DOI: 10.31577/ahs-2025-0026.02.0035

ACTA HYDROLOGICA SLOVACA

Volume 26, No. 2, 2025, 347 – 357

Comparative study between classical methods and machine learning models to estimate missing monthly rainfall, case of Coastal Oranians Center watershed

Mehdi KEBLOUTI*, Bojan ĐURIN, Mohamed Walid AZIZI, Durga SHARMA, Soraya BOUSLAH, Rayane SACI, Kenza BOUALI, Faiza BOUZAHAR, Mohamed Amine HABBACHE

The accurate estimate of missing rainfall data is essential for hydrological modeling, climate research, and water resource management. This study evaluates and compares the effectiveness of conventional interpolation techniques. Using machine learning (ML) techniques to predict missing monthly rainfall data in the coastal Oranians center watershed, Algeria. Four classical methods, such as linear regression, inverse distance weighting (IDW), spline interpolation, and local polynomial interpolation, were assessed alongside four ML models: Multilayer Perceptron (MLP), Modified Nearest Neighbor (MNN), Normalized Nearest Neighbor (NNN), and Least Squares Support Vector Machine (LSVM). Several statistical metrics, including RMSE, MAE, MRE, R^2, and Taylor diagrams, were used to assess the models and conduct comparative analysis. LSVM has the maximum performance (R^2=0.99, RMSE=9.13), according to the data, which implies that ML models exhibit superior performance compared to conventional methodologies. Even though local polynomial interpolation was less accurate than machine learning models, it was still superior to traditional techniques. These results suggest that incomplete rainfall datasets can be enhanced by AI-driven techniques, making them suitable for use in environmental and hydrological studies.

KEY WORDS: Algeria, machine learning, monthly precipitation, spatial interpolation methods, missing data

Introduction

The availability of hydrological data is critical for long-term water resource management, disaster prevention, and economic development (Bouzahar et al., 2018; Waqas et al., 2023; Katipoğlu and Keblouti, 2024; Saci et al., 2025). These data are critical for determining water levels, flows, quality, and hydrological trends. They allow for the planning of water resource use, the monitoring and mitigation of flood and drought hazards (Keblouti et al., 2012, 2022, Zhou et al. 2021), the support of potable water supply and agricultural irrigation, and the formulation of informed decisions for the conservation of aquatic ecosystems. Additionally, they are crucial for modeling hydrological characteristics and forecasting extreme hydrological events, such as floods. For this reason, the availability of hydrological data is crucial for managing water sustainably and mitigating the negative impacts of hydrological events on communities and the environment.

Rainfall is a critical climatic variable frequently used for water resource management, hydrological and ecological modeling, recharge evaluation, and irrigation planning. Significant spatiotemporal variability of precipitation is the primary source of uncertainty in hydrological modeling, meaning that precipitation data with high spatial and temporal resolution are required for urban watersheds (Ochoa-Rodriguez et al., 2015; Lu et al., 2025; Bruni et al., 2015; Woodson et al., 2019; Liu et al., 2021; Holko et al., 2025; Markovič and Faško, 2025).

Through the use of meteorological networks, hydrological data, particularly rainfall, are continually measured and stored. There are two primary reasons why these data are frequently fragmented or imperfect: the first reason is that there are not enough measurement stations, and the second is a considerable gap in these data values. This deficiency is due to several reasons, including reading, measurement, and copying errors, in addition to changes in observers, as well as malfunctions in measuring devices (Coulibaly and Evora, 2007; Louamri 2009; Wagner et al., 2012; Falck et al., 2015; Camuffo et al., 2022; Duarte et al., 2022), and finally the irregular distribution of measuring stations, especially in mountainous and remote areas, which affects our understanding of the spatial variation of rainfall.

Due to unreliable rainfall data, its use in several fields is limited. Public safety may be at stake if weather predictions become less accurate, especially during extreme weather events. The absence of complete rainfall data can lead to poor decision-making in agriculture, water resource management, and the prevention of

hydrological disasters, such as floods and droughts, with significant economic and environmental consequences. In addition, insufficient data can limit our comprehension of hydrological models and meteorological patterns in scientific investigation (Diez-Sierra and Del Jesus, 2020; Esmaeili et al., 2023; Sheikhi et al., 2023).

The demand for more accurate and dependable meteorological data makes spatial interpolation important in water resource management. To address the deficiencies in climate data, the missing climatic data can be filled in using interpolation techniques that estimate data values based on measurements from several sites (Keblouti et al., 2012; De Amorim Borges et al., 2015). Spatial interpolation enables a more comprehensive and detailed understanding of climate fluctuations within a region, which is crucial for informed water management decisions. This method is helpful in making water management systems more efficient and resistant to the effects of climate change, whether irrigation planning, flood prevention, or water resource management in agricultural and urban sectors (Konca-Kedzierska et al., 2023). These classical methods may suffer from many limitations that directly affect the accuracy of their estimates, including climatic diversity, long distances between measuring stations, and, finally, the failure to take into account complex temporal relationships, which will reduce their reliability, particularly in dynamic hydrological environments (Barbulescu, 2015; Cantet, 2015).

In recent years, considerable progress has been made in applying machine learning to estimate missing rainfall data. (Agrawal and Das, 2023; Aderyani et al., 2022). As a means of compensating for data deficiencies, these machine learning and neural network algorithms provide rainfall forecasts that are accurate and flexible (Djerbouai, 2022; Diez-Sierra and Del Jesus, 2020). These models enhance decision-making in critical areas, such as meteorology, water management, and disaster prevention, due to their ability to learn from historical and real-time data (Gu et al., 2022). Applying ML models to estimate missing rainfall data strengthens our ability to address climate and water-related challenges, significantly benefiting society and the environment. It enables better water resource management, more precise weather forecasting, and reduced risks associated with extreme weather events.

The present study compares the effectiveness of classical spatial interpolation methods and machine learning models for estimating missing rainfall data in the coastal Oranians center watershed in Northwestern Algeria. This is a significant problem in Algeria since there are often big holes in hydrological data that make it hard to analyze and manage water resources appropriately. This paper aims to determine the most effective technique for reconstructing hydrological data by comparing the two methods using statistical performance measures, including RMSE, MAE, MRE, and the coefficient of determination (R^2). The results should help semi-arid regions with better rainfall prediction and water resource management.

Material and method

Study area

In the northeast of Algeria is the Oranie Coastal Center basin, which includes the areas of Oran, Arzew, Mostaganem and Bethioua. The Mediterranean Sea surrounds it, covers 4,665 km^2, has an average altitude of 380 m and has a semi-arid climate. This region has a drainage density of 0.15 km^{-1} and contains several codified sub-basins (0402 to 0406). The main waterways are the El Maleh river and its tributaries, the Mategeur and Es Senane rivers. The estimated annual water resources are 47 million cubic meters, but these resources require complex management due to urbanization, intensive agriculture, and salt intrusion (Fig. 1).

The basin is characterized by lush plains, coastal reliefs and wetlands such as the Sebkha of Oran and the Salt Lake of Arzew. It encompasses strategic economic activities such as marine, agricultural and industrial activities while addressing important hydrological issues such as the preservation of the groundwater and flood risk management.

Description of the data sets

The majority of rainfall measurement stations in Algerian watersheds are managed by the National Water Resources Agency. Incomplete rainfall data is a known problem at several stations, and this can affect the reliability of hydrological study results (drought and

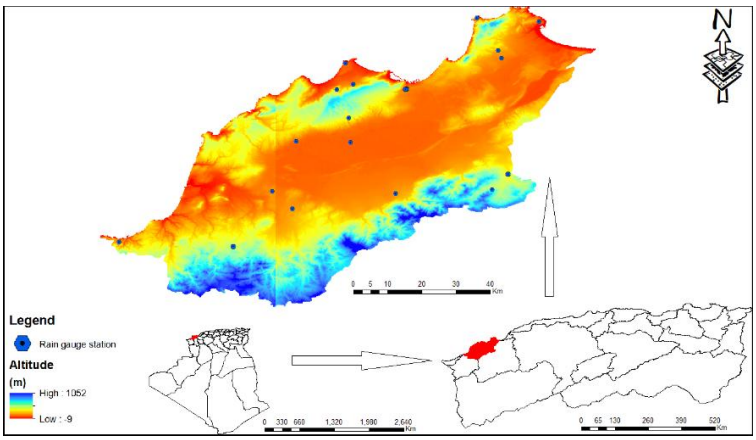


Fig. 1. Location of the study area and positions of the rain gauge stations.

flooding). For our study area, the number of missing monthly rainfall events is 120, over a period (1913–2009).

Classical methods

Inverse distance weighting

Nearly all interpolation approaches are based on the concepts of spatial autocorrelation, which suggests that items in proximity exhibit similarities to more distant ones (Setianto and Triandini, 2013). Using a weighted average approach (Shepard, 1968), the 2D interpolation technique for irregularly spaced data is further grounded in the concepts of spatial autocorrelation.

Additionally, using known data from nearby sites, it can be used to estimate unknown spatial rainfall data (Bedient and Huber, 1992; Burrough and McDonnell, 1998; Goovaerts, 2000; Li and Heap, 2008). This way of imputing rainfall data is referred to as the Inverse Distance Weighting Method (IDW). It can be determined by

$$R_u = \sum_{i=1}^{N} w_i R_i, \quad (1)$$

where, R_u is the unknown rainfall data, w_i is the weight assigned to each rainfall station, and R_i is the rainfall of a known station. w_i can be calculated as $d_i^{-\alpha} / \sum_{i=1}^{N} d_i^{-\alpha}$, d_i is the distance between rainfall station *i* and unknown rainfall station, α is the power parameter that controls how strongly the distance influences the weighting. The value of α ranges typically between 1 and 3, with larger values boosting the influence of nearby stations and lowering the impact of more distant ones. This formula guarantees that the weights are normalized ($w_i = 1$) and that the projected rainfall value is more strongly influenced by closer stations. IDW assumes that the variable of interest, in this case, rainfall, exhibits spatial autocorrelation, meaning that nearby locations are more likely to exhibit comparable values. Despite its simplicity, IDW offers a computationally effective method for interpolating geographical data, and it works particularly well when there is a dense network of observation locations (Lu and Wong, 2008).

Spline

By fitting a piecewise continuous polynomial (of degree three) across a set of known data points, cubic spline interpolation is a potent numerical method for estimating missing data values. This approach creates a smooth curve that approximates natural events, such as rainfall, realistically by maintaining continuity in the first and second derivatives, as well as in function values. Cubic spline interpolation is especially well-suited for hydrological applications since it can simulate natural variations and smooth transitions in rainfall data, as noted by Hamzah et al. (2022). Furthermore, cubic splines provide a more continuous and aesthetically cohesive representation of rainfall behavior over time or space than linear interpolation, which can result in abrupt changes at the data boundaries. The cubic spline method constructs a smooth curve passing through known data points using third-degree polynomials. The cubic spline interpolation algorithm for imputing missing rainfall data begins with an input time series (x_i, y_i), where $x_i = time$, $y_i =$ rainfall. For each missing y_j at x_j, the nearest available points (x_{j-1}, y_{j-1}) and (x_{j+1}, y_{j+1}) are identified and then the cubic polynomial is constructed as:

$$S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \quad (2)$$

We need to ensure continuity $S_i(x_{i+1}) = S_{i+1}(x_{i+1})$ and smoothness (matching first and second derivatives at knots). Apply natural spline conditions ($S''(x_1) = S''(x_n) = 0$) at end points to ensure a smooth and continuous interpolation, the polynomial must satisfy two key conditions: continuity $S_i(x_{i+1}) = S_{i+1}(x_{i+1})$ guaranteeing that adjacent splines meet at the knots, and the second one is smoothness, requiring that the first and second derivatives of adjacent splines match at each knot to avoid abrupt changes in curvature. Additionally, natural spline boundary conditions are imposed by setting the second derivatives at the endpoints to zero, ($S''(x_1) = S''(x_n) = 0$), which ensures a stable and smooth interpolation near the edges of the dataset. The coefficients a_i, b_i, c_i and d_i are determined by solving a tridiagonal system of linear equations derived from these constraints and the resulting spline $S_{(x)}$ is evaluated at x_j to estimate y_j. This process is repeated for all missing data points, yielding a complete and smoothly interpolated rainfall time series. To carry out the interpolation using a cubic spline, it is necessary to carefully check the estimated rainfall values, as this type of interpolation can generate negative values. To maintain consistency, all negative values are replaced with a value of zero.

Linear regression

Techniques such as IDW and Spline are deterministic, meaning they always yield the same result with no unpredictability. In contrast, linear regression is a probabilistic technique that makes certain assumptions about randomness. In hydrology and meteorology, linear regression is frequently used to estimate missing precipitation data. Research has shown that it is effective for spatial interpolation, particularly when there are sufficient historical records from nearby stations. In contrast to the probabilistic approach of linear regression, which integrates randomness to simulate real-world variability better, (Farmer and Vogel, 2016) highlighted the deterministic nature of techniques such as Spline and Inverse Distance Weighting (IDW). (Chen et Al., 2017) emphasized the benefits of estimating precipitation using regression approaches, pointing out that spatial dependencies between stations are effectively taken into account by linear regression.

Improvements to conventional linear regression models have been investigated in other studies. To increase prediction accuracy, for instance, (St-Hilaire et al., 2003) included climatic and geographic factors. Comparing several interpolation techniques, (Vicente-Serrano et al.,

2003) discovered that linear regression worked best in areas with extensive meteorological networks. Historical rainfall data from several weather stations is gathered in order to apply linear regression for spatial rainfall interpolation. A thorough spatial coverage is ensured by the dataset's continual rainfall recordings from nearby stations. The following is the expression for the general linear regression model that is used to estimate missing rainfall measurements at a specific station:

$$Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ \ldots\ldots + \beta_n X_n + \varepsilon \qquad (3)$$

where, Y_1 is the estimated rainfall at the station with missing data, $X_1, X_2, X_3, \ldots\ldots X_4$ represent the recorded rainfall values from neighboring stations, β_0 is the intercept term, $\beta_1, \beta_2, \beta_3, \ldots\ldots \beta_n$ are the regression coefficients that quantify the influence of each neighboring station, ε is the error term capturing unmodeled variability.

Local polynomial interpolation

Locally Weighted Polynomial Regression (LWPR) is a non-parametric approach that makes few assumptions about the underlying relationship between variables, in contrast to parametric techniques such as linear regression, which assume a fixed functional form (Cleveland and Devlin, 1988). Because of its adaptability, LWPR is especially useful for simulating intricate, non-linear connections, such as those involving missing rainfall data (the dependent variable) and the variables that influence it (independent variables, such as elevation, temperature, humidity, and measurements from surrounding stations). With the dataset containing observed rainfall values y_i at location x_i, LPWR estimates the missing value $\hat{y}_j$ at a target point x_j by fitting a weighted polynomial regression using nearby observations. The key steps are:

i) Kernal Weighting Function

A symmetrical kernel function K, such as the Gaussian or tricube kernel, assigns weights to neighboring points based on their distance from x_j:

$$w_i = K\left(\frac{\|x_i - x_j\|}{h}\right) \qquad (4)$$

where h is the bandwidth parameter, controlling the locality of the regression.

ii) Local Polynomial Regression

A low-degree polynomial, usually cubic or quadratic, is fitted to the weighted data for each missing point

$$\hat{y}_j = \beta_0 + \beta_1(x_j - x_i) + \beta_2(x_j - x_i)^2 + \cdots + \beta_p \qquad (5)$$

The coefficient β are estimated via weighted least squares (WLS):

$$\hat{\beta} = arg\min_{\beta} \sum_{i=1}^{n} w_i \,(y_i - \sum_{k=0}^{p} \beta_k (x_j - x_i)^k)^2 \qquad (6)$$

iii) Bandwidth Selection and Optimization

The selection of h is essential since a value that is too small causes overfitting, while a value that is too big causes over-smoothing. Cross-validation is frequently employed to optimize h:

$$h_{opt} = \arg\min_{h} \sum_{i=1}^{n} (y_i - \hat{y}_{i-1} h)^2 \qquad (7)$$

where $\hat{y}_{i-1}$ is the prediction at x_i when (x_i, y_i) is excluded.

Machine learning models

Linear support vector machine

The basic idea behind Linear Support Vector Machines (LSVM) is to project high-dimensional input vectors into a feature space, where an ideal linear decision boundary is built to optimize class separation (Vapnik, 1995). Although Support Vector Machines (SVMs) were initially created for classification tasks, their application to regression problems, known as Support Vector Regression, or SVR, has shown remarkable efficacy in predictive modeling (Drucker et al., 1997). The SVR formulation seeks a regression function *f(x)* that, while being robust to noise, accurately represents the relationship between input variables *x* and target outputs *y*. The challenge of optimization is stated as follows:

$$\min_{w,b,\varepsilon_i \varepsilon_i^*} \frac{1}{2}\|w\|^2 + C\ \sum_{i=1}^{n} (\xi_i + \xi_i^*) \qquad (8)$$

subject to:

$$\begin{cases} y_i - \langle w, \emptyset(x_i)\rangle - b \leq\ \varepsilon + \xi_i \\ \langle w, \emptyset(x_i)\rangle + b - y_i \leq\ \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \qquad (9)$$

- $\boldsymbol{w}$ represents the weight vector in feature space,
- $\phi(\boldsymbol{x})$ denotes a nonlinear mapping to a higher-dimensional feature space,
- defines the insensitivity zone where errors are not penalized,
- C controls the trade-off between model complexity and training error tolerance,
- ξ_i, ξ_i^* are slack variables accommodating outliers.

When appropriately used, linear SVR balances simplicity and accuracy, making it a viable option for rainfall imputation when capturing incredibly complex spatial relationships is not as crucial as computational efficiency and model interpretability (Taillardat et al., 2016). According to comparative research, linear SVR remains a dependable baseline technique, particularly in situations with limited training data, even though it may not always outperform sophisticated nonlinear techniques (Wangwongchai et al., 2023).

Modular neural network

A sophisticated machine learning framework for predicting missing rainfall records is provided by modular neural networks (MNNs), which break down the challenging rainfall prediction task into smaller, more focused subtasks. In comparison to monolithic neural architectures, this modular method improves model performance by enabling various neural network

components to concentrate on different meteorological patterns, resulting in more reliable and accurate imputations (Kajornrit et al., 2012). Where rainfall patterns exhibit notable geographical and temporal diversity, the technique is particularly effective in areas with a range of meteorological variables. To learn specific rainfall patterns, the MNN framework comprises several separate yet interconnected Artificial Neural Networks (ANNs). Three main phases comprise the system's operation:

i) ClusteringData:
Fuzzy C-Means (FCM) clustering is used to divide input data (weather variables, temporal aspects, and spatial coordinates) into homogenous clusters (Kajornrit, 2012). Each data point is given a membership grade by FCM, enabling gentle grouping that takes transitional weather patterns (such monsoon onset periods) into account. Validity indices, such as the Xie-Beni index, are used to calculate the ideal cluster count.

ii) Module Training:
During the module training phase, each cluster is assigned a customized Artificial Neural Network (ANN) model tailored to the specific properties of its data subset. Each ANN's input layer includes a combination of meteorological predictors (such as temperature, humidity, and pressure) as well as spatiotemporal variables that represent rainfall variability in both space and time. The network architecture consists of one to three hidden layers, each with the ReLU (Rectified Linear Unit) activation function, which effectively captures non-linear patterns in the data. Finally, the output layer utilizes a linear activation function to generate continuous rainfall forecasts, thereby preparing the model for regression-based forecasting tasks.

iii) Gating Network Integration:
The gating network integration is crucial in merging predictions from numerous ANN modules adapted to distinct data clusters. A meta-learner, like a SoftMax classifier, assigns weights (w_k) depending on input membership grades to determine module contribution. The final rainfall prediction $\hat{y}$ is produced by the weighted average outputs from the modules.

$$\hat{y} = \sum_{k=1}^{K} w_k f_k(x) \quad (10)$$

where w_k is the gating weight for module k and f_k is its prediction.

Neural network node

The perceptron, invented by McCulloch and Pitts in 1943, is the fundamental computational unit in Artificial Neural Networks (ANNs). A perceptron takes a set of input features, such as meteorological data (e.g., temperature, humidity), and applies trainable weights to them. These weighted inputs are combined with a bias term, and the output is then processed using an activation function, most commonly the Rectified Linear Unit (ReLU), to introduce nonlinearity. The mathematical operation performed at each node can be described as:

$$z = \sum_{i=1}^{n} w_i x_i + b \quad (11)$$

$$a = \emptyset(z) \quad (12)$$

where, x_i represents the input features, w_i are the associated weights, b is the bias, z is the linear combination of inputs, w_i are the associated weights, w_i are the associated weights, b is the bias, z is the linear combination of inputs, $\phi(z)$ is the activation function (e.g., ReLU: $\phi(z) = max(0, z)$), and a is the activated output passed to the next layer.

These networks create hierarchical representations of rainfall patterns by modifying link weights during training via backpropagation, reducing prediction errors compared to observed data (St-Hilaire et al., 2003). For rainfall imputation, the ANN undergoes supervised training on the entire historical dataset, where it learns to associate input parameters (such as surrounding station readings, elevation, and temporal indicators) with corresponding rainfall values. After training, the network can estimate missing values by processing existing predictors using its tiered architecture, with each layer converting inputs to extract increasingly abstract information important to rainfall prediction. The networks' ability to integrate different predictor types, including both point measurements and derived spatial-temporal variables, makes them especially useful for filling gaps in heterogeneous rainfall datasets. However, their performance is mainly reliant on good architectural design (usually 1–3 hidden layers for rainfall applications) and regularization to avoid overfitting to noisy meteorological data. When applied with proper input feature selection and model validation, ANNs offer a robust data-driven approach to rainfall imputation, capturing complex, non-linear relationships that simpler statistical methods often overlook.

Multilayer perceptron

A Multi-Layer Perceptron Neural Network (MLPNN) is an effective solution for missing rainfall imputation, utilizing the Universal Approximation Theory (Hornik et al., 1989) and the Backpropagation Learning Algorithm (Rumelhart et al., 1986). Any continuous function can be accurately approximated by a feedforward neural network with a nonlinear activation function and at least one hidden layer, according to Universal Approximation Theory. Due to this feature, MLPNN is well-suited for modeling the intricate, nonlinear relationships between rainfall data at different locations. By adjusting synaptic weights based on the discrepancy between expected and actual rainfall amounts, the Backpropagation method, a supervised learning technique, enhances the network and increases the model's ability to retrieve missing data. Based on fundamental neural network theory, the MLPNN method for rainfall imputation has the following formal definition:

i) Forward Propagation:
Given an input vector x (rainfall observations from neighboring stations), the output is computed as:

$$\hat{y} = f_0(W^{(2)}.f_h\left(W^{(1)} + b^{(1)}\right) + b^{(2)}) \quad (13)$$

where, $W^{(1)}$ and $b^{(1)}$ are the weights and biases of the hidden layer, $W^{(2)}$ and $b^{(2)}$ are the weights and biases of the hidden output layer, f_h is the hidden layer activation function, f_0 is the output activation function.

ii) Backward Propagation:
The weights are updated via gradient descent to minimize the mean squared error (MSE) loss:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (14)$$

The weight updates are computed as:

$$\Delta W^{-(l)} = -\eta \frac{\partial \mathcal{L}}{\partial W^l} \quad (15)$$

where, η is the learning rate, $\frac{\partial \mathcal{L}}{\partial W^l}$ is the gradient obtained via the chain rule.

iii) Universal Approximation Theorem: The MLPNN, with at least one hidden layer and a nonlinear activation function, can approximate any continuous function $f(x)$ with arbitrary precision:

Comparison of methods
For our case and to study the performance of the methods used to estimate the monthly rain, four performance metrics were used:

R-squared (R^2)
R-squared is a popular diagnostic measure for evaluating the effectiveness of various missing data imputation algorithms, allowing comparisons amongst methods to discover which one best predicts missing precipitation values. A high R-squared value indicates that the chosen method effectively explains the relationship between the observed precipitation data and the imputed values, implying that the imputation model is performing well in predicting missing precipitation data based on available information. A higher R-squared value indicates greater accuracy in imputation.

$$R^2 = 1 - \sum \frac{(R_o - R_p)^2}{(R_o - R_{av})^2}, \quad (16)$$

where R_o is the observed rainfall data, R_p is the predicted rainfall and R_{av} is the average value of the observed rainfall.

Root Mean Squared Error (RMSE)
The discrepancy between expected and actual values is indicated by the root mean square error (RMSE), which calculates the standard deviation (SD) of the prediction error. The dependent variable's units are the same, and its range is 0 to positive infinity. It can be supplied by:

$$RMSE = \sqrt{\sum \frac{(R_o - R_p)^2}{N}} \quad (17)$$

Mean Absolute Error (MAE)
The mean absolute error (MAE) characterizes the alteration between the original and predicted values and is defined as the dataset's total alteration mean. It is given by:

$$MAE = \sum \frac{(R_p - R_o)}{N} \quad (18)$$

Here *N* is the total number of observations.

Mean Relative Error (MRE)
A measure of how far a model's predictions deviate from the actual data is called mean relative error, or MRE. Depending on whether the forecasts are consistently higher or lower than the actual figures, the outcome can be either positive or negative.

$$MRE = \frac{1}{N}\sum \frac{|R_o - R_p|}{R_o} \quad (19)$$

Results and discussions

Among the classical methods, the local polynomial method performed best with the lowest RMSE (10.89), MAE (7.79), and MRE (27.51), along with a high coefficient of determination (R^2=0.98). Conversely, the spline approach displayed the worst performance, displaying the most significant error values (RMSE=46.62, MAE=32.24, MRE=178.49) and a very low R^2=0.04, which indicates poor predictive accuracy (Table 1).
For the ML models, all four models (MNN, NNN, MLP, LSVM) outperformed the classical methods in terms of lower error metrics and higher correlation values. The LSVM (Least Squares Support Vector Machine) achieved the best results overall, with the lowest RMSE (9.13), MAE (5.37), and MRE (19.82), and the highest R^2 (0.99), indicating a strong capability in reconstructing the missing rainfall data.
The Taylor diagrams visually confirm these findings:
The ML models cluster closely around the reference point, exhibiting high correlation coefficients and lower standard deviations, particularly for LSVM and NNN.
Classical methods, while showing acceptable performance in some cases (notably the polynomial method), remain less accurate, particularly the spline method, which stands out for its high deviation and weak correlation (Fig. 2).
The LSVM model provides estimates that closely follow the observed monthly rainfall values. The predicted curve aligns very well with the observed data, capturing both seasonal trends and monthly fluctuations, indicating high model precision and generalization capability (Fig. 3).
The local polynomial interpolation method shows a reasonable fit to the observed rainfall data. While it generally follows the seasonal pattern, some discrepancies appear in capturing peak and low values, suggesting moderate estimation accuracy compared to models (Fig. 4).
The comparison between classical interpolation methods and machine learning techniques models shows a clear advantage of ML-based approaches in estimating missing monthly rainfall data for the coastal Oranians center watershed.

Table 1. Performance criteria of the used models

	RMSE	MAE	MRE	R^2
		Classical methods		
Linéaire Régression	17.21	11.41	51.54	0.81
IDW	18.13	10.51	43.13	0.69
Spline	46.62	32.24	178.49	0.04
Local polynomial	10.89	7.79	27.51	0.98
		ML models		
MNN	9.78	7.40	29.43	0.95
NNN	9.99	6.55	27.13	0.91
MLP	12.68	9.06	31.52	0.98
LSVM	9.13	5.37	19.82	0.99

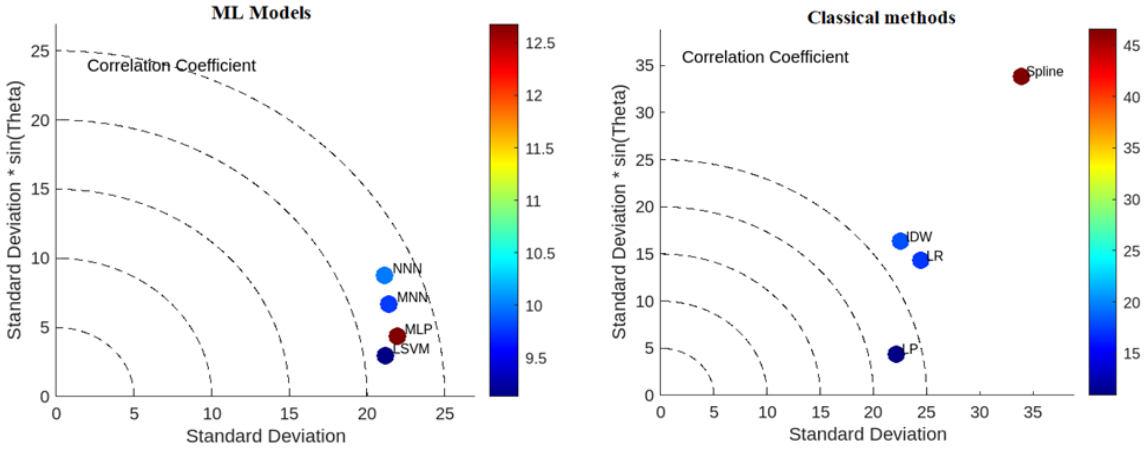
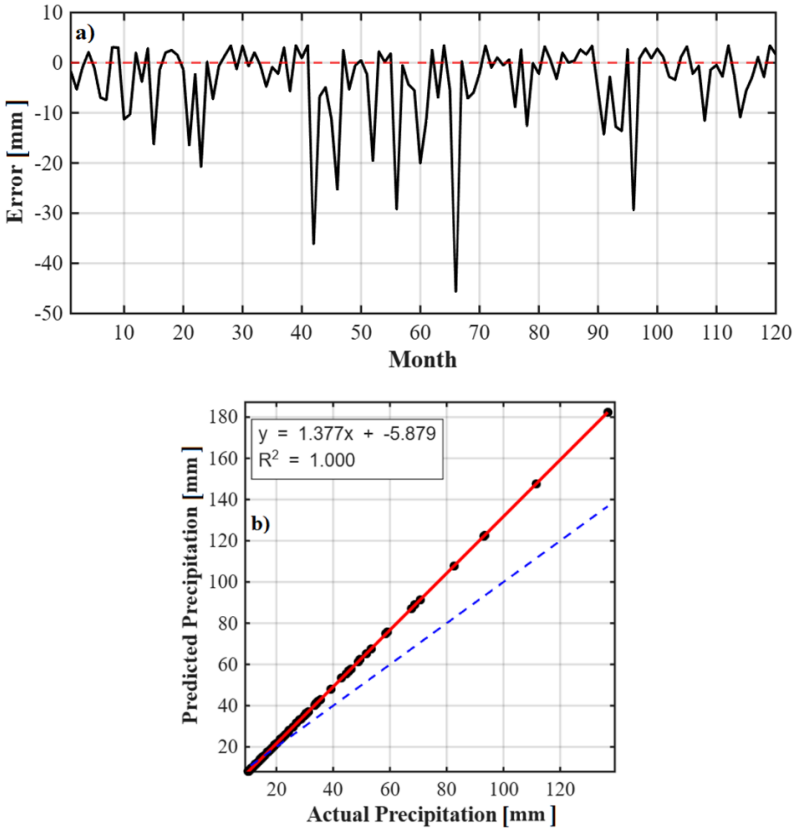


Fig. 2. Taylor diagram of test results ML models, Classical methods.



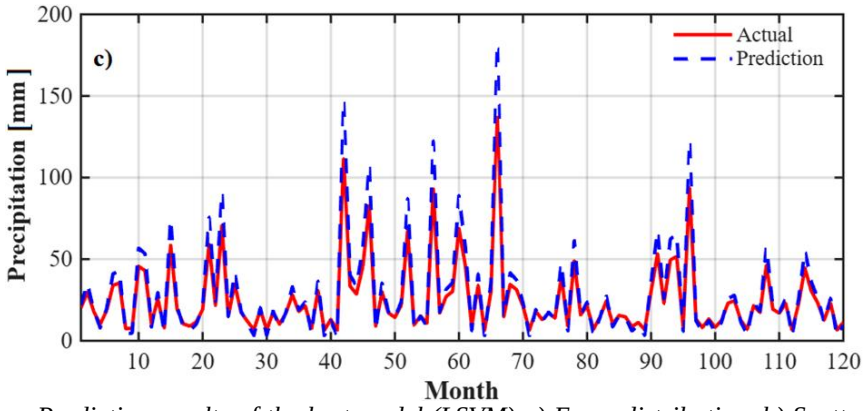


Fig. 3. Prediction results of the best model (LSVM) a) Error distribution, b) Scatter diagram, c) Time series.

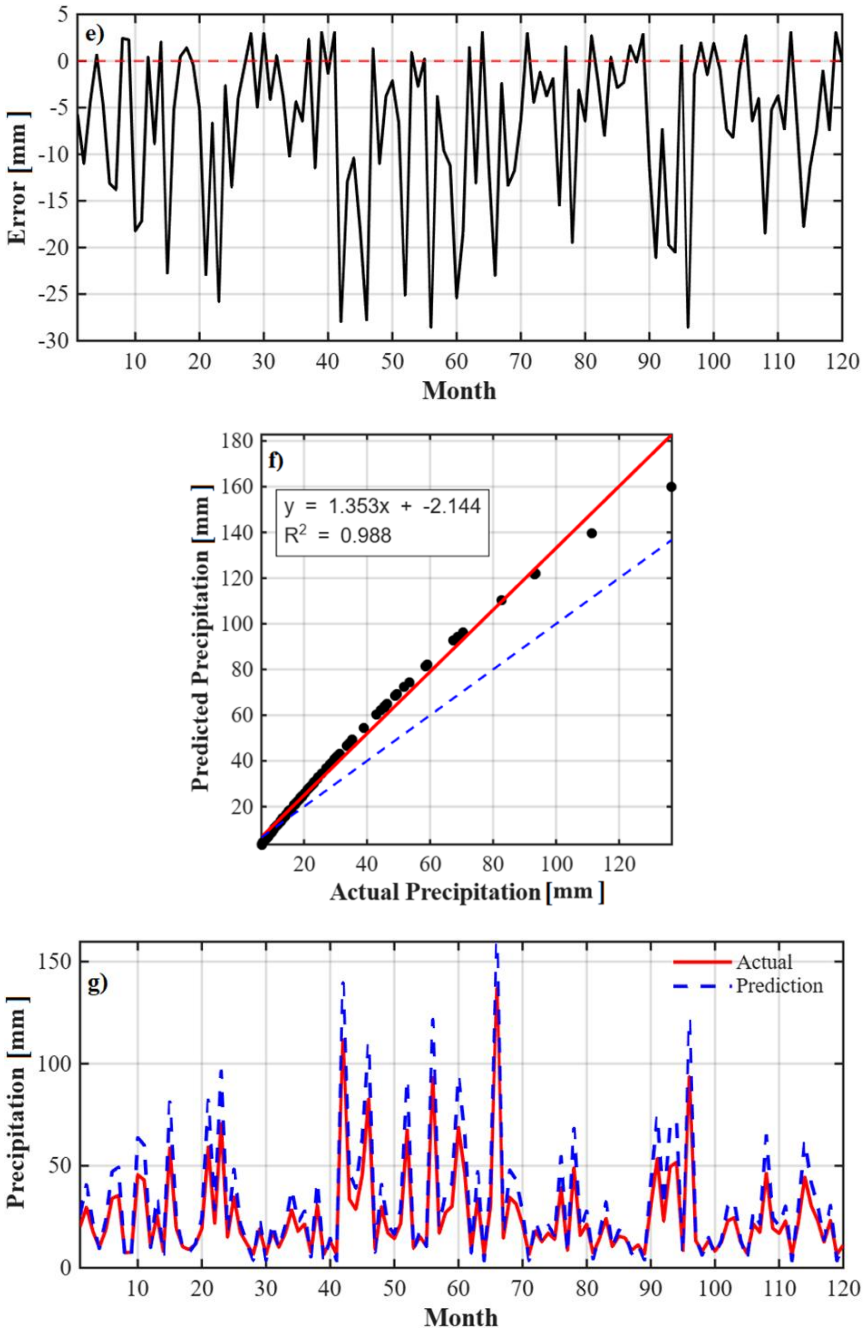


Fig. 4. Prediction results of the best model (Local polynomial interpolation) e) Error distribution, f) Scatter diagram, g) Time series.

Conclusion

This study demonstrates that when it comes to predicting missing monthly rainfall data in the coastal Oranians center watershed, machine learning models perform better than conventional interpolation techniques. Although local polynomial interpolation, inverse distance weighting (IDW), and linear regression are often used due to their simplicity and interpretability, these statistical methods are incapable of accurately reflecting the intricate and nonlinear variability of rainfall patterns. The results indicate that the LSVM model achieved the best overall performance (R^2=0.99, RMSE=9.13), outperforming most conventional methods. While local polynomial interpolation was strong (R^2=0.98), other classical techniques, such as IDW (R^2=0.69) and spline (R^2=0.04), were significantly less accurate. Machine learning models, such as LSVM, MLP, MNN, and NNN, exhibit higher accuracy, lower error metrics (RMSE, MAE, MRE), and better correlations. These results show that ML models can generalize rainfall behavior from complicated watershed data.

Hydrological modeling, drought analysis, and reservoir management need accurate rainfall data. ML improves data processing, enhancing water resource management, climate adaptability, and early warning systems.

This study proposes a way to increase ML usage in hydrometeorology to alleviate data shortages. Adding meteorological variables, deep learning, and hybrid models might improve spatiotemporal rainfall reconstruction.

Ultimately, integrating classical knowledge with current ML approaches may lead to more robust and data-driven environmental decision-making.

Acknowledgement

The Authors would like to express gratitude for the support of the University North during this research, within the scientific project 'Hydrological and geodetic analysis of the watercourse-second part', 2025.

References

Adaryani, F. R., Mousavi, S. J., Jafari, F. (2022): Short term rainfall forecasting using machine learning-based approaches of PSO-SVR, LSTM and CNN. Journal of Hydrology, 614, 128463. ISSN 0022-1694. DOI: https://doi.org/10.1016/j.jhydrol.2022.128463.

Agrawal, J., Das, S. (2023): ANN in forecasting missing rainfall data. E3S Web of Conferences, 405, 01010.

Barbulescu, A. (2015): A new method for estimation the regional precipitation. Water Resources Management, 30(1), 33–42. https://doi.org/10.1007/s11269-015-1152-2.

Bedient, P. B., Huber, W. C. (1992): Hydrology and floodplain analysis. Addison-Wesley.

Bouzahar, F. H., Ouerdachi, L., Keblouti, M., Seddiki, A. (2018): The contribution of remote sensing in hydraulics and hydrology, analysis and evaluation of digital terrain model for flood risk mapping. Journal of Water and Land Development, 39(1), 17–26. https://doi.org/10.2478/jwld-2018-0055.

Bruni, G., Reinoso, R., Van De Giesen, N. C., Clemens, F. H. L. R., Veldhuis, J. A. E. T. (2015): On the sensitivity of urban hydrodynamic modelling to rainfall spatial and temporal resolution. Hydrology and Earth System Sciences, 19. https://doi.org/10.5194/hess-19-691-2015.

Burrough, P. A., McDonnell, R. A. (1998): Spatial analysis using continuous fields. Principles of geographical information systems, 190–195. http://infoscience. epfl.ch/record/51429.

Camuffo, D., Becherini, F., della Valle, A., Zanini, V. (2022): A comparison between different methods to fill gaps in early precipitation series. Environmental Earth Sciences, 81, 1–14. https://doi.org/10.1007/s12665-022-10467-w.

Cantet, P. (2015): Mapping the mean monthly precipitation of a small island using kriging with external drifts. Theoretical and Applied Climatology, 127(1–2), 31–44. https://doi.org/10.1007/s00704-015-1610-z.

Chen, T., Ren, L., Yuan, F., Yang, X., Jiang, S., Tang, T., Zhang, L. (2017): Comparison of spatial interpolation schemes for rainfall data and application in hydrological modeling. Water, 9. https://doi.org/10.3390/w9050342.

Cleveland, W. S., Devlin, S. J. (1988): Locally weighted regression: an approach to regression analysis by local fitting. Journal of the American Statistical Association, 83. https://doi.org/10.1080/01621459.1988.10478639.

Coulibaly, P., Evora, N. D. (2007): Comparison of neural network methods for infilling missing daily weather records. Journal of Hydrology, 341, 27–41. https://doi.org/10.1016/j.jhydrol.2007.04.020.

De Amorim Borges, P., Franke, J., Da Anunciação, Y. M. T., Weiss, H., Bernhofer, C. (2015): Comparison of spatial interpolation methods for the estimation of precipitation distribution in Distrito Federal, Brazil. Theoretical and Applied Climatology. 123(1–2), 335–348. https://doi.org/10.1007/s00704-014-1359-9.

Diez-Sierra, J., Del Jesus, M. (2020): Long-term rainfall prediction using atmospheric synoptic patterns in semi-arid climates with statistical and machine learning methods. Journal of Hydrology, 586, 124789. https://doi.org/10.1016/j.jhydrol.2020.124789.

Djerbouai, S. (2022): Missing precipitation data estimation using Long Short-Term Memory deep neural networks. Journal of Ecological Engineering, 23(5), 216–225. https://doi.org/10.12911/22998993/147322.

Duarte, L. V., Formiga, K. T. M., Costa, V. A. F. (2022): Comparison of methods for filling daily and monthly rainfall missing data: Statistical models or imputation of satellite retrievals? Water, 14(19), 3144. DOI: https://doi.org/10.3390/w14193144.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J. Vapnik, V. (1996): Support vector regression machines. Fraunhofer-Publica (Fraunhofer-Gesellschaft), 9, 155–161.

Esmaeili, F., Shabanlou, S., Saadat, M. (2023): Novel reliable model by integrating the adaptive neuro-fuzzy inference systems with wavelet transform and firefly algorithms for rainfall forecasting in the north of Iran. Applied Water Science, 13, 1–16. https://doi.org/10.1007/s13201-022-01859-1.

Falck, A. S., Maggioni, V., Tomasella, J., Vila, D. A., Diniz, F. L. (2015): Propagation of satellite precipitation uncertainties through a distributed hydrologic model: A case study in the Tocantins-Araguaia basin in Brazil. Journal of Hydrology, 527, 943–957. https://doi.org/10.1016/j.jhydrol.2015.05.042.

Farmer, W. H., Vogel, R. M. (2016): On the deterministic and stochastic use of hydrologic models. Water Resources

Research, 52(7), 5619–5633. https://doi.org/10.1002/2016wr019129.

Goovaerts, P. (2000): Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. Journal of Hydrology, 228(1–2), 113–129. https://doi.org/10.1016/s0022-1694(00)00144-x.

Gu, J., Liu, S., Zhou, Z., Chalov, S. R., Zhuang, Q. (2022): A stacking ensemble learning model for monthly rainfall prediction in the Taihu Basin, China. Water, 14(3), 492. https://doi.org/10.3390/w14030492.

Hamzah, F. B., Mohamad Hamzah, F., Mohd Razali, S. F. El-Shafie, A. (2022): Multiple imputations by chained equations for recovering missing daily streamflow observations: a case study of Langat River basin in Malaysia. Hydrological Sciences Journal, 67(1), 137–149. https://doi.org/10.1080/02626667.2021.2001471.

Holko, L., Danko, M., Jančo, M., Sleziak, P., Varšová, S. (2025): Precipitation characteristics estimated from the data given by recording raingauges and automatic weather stations. Acta Hydrologica Slovaca, 26(1), 27–39. DOI: https://doi.org/10.31577/ahs-2025-0026.01.0003.

Hornik, K., Stinchcombe, M., White, H. (1989): Multilayer feedforward networks are universal approximators. Neural Networks, 2(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8.

Kajornrit, J., Wong, K. W., Fung, C. C. (2012): Estimation of missing precipitation records using modular artificial neural networks. Neural Information, Processing: 19th International Conference, ICONIP 2012, Doha, Qatar, November 12–15, 2012, Proceedings, Part IV, 19, 52–59.

Kajornrit, P. (2012): Clustering meteorological data using fuzzy C-means and its application in neural networks. Applied Soft Computing, 12(1), 204–209.

Katipoğlu, O. M., Keblouti, M. (2024): Comparative analysis of data-driven models and signal processing techniques in the monthly maximum daily precipitation prediction of El Kerma station, northeast of Algeria. Soft Computing, 28, 10751–10765. https://doi.org/10.1007/s00500-024-09860-3.

Keblouti, M., Boutaghane, H., Bouzahar, F., Habbeche, M. A. (2022): Detection and correction of suspicious pixels of rainy images for Sétif weather radar, northeast of Algeria. International Journal of Hydrology Science and Technology, 13, 335–338. https://doi.org/10.1504/ijhst.2022.121965.

Keblouti, M., Ouerdachi, L., Boutaghane, H. (2012): Spatial interpolation of annual precipitation in Annaba, Algeria: Comparison and evaluation of methods. Energy Procedia, 18, 468–475. DOI: https://doi.org/10.1016/j.egypro.2012.05.058.

Konca-Kedzierska, K., Wibig, J., Gruszczyńska, M. (2023): Comparison and combination of interpolation methods for daily precipitation in Poland: evaluation using the correlation coefficient and correspondence ratio. Meteorology Hydrology and Water Management, 11(2), 1–27. https://doi.org/10.26491/mhwm/171699.

Li, J., Heap, A. D. (2008): A review of spatial interpolation methods for environmental scientists. Geoscience Australia Record.

Liu, Y., Huang, Y., Liu, Y., Li, K., Li, M. (2021): The impact of rainfall movement direction on urban runoff cannot be ignored in urban hydrologic management. Water, 13(20), 2923. https://doi.org/10.3390/w13202923.

Louamri, A. (2009): Cartographie des pluies annuelles dans le bassin versant de l'oued Seybouse (Nord-Est Algérien). Sciences & Technologie D, Sciences de la Terre, 0, 43–52.

Lu, G. Y., Wong, D. W. (2008): An adaptive inverse-distance weighting spatial interpolation technique. Computers & Geosciences, 34(9), 1044–1055. https://doi.org/10.1016/j.cageo.2007.07.010.

Lu, X., Wen, T., Zhang, L., Zhang, Q. (2025): Impact of spatiotemporal rainfall distribution and underlying surface changes on flood processes in Meijiang River Basin, China. Water, 17(4), 466. https://doi.org/10.3390/w17040466.

Markovič, L., Faško, P. (2025): Extreme precipitation events in Slovakia: A 50-year retrospective of the October 1974 hydro-meteorological event. Acta Hydrologica Slovaca, 26(1), 3–15. DOI: https://doi.org/10.31577/ahs-2025-0026.01.0001.

Ochoa-Rodriguez, S., Wang, L., Gires, A., Pina, R. D., Reinoso-Rondinel, R., Bruni, G., Ichiba, A., Gaitan, S., Cristiano, E., Van Assel, J., Kroll, S., Murlà-Tuyls, D., Tisserand, B., Schertzer, D., Tchiguirinskaia, I., Onof, C., Willems, P., Veldhuis, M. T. (2015): Impact of spatial and temporal resolution of rainfall inputs on urban hydrodynamic modelling outputs: A multi-catchment investigation. Journal of Hydrology, 531, 389–407. https://doi.org/10.1016/j.jhydrol.2015.05.035.

Rumelhart, D., Hinton, G. & Williams, R. (1986): Learning representations by back-propagating errors. Nature, 323, 533–536. https://doi.org/10.1038/323533a0.

Saci, R., Keblouti, M., Katipoğlu, O. M., Đurin, B., Majour, H., Sayad, L., Bouzahar, F., Benchaiba, L. (2025): Assessing the efficacy of various predictive models in simulating monthly reference evapotranspiration patterns and its impact on water resource management for agriculture in the Kebir-West watershed, north-east of Algeria. Journal of Hydrology and Hydromechanics, 73(3), 284–294. https://doi.org/10.2478/johh-2025-0022.

Setianto, A., Triandini, T. (2013): Comparison of kriging and inverse distance weighted (IDW) interpolation methods in lineament extraction and analysis. Journal of Applied Geology and Geophysics, 1(1), 1–7. https://doi.org/10.22146/jag.7204.

Sheikhi, Y., Ashrafi, S. M., Nikoo, M. R., Haghighi, A. (2023): Enhancing daily rainfall prediction in urban areas: A comparative study of hybrid artificial intelligence models with optimization algorithms. Applied Water Science, 13, 1–19.

Shepard, D. (1968): A two-dimensional interpolation function for irregularly-spaced data. Proceedings of the 1968 23rd ACM National Conference, 517–524.

St-Hilaire, A., Ouarda, T. B. M. J., Lachance, M., Bobée, B., Gaudet, J., Gignac, C. (2003): Assessment of the impact of meteorological network density on the interpolation of daily precipitation. Journal of Hydrology, 281(1–2), 84–96.

Taillardat, M., Mestre, O., Zamo, M., Naveau, P. (2016): Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. Monthly Weather Review, 144(6), 2375–2393. https://doi.org/10.1175/mwr-d-15-0260.1.

Vapnik, V. (1995): The nature of statistical learning theory. Springer eBooks. https://doi.org/10.1007/978-1-4757-2440-0.

Vicente-Serrano, S. M., Saz-Sánchez, M. A., Cuadrat, J. M. (2003): Comparative analysis of interpolation methods in the middle Ebro Valley (Spain): application to annual precipitation and temperature. Climate Research, 24(2), 161–180. https://doi.org/10.3354/cr024161.

Wagner, P. D., Fiener, P., Wilken, F., Kumar, S., Schneider, K. (2012): Comparison and evaluation of spatial interpolation schemes for daily rainfall in data-scarce regions. Journal of Hydrology, 464–465, 388–400. https://doi.org/10.1016/j.jhydrol.2012.07.026.

Wangwongchai, A., Waqas, M., Dechpichai, P., Hlaing, P. T.,

Ahmad, S., Humphries, U. W. (2023): Imputation of missing daily rainfall data; a comparison between artificial intelligence and statistical techniques. MethodsX, 11, 102459. https://doi.org/10.1016/j.mex.2023.102459.

Waqas, M., Humphries, U. W., Wangwongchai, A., Dechpichai, P., Ahmad, S. (2023): Potential of Artificial Intelligence-Based Techniques for Rainfall Forecasting in Thailand: A Comprehensive Review. Water, 15(16), 2979. https://doi.org/10.3390/w15162979.

Woodson, D., Adams, T. E., Dymond, R. (2019): Precipitation estimation methods in continuous, distributed urban hydrologic modeling. Water, 11(7), 1340. https://doi.org/10.3390/w11071340.

Zhou, Z., Smith, J. A., Baeck, M. L., Wright, D. B., Smith, B. K., Liu, S. (2021): The impact of the spatiotemporal structure of rainfall on flood frequency over a small urban watershed: an approach coupling stochastic storm transposition and hydrologic modeling. Hydrology and Earth System Sciences, 25(9), 4701–4717. https://doi.org/10.5194/hess-25-4701-2021.

Mehdi Keblouti, Class A Assistant Professor (*corresponding author, e-mail: mehdi.keblouti@univ-annaba.dz)
Mohamed Amine Habbache, PhD student
Department of Hydraulic, Faculty of Technology
Badji Mokhtar-Annaba University, P.O. Box 12, 23000, Annaba
Algeria

Soraya Bouslah, Class A Assistant Professor
Laboratory of Soils and Hydraulic, Department of Hydraulic, Faculty of Technology
Badji Mokhtar-Annaba University, P.O. Box 12, 23000, Annaba
Algeria

Rayan Saci, PhD student
Geology Research Laboratory (GRL), Department of Geology, Faculty of Earth Sciences
Badji Mokhtar-Annaba University
P.O. Box 12, 23000, Annaba
Algeria

Mohamed Walid Azizi, Class A Assistant Professor
Kenza Bouali, PhD student
Environmental Engineering and Technology Laboratory
Abdelhafid Boussouf University Center
P.O. Box 26, 43000, Mila
Algeria

Bojan Đurin, Associate Professor
Department of Civil Engineering, University North
Jurja Križanića 31b
42000 Varaždin
Croatia

Durga Sharma, Academic researcher
Impact Hub Hyderabad, University of Hyderabad
1st Floor, C9MQ+G2V A&A Lakefront, opposite to Durgam Cheruvu Lake
Software Units Layout, Madhapur, Hyderabad
Telangana 500081
India

Faiza Bouzahar, Class B Assistant Professor
Kasdi Merbah University
Ouargla Ghardaia Road, BP. 511, 30,000
Algeria