Acta Sci. Pol.
Formatio Circumiectus 25 (1) 2026, 3–16
ENVIRONMENTAL PROCESSES

www.acta.urk.edu.pl ISSN 1644-0765 DOI: http://dx.doi.org/10.15576/ASP.FC/210976

ORIGINAL PAPER Accepted: 16.09.2025

MACHINE LEARNING-BASED EVAPORATION PREDICTION AND KERNEL FUNCTION ANALYSIS: A CASE STUDY OF THE BOUKOURDANE DAM, ALGERIA

Marouane Boudjerda[1] 0000-0003-2780-3842, Issam Rehamnia[2], Parveen Sihag[3] 0000-0002-7761-0603, Andrea Petroselli[4]✉ 0000-0003-4943-0928

[1] University of Jijel, Algeria
[2] Badji-Mokhtar Annaba University, Algeria
[3] Chandigarh University, Punjab, India
[4] University of Tusca, Viterbo, Italy

ABSTRACT

Aim of the study

The objective of this study is to evaluate the impact of three kernel functions − Pearson VII, radial basis function (RBF), and polynomial − on the predictive performance of Support Vector Regression (SVR) and Gaussian Process Regression (GPR) models.

Materials and methods

Three machine learning models − Random Forest (RF), Support Vector Regression (SVR), and Gaussian Process Regression (GPR) − were applied to estimate monthly evaporation at Boukourdane Dam, Algeria. The dataset included 240 observations over 20 years, with the following inputs: max./min. air temperature, relative humidity, wind speed, and water temperature; the output being: evaporation.

Results and conclusions

Model performance was evaluated via Correlation Coefficient (*CC*), Root Mean Square Error (*RMSE*), and Mean Absolute Error (*MAE*). RF outperformed GPR and SVR across kernels, achieving MAE = 1.01 mm, RMSE = 1.29 mm, and *CC* = 0.81 in testing. Moreover, the Pearson VII kernel delivered the highest accuracy within both the GP and SVM frameworks. Sensitivity analysis highlighted relative humidity as the most influential factor in evaporation forecasting.

Keywords: evaporation forecasting, gaussian process, random forest regression, support vector machines, sensitivity analysis

INTRODUCTION

Water is a vital yet finite resource, crucial for ecosystems, human welfare, and industry (Giordano et al., 2010; Pahl-Wostl, 2017). Mismanagement of water resources threatens shared sustainability (Garrick et al., 2020), while rising water scarcity – especially in arid regions – disrupts agriculture and food security. By 2025, over 30 nations may face severe shortages (Jasmine et al., 2022). Improving water use efficiency, particularly via agricultural technologies, is essential (Benzagtha, 2014).

✉e-mail: petro@unitus.it

Evaporation, a key process in the hydrological cycle, affects diverse sectors (Kamienski et al., 2019), driven by solar energy and influenced by climatic factors (Dimitriadis et al., 2021). Accurate estimation is vital for hydrology, irrigation, flood control, and ecosystem modeling (Rajput et al., 2024; Zhao et al., 2024).

Direct methods (e.g., pans) offer precision but are labor-intensive and climate-sensitive (Kumar and Arakeri, 2021; Melišová et al., 2021). Indirect approaches use equations like Penman-Monteith but face limitations in modeling evaporation's non-linear behavior (Seifi and Soroush, 2020).

Machine learning (ML) presents a promising alternative, enhancing prediction through real-time weather integration (LeCun et al., 2015; Cappelli et al., 2023; Grimaldi et al., 2024). ML has shown high accuracy in evaporation forecasting (Adnan et al., 2022; Boudjerda et al., 2024). For instance, Shabani et al. (2020) applied four techniques for potential evapotranspiration (PE) estimation across three stations in Iran, reporting that Gaussian Process Regression (GPR) achieved higher predictive accuracy compared to K-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Regression (SVR). Similarly, Sattari et al. (2021) investigated the estimation of reference evapotranspiration in the Çorum region of Turkey and found that the Broyden–Fletcher–Goldfarb–Shanno Artificial Neural Network (BFGS-ANN) outperformed GPR, SVR, and Long Short-Term Memory (LSTM) models in terms of prediction accuracy. While prior work has highlighted key features like temperature and humidity (Tao et al., 2018; Wu et al., 2020), limited attention has been given to model structure and kernel selection in ML algorithms.

The primary objective of this study is to systematically evaluate the influence of kernel functions on the predictive performance of Support Vector Regression (SVR) and Gaussian Process Regression (GPR) models for estimating daily evaporation rates at the Boukourdane Dam, Algeria. By comparing different kernels − Pearson VII, radial basis function (RBF), and polynomial − this study investigates how kernel selection affects model accuracy and generalization in a water-scarce region. Although machine learning methods such as SVM (Sharafi et al., 2023), GP (De Caro et al., 2023), and ANN (Singh et al., 2019) are widely applied in hydrological modeling, few studies have specifically addressed kernel-dependent performance in evaporation forecasting. By filling this gap, the present work provides novel insights for improving ML-based predictions under conditions of climatic variability and limited observational data, thereby contributing to more informed water resource management.

MATERIALS AND METHODS

The study area and the available data

This case study examines Boukourdane Dam in northern Algeria (36°31′N, 2°18′E, 119.5 m elevation), with a reservoir capacity of 105 million m³ and a regulated volume of 50 million m³ (Fig. 1). The Boukourdane Dam serves two primary functions: (i) providing domestic water supply to surrounding communities, and (ii) supporting agricultural irrigation in the region. This site offers a valuable context for analyzing water resource management and evaporation estimation.

Monthly meteorological data from the National Agency of Dams (ANBT), spanning September 1996 to August 2016, include observed evaporation [mm/month], maximum and minimum air temperatures [°C], relative humidity [%], wind speed [km/h], and maximum and minimum water temperatures [°C]. The dataset, comprising 240 monthly observations over the 20-year period, was obtained directly from a meteorological station installed at the Boukourdane Dam site. These variables serve as inputs for modeling, with evaporation [mm/day] as the primary output. Through regression analysis between evaporation and the relevant climatic parameters, the optimal predictive formula was derived as follows:

$$Ev = 0.076 \cdot T_{\max} + 0.05 \cdot T_{\min} + 0.0234 \cdot U - 0.007 \cdot H + \\ + 0.102 \cdot Tw_{\max} + 0.096 \cdot Tw_{\min} - 2.16 \qquad (1)$$

with $R^2 = 0.71$, and NSE = 0.57

Figure 2 illustrates the observed monthly variation in evaporation levels at the Boukourdane Dam. This extensive dataset serves as the basis for our analysis, allowing us to explore the complex relationships between meteorological factors and evaporation rates in this particular geographical setting.

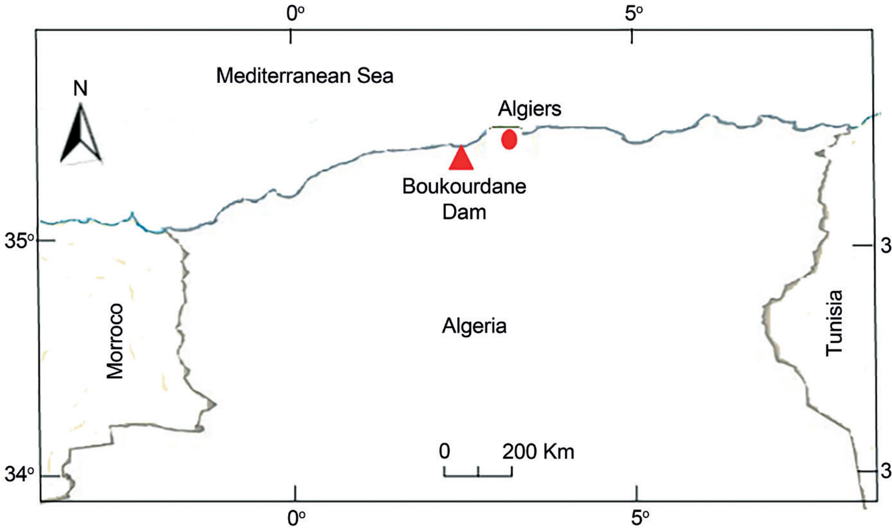

Fig. 1. Location of case study area (Source: own elaboration)

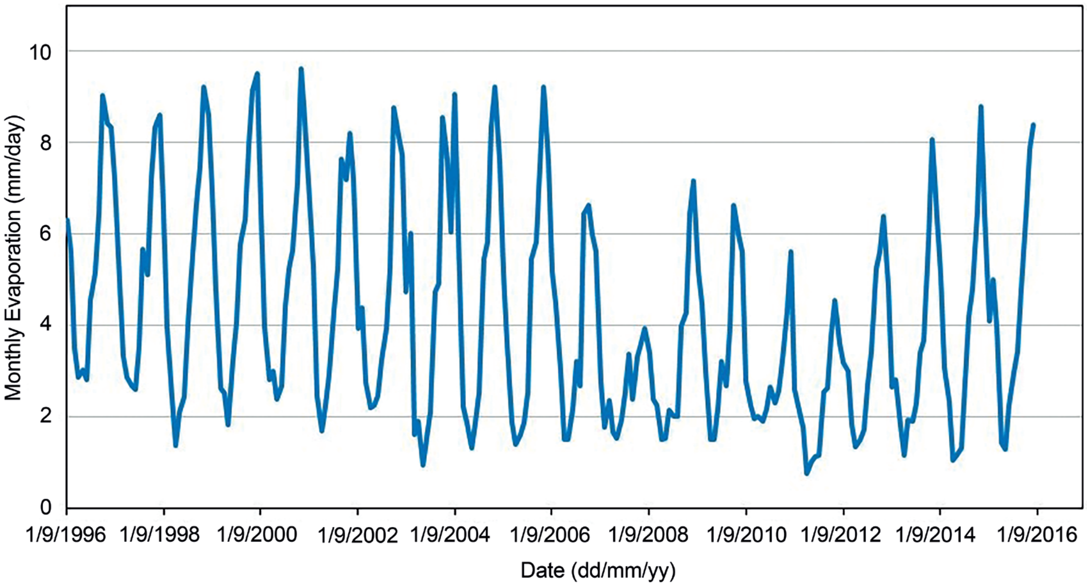

Fig. 2. Time series of the observed evaporation (Source: own elaboration)

To meet our research goals, meteorological data was split into training (Sept. 1996–Aug. 2012) and testing (Sept. 2012–Aug. 2016) sets, as shown in Table 1. This temporal division allows model training on historical data and evaluation on unseen data, ensuring robust, generalizable predictions of evaporation rates.

Random Forest regression (RF)

Random Forest (RF), introduced by Breiman (2001), is a robust ensemble learning method based on classification and regression trees (CART). It employs bootstrap sampling and bagging to build multiple decision trees using random subsets of the training data, with the remaining out-of-bag (OOB) samples used to esti-

Table 1. Statistical properties of investigated variables (Source: own elaboration)

Parameters	Units	Training (Sept. 1996–Aug. 2012)				Testing (Sept. 2012–Aug. 2016)			
		Min.	Max.	Mean	Standard deviation	Min.	Max.	Mean	Standard deviation
$T_{max.}$	°C	10	32	21.2	5.9	11	32	22.0	6.1
$T_{min.}$	°C	4.5	25	13.9	5.5	4	25	14.3	5.8
U	km/h	0	9	4.5	2.4	1.05	4.94	2.5	0.7
H	%	32.5	76.5	50.6	6.5	25	66.5	48.1	8.2
$Tw_{max.}$	°C	10	34	21.0	6.0	8	31	18.7	6.0
$Tw_{min.}$	°C	5	26.4	14.7	5.9	2	25	12.5	5.6
Evaporation	mm/day	1.0	9.6	4.5	2.4	0.8	8.8	3.4	2.0

mate prediction error. In classification, final decisions are based on majority voting, guided by the Gini index to select optimal splits.

Formulation of the Gini index in classification is given as:

$$Gini\ (node) = 1 - \sum\nolimits_{i=1}^{C} (pi)^2 \qquad (2)$$

where C is the number of classes and pi is the proportion of samples belonging to class i.

Depending on the dataset, Random Forest trees construction may be led to a large forest. To avoid this inconvenience, two parameters should be adjusted: the number of predictors randomly selected at each node (mtry), and the number of ensemble trees (ntree).

Gaussian Process (GP)

The Gaussian Process (GP) model (Rasmussen and Williams, 2006) is a non-parametric Bayesian regression method that defines a distribution over functions, extending the concept of Gaussian distributions from variables to functions. For a given dataset S (x, y), where x represents the input vector, y represents the output vector (the target to be predicted), and N is the number of observations, the GP model is expressed by the following equation:

$$y = f(x) + \varepsilon \sim N(m(x), k(x, x')) + \varepsilon \qquad (3)$$

where ε is the Gaussian distribution noise value with the mean zero and σ^2 variance $\in \sim (0, \sigma^2)$.

GP process $f(x)$ is fully defined by its mean $m(x)$ and covariance (kernel) $k(x, x')$ functions that are represented respectively by a vector and a matrix in the following equations:

$$m(x) = E(f(x)) \qquad (4)$$

$$k(x, x') = E((f(x) - m(x))(f(x') - m(x'))) \qquad (5)$$

In the above equations, f(x) denotes the random function value of the Gaussian Process (GP) at the input x, representing the modeled quantity (e.g., evaporation). The operator E[·] denotes the expected value (or statistical mean) of the quantity inside the brackets. Specifically, Equation (4) represents the mean function of the GP, while Equation (5) defines the covariance (or kernel) function, which measures the expected joint variability between the function values at inputs x and x'. According to Rasmussen and Williams (2006), the kernel function can be assumed as exponential, squared exponential, rational quadratic, or using a Matern kernel.

Support Vector Machines (SVM)

The Support Vector Machine (SVM) is a supervised learning method commonly used for classification and regression tasks. Developed by Cortes and Vapnik (1995), SVM is based on statistical learning theory. Considering a given training data set ($x1$, $y1$), … , (xN, yN), where x represents the input vector, y is the output value and N is the dimension of the data set, the regression function of SVM can be described as:

$$y = w\,\psi(x) + b \tag{6}$$

where *w* and *b* are the weights and bias, respectively, and $\psi(x)$ is a nonlinear transfer function (kernel) in a feature space.

The regression issue can be expressed as an optimization process that should be minimized with an ε − insensitivity loss function given as follows:

$$\text{minimize } \frac{1}{2}\|w\|^2 + C\left(\sum_{i}^{N}\left(\xi_i + \xi_i^*\right)\right)$$

$$\text{Subject to}\begin{cases} y_i - b - \left[w \cdot \varphi(x_i)\right] \le \varepsilon + \xi_i \\ \left[w \cdot \varphi(x_i) + b - y_i \le \varepsilon + \xi_i^*\right] \\ \xi_i, \xi_i^* \ge 0 \; i = 1, 2, 3, \dots N \end{cases} \tag{7}$$

where *C* is a positive parameter that determines the degree of penalized loss for prediction error, while ξ_j and ξ_i^* represent the slack variables, respectively.

Various kernel including linear, polynomial, sigmoid, and radial basic functions, have been used in SVM to determine the most efficient model for a given type of dataset (Gu et al., 2010; Kisi and Parmar, 2016).

The employed performance metrics

The main goal of this research was to thoroughly evaluate and compare different modeling techniques. To assess the effectiveness of these approaches, we used key performance metrics such as the Correlation Coefficient (*CC*), Mean Absolute Error (*MAE*), and Root Mean Square Error (*RMSE*), calculated using both the training and testing datasets. The predictive performance of the regression models − specifically RF, GP and SVM − depended on identifying the optimal values for user-defined parameters.

Root Mean Square Error (*RMSE*):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(Y_{Oi} - \widehat{Y}_{Ci}\right)^2}{N}} \tag{8}$$

Mean Absolute Error *(MAE*):

$$MAE = \frac{\sum_{i=1}^{N}\left(Y_{Oi} - Y_{Ci}\right)}{N} \tag{9}$$

Correlation Coefficient (*CC*):

$$CC = \frac{\sum_{i=1}^{N}\left(\widehat{Y}_{Ci} - \bar{Y}_{Oi}\right)\left(Y_{Ci} - \bar{Y}_{Oi}\right)}{\sqrt{\left(\sum_{i=1}^{N}\left(Y - \bar{Y}_{CI}\right)^2\right)\left(\sum_{i=1}^{N}\left(Y_{Ci} - \bar{Y}_{Oi}\right)^2\right)}} \tag{10}$$

where (Y_C), (Y_O), ($\bar{Y}_{Oi}$) and (*N*) are the calculated, measured, mean of the observed pan evaporation, and quantity of data, respectively.

We assessed model effectiveness using three key metrics: Correlation Coefficient (*CC*), Mean Absolute Error (*MAE*), and Root Mean Square Error (*RMSE*). The *CC*, ranging from –1 to 1, indicates the strength and direction of a linear relationship, and *MAE* and *RMSE* measure prediction accuracy by quantifying average errors. Whereas *RMSE* penalizes larger errors more heavily, *MAE* treats all errors equally. Lower values for both signify better model performance. These metrics together provide a comprehensive evaluation of model reliability. Table 2 summarizes the optimized parameters used, including kernel functions − Pearson VII, radial basis function, and polynomial − for GP and SVM models, which play a key role in capturing complex data relationships and enhancing prediction accuracy, as further detailed below.

Pearson VII function kernel (PUK):

$$K(P,Q) = \left(\frac{1}{\left[1 + \left[2\frac{\sqrt{\|P-Q\|^2\left(\sqrt{2^{\left(\frac{1}{\omega}\right)}} - 1\right)}}{\sigma}\right]^2\right]^{\omega}}\right) \tag{11}$$

Radial basis kernel (RBF):

$$K(P,Q) = e^{\left(-\gamma\|P-Q\|^2\right)} \tag{12}$$

Polynomial:

$$K(P,Q) = (1 + (P,Q))^d \tag{13}$$

This study used key variables for modeling: "*K*" (kernel), "*P*" (training input), "*Q*" (unlabeled input), and kernel-specific parameters − γ (kernel width), ω (scale), σ (variance), and *d* (polynomial degree).

These parameters significantly influence model behavior and accuracy.

A physical, trial-based approach was used to optimize user-defined parameters for each model − RF, SVM, and GP. For RF, we tuned "*m*" (features per split) and "*K*" (number of trees). For SVM, we adjusted the error-insensitive zone, keeping "*C*" and kernel parameters constant. Identical kernel parameters were used across SVM and GP models for consistency.

Parameter tuning aimed to minimize *RMSE* and maximize *CC*, ensuring accurate, reliable predictions. Ultimate values, selected for optimal performance, are listed in Table 2.

Table 2. Optimal value of user defined parameters (Source: own elaboration)

No	Classifiers used	User defined parameters
1.	RF	*K* = 10, *m* = 1
2.	GP with Poly kernel	Noise = 0.01, *d* = 2
3.	GP with PUK kernel	Noise = 0.01, σ = 1, ω = 0.1
4.	GP with RBF kernel	Noise = 0.01, γ = 1
5.	SVM with Poly kernel	C = 2, *d* = 2
6.	SVM with PUK kernel	C = 2, σ = 1, ω = 0.1
7.	SVM with RBF kernel	C = 2, γ = 1

RESULTS

To evaluate model performance in predicting evaporation rates, we used key performance metrics, with results summarized in Table 3. These outcomes highlight the accuracy and efficiency of each approach, and form the basis for the following discussion on the implications for evaporation forecasting. In the following, *CC* values are dimensionless, while *MAE* and *RMSE* are expressed in mm/day.

Assessment of SVM based models

Figure 3, which refers to the test set, compares observed and predicted evaporation [mm/day] values using SVM with three kernels: polynomial, Pearson VII, and radial basis. The model was trained and tested on carefully selected datasets to ensure robust evaluation.

In training, the Pearson VII kernel outperformed others with a *CC* of 0.9988, *MAE* of 0.0082, and *RMSE* of 0.0092. The radial basis and polynomial kernels showed lower CCs (0.8969 and 0.8992) and higher errors.

In testing, Pearson VII remained superior (*CC*: 0.7733, *MAE*: 1.1575, *RMSE*: 1.4037), outperforming the radial basis and polynomial kernels, which had lower correlation and higher error metrics.

These findings confirm the Pearson VII kernel's effectiveness in enhancing SVM-based evaporation predictions, offering greater accuracy and reliability across both training and testing phases.

To assess GP regression performance, three kernels − polynomial, Pearson VII, and radial basis − were tested. Figure 4 illustrates the comparison between the actual and the predicted evaporation values based on rigorous training and testing.

In training data set, the Pearson VII kernel achieved near-perfect results (*CC*: 0.9999, *MAE*: 0.0002, *RMSE*:

Table 3. Detail of performance evaluation parameters (Source: own elaboration)

Techniques	Training			Testing		
	CC	*MAE*	*RMSE*	*CC*	*MAE*	*RMSE*
RF	0.9874	0.2979	0.3883	0.8066	1.0073	1.2872
GP_PUK	0.9999	0.0002	0.0004	0.7732	1.1572	1.4038
GP_RBF	0.9156	0.7117	0.9481	0.7586	1.1304	1.4161
GP_Poly	0.8753	0.9174	1.1618	0.4842	1.6647	2.1306
SVM_PUK	0.9988	0.0082	0.0092	0.7733	1.1575	1.4037
SVM_RBF	0.8969	0.7937	1.0449	0.7670	1.1634	1.4926
SVM_Poly	0.8992	0.7749	1.0363	0.7422	1.2468	1.6039

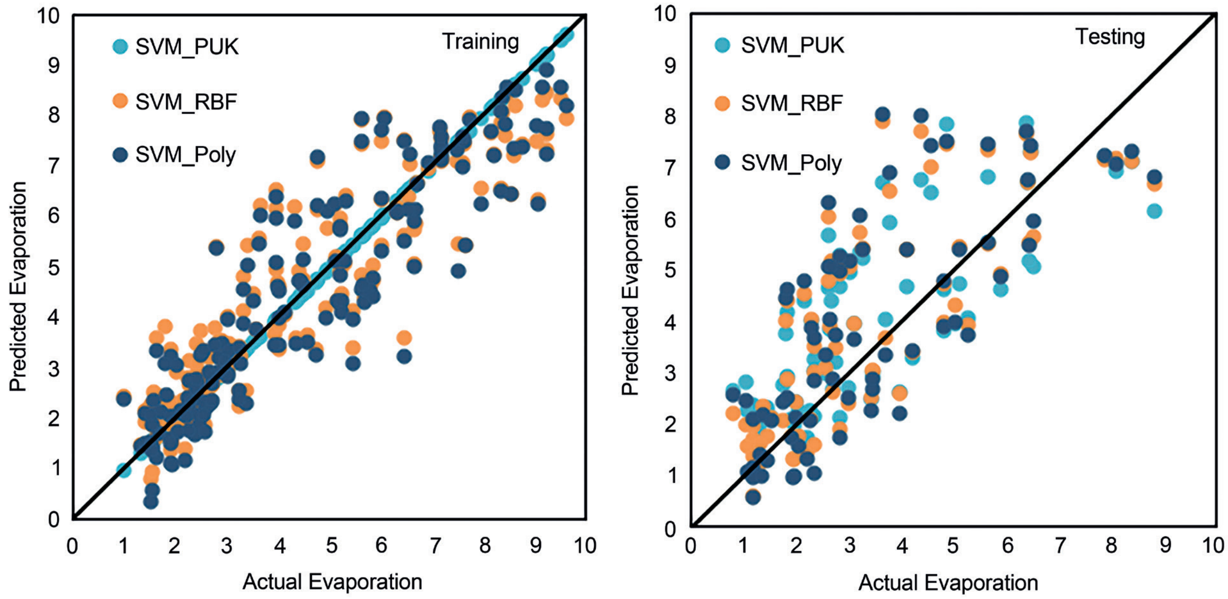

Fig. 3. Actual vs predicted values of evaporation using SVM with poly, PUK and RBF for training and testing data sets (Source: own elaboration)

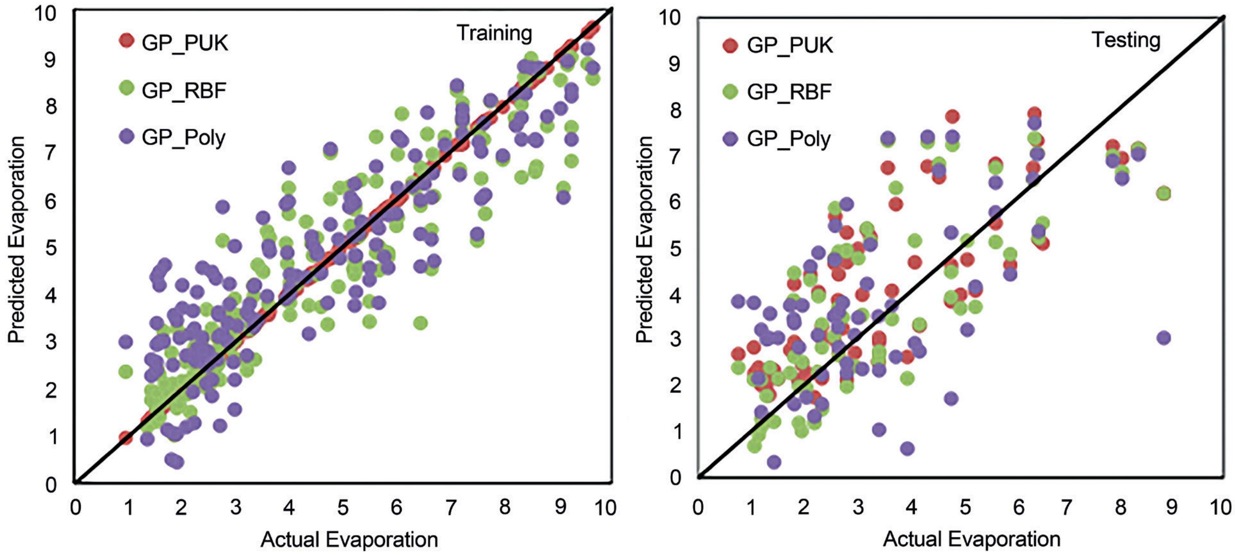

Fig. 4. Actual vs predicted values of evaporation using Gaussian process with poly, PUK and RBF for training and testing data sets (Source: own elaboration)

0.0004), clearly outperforming the radial basis (*CC*: 0.9156) and polynomial kernels (*CC*: 0.8753), both with notably higher error metrics.

In testing data set, Pearson VII again led with a *CC* of 0.7732, *MAE* of 1.1572, and *RMSE* of 1.4037. Radial basis and polynomial kernels followed with lower CCs and higher errors.

These findings highlight the Pearson VII kernel's superior accuracy and robustness in GP regression, confirming its value for reliable evaporation forecasting.

Assessment of RF based models

To evaluate evaporation prediction, we applied a Random Forest (RF) regression model to both training and testing datasets. Figure 5 compares observed and predicted values, illustrating model performance.

In training, RF showed strong accuracy with a *CC* of 0.9874, *MAE* of 0.2979, and *RMSE* of 0.3883. On the testing dataset, it maintained solid performance (*CC*: 0.8066, *MAE*: 1.0073, *RMSE*: 1.2870), with a moderate increase in error.

These results confirm RF's effectiveness in capturing evaporation patterns and delivering reliable predictions, particularly during training, with consistent generalization in testing.

Inter-comparison of the best developed models

To enhance evaporation prediction, we applied SVM with the Pearson VII kernel, GP, and RF models to both training and testing datasets. Figure 6 compares observed and predicted values, showing all three models effectively captured key data patterns.

As detailed in Table 3, the Pearson VII kernel delivered strong results in both GP and SVM, with nearly identical performance (*CC*: 0.7732 and 0.7733; *MAE*: 1.1572 and 1.1575; *RMSE*: 1.4037 for both). While SVM slightly outperformed GP in CC, the RF model achieved the highest *CC* (0.8066) and the lowest *RMSE* (1.2870), indicating the strongest predictive performance overall.

These results confirm the effectiveness of all three models, with RF showing the best generalization, and the Pearson VII kernel proving highly reliable in kernel-based approaches.

To evaluate model performance in capturing complex patterns, we used the Taylor diagram, which quantifies correlation, standard deviation, and *RMSE*. Generated via the Weka platform, the Taylor diagram (Figure 7) visually compares the accuracy of different soft computing techniques in predicting evaporation rates.

Results confirm earlier findings: the RF model showed the highest accuracy and reliability, clearly outperforming other models. The results yielded by the RF model were closer to the observed point, which provided higher SD and Correlation Coefficient (equal to 0.8066) and lower RMSE (equal to 1.2870). In contrast, the GP_Poly model had a lower correlation

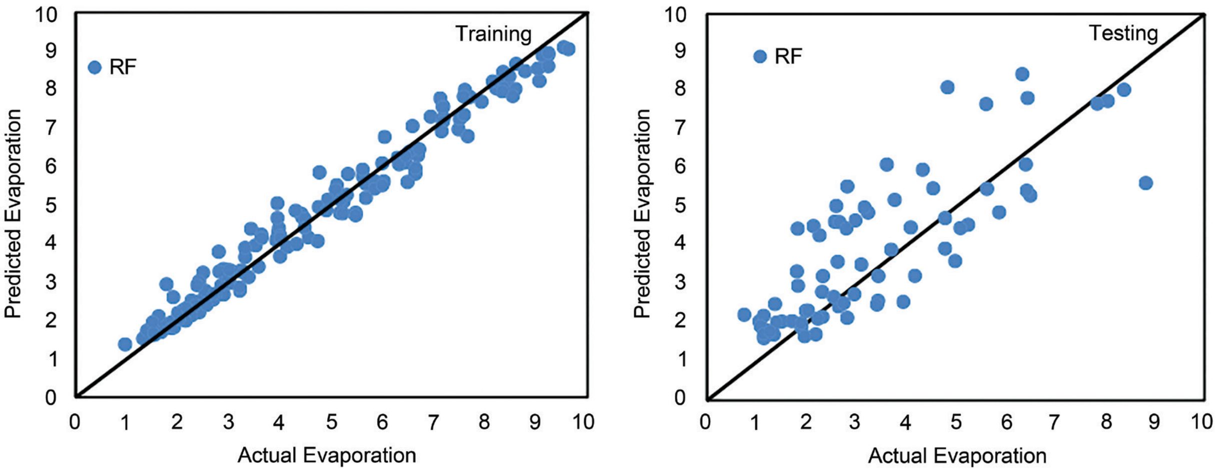

Fig. 5. Actual vs predicted values of evaporation using Random Forest regression for training and testing data sets (Source: own elaboration)

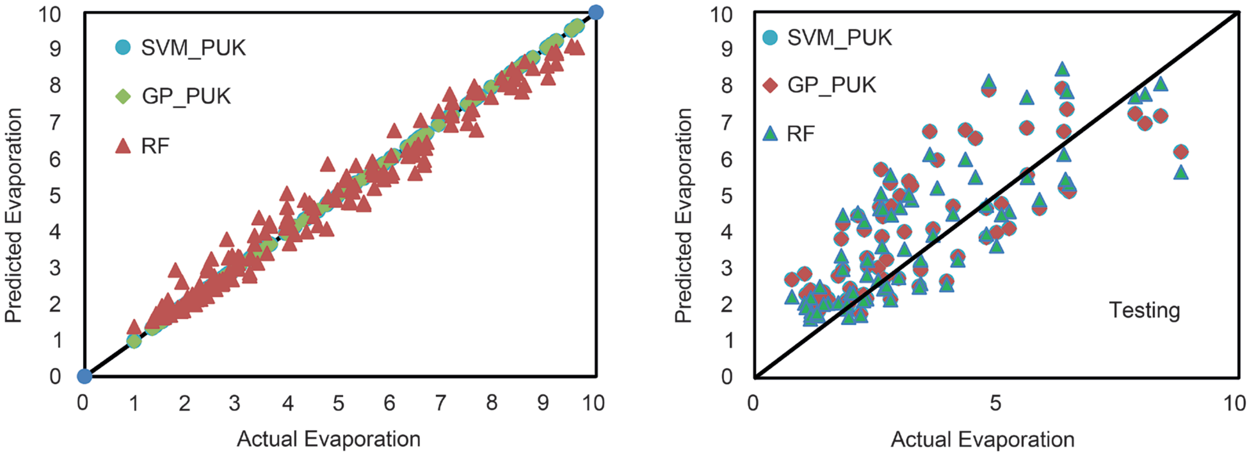

Fig. 6. Actual vs predicted values of evaporation using various random forest, Pearson VII kernel based SVM and GP for training and testing data sets (Source: own elaboration)

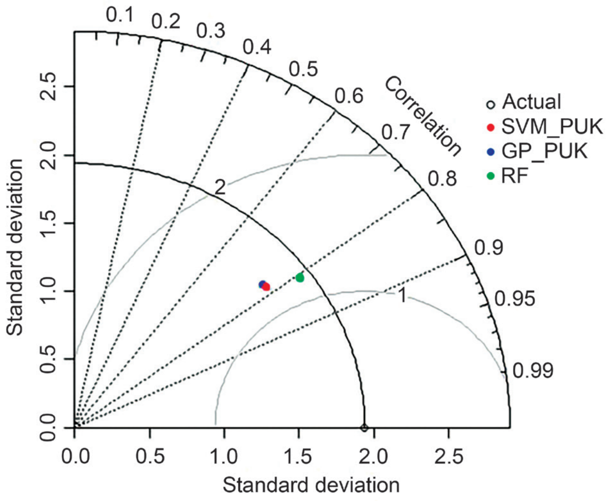

Fig. 7. Taylor's diagram for evaporation rate using the RF, GP, and SVM modelling techniques (Source: own elaboration)

(*CC* = 0.48), reinforcing its weaker predictive performance in this context.

Additionally, we used a box plot (Figure 8) to visualize the distribution and variability of predicted evaporation rates across RF, GP, and SVM models. The RF model displayed the most consistent results compared with the other models, with median and mean values closely matching the observed data, highlighting its robustness and precision, whereas the GP and SVM models showed little similarity.

A violin plot, which combines box plot and density plot features, was used to visualize the distribution of evaporation rate predictions across different models and parameter settings. Each "violin" reflects the data spread and density, highlighting central tendencies such as median and quartiles.

As shown in Figure 9, the RF model displays the most concentrated distribution, with relative errors clustering closer to zero − indicating higher consistency and lower variability compared to GP and SVM models.

Sensitivity analysis

To assess the influence of input variables on evaporation prediction, we performed sensitivity analysis using the RF model. Key factors − T_{max}, T_{min}, U, H, Tw_{max}, and Tw_{min} − were tested in various combinations to isolate their individual impact.

Whether we use the *CC*, *MAE*, or *RMSE* as evaluation metrics, results (Table 4) show that relative humidity (H) plays a critical role. While removing other variables had minimal effect ($CC \approx 0.80$), excluding H dropped the *CC* to 0.77 and increased both *MAE* and *RMSE*, underscoring its importance in accurate evaporation forecasting.

These results underscore the dominant role of relative humidity (*H*) in shaping the evaporation process, reinforcing its critical influence among the parameters considered. This insight enhances our understanding

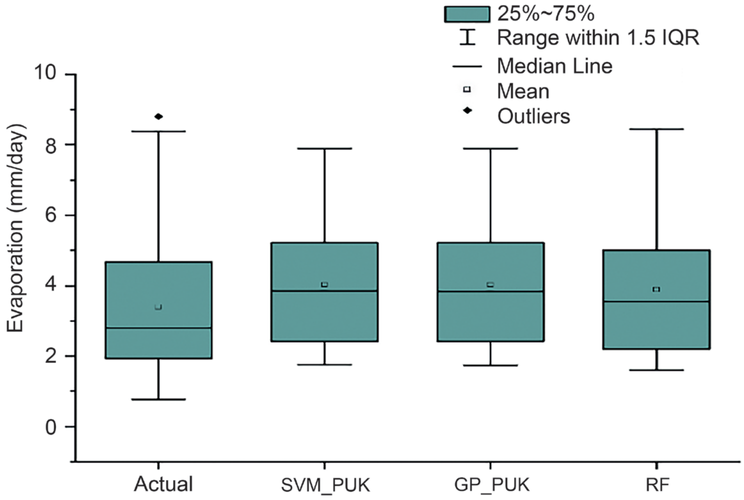

Fig. 8. Boxplot for evaporation rate using the actual, RF, GP, and SVM modelling techniques (Source: own elaboration)

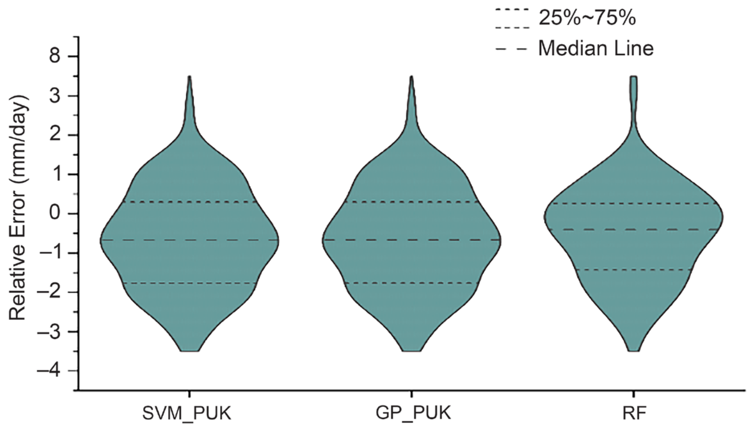

Fig. 9. Violin plots for evaporation rate using RF, GP, and SVM modelling techniques (Source: own elaboration)

Table 4. Sensitivity analysis using random forest regression (Source: own elaboration)

Input parameter						Output	RF regression		
T_{max}	T_{min}	U	H	Tw_{max}	Tw_{min}	Evaporation	CC	MAE	RMSE
✓	✓	✓	✓	✓	✓	✓	0.8066	1.0073	1.2870
x	✓	✓	✓	✓	✓	✓	0.7943	1.0323	1.2946
✓	x	✓	✓	✓	✓	✓	0.8013	1.0010	1.2887
✓	✓	x	✓	✓	✓	✓	0.8144	1.0381	1.3843
✓	✓	✓	x	✓	✓	✓	0.7768	1.0569	1.3538
✓	✓	✓	✓	x	✓	✓	0.8076	1.0942	1.3800
✓	✓	✓	✓	✓	x	✓	0.8067	1.0196	1.3016

Explanations: ✓ denotes the input parameter considered, × denotes the input parameter removed of the factors driving evaporation and provides valuable implications for modeling and predicting this essential phenomenon.

DISCUSSION

The results of the present study have practical significance for the management of the Boukourdane Dam, as accurate daily evaporation estimates are essential for optimizing water allocation for both domestic supply and irrigation purposes. By identifying the most effective kernel functions for SVR and GPR models, the study provides water managers with reliable predictive tools to anticipate water losses due to evaporation. This can support more informed decision-making, improve reservoir operation efficiency, and help mitigate the impacts of water scarcity in the region.

The study assessed the predictive performance of three machine learning models − Support Vector Regression (SVR), Gaussian Process Regression (GPR), and Random Forests (RF) − for estimating monthly evaporation rates, with a focus on kernel function influence in SVR and GPR. Model accuracy was evaluated using Correlation Coefficient (*CC*), Mean Absolute Error (*MAE*), and Root Mean Square Error (*RMSE*).

Among SVR and GPR models, those using the Pearson VII kernel consistently outperformed their radial basis function (RBF) and polynomial counterparts, highlighting the kernel's adaptability to complex, non-linear evaporation patterns. Its flexible shape parameters allow for better modeling of diverse data structures, while improving similarity assessment and overall predictive performance.

The RF model achieved the highest *CC* and lowest *RMSE* on the test set, demonstrating superior generalization and robustness. Taylor diagram and violin plot analyses further confirmed RF's consistent and low-error predictions across the dataset.

These findings are consistent with those of Sattari et al. (2021), who reported that the Pearson VII function-based Support Vector Machine (SVM) achieved the highest predictive accuracy for evapotranspiration estimation in Turkey. In contrast, Shabani et al. (2024) found that Gaussian Process Regression (GPR) outperformed both SVM and Random Forest (RF) models in a similar application in Iran. Such discrepancies highlight the critical influence of regional climatic conditions, data resolution, and local environmental factors on model performance. They further underscore the importance of conducting site-specific model validation to ensure reliable predictions, as the relative effectiveness of machine learning approaches can vary substantially across different geographic and hydrological contexts.

Due to limited meteorological data, traditional empirical methods (e.g., Penman, Thornthwaite) couldn't be applied. Nonetheless, the strong performance of RF and Pearson VII-based SVR and GPR models shows their value in data-scarce environments, offering viable alternatives when conventional variables like solar radiation or vapor pressure are unavailable.

A sensitivity analysis using RF identified relative humidity (H) as the most influential variable. Removing H led to the sharpest drop in performance, while other exclusions had minor effects. Although this study used monthly data, future work will explore higher-resolution datasets and handle missing data systematically to enhance accuracy.

Lastly, while known drivers such as solar radiation and vapor pressure were excluded due to data gaps, their inclusion could further improve the quality of results. Future research will extend this work to sites with richer datasets to evaluate the added value of these variables in evaporation modeling.

CONCLUSION

The present study highlights the effectiveness of machine learning models − Random Forest (RF), Gaussian Process (GP), and Support Vector Machine (SVM) − for predicting evaporation, using six key meteorological variables. Among them, RF achieved the best overall performance (*CC* = 0.81, *MAE* = 1.01 mm, RMSE = 1.29 mm), while the Pearson VII kernel delivered the highest accuracy within both GP and SVM frameworks. Sensitivity analysis confirmed relative humidity as the most influential predictor.

From an applied perspective, these results support the integration of ML models into national evaporation monitoring systems, particularly in data-scarce areas. With access to real-time meteorological data, such models can enable reliable operational forecasting for water resource management, agriculture, and climate resilience planning.

Nonetheless, several limitations remain. The study's single-site focus and use of monthly data reduce generalizability and overlook short-term variability. Moreover, the lack of key inputs − solar radiation and vapor pressure − limited both the predictive accuracy and the possibility of comparing with traditional empirical models. These variables are known to drive evaporation by influencing energy input and atmospheric moisture, and their inclusion would likely enhance model performance.

Future work should address these constraints by using higher-resolution data, robust methods for handling missing inputs, and expanding to multiple locations with diverse climates. Incorporating additional variables, applying deep learning techniques, and leveraging real-time data assimilation could further improve model robustness and broaden its applicability.

REFERENCES

Adnan, R.M., Mostafa, R.R., Elbeltagi, A., Yaseen, Z.M., Shahid, S., Kisi, O. (2022). Development of new machine learning model for streamflow prediction: case studies in Pakistan. Stoch. Environ. Res. Risk Assess., 36, 999–1033.

Benzagtha, M.A. (2014). Estimation of evaporation from a reservoir in semi-arid environments using artificial neural network and climate based models. Br. J. Appl. Sci. Technol., 4, 3501–3518.

Boudjerda, M., Mu'azu, M.A., Petroselli, A. (2024). Prediction of reservoir evaporation considering water temperature and using ANFIS hybridized with metaheuristic algorithms. Earth Sci. Inform., 17, 1779–1798.

Breiman, L. (2001). Random forests. Mach. Learn., 45(1), 5–32.

Cappelli, F., Tauro, F., Apollonio, C., Petroselli, A., Borgonovo, E., Grimaldi, S. (2023). Feature importance measures to dissect the role of sub-basins in shaping the catchment hydrological response: a proof of concept. Stoch. Environ. Res. Risk Assess., 37(4), 1247–1264.

Cortes, C., Vapnik, V. (1995). Support-vector networks. Mach. Learn., 20(3), 273–297.

De Caro, D., Ippolito, M., Cannarozzo, M., Provenzano, G., Ciraolo, G. (2023). Assessing the performance of the Gaussian Process Regression algorithm to fill gaps in the time-series of daily actual evapotranspiration of different crops in temperate and continental zones using ground and remotely sensed data. Agric. Water Manag., 290, 108596.

Dimitriadis, P., Tegos, A., Koutsoyiannis, D. (2021). Stochastic analysis of hourly to monthly potential evapotranspiration with a focus on the long-range dependence and application with reanalysis and ground-station data. Hydrology, 8(4), 177. DOI:10.3390/hydrology8040177.

Garrick, D.E., Hanemann, M., Hepburn, C. (2020). Rethinking the economics of water: an assessment. Oxf. Rev. Econ. Policy, 36, 1–23.

Giordano, R., Milella, P., Portoghese, I., Vurro, M., Apollonio, C., D'Agostino, D. (2010). An innovative monitoring system for sustainable management of groundwater resources: objectives, stakeholder acceptability and implementation strategy. 2010 IEEE Workshop on Environmental Energy and Structural Monitoring Systems, Taranto, Italy, 32–37.

Grimaldi, S., Cappelli, F., Papalexiou, S.M., Petroselli, A., Nardi, F., Annis, A., Piscopia, R., Tauro, F., Apollonio, C. (2024). Optimizing sensor location for the parsimonious design of flood early warning systems. J. Hydrol. X, 100182.

Gu, Y., Zha, W., Wu, Z. (2010). Least squares support vector machine algorithm. J. Tsinghua Univ. Sci. Technol., 7, 1063–1066.

Jasmine, M., Mohammadian, A., Bonakdari, H. (2022). On the prediction of evaporation in arid climate using machine learning model. Math. Comput. Appl., 27(2), 32. DOI:10.3390/mca27020032.

Kamienski, C., Soininen, J.P., Taumberger, M., Dantas, R., Toscano, A., Salmon Cinotti, T., Filev Maia, R., Torre Neto, A. (2019). Smart water management platform: IoT-based precision irrigation for agriculture. Sensors, 19(2), 276. DOI:10.3390/s19020276.

Kisi, O., Parmar, K.S. (2016). Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution. J. Hydrol., 534, 104–112.

Kumar, N., Arakeri, J.H. (2021). A fast method to measure the evaporation rate. J. Hydrol., 594, 125642.

LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

Melišová, E., Vizina, A., Hanel, M., Pavlík, P., Šuhájková, P. (2021). Evaluation of evaporation from water reservoirs in local conditions at Czech Republic. Hydrology, 8, 153, 93–109.

Pahl-Wostl, C. (2017). An evolutionary perspective on water governance: From understanding to transformation. Water Resour. Manag, 31, 2917–2932.

Rajput, J., Kushwaha, N.L., Srivastava, A., Pande, C.B., Suna, T., Sena, D.R., Singh, D.K., Mishra, A.K., Sahoo, P.K., Elbeltagi, A. (2024). Development of machine learning models for estimation of daily evaporation and mean temperature: a case study in New Delhi, India. Water Pract. Technol., wpt2024144.

Rasmussen, C.E., Williams, C.K. (2006). Gaussian processes for machine learning. MIT Press.

Sattari, M.T., Apaydin, H., Band, S.S., Mosavi, A., Prasad, R. (2021). Comparative analysis of kernel-based versus ANN and deep learning methods in monthly reference evapotranspiration estimation. Hydrol. Earth Syst. Sci., (25), 603–618.

Seifi, A., Soroush, F. (2020). Pan evaporation estimation and derivation of explicit optimized equations by novel hybrid meta-heuristic ANN based methods in different climates of Iran. Comput. Electron. Agric., 173, 105418.

Shabani, M., Asadi, M.A., Fathian, H. (2024). Improving the daily pan evaporation estimation of long short-term memory and support vector regression models by using the Wild Horse Optimizer algorithm. Water Supply, 24(4), 1315–1334.

Shabani, S., Samadianfard, S., Sattari, M.T., Mosavi, A. (2020). Shamshirband S., Kmet T.,Várkonyi Kóczy A.R. Modeling Pan Evaporation Using Gaussian Process Regression KNearest Neighbors Random Forest and Support Vector Machines; comparative analysis. Atmosphere, 11(66).

Sharafi, M., Samadianfard, S., Behmanesh, J., Prasad, R. (2023). Integration of fruit fly and firefly optimization algorithm with support vector regression in estimating daily pan evaporation. International Journal of Biometeorology, 68(2), 237–251.

Singh, A., Singh, R.M., Kumar, A.R., Kumar, A., Hanwat, S., Tripathi, V.K. (2019). Evaluation of soft computing and regression-based techniques for the estimation of evaporation. J. Water & Clim. Change, 12(2), 32–43. DOI:10.2166/wcc.2019.101.

Tao, H., Diop, L., Bodian, A., Djaman, K., Ndiaye, P.M., Yaseen, Z.M. (2018). Reference evapotranspiration prediction using hybridized fuzzy model with firefly algorithm: regional case study in Burkina Faso. Agric. Water Manag., 208, 140–151.

Wu, M., Feng, Q., Wen, X., Deo, R., Yin, Z., Yang, L., Sheng, D. (2020). Random Forest predictive model development with uncertainty analysis capability for the estimation of evapotranspiration in an arid oasis region. Hydrol. Res., 51(4), 648–665.

Zhao, B., Huntington, J., Pearson, C., Zhao, G., Ott, T., Zhu, J., Weinberg, A., Holman, K.D., Zhang, S., Anderson, R., Strickler, M., Cotter, J., Nelun, F., Nowak, K., Gao, H. (2024). Developing a general daily lake evaporation model and demonstrating its application in the state of Texas. Water Resour. Res., 60, e2023WR036181.

PROGNOZOWANIE EWAPORACJI OPARTE NA UCZENIU MASZYNOWYM ORAZ ANALIZA FUNKCJI JĄDROWYCH: STUDIUM PRZYPADKU ZAPORY BOUKOURDANE W ALGIERII

ABSTRAKT

Cel badania

Celem niniejszego badania jest ocena wpływu trzech funkcji jądrowych – Pearsona VII, radialnej funkcji bazowej (RBF) i wielomianu – na wydajność predykcyjną modeli regresji wektorów nośnych (SVR) oraz regresji procesu gaussowskiego (GPR).

Materiały i metody

Do oszacowania miesięcznej ewaporacji na zaporze Boukourdane w Algierii zastosowano trzy modele uczenia maszynowego: model lasu losowego (RF), regresję wektorów nośnych (SVR) i regresję procesu gaussowskiego (GPR). Zbiór danych obejmował 240 obserwacji dokonanych na przestrzeni 20 lat, z następującymi danymi wejściowymi: maks./min. temperatura powietrza, wilgotność względna, prędkość wiatru i temperatura wody; przy czym dane wyjściowe dotyczyły ewaporacji.

Wyniki i wnioski

Wydajność modelu oceniono za pomocą współczynnika korelacji (*CC*), pierwiastka średniokwadratowego błędu (*RMSE*) i średniego błędu bezwzględnego (*MAE*). Wydajność modelu RF przewyższyła wydajność modeli GPR i SVR w przypadku wszystkich funkcji jądrowych, osiągając w testach następujące wskaźniki: *MAE* = 1,01 mm, *RMSE* = 1,29 mm, i *CC* = 0,81. Co więcej, funkcja jądrowa Pearson VII zapewniła najwyższą dokładność zarówno w ramach procesu gaussowskiego, jak i maszyny wektorów nośnych. Analiza wrażliwości wskazała, że wilgotność względna jest najbardziej wpływowym czynnikiem w prognozowaniu ewaporacji.

Słowa kluczowe: prognozowanie parowania, proces gaussowski, regresja lasu losowego, maszyny wektorów nośnych, analiza wrażliwości