Analysis with Isolation Forest Method: Impact of the Contamination Hyperparameter

In this section, we apply the unsupervised Isolation Forest algorithm [17] to detect outliers across different values of the contamination parameter, which represents the expected proportion of anomalies in the dataset. Using the implementation

from the Python scikit-learn package, we first set the contamination parameter to "auto". Figure 10 presents the results for the Bobo-Dioulasso station, where outlier regions can be distinguished based on the decision function score: negative values indicate likely anomalies, while positive values correspond to normal data points. For each variable, the number and proportion of potential outliers are reported. When this procedure is repeated across all nine stations, the resulting proportions are summarized in the heatmap shown in Figure 11.

The results show a relatively high proportion of outliers compared to typical meteorological datasets, where less than 10% is expected. In our case, validating this proportion would imply inspecting over 4400 rows per variable, which is not practical. To address this, we manually varied the contamination rate from 1% to 10%. The results, shown in Figure 12, indicate that the detection stabilizes around each rate of expected outliers in the database.

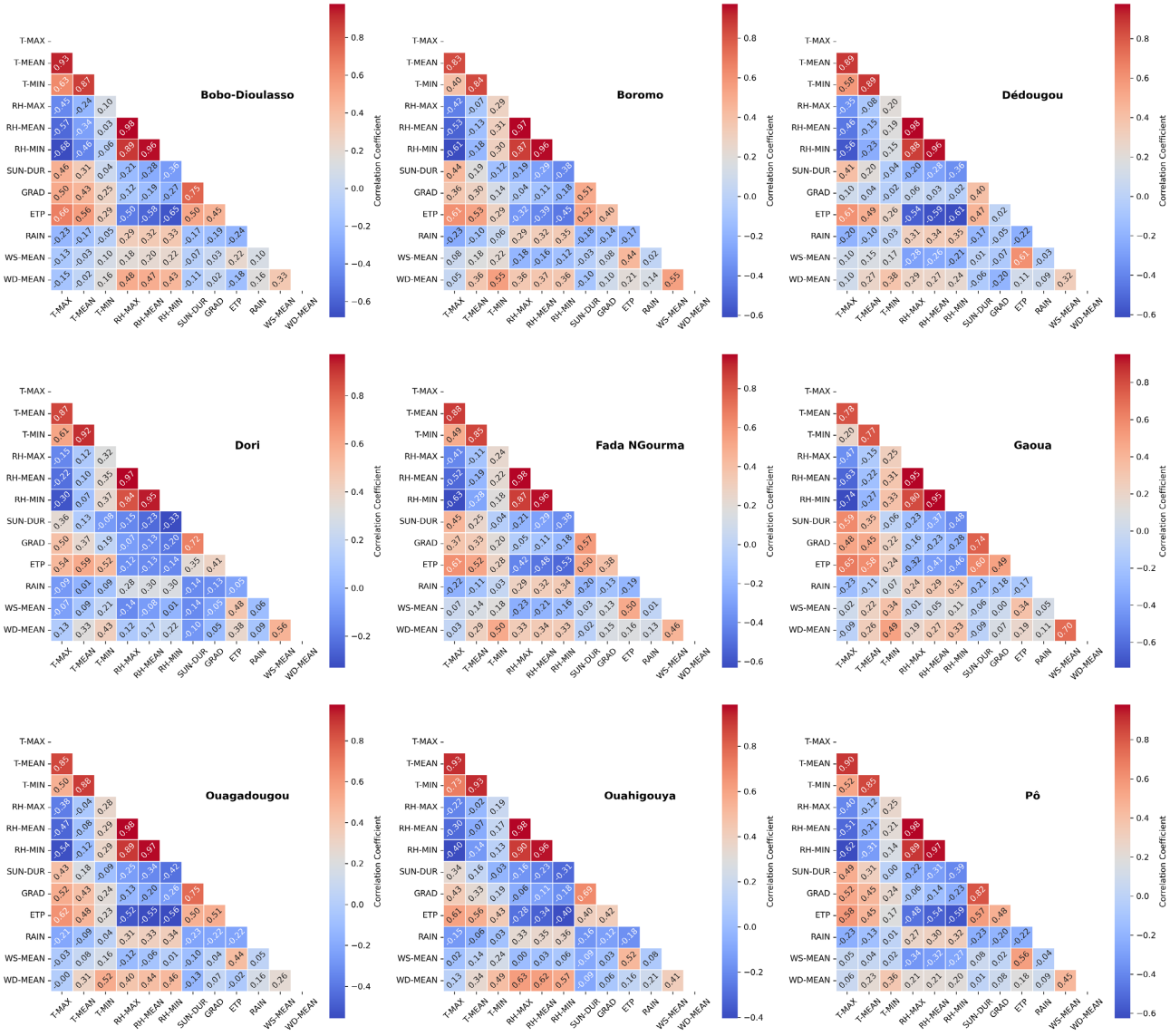


Figure 9. Station-wise correlation matrices of meteorological variables.

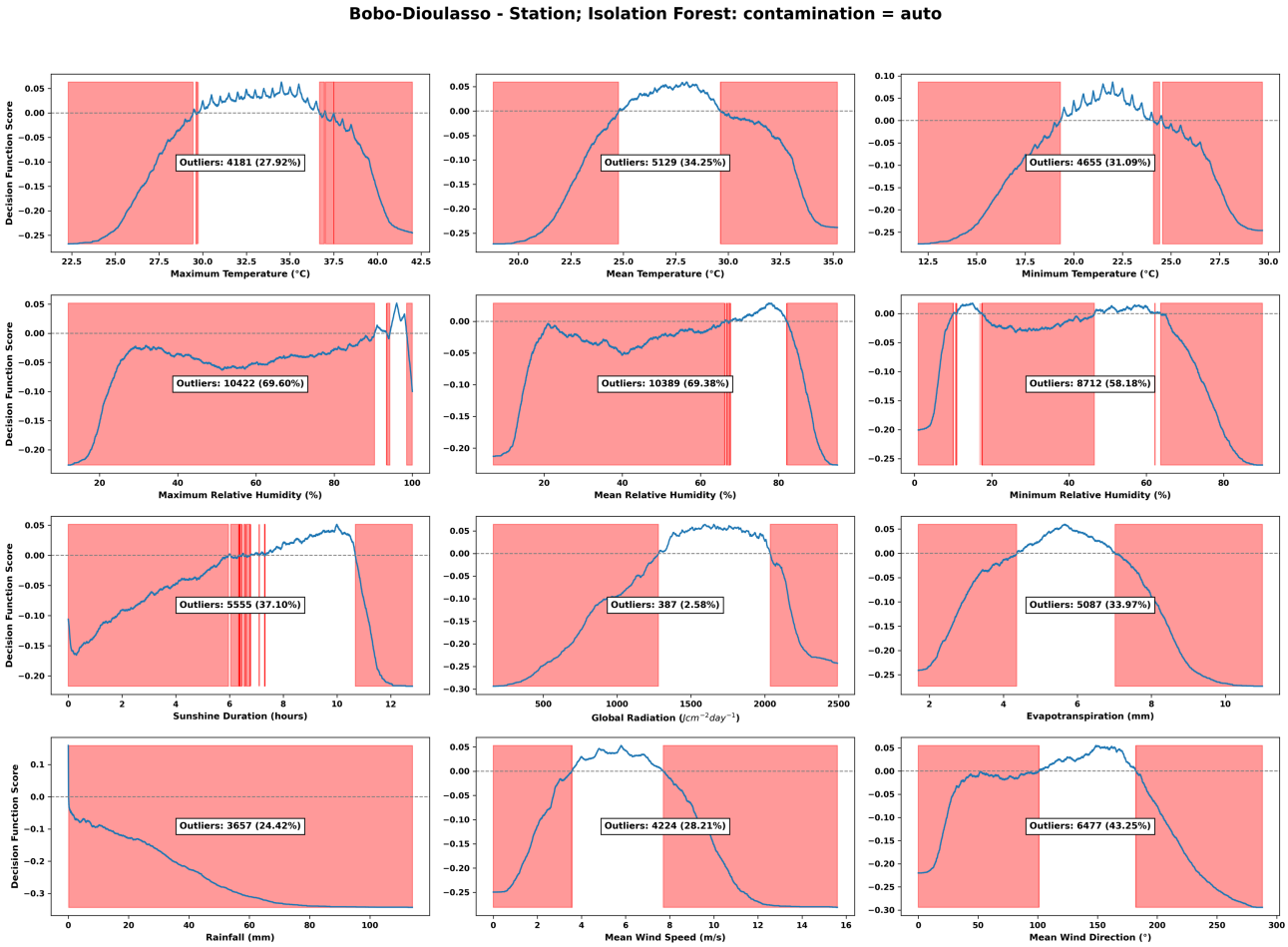


Figure 10. Isolation forest-based outlier detection of meteorological variables at Bobo-Dioulasso (Contamination = 'auto').

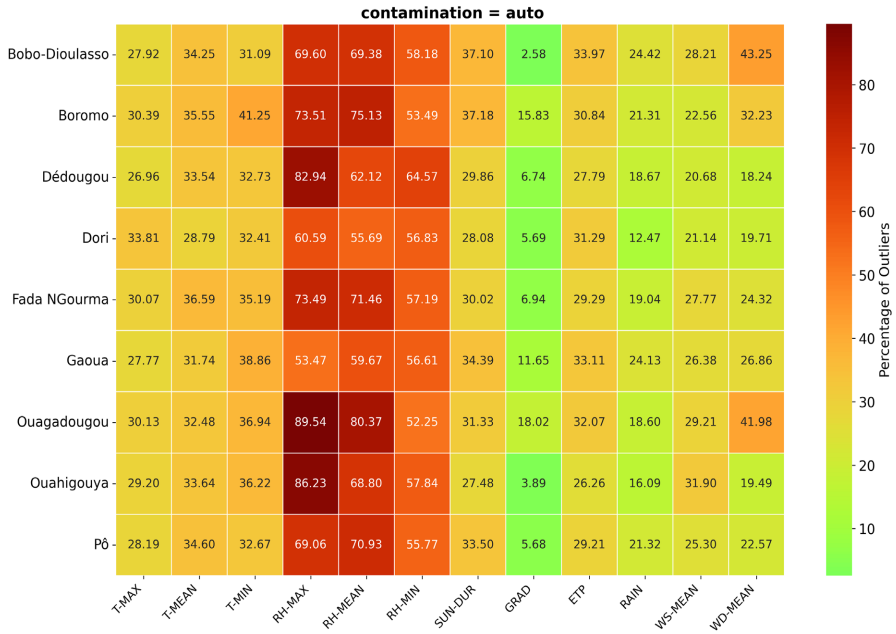


Figure 11. Proportion of outliers per variable and station with isolation forest (Contamination = 'auto').

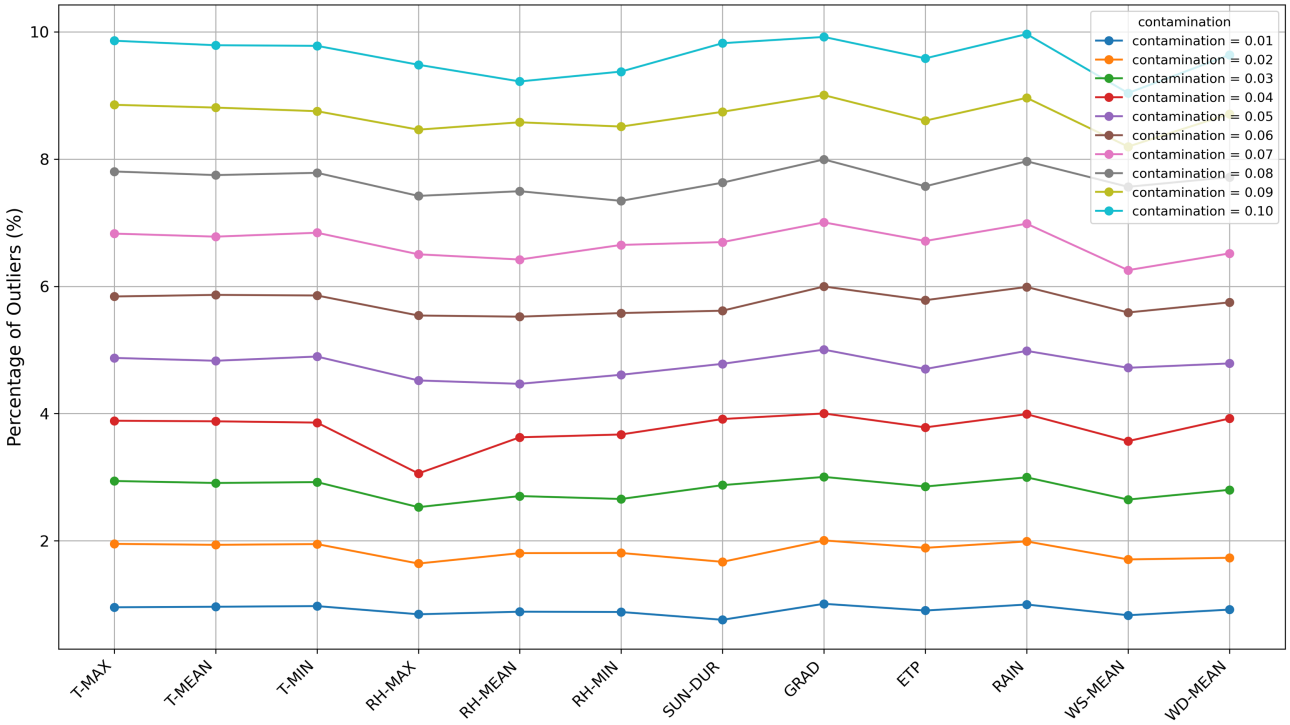


Figure 12. Effect of contamination parameter on the percentage of detected outliers across climatic variables using isolation forest.

Since optimizing the contamination parameter in Isolation Forest requires labeled data, which is rarely available in climate datasets, we explored alternative approaches less sensitive to this hyperparameter. As noted by [17], the algorithm remains robust when using raw anomaly scores instead of the default binary classification using the contamination parameter.

In this context, thresholding can be applied a posteriori using robust statistical methods, such as extreme quantiles (e.g., 95th or 99th percentile). In our case, we used the 95th percentile of the scores, as illustrated in **Figure 13**. This approach results in an outlier detection rate close to 5% of the dataset.

To further our analysis, we explored the Extended Isolation Forest (EIF) method developed by [22], which enhances the classical Isolation Forest by correcting its structural bias, particularly in detecting anomalies with curved or sinusoidal patterns. By using random hyperplanes instead of axis-parallel splits, EIF produces more robust and less biased anomaly scores [22]. We then apply EIF to our climate dataset using the 95th percentile as the threshold for anomaly detection. The results, presented in **Figure 14**, show a detection rate again close to 5%.

In conclusion, the univariate analyses performed with the Isolation Forest model and its extended version (EIF) reveal an outlier proportion of approximately 5% to 6% for each variable. This proportion presents a challenge, as removing these data points would result in a significant loss of information. Furthermore, analyzing anomalies in Isolation Forest model may be misleading, for instance, an extreme weather event such as a sudden storm could simultaneously

impact rainfall, wind speed, and temperature. To address this, we next turn to the PyOD library to explore multivariate anomaly detection, aiming for a more integrated and realistic identification of outliers.

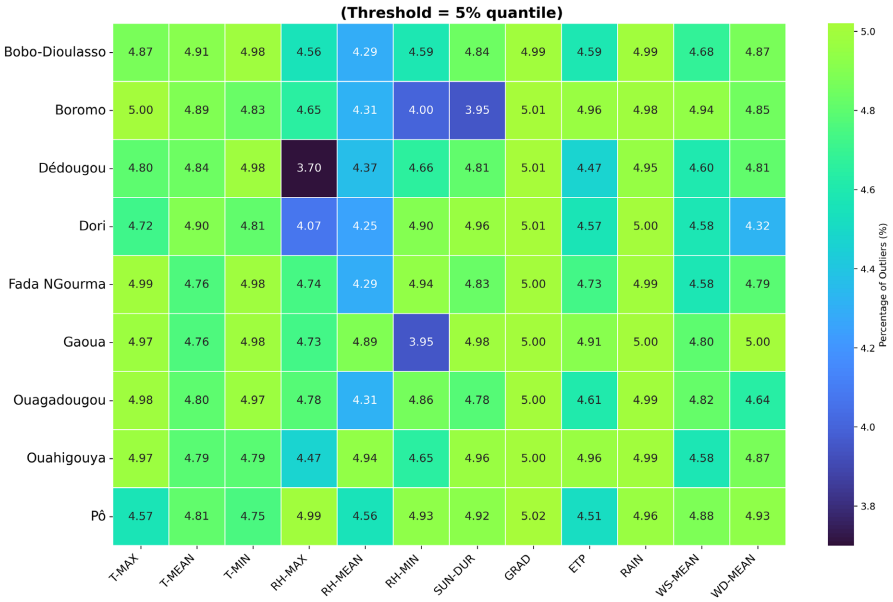


Figure 13. Proportion of outliers per variable and station (Threshold = 5% quantile).

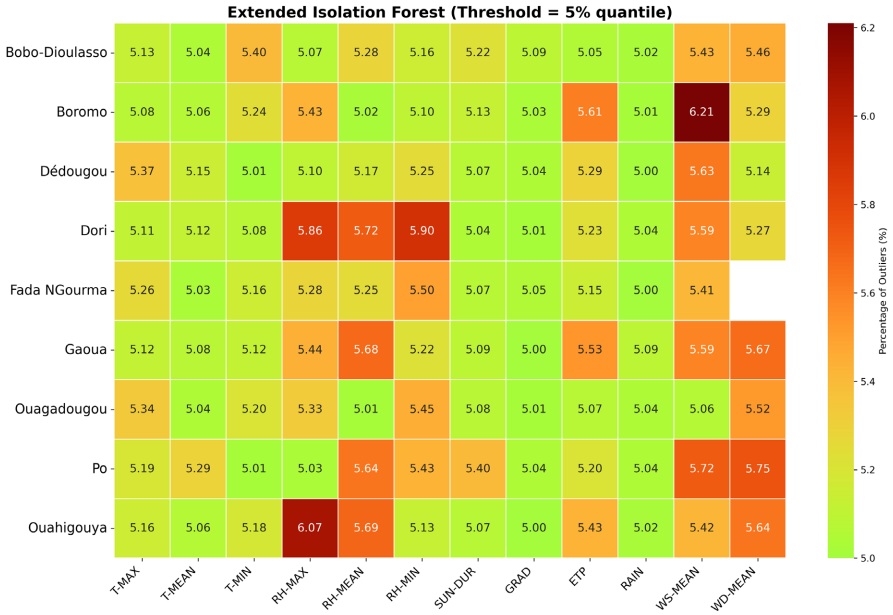


Figure 14. Proportion of outliers per variable and station with extended isolation forest (Threshold = 5% quantile).

5.4. Multivariate Evaluation of PyOD Outlier Detection Models

As previously mentioned in the correlation analysis between variables, we now apply the 18 models (see Table 1) from the PyOD package to our dataset. Given that most unsupervised models in PyOD require an estimated contamination rate

to indicate the proportion of anomalous data, and based on the results obtained from the Isolation Forest and Extended Isolation Forest methods discussed above, we fix the contamination rate at 5%. The detection is performed on all stations individually. In particular, special attention is given to the variables T-MAX (Maximum Temperature) and RH-MIN (Minimum Relative Humidity), which exhibit a strong negative correlation across all stations, suggesting a strong climatic dependency between extreme temperature and humidity conditions. The outlier detection results for all models and stations are presented in Figure 15.

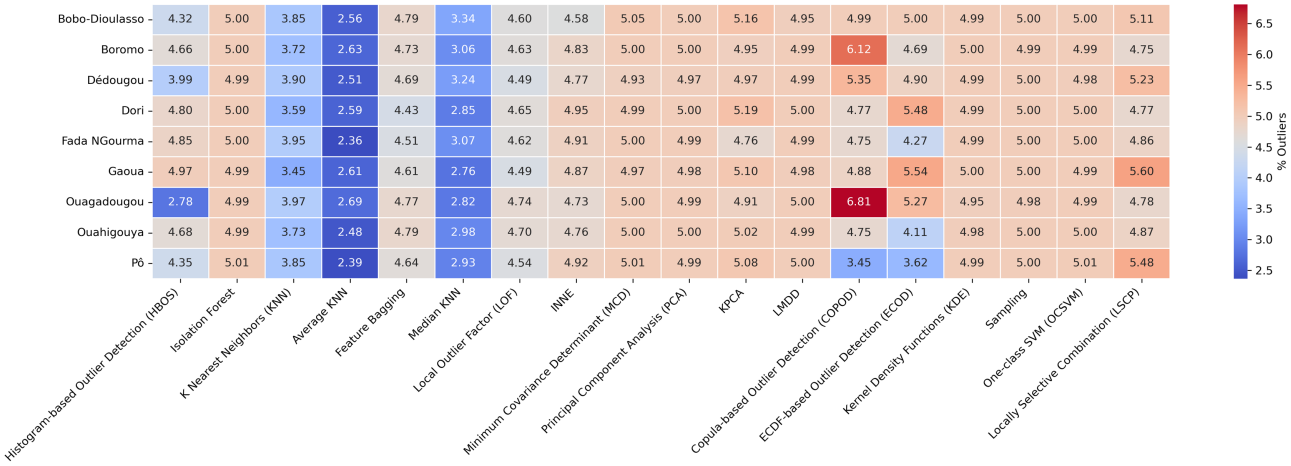


Figure 15. Proportion of outliers Detected by 18 PyOD models across synoptic stations (Contamination = 5%) applied to T-MAX and RH-MIN variable pair.

The results show a consistent pattern of outlier proportions around the 5% threshold across most unsupervised models, including Isolation Forest, HBOS, and PCA-based methods. However, the K-Nearest Neighbors (KNN)-based models (KNN, Average KNN, and Median KNN), which are semi-supervised, display lower outlier rates that deviate from the general trend observed across other methods. These findings align with [23], who emphasized that supervised approaches are often less suitable for climatic data due to the lack of reliable ground truth labels, which complicates the calibration of such models for outlier detection.

5.5. How to Select the Most Suitable Detection Method for the Variable Pair at this Stage?

Figure 16 and Figure 17 display the learned decision function maps for 18 outlier detection models from the PyOD package, applied to a bivariate correlation (T-MAX and RH-MIN) for Bobo-Dioulasso climatic dataset. These visualizations provide valuable insight into how each model defines the boundary between inliers and potential outliers. Models that produce clear, compact, and smooth contours surrounding the main data cloud while effectively isolating sparse regions are generally more effective at identifying meaningful anomalies. Based on this visual inspection, Kernel PCA (KPCA), Locally Selective Combination (LSCP), Feature Bagging, and Local Outlier Factor (LOF) emerge as the most suitable for

our case. These models delineate inlier regions with high precision and capture complex data structures, which is crucial for detecting subtle anomalies in multivariate climatic time series.

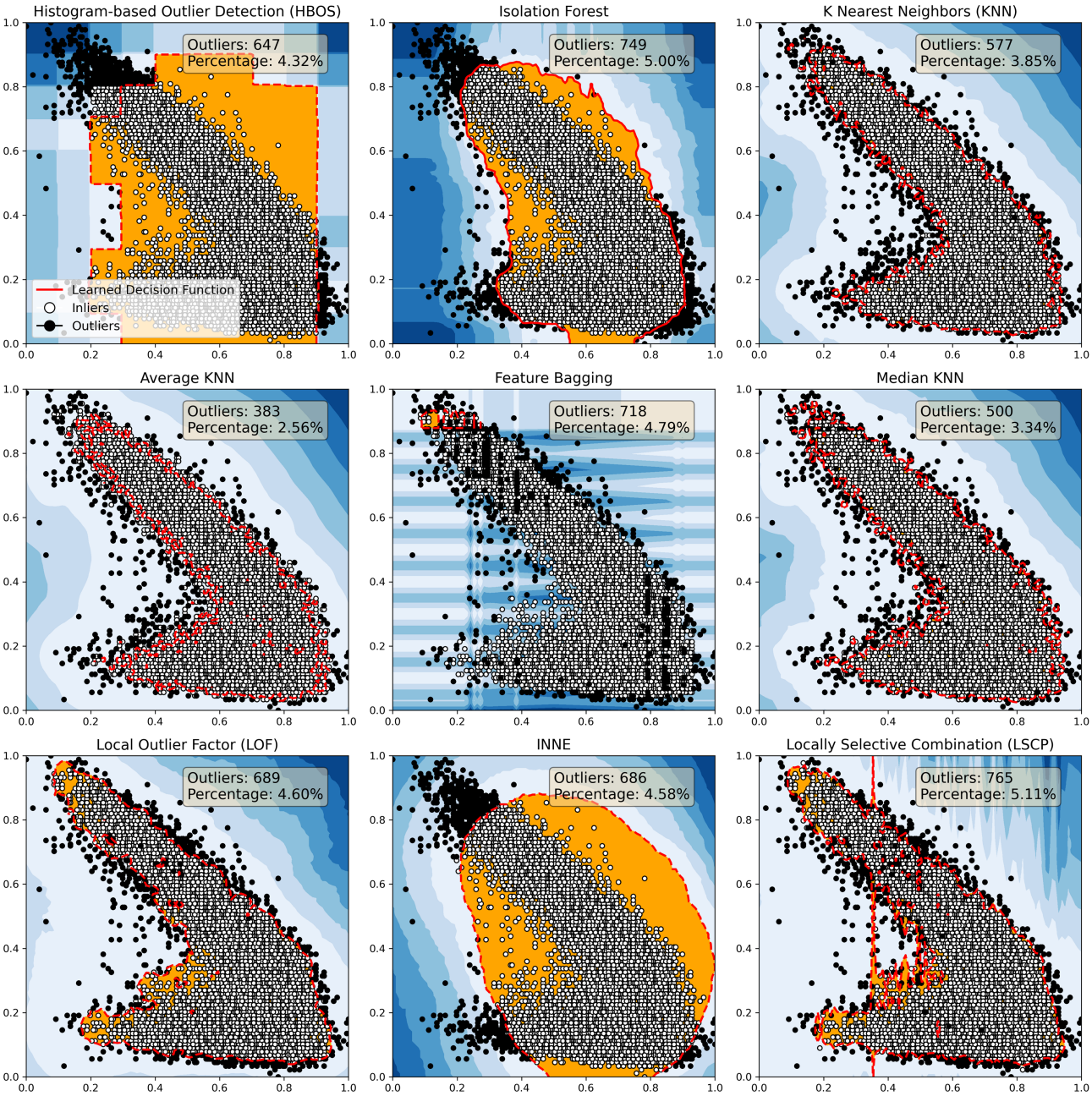


Figure 16. Decision function maps of the first nine PyOD models applied to the T-MAX and RH-MIN variable pair at Bobo-Dioulasso Station (Contamination = 5%).

In contrast, models such as HBOS, INNE, MCD, LMDD, KDE, PCA, COPOD, and OCSVM tend to produce overly broad or rigid decision boundaries, often closely matching the imposed contamination rate without effectively adapting to the intrinsic structure of the data. This can lead to misidentification of relevant

anomalies, particularly in complex and high-dimensional time series. These limitations are especially pronounced when working with unlabeled temporal datasets, where unsupervised models must infer structure without supervision or temporal continuity constraints. As noted by [24], traditional density- and projection-based models may struggle to capture temporal dependencies or localized irregularities common in environmental data. Similarly, [25] argue that outlier detection in time series requires models capable of understanding not only the statistical distribution but also the sequential and contextual nature of the data, a feature missing in many classical unsupervised methods.

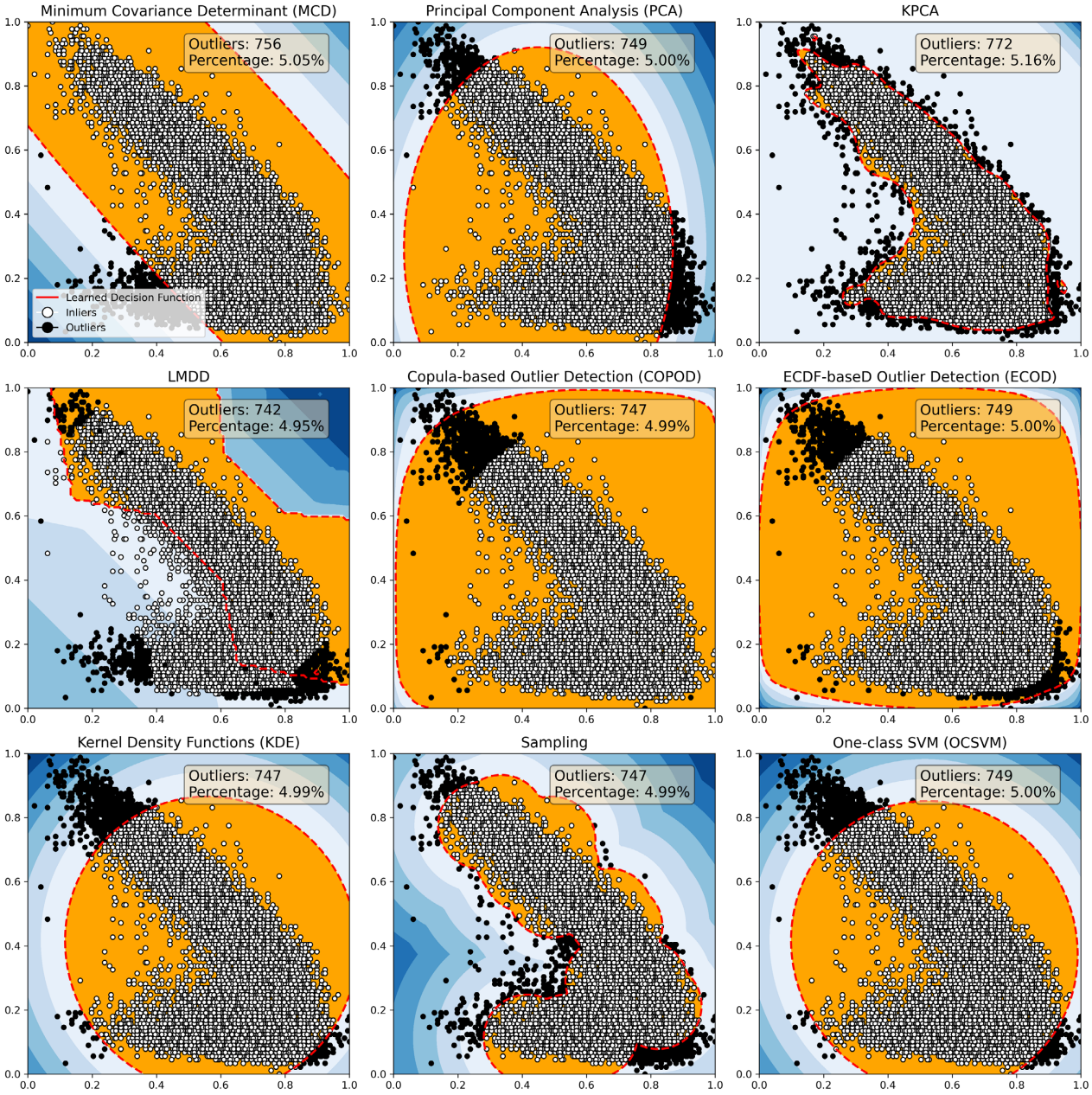


Figure 17. Decision function maps of the remaining nine PyOD models applied to the T-MAX and RH-MIN variable pair at Bobo-Dioulasso Station (contamination = 5%).

The behavior of KNN-based models, such as Average KNN and Median KNN, which show refined boundaries and underestimating the proportion or lead to a poor representation of rare meteorological events. Such limitations are consistent with the observations of [26], who noted that certain distance or density-based models can be ill-suited to complex, high-dimensional environmental datasets due to the difficulty in defining effective neighborhood thresholds without labeled data.

Consequently, for robust and interpretable outlier detection in climate-related applications, KPCA, LSCP, LOF, and Feature Bagging offer promising potential.

At the end of this study, we used the results from KPCA to record in a spreadsheet the rows (*dates*) where a given variable was identified as anomalous. A visual inspection using interactive graph representations generated with the Plotly package [27] will allow for filtering, comparison with recorded extreme climatic events during the period, and ultimately support decision-making regarding the removal of potentially erroneous values.

6. Conclusions

Time series outlier detection, also known as time series anomaly detection, is the process of identifying anomalies in a time series dataset. It involves several techniques, including statistical approaches (e.g., z-score, percentiles, Density-Based Spatial Clustering of Applications with Noise-DBSCAN), time series models (e.g., moving averages, ARIMA), machine learning techniques (e.g., Isolation Forest, One-Class SVM), and hybrid methods that take into account multivariate relationships, seasonal patterns, or historical data comparisons.

In this work, we demonstrate the effectiveness of various machine learning techniques for the detection of outliers in complex multivariate climate data. Applying 18 models from the PyOD library, the results show a consistent trend of outlier proportions around the 5% threshold in most unsupervised models, including methods based on Isolation Forest, HBOS, and PCA. However, models which are based on K-nearest neighbors (KNN) (KNN, mean KNN, and median KNN), and are semi-supervised, show lower outlier rates, and deviate from the general trend observed with the other methods. The results of visual inspection show that Kernel PCA (KPCA), Local Selective Combination (LSCP), Feature Clustering, and Local Outlier Factor (LOF) appear to be the most suitable for our case. These models delineate outlier regions with great precision and capture complex data structures, a crucial result for detecting subtle anomalies in multivariate climate time series.

The approach presented in this paper represents a pioneering effort in Burkina Faso for improving the quality of climatic datasets and enhancing the reliability of models used to analyze extreme weather events. The results obtained suggest promising avenues for better managing climate-related risks and understanding environmental dynamics in data-scarce contexts. However, further investigations are required, particularly for the Bogandé station, which was excluded from the

current analysis due to a high proportion of missing values. Addressing this challenge will require integrating complementary approaches, notably Physics-Informed Neural Networks (PINNs), which are increasingly recognized as powerful tools for ensuring data consistency in climate databases, given that climate variables are inherently governed by physical laws.

In addition, solar global radiation will be the subject of a separate, in-depth analysis due to its specific behavior and relevance. More broadly, future work should consider extending the current framework by exploring time series-specific outlier detection techniques, including advanced machine learning approaches (e.g., deep learning-based models, hybrid multivariate detectors). These techniques can help capture seasonal patterns, historical trends, and multivariate interactions, thereby improving the robustness of anomaly detection in climate time series.

Acknowledgements

We gratefully acknowledge the support of the Higher Education Support Project (PAES) in Burkina Faso, funded by the World Bank, for providing a doctoral research equipment grant to Mr. Ki.

Conflicts of Interest

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

References

[1] Blázquez-García, S., Conde, A., Mori, U. and Lozano, J.A. (2021) A Review on Outlier/Anomaly Detection in Time Series Data. *ACM Computing Surveys*, **54**, 1-36. https://doi.org/10.1145/3444690

[2] Choi, K., Yi, J., Park, C. and Yoon, S. (2021) Deep Learning for Anomaly Detection in Time-Series Data: Review, Analysis, and Guidelines. *IEEE Access*, **9**, 120043-120065. https://doi.org/10.1109/ACCESS.2021.3107975

[3] Schmidl, S., Wenig, P. and Papenbrock, T. (2022) Anomaly Detection in Time Series: A Comprehensive Evaluation. *Proceedings of the VLDB Endowment*, **15**, 1779-1797. https://doi.org/10.14778/3538598.3538602

[4] Srinivasan, R., Wang, L. and Bulleid, J.L. (2020) Machine Learning-Based Climate Time Series Anomaly Detection Using Convolutional Neural Networks. *Weather and Climate*, **40**, 16-31. https://doi.org/10.2307/27031377

[5] Wu, R. and Keogh, E.J. (2021) Current Time Series Anomaly Detection Benchmarks Are Flawed and Are Creating the Illusion of Progress. *IEEE Transactions on Knowledge and Data Engineering*, **35**, 2421-2429. https://doi.org/10.1109/ICDE53745.2022.00116

[6] Wahyono, T., Heryadi, Y., Soeparno, H. and Abbas, B.S. (2020) Anomaly Detection in Climate Data Using Stacked and Densely Connected Long Short-Term Memory Model. *Journal of Computers*, **31**, 42-53. https://doi.org/10.3966/199115992020083104004

[7] Bâra, A., Văduva, A.G. and Oprea, S.V. (2024) Anomaly Detection in Weather Phenomena: News and Numerical Data-Driven Insights into the Climate Change in Romania's Historical Regions. *International Journal of Computational Intelligence Sys-*

tems, **17**, Article No. 134. https://doi.org/10.1007/s44196-024-00536-2

[8] Shen, J., Yang, M., Zou, B., Wan, N. and Liao, Y. (2012) Outlier Detection of Air Temperature Series Data Using Probabilistic Finite State Automata-Based Algorithm. *Complexity*, **17**, 48-57. https://doi.org/10.1002/cplx.21390

[9] Esmaeili, F., Cassie, E., Nguyen, H.P.T., Plank, N.O.V., Unsworth, C.P. and Wang, A. (2023) Anomaly Detection for Sensor Signals Utilizing Deep Learning Autoencoder-Based Neural Networks. *Bioengineering* (*Basel, Switzerland*), **10**, Article 405. https://doi.org/10.3390/bioengineering10040405

[10] Tinawi, I. (2019) Machine Learning for Time Series Anomaly Detection. Doctoral Dissertation, Massachusetts Institute of Technology.

[11] Han, S., Hu, X., Huang, H., Jiang, M. and Zhao, Y. (2022) Adbench: Anomaly Detection Benchmark. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4266498

[12] Kawale, J., Chatterjee, S., Kumar, A., Liess, S., Steinbach, M. and Kumar, V. (2011) Anomaly Construction in Climate Data: Issues and Challenges. *Proceedings of the* 2011 *Conference on Intelligent Data Understanding*, California, 19-21 October 2011.

[13] Zhao, Y., Nasrullah, Z. and Li, Z. (2019) PyOD: A Python Toolbox for Scalable Outlier Detection. *Journal of Machine Learning Research*, **20**, 1-7. https://github.com/yzhao062/pyod. http://jmlr.org/papers/v20/19-011.html

[14] Chen, S., Qian, Z., Siu, W., Hu, X., Li, J., Li, S. and Zhao, Y. (2024) PyOD 2: A Python Library for Outlier Detection with LLM-powered Model Selection.

[15] Gregoire, T., Ayala Solares, H.A., Coutu, S., Cowen, D., DeLaunay, J.J., Fox, D.B., Keivani, A., Krauss, F., Mostafá, M., Murase, K., Neights, E. and Turley, C.F. (2021) Model Independent Search for Transient Multimessenger Events with AMON Using Outlier Detection Methods. 37*th International Cosmic Ray Conference*, Berlin, 15-22 July 2021, 934.

[16] Li, Y., Zha, D., Venugopal, P., Zou, N. and Hu, X. (2020) PyODDS: An End-to-End Outlier Detection System with Automated Machine Learning. *Companion Proceedings of the Web Conference* 2020, Taipei, 20-24 April 2020, 153-157.

[17] Liu, F.T., Ting, K.M. and Zhou, Z.H. (2008) Isolation Forest. 2008 *Eighth IEEE International Conference on Data Mining* (*ICDM*), Pisa, 15-19 December 2008, 413-422. https://doi.org/10.1109/ICDM.2008.17

[18] Lai, K.H., Zha, D., Xu, J., Zhao, Y., Wang, G. and Hu, X. (2021) Revisiting Time Series Outlier Detection: Definitions and Benchmarks.

[19] Teixeira, C.F. (2024) Outlier Explanations in Data Streams-Applications for Environmental Data. Master's Thesis, Universidade do Porto (Portugal).

[20] Dembélé, M. and Zwart, S.J. (2016) Evaluation and Comparison of Satellite-Based Rainfall Products in Burkina Faso, West Africa. *International Journal of Remote Sensing*, **37**, 3995-4014. https://doi.org/10.1080/01431161.2016.1207258

[21] Ki, Z.G. (2020) Données Climatiques: Analyses de corrélations, de régressions et prédiction de données manquantes. Master degree thesis, Université Josep KI-ZERBO (Burkina Faso).

[22] Hariri, S., Kind, M.C. and Brunner, R.J. (2021) Extended Isolation Forest. *IEEE Transactions on Knowledge and Data Engineering*, **33**, 1479-1489. https://doi.org/10.1109/tkde.2019.2947676

[23] Wolpher, M. (2018) Anomaly Detection in Unstructured Time Series Datausing an LSTM Autoencoder. Master of Science, Engineering Physics in the School of Electrical Engineering and Computer Science, Kth Royal Institute of Technology (Sweden).

[24] Pang, G., Shen, C., Cao, L. and Hengel, A.V.D. (2021) Deep Learning for Anomaly Detection: A Review. *ACM Computing Surveys*, **54**, 1-38. https://doi.org/10.1145/3439950

[25] Chauhan, S. and Vig, L. (2015) Anomaly Detection in ECG Time Signals via Deep Long Short-Term Memory Networks. 2015 *IEEE International Conference on Data Science and Advanced Analytics* (*DSAA*), Paris, 19-21 October 2015, 1-7. https://doi.org/10.1109/dsaa.2015.7344872

[26] Hodge, V.J. and Austin, J. (2004) A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, **22**, 85-126. https://doi.org/10.1023/B:AIRE.0000045502.10941.a9

[27] Plotly Technologies Inc. (2015) Collaborative Data Science. Plotly Technologies Inc. https://plotly.com/python