MAUSAM, **74**, 4 (October 2023), 951-962

MAUSAM

DOI : https://doi.org/10.54302/mausam.v74i4.5381
Homepage: https://mausamjournal.imd.gov.in/index.php/MAUSAM

UDC No. 551.588.6 : 551.573

Efficient prediction of evaporation using ensemble feature selection techniques

RAKHEE SHARMA, ARCHANA SINGH* and MAMTA MITTAL**

Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi, India

**ASET, Amity University, Noida, India*

***Delhi Skill and Entrepreneurship University, New Delhi, India*

(*Received 22 February 2022, Accepted 12 June 2023*)

e mail : rakheesharma234@gmail.com

सार — **जल संसाधनों की समयबद्ध योजना और प्रबंधन के लिए, वाष्पीकरण के पूर्वानुमान का सही अनुमान लगाना आवश्यक है, खासकर सूखे की संभावना वाले क्षेत्र जहां वाष्पीकरण सीधे कीटों की आबादी को प्रभावित करता है। तापमान, सापेक्षिक आर्द्रता, सौर विकिरण, वर्षा जैसे मौसम परिवर्ती में परिवर्तन का वाष्पीकरण प्रक्रिया पर बहुत प्रभाव पड़ता है। परिवर्ती का पूर्वानुमान लगाने के लिए, विभिन्न मशीन लर्निंग तकनीकों के साथ-साथ सामूहिक विशेषता के चयन की तकनीकों की जांच की गई।** ICRISAT **से** 1974 **से** 2021 **की अवधि का साप्ताहिक मौसम संबंधी डेटा एकत्र किया गया। इन विकसित मॉडलों की विश्वसनीयता सांख्यिकीय दृष्टिकोणों नामत: मीन एब्सोल्यूट एरर, रूट मीन स्क्वायर एरर, निर्धारण का गुणांक, नैश-सटक्लिफ दक्षता गुणांक और विभिन्न ग्राफिकल सहायता विलमॉट के समझौते सूचकांक पर आधारित थी। परिणाम बताते हैं कि लैस्सो समाश्रयण अन्य सभी मशीन लर्निंग दृष्टिकोणों से बेहतर है और परिणाम हाल के डेटा** (2020-2021) **का उपयोग करके मान्य किए गए हैं। परिणामों की बेहतर समझ के लिए, इन मान्य परिणामों की तुलना स्थापित रैखिक समाश्रयण विधि और कृत्रिम तंत्रिका नेटवर्क से प्राप्त परिणामों से भी की गई। यह भी देखा गया कि लैस्सो समाश्रयण ने रैखिक समाश्रयण (आर**^{2} = 0.871) **और कृत्रिम तंत्रिका नेटवर्क (आर**^{2} = 0.889) **की तुलना में बेहतर प्रदर्शन (आर**^{2} = 0.929) **किया।**

ABSTRACT. For the timely planning and management of water resources, prediction of evaporation is required to be estimated properly, especially in regions that are prone to drought and where evaporation directly affects the pest population. Changes in meteorological variables such as temperature, relative humidity, solar radiation, rainfall have a great impact on the evaporation process. In order to forecast the variable, techniques for selection of ensemble feature along with various machine learning techniques were investigated. Weekly meteorological weather data were collected from the ICRISAT over a period from 1974 to 2021. The reliability of these developed models was based on statistical approaches namely Mean Absolute Error, Root Mean Square Error, Coefficient of Determination, Nash-Sutcliffe Efficiency coefficient, and Willmott's Index of agreement along with several graphical aids. The results indicate that lasso regression outperforms all other machine learning approaches and the results are validated using recent data (2020-2021). For a better understanding of the results, these validated results were also compared with results obtained from the established linear regression method and artificial neural network. It was further found that lasso regression showed an improved performance (R^2 = 0.929) over linear regression (R^2 = 0.871) and artificial neural network (R^2 = 0.889).

Key words – Meteorological parameters, Predictions, Evaporation, Machine Learning, Feature selection.

1. Introduction

Evaporation is the fundamental component of the hydrologic cycle which has direct effect on irrigation and pest population outbreak. Since the last ten years (Molle *et al*., 2012; Bournaris *et al*., 2015; Rizwan *et al*., 2018; El Bilali and Taleb, 2020) there has been a rapid increase in research into how irrigation water is lost and what causes the same. The rate of evaporation is extremely high in semi-arid regions with low rainfall compared to other regions. In developing countries such as India, where some stations are experiencing drought and others are experiencing excessive rain, accurate and timely monitoring of evaporation is required. Evaporation can be predicted using both direct and indirect methods; the indirect methods include stochastic methods such as water budget and aerodynamic approaches, as well as empirical methods. Indirect methods, on the other hand, may

notaccurately predict evaporation. Furthermore, they cannot be applied effectively to different climates. The direct methods, such as the evaporation pan, are based on field observations. Evaporation is a nonlinear and complex process. As a result, developing a physical-based formula for predicting evaporation is difficult. Furthermore, modeling researchers may lack access to the tools required for direct measurement of evaporation based on field data. There are different soft computing methods used to monitor and predict various hydrological variables such as models to predict rainfall (Salih *et al*., 2020), drought (Malik *et al*., 2020; Mokhtarzad *et al*., 2017), surface water quality (Rezaie-Balf *et al*., 2020; Tao *et al*., 2019).

At the ICRISAT (Telangana) agricultural regions and other sub regions, evaporation from water bodies has been a major source of concern. In the semi-arid tropics, tanks are commonly used as water harvesting reservoirs. These are minor components that can be used to store water near dams or ponds during the dry season in order to collect runoff during heavy rain. Only a small amount of water is available for irrigation from these tanks due to significant evaporation and sometimes large seepage losses caused by the enormous water surface area. Furthermore, various studies have been conducted to reduce evaporation losses, but none of them yielded the desired results. Unfortunately, little progress has been made in this aspect of research while, their costs render them unsuitable for long-term usage. It would be extremely beneficial if the evaporation rate for the coming weeks could be forecasted ahead of time, allowing for the appropriate actions to be made. Predictions are especially crucial in places like the semi-arid tropics, where capital resources are limited. Various weather conditions are taken into account while predicting evaporation. The data becomes extremely complex as a result of all of the factors. In high-dimensional datasets, feature selection is critical for preventing over-fitting of prediction/ classification models and reducing computation time and resources.

In recent years, the domain of variables or features utilised in machine learning or pattern recognition applications have grown several times. Several strategies have been developed to solve the challenge of minimising irrelevant and superfluous variables that slow down difficult activities. Feature selection,*i.e*., variable elimination aims at improving predictor performances and reduces the computation time as well. The goal of feature selection is to select a subset of variables from the input that can accurately characterise the data while reducing the influence of noise and irrelevant variables while still delivering high prediction results. A review of such methods has been provided by (Piles *et al*., 2021). The research work states that every technique for selecting features requires a matrix of independent variables for a collection of samples that have distinct outputs or targets. The method then generates a set of preferred features, whose size can be either specified by the user or fine-tuned by the method.There are broadly three categories of feature selection namely filter method, wrapper and embedded methods, respectively. Filtering methods are frequently used as part of the pre-processing process. The selection of features is unaffected by any machine learning algorithms. Rather, features are chosen based on their relationship to the result variable, as determined by various statistical tests. The performance of the predictor is used as the criterion in wrapper methods, *i.e*., the predictor is wrapped around a search algorithm that finds the subset with the best predictor performance. Without splitting the data into training and testing sets, embedded methods include variable selection as part of the training process. In general, the wrapper approaches outperform the filter methods in terms of accuracy, but the former are more computationally expensive.

Evolutionary computing, kernel models, classical neural networks, fuzzy logic, decision trees, deep learning, complementary wavelet-machine learning, and hybrid machine learning are some of the Machine Learning (ML) models that have been created for evaporation modelling. In terms of prediction accuracy, these models and their hybrid combinations have performed admirably (Ghorbani *et al*., 2018;). However, because each environment has its own characteristics of stochasticity and non-stationarity, the majority of these studies primarily focus on examining the generic capabilities of ML models in diverse climates. Recent evaporation prediction research has revealed a considerable improvement in more reliable generalised predictive models. They used feature selection approaches to exclude undesirable features, improving the accuracy of machine learning models for predicting evaporation. Using experimental or virtual data sets, several research from various domains assessed the prediction ability of feature selection approaches integrated with machine learning methods. However, only a few employed extensive evaporation data, and none assessed the stability of the feature selectionalgorithm's performance. In the research paper by (Wu *et al*., 2020) and (Yaseen *et al*., 2020), evaporation has been forecasted using machine learning techniques, although no feature selection strategy has been applied. To forecast evaporation, Moazenzadeh *et al*., (2018) used a hybrid firefly and support vector machine approach, whereas(Mohamadi, Ehteram, and El-Shafie 2020) have employed evolutionary algorithms to predict evaporation in real time, but none of them have validated the data to ensure the reliability of such a model in the future.

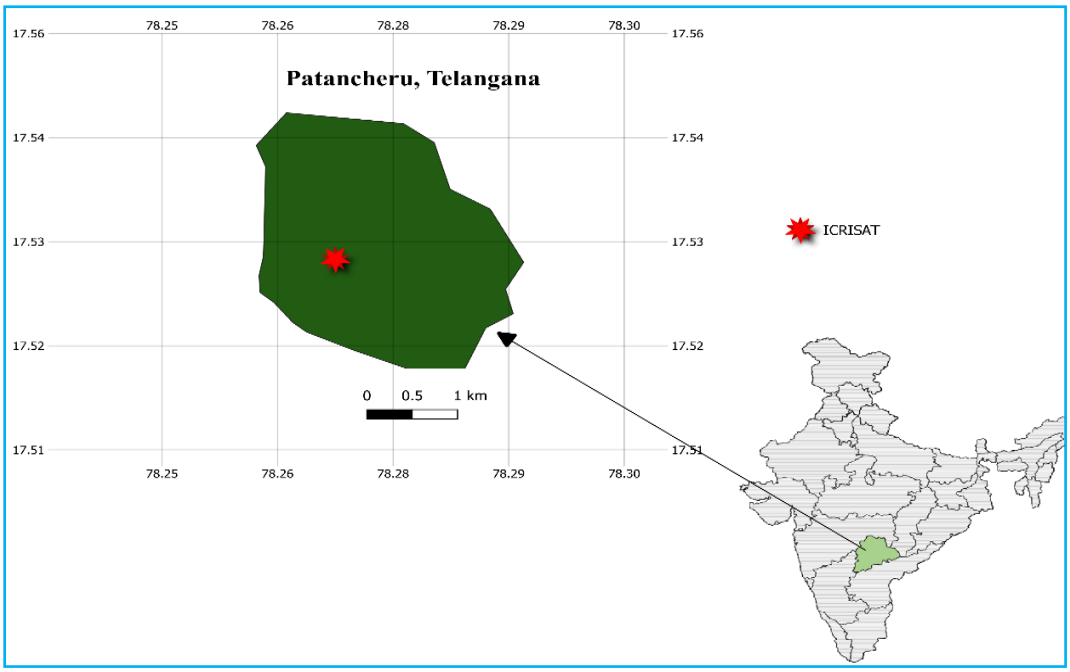

Fig. 1. Location map of the study area

TABLE 1

Statistical description of meteorological parameters

Parameters	Min	Max	Mean	St. Dev	CV	Skewness
Maximum Temperature (°C)	23.32	42.51	32.09	3.87	0.12	0.72
Minimum Temperature (°C)	6.87	28.57	19.61	4.22	0.22	-0.62
Relative Humidity1 (Morning) (%)	23.71	97.71	81.94	13.01	0.16	-1.29
Relative Humidity2 (Evening) (%)	11.00	93.57	44.40	17.71	0.40	0.29
Wind Velocity (km/h)	1.12	32.45	8.51	4.20	0.49	1.22
Solar Radiation (MJ/m^2)	6.34	26.37	17.76	3.41	0.19	-0.05
Bright Sunshine Hours (hrs)	0.15	11.65	7.40	2.57	0.35	-0.74
Rainfall mm	0.00	517.29	17.45	34.39	1.97	4.09
Evaporation (mm)	10.60	121.60	45.12	19.81	0.44	0.98

Based on the existing literature, it can be observed that the deep learning and machine learning methods have some advantages in the field of evaporation prediction, but they also have some drawbacks. First, the model's input parameters were chosen using correlation analysis in a large number of studies, and evaporation was considered the model's output. However, because of the differences in climate at different sites, the correlation between meteorological parameters and evaporation was not the same, which means that the input parameters chosen at one site shall not be applicable at another. To put it the other way, the model that considers a single correlation between meteorological parameters and evaporation across multiple sites has obvious flaws. Second, predicting evaporation on a weekly basis can help the model achieve high accuracy, which is beneficial to the irrigation cycle. However, a large number of studies focus on predicting evaporation on a monthly or annual basis,

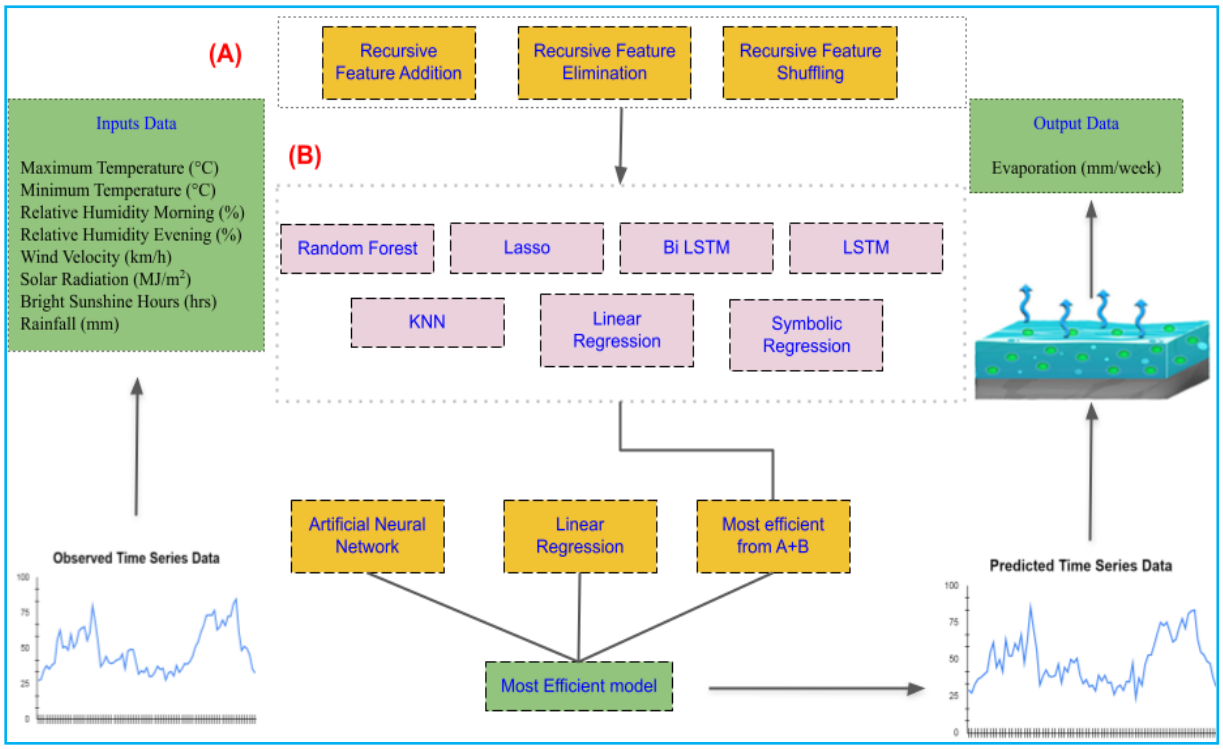

Fig. 2. Schematic structure of the proposed study used for evaporation prediction

whereas predicting evaporation on a weekly basis is less common.

The goal of this research is to look into three feature selection techniques namely recursive feature addition, elimination and shuffling and then the feasibility of five different machine learning models and two deep learning models for modelling weekly evaporation at ICRISAT in India. At a later stage of validation, the performance of the best selected model is compared to that of traditional algorithms such as artificial neural networks and linear regression.

2. Study area and data description

In this study evaporation data of the station International Crops Research Institute for the Semi-AridTropics (ICRISAT), Patancheru region of India was considered for modeling purpose (Fig. 1). ICRISAT is located at 17.53°N 78.27°E, having an average elevation of 522 meters. This research center has a land area of 1,390 ha, of which 800 ha is arable. The station experiences extreme seasonal variation in the perceived humidity. The temperature here averages 25.7 °C | 78.2 °F. Precipitation here is about 846 mm | 33.3 inch per year. Different variables are considered for estimation of weekly evaporation (mm) including weekly values of maximum and minimum temperature (°C), relative humidity morning and evening (%), wind velocity (km/h), solar radiation (MJ/m^2), bright sunshine hours (hrs) and rainfall (mm). The data were collected from 1974-2021 (48 years). The weekly data from 1974-2019 were used for training (70%) and testing set (30%). The data from years 2020 and 2021 were used for validation purpose. It is always practical to take care of the data before proceeding to modeling hence, a rigorous quality check was performed on the collected data. It was observed that only 0.04% of the whole data was missing, Artificial Neural Network (ANN) method was used to fill the missing information and ensure that all the statistical properties of the dataset were preserved.

Table 1 presents a statistical description of the predictors and predict and parameters. It illustrate the statistical parameters of maximum, minimum, mean, standard deviation, coefficient of variation and skewness for the predictors (maximum and minimum temperature, relative humidity morning and evening, wind velocity, solar radiation, bright sunshine hours, rainfall) and the predict and (evaporation).

During the analysis, statistical correlation between the variables were also computed and it is interesting to note that evaporation is highly correlated with maximum

temperature (0.91), relative humidity morning (-0.88), relative humidity evening (-0.65) and solar radiation (0.77). It means that evaporation increase significantly when there is rise in maximum temperature and solar radiation while evaporation decreases with rise in morning and evening relative humidity. It indicated a very week correlation with rainfall (-0.29), wind velocity (0.32), bright sunshine hours (0.46) and minimum temperature (0.47).

3. Proposed predictor framework

In the proposed study, weekly weather variables from 1974-2021 were collected from ICRISAT, Patancheru station, considering evaporation as output and all other weather variables were used as input to the model. The methods utilized in this study are Random Forest, Lasso (Least Absolute Shrinkage and Selection Operator), LSTM (Long Short-Term Memory), Bi-LSTM, KNN (K-Nearest Neighbor), Linear Regression, Symbolic Regression along with three feature selection techniques namely Recursive feature addition, elimination and shuffling. The most efficient algorithm for evaporation prediction s also compared with the traditional established algorithm. Fig. 2 shows the schematic diagram representing the entire modeling process used in this study. The subsection briefly describes the techniques.

3.1. *Feature selection techniques*

In Recursive feature addition, features are ranked according to their importance derived from a machine learning algorithm. Initially, a model is constructed by utilizing a single important feature, and performance metrics for the model are evaluated. Subsequently, the most significant feature is incorporated into the existing model, and the models' efficacy is assessed by considering both the newly added feature and the previously used feature. Next step is to calculate the accuracy and the error rate, if the metric increases by more than an arbitrarily set threshold, then that feature is important and should be kept. Otherwise, the feature can be removed. This is a hybrid approach as it derives the importance from the machine learning algorithm like embedded methods and it builds several machine learning models like wrapper methods. The Recursive feature elimination selection method (Guyon *et al.*, 2002) is essentially a recursive procedure that ranks features based on some metric of relevance. At each iteration, the importance of each feature is assessed, and the one that is not so important is removed. Another option, which was not used in this case, is to eliminate a group of characteristics at a time to speed up the process. Because the relative value of each feature can change significantly when evaluated over a new group of characteristics throughout the stepwise elimination process (especially for strongly linked features), the recursion is required. The final ranking is based on the (inverse) order in which features are eliminated. Only the first n characteristics from this rating are used in the feature selection procedure. A frequent way of feature selection is to shuffle the values of a certain variable at random and see how that permutation impacts the machine learning algorithm's performance metric. To put it the other way, the goal is to permute the values of each characteristic one at a time and see how much the permutation (or shuffling of its values) increases accuracy or affects MSE of the machine learning models. If the variables are significant, a random permutation of their values will reduce any of these measurements considerably. The permutation or shuffling of values, on the other hand, should have a little to no impact on the model performance indicator we are evaluating. The process of recursive feature shuffling goes like this: (a) Build a machine learning model and store its performance metric, (b) Shuffle one feature and make a new prediction using the previous model, (c) Determine the performance of this prediction, (d) Determine the change in the performance of the prediction with the shuffled feature and the original one, (e) Repeat the above steps for each feature. To select features we chose those that induced rise in model performance. We chose regression and classification problem to select feature based on random shuffling.

3.2. *Machine learning algorithm*

Machine learning algorithms are used to obtain better predictive performance by combining the advantages of several different algorithms. In this study, different algorithms are evaluated which are well-equipped to solve non-linear problems through feature selection such as prediction of evaporation.

3.2.1. *Random forest*

The algorithm is used for general purpose classification and regression method. Each individual tree votes for one class for each observation, and the forest forecasts the class with the most votes. The number of randomly picked variables (mtry) to be searched through the optimal split at each node must be specified by the user. The dividing criterion is the Gini index. The tree is grown to its maximum size and is not pruned. As the training set, each tree in the forest has a bootstrap sample from the original data at its root node. For each observation, each individual tree votes for one class and the forest. By combining the predictions given by the T single trees, a prediction for a new observation is obtained. In the case of regression RF, the most basic and typical technique is to average the predictions of

individual trees, whereas aggregate classification trees are normally voted on by majority. This signifies that the new observation is assigned to the T tree's most often predicted class.

3.2.2. *Least Absolute Shrinkage and Selection Operator*

LASSO is an innovative variable selection method for regression presented by Tibshirani, (1996) that minimizes the residual sum of squares subject to the sum of the absolute values of the coefficients being less than a constant, and it is a well-known sparse regression method that regularizes the parameter under sparse assumption. It was first used in the area of least squares problems. Lasso regression is often used in agriculture areas, for example in detection of soil nutrients (Erler *et al*., 2020), crop yield prediction (Panda *et al*., 2010), rainfall prediction (Pham *et al*., 2020) and air quality prediction (Sethi and Mittal 2021).

3.2.3. *K-Nearest Neighbor*

Both classification and regression problems are solved using KNN. It is one of the most basic categorization methods available. It works by determining the number of nearest neighbors, or parameter k. When a new data point needs to be classified, the training data is used to determine its k-nearest neighbors by computing the distance between the input variable and all of the data points in the dataset. This distance is determined using a variety of methods, including Euclidean, Minkowski and Mahalanobis distances. The higher the value of k, the better the classification. The system keeps track of all eligible qualities and categorizes the new ones according to their likeness measure. It uses a tree-like data structure to calculate the distance between points of interest and points in the training data set. The attribute is categorized based on its surroundings. The value of k in a classification algorithm is always a positive integer of nearest neighbor.

3.2.4. *Linear regression*

For exploring any relationship between small sample sizes of dependent and independent variables, statistical approaches such as regression models are the best instruments (Razi and Athappilly 2005). To model evaporation data in terms of local climatological characteristics such as temperature, relative humidity, and wind speed, multiple linear regression techniques can be utilized. Linear regression is applied in different studies such as in prediction of pan evaporation (Malik *et al*., 2020; Wu *et al*., 2020), forecasting of pest infestation (Fuentes *et al*., 2021) and for predicting crop yields (Zhou *et al*.**,** 2017; Hassan *et al*., 2019).

3.2.5. *Symbolic regression*

Symbolic regression techniques use a population of randomly generated candidate solutions to explore a function space that is generally constrained by a preselected set of mathematical operators and operands (variables, constants, etc.). Each candidate solution encoded as a tree effectively functions as a function and is rated on its fitness, or ability to match the observed output. A fitness-weighted selection mechanism and several recombination and variation operators are used to generate these candidate solutions. Symbolic regression has previously been applied to the prediction and identification of rainfall-runoff models (Davidson *et al*., 2003; Hyeon *et al*., 2014; Phukoetphim *et al*., 2016) and, more recently, to the prediction of evaporation (Xu *et al*., 2016). It is demonstrated in the aforesaid literature how to find symbolic equations in a very broad form.

3.3. *Deep Neural Networks*

Deep ANNs are classified as either feed-forward neural networks (FNNs) or feedback-based neural networks (RNNs) based on their structure. In FNNs, information moves directly from the input layer to the output layer, passing through any hidden layers without any cycles or loops. An input layer, hidden layers, and an output layer comprise the basic FNN. RNNs, on the other hand, rely on memory to allow different layers to cycle back and forth to influence previous layers.

3.3.1. *Long short-term memory network*

The LSTM network is a type of recursive neural network that is made up of units such as cell, an input gate, an output gate, and a forget gate. The three gates control the flow of information into and out of the cell, and the cell remembers values over arbitrary time intervals. The LSTM is more capable of dealing with exploding and vanishing gradient problems than traditional RNNs. The structure of LSTMs can be used to tackle this problem (Hochreiter and Schmidhuber, 1997).

3.3.2. *Bi-Long short-term memory network*

Bi LSTM networks are bidirectional LSTM which means the signal propagates forward and backward both in time. The LSTMs network are successfully applied to disease inference, rainfall prediction (Poornima and Pushpalatha 2019), pest forecasting (Wahyono *et al*., 2020) and temperature analysis (Zhang *et al*., 2018).

TABLE 2

Performance evaluation of different evaporation models with feature selection techniques

Models	Recursive Feature Elimination			Recursive Feature Addition			Recursive Feature Shuffling		
	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
(a) Training Phase									
Random Forest	0.194	0.264	0.925	0.186	0.245	0.931	0.275	0.382	0.831
KNN	0.243	0.341	0.879	0.226	0.331	0.882	0.346	0.491	0.754
Lasso	**0.184**	**0.238**	**0.927**	**0.176**	**0.238**	**0.941**	**0.267**	**0.387**	**0.852**
LR	0.214	0.298	0.902	0.193	0.297	0.912	0.267	0.387	0.851
Symbolic Regression	0.755	1.084	0.720	0.776	1.074	0.725	0.796	1.078	0.641
BiLSTM	1.170	1.313	0.631	1.155	1.294	0.634	0.765	0.897	0.615
LSTM	0.733	1.016	0.652	0.780	0.969	0.657	0.779	0.980	0.647
(b) Testing Phase									
Random Forest	0.132	0.248	0.921	0.166	0.240	0.919	0.260	0.351	0.820
KNN	0.180	0.326	0.875	0.206	0.327	0.870	0.331	0.459	0.743
Lasso	**0.121**	**0.223**	**0.923**	**0.156**	**0.233**	**0.929**	**0.252**	**0.355**	**0.841**
LR	0.151	0.283	0.898	0.174	0.285	0.900	0.252	0.355	0.840
Symbolic Regression	0.693	1.068	0.716	0.756	1.061	0.713	0.781	1.047	0.630
BiLSTM	1.108	1.298	0.627	1.135	1.281	0.622	0.750	0.865	0.604
LSTM	0.671	1.001	0.648	0.760	0.957	0.645	0.764	0.948	0.636

4. Assessment criteria

The metrics to evaluate the ability of each feature selection technique and then of each predictor model is observed through mean absolute error (MAE), root mean square error (RMSE), coefficient of determination (R^2), Nash Sutcliffe efficiency coefficient (NSE), Willmott's index of agreement (WI). These metrics are represented as following equations:

$$\mathrm{MAE} = \sum \frac{\left(|x_i - y_i|\right)}{n}$$

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

$$\mathrm{R}^2 = \left[\frac{1}{n} * \frac{\sum \left(x_i - \bar{x}_i\right)\left(y_i - \bar{y}_i\right)}{\sigma_x \sigma_y}\right]$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n} \left(x_i - y_i\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{y}_i\right)^2}, -\infty \le \mathrm{NSE} \le 1$$

$$\mathrm{WI} = 1 - \frac{\sum_{i=1}^{n} \left(x_i - y_i\right)^2}{\sum_{i=1}^{n} \left(\left|x_i - \bar{x}_i\right| + \left|y_i - \bar{y}_i\right|\right)^2}, 0 \le \mathrm{WI} \le 1$$

Where x_i is the observed value, $\bar{x}$ is the mean of observed values, y_i is the predicted value, $\bar{y}$ is the mean of predicted values, σ_x and σ_y are the standard deviations of the observed and predicted data, and n is the number of observations. Apart from these metrics the results are also evaluated through scatter plots and line charts.

5. Results and discussion

The inputs to prediction model are the variables which are considered to be best features for prediction of evaporation. On completion of recursive feature elimination method it was observed that three most important features for predicting evaporation are maximum temperature, relative humidity morning and rainfall. Obtaining the most relevant features from recursive feature addition are maximum temperature, solar radiation and relative humidity morning while recursive feature shuffling gives three features namely maximum temperature, solar radiation and rainfall. In the modeling phase, different algorithm is implemented to generate

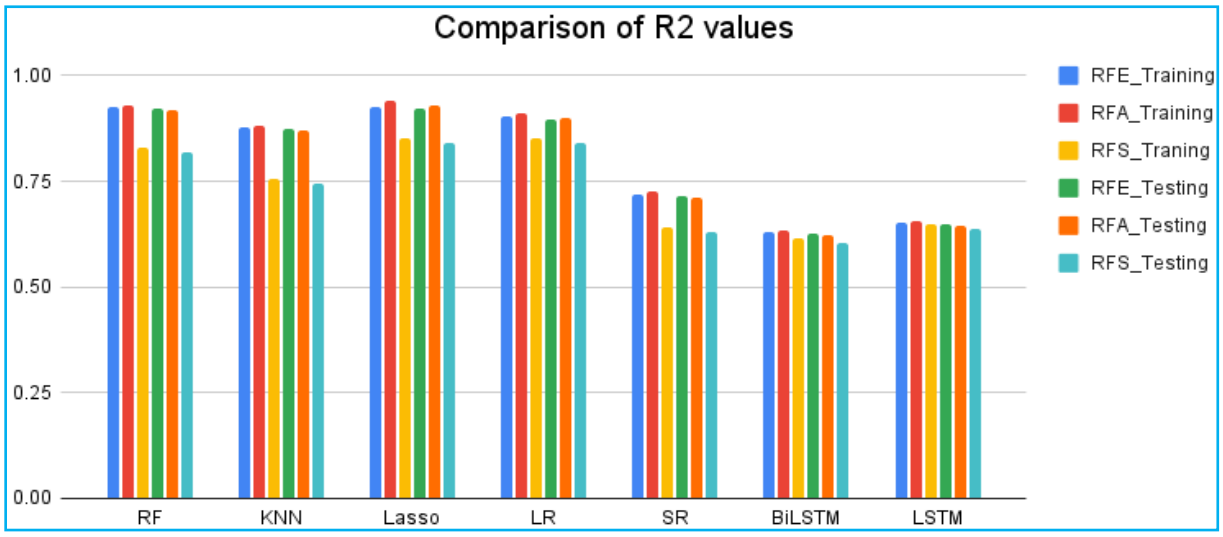

Fig. 3. Comparison of R^2 values

TABLE 3

Performance evaluation of selected model with other traditional models for validation phase

Models	MAE	RMSE	R^2	NSE	WI
Artificial Neural Network	4.245	5.159	0.889	0.879	0.970
Linear Regression	4.927	7.020	0.871	0.781	0.952
Recursive Feature Addition-Lasso	3.022	4.217	0.929	0.918	0.981

prediction-based model using three feature selection techniques. To attain the consistency each data point is normalized between 0 and 1 before providing it to the model for training and testing purpose. The data are then renormalized to its original unit for final comparison between the actual values and predicted one. In all the modeling process 70% of data is considered to be training set and remaining 30% data is considered to be testing set. The recent year data (2020-2021) is then used to validate the selected model. The performance of the model is done using assessment criteria presented in section 4. The results of performance evaluation statistics for different algorithms to predict evaporation is presented in Table 2. The table also compares the three feature selection techniques namely, recursive feature elimination, recursive feature addition and recursive feature shuffling.

It is evident from the table that the two deep learning approaches namely BiLSTM and LSTM provide low R^2 value with high error, while the machine learning approach performed better. There are several possible explanations for this finding. But the prime one is that this considered data size of the study is not that large where deep learning approaches can outperform any other technique. In such data set, traditional machine learning approaches are followed, *e.g.*, random forest, linearregression and lasso. Comparing these top three machine learning techniques for the given result, it can be seen that for training set model, R^2 of lasso is 0.21% more than random forest and 2.77% more than linear regression while adapting recursive feature elimination. If we use recursive feature addition then, lasso R^2 is increased by 1.07% than random forest while it is 3.17% more in comparing with linear regression. Lastly using recursive feature shuffling method, lasso R^2 is 2.52% more than random forest and 0.11% more than linear regression. Recursive feature shuffling method is the only method adapting, where linear regression is providing better results than random forest. However, recursive feature addition along with lasso performed well in all case comparison. Coming to the testing set, for recursive feature elimination it is observed that there is sharp increase of 2.78% and 0.21% in R^2 with lasso model as compared to random forest and linear regression model

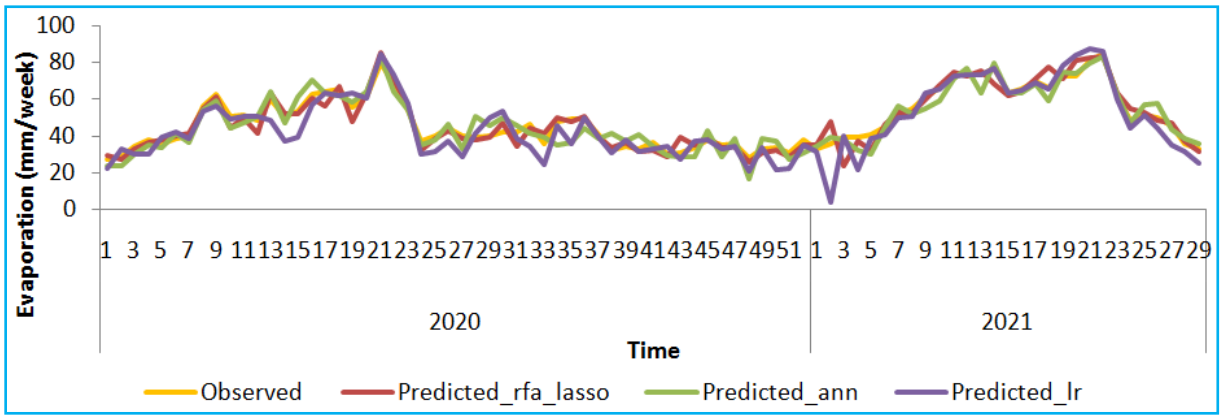

Fig. 4. Time series graph of observed and predicted values obtained through different algorithms

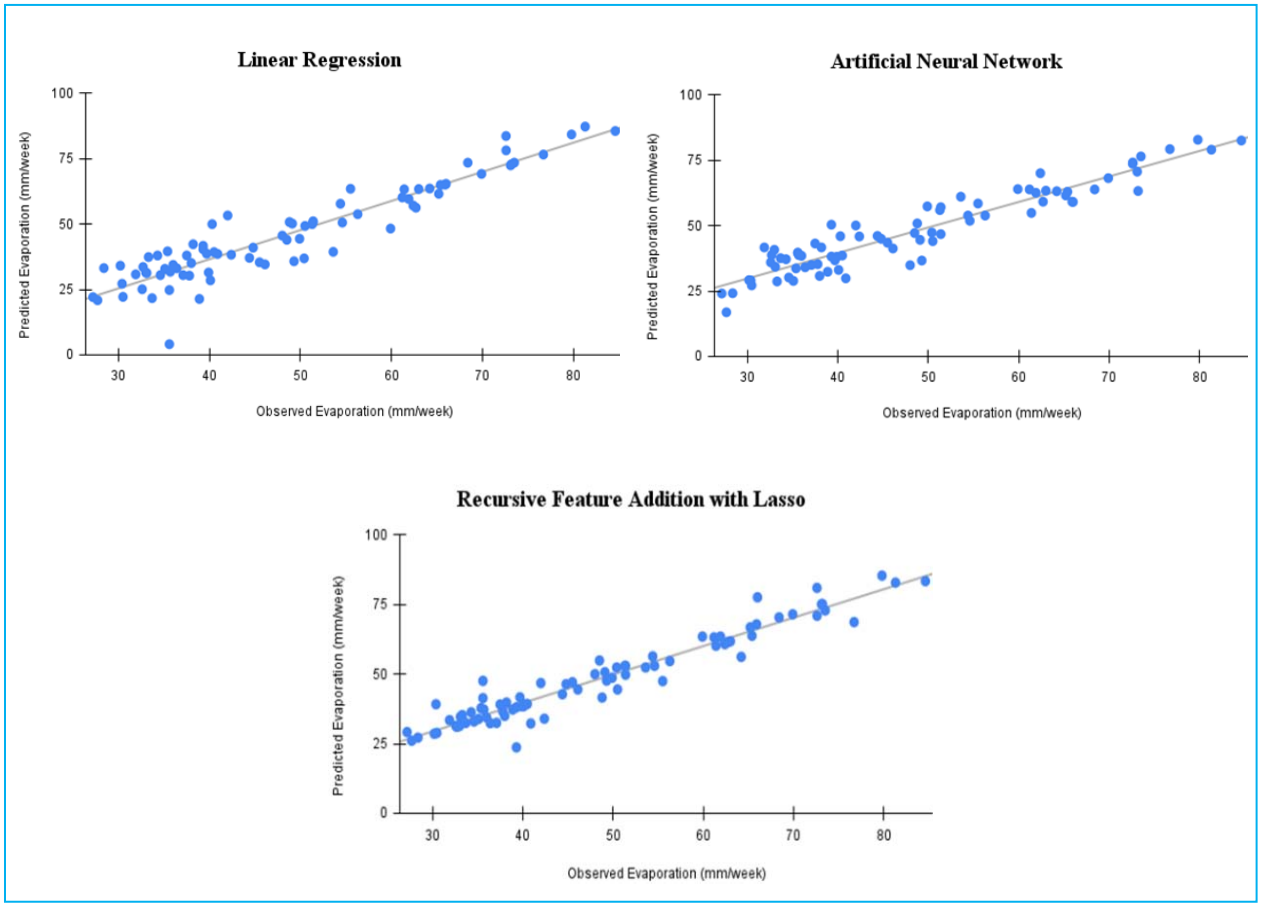

Fig. 5. Scatter plots of algorithms for validated period

respectively. Significant increase in lasso model R^2 is shown while using recursive feature addition having 3.22% compared to random forest and 1.08% compared to linear regression. Lastly, when attempting to select feature with recursive feature shuffling, lasso performed well overall with 0.11% and 2.56% increase in R^2 value

compared with random forest and linear regression respectively. It is noteworthy that recursive feature addition along with lasso algorithm showed good values of R^2 and very less error in both training and testing set, respectively. Fig. 3 gives clear picture of the model evaluation with different feature selection techniques for both training and testing sets.

In order to get a better understanding of the selected prediction model, the recursive feature addition with lasso algorithm is compared with well-established traditional models in prediction such as artificial neural network and linear regression.

The model is validated with the data from years 2020 (1-52 SMW) and 2021 (1-29 SMW). Table 3 highlights the results obtained through these three models. The higher value of R^2, NSE and WI (close to 1) and lower values of different error measures (such as MAE and RMSE) indicates better performance of the model. It is noted from the table that recursive feature addition with lasso algorithm shows an improved performance (R^2 = 0.929, RMSE = 4.217) as compared with artificial neural network (R^2 = 0.889, RMSE = 5.159) and linear regression (R^2 = 0.871, RMSE = 7.020). This validation of results has further strengthened the confidence in timely evaporation prediction with lasso algorithm. Fig. 4 shows the time series graph (validation data) of observed data with prediction data of three different approaches. It is observed that prediction line with recursive feature addition-lasso has close approximation with the observed line. Thus, the model may be applied to get timely prediction of evaporation so that proper remedial measures can be taken in advance.

Fig. 5 presents a scatter plot of estimated evaporation based on linear regression, artificial neural network and selected recursive feature addition-lasso algorithm. The graphs above show that the predicted values are closer to the trend line in recursive feature addition with lasso as compared to the other two traditional algorithms.

6. Conclusion

This study was carried out to analyse feature selection technique and to assess the potentiality of different machine learning algorithms for estimation of weekly evaporation under different meteorological variables. The investigation has led to the following conclusions:

(*i*) Recursive feature addition selected the best set of feature for further development of prediction model in comparison to recursive feature elimination and recursive feature shuffling.

(*ii*) The machine learning algorithms such as random forest, KNN, lasso, linear regression, symbolic regression, BiLSTM, LSTM were studied for model development out of which Lasso outperformed all other algorithms.

(*iii*) In both training and testing set recursive feature addition with lasso algorithm provided better results with least error rate.

(*iv*) For better understanding of validity of the developed prediction model, the prediction results for validation phase was compared with well-established tradition prediction model such as artificial neural network and linear regression.

(*v*) Depending upon the availability of meteorological variables appropriate machine learning model can be adopted for estimating evaporation especially in the stations where measurement of evaporation is not done. Future studies should target other algorithms as well to capture the nonlinear process such as prediction of evaporation.

Disclaimer: The contents and views expressed in this study are the views of the authors and do not necessarily reflect the views of the organizations they belong to.

References

Bournaris, Th., J. Papathanasiou, B. Manos, N. Kazakis and K. Voudouris. 2015, "Support of Irrigation Water Use and Eco-Friendly Decision Process in Agricultural Production Planning", *Operational Research*, **15**, 2, 289-306. doi : 10.1007/s12351-015-0178-9.

Davidson, J. W., Savic, D. A. and Walters, G. A., 2003, "Symbolic and Numerical Regression : Experiments and Applications", *Information Sciences*, **150**, 1, 95-117. doi : 10.1016/S0020-0255(02)00371-7.

El Bilali, Ali and Abdeslam Taleb, 2020, "Prediction of Irrigation Water Quality Parameters Using Machine Learning Models in a Semi-Arid Environment", *Journal of the Saudi Society of Agricultural Sciences*, **19**, 7, 439-51. doi : 10.1016/j.jssas.2020.08.001.

Erler, Alexander, Daniel Riebe, Toralf Beitz, Hans-Gerd Löhmannsröben and Robin Gebbers, 2020, "Soil Nutrient Detection for Precision Agriculture Using Handheld Laser-Induced Breakdown Spectroscopy (LIBS) and Multivariate Regression Methods (PLSR, Lasso and GPR)", *Sensors*, **20**, 2, 418. doi : 10.3390/s20020418.

Fuentes, Sigfredo, Eden Tongson, Ranjith R. Unnithan and Claudia Gonzalez Viejo, 2021, "Early Detection of Aphid Infestation and Insect-Plant Interaction Assessment in Wheat Using a Low-Cost Electronic Nose (E-Nose), Near-Infrared Spectroscopy and Machine Learning Modeling", *Sensors*, **21**, 17, 5948. doi : 10.3390/s21175948.

Ghorbani, Mohammad Ali, Ravinesh C. Deo, Vahid Karimi, Zaher Mundher Yaseen and Ozlem Terzi, 2018, "Implementation of a Hybrid MLP-FFA Model for Water Level Prediction of Lake Egirdir, Turkey", *Stochastic Environmental Research and Risk Assessment*,**32**, 6, 1683-97. doi : 10.1007/s00477-017-1474-0.

Guyon, Isabelle, Jason Weston, Stephen Barnhill and Vladimir Vapnik. 2002, "Gene Selection for Cancer Classification Using Support Vector Machines",*Machine Learning*, **46**, 1, 389-422. doi : 10.1023/A:1012487302797.

Hassan, Muhammad Adeel, Mengjiao Yang, Awais Rasheed, Guijun Yang, Matthew Reynolds, Xianchun Xia, Yonggui Xiao and Zhonghu He, 2019, "A Rapid Monitoring of NDVI across the Wheat Growth Cycle for Grain Yield Prediction Using a Multi-Spectral UAV Platform", *Plant Science*, **282**, 95-103. doi : 10.1016/j.plantsci.2018.10.022.

Hochreiter, Sepp and Jürgen Schmidhuber, 1997, "Long Short-Term Memory", *Neural Computation*, **9**, 8, 1735-80. doi : 10.1162/neco.1997.9.8.1735.

Hyeon, Byeongyong, Yong-Hee Lee and KisungSeo, 2014, "A Prediction Algorithm for a Heavy Rain Newsflash using the Evolutionary Symbolic Regression Technique",*Journal of Institute of Control, Robotics and Systems*, **20**, 7, 730-35. doi : 10.5302/J.ICROS.2014.13.1984.

Malik, Anurag, Anil Kumar, Sinan Q. Salih, Sungwon Kim, Nam Won Kim, Zaher Mundher Yaseen and Vijay P. Singh, 2020, "Drought Index Prediction Using Advance Fuzzy Logic Model: Regional Case Study over Kumaon in India", *PLOS ONE*, **15**, 5, e0233280. doi : 10.1371/journal.pone.0233280.

Malik, Anurag, Priya Rai, Salim Heddam, Ozgur Kisi, Ahmad Sharafati, Sinan Q. Salih, Nadhir Al-Ansari and Zaher Mundher Yaseen, 2020, "Pan Evaporation Estimation in Uttarakhand and Uttar Pradesh States, India: Validity of an Integrative Data Intelligence Model", *Atmosphere*, **11**, 6, 553. doi : 10.3390/ atmos11060553.

Moazenzadeh, Roozbeh, Babak Mohammadi, Shahaboddin Shamshirband and Kwok-wing Chau, 2018, "Coupling a Firefly Algorithm with Support Vector Regression to Predict Evaporation in Northern Iran", *Engineering Applications of Computational Fluid Mechanics*, **12**, 1, 584-97. doi : 10.1080/19942060.2018.1482476.

Mohamadi, S., M. Ehteram and A. El-Shafie, 2020, "Correction to: Accuracy Enhancement for Monthly Evaporation Predicting Model Utilizing Evolutionary Machine Learning Methods", *International Journal of Environmental Science and Technology*. doi : 10.1007/s13762-020-02800-2.

Mokhtarzad, Maryam, Farzad Eskandari, Nima Jamshidi Vanjani and Alireza Arabasadi, 2017, "Drought Forecasting by ANN, ANFIS, and SVM and Comparison of the Models", *Environmental Earth Sciences*, **76**, 21, 729. doi : 10.1007/s12665-017-7064-0.

Molle, B., S. Tomas, M. Hendawi and J. Granier, 2012, "Evaporation and Wind Drift Losses During Sprinkler Irrigation Influenced by Droplet Size Distribution", *Irrigation and Drainage*, **61**, 2, 240-50. doi : 10.1002/ird.648.

Panda, Sudhanshu Sekhar, Daniel P. Ames and Suranjan Panigrahi, 2010, "Application of Vegetation Indices for Agricultural Crop Yield Prediction Using Neural Network Techniques", *Remote Sensing*, **2**, 3, 673-96. doi : 10.3390/rs2030673.

Pham, Binh Thai, Lu Minh Le, Tien-Thinh Le, Kien-Trinh ThiBui, Vuong Minh Le, Hai-Bang Ly and Indra Prakash, 2020, "Development of Advanced Artificial Intelligence Models for Daily Rainfall Prediction", *Atmospheric Research*, **237**, 104845. doi : 10.1016/j.atmosres.2020.104845.

Phukoetphim, Phanida, Asaad Y. Shamseldin and Keith Adams, 2016, "Multimodel Approach Using Neural Networks and Symbolic Regression to Combine the Estimated Discharges of Rainfall-Runoff Models", *Journal of Hydrologic Engineering*, **21**, 8, 04016022. doi : 10.1061/(ASCE)HE.1943-5584.0001332.

Piles, Miriam, Rob Bergsma, Daniel Gianola, Hélène Gilbert and Llibertat Tusell, 2021, "Feature Selection Stability and Accuracy of Prediction Models for Genomic Prediction of Residual Feed Intake in Pigs Using Machine Learning", *Frontiers in Genetics*, **12**, 137. doi : 10.3389/fgene. 2021. 611506.

Poornima, S. and M. Pushpalatha, 2019, "Prediction of Rainfall Using Intensified LSTM Based Recurrent Neural Network with Weighted Linear Units", *Atmosphere*, **10**, 11, 668. doi : 10.3390/atmos10110668.

Razi, Muhammad A. and Kuriakose Athappilly, 2005, "A Comparative Predictive Analysis of Neural Networks (NNs), Nonlinear Regression and Classification and Regression Tree (CART) Models", *Expert Systems with Applications*, **29**, 1, 65-74. doi : 10.1016/j.eswa.2005.01.006.

Rezaie-Balf, Mohammad, Nasrin Fathollahzadeh Attar, Ardashir Mohammadzadeh, Muhammad Ary Murti, Ali Najah Ahmed, Chow Ming Fai, Narjes Nabipour, Sina Alaghmand and Ahmed El-Shafie, 2020, "Physicochemical Parameters Data Assimilation for Efficient Improvement of Water Quality Index Prediction: Comparative Assessment of a Noise Suppression Hybridization Approach", *Journal of Cleaner Production*, **271**, 122576. doi : 10.1016/j.jclepro.2020.122576.

Rizwan, Muhammad, Allah Bakhsh, Xin Li, Lubna Anjum, Kashif Jamal and Shanawar Hamid, 2018, "Evaluation of the Impact of Water Management Technologies on Water Savings in the Lower Chenab Canal Command Area, Indus River Basin", *Water*, **10**, 6, 681. doi : 10.3390/w10060681.

Salih, Abubakr A. M., Marta Baraibar, Kenneth Kemucie Mwangi and Guleid Artan, 2020, "Climate Change and Locust Outbreak in East Africa", *Nature Climate Change*, **10**, 7, 584-85. doi : 10.1038/s41558-020-0835-8.

Sethi, Jasleen Kaur and Mamta Mittal, 2021, "An Efficient Correlation Based Adaptive LASSO Regression Method for Air Quality Index Prediction", *Earth Science Informatics*, **14**, 4, 1777-86. doi : 10.1007/s12145-021-00618-1.

Tao, Hai, Aiman M. Bobaker, Majeed Mattar Ramal, Zaher Mundher Yaseen, Md Shabbir Hossain and Shamsuddin Shahid, 2019, "Determination of Biochemical Oxygen Demand and Dissolved Oxygen for Semi-Arid River Environment: Application of Soft Computing Models", *Environmental Science and Pollution Research*, **26**, 1, 923-37. doi : 10.1007/s11356-018-3663-x.

Tibshirani, Robert, 1996, "Regression Shrinkage and Selection via the Lasso", *Journal of the Royal Statistical Society*, Series B (Methodological) **58**, 1, 267-88.

Wahyono, Teguh, Yaya Heryadi, Haryono Soeparno and Bahtiar Saleh Abbas, 2020, "Enhanced LSTM Multivariate Time Series Forecasting for Crop Pest Attack Prediction".

Wu, Lifeng, Guomin Huang, Junliang Fan, Xin Ma, Hanmi Zhou and Wenzhi Zeng, 2020, "Hybrid Extreme Learning Machine with Meta-Heuristic Algorithms for Monthly Pan Evaporation Prediction", *Computers and Electronics in Agriculture*, **168**, 105115. doi : 10.1016/j.compag.2019.105115.

Xu, Junzeng, Junmei Wang, Qi Wei and Yanhua Wang, 2016, "Symbolic Regression Equations for Calculating Daily Reference Evapotranspiration with the Same Input to Hargreaves-Samani in Arid China", *Water Resources Management*, **30**, 6, 2055-73. doi : 10.1007/s11269-016-1269-y.

Yaseen, Zaher Mundher, Anas Mahmood Al-Juboori, Ufuk Beyaztas, Nadhir Al-Ansari, Kwok-Wing Chau, Chongchong Qi, Mumtaz Ali, Sinan Q. Salih and Shamsuddin Shahid, 2020, "Prediction of Evaporation in Arid and Semi-Arid Regions: A Comparative Study Using Different Machine Learning Models",*Engineering Applications of Computational Fluid Mechanics*, **14**, 1, 70-89. doi : 10.1080/19942060.2019.1680576.

Zhang, Xike, Qiuwen Zhang, Gui Zhang, Zhiping Nie, ZifanGui and Huafei Que, 2018, "A Novel Hybrid Data-Driven Model for Daily Land Surface Temperature Forecasting Using Long Short-Term Memory Neural Network Based on Ensemble Empirical Mode Decomposition", *International Journal of Environmental Research and Public Health*, **15**, 5, 1032. doi : 10.3390/ijerph15051032.

Zhou, Xiang, Hengbiao Zheng, X. Q. Xu, Jiaoyang He, X. K. Ge, Xia Yao, Tao Cheng, Yu Zhu, W. X. Cao and Y. C. Tian, 2017, "Predicting Grain Yield in Rice Using Multi-Temporal Vegetation Indices from UAV-Based Multispectral and Digital Imagery", *ISPRS Journal of Photogrammetry and Remote Sensing*, **130**, 246-55. doi : 10.1016/j.isprsjprs.2017.05.003.