Environmental Data Science (2025), 4: e50, 1–31
doi:10.1017/eds.2025.10025

CAMBRIDGE UNIVERSITY PRESS

APPLICATION PAPER

Prediction and uncertainty quantification of drought in North Benin

Bernardin Marie Augustin Sèdjro Ligan[1], Gbègninougbo Aurel Davy Tchokponhoue[1], Alain Thierry Iliho Manzi[1], Kora B. Désiré Simperegui[2] and Rodrigue Bignon Wilfried Vodounon[3]

[1]Université Mohammed VI Polytechnique (UM6P), Ben Guerir, Marrakesh-Safi, Morocco
[2]International Fertilizer Development Center (IFDC), Accra, Greater Accra, Ghana
[3]Université d'Abomey-Calavi Institut de Mathématiques et des Sciences Physiques (IMSP), Dangbo, Ouémé, Bénin
Corresponding author: Bernardin Marie Augustin Sèdjro Ligan; Email: bernardin.ligan@um6p.ma

Received: 11 September 2024; **Revised:** 03 October 2025; **Accepted:** 09 October 2025

Keywords: conformal prediction; deep learning (DL); drought forecasting; machine learning (ML); uncertainty quantification

Abstract

Drought forecasting is a critical tool for mitigating the severe impacts of water scarcity, particularly in regions like North Benin, where agriculture is a cornerstone of livelihoods. Despite the vital importance of its accurate prediction in resource management, the ability to quantify uncertainties in forecasts is a significant pain point to enable more informed and trustworthy decision-making. So, this study aims to develop an uncertainty-aware prediction model for drought forecasting in six key localities within the Alibori department—Banikoara, Gogounou, Kandi, Karimama, Malanville, and Segbana—each facing unique challenges due to drought. To achieve this, we conducted a comprehensive experiment involving six machine learning models (linear regression, ridge regression, random forest, Xgboost, LightGBM, and SVM) and four deep learning models (Conv1D, LSTM, GRU, and Conv1D-LSTM) using the Standardized Precipitation Index at a 6-month scale. To address the uncertainty quantification challenge, we employed the Ensemble Batch Prediction Interval, a conformal prediction method specifically designed for time series data. Our comparative analysis, framed within the Borda count methodology, utilized performance metrics such as R^2, RMSE, MSE, and carbon footprint, as well as uncertainty quantification metrics, including empirical coverage and the width of prediction intervals. The top-performing models achieved R^2 scores of 98.29, 97.84, 97.76, 97.42, 96.61, and 97.07%, and prediction interval coverages of 0.94, 0.79, 0.93, 0.77, 0.73, and 0.93, respectively, for Banikoara, Gogounou, Malanville, Kandi, Segbana, and Karimama. The Conv1D-LSTM model stood out as the most effective, offering an optimal balance between predictive accuracy and uncertainty coverage.

Impact Statement

This study, one of the first of its kind in a low-income country like Benin, underscores the importance of uncertainty quantification in drought forecasting within these specific localities and can help decision-makers, agricultural researchers, and local populations to take early precautions for drought occurrence based on some possible scenarios.

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

1. Introduction

Over recent decades, human activities such as industrialization, deforestation driven by agricultural expansion, and rapid urbanization have intensified the pace of climate change (Mir et al., 2025). These anthropogenic pressures have resulted in shifting rainfall patterns, rising temperatures, and prolonged droughts—phenomena that significantly threaten agricultural productivity. Among the various climate-related hazards, drought stands out as one of the most pervasive and damaging in sub-Saharan Africa (SSA) (Aryal et al., 2023). Between 1900 and 2013, Africa recorded 291 drought events, affecting more than 362 million people and causing 847,000 deaths, with an estimated $3 billion in economic losses (Masih et al., 2014). Droughts can be categorized into four main types: meteorological, hydrological, agricultural, and socioeconomic. They are typically characterized using indices such as the Standardized Precipitation Index (SPI), Crop Moisture Index, Palmer Drought Severity Index, and the Standardized Precipitation Evapotranspiration Index (Yihdego et al., 2019).

In SSA, where agriculture is predominantly rainfed, droughts have devastating effects on crop yields, threatening food availability and nutritional security for the region's 1.15 billion people (Pison et al., 2022; United Nations Department of Economic and Social Affairs Population Division, 2022). This has led to alarming malnutrition rates of 20.4% (FAO et al., 2021). Staple crops, such as maize (*Zea mays* L.), often fall short of their yield potential, further exacerbating food insecurity (Scheiterle and Birner, 2018). This productivity gap poses a substantial challenge for over 60% of the SSA population, who are directly or indirectly involved in agriculture and heavily reliant on the income generated from their agricultural activities (Oluwatayo and Ojo, 2016). In Benin, agriculture plays a central economic and social role, employing ~70% of the active labor force and contributing 33% to the national Gross Domestic Product. The northern region—especially the "Pôle de Développement Agricole 2 Borgou-Alibori"—is critical to national food security, accounting for 36% of Benin's agricultural output (MAEP, 2017). However, this region is particularly vulnerable to climate change, with drought expected to reduce agricultural production by 5–20% by 2025 (Paeth et al., 2008). This underscores the urgent need for improved drought monitoring and adaptation strategies in rainfed agricultural systems. Machine learning (ML) has emerged as a powerful tool for modeling complex environmental processes, including climate variability. It has been successfully applied in diverse contexts, such as flood forecasting (Kazadi et al., 2024), precipitation pattern analysis (Mooers et al., 2024), and rainfall prediction (Aïzansi et al., 2024).

In Benin, recent studies have applied ML and statistical approaches to assess drought risks. For instance, Vodounon et al. (2022), Halissou et al. (2022), Oyerinde and Olowookere (2018), Emmanuel et al. (2019), and Hounkpè et al. (2016) analyzed hydrometeorological trends in the Ouémé watershed and the Beninese Niger River basin using indices such as SPI and the Standardized Streamflow Drought Index. These studies revealed substantial shifts in rainfall cycles and documented moderate to severe drought episodes between 1970 and 2016. Halissou et al. (2022) further identified Kompongou (Mékrou), Couberi (Sota), Yakin (Alibori), and Gbasse (Sota) as high-risk sites for hydrological droughts. Projections suggest these patterns will persist and intensify through 2050. Vodounon et al. (2022) demonstrated that ML models could effectively predict SPI at a 12-month scale, achieving a coefficient of determination (R^2) of 0.99 and a root mean square error (RMSE) of 0.07. Despite these encouraging results, most existing studies, including those focused on the agriculturally critical northern region, lack quantification of the uncertainties associated with their forecasts. This omission reduces the practical value of ML models for decision-makers, who need to understand how confident the models are in their predictions to formulate adapted policies (Jalaian et al., 2019; Poggio et al., 2021; Gruber et al., 2023; Lilburne et al., 2024; Simperegui et al., 2025).

This study aims to develop a robust framework for predicting meteorological drought—specifically the 6-month SPI-6—using ML and deep learning (DL) techniques, while also quantifying the associated prediction uncertainties in the Alibori department of northern Benin. Using gridded meteorological datasets from six towns in the region, we intend to: (i) build multiple ML and DL models to predict SPI-6 values, (ii) quantify the uncertainty associated with each model's predictions, and (iii) compare the performance of ML and DL approaches in terms of accuracy and reliability. The outcomes of this research

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

will support climate risk management by providing stakeholders—particularly farmers and policy-makers with actionable insights to enhance drought preparedness and resilience.

2. Materials and methods

In this section, we present the methodology adopted in this study, as illustrated in Figure 1. The process begins with a multivariate time series dataset and involves several preprocessing steps, including feature selection, creation of lagged variables, and normalization. The dataset is then split into training and testing subsets. Multiple ML and DL models are then trained, and their predictions are further processed using the Ensemble Batch Prediction Interval (EnbPI) method for uncertainty quantification. Finally, the performance of the models is evaluated using various statistical metrics.

2.1. Study area

Geographically situated between 11°19′ north latitude and 2°55′ east longitude, Alibori is a department in the northeast of the Republic of Benin. It is bordered to the north by the Republic of Niger, to the northwest by the Republic of Burkina Faso, to the east by the Federal Republic of Nigeria, to the west by Atacora, and the south by the department of Borgou. Covering an area of $26242 km^2$, Alibori is subdivided into six municipalities: Malanville, Karimama, Segbana, Gogounou, Banikoara, and Kandi (Figure 2). For each municipality, one location was selected based on cropland cover as indicated by the MODIS (Moderate Resolution Imaging Spectroradiometer) land cover map of Benin.

2.2. Data collection

To model the SPI and conduct our analysis, we obtained monthly weather data spanning a range of climate variables. These included temperature, total precipitation, wind speed and direction, surface pressure, relative humidity, cloud cover, soil moisture, top-of-atmosphere shortwave downward irradiance, and all-sky surface shortwave downward irradiance (Table 1). Although the national meteorological agency provides some of these datasets, their spatial coverage is limited and does not extend to many of the locations included in our study. Moreover, the available records do not span the 30-year baseline period recommended by the World Meteorological Organization (WMO) for climate assessments, especially those concerning drought characterization (WMO, 2017). To address these limitations and ensure spatially consistent and temporally comprehensive data, we relied on gridded satellite-based datasets. Two main sources were considered: NASA Prediction of Worldwide Energy Resources (POWER) dataset, with a spatial resolution of 0.5° latitude by 0.625° longitude (∼ 50 km), and the ERA5 reanalysis dataset, offering finer resolution at 11 km. Given that our analysis is conducted at the municipality level, we visualized both 50- and 11-km grids using QGIS (Quantum Geographic Information System) to assess their spatial alignment with administrative boundaries (Figure 2). While ERA5 provided higher spatial resolution, it resulted in multiple grid cells per municipality, which would have required aggregation and potentially introduced uncertainty. In contrast, the coarser NASA POWER grid is more closely aligned with municipality extents, thereby minimizing the need for spatial averaging. Consequently, we selected NASA POWER as the data source for this study. For each municipality, a single representative location was chosen based on the spatial distribution of cropland, ensuring that the extracted coordinates accurately reflect the areas of dominant agricultural activity (Figure 2). The dataset spans 41 years (1981–2021) and was extracted for the six municipalities under investigation (Malanville, Karimama, Segbana, Gogounou, Banikoara, and Kandi).

2.2.1. Standardized Precipitation Index

The SPI is a functional and quantitative definition of drought. It can be calculated at different timescales (3, 6, 12, 24, and 48) using the monthly rainfall data (Lawin et al., 2019; McKee et al., 1993). The SPI is defined as:

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

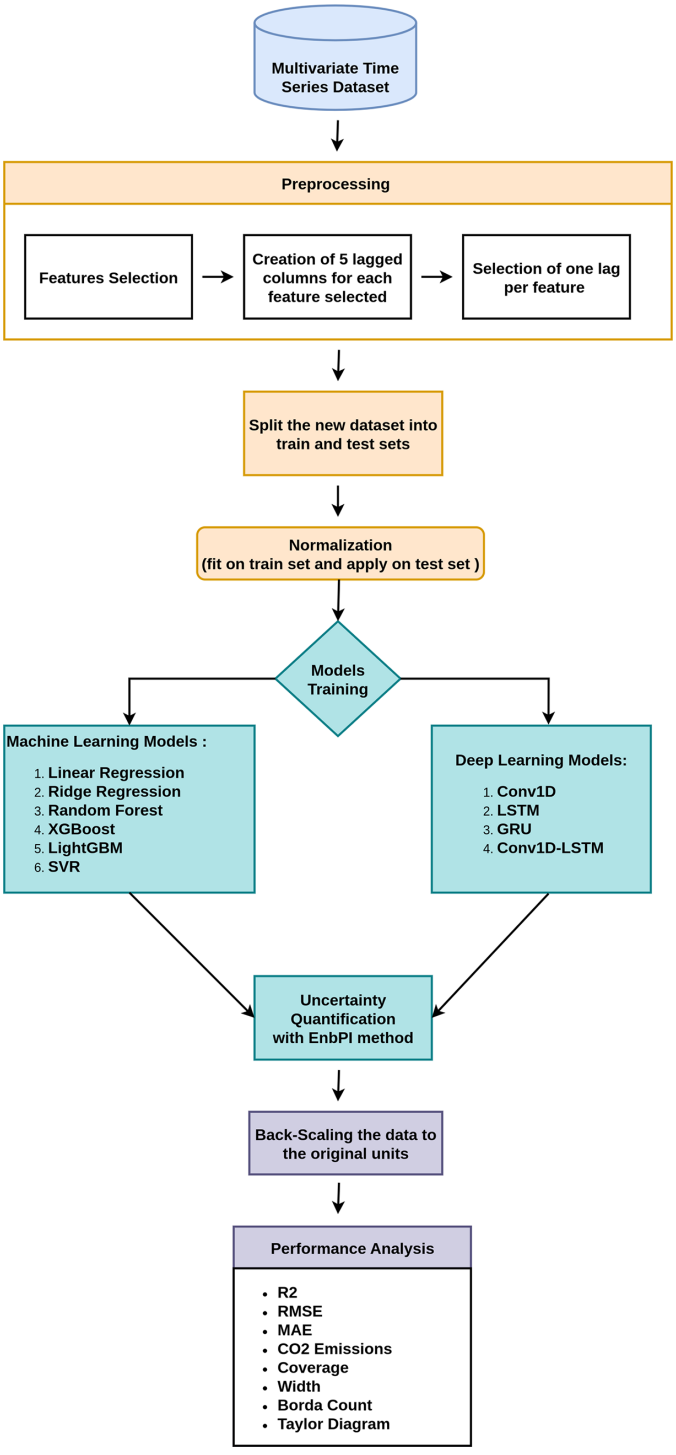


Figure 1. *Flowchart of the proposed methodology: From multivariate time series preprocessing to model training, uncertainty quantification with EnbPI, and final performance evaluation using various statistical metrics.*

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

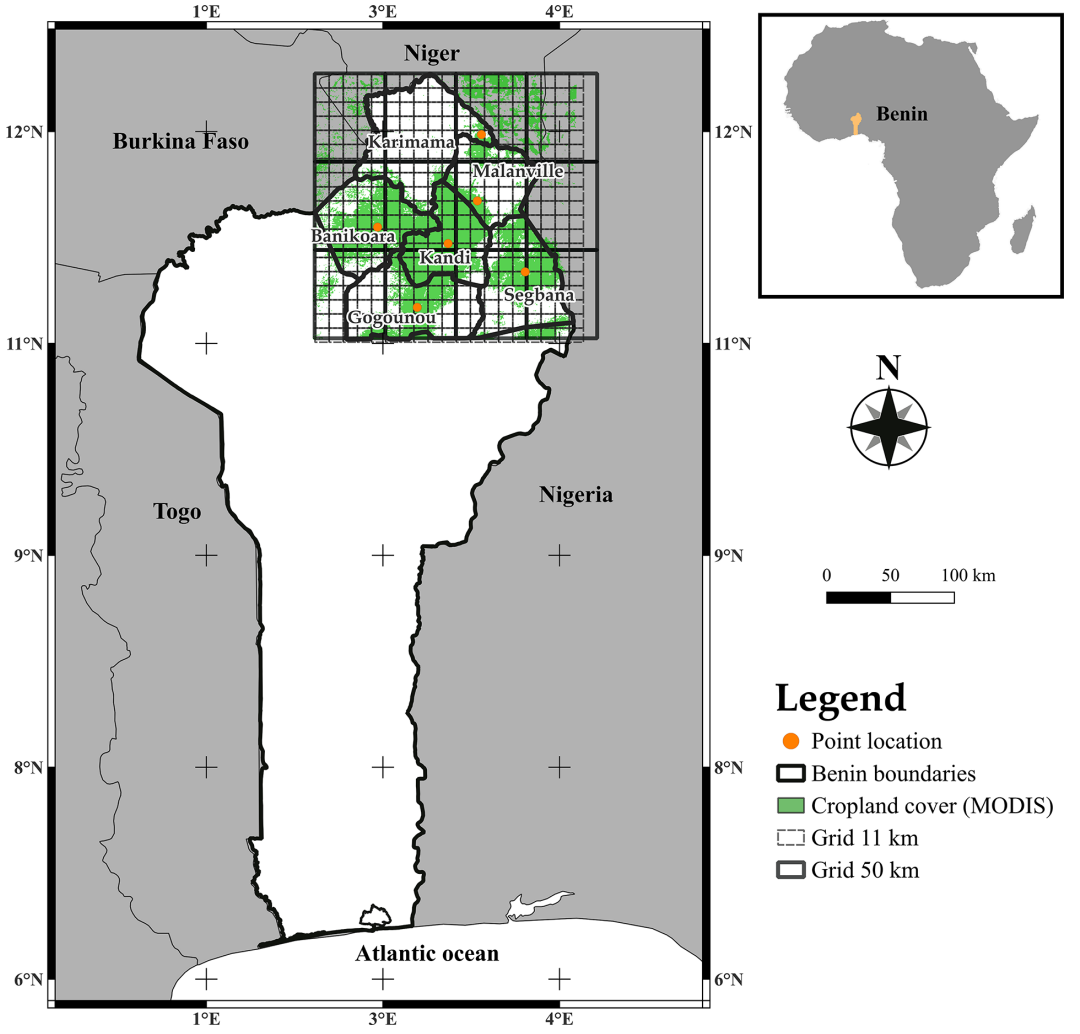


Figure 2. *Alibori department cartography and geographic distribution of the locations. The cropland cover is derived from the MODIS land cover.*

$$SPI_i = \frac{P_i - \overline{P}}{\sigma} \tag{1}$$

where SPI_i is the SPI of the month i, P_i is the rainfall of the month i, $\overline{P}$ and σ are the mean annual rainfall and the standard deviation on the timescale considered. In the present study, the SPI6 considers a 6-month timescale. SPI6 is used to assess drought occurrence on a semester basis. Table 2 provides a guideline for interpreting the SPI value.

2.2.2. Data preprocessing

Before modeling the SPI6, the data underwent a quality control process. Among the variables, CLOUD_AMT, TOA_SW_DWN, and ALLSKY_SFC_SW_DWN had 36 missing values (∼8%). Additionally, due to the 6-month timescale required to calculate SPI6, the first five observations of this variable were missing. To prevent these gaps from impacting the model performance, we excluded the CLOUD_AMT, TOA_SW_DWN, and ALLSKY_SFC_SW_DWN variables and removed the first 5 observations from the dataset, resulting in 487 observations across the datasets.

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

Table 1. *Description of weather variables used in the study*

Variable name	Abbreviation	Unit	Resolution	Citation
Temperature	T2M	◦C	∼50 km	https://power.larc.nasa.gov/data-access-viewer/
Total precipitation	PRECTOTCORR_SUM	mm	∼50 km	
Wind speed	WS2M	m s^{-1}	∼50 km	
Wind direction	WD2M	Degrees	∼50 km	
Surface pressure	PS	kPa	∼50 km	
Relative humidity	RH2M	%	∼50 km	
Cloud cover	CLOUD_AMT	%	∼50 km	
Soil moisture	GWETPRO	–	∼50 km	
TOA shortwave downward irradiance	TOA_SW_DWN	MJ m^{-2} day^{-1}	∼50 km	
All-sky surface shortwave irradiance	ALL_SKY_SFC-SW_DWN	MJ m^{-2} day^{-1}	∼50 km	

Table 2. SPI interpretation (*Lloyd-Hughes and Saunders, 2002*)

SPI value	Significance
2.0 and +	Extremely wet
1.5 1.99	Very wet
1.0 to 1.49	Moderately wet
−0.99 to 0.99	Near normal
−1 to −1.49	Moderately dry
−1.5 to −1.99	Severely dry
−2 and less	Extremely dry

Moreover, using Pearson correlation for feature selection, variables with an absolute value of <10% were removed, leading to datasets of seven variables (Malanville, Karimama, and Banikoara) and eight variables (Segbana, Kandi, and Gogounou). Furthermore, time series data exhibit relationships over historical periods (Stübinger and Adler, 2020). To capture these patterns, we generated lagged versions of each variable using a window size of 5 (considering the use of SPI6 of the last 5 months to predict the sixth), creating 5 new lagged variables (t−5, t−4, t−3, t−2, and t−1) for each original covariate, leading to a total of 42 (for 7 variables) and 48 (for 8 variables) variables when including current time (t) observations.

Subsequently, using the Pearson correlation matrix, we performed feature selection by retaining the lag (t−5, t−4, t−3, t−2, t−1, or t) with the highest absolute Pearson correlation value with SPI6 (Table 3 and Figure 3). Finally, to normalize the scale of all covariates and eliminate the influence of variables with larger magnitudes, we applied a min-max scaler, bringing all covariates into the same range $[-1, 1]$.

2.3. Modeling

We employed a range of models to forecast time series data, encompassing both traditional ML methods and DL architectures.

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

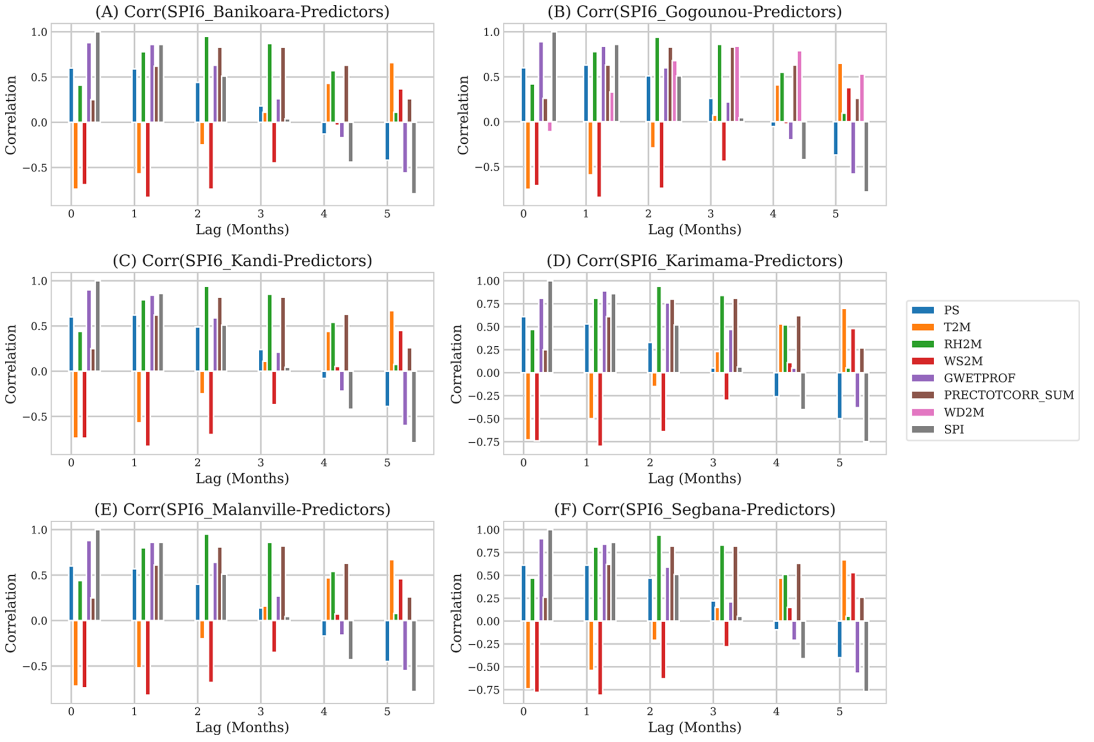


Figure 3. *Barplots of the correlations between SPI6 and lagged weather variables.*

Table 3. *Lag retained per covariate and locality*

	PS	T2M	RH2M	WS2M	GWETPROF	PRECTOTCORR_SUM	SPI6	WD2M
Banikoara	t	t	t−2	t−1	t	t−3	t−1	
Gogounou	t−1	t	t−2	t−1	t	t−3	t−1	t−3
Kandi	t−1	t	t−2	t−1	t	t−3	t−1	t−3
Karimama	t	t	t−2	t−1	t−1	t−3	t−1	
Malanville	t	t	t−2	t−1	t	t−3	t−1	
Segbana	t−1	t	t−2	t−1	t	t−3	t−1	t−3

Note. Blank cells indicate that none of the lags for that covariate had a Pearson correlation coefficient >0.1, leading to their removal.

2.3.1. Presentation of models

- **Linear regression**

  Linear regression by Huang (2020) is a fundamental statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship of the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

  where y represents the predicted value of the target variable, β_0 is the intercept term (the value of y when all predictors are zero), β_i are the coefficients associated with each independent variable x_i (for $i = 1, \ldots, n$), which quantify the effect of each predictor on the target, and ε is the error term that captures the unexplained variability or noise in the data. The model estimates coefficients that minimize the difference between predicted and actual values. It is widely used as a baseline due to its

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

simplicity, interpretability, and low computational cost. However, in the context of time-series forecasting, its performance can be limited. It does not account for temporal dependencies or autocorrelation between observations, and it cannot model seasonality or long-term trends unless these are explicitly encoded as features. Consequently, its relevance is constrained when the data exhibit complex time-based dynamics.

- **Support vector regression (SVR)**
  SVR by Basak et al. (2007) adapts the principles of support vector machines to regression problems. It aims to find a function of the form:

$$f(x) = \langle w, x \rangle + b$$

  that approximates the data within a specified tolerance level ε, while keeping the model as flat as possible. In this expression, $\langle w, x \rangle$ denotes the dot product between the weight vector w and the input vector x, which defines the orientation of the regression hyperplane. The scalar b is the bias term that shifts the hyperplane vertically. The parameter ε defines an ε-insensitive margin around the predicted values, within which no penalty is assigned for prediction errors. The objective of SVR is to find the flattest possible function (i.e., minimizing $\|w\|^2$) that fits as many data points as possible within this margin. SVR is known for its robustness to outliers and its ability to model nonlinear relationships through kernel functions. Despite these strengths, SVR has several drawbacks in time-series applications. It is computationally expensive for large datasets and sensitive to the choice of kernel and hyperparameters. Moreover, SVR does not inherently capture sequential dependencies in data, making it less effective for long-range forecasting unless temporal structures are carefully embedded into the feature space.
- **Random forest**
  Random forest by Breiman (2001) is an ensemble learning method that constructs multiple decision trees using bootstrapped samples of the data and aggregates their outputs for prediction. For regression tasks, the final output is typically the average of all tree predictions:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(x)$$

  where each f_t is the prediction from an individual decision tree. This approach reduces variance and overfitting while capturing nonlinear relationships between inputs and outputs. While random forest is powerful and robust, it lacks temporal awareness. It does not consider the order of data points, and it cannot extrapolate future trends beyond the training range. To be effective for time series, it relies heavily on the inclusion of lag features or engineered time-based inputs.
- **eXtreme gradient boosting (XGBoost)**
  XGBoost by Chen and Guestrin (2016) is an optimized implementation of gradient boosting that combines decision trees in a sequential manner, where each new tree corrects the residuals of the previous ones. It is known for its speed, regularization capabilities, and strong performance on structured data. It handles missing values well and includes several techniques to prevent overfitting. However, XGBoost is not specifically designed for time-series data. Without careful feature engineering (e.g., time lags and trend indicators), it cannot model temporal dependencies directly. Additionally, it may overfit if the model complexity is not well-controlled, especially when dealing with small or noisy datasets.
- **Light gradient boosting machine (LightGBM)**
  LightGBM by Ke et al. (2017) is a gradient-boosting framework that improves upon earlier methods like XGBoost through optimizations, such as Gradient-based One-Side Sampling and Exclusive Feature Bundling. These enhancements make it highly efficient in terms of training speed and memory usage, mainly for large datasets with many features. Like other tree-based models, LightGBM can capture nonlinear interactions and works well with heterogeneous data. However,

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

it does not intrinsically understand time-based relationships and, like XGBoost, depends on handcrafted temporal features. Its aggressive sampling strategy may also bias the model toward frequent or recent events, which can be problematic for forecasting rare or long-term patterns.

- **One-dimensional convolutional neural network (Conv1D)**
  Conv1D by Kiranyaz et al. (2021) is a type of Convolutional Neural Network (CNN) model designed to process sequential data using one-dimensional convolutional filters. These filters slide over the time dimension to detect local patterns and short-term dependencies. The model can extract meaningful features from fixed-length windows and is typically composed of multiple layers, including convolutions, nonlinear activation functions, and pooling operations, followed by fully connected layers for prediction. While Conv1D is computationally efficient and effective at identifying localized temporal patterns, it is naturally limited in its ability to capture long-term dependencies. It also requires consistent input lengths and may struggle with nonstationary data unless combined with deeper or recurrent structures.
- **Long short-term memory networks (LSTMs)**
  LSTM by Hochreiter and Schmidhuber (1997) is a specialized type of recurrent neural network (RNN) designed to handle the limitations of traditional RNNs in learning long-term dependencies. It introduces memory cells and gating mechanisms that control the flow of information across time steps, allowing the model to retain and forget information as needed. LSTMs are well-suited for sequential data with complex temporal relationships, such as time series with trends and seasonality. However, they come with high computational costs, especially for long sequences or deep architectures. They also require large datasets to generalize well and are prone to overfitting if regularization is not carefully applied.
- **Gated recurrent units (GRUs)**
  GRU by Chung et al. (2014) is a variant of LSTM that simplifies the network architecture by merging the input and forget gates into a single update gate. This leads to faster training and reduced memory usage while still allowing the model to capture temporal dependencies. GRUs have been shown to perform comparably to LSTMs on many time-series tasks, especially when computational efficiency is a concern. However, their simplified structure may limit their ability to model very long-term dependencies. Like LSTMs, they also require careful tuning and regularization to avoid overfitting, specifically when training data are scarce or highly variable.
- **Conv1D-LSTM**
  Conv1D-LSTM by Shi et al. (2015) is a hybrid model that combines one-dimensional convolutional layers with LSTM units. The convolutional layers serve to extract local features and reduce the dimensionality of the input sequence before passing it to the LSTM layers, which model the temporal evolution of these features. This architecture is particularly effective for spatiotemporal forecasting, where both short-term patterns and long-term dependencies are essential. It leverages the strengths of both CNNs and RNNs, but also inherits their complexities. The model requires more computational resources and careful tuning, and its interpretability is lower compared to simpler models. Overfitting can also be an issue if the dataset is not adequately large or diverse.

2.3.2. Experimental setup

For our experiments, across all localities, the dataset was split at the date January 31, 2014, resulting in a training dataset of length 387 and a test dataset of 96. Before introducing the ML models, all features and the target variable were normalized using min-max scaling, applied only on the training set to prevent data leakage. The fitted scaler was then applied to the test set for evaluation of the models. The transformation follows the exact formula implemented in scikit-learn's `MinMaxScaler` (Pedregosa et al., 2011:

$$X_{std} = \frac{X - X_{min}}{X_{max} - X_{min}}; \quad X_{\text{scaled}} = X_{std} \times (\text{max_range} - \text{min_range}) + \text{min_range}$$

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

In our case, we used min_range = −1 and max_range = 1. The ML models built for these datasets are linear regression, ridge regression, random forest, XGBoost, and LightGBM, using the default parameters from the scikit-learn, XGBoost, and LightGBM libraries.

Before implementing DL models, the dataset was reshaped into the format $(n_{\text{elements}}, \text{sequence_length}, n_{\text{features}})$, where n_{elements} is the number of samples, sequence_length is the fixed window size, and n_{features} is the number of predictors. In this case, the window size was set to 5, and seven features were used, resulting in an input shape of $(n,5,7)$.

To ensure robust training, 20% of the training set was reserved for validation. Normalization was performed using a MinMaxScaler fitted on the training set, with the same scaling parameters applied to the validation and test sets. All DL models were developed using the Keras API within TensorFlow. A consistent training setup was applied across all models and localities: a batch size of 64, the Adam optimizer with a learning rate of 0.001, mean squared error as the validation loss function, and RMSE as the evaluation metric. Monitoring the validation loss during training showed that convergence was typically achieved just before the 100th epoch across all localities and models, so the number of epochs was set to 100.

The implemented architectures include a Conv1D model consisting of a 1D convolutional layer with 64 filters and a kernel size of 1, followed by dense layers (see Table 4); an LSTM model composed of two stacked LSTM layers with 64 and 32 units (Table 5); a GRU variant using the same configuration as the LSTM but with GRU layers (Table 6); and a hybrid Conv1D-LSTM model that starts with a Conv1D layer (32 filters and kernel size 3), followed by two LSTM layers with 64 and 32 units, and dense layers (Table 7).

All experiments were conducted on a machine running Ubuntu 24.04.2 LTS (64-bit), with Python 3.12.3 and TensorFlow 2.16.2. The system features a 13th Gen Intel® Core™ i7−13620H CPU (Central Processing Unit) with 16 GB RAM (Random Access Memory) and an NVIDIA GeForce RTX™ 4060 Laptop GPU (Graphics Processing Unit) with 8 GB of memory.

2.4. Uncertainty quantification and model ranking

After carefully selecting ML and DL algorithms for drought forecasting, we sought to quantify the uncertainty associated with their predictions rigorously. Reliable uncertainty quantification is crucial for

Table 4. *Architecture of the Conv1d model used in the study*

Layer type	Key parameters	Output shape
Input	input_shape = (5, 7)	(None, 5, 7)
Conv1D	filters = 64, kernel_size = 5, activation = ReLU	(None, 5, 64)
Flatten	–	(None, 320)
Dense_0	units = 64, activation = ReLU	(None, 64)
Dense_1	units = 32, activation = ReLU	(None, 32)
Dense_2	units = 1, activation = Linear	(None, 1)

Table 5. *Architecture of the LSTM model used in the study*

Layer type	Key parameters	Output shape
Input	input_shape = (5, 7)	(None, 5, 7)
LSTM_0	units = 64, return_sequences = True	(None, 5, 64)
LSTM_1	units = 32, return_sequences = False	(None, 32)
Dense_0	units = 32, activation = ReLU	(None, 32)
Dense_1	units = 1, activation = Linear	(None, 1)

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

Table 6. *Architecture of the GRU model used in the study*

Layer type	Key parameters	Output shape
Input	input_shape = (5, 7)	(None, 5, 7)
GRU_0	units = 64, return_sequences = True	(None, 5, 64)
GRU_1	units = 32, return_sequences = False	(None, 32)
Dense_0	units = 32, activation = ReLU	(None, 32)
Dense_1	units = 1, activation = Linear	(None, 1)

Table 7. *Architecture of the Conv1D-LSTM hybrid model used in the study*

Layer type	Key parameters	Output shape
Input	input_shape = (5, 7)	(None, 5, 7)
Conv1D	filters = 64, kernel_size = 5, activation = ReLU	(None, 5, 64)
LSTM	units = 64, return_sequences = False	(None, 64)
Dense_0	units = 32, activation = ReLU	(None, 32)
Dense_1	units = 1, activation = Linear	(None, 1)

informed decision-making in drought management, allowing stakeholders to assess forecast confidence and anticipate possible variations. To achieve this goal, we integrated the selected predictive algorithms into the conformal prediction (CP) framework initially introduced by Vovk et al. (2005).

However, traditional CP methods for regression tasks assume data points are exchangeable, an assumption inherently violated by time-series data due to their temporal dependencies (Barber et al., 2023; Xu and Xie, 2023). Therefore, applying standard conformal methods would yield unreliable uncertainty estimates for our drought-forecasting problem.

To overcome this fundamental challenge, we adopted the EnBPI CP method proposed by Xu and Xie (2023), specifically tailored for time-series prediction tasks with nonexchangeable data points. This method leverages ensemble modeling and bootstrap resampling to construct robust prediction intervals without relying on the strict assumption of exchangeability.

2.4.1. Conformal prediction

CP (Vovk et al., 2005) is a flexible statistical framework for constructing prediction intervals or sets that contain the true outcome with a guaranteed probability. Given a training dataset $\{(x_i, y_i)\}_{i=1}^{n}$ and a regression model $\hat{f}$, CP assigns a nonconformity score to each training example, typically computed as the absolute residual:

$$\epsilon_i = |y_i - \hat{f}_{-i}(x_i)|, \tag{2}$$

where $\hat{f}_{-i}$ denotes the model trained with the ith data point excluded (leave-one-out). For a new input x_{n+1}, the prediction interval is given by:

$$C_{1-\alpha}(x_{n+1}) = \left[\hat{f}(x_{n+1}) \pm q_{1-\alpha}\right], \tag{3}$$

where $q_{1-\alpha}$ is the $(1-\alpha)$ quantile of the empirical distribution of $\{\epsilon_i\}$. Under the assumption of exchangeability, this interval satisfies the marginal coverage guarantee:

$$\mathbb{P}(y_{n+1} \in C_{1-\alpha}(x_{n+1})) \geq 1 - \alpha. \tag{4}$$

However, in time-series applications, the assumption of exchangeability is typically violated, requiring more robust methods.

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

2.4.2. Ensemble batch prediction interval

The EnBPI (Xu and Xie, 2023) method modifies traditional conformal inference by incorporating ensemble methods and bootstrap techniques, effectively addressing the sequential dependencies in time-series data. The algorithm operates as follows:

- **Bootstrap resampling**: The training dataset is bootstrapped multiple times (30 bootstrap samples in our study). Each bootstrap sample simulates different potential outcomes of the underlying data-generating process.
- **Ensemble model construction**: The selected ML and DL algorithms are independently trained on each bootstrap sample, creating an ensemble of diverse predictive models. This ensemble captures a broad spectrum of possible data behaviors and prediction outcomes.
- **Batch-wise conformalization**: Prediction intervals are computed batch-wise, considering historical residuals from previous batches. Residuals (absolute errors between observed and predicted values) from previous batches are used to construct empirical quantiles, resulting in robust and adaptive prediction intervals for future observations.

For this study, we implemented EnBPI using the Python library `Puncc`, with a significance level $\alpha = 0.1$, corresponding to a 90% coverage level, meaning theoretically 90% of prediction intervals should contain true observed values. While EnBPI offers a practical and adaptive solution for generating prediction intervals in time-series forecasting, it is not without limitations. A key drawback is its computational cost: since the method requires training multiple models on bootstrapped datasets, it may face scalability issues when applied to very large datasets or high-dimensional input spaces.

2.4.3. Evaluation metrics

To evaluate model performance, we consider both classical regression metrics and uncertainty quantification-specific criteria. This dual evaluation allows us to assess not only the accuracy but also the trustworthiness and informativeness of predictions.

2.4.4. Regression metrics

- ***R*²** by Wright (1921): The R^2 metric measures the proportion of variance in the dependent variable that is predictable from the independent variables. It is calculated as:

$$R^2 = 1 - \frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2}, \tag{5}$$

  where y_t are the actual values, $\hat{y}_t$ are the predicted values, and $\bar{y}$ is the mean of the actual values. R^2 ranges from 0 to 1, with higher values indicating a better fit.
- **RMSE** by Adhikari and Agrawal (2013): RMSE is a widely used metric that measures the square root of the average squared differences between predicted and actual values. It is given by:

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}. \tag{6}$$

  RMSE is sensitive to large errors and provides a measure of the magnitude of the prediction errors.
- **Mean absolute error (MAE)** by Adhikari and Agrawal (2013): MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It is calculated as:

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}|y_t - \hat{y}_t|. \tag{7}$$

  MAE gives an easily interpretable measure of the average prediction error in the same units as the target variable.

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

Uncertainty quantification metrics.

- **Coverage.** The fraction of true targets that fall within the predicted intervals. A well-calibrated model should satisfy:

$$\text{Coverage} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{y_t \in C_{1-\alpha}(x_t)\}. \tag{8}$$

- **Prediction interval width (Width).** This measures the average width of the prediction intervals. Narrower intervals are more informative:

$$\text{Width} = \frac{1}{T}\sum_{t=1}^{T} (\text{upper}_t - \text{lower}_t). \tag{9}$$

Coverage and interval width quantify (Stankevičiutė et al., 2021; Auer et al., 2023; Xu and Xie, 2023; Lee et al., 2024) the trade-off between reliability (i.e., how often intervals contain the true value) and informativeness (i.e., how narrow the intervals are). A trustworthy model should maintain nominal coverage while producing tight intervals.

Computational sustainability metric.

- **Carbon footprint (CO_2e):** We report the environmental cost of computation separately from predictive metrics. Emissions are estimated with `CodeCarbon` (Courty et al., 2024), developed by a consortium including MILA (Quebec AI Institute), Haverford College, Comet.ml, and BCG GAMMA. We use this library to monitor hardware power use (CPU/GPU/RAM) and record, for each model, the total emissions in kg CO_2e.

2.4.5. Model ranking via *Borda Count method*

To systematically compare and rank predictive model performance, we used the Borda Count ranking system, like many studies (Cherif and Idri, 2025; Hamdouchi and Idri, 2025; Nakach et al., 2025; Zerouaoui et al., 2025). The Borda Count is a consensus-based ranking method effective in combining multiple performance metrics. The method works as follows:

- Each performance metric (R^2, RMSE, MAE, Coverage, Width, and CO emissions) acts as a voter, ranking candidate models.
- For n candidate models, each model receives points based on its position in each metric: n points for first place, $n-1$ points for second, down to 1 point for last place.
- Points from all metrics are summed for each model; the model with the highest total score is considered the most effective and reliable across multiple evaluation dimensions.

This structured, transparent approach ensures comprehensive, balanced, and interpretable evaluations of model performance, explicitly accounting for accuracy, reliability, precision, and environmental considerations.

2.4.6. Taylor Diagram

This study uses the Taylor Diagram (Taylor, 2001) to provide a comprehensive statistical evaluation of the performance of ML and DL models. The Taylor Diagram framework enables simultaneous visualization of three critical performance metrics on a single polar coordinate plot. Unlike single-metric evaluation, the Taylor Diagram reveals trade-offs between correlation, centered root-mean-squared-difference, and standard deviation. It gives a simultaneous assessment of multiple models across complementary statistical dimensions on a single two-dimensional plot. The diagram focuses on centered pattern differences after removing overall bias, isolating structural model differences

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

from mean field errors. However, the Borda Count ranking method uses more metrics compared to the Taylor Diagram.

3. Results and discussion

Table 8 presents a comprehensive evaluation of various predictive models for drought forecasting in Malanville, integrating multiple performance dimensions including predictive accuracy (R^2, RMSE, and MAE), uncertainty quantification (Coverage and Width), and environmental impact (CO_2 emissions). The Borda Count ranking method synthesizes these metrics, providing a clear and balanced assessment framework.

The results clearly demonstrate the superior performance of the **Conv1D** model, which achieved the highest overall rank. With the best predictive accuracy metrics (highest $R^2 = 97.76\%$, lowest RMSE = 0.140, and MAE = 0.097), the Conv1D model effectively captures complex temporal and nonlinear patterns characteristic of drought time series. Furthermore, the high coverage rate (92.8%) indicates that this model provides reliable uncertainty quantification, ensuring that decision-makers can confidently interpret its drought predictions. Although the Conv1D model exhibited moderate interval widths and slightly elevated carbon emissions relative to simpler statistical models, the significant gains in accuracy and reliability clearly outweigh these trade-offs.

Furthermore, the quality of the uncertainty estimates produced by Conv1D is visually confirmed in Figure 4, which presents the prediction intervals generated by the model using the EnBPI method. The vast majority of the test set observations fall within the predicted intervals, with only a few instances lying outside. This indicates that the prediction intervals are not only well-calibrated but also adaptive to temporal fluctuations in the data. The intervals effectively capture the seasonal dynamics of the SPI/6 drought index, adjusting their width according to local uncertainty. This reinforces the robustness of the Conv1D model in providing accurate point predictions and reliable uncertainty quantification.

Closely following, the **Conv1D-LSTM** hybrid model ranks second, showcasing a compelling blend of convolutional and sequential learning strengths. It achieves strong predictive performance ($R^2 = 97.698\%$, RMSE = 0.142, and MAE = 0.098) and delivers slightly narrower prediction intervals compared to the Conv1D model, indicating higher precision. However, its marginally lower coverage (89.7%) suggests a small risk of missing some valid observations. The hybrid approach thus represents a robust alternative that effectively balances accuracy, uncertainty quantification, and moderate environmental impact.

Traditional statistical models, such as linear regression, ridge regression, and random forest, ranked among the lowest performers. Despite their lower computational demands and reduced CO_2

Table 8. *Performance of different models for Malanville*

Model	R^2(%)	RMSE	MAE	Coverage	Width	CO_2 EMISSIONS	Rank
Linear regression	96.008	0.187	0.157	0.906	0.592	3.44E−06	8
Ridge	96.061	0.186	0.157	0.906	0.600	3.58E−06	9
Random forest	95.745	0.194	0.150	0.927	0.701	1.13E−05	10
XGBoost	97.326	0.153	0.125	0.906	0.508	1.25E−05	4
LightGBM	97.252	0.156	0.130	0.927	0.524	1.62E−05	7
SVR	97.282	0.155	0.129	0.906	0.499	5.58E−06	5
Conv1D	**97.760**	**0.140**	**0.097**	**0.928**	**0.469**	**4.85E−04**	**1**
LSTM	97.562	0.146	0.101	0.835	0.405	8.73E−04	3
GRU	97.311	0.153	0.103	0.866	0.418	9.62E−04	6
Conv1D-LSTM	97.698	0.142	0.098	0.897	0.452	6.15E−04	2

Note. The best model according to the Borda Count ranking is highlighted in bold.

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

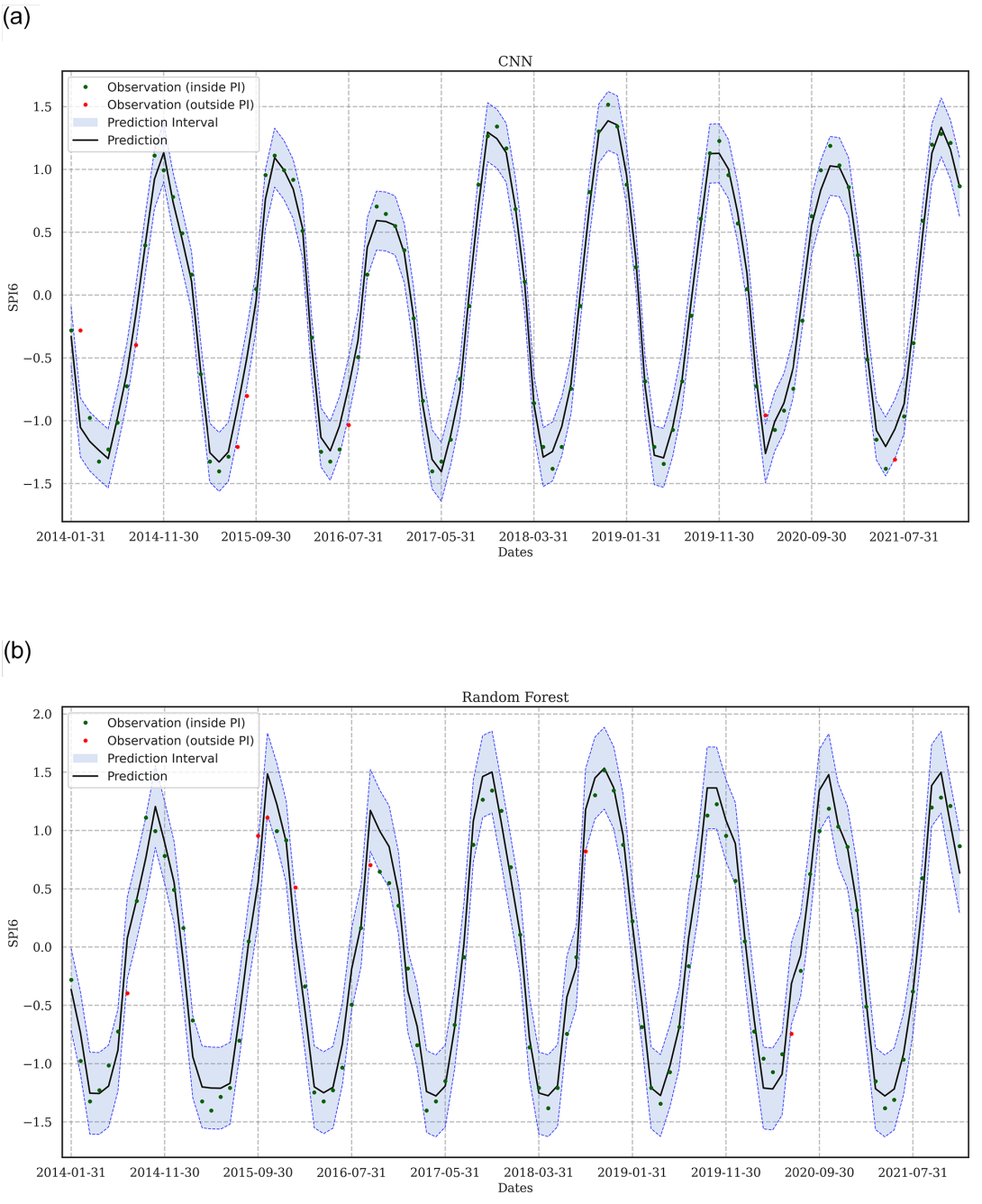


Figure 4. *Malanville's prediction intervals and observations of the most effective algorithm (a) and the least effective algorithm (b) using the EnBPI conformal prediction method for the SPI6. The shaded blue area represents the 90% prediction interval, the black line indicates the model's point predictions, green dots show observations within the interval, and red dots denote observations outside the interval.*

emissions, these methods failed to achieve comparable predictive accuracy and produced wider prediction intervals. Consequently, their practical utility for precise drought forecasting in critical applications is limited.

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

Moreover, DL architectures, such as LSTM and GRU, displayed promising predictive accuracy but incurred substantially higher carbon footprints. This increased environmental cost poses sustainability challenges, particularly relevant in resource-limited or environmentally sensitive contexts. Therefore, although these models have the potential to deliver strong predictive capabilities, their deployment must be carefully evaluated against available computational resources and environmental considerations.

Overall, the results in Figure 5 confirm the results obtained using the Borda Count ranking method, specifically in terms of the most effective and least effective models, and Conv1D shows better performance across all three metrics. DL architecture models show little variability among them and align better with observations compared to the ML models. The analysis underscores the effectiveness of convolutional-based DL models—particularly Conv1D and Conv1D-LSTM—for accurate, reliable, and environmentally balanced drought forecasts in Malanville.

In Karimama, the evaluation result summarized in Table 9 reveals two top-performing models with the top Borda Count rank: **Conv1D** and **Conv1D-LSTM**.

The **Conv1D** model exhibited the strongest overall predictive accuracy, achieving the highest R^2 ($R^2 = 97.072\%$), the lowest RMSE (0.153), and a notably low MAE (0.109). These results underscore the model's capability to model complex drought dynamics and provide reliable forecasts effectively. Additionally, its high coverage rate (92.8%) indicates robust uncertainty quantification, ensuring the constructed prediction intervals reliably capture the actual observations (see Figure 6). However, the interval width (0.475) and the model's relatively higher carbon footprint (1.17E−03) suggest inevitable trade-offs regarding precision and computational sustainability.

Equally ranked first, the **Conv1D-LSTM** hybrid model provided similarly impressive predictive performance, slightly trailing in terms of R^2 (96.917%) and RMSE (0.157), but surpassing Conv1D with the lowest MAE (0.107). Importantly, this model generated narrower prediction intervals (width = 0.463), improving precision, although with a slight reduction in coverage (89.7%). A key advantage of Conv1D-

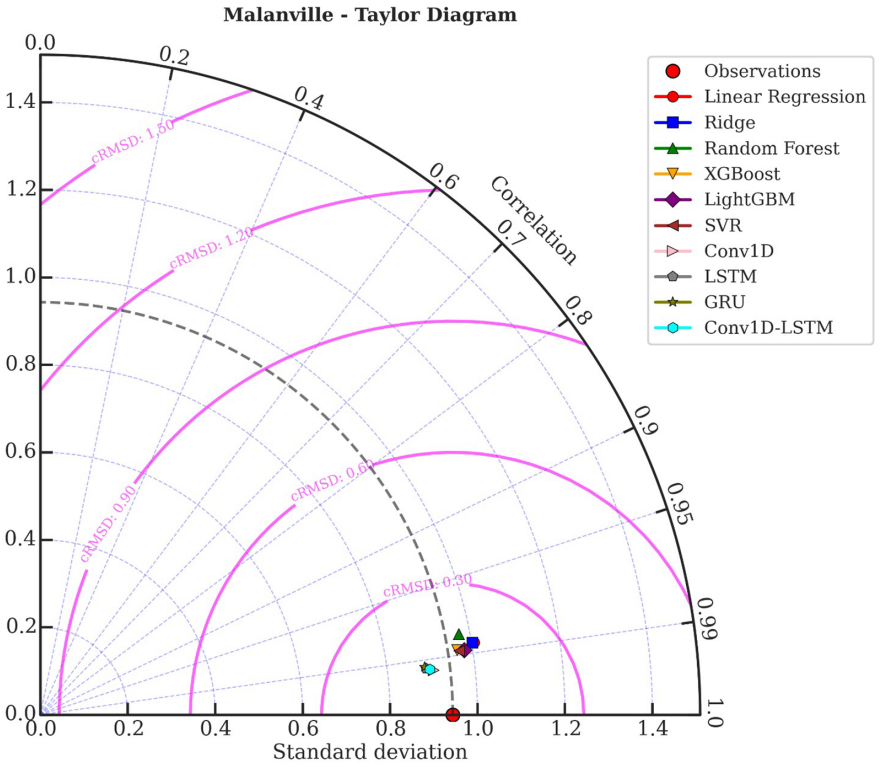


Figure 5. *Malanville's Taylor Diagram.*

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

Table 9. *Performance of different models for Karimama*

Model	R^2(%)	RMSE	MAE	Coverage	Width	CO_2 EMISSIONS	Rank
Linear regression	91.089	0.268	0.216	0.875	0.792	2.77E−06	10
Ridge	91.155	0.267	0.216	0.885	0.798	3.25E−06	9
Random forest	93.362	0.232	0.186	0.885	0.765	1.68E−05	8
XGBoost	96.031	0.179	0.144	0.906	0.620	5.23E−05	5
LightGBM	95.953	0.181	0.143	0.948	0.635	2.76E−05	4
SVR	95.552	0.190	0.154	0.938	0.664	4.63E−06	7
Conv1D	**97.072**	**0.153**	**0.109**	**0.928**	**0.475**	**1.17E−03**	**1**
LSTM	96.572	0.166	0.112	0.856	0.433	1.76E−03	6
GRU	96.581	0.165	0.113	0.876	0.457	1.17E−03	3
Conv1D-LSTM	**96.917**	**0.157**	**0.107**	**0.897**	**0.463**	**6.63E−04**	**1**

Note. The best model according to the Borda Count ranking is highlighted in bold.

LSTM is its lower environmental impact, with substantially reduced CO_2 emissions (6.63E−04), making it a particularly attractive option for sustainable forecasting applications.

Traditional models, such as linear regression and ridge regression, underperformed considerably across all metrics. Their lower accuracy (~91% R^2), wider prediction intervals, and suboptimal coverage highlight their limitations in addressing the complexity of drought-forecasting tasks. While these models have suitable computational efficiency, their relatively poor performance significantly constrains their practical use.

Intermediate approaches, such as **XGBoost**, **LightGBM**, and **SVR**, demonstrated moderate predictive accuracy and acceptable uncertainty quantification, but were consistently outperformed by the DL-based methods. Their intermediate rankings reflect a trade-off between computational simplicity and predictive robustness.

In conclusion, Figure 7 validates the results of the Borda Count ranking method. DL architecture models show little variability among them, and align better with observations compared to the ML models. The Karimama case study reinforces the effectiveness of convolution-based DL models, particularly **Conv1D** and **Conv1D-LSTM**, as optimal methods for accurate, reliable, and sustainable drought forecasting. Coupled with the EnBPI uncertainty framework, these models offer significant promise for practical decision-making and drought management strategies in regions facing similar environmental challenges.

In Kandi, the evaluation results summarized in Table 10 provide critical insights into the performance and suitability of various predictive models for drought forecasting, based on predictive accuracy (R^2, RMSE, and MAE), uncertainty quantification (Coverage and Width), and environmental impact (CO_2 emissions). Using the Borda Count ranking method, the **Conv1D-LSTM** hybrid model emerged as the highest-ranked algorithm, primarily due to superior regression performance metrics (highest R^2 = 97.417%, lowest RMSE = 0.173, and lowest MAE = 0.119). This highlights the strong capability of hybrid DL architectures in capturing complex drought patterns.

However, despite its high predictive accuracy, the Conv1D-LSTM model did not meet the targeted 90% coverage, achieving only 77.3%. This deficiency indicates a potential overconfidence in its predictions, as further suggested by its notably narrow prediction intervals (width = 0.391). Such narrow intervals, although indicative of high precision, may not accurately reflect the true predictive uncertainty and hence limit the model's reliability for practical decision-making. This limitation is visually confirmed in Figure 8, which displays the prediction intervals generated by the Conv1D-LSTM model using the EnBPI method. Figure 8 reveals a significant number of test set observations lying outside the prediction intervals, confirming the model's inability to encapsulate true values. Although the intervals appear adaptively narrow, they fail to provide adequate coverage, further underscoring the model's overconfidence in this specific locality.

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

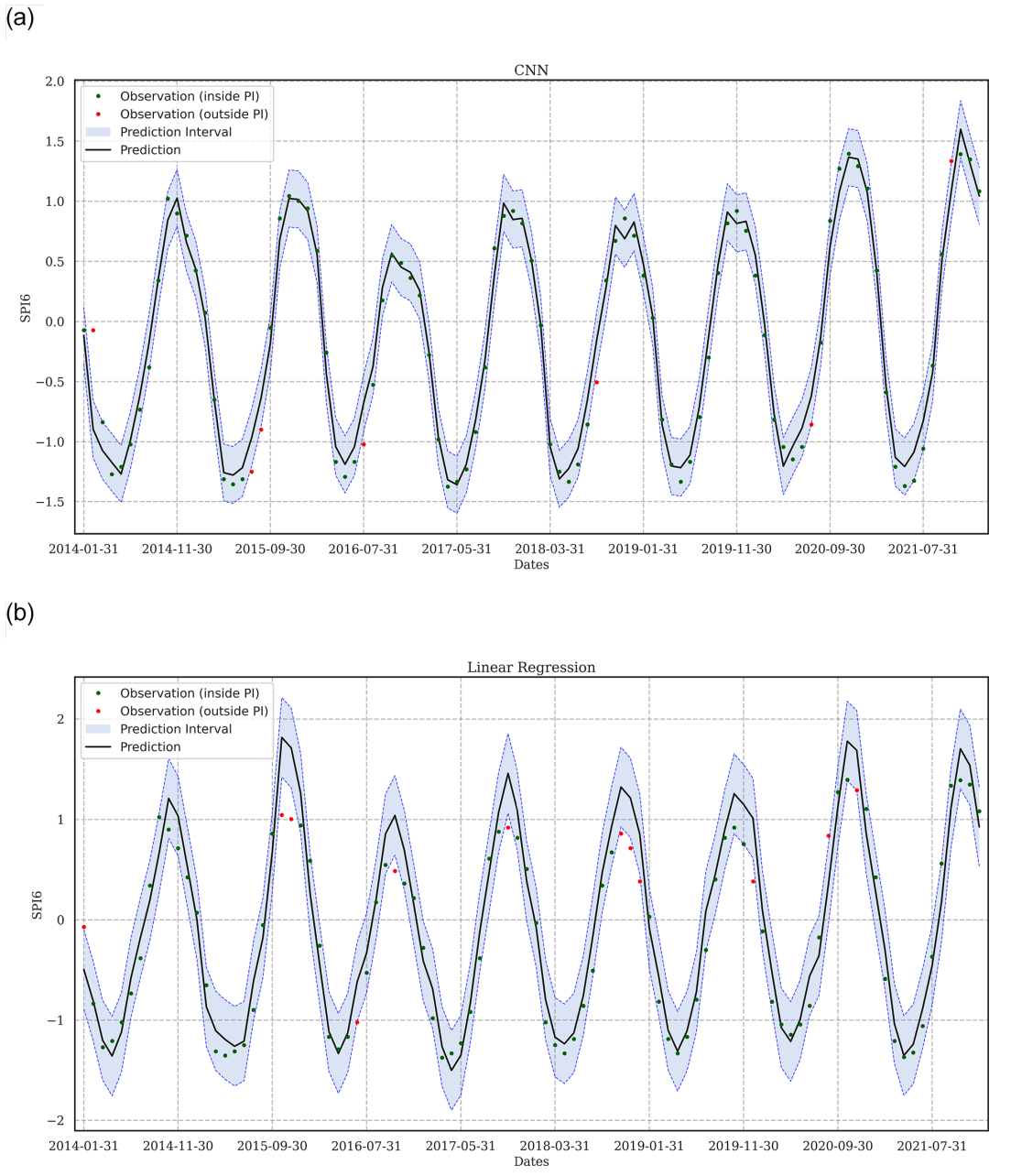


Figure 6. *Karimama's prediction intervals and observations of the most effective algorithm (a) and the least effective algorithm (b) using the EnBPI conformal prediction method for the SPI6. The shaded blue area represents the 90% prediction interval, the black line indicates the model's point predictions, green dots show observations within the interval, and red dots denote observations outside the interval.*

Conversely, traditional regression methods such as linear regression and ridge regression demonstrated relatively balanced performance, achieving coverage rates very close to the ideal 90% (~89.6%). Despite having moderately lower predictive accuracy metrics compared to DL models, these statistical methods exhibit desirable reliability in their uncertainty quantification. Their wider, more adaptive prediction

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

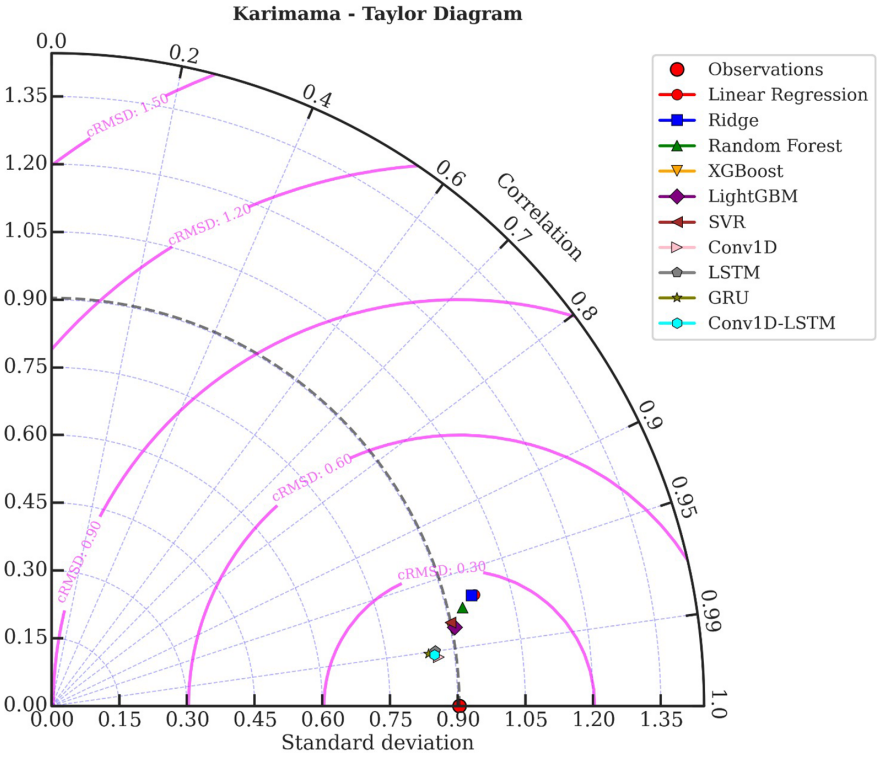


Figure 7. *Karimama's Taylor Diagram.*

Table 10. *Performance of different models for Kandi*

Model	R^2(%)	RMSE	MAE	Coverage	Width	CO_2 EMISSIONS	Rank
Linear regression	96.597	0.200	0.162	0.896	0.596	3.55E−06	7
Ridge	96.718	0.196	0.158	0.896	0.583	9.89E−06	5
Random forest	95.594	0.227	0.166	0.885	0.670	2.33E−05	10
XGBoost	96.883	0.191	0.143	0.812	0.466	5.77E−05	3
LightGBM	96.520	0.202	0.150	0.812	0.491	3.28E−05	8
SVR	95.947	0.218	0.152	0.812	0.482	4.24E−06	9
Conv1D	96.995	0.187	0.133	0.763	0.373	1.16E−03	2
LSTM	96.872	0.191	0.139	0.773	0.402	1.79E−03	6
GRU	97.164	0.181	0.126	0.742	0.393	2.12E−03	4
Conv1D-LSTM	**97.417**	**0.173**	**0.119**	**0.773**	**0.391**	**1.25E−03**	**1**

Note. The best model according to the Borda Count ranking is highlighted in bold.

intervals provide a realistic representation of prediction uncertainty, enhancing the confidence stakeholders can place in their forecasts.

Moreover, these traditional models offer significant computational advantages, reflected by substantially lower carbon footprints (~3.55E−06 and 9.89E−06 for linear regression and ridge regression, respectively), emphasizing their suitability for sustainable and efficient deployment in operational settings.

In summary, the Conv1D-LSTM model ranked first based on overall performance metrics; traditional regression approaches (linear and ridge regressions) might represent more suitable alternatives for

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

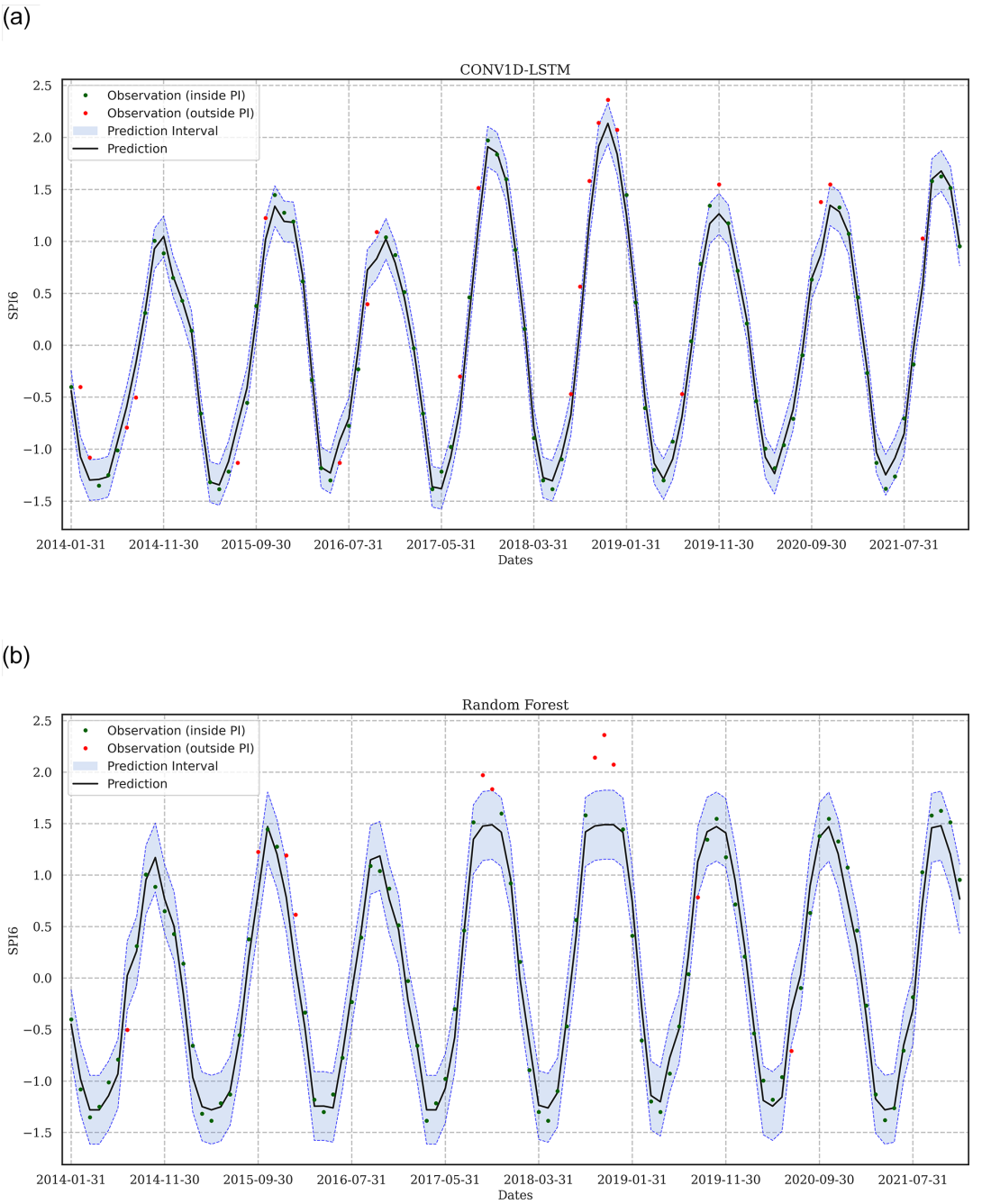


Figure 8. *Kandi's prediction intervals and observations of the most effective algorithm (a) and the least effective algorithm (b) using the EnBPI conformal prediction method for the SPI6. The shaded blue area represents the 90% prediction interval, the black line indicates the model's point predictions, green dots show observations within the interval, and red dots denote observations outside the interval.*

practical deployment. Additionally, Figure 9 validates the results of the Borda Count ranking method. The Conv1D-LSTM model aligns better with the observations than the other models, as it performs better across all three metrics with only marginal variability regarding the other DL models. Their reliable

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

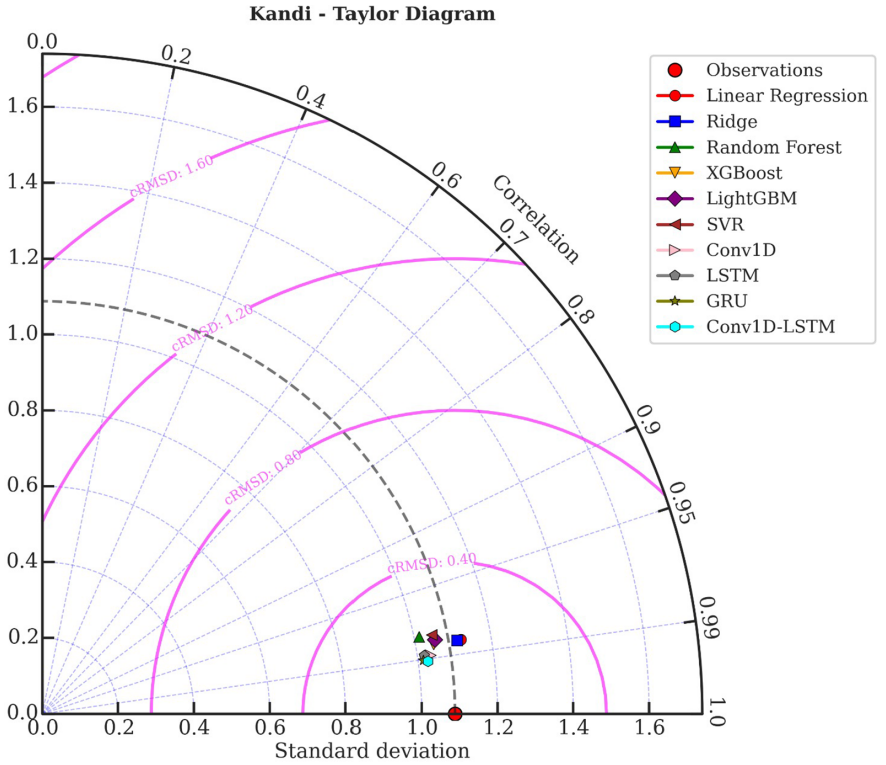


Figure 9. *Kandi's Taylor Diagram.*

uncertainty quantification, near-ideal coverage rates, computational efficiency, and environmental sustainability collectively provide stronger assurance for informed and confident decision-making in drought management applications.

In Gogounou, the evaluation results presented in Table 11 provide an in-depth comparison of different models' performances based on predictive accuracy (R^2, RMSE, and MAE), uncertainty quantification (Coverage and Width), and environmental impact (CO_2 emissions). According to the Borda Count ranking, the **Conv1D-LSTM** model secured the top position overall.

Table 11. *Performance of different models for Gogounou*

Model	R^2(%)	RMSE	MAE	Coverage	Width	CO_2 EMISSIONS	Rank
Linear regression	97.605	0.172	0.133	0.875	0.540	4.00E−06	2
Ridge	97.598	0.172	0.134	0.896	0.540	3.57E−06	3
Random forest	93.767	0.277	0.193	0.833	0.590	2.06E−05	10
XGBoost	96.464	0.208	0.152	0.812	0.482	5.96E−05	8
LightGBM	96.010	0.221	0.154	0.792	0.465	3.42E−05	9
SVR	94.943	0.249	0.151	0.812	0.459	6.37E−06	7
Conv1D	97.332	0.180	0.125	0.814	0.395	1.16E−03	6
LSTM	97.675	0.168	0.120	0.784	0.367	2.01E−03	5
GRU	97.767	0.165	0.112	0.763	0.347	2.08E−03	4
Conv1D-LSTM	**97.842**	**0.162**	**0.109**	**0.794**	**0.399**	**1.55E−03**	**1**

Note. The best model according to the Borda Count ranking is highlighted in bold.

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

The **Conv1D-LSTM** model achieved the highest predictive accuracy among all models ($R^2 = 97.842\%$), with the lowest RMSE (0.162) and MAE (0.109), demonstrating its superior capability in capturing complex patterns in drought time series. These accuracy metrics are notable. However, its uncertainty quantification was less satisfactory. The model recorded a coverage of only 79.4%, substantially below the 90% threshold typically expected from reliable prediction intervals. While it maintained a relatively narrow interval width (0.399), this may indicate overconfident predictions and a lack of adaptability in the uncertainty estimates—raising concerns about the robustness of its predictive intervals in operational settings (see Figure 10).

By contrast, traditional models such as **linear regression** and r**idge regression** ranked second and third, respectively, and showed a more balanced trade-off between predictive performance and uncertainty reliability. Despite slightly lower R^2 values (97.605 and 97.598%) and higher MAEs (0.133 and 0.134), both models offered good coverage (87.5 and 89.6%) and acceptable width (0.540), more faithfully capturing the uncertainty around their forecasts. Additionally, their carbon emissions were minimal (on the order of 10–6), making them significantly more computationally efficient and environmentally sustainable than DL-based alternatives.

Other DL models, such as **LSTM**, **GRU**, and **Conv1D**, exhibited strong predictive accuracy, with R^2 scores all above 97.3%, but they also suffered from insufficient coverage (ranging from 76.3 to 81.4%), reinforcing the concern that these architectures may generate overly narrow, overconfident prediction intervals. In particular, GRU and LSTM demonstrated relatively high emissions and low coverage, making them less suitable for operational deployment despite their precision.

To conclude, Figure 11 presents outcomes that confirm the results of the Borda Count ranking method, where the Conv1D-LSTM model aligns better with observations compared to the rest of the models. The **Conv1D-LSTM** model led in overall predictive performance and Borda ranking in Gogounou; its suboptimal coverage and overconfident uncertainty estimates call into question its trustworthiness in risk-sensitive applications. Alternatively, traditional models like l**inear regression** and r**idge regression** offer a better balance of accuracy, reliable uncertainty quantification, and minimal environmental impact, making them more pragmatic choices for reliable and sustainable drought forecasting in Gogounou.

In Banikoara, the results presented in Table 12 reveal the outstanding performance of DL models, particularly the hybrid **Conv1D-LSTM**, which ranked first according to the Borda Count method. This model achieved the highest $R^2 = 98.285\%$, the lowest RMSE (0.130), and the lowest MAE (0.084), marking it as the most accurate model for drought forecasting in this locality. In addition to its strong accuracy, it maintained a coverage level of 93.8% well above the 90% threshold, indicating a reliable uncertainty estimation. Its prediction interval width (0.435) was also among the narrowest, suggesting both high precision and trustworthy predictive intervals (see Figure 12).

This result is particularly noteworthy because it contrasts with observations in other regions where DL models tended to underperform in coverage. Here, the Conv1D-LSTM model demonstrated both high predictive power and reliable uncertainty quantification, making it a strong candidate for real-world deployment, despite its higher computational cost (CO_2 emissions of 1.47E−03). Its ability to maintain a balanced trade-off between accuracy and reliability sets it apart as a robust and practical solution for forecasting in Banikoara.

Other DL models, such as **Conv1D**, **LSTM**, and **GRU**, also performed competitively. Conv1D, for instance, achieved a very high $R^2 = 98.164\%$ and excellent uncertainty coverage (94.8%), reinforcing the strength of convolutional architectures in handling temporal data. However, it had a slightly wider interval (0.432) and marginally higher emissions than GRU and LSTM, which delivered similar levels of accuracy and competitive interval widths (both below 0.37), although with slightly lower coverage.

Traditional models like **SVR**, r**idge**, and **linear regressions** provided satisfactory results in terms of coverage (all around or above 90%) but trailed behind in predictive accuracy. Notably, SVR combined solid performance ($R^2 = 98.235\%$) with good coverage (91.7%) and moderate width (0.451), making it a reasonable choice when considering computational efficiency (emissions of 1.11E−05).

In summary, Figure 13 corroborates the findings of the Borda Count ranking method. The Conv1D-LSTM model aligns more closely with the observations than the other models, and the DL models show

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

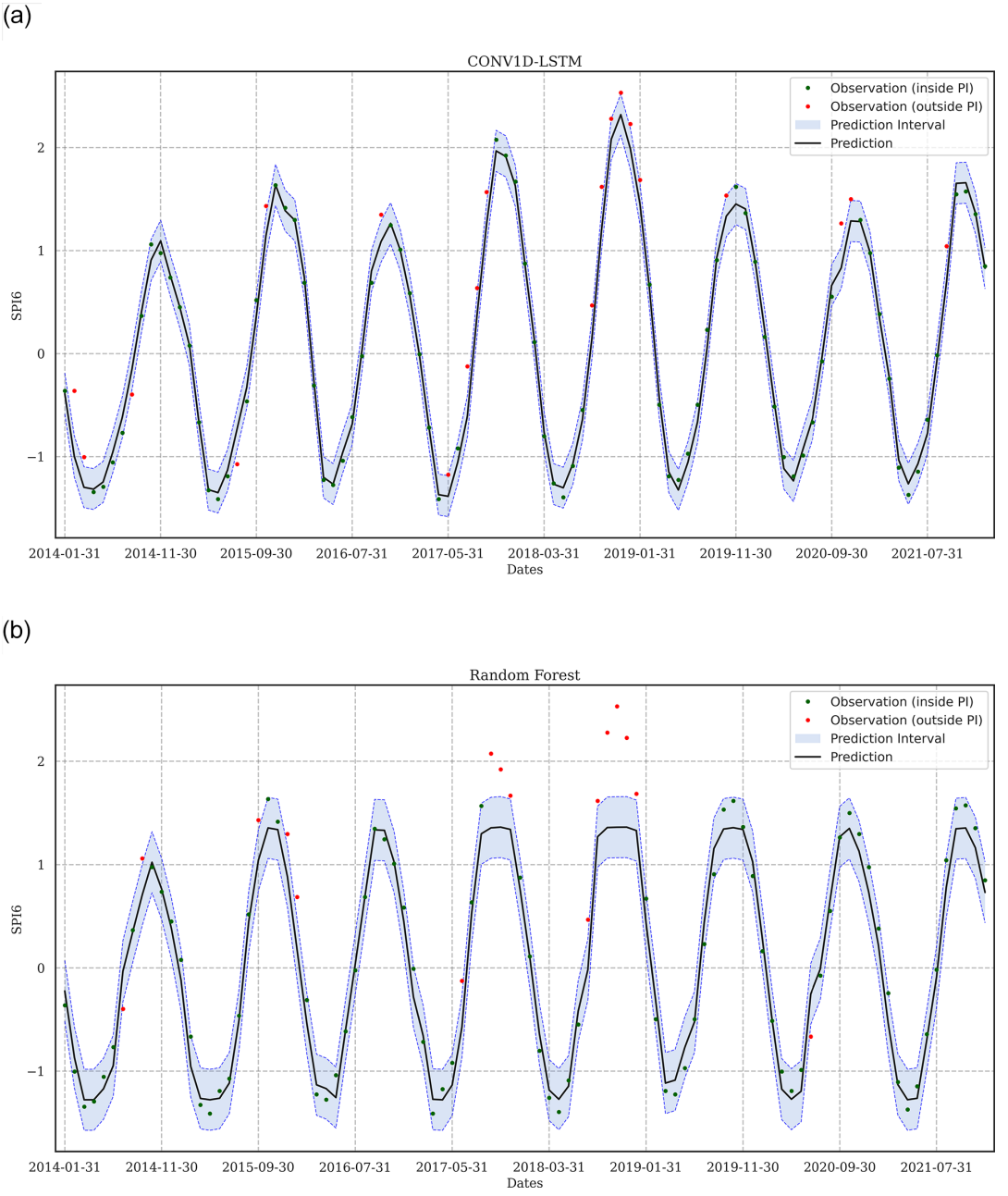


Figure 10. *Gogounou's prediction intervals and observations of the most effective algorithm (a) and the least effective algorithm (b) using the EnBPI conformal prediction method for the SPI6. The shaded blue area represents the 90% prediction interval, the black line indicates the model's point predictions, green dots show observations within the interval, and red dots denote observations outside the interval.*

little variability. Banikoara is the region in which the Conv1D-LSTM model achieves near-ideal performance across all evaluation criteria—predictive accuracy, interval reliability, precision, and consistency—thus validating its suitability for deployment in data-driven drought forecasting. While simpler models such as SVR offer good trade-offs for environments with fewer resources, DL models demonstrated their full potential in this context.

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

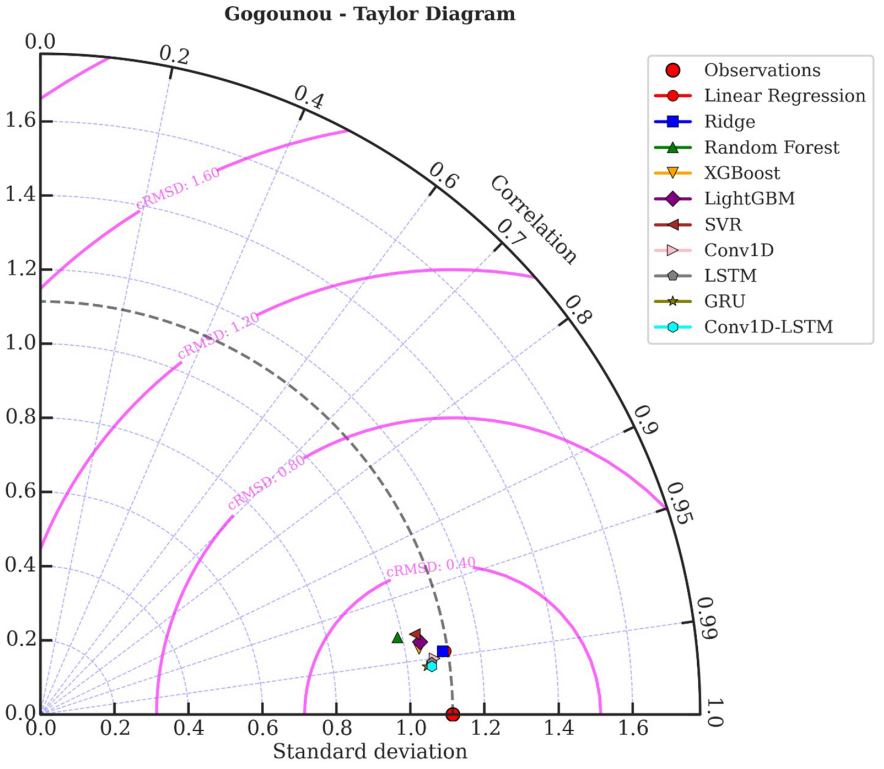


Figure 11. *Gogounou's Taylor Diagram.*

Table 12. *Performance of different models for Banikoara*

Model	R^2(%)	RMSE	MAE	Coverage	Width	CO_2 EMISSIONS	Rank
Linear regression	97.544	0.156	0.126	0.917	0.562	4.37E−06	8
Ridge	97.581	0.155	0.126	0.906	0.552	1.22E−05	9
Random forest	97.305	0.163	0.119	0.948	0.666	2.03E−05	10
XGBoost	97.989	0.141	0.105	0.906	0.505	6.50E−05	7
LightGBM	98.018	0.140	0.103	0.885	0.493	3.01E−05	6
SVR	98.235	0.132	0.104	0.917	0.451	1.11E−05	2
Conv1D	98.164	0.134	0.095	0.948	0.432	1.16E−03	3
LSTM	98.174	0.134	0.091	0.887	0.370	1.97E−03	4
GRU	98.128	0.135	0.090	0.856	0.365	1.97E−03	5
Conv1D-LSTM	**98.285**	**0.130**	**0.084**	**0.938**	**0.435**	**1.47E−03**	**1**

In Segbana, Table 13 reports the performance of multiple predictive models across regression metrics (R^2, RMSE, and MAE), uncertainty quantification (Coverage and Width), and environmental impact (CO_2 emissions). The Borda Count ranking identifies both **Conv1D** and **Conv1D-LSTM** as top-performing models, sharing the first position.

The **Conv1D-LSTM** model achieved the highest $R^2 = 96.613\%$, the lowest RMSE (0.209), and a competitive MAE (0.145), confirming its superior predictive capability in this setting. However, its uncertainty quantification was less reliable, with a coverage rate of only 72.2%, significantly below the standard 90% expectation. Despite producing relatively narrow intervals (width = 0.409), this suggests the model is potentially overconfident—failing to capture true outcomes consistently within its predictive

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

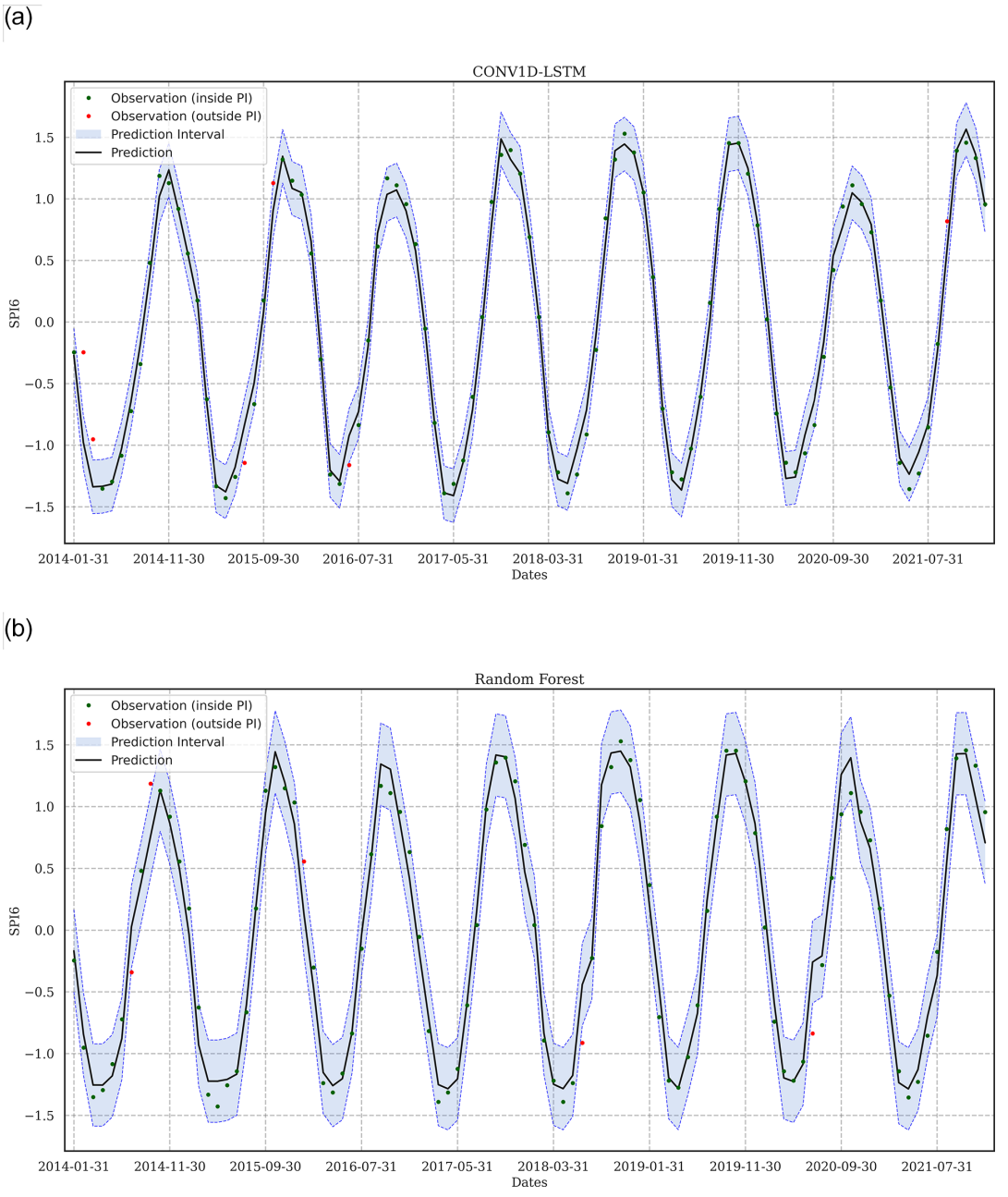


Figure 12. *Banikoara's prediction intervals and observations of the most effective algorithm (a) and the least effective algorithm (b) using the EnBPI conformal prediction method for the SPI6. The shaded blue area represents the 90% prediction interval, the black line indicates the model's point predictions, green dots show observations within the interval, and red dots denote observations outside the interval.*

bounds (see Figure 14). Moreover, its CO_2 emissions (1.42E−03) indicate a high computational cost, which may be limiting in real-world, resource-constrained applications.

Similarly, the **Conv1D** model performed very well in terms of predictive accuracy ($R^2 = 96.466\%$, RMSE = 0.214, and MAE = 0.148) and had the narrowest prediction intervals (width = 0.369). However,

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

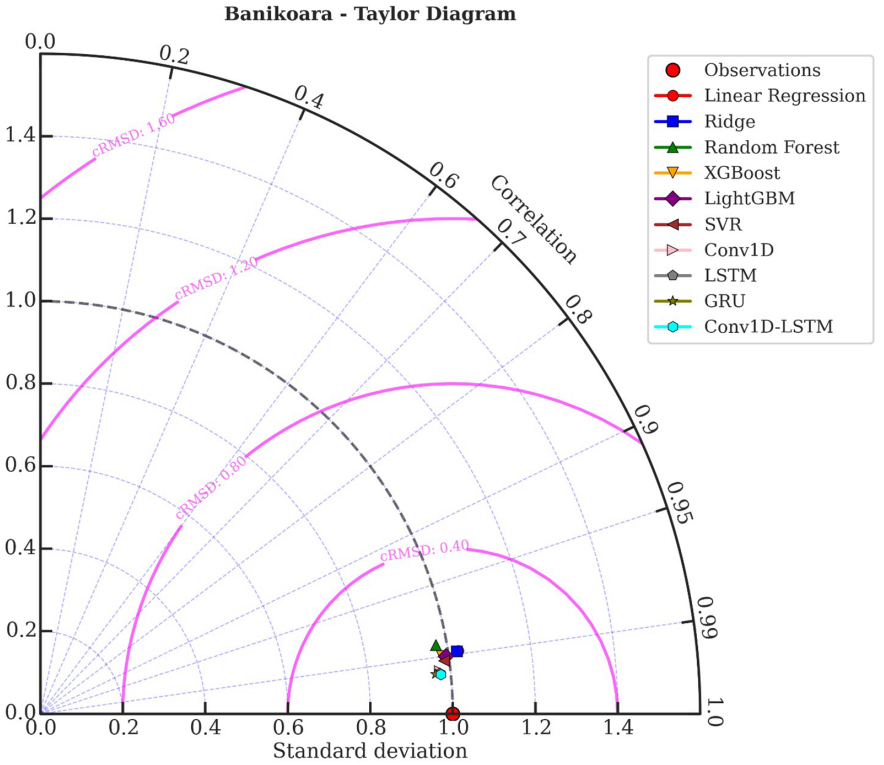


Figure 13. *Banikoara's Taylor Diagram.*

Table 13. *Performance of different models for Segbana*

Model	R^2(%)	RMSE	MAE	Coverage	Width	CO_2 EMISSIONS	Rank
Linear regression	95.480	0.243	0.192	0.802	0.590	4.41E−06	8
Ridge	95.743	0.236	0.185	0.823	0.601	8.44E−06	4
Random forest	94.862	0.259	0.188	0.833	0.621	1.64E−05	10
XGBoost	96.503	0.214	0.157	0.792	0.491	2.39E−05	3
LightGBM	95.643	0.239	0.175	0.771	0.484	2.43E−05	6
SVR	94.741	0.262	0.183	0.698	0.452	3.07E−06	9
Conv1D	**96.466**	**0.214**	**0.148**	**0.732**	**0.369**	**1.11E−03**	**1**
LSTM	95.874	0.231	0.162	0.691	0.387	1.68E−03	5
GRU	96.159	0.223	0.153	0.680	0.388	1.93E−03	4
Conv1D-LSTM	**96.613**	**0.209**	**0.145**	**0.722**	**0.409**	**1.42E−03**	**1**

Note. The best model according to the Borda Count ranking is highlighted in bold.

its coverage rate (73.2%) was also unsatisfactory, reaffirming the issue of overconfidence. Although this model has a slightly lower carbon footprint (1.11E−03) than Conv1D-LSTM, it still ranks among the more computationally expensive options.

On the other hand, traditional models such as r**idge regression** and l**inear regression** demonstrated more balanced performances. While their R^2 values were lower (95.7 and 95.5%, respectively), they produced higher coverage rates (82.3 and 80.2%, respectively), more consistent with trustworthy uncertainty quantification, although still below the desired 90% level. Their predictive intervals were

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

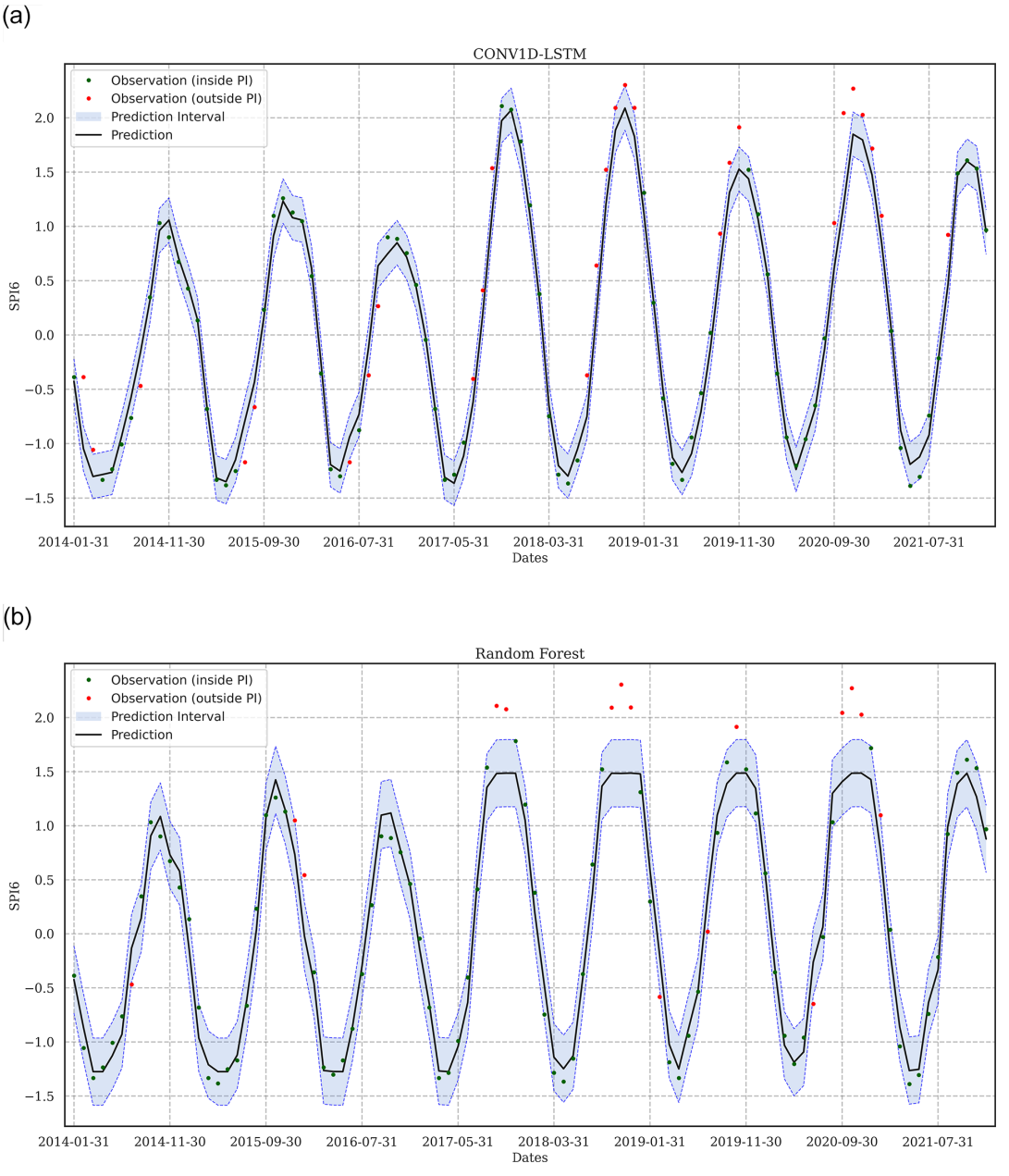


Figure 14. *Segbana's prediction intervals and observations of the most effective algorithm (a) and the least effective algorithm (b) using the EnBPI conformal prediction method for the SPI6. The shaded blue area represents the 90% prediction interval, the black line indicates the model's point predictions, green dots show observations within the interval, and red dots denote observations outside the interval.*

wider, but better reflected true variation in the data. Additionally, these models remain computationally efficient, with significantly lower emissions (e.g., 4.41E-06 for linear regression).

Models such as **XGBoost** and **LightGBM** delivered reasonable regression metrics but also struggled with low coverage, while **GRU**, **LSTM**, and **SVR** exhibited both low uncertainty reliability and higher emissions—making them less practical for deployment.

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

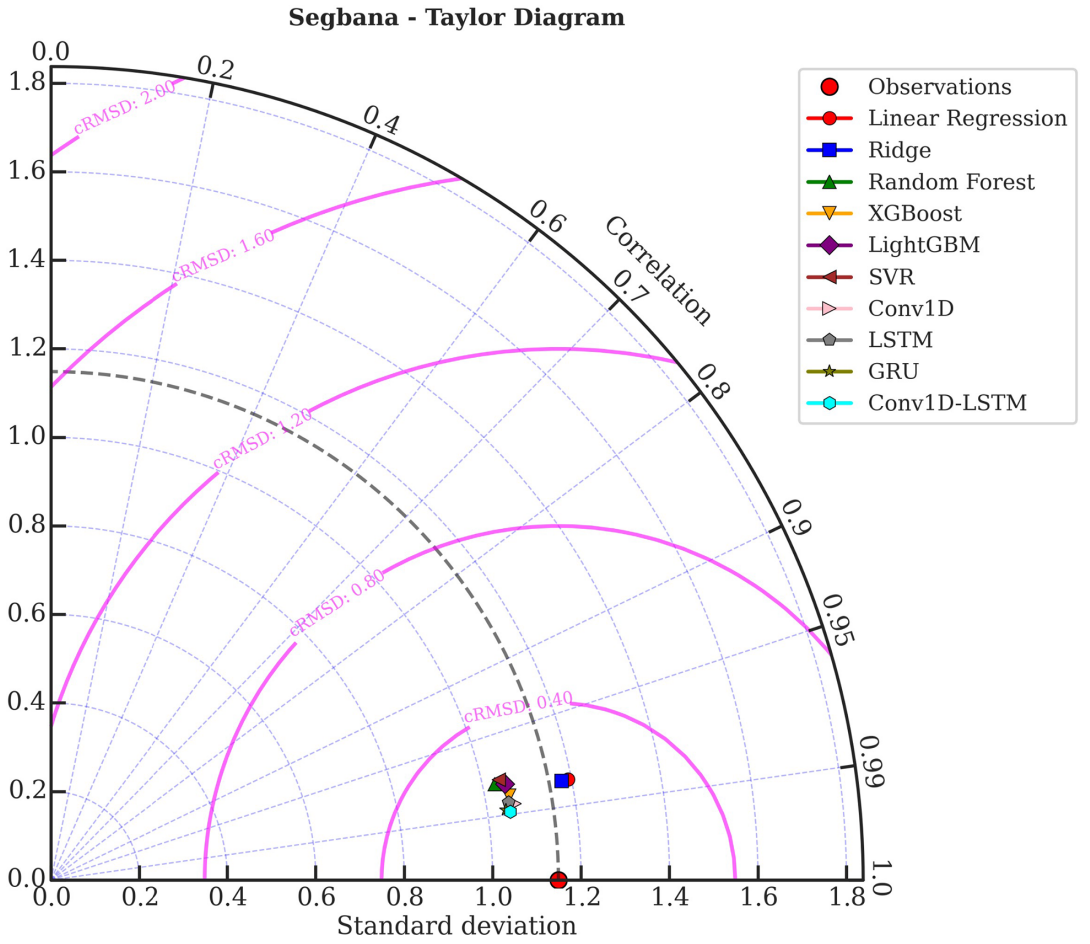


Figure 15. *Segbana's Taylor Diagram.*

In conclusion, while **Conv1D** and **Conv1D-LSTM** are ranked first in terms of overall performance, their low coverage values raise concerns about their robustness for uncertainty-aware forecasting. In contrast, simpler regression models like **ridge** or **linear regression**—although slightly less accurate—may offer more dependable and computationally efficient solutions for deployment in Segbana, especially when reliable uncertainty quantification is a priority. Figure 15 confirms the results of the Borda Count ranking method. Conv1D and Conv1D-LSTM models align better with observations compared to the other models.

4. Discussion and limitations

The prediction of SPI6 across several regions achieved impressive R^2 values of 98.29, 97.84, 97.76, 97.42, 96.61, and 97.07% in Banikoara, Gogounou, Malanville, Kandi, Segbana, and Karimama, respectively. A similar study conducted in these regions by Vodounon et al. (2022), using the same data, reported a significantly lower R^2 of 80.17% for SPI6, highlighting the superior performance of our models. This discrepancy may be due to the spatio-temporal aspects of the data not being fully accounted for in their study, potentially leading to less reliable results (Botache et al., 2023). When assessed across all locations, the DL Conv1D-LSTM model emerged as the best overall performer. Its combination of convolutional layers and recurrent networks enables the capture of both local patterns and temporal

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

dependencies, which is critical in weather forecasting where complex dynamics are involved (Shi et al., 2015). However, from a sustainability perspective, traditional ML models may be preferable in some cases. They often achieve comparable performance to DL models while generating far lower CO_2 emissions. For example, in Kandi, Conv1D-LSTM outperformed XGBoost by only 0.7% in R^2 but produced 126 times more CO_2 emissions. This raises the question of whether marginal accuracy gains justify significantly higher environmental costs, especially in the context of climate change mitigation. Beyond accuracy, this study emphasizes the importance of quantifying predictive uncertainty in agricultural applications. High accuracy does not imply high certainty; decision-makers should be informed of model uncertainty to fully understand the implications of using the forecasts. Several limitations should be noted. First, the analysis does not cover all agricultural regions of the country, leaving some areas without information. Second, the study focuses on retrospective predictions, whereas stakeholders require forward-looking drought forecasts to enable proactive measures. Third, the analysis is limited to short-term drought assessment using the 6-month SPI, without addressing longer-term drought conditions. Finally, the use of gridded datasets introduces inherent uncertainties, which may increase the overall error of the predictive models.

5. Conclusion

Droughts pose significant challenges to food security, particularly when they occur in major agricultural regions. This study focused on quantifying the uncertainty of drought predictions using ML and DL techniques across the six localities in the Alibori department of Benin. The results demonstrated high predictive accuracy for SPI6, with R^2 values of 98.29, 97.84, 97.76, 97.42, 96.61, and 97.07% in Banikoara, Gogounou, Malanville, Kandi, Segbana, and Karimama, respectively. The Borda count rankings highlighted the Conv1D-LSTM model as the top performer in two localities and second in another two, suggesting its strong suitability for spatiotemporal drought prediction. This analysis underscores the importance of considering model uncertainty when selecting the best predictive approach, as it allows for a balance between accuracy and robustness. Given the coarse resolution of the data used in this study (~50 km), future research should consider utilizing higher-resolution datasets to better capture the spatial variability within each locality, which may improve model performance and reliability.

Open peer review. To view the open peer review materials for this article, please visit http://doi.org/10.1017/eds.2025.10025.

Author contribution. Conceptualization: B.L., A.D.T., R.V., K.B.D.S., and A.I. Data curation: A.I., B.L., R.V., K.B.D.S., and A.D.T. Formal analysis: B.L., A.D.T., and A.I. Methodology: B.L., A.D.T., R.V., K.B.D.S., and A.I. Software: B.L., A.D.T., and A.I. Writing—original draft: A.D.T., K.B.D.S., B.L., and A.I. Writing—review and editing: K.B.D.S., B.L., A.I., and A.D.T. All authors approved the final submitted draft.

Competing interests. The authors declare none.

Data availability statement. The code for this study is archived on Zenodo at https://zenodo.org/records/16815747.

Funding statement. The authors did not receive any support from any organization for the submitted work.

References

Adhikari R and Agrawal RK (2013) An introductory study on time series modeling and forecasting. arXiv preprint, arXiv: 1302.6613.

Aïzansi AN, **Ogunjobi KO and Ogou FK** (2024) Monthly rainfall prediction using artificial neural network (case study: Republic of Benin). *Environmental Data Science 3*, e11.

Aryal JP, **Rahut DB and Marenya P** (2023) Do ex-ante and ex-post strategies to manage drought differ at the farm household level? A case of sub-Saharan Africa. *Environment, Development and Sustainability 25*(10), 10145–10169. https://doi.org/10.1007/s10668-023-03660-9.

Auer A, **Gauch M**, **Klotz D and Hochreiter S** (2023) Conformal prediction for time series with modern hopfield networks. *Advances in Neural Information Processing Systems*.

Barber RF, **Candès EJ**, **Ramdas A and Tibshirani RJ** (2023) Conformal prediction beyond exchangeability.

Basak D, **Pal S**, **Ch D and Patranabis R** (2007) Support vector regression. *Neural Information Processing—Letters and Reviews 11*, 203–224.

Botache D, **Dingel K**, **Huhnstock R**, **Ehresmann A and Sick B** (2023) Unraveling the complexity of splitting sequential data: tackling challenges in video and time series analysis. arXiv e-prints. arXiv:2307.14294.

Breiman L (2001) Random forests. *Machine Learning 45*, 5–32.

Chen T and Guestrin C (2016) Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*. ACM, pp. 785–794.

Cherif Y and Idri A (2025) Single and ensemble based filters in environmental data. *Expert Systems 42*(7), e70076. https://onlinelibrary.wiley.com/doi/pdf/10.1111/exsy.70076.

Chung J, **Gulcehre C**, **Cho K and Bengio Y** (2014) Empirical evaluation of gated recurrent meural networks on sequence modeling. https://doi.org/10.48550/arXiv.1412.3555.

Courty B, **Schmidt V**, **Luccioni S**, **Goyal-Kamal**, **Marion Coutarel**, **Feld B**, **Lecourt J**, **LiamConnell**, **Saboni A**, **Inimaz**, **supatomic**, **Léval M**, **Blanche L**, **Cruveiller A**, **Ouminasara**, **Zhao F**, **Joshi A**, **Bogroff A**, **de Lavoreille H**, **Laskaris N**, **Abati E**, **Blank D**, **Wang Z**, **Catovic A**, **Alencon M**, **Michał S**, **Bauer C, de**, **Araújo LON**, **JPW**, and **MinervaBooks** (2024) mlco2/codecarbon: v2.4.1.

Emmanuel LA, **Hounguè NR**, **Biaou CA and Badou DF** (2019) Statistical analysis of recent and future rainfall and temperature variability in the mono river watershed (Benin, Togo). *Climate 7*(1), 8.

FAO, IFAD, UNICEF, WFP, and WHO (2021) The state of food security and nutrition in the world 2021.

Gruber C, **Schenk PO**, **Schierholz M**, **Kreuter F and Kauermann G** (2023) Sources of uncertainty in supervised machine learning – A statisticians' view. arXiv preprint, arXiv:2305.16703. https://arxiv.org/abs/2305.16703

Halissou Y, **Eric AA**, **Eliézer BI**, **Ezéchiel O**, **Bio TD and Abel A** (2022) History and projection of hydrological droughts in the Benin basin. *Journal of Atmospheric Science Research 5*(2), 33–51.

Hamdouchi A and Idri A (2025) Empowering IoT security: Deploying TinyML ensemble techniques for cyberattack detection. *Scientific African 29*, e02809.

Hochreiter S and Schmidhuber J (1997) Long short-term memory. *Neural Computation 9*, 1735–1780.

Hounkpè J, **Diekkrüger B**, **Badou DF and Afouda AA** (2016) Change in heavy rainfall characteristics over the ouémé river basin. *Climate 4*(1), 15.

Huang M (2020) Theory and implementation of linear regression. In *2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, pp. 210–217.

Jalaian B, **Lee M and Russell S** (2019) Uncertain context: Uncertainty quantification in machine learning. *AI Magazine 40*(4), 40–49.

Kazadi A, **Doss-Gollin J**, **Sebastian A and Silva A** (2024) FloodGNN-GRU: A spatio-temporal graph neural network for flood prediction. *Environmental Data Science 3*, e21.

Ke G, **Meng Q**, **Finley T**, **Wang T**, **Chen W**, **Ma W**, **Ye Q and Liu T-Y** (2017) LightGBM: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, December *2017*, pp. 3149–3157.

Kiranyaz S, **Avci O**, **Abdeljaber O**, **Ince T**, **Gabbouj M and Inman DJ** (2021) 1D convolutional neural networks and applications: a survey. *Mechanical Systems and Signal Processing 151*, Article 107398. https://doi.org/10.1016/j.ymssp.2020.107398.

Lawin EA, **Houngue NR**, **Biaou CA and Badou DF** (2019) Statistical analysis of recent and future rainfall and temperature variability in the Mono River watershed (Benin, Togo). *Climate 7*(1), 8. https://doi.org/10.3390/cli7010008.

Lee J, **Xu C and Xie Y** (2024) Transformer conformal prediction for time series. arXiv preprint, arXiv:2406.05332. https://arxiv.org/abs/2406.05332.

Lilburne L, **Helfenstein A**, **Heuvelink GBM and Eger A** (2024) Interpreting and evaluating digital soil mapping prediction uncertainty: A case study using texture from soilgrids. *Geoderma 450*, 117052.

Lloyd-Hughes B and Saunders MA (2002) A drought climatology for europe. *International Journal of Climatology 22*, 1571–1592.

McKee TB, **Doesken NJ and Kleist J** (1993) The relationship of drought frequency and duration to time scales. In *Proceedings of the 8th Conference on Applied Climatology*, volume *17*(22), pp. 179-–183. https://climate.colostate.edu/pdfs/relationshipofdroughtfrequency.pdf.

Masih I, **Maskey S**, **Mussá FE and Trambauer P** (2014) A review of droughts on the african continent. *Hydrology and Earth System Sciences 18*, 3635–3649.

Mir YH, **Mir S**, **Ganie MA**, **Bhat JA**, **Shah AM**, **Mushtaq M and Irshad I** (2025) Overview of land use and land cover change and its impacts on natural resources. In Jatav HS, Raiput VD and Minkina T (eds), *Ecologically Mediated Development, Volume 41 of Sustainable Development and Biodiversity*, Singapore: Springer.

Mooers G, **Beucler T**, **Pritchard M and Mandt S** (2024) Understanding precipitation changes through unsupervised machine learning. *Environmental Data Science 3*, e3.

Nakach F-Z, **Idri A and Tchokponhoue GAD** (2025) Multimodal random subspace for breast cancer molecular subtypes prediction by integrating multi-dimensional data. *Multimedia Tools and Applications 84*(27), 32671–32703.

Oluwatayo I and Ojo A (2016) Is africa's dependence on agriculture the cause of poverty? *The Journal of Developing Areas 50*, 93–102.

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press

Oyerinde GT and Olowookere BT (2018) Observed shift and merge of hydrological regimes in the sota catchment. *International Journal of Research-GRANTHAALAYAH 6*(6), 205–211.

Paeth H, **Capo-Chichi A and Endlicher W** (2008) Climate change and food security in tropical West Africa. *Erdkunde 62*(2), 101–115.

Pedregosa F, **Varoquaux G**, **Gramfort A**, **Michel V**, **Thirion B**, **Grisel O**, **Blondel M**, **Prettenhofer P**, **Weiss R**, **Dubourg V**, **Vanderplas J**, **Passos A**, **Cournapeau D**, **Brucher M**, **Perrot M and Duchesnay E** (2011) Scikit-learn: Machine learning in python. *Journal of Machine Learning Research 12*, 2825–2830.

Pison G, **Couppié E and Caporali A** (2022) The population of the world, 2022. *Population Societies 603*(8), 1–8.

Poggio L, **De Sousa LM**, **Batjes NH**, **Heuvelink GBM**, **Kempen B**, **Ribeiro E and Rossiter D** (2021) Soilgrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *The Soil 7*, 217–240.

République du Bénin – Ministère de l'Agriculture, de l'Élevage et de la Pêche (MAEP) (2017) Plan Stratégique de Développement du Secteur Agricole (PSDSA): Orientations stratégiques 2025. *Plan National d'Investissements Agricoles et de Sécurité Alimentaire et Nutritionnelle (PNIASAN) 2017–2021. Version finale*. Cotonou, Bénin: Ministère de l'Agriculture, de l'Élevage et de la Pêche (MAEP) et NEPAD.

Scheiterle L and Birner R (2018) Assessment of Ghana's comparative advantage in maize production and the role of fertilizers. *Sustainability 10*(11), 4181.

Shi XJ, **Chen ZR**, **Wang H**, **Yeung D-Y**, **Wong W-K and Woo W-C** (2015) Convolutional LSTM network: a machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems 28*, 802–810.

Simperegui KBD, **Kouame AKK**, **Kwesie B**, **Bindraban PS**, **Adzawla W**, **Asamoah E and El Gharous M** (2025) Digital mapping of Ghana's soil properties and nutrients: Performance of spline and weighted average approaches. *Geoderma 459*, 117365.

Stankevičiūtė K, **Alaa AM and van der Schaar M** (2021) Conformal time-series forecasting. *Advances in Neural Information Processing Systems 34*, 6216–6228.

Stübinger J and Adler K (2020) How to identify varying lead–lag effects in time series. *Algorithms 13*(4), 95.

Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres 106*(D7), 7183–7192.

United Nations Department of Economic and Social Affairs Population Division (2022) World population prospects 2022.

Vodounon RBW, **Soude H and Mamadou O** (2022) Drought forecasting in Alibori Department in Benin using the standardized precipitation index and machine learning approaches. *International Journal of Advanced Computer Science and Applications (IJACSA) 13*(12). https://doi.org/10.14569/IJACSA.2022.01312113

Vovk V, **Gammerman A and Shafer G** (2005) *Algorithmic Learning in a Random World*. New York: Springer.

WMO (2017) Guidelines on the calculation of climate normals (wmo-no. 1203). Technical Report, World Meteorological Organization, Geneva.

Wright S (1921) Correlation and causation. *Journal of Agricultural Research 20*, 557–585.

Xu C and Xie Y (2023) Conformal prediction for time series. *IEEE Transactions on Pattern Analysis and Machine Intelligence 45* (10), 11575–11587. https://doi.org/10.1109/TPAMI.2023.3272339.

Yihdego Y, **Vaheddoost B and Al-Weshah RA** (2019) Drought indices and indicators revisited. *Arabian Journal of Geosciences 12*, 69.

Zerouaoui H, **Idr A and El Alaoui O** (2025) DHHoE: Deep hybrid homogenous ensemble for digital histological breast cancer classification. *Expert Systems 42*(1),e13397https://onlinelibrary.wiley.com/doi/pdf/10.1111/exsy.13397.

Cite this article: Ligan BMAS, Tchokponhoue GAD, Iliho Manzi AT, Simperegui KBD and Vodounon RBW (2025). Prediction and uncertainty quantification of drought in North Benin. *Environmental Data Science*, 4: e50. doi:10.1017/eds.2025.10025

https://doi.org/10.1017/eds.2025.10025 Published online by Cambridge University Press