Journal of Hydroinformatics

Journal of Hydroinformatics Vol 26 No 7, 1661 doi: 10.2166/hydro.2024.016

Forecasting daily rainfall in a humid subtropical area: an innovative machine learning approach

Miran Hikmat Mohammed [a] and Sarmad Dashti Latif [b,c,*]

[a] Basic Science Department, College of Dentistry, University of Sulaimani, Sulaymaniyah, Kurdistan Region, Iraq
[b] Civil Engineering Department, College of Engineering, Komar University of Science and Technology, Sulaimany, Kurdistan Region, Iraq
[c] Scientific Research Center, Soran University, Soran, Erbil, Kurdistan Region, Iraq
*Corresponding author. E-mail: sarmad.latif@komar.edu.iq

MHM, 0000-0003-1368-6374; SDL, 0000-0002-0417-3545

ABSTRACT

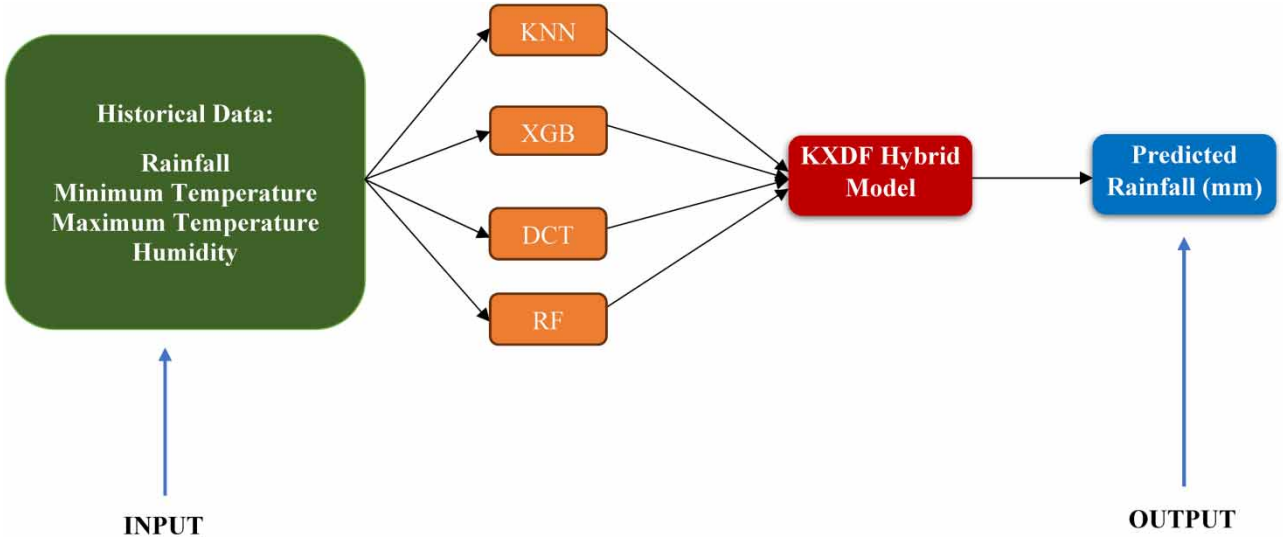
Hydrological modeling is one of the most complicated tasks in sustainable water resources management, particularly in terms of predicting rainfall. Predicting rainfall is critical to build a sustainable society in terms of hydropower operations, agricultural planning, and flood control. In this study, a hybrid model based on the integration of k-nearest neighbor (KNN), XGBoost (XGB), decision tree (DCT), and Random Forest (RF) has been developed and implemented for forecasting daily rainfall for the first time at Sydney airport, Australia. Daily rainfall, temperature, evaporation, and humidity have been selected as input parameters. Three statistical measurements, namely, root mean square error (RMSE), coefficient of determination (R^2), mean absolute error (MAE), and Normalized Root Mean Square Error (NRMSE) have been utilized in order to check the accuracy of the proposed model. A sensitivity analysis was conducted, and the results indicated that for the purpose of prediction, the temperature, humidity, and evaporation were highly sensitive to the rainfall data. According to the results, the developed hybrid model was capable of predicting daily rainfall with high performance for both training and testing parts with RMSE = 0.124, R^2 = 0.999, MAE = 0.007, NRMSE = 0.04 and RMSE = 1.246, R^2 = 0.991, MAE = 0.109, NRMSE = 0.339, respectively.

Key words: hybrid model, machine learning, rainfall prediction, sustainable development, Sydney airport, water resources management

HIGHLIGHTS

- A hybrid novel model has been developed for rainfall prediction for the first time.
- Accurate rainfall prediction will lead to better management of water resources.
- Rainfall prediction is an important hydrological tool due to global climate change.
- Farmers use rainfall forecast methods to plan for irrigation decisions.

GRAPHICAL ABSTRACT



ABBREVIATIONS

KNN	k-nearest neighbor
XGB	XGBoost
DCT	decision tree
RMSE	random forest
R^2	root mean square error
MAE	mean absolute error
NRMSE	normalized root mean square error
LSTM	long short-term memory
KXDF	the developed hybrid model

1. INTRODUCTION

For human life, societal growth, and ecological (natural, biological, and environmental) health, water is a vital natural resource. Water is essential for manufacturing, agriculture, and biotransformation, whether for drinking or personal hygiene. Since the global population is rapidly increasing, the demand for water, especially fresh water, is increasing as well. The demand for food increases and it all depends on irrigation, which requires a huge amount of water. As a result, humans will face more water problems. Water shortage has become a major issue worldwide in recent decades, particularly in developing countries (Gohari *et al.* 2013; Sowby 2023; Villanueva *et al.* 2023).

Rainfall from the atmosphere is an important component of the global water cycle. The main source of water supply, especially in dry and semi-arid regions, is rainfall. The richness of water resources in any area depends on the amount of rainfall. Rainfall is another crucial indicator for assessing the ecological integrity of a place because it can significantly reflect the dynamic changes brought on by drought. As a result, predicting rainfall is commonly needed in hydrological prediction and water resource management. Forecasting rainfall also has a big effect on how the economy and people's lives grow. Forecasting rainfall can assist individuals in developing plausible strategies to lessen the effects of unforeseen climatic disasters as well as in predicting the occurrence of disasters. For instance, Tesco, a British grocery chain, decreased costs by 30% and saved over 6 million pounds in 2013 by modifying warehouse inventory and sales methods based on weather forecast information. However, because of how unpredictable, diverse, and complex meteorological circumstances are, there are a lot of uncertainties and unpredictability in the process of rainfall. It has been challenging to pinpoint the precise amount of rain that will fall at a certain location and time in the future due to physical factors so far. The hydrological field faces more challenges as a result of the rising demand for rainfall prediction than just the atmosphere's inherent complexity and its associated dynamic processes do, which has the machine learning community interested in their research (Zhang *et al.* 2020; Latif *et al.* 2023).

With 70% of the land classed as desert or semi-desert, Australia is the driest inhabited continent. This island continent's water supply is dependent on rainfall. Variation in rainfall, like variation in any other region, frequently influences water availability across Australia. The continent experiences a variety of rainfall patterns, including dryness (lack of rain), food (excess rain), and droughts. Some examples include the Millenium drought, which lasted from 1995 to 2009, the 1970s dry shift in southwest Australia, and widespread flooding in eastern Australia from 2009 to 2012. The tropical regions in the north receive the most rainfall, while the interior is dry and deserted. These variations in rainfall raise serious concerns about water availability, management, and future resource planning (Raval *et al.* 2021).

Many studies focused on developing machine learning techniques to forecast hydrological parameters (Praveen *et al.* 2020; Koppa *et al.* 2022; Mohammadi *et al.* 2022), especially rainfall (Dash *et al.* 2018; Xiang *et al.* 2018). For instance, Chatterjee *et al.* (2018) proposed a study to develop a novel method for forecasting rainfall in India. They have applied two algorithms, namely, Greedy forward selection and k-means algorithms with neural networks in order to develop a hybrid model. They have also applied a multilayer perceptron hybrid model for comparison purposes with the proposed hybrid model. According to their findings, it is revealed that the proposed hybrid model outperformed the conventional model with a huge difference in terms of accuracy. Moreover, Kumar *et al.* (2021) developed two hybrid models for predicting rainfall in India. They have also applied three other machine learning algorithms for comparison purposes with their hybrid models. According to their results, their proposed hybrid models performed better than single machine learning algorithms. Furthermore, Bellido-Jiménez *et al.* (2021) implemented several artificial intelligence techniques to forecast rainfall in different areas in Spain. Based on their findings, the utilization of neighbor data within a 50-km radius outperformed the other options tested. Furthermore, inland areas outperformed coastal areas in most locations, indicating that efficiency effects based on distance to the sea exist. For their proposed models, MLP outperformed simple RF. In addition, Xiang *et al.* (2020) proposed a study for estimating hourly rainfall-runoff in Clear Creek and Upper Wapsipinicon River in Iowa. They have applied a prediction model based on long short-term memory (LSTM) and the seq2seq structure. Based on their results, the LSTM-seq2seq outperformed other proposed machine learning models including single LSTM. According to their findings, the LSTMseq2seq model has enough predictive capacity to enhance forecast accuracy in short-term flood forecasting implementations. On the other hand, Aggarwal *et al.* (2023) showed the influence of uncertainty of climate change in small regions. They have mentioned that there is a high correlation between temperature and rainfall parameters in Ludhiana district, Punjab, India. This shows that hydrological and meteorological parameters should be taken into consideration for sustainable water resources management. Moreover, different models could be used for different fields in water resources management. Other models were also used for various fields in water resources management. For instance, Eltarabily *et al.* (2023) utilized the Slide2 model for estimating seepage loss.

In this study, a hybrid model based on the integration of four machine learning algorithms has been developed and applied in order to forecast daily rainfall at Sydney Airport, Australia. Several statistical indices were utilized to check the accuracy of the proposed models. The conventional machine learning model could not successfully predict the rainfall parameter in the humid subtropical areas due to the complexity of the weather pattern. Therefore, the novelty of this study is that it is the first time these four algorithms, namely, k-nearest neighbor (KNN), XGBoost (XGB), decision tree (DCT), and RF have been combined as a hybrid model in order to predict daily rainfall. The developed model could be practically used by the water sectors to achieve sustainable water resources management.

2. MATERIALS AND METHODS

2.1. Study area and data

Due to the city's closeness to the water, Sydney has a temperate climate with warm, occasionally blistering summers and mild winters with no discernible seasonal differences (Imteaz & Moniruzzaman 2018). Sydney Airport is Australia's main airport and the city's most important piece of infrastructure (Bowyer & Chapman 2014). The Alexandra Canal forms the northern boundary of Sydney Airport, the Cooks River forms the western boundary, Botany Bay forms the southern boundary, and major wetlands comprise the eastern boundary. Furthermore, groundwater runs beneath the location. Despite the fact that total water demand at Sydney Airport is expected to increase, the airport is dedicated to reducing potable water consumption per passenger (Sydney Airport 2019; Latif 2023). Since evaporation allows molecules to form water vapor and this vapor forms clouds then again it will transfer to form water, forecasting evaporation is crucial for estimating future water amounts. Figure 1 shows the location of the study area based on the Google Map.

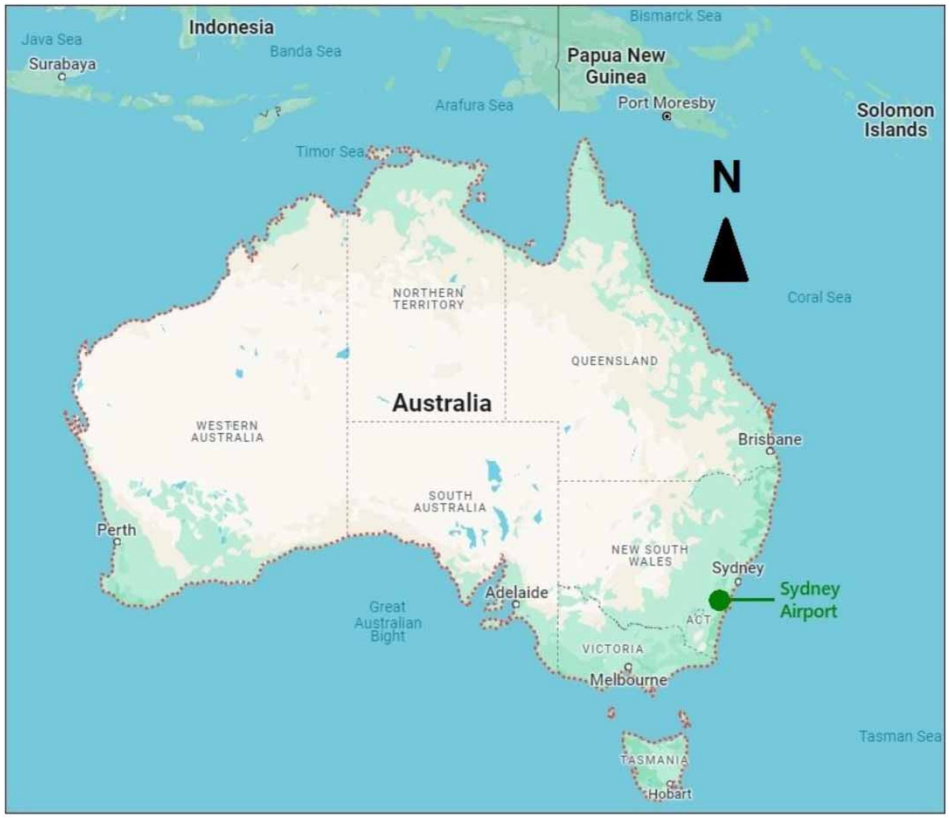


Figure 1 | Study area.

In this study, four hydrological parameters have been utilized for developing the proposed hybrid model. The parameters were rainfall, minimum temperature, maximum temperature, humidity, and evaporation. The humidity data was in two different parts: the first one was recorded at 9:00 am, and the second one was recorded at 3:00 pm. Also, the collected daily temperature was two different data, the first one was minimum temperature, and the second one was maximum temperature. To eliminate limitations, 1-day ahead, 2-day ahead, 3-day ahead and 4-day ahead, we considered the proposed prediction model. The utilized data at Sydney Airport was collected by the Australian Government Bureau of Meteorology. The duration of the utilized data was from 1 January 2008 to 25 June 2017. Table 1 represents a descriptive analysis of the observed data. Figure 2 represents the daily rainfall time-series data at Sydney Airport. For instance, rainfall and evaporation units are in mm, however, temperature units are in degrees celsius, °C.

2.2. KXDF hybrid model

In this study, a hybrid model for predicting rainfall is developed based on the integration of KNN, XGB, DCT, and RF. The developed hybrid model was named KXDF, referring to the four utilized machine learning algorithms. KNN regression is a

Table 1 | Descriptive analysis of the data observed from Sydney airport

Data	Mean	Median	Mode	SD	Min	Max
Min. Temperature (C°)	14.87	14.9	11.1	4.55	4.3	27.6
Max. Temperature (C°)	23.00	22.8	19.6	4.49	11.7	45.8
Rainfall (mm)	3.32	0	0	9.88	0	119.4
Evaporation (mm)	5.17	4.8	4	2.77	0	18.4
Humidity (9:00 am)	68.18	69	71	15.1	19	100
Humidity (3:00 pm)	54.68	56	59	16.3	10	99

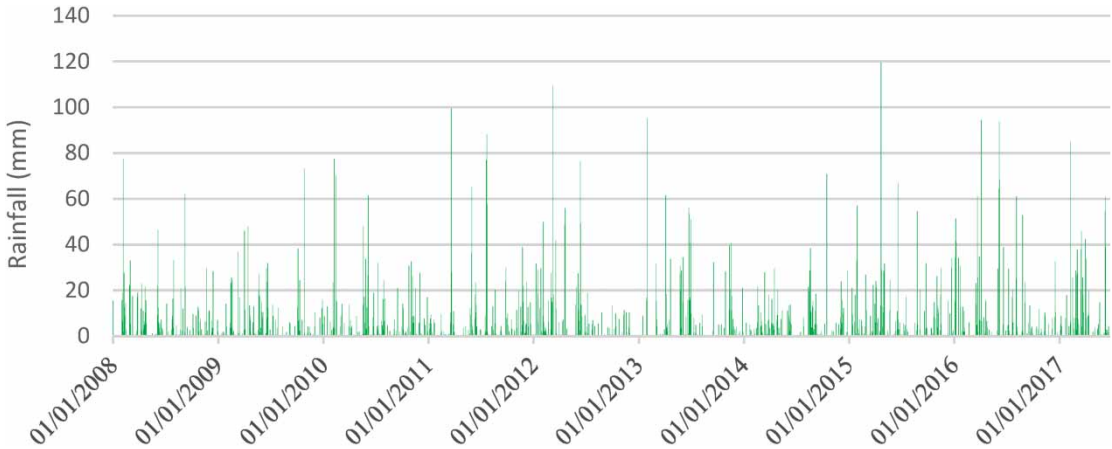


Figure 2 | Daily rainfall at Sydney Airport from 1 January 2008 to 25 June 2017.

non-parametric method for approximating the relationship between independent variables and continuous outcomes in an understandable manner by averaging data in the same neighborhood (Song *et al.* 2017). The gradient boosted trees method is effectively implemented by the open-source tool XGBoost. Gradient boosting is a supervised learning technique that combines the forecasts of a number of weaker, simpler models in an effort to accurately predict a target variable (Yu *et al.* 2020). Because they are easy to comprehend and efficient, decision trees are frequently the tool of choice for predictive modeling. The primary function of a decision tree is to partition a large amount of data into more manageable chunks (Mehraein *et al.* 2022). RF is a decision tree algorithm that is widely used for regression time-series prediction modelling especially in the hydrological area (Saadi *et al.* 2019). Performance assessment has been carefully accomplished to check the accuracy of the proposed model (Krishnaraj & Honnasiddaiah 2022).

The techniques used for predictions in this study are machine learning algorithms using Python programming language version 3.9, the latest Python version with the updated list of packages. The reason for using Python is that this programming language provides many available packages in different programming areas, especially in machine learning fields, which use CSV datasets for data preprocessing and predictions.

The process starts with preparing the dataset and calling the CSV file into the Python programming language. After that, the proposed dataset will pass through several steps in the Python machine learning models. The data preparation techniques include managing null values and handling data to the numerical data type. In addition, the dataset passes through testing of any possibility of outliers among the columns; if any are found, the normalization technique is used by employing a min-max scaler. The normalization technique works on scaling the values in the dataset into the same range.

Pandas, Matplotlib, Seaborn, and Sklearn for model prediction and model metrics were used in the programming part. The Pandas package is used to read and manage different operations on the dataset, such as adding, removing, and updating the values of rows. Moreover, Matplotlib and Seaborn are used for plotting and visualizing the data before and after manipulation to show the most relevant changes in data. Also, the most important package used in this study is Sklearn which calls machine learning models and embeds different evaluation metrics such as RMSE, MAE, NRMSE, and R^2 score. Also, with the Sklearn package, a cross-validation package is used. Figure 3 shows the structure of the proposed hybrid model.

KXDF is a hybrid model, consisting of KNN, X-GBoost, Decision Tree, and RF. These models work together to make predictions as base-0. Also, the process works as each model makes its prediction separately, then at the stage called stacking regressor model all the outcomes of predictions are combined, at the last stage they passed to the final model as base –1, and the model that used as final stage is RF. In addition, the output from the stacking regressor model is the same as a table with four columns, and each column is a prediction of KNN, DCT, XGB, and RF. So, the final RF implements the algorithm process in this table and makes the final output prediction.

The reason behind choosing KXDF as a hybrid model in this work, is because the dataset value is not linearly distributed, though the most non-linear model with high capability to go into further numerical calculation is chosen. KNN and support vector machines can work on linear and non-linear data distributions, and they perform better on non-linear data. The same idea can be applied to DCT and RF models. Moreover, the chosen models are less sensitive to outliers and unnormalized data,

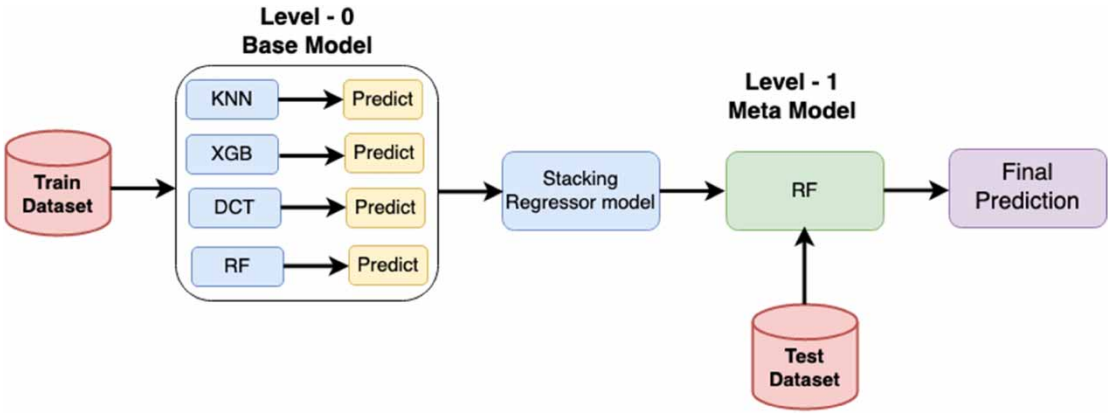


Figure 3 | Structure of the developed KXDF hybrid model.

especially in the case of DCT, XGB, and RF. It makes the step of normalizing data, and the process of removing outliers is not required, which makes the process model slow and time-consuming.

The method utilized had some limitations. For instance, training the dataset took a very long time since four different machine-learning techniques were combined to train the data. KNN was struggling to train the large dataset, but, with the help of other algorithms, it could proceed successfully. XGB was extremely sensitive to the outliers. Furthermore, the large number of trees in the RF model made the algorithm too slow; however, the algorithm is generally fast to train. Finally, the calculation for the final prediction in the decision tree was complex.

The composition of the proposed hybrid KXDF model includes two stages. The first stage is that the network is trained about input–output patterns using a training set. The second stage is testing, in which the performance of the network is assessed when the unknown pattern was not discovered during the training stage. A hybrid model consisting of the integration of four machine learning algorithms, namely KNN, XGB, DT, and LR has been successfully developed for forecasting daily rainfall at Sydney Airport, Australia. Two conventional machine learning algorithms have been implemented in order to compare the accuracy of the proposed hybrid KXDF model. The two conventional machine learning algorithms are RF and artificial neural networks.

The primary model used is the stacking model, which is a type of hybrid model that works on combining different models into two levels: level-0 as the base model and level-1 as the meta-model. On a classification or regression issue, the hybrid model can combine the capabilities of many highly effective models to produce results that are superior to any individual model in the ensemble. Using a hybrid model increases the performance of the proposed model. The main reason is that the model works on predicting the data at a different level with different models, one after another. If one of the models does not perform well or provides a low prediction score, the following model works on it and enhances the score value. The base model and model levels, which fit the training data and whose predictions are created, make up the model's primary structure. The second level is the meta-model, a model that develops the most effective way to combine the predictions of the base models, working on the ultimate decider in cases of accurately forecasting the outcomes.

In the hybrid model, the first algorithm used in the base model list is KNN, and the reason for using it is that it does not require any hyperparameter tuning; the only one that is used is the value of K, which is the number of neighbors, and it should be an odd number. Also, because there is no explicit training step, the prediction is modified as new data is added to the dataset without retraining a new model. After KNN, the XGB model is used as the second level in the hybrid model, performs faster, and can work with a dataset of high dimensionality. The dataset that is used in this study is around three thousand samples. Also, the main hyperparameters that are used with XGB are the type of the kernel, which is linear and works on finding the best estimators in linear solution, and gamma with the value of two, which works on the distance of separators of the data points to predict values of Rainfall with a high score and fewer possible errors. The third level used is the Decision Tree model; this model works on predicting values into n numbers of estimators and depth of tree separation branches. So, increasing the number of separations makes the result of prediction more accurate with minimum error. Also, it performs well even if there is some low data accuracy, which means in this case if the previous model didn't perform well, the DT tuned the

data and gave a better result. The last stage is Random, the last model in the level-0 based model, which gets all the predicted data as input from previous models and predicts the output. The RF model works on arranging the inputs, and it can reduce the overfit by dealing with a regularization technique that can capture the noise data.

An extremely popular supervised machine learning approach called the RF algorithm is utilized to solve classification and regression issues. A forest is made up of numerous different species of trees, and the forest will be more vigorous the more trees there are. Similar to this, the number of trees in an RF algorithm enhances the algorithm's accuracy and ability to solve problems.

The final level of the hybrid algorithm (level-1) is called the meta-model, which works on the final decision for result prediction. This model takes all the previous predictions, which were combined into the base model, and predicts new results. In this level model, a decision tree is used because DT has resulted in a good score from the base model as well, and it can work on giving a more in-depth division of the tree to provide higher results for predictions. In the current study, the data has been divided into two subsets: 60% for training and 40% for testing. The optimized weights and their sensitivity to rainfall were assessed. Some input variables have been removed from the model since they was not correlated to the output.

2.3. Statistical measurements

2.3.1. Root mean square error

Root mean square error (RMSE) is the standard deviation of the residuals (predictive errors). The distance between the data points and the regression line determines the residuals. The RMSE is a measurement of how evenly specific residuals are spread. In other words, it demonstrates how closely the data is clustered along the best-fit line.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

where x_i and y_i are the actual and expected rainfall amounts, n is the mean value of the observed rainfall. The closer the RMSE value is to zero, the higher the accuracy.

2.3.2. Coefficient of determination (R^2)

The coefficient of determination is an important feature of regression analysis (denoted by R^2). This is the percentage of variation in the dependent variable that the independent variable predicts.

$$R^2 = \{(1/N)*\sum[(xi - X)*(yi - Y)/(\sigma x*\sigma y\}^2 \tag{2}$$

2.3.3. Mean absolute error (MAE)

The mean absolute error is the model evaluation metric in regression analysis. The absolute values of each individual prediction error in the test set are added to determine the mean absolute error of the model with regard to the test set. The variation between the example's true value and predicted value demonstrates a growing prediction error. The units of MAE and RMSE are the same as the rainfall variables which are millimeters (mm).

$$MAE = \frac{\sum_{i=1}^{n} ABS(yi - \lambda(xi))}{\mathrm{n}} \tag{3}$$

2.3.4. Normalized root mean square error

Normalized root mean square error (NRMSE) is a metric used to relate the RMSE to the mean of the actual value.

$$\mathrm{NRMSE} = \frac{RMSE}{\bar{O}} \tag{4}$$

where $\bar{O}$ is the mean of the actual value.

3. RESULTS AND DISCUSSION

According to the results from statistical indices, the proposed hybrid KXDF model achieved sufficient accuracy for forecasting daily rainfall in Sydney Airport, Australia. Regarding the training part, RMSE = 0.124, R^2 = 0.999, MAE = 0.007, and NRMSE = 0.04. For the testing part, RMSE = 1.246, R^2 = 0.991, MAE = 0.109, and NRMSE = 0.339. Table 2 shows the performance of the training and testing parts of the proposed hybrid KXDF model. Figure 4 represents the tested and trained performance of the hybrid KXDF model.

According to the training and testing performance of the hybrid KXDF model, it shows that the hybrid model was capable of predicting daily rainfall accurately. It could be mentioned that only in extreme events were very few errors in the testing session. It is a standard condition that the hybrid model could perform better in the training part because the model should consistently outperform the testing set on the training set. This is due to the model being trained on training data rather than testing data. In terms of comparison, the current study has similar results with the previous studies in the same field. Chatterjee *et al.* (2018) developed a novel hybrid model based on Greedy forward selection and k-means algorithms with neural networks for predicting rainfall in India and their proposed hybrid model performed well. Moreover, Kumar *et al.* (2021) developed two hybrid models, namely, BBO-ELM and DNN for rainfall prediction in India and they achieved high performance of their hybrid models in terms of accuracy. In addition, Xiang *et al.* (2018) applied LSTM and LSTM-seq2seq in order to estimate hourly rainfall-runoff in Clear Creek and Upper Wapsipinicon River in Iowa. According to their findings, the LSTM-seq2seq model outperformed other proposed machine learning models, including a single LSTM. Their findings indicate that the LSTMseq2seq model has sufficient predictive capacity to improve forecast accuracy in short-term flood forecasting applications. Therefore, it can be concluded that the integration machine learning model in the current study has been developed and applied for the first time for forecasting daily time-series rainfall as one of the major hydrological parameters. It is completely different from all similar previous studies in the same field. Figure 5 represents the scatter plot for both the training and testing dataset of the hybrid KXDF prediction model.

Based on Fig. 5, the training set of the predicted model is very accurate, and it can be realized that there is a linear relationship between observed and predicted data. However, there are minor errors in the testing set of the predicted model. Therefore, in some extreme events, the model was not capable of predicting it accurately. The dataset has very few outliers, and that is the reason that some of the data points are out of range, and this does not affect the performance of the proposed model; additionally, the size of the data is around 3,000 data points, and only a few are detected as outliers. Also, the chosen models are not sensitive to outliers and unnormalized data, because they have the ability to work on non-linear data. But, overall, the performance of both the training and testing datasets was reliable and acceptable. The results of the currently developed hybrid model in this study are very important for managing the water resources in Australia since predicting rainfall is extremely important for human beings because heavy and irregular rainfall will cause the destruction of many crops, farms and materials of the farmers. So, the results of the currently developed model for forecasting rainfall are crucial in order to give an early warning so the risks of heavy rainfall for humans, animals, and plants can be reduced and managing the water resources will be effective. There are many different parameters for forecasting rainfall including weather conditions such as temperature, humidity, and evaporation. Therefore, in the current study, humidity, temperature, evaporation and rainfall have been utilized for developing KXDF hybrid models. However, the most common model for forecasting weather is the Numerical Weather Prediction (NWP) model. The model that is used in this work is proposed giving careful consideration to time-consuming, central processing unit (CPU) and memory utilization. CPU is utilized to perform the proposed machine learning model. Also, the models are tested on their ability to give high accuracy. In addition, the hybrid model is a model that combines other, different machine learning models. So, the chosen model should be considered based on the dataset type and the distributions of data. Hence, it has been found that the data are more non-linear, which in return those models should choose those that are more likely to fit to those dataset types. Sensitivity analysis was performed, and it showed that the temperature, humidity, and evaporation were very sensitive to the rainfall data for prediction purposes.

Table 2 | The performance of training and testing parts of the hybrid KXDF model

	RMSE	R^2	MAE	NRMSE
Training	0.124	0.999	0.007	0.04
Testing	1.246	0.991	0.109	0.339

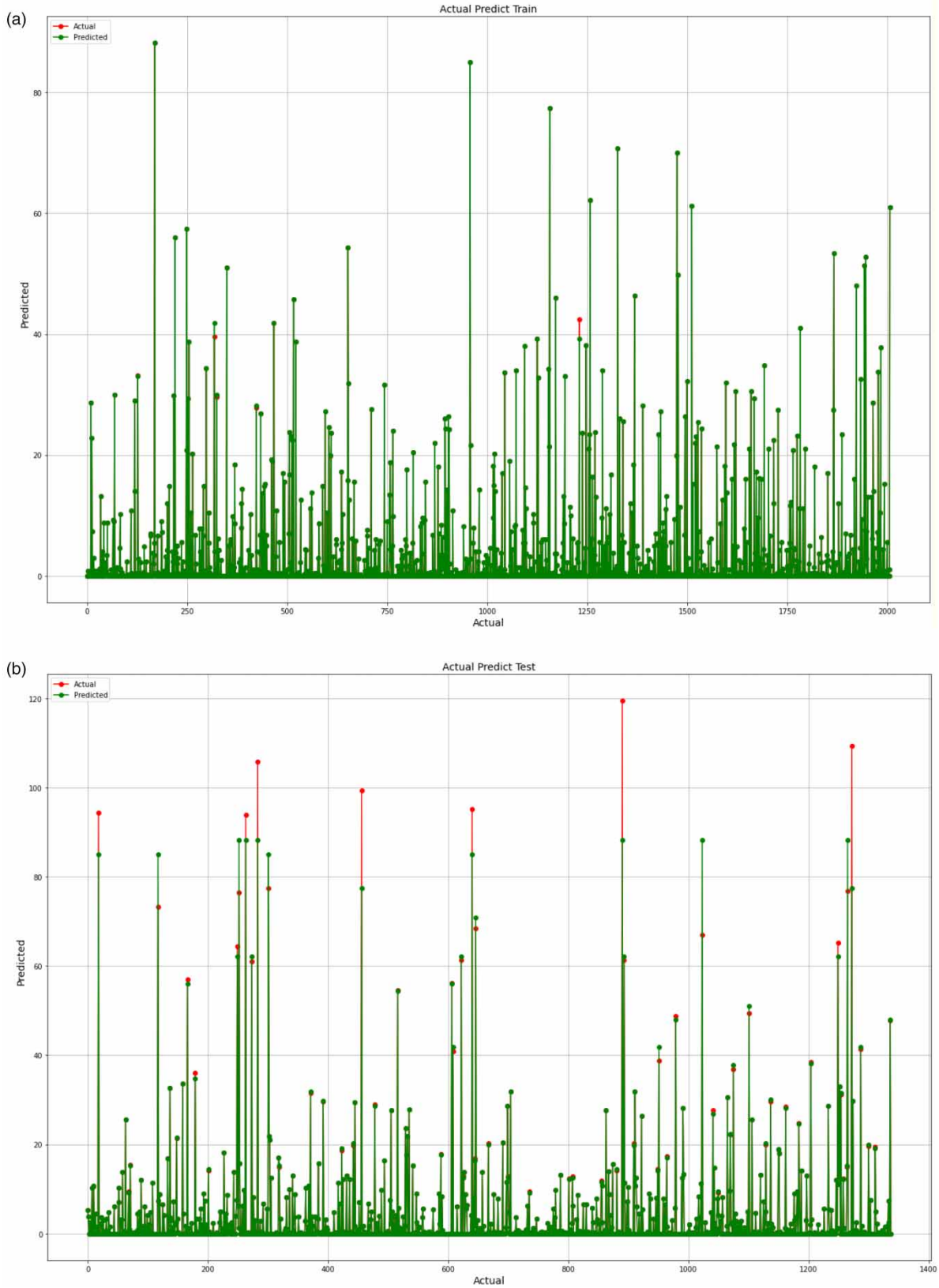


Figure 4 | (a) Actual vs. predicted rainfall (a) training performance and (b) testing performance.

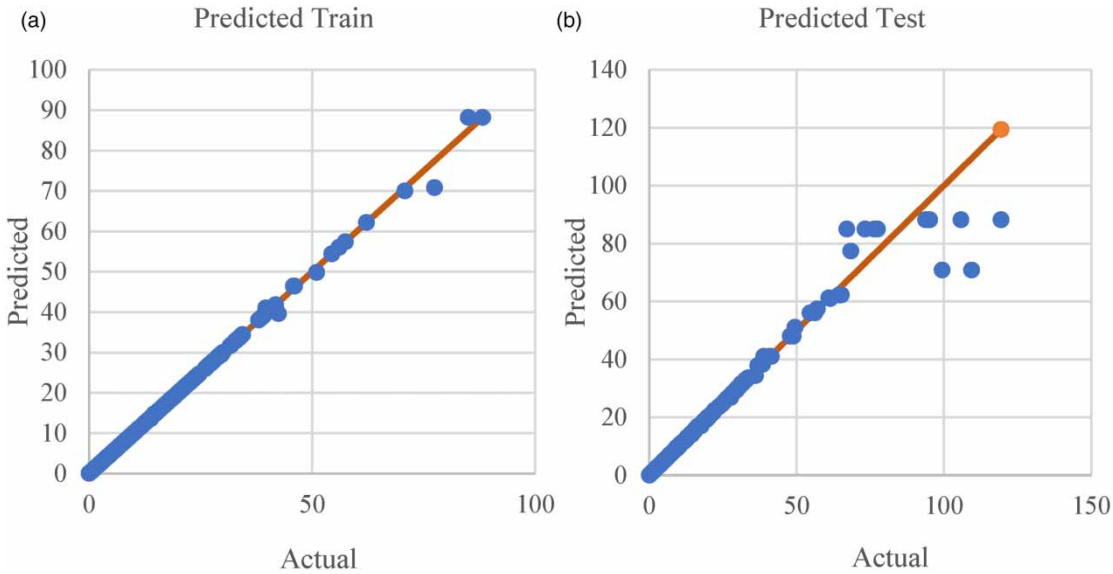


Figure 5 | Scatter plots for (a) predicted training set and (b) predicted testing set.

The results of the currently developed model can help decision-makers to prepare for unwanted fundamentals such as flooding in Sydney. There are two ways of predicting rainfall: short-term and long-term. Most of the time short-term predictions will be more accurate. One of the most complex issues that researchers face is that they cannot build an accurate model for long-term prediction. Heavy precipitation is one of the major issues that scientists struggle with since this problem is related to the economy and social security. The consequence of the heavy rainfall leads to many disasters. For example, drought, water shortages, and floods happen yearly in many countries around the world. Recording rainfall is very important for developed countries such as Australia since they rely on agricultural production, and import many agricultural products to many countries in the world. Therefore, it can be concluded that the accuracy of the current developed model could be considered by the authority of Sydney in order to manage its water resources effectively. Furthermore, the potential source of error is that when a single machine learning model is used for predicting rainfall, the accuracy will be very low. As a consequence of this error, policymakers will wrongly plan for the water resources sector. Finally, even though the result of the current study is achieved in a case study of Sydney, the developed hybrid model could be generalized and applied to other regions around the world since it consists of four different machine learning models.

4. CONCLUSION

Forecasting rainfall is important because it can lead to a range of negative effects, such as agricultural and farm destruction and property damage. For early warnings that can reduce risks to life and property while also better managing agricultural farms, a suitable prediction model is essential. The main objective of this study is to establish and develop a hybrid model based on the integration of four machine learning algorithms (KNN, XGB, DCT, and LR) for forecasting daily rainfall. The proposed hybrid KXDF model was capable of predicting daily rainfall at Sydney Airport accurately. A sensitivity analysis was conducted, and the results indicated that for the purpose of prediction, the temperature, humidity, and evaporation were highly sensitive to the rainfall data. One of the limitations is that it required a very long time to train the dataset because four distinct machine-learning techniques were used in combination. Although KNN was having trouble training the big dataset, it was able to do it with the assistance of other methods. XGB was highly susceptible to the anomalies. In addition, the RF model's high tree count made the process excessively slow; yet the technique trains quickly overall. Ultimately, the decision tree's prediction required a complicated computation. This research has aided the area of water resources engineering by drawing the attention of academics, teachers, and policymakers to forecast models. In future studies, it is suggested that the proposed hybrid KXDF model be used in diverse climate zones and can be generalized and applied to other regions around the world since it consists of four different machine-learning models. Due to the intricate weather pattern, the traditional machine learning model was unable to accurately estimate the rainfall parameter in the humid subtropical

regions. Thus, what makes this work new is that it is the first time a hybrid model combining KNN, XGB, DCT, and RF has been used to forecast daily rainfall. The water sectors could put the created model to practical use to manage water resources in a sustainable manner. Finally, further studies might be needed to use graphics processing units (GPU) for time-series forecasting models for predicting rainfall data in other regions around the world.

ACKNOWLEDGEMENT

The authors would like to thank the Australian Government for providing data through the Bureau of Meteorology.

AUTHOR CONTRIBUTIONS

S.D.L. wrote the original draft, performed the methodology, and did analysis; M.H.M. wrote the review and edited, performed the methodology, and did analysis.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Aggarwal, R., Kaur, S., Dar, M. U. D. & Kuriqi, A. 2023 Uncertainties in climate change scenarios for determining temperature and rainfall patterns in regions with mixed climate conditions. *Acta Sci. Pol. Form. Circumiectus* **22**, 91–106. https://doi.org/10.15576/ASP.FC/2023.22.1.91.

Bellido-Jiménez, J. A., Gualda, J. E. & García-Marín, A. P. 2021 Assessing machine learning models for Gap filling daily rainfall series in a semiarid region of Spain. *Atmosphere (Basel)* **12**. https://doi.org/10.3390/atmos12091158.

Bowyer, D. & Chapman, R. L. 2014 Does privatisation drive innovation? Business model innovation through stakeholder viewpoints: The case of Sydney Airport 10 years post-privatisation. *J. Manage. Organ.* **20**, 365–386. https://doi.org/10.1017/jmo.2014.16.

Chatterjee, S., Datta, B., Sen, S., Dey, N. & Debnath, N. C. 2018 Rainfall Prediction using Hybrid Neural Network approach. In: *2018 2nd International Conference on Recent Advances in Signal Processing, Telecommunications & Computing (SigTelCom)*, 29–31 January 2018. IEEE, New York, pp. 67–72. https://doi.org/10.1109/SIGTELCOM.2018.8325807.

Dash, Y., Mishra, S. K. & Panigrahi, B. K. 2018 Rainfall prediction for the Kerala state of India using arti fi cial intelligence approaches. *Comput. Electr. Eng.* **70**, 66–73. https://doi.org/10.1016/j.compeleceng.2018.06.004.

Eltarabily, M. G., Elshaarawy, M. K., Elkiki, M. & Selim, T. 2023 Modeling surface water and groundwater interactions for seepage losses estimation from unlined and lined canals. *Water Sci.* **37**, 315–328. https://doi.org/10.1080/23570008.2023.2248734.

Gohari, A., Eslamian, S., Mirchi, A., Abedi-Koupaei, J., Massah Bavani, A. & Madani, K. 2013 Water transfer as a solution to water shortage: A fix that can backfire. *J. Hydrol.* **491**, 23–39. https://doi.org/10.1016/j.jhydrol.2013.03.021.

Imteaz, M. A. & Moniruzzaman, M. 2018 Spatial variability of reasonable government rebates for rainwater tank installations: A case study for Sydney. *Resour. Conserv. Recycl.* **133**, 112–119. https://doi.org/10.1016/j.resconrec.2018.02.010.

Koppa, A., Rains, D., Hulsman, P., Poyatos, R. & Miralles, D. G. 2022 A deep learning-based hybrid model of global terrestrial evaporation. *Nat. Commun.* **13**, 1–11. https://doi.org/10.1038/s41467-022-29543-7.

Krishnaraj, A. & Honnasiddaiah, R. 2022 Remote sensing and machine learning based framework for the assessment of spatio-temporal water quality in the Middle Ganga Basin. *Environ. Sci. Pollut. Res.* **29**, 64939–64958. https://doi.org/10.1007/s11356-022-20386-9.

Kumar, R., Prasad, M. & Bishwajit, S. 2021 A comparative assessment of metaheuristic optimized extreme learning machine and deep neural network in multi - step - ahead long - term rainfall prediction for All - Indian regions. *Water Resour. Manage.* 1927–1960. https://doi.org/10.1007/s11269-021-02822-6.

Latif, S. D. 2023 Evaluating deep learning and machine learning algorithms for forecasting daily pan evaporation during COVID-19 pandemic. *Environ. Dev. Sustainability* https://doi.org/10.1007/s10668-023-03469-6.

Latif, S. D., Alyaa Binti Hazrin, N., Hoon Koo, C., Lin Ng, J., Chaplot, B., Feng Huang, Y., El-Shafie, A. & Najah Ahmed, A. 2023 Assessing rainfall prediction models: Exploring the advantages of machine learning and remote sensing approaches. *Alexandria Eng. J.* **82**, 16–25. https://doi.org/10.1016/j.aej.2023.09.060.

Mehraein, M., Mohanavelu, A., Naganna, S. R., Kulls, C. & Kisi, O. 2022 Monthly streamflow prediction by metaheuristic regression approaches considering satellite precipitation data. *Water* **14**, 3636. https://doi.org/10.3390/w14223636.

Mohammadi, B., Safari, M. J. S. & Vazifehkhah, S. 2022 IHACRES, GR4J and MISD-based multi conceptual-machine learning approach for rainfall-runoff modeling. *Sci. Rep.* **12**, 1–21. https://doi.org/10.1038/s41598-022-16215-1.

Praveen, B., Talukdar, S., Shahfahad, Mahato, S., Mondal, J., Sharma, P., Islam, A. R. M. T. & Rahman, A. 2020 Analyzing trend and forecasting of rainfall changes in India using non-parametrical and machine learning approaches. *Sci. Rep.* **10**, 1–21. https://doi.org/10.1038/s41598-020-67228-7.

Raval, M., Sivashanmugam, P., Pham, V., Gohel, H. & Kaushik, A. 2021 Automated predictive analytics tool for rainfall forecasting. *Sci. Rep.* 1–13. https://doi.org/10.1038/s41598-021-95735-8.

Saadi, M., Oudin, L. & Ribstein, P. 2019 Random forest ability in regionalizing hourly hydrological model parameters. *Water (Switzerland)* https://doi.org/10.3390/w11081540.

Song, Y., Liang, J., Lu, J. & Zhao, X. 2017 An efficient instance selection algorithm for k nearest neighbor regression. *Neurocomputing* **251**, 26–34. https://doi.org/10.1016/j.neucom.2017.04.018.

Sowby, R. B. 2023 Making waves: Research to support water and wastewater utilities in the transition to a clean-energy future. *Water Res.* **233**, 119739. https://doi.org/10.1016/j.watres.2023.119739.

Sydney Airport 2019 *Water Quality and Water use [WWW Document]*. URL Available from: https://www.sydneyairport.com.au/corporate/sustainability/environment/water-quality-and-water-use.

Villanueva, C. M., Evlampidou, I., Ibrahim, F., Donat-Vargas, C., Valentin, A., Tugulea, A. M., Echigo, S., Jovanovic, D., Lebedev, A. T., Lemus-Pérez, M., Rodriguez-Susa, M., Luzati, A., de Cássia dos Santos Nery, T., Pastén, P. A., Quiñones, M., Regli, S., Weisman, R., Dong, S., Ha, M., Phattarapattamawong, S., Manasfi, T., Musah, S. I. E., Eng, A., Janák, K., Rush, S. C., Reckhow, D., Krasner, S. W., Vineis, P., Richardson, S. D. & Kogevinas, M. 2023 Global assessment of chemical quality of drinking water: The case of trihalomethanes. *Water Res.* **230**. https://doi.org/10.1016/j.watres.2023.119568.

Xiang, Y., Gou, L., He, L., Xia, S. & Wang, W. 2018 A SVR – ANN combined model based on ensemble EMD for rainfall prediction. *Appl. Soft Comput. J.* **73**, 874–883. https://doi.org/10.1016/j.asoc.2018.09.018.

Xiang, Z., Yan, J. & Demir, I. 2020 Water resources research - 2020 - Xiang – A rainfall-Runoff model with LSTM-Based sequence-to-Sequence learning.pdf. *Water Resour. Res.* https://doi.org/10.1029/2019WR025326.

Yu, X., Wang, Y., Wu, L., Chen, G., Wang, L. & Qin, H. 2020 Comparison of support vector regression and extreme gradient boosting for decomposition-based data-driven 10-day streamflow forecasting. *J. Hydrol.* **582**, 124293. https://doi.org/10.1016/j.jhydrol.2019.124293.

Zhang, P., Cao, W. & Li, W. 2020 Surface and high-altitude combined rainfall forecasting using convolutional neural network. *Peer-to-Peer Networking Appl.* https://doi.org/10.1007/s12083-020-00938-x.

First received 14 January 2024; accepted in revised form 2 June 2024. Available online 11 June 2024