

# Machine learning algorithms for predicting rainfall in India

Sandip Garai<sup>1,2</sup>, Ranjit Kumar Paul<sup>3,\*</sup>, Md. Yeasin<sup>3</sup>, H. S. Roy<sup>3</sup> and A. K. Paul<sup>3</sup>

<sup>1</sup>The Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi 110 012, India

<sup>2</sup>ICAR-Indian Institute of Agricultural Biotechnology, Ranchi 834 003, India

<sup>3</sup>ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110 012, India

**Due to the changing climate and frequent occurrence of extreme events, farmers face significant challenges. Precise rainfall prediction is necessary for proper crop planning. The presence of nonlinearity and chaotic structure in the historical rainfall series distorts the performances of the usual prediction models. In the present study, algorithms based on complete ensemble empirical mode decomposition with adaptive noise combined with stochastic models like autoregressive integrated moving average and generalized autoregressive conditional heteroscedasticity; machine learning techniques like random forest, artificial neural network, support vector regression and kernel ridge regression (KRR) have been proposed for predicting rainfall series. KRR has been considered to combine predicted intrinsic mode functions and residuals generated by various algorithms to capture the volatility in the series. The proposed algorithms have been applied for predicting rainfall in three selected subdivisions of India, namely, Assam and Meghalaya, Konkan and Goa, and Punjab. An empirical comparison of the proposed algorithms with the existing models revealed that the developed models have outperformed the latter.**

**Keywords:** Climate change, crop planning, empirical comparison, machine learning, prediction, rainfall.

ACCURATE and precise forecasts of climatic variables are of utmost importance for an agriculture-based economy like India. Dutta *et al.*<sup>1</sup> have provided probabilistic flood hazard maps using an ensemble method of hydrodynamic model and frequency analysis of water discharge. Lots of work can be found in linear and nonlinear domains of time-series prediction using various conventional econometric techniques and statistical models, including autoregressive integrated moving average (ARIMA)<sup>2</sup>, random walk<sup>3</sup>, generalized autoregressive conditional heteroscedasticity (GARCH)<sup>4</sup>, error correction model<sup>5</sup> and vector autoregressive models<sup>6</sup>. Singh *et al.*<sup>7</sup> performed a copula-based regional and local analysis of meteorological drought in the Netherlands. ARIMA model is known for its performance in time-series prediction in linear dynamics<sup>8</sup>. However, this model cannot work satisfactorily in an environment where

the phenomena of nonstationarity and nonlinearity distort the inherent functional form of the underlying series<sup>9</sup>.

Therefore, several artificial intelligence (AI) techniques have been utilized to tackle such complexities in a dataset. The most widely used techniques are artificial neural networks (ANNs)<sup>10,11</sup> and SVM<sup>3</sup>. ANN has been proven to provide accurate prediction when used singly or combined with other methodologies. With a few exceptions, ANN-based hybrid models outperform the singular ANN models as they help reduce prediction failures in many real-world problems<sup>12</sup>. A combination of various techniques can be an alternative to achieve more efficiency in forecasting<sup>13</sup>.

Wang and Ma<sup>14</sup> proposed two ensemble strategies, i.e. random subspace (RS) and bagging (B), and utilized support vector regression (SVR) as the base model to propose the RSB-SVR model for credit risk assessment. Most neural networks (NNs) do not possess a memory for the individual inputs fed to them, and there is no state-space maintained for them<sup>15</sup>.

Changes in historical rainfall patterns may be due to climatic change, which alters cropping patterns, causes extreme weather behaviours like drought and flood, and is problematic for water resource management authorities<sup>16</sup>. So, rainfall risk mitigation is important for policymakers, hydrologists, and ultimately farmers. To improve water supply management, Sharma<sup>17</sup> proposed seasonal to inter-annual rainfall probabilistic forecast. ARIMA and multiple linear regression (MLR) models have successfully been utilized to predict the rainfall trend<sup>18,19</sup> and seasonal run-off<sup>20</sup>. AI-based algorithms like ANN do not require sophisticated knowledge of the physical and hydrological behaviours of a watershed for forewarning about any extreme rainfall event. They can efficiently handle nonlinear input features<sup>21</sup>. Liyew and Melese<sup>22</sup> studied several atmospheric attributes correlated with rainfall using machine learning (ML) algorithms, like MLR, random forest (RF) and extreme gradient boosting (XGBoost).

It is evident from the literature that forecasting rainfall trends is mainly done through the use of regression and ML-based models. A drastic change in climatic conditions has altered rainfall trends over the years. To cope with the chaotic behaviour and nonlinearity in the historic rainfall pattern, a dire need is to develop data-driven models for forecasting rainfall with better precision. A multi-resolution

\*For correspondence. (e-mail: ranjitstat@gmail.com)

decomposition-based analytical tool is useful in separating large-frequency signals from smaller ones<sup>23</sup> to determine climate-introduced volatility inside rainfall series. Being data-driven and self-adaptive in nature, complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN)-based decomposition is popular in time-series forecasting<sup>24</sup>.

The main objective of the present study is to propose a novel CEEMDAN-based ML algorithm, including ANN, SVR, kernel ridge regression (KRR) and RF, to predict rainfall in three subdivisions of India, namely Assam and Meghalaya (ASMEG), Konkan and Goa (KNGOA), and Punjab (PUNJB) during the period 1871–2016. RF, which is considered a rigorous bootstrapping aggregation framework, can efficiently be implemented in the domain of time-series forecasting as a robust methodology. KRR can be used as an aggregation method for the forecasted results obtained from the decomposed series by the RF-based data intelligent model.

## Theoretical background

### *Empirical model decomposition and improvements*

Based on the noise occurring due to the change in climatic conditions and human interventions over time, the decomposition of rainfall time series provides better feature extraction for statistical modelling and AI techniques. Empirical mode decomposition (EMD) does not use any basis function, unlike wavelet-based decomposition, and is fully dependent on the timescale properties of the noisy time series for decomposition. The relevance of its use is in the transformation of nonstationary and nonlinear signals into stationary and linear ones. However, the EMD proposed by Huang *et al.*<sup>25</sup> has the problem of intrinsic mode mixing due to the intermittent signals. CEEMDAN, proposed by Torres *et al.*<sup>26</sup> is an improvement in the succession of EMD-based decomposition after ensemble empirical mode decomposition (EEMD)<sup>27</sup> and complementary EEMD (CEEMD)<sup>28</sup>. CEEMDAN-based decomposition solves the problem of mode mixing.

EMD decomposes a signal  $x(t)$  into several intrinsic mode functions (IMFs). First, cubic spline interpolation is applied to draw the upper and lower envelopes by joining consecutive maximas and minimas (extremas) respectively. The mean envelopes  $M(t)$  is the average of each of the upper and lower envelopes. If this mean envelope is subtracted from the original signal, one can get the intermediary signal as  $h^1(t)$ . A decomposed signal is considered an IMF if it fulfils the following two conditions: (i) During the whole range of time-series signals, the number of local extremas and zero-crossing points must be equal, or the difference may be at most one. (ii) The local mean at a certain point, i.e. the average of the upper and lower envelopes, must be zero.

To mitigate the mode-mixing effect induced by the EMD process, Wu and Huang<sup>27</sup> proposed the addition of normal white noise (WN) with the original signal before decomposition to make it smoothly distributed at the extreme points throughout the band and introduced the EEMD process. However, EEMD causes another problem of incompleteness in the original signal. Due to the addition of noise to the original signal, error is generated in the reconstruction process, which is the difference between the original and reconstructed signals. CEEMDAN was introduced by Torres *et al.*<sup>26</sup> by adding finite adaptive WN based on the EEMD process to overcome the problems of incompleteness in the reconstruction process.

First, a standard normal WN  $[v^i(t)]$  is added to  $x(t)$ . The signal at the  $i$ th iteration of the EMD process is computed as

$$x^i(t) = x(t) + v^i(t), \quad i = 1, 2, \dots, I,$$

The intermittent IMF, i.e.  $\text{IMF}_1^i$  is obtained by EMD at the  $i$ th iteration. Therefore, the first IMF is calculated as the average of all the intermittent IMFs obtained in  $I$  number of iterations as

$$\text{IMF}_1 = \frac{1}{I} \sum_{i=1}^I \text{IMF}_1^i,$$

and the residual is  $r_1 = x(t) - \text{IMF}_1$ . This process is continued till we end up with a signal ( $r_J$ ) which is a monotonous function and cannot be further decomposed by the EMD method. Then, the original signal can be reconstructed as

$$x(t) = \sum_{j=1}^J \text{IMF}_j + r_J.$$

### *Random forest*

Random forest is a decision tree (DT)-based ML approach. It is based on the popular bootstrapping and bagging procedure<sup>29,30</sup>. RF randomly chooses a bagging method to identify and adopt a feature. A node is forked by choosing the most important and dominating features or predictors. This bagging helps improve the result without overfitting the model<sup>31</sup>. Moore *et al.*<sup>32</sup> have utilized RF in hydrological and environmental management applications. Bootstrapping ensembles ( $n_{\text{trees}}$ ) are generated using the input predictor variables, where  $n$  indicates the number of trees generated in the bootstrapping process. Maximum predictors split is chosen by defining a random input variable sample represented as  $m_{\text{tree}}$ . All the predictions from the bootstrapping ensembles are combined (bagged) to get the forecast of the response variable itself.

### Kernel ridge regression

Ridge regression was first introduced by Hoerl and Kennard<sup>33</sup> to combat the multicollinearity problem in the MLR models. KRR is a combination of kernel trick with ridge regression. A small bias ( $0 < k \leq 0.3$ ) is added with the diagonal elements of the correlation matrix of the predictors. As the diagonal elements of this matrix are one, they may be considered a ridge. The graphical method ridge trace is used to find the appropriate value of  $k$  for which the mean square error (MSE) of the ridge estimators becomes less than the ordinary least square (OLS) estimators. In the ridge trace graph, OLS estimates of the regression coefficients are plotted along the vertical axis against the  $k$  values on the horizontal axis (Figure 1).  $k$  values vary from 0 and increase gradually. With the increase of  $k$ , regression coefficients vary drastically at first and then get stabilized around the  $X$  axes. The smaller the bias, the better the estimate with stabilized regression coefficient estimates.

### Proposed algorithm

Figure 2 presents a flowchart of the proposed approach to predict time-series data. The original series is modelled using ARIMA and GARCH models, and the corresponding prediction is obtained. The noise obtained from fitting the above models is again fitted by ML techniques, like ANN, SVR and RF, and the predicted values are computed. Different ML techniques have been used for modelling the CEEMDAN-decomposed subseries. Predictions obtained for every decomposed series from the ML models are combined together to get final predictions. KRR is used to formulate a hybrid CEEMDAN\_RF\_KRR model utilizing the predictions of the subseries obtained from the RF method.

The steps of the proposed algorithm are summarized below:

Step 1: The underlying series is predicted using stochastic models like ARIMA and GARCH, and ML techniques like ANN, SVR and RF.

Step 2: CEEMDAN decomposition is carried out on the original series to compute IMFs and residual series.

Step 3: ANN, SVR and RF are applied on individual IMFs, and residual series are obtained in step 2 to result in CEEMDAN\_ANN, CEEMDAN\_SVR and CEEMDAN\_RF.

Step 4: The predicted series obtained through RF applied on each of the IMFs and residuals are considered as input in KRR to result in the hybrid CEEMDAN\_RF\_KRR model.

Step 5: The prediction accuracy is compared empirically for each of the above-mentioned algorithms.

Table 1 shows a comprehensive set of models utilized to predict the series under consideration. KRR with linear, polynomial, radial basis function and sigmoid kernel has been represented by KRR\_Lk, KRR\_Pk, KRR\_Rk and

KRR\_Sk respectively. The CEEMDAN-ML models depicted in Table 1 are CEEMDAN\_ANN, CEEMDAN\_SVR and CEEMDAN\_RF to represent CEEMDAN decomposition-based ANN, SVR and RF models respectively. KRR works as a powerful linkage in the CEEMDAN-based KRR-RF hybrid models, i.e. CEEMDAN\_RF\_KRR\_Lk, CEEMDAN\_RF\_KRR\_Pk, CEEMDAN\_RF\_KRR\_Rk and CEEMDAN\_RF\_KRR\_Sk.

### Empirical analysis

Annual rainfall data of the three subdivisions of India, ASMEG, KNGOA and PUNJB were obtained from the Indian Institute of Tropical Meteorology (IITM), Pune (<https://www.tropmet.res.in>) for the period 1871 to 2016. These subdivisions represent northeast, southwest and northwest India respectively (Figure 3).

### Data description

Descriptive statistics are presented in Table 2 for the historical annual rainfall data of the three subdivisions. The ASMEG subdivision, in particular, receives rainfall of about 132.3 cm. The data are positively skewed and leptokurtic. The variation in the dataset represented through coefficient of variation (CV) depicts an 11% disparity compared to the average rainfall. The KNGOA subdivision received a minimum rainfall of 130.30 cm in 1899 and a maximum rainfall of 397.50 cm in 1878. The average rainfall in this subdivision over a period of 146 years was 258.20 cm, with a standard deviation of 50.10 cm. The rainfall is negatively skewed and leptokurtic in nature. Among the three subdivisions discussed here, minimum rainfall was received in PUNJB (23.32 cm). The maximum rainfall received in the PUNJB subdivision was also lower than the minimum rainfall received in the other two

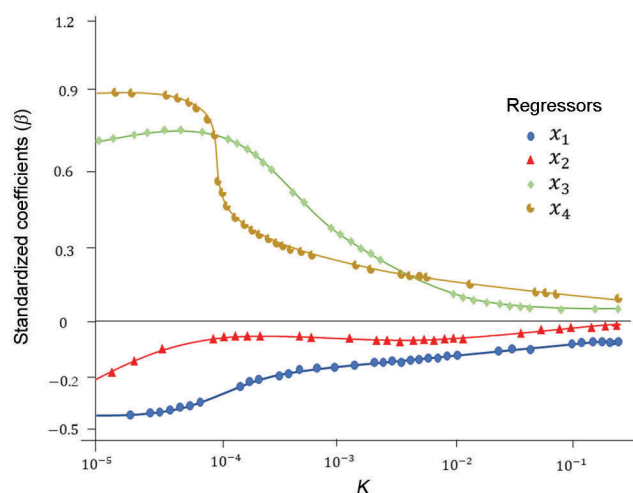


Figure 1. Ridge trace.

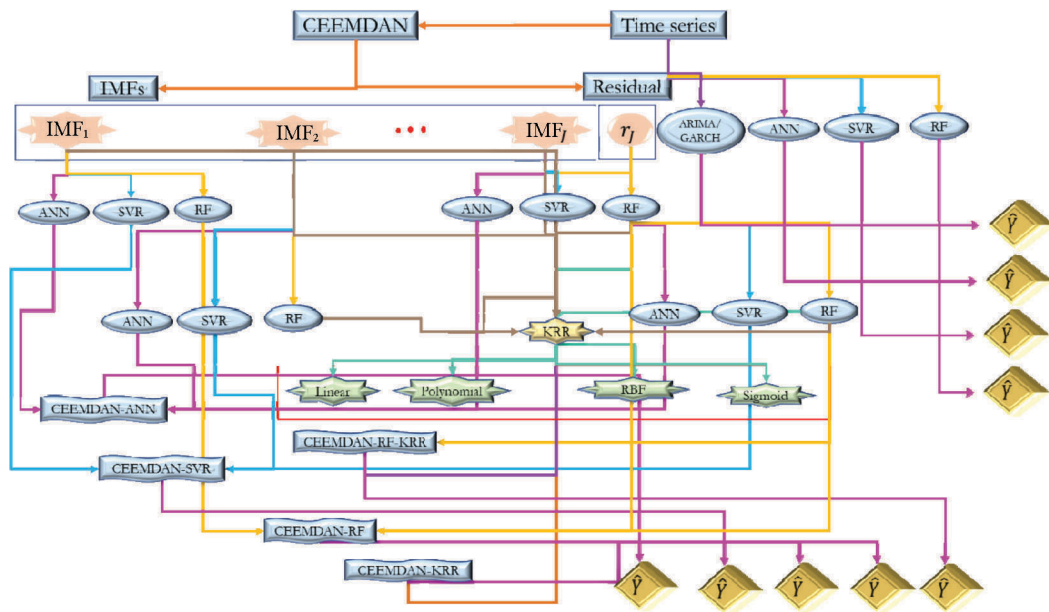


Figure 2. Proposed CEEMDAN-ML algorithm.

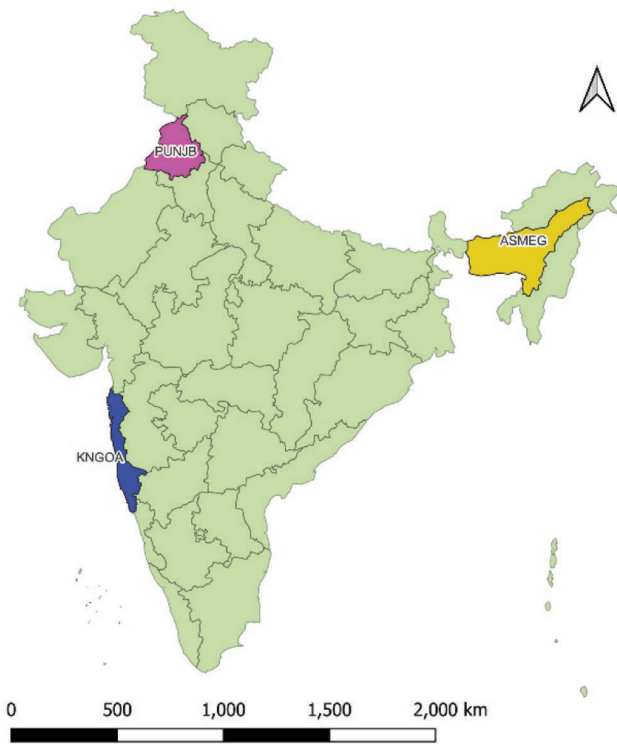


Figure 3. Rainfall subdivisions selected for analysis.

subdivisions. However, the variation in rainfall was the highest in PUNJB among all other subdivisions, with a CV of 28.09%. The skewness and kurtosis were also the highest in rainfall patterns in the PUNJB subdivision. It can be seen that the historical annual rainfall in PUNJB is highly positively skewed and leptokurtic in nature.

Table 1. Set of models used for prediction

Stochastic model	ML	CEEMDAN-ML
ARIMA	ANN	CEEMDAN_ANN
GARCH	SVR	CEEMDAN_SVR
	RF	CEEMDAN_RF
	KRR_Lk	CEEMDAN_RF_KRR_Lk
	KRR_Pk	CEEMDAN_RF_KRR_Pk
	KRR_Rk	CEEMDAN_RF_KRR_Rk
	KRR_Sk	CEEMDAN_RF_KRR_Sk

Table 2. Description of data

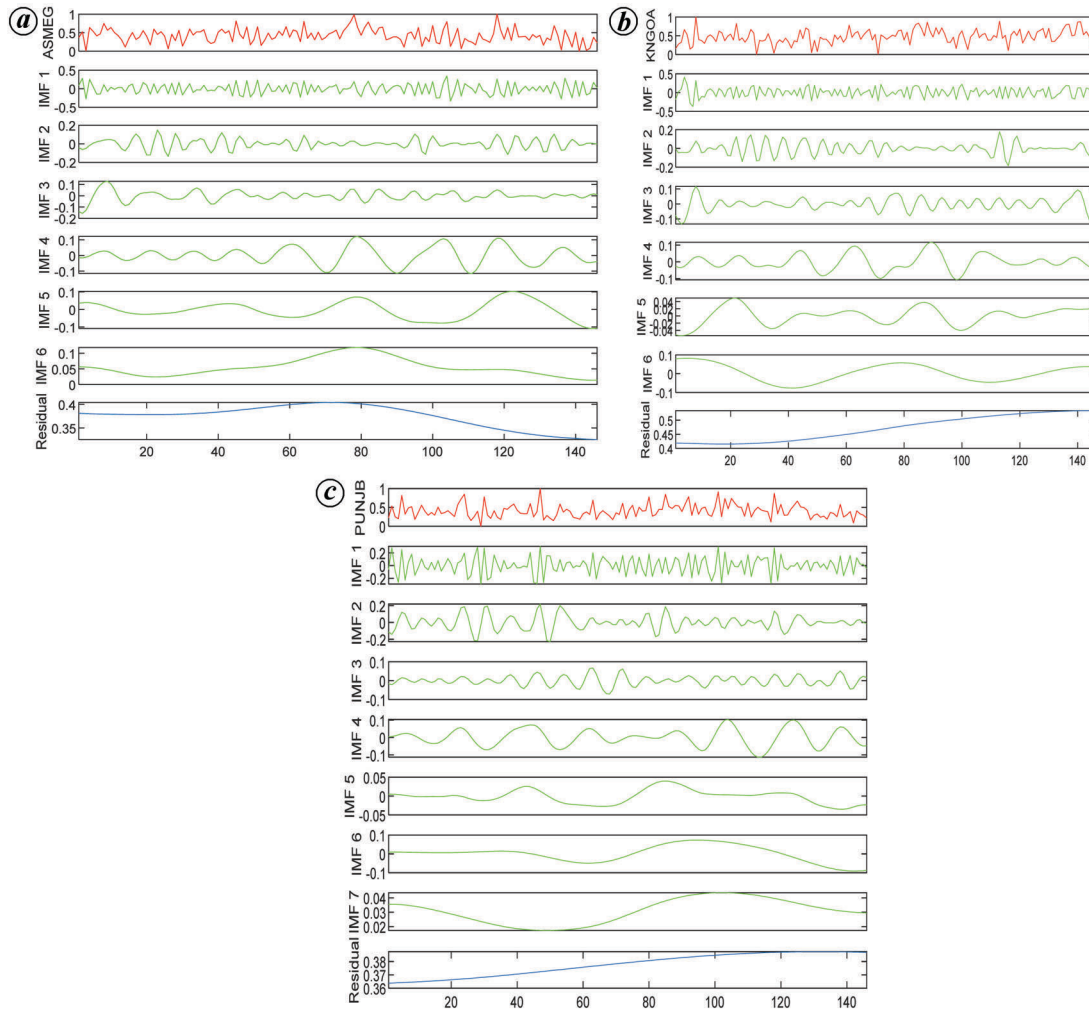
Statistics	ASMEG	KNGOA	PUNJB
Minimum (cm)	178.0	130.27	23.32
Maximum (cm)	310.3	397.50	119.55
Mean (cm)	234.0	258.20	62.38
SD (cm)	25.77	50.10	17.52
CV (%)	11.01	19.40	28.09
Skewness	0.18	-0.06	0.80
Kurtosis	0.02	0.22	0.51
Shapiro-Wilk	0.991	0.990	0.955***

\*\*\*Indicates significant at 1% level of significance.

The Shapiro-Wilk test was conducted to check whether the series under consideration were normal or not. The results in Table 2 confirm that historical rainfall data of ASMEG and KNGOA subdivisions are normally distributed, but the rainfall in PUNJB subdivision is non-normal in nature. The stationarity of the rainfall series has been ensured by means of the augmented Dickey-Fuller (ADF) test.

Data decomposition

MATLAB and R software were used for data analysis. Figure 4 a-c represents the CEEMDAN decomposition of



**Figure 4.** Original and CEEMDAN-decomposed annual rainfall (1871–2016) series of (a) ASMEG, (b) KNGOA and (c) PUNJB.

the subdivisional rainfall. Six IMFs and one residual were generated by the CEEMDAN process for rainfall data in the subdivisions of ASMEG and KNGOA, whereas for the PUNJB subdivision, seven IMFs and one residual were produced.

### Performance measure

To empirically compare the performance of the proposed algorithm with that of the existing methods, a number of statistical indicators as discussed below were used.

Root mean square error

$$RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^N e_t^2}.$$

Relative root mean square error

$$RRMSE = 100 * \frac{\sqrt{\frac{1}{N} \sum_{t=1}^N e_t^2}}{\bar{Y}}.$$

Mean absolute error

$$MAE = \frac{1}{N} \sum_{t=1}^N |e_t|.$$

Mean absolute percentage error

$$MAPE = \left( \frac{1}{N} \sum_{t=1}^N \frac{|e_t|}{y_t} \right) * 100\%.$$

Here  $e_t$  and  $\bar{Y}$  are the residuals and the mean of actual series respectively.  $N$  denotes the number of observations used for validation of the models.

## Results and discussion

The rainfall series was split into training and testing sets in the ratio 90 : 10. The training sets were fitted with various stochastic and AI-based models. In the CEEMDAN-RF-KRR hybrid approach, the decomposed series were first predicted using RF, and then each predicted series was transferred to the KRR model to predict the original series. The number of trees was set as 500 for training the RF for the case, and accordingly, the number of split predictors was determined after the completion of the training phase. ANN models have been trained with a learning rate 0.4 and one hidden layer with five neurons. MSE was used as the loss measure for each of them.

The results of model validation for both the stochastic models and ML techniques, including the decomposition-based ML methods, are provided for ASMEG, KNGOA and PUNJB in Tables 3–5 respectively. It may be seen that KRR\_Pk outperforms the other stochastic, ML or de-

composition-based ML models to predict rainfall in the PUNJB subdivision. CEEMDAN-based decomposition also enriches prediction performance of ANN and RF models. The CEEMDAN\_ANN and CEEMDAN\_SVR methods provide the best prediction of rainfall in the ASMEG subdivision. CEEMDAN\_RF is the best-performing model for predicting rainfall in the KNGOA subdivision. CEEMDAN\_ANN and CEEMDAN\_SVR models can also attenuate the prediction of annual rainfall in the KNGOA subdivision.

MAPE values obtained from CEEMDAN-ANN and CEEMDAN-RF models were less than 10%, indicating better prediction by this decomposition-based ML algorithm for the rainfall dataset of the ASMEG subdivision.

SVR performed well in predicting rainfall in the PUNJB subdivision, which is characterized by high volatility. Moreover, based on the MAPE value, CEEMDAN\_SVR performed better for PUNJB. KRR methods also performed well in this dataset, having a non-normal distribution. Overall, in terms of MAPE, it can be mentioned that the gain in prediction accuracy in the best-fitted model over the usual ARIMA model was more than 50%, 30% and 25% respectively, in the ASMEG, KNGOA and PUNJB subdivisions. Figure 5 depicts the actual versus predicted rainfall using best-fitted models for the three subdivisions.

## Conclusion

In this study, several stochastic and AI-based models have been used to predict rainfall in three subdivisions of India representing the northeast, southwest and northwest regions respectively. Novel CEEMDAN decomposition-based hybrid AI models have also been introduced to attain greater efficiency in predicting rainfall. The residuals have been used to validate the performance of various models. In total, 16 models have been utilized to predict the rainfall series. Various validation measures, namely RMSE, RRMSE,

**Table 3.** Validation results of all the models in ASMEG

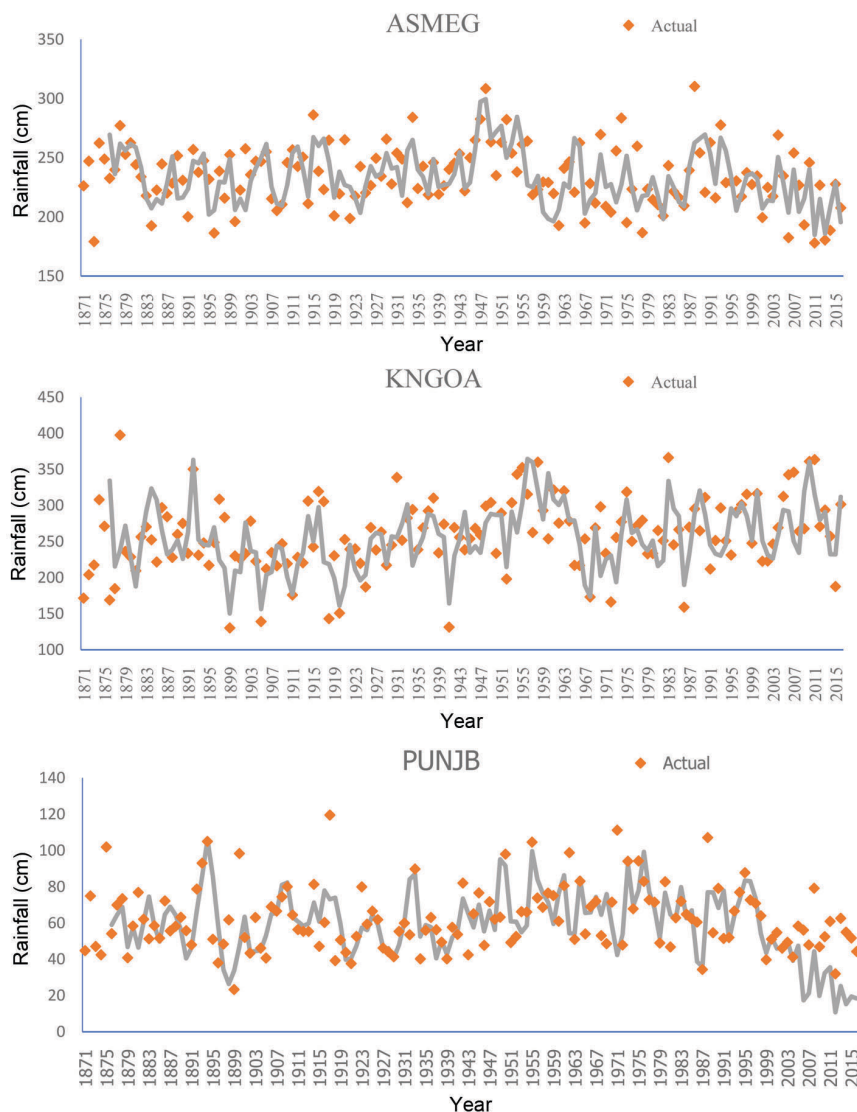
Model	RMSE	RRMSE	MAE	MAPE
ARIMA	36.83	17.52	31.00	13.13
GARCH	36.85	17.53	31.02	13.13
ANN	30.73	14.62	26.66	12.02
CEEMDAN_ANN	18.47	8.79	15.34	7.27
SVR	35.22	16.76	26.93	11.43
CEEMDAN_SVR	14.90	7.09	13.01	6.22
RF	35.26	16.78	29.11	12.42
CEEMDAN_RF	21.21	10.09	18.71	8.69
KRR_Lk	52.05	24.77	45.10	17.64
KRR_Pk	53.81	25.60	46.61	18.10
KRR_Rk	49.93	23.75	42.62	16.86
KRR_Sk	49.55	23.57	42.54	16.86
CEEMDAN_KRR_RF_Lk	39.65	18.86	33.38	13.91
CEEMDAN_KRR_RF_Pk	51.65	24.57	44.57	17.50
CEEMDAN_KRR_RF_Rk	50.34	23.95	43.19	17.06
CEEMDAN_KRR_RF_Sk	48.91	23.27	41.82	16.63

**Table 4.** Validation results of all the models in KNGOA

Model	RMSE	RRMSE	MAE	MAPE
ARIMA	59.29	20.03	45.92	17.22
GARCH	64.32	21.73	51.52	20.02
ANN	61.76	20.86	48.06	18.36
CEEMDAN_ANN	43.99	14.86	35.06	13.46
SVR	66.10	22.33	52.68	20.84
CEEMDAN_SVR	43.15	14.58	33.48	12.49
RF	62.68	21.17	49.52	18.59
CEEMDAN_RF	40.21	13.58	33.17	11.43
KRR_Lk	58.54	19.78	51.81	18.51
KRR_Pk	49.12	16.59	44.20	15.33
KRR_Rk	51.91	17.54	42.86	15.23
KRR_Sk	54.32	18.35	44.56	15.90
CEEMDAN_KRR_RF_Lk	65.20	22.02	52.60	20.57
CEEMDAN_KRR_RF_Pk	52.06	17.59	42.82	15.01
CEEMDAN_KRR_RF_Rk	53.14	17.95	43.37	15.37
CEEMDAN_KRR_RF_Sk	54.32	18.35	43.82	15.72

**Table 5.** Validation results of all the models in PUNJB

Model	RMSE	RRMSE	MAE	MAPE
ARIMA	15.10	28.20	12.75	20.10
GARCH	15.11	28.22	12.76	20.12
ANN	15.46	28.87	13.64	21.35
CEEMDAN_ANN	34.10	63.70	33.45	27.48
SVR	12.94	24.17	11.83	19.69
CEEMDAN_SVR	30.62	57.18	29.88	14.81
RF	15.50	28.95	14.33	22.14
CEEMDAN_RF	32.58	60.85	31.77	16.39
KRR_Lk	12.80	23.90	9.65	20.17
KRR_Pk	10.86	20.29	7.88	15.22
KRR_Rk	11.26	21.03	8.29	16.22
KRR_Sk	11.79	22.02	8.65	17.15
CEEMDAN_KRR_RF_Lk	13.57	25.35	10.28	22.25
CEEMDAN_KRR_RF_Pk	11.56	21.59	8.36	16.19
CEEMDAN_KRR_RF_Rk	11.69	21.83	8.53	16.71
CEEMDAN_KRR_RF_Sk	11.84	22.10	8.70	17.25



**Figure 5.** Actual versus predicted plot of annual rainfall in the three subdivisions of India.

MAE and MAPE were used to compare the prediction accuracy of different models. It has been found that CEEMDAN-based decomposition is efficient in capturing the rainfall pattern in the datasets. The CEEMDAN-based decomposition followed by the application of ML algorithms resulted in significant improvement in the prediction accuracy.

The proposed algorithm may also be used in predicting high-frequency rainfall data. Other decomposition techniques like wavelet can be employed to preprocess the data before modelling in future studies.

**Conflict of interest:** The authors declare that there is no conflict of interest.

1. Dutta, S., Medhi, H., Karmaker, T., Singh, Y., Prabu, I. and Dutta, U., Probabilistic flood hazard mapping for embankment breaching. *ISH J. Hydraul. Eng.*, 2010, **16**, 15–25.

2. Paul, R. K., Prajneshu and Ghosh, H., Statistical modelling for forecasting of wheat yield based on weather variables. *Indian J. Agric. Sci.*, 2013, **83**, 180–183.
3. Yu, L., Zhang, X. and Wang, S., Assessing potentiality of support vector machine method in crude oil price forecasting. *EURASIA J. Math. Sci. Technol. Educ.*, 2017, **13**, 7893–7904.
4. Wei, Y., Wang, Y. and Huang, D., Forecasting crude oil market volatility: further evidence using GARCH-class models. *Energy Econ.*, 2010, **32**, 1477–1484.
5. Brigida, M., The switching relationship between natural gas and crude oil prices. *Energy Econ.*, 2014, **43**, 48–55.
6. Ramyar, S. and Kianfar, F., Forecasting crude oil prices: a comparison between artificial neural networks and vector autoregressive models. *Comput. Econ.*, 2019, **53**, 743–761.
7. Singh, S., Griffiths, G. and Pham, H., Meteorological drought in Northland, New Zealand: a regional and local analysis using copulas. *J. Hydrol.*, 2021, **60**, 1–16.
8. Grzegorowski, M., Zdravetski, E., Janusz, A., Lameski, P., Apanowicz, C. and Ślęzak, D., Cost optimization for big data workloads based on dynamic scheduling and cluster-size tuning. *Big Data Res.*, 2021, **25**, 100203; <https://doi.org/10.1016/j.bdr.2021.100203>.

9. Jia, X., Shao, M., Zhu, Y. and Luo, Y., Soil moisture decline due to afforestation across the Loess Plateau, China. *J. Hydrol.*, 2017, **546**, 113–122.
10. Paul, R. K. and Garai, S., Wavelets based artificial neural network technique for forecasting agricultural prices. *J. Indian Soc. Probab. Stat.*, 2022, **23**, 47–61.
11. Garai, S., Paul, R. K., Kumar, M. and Choudhury, A., Intra-annual national statistical accounts based on machine learning algorithm. *J. Data Sci. Intell. Syst.*, 2023, 1–8; doi:10.47852/bonviewJDSIS-3202870.
12. Wong, K. K. F., Song, H., Witt, S. F. and Wu, D. C., Tourism forecasting: to combine or not to combine? *Tour. Manage.*, 2007, **28**, 1068–1078.
13. Paul, R. K., ARIMAX–GARCH–WAVELET model for forecasting volatile data. *Model Assist. Stat. Appl.*, 2015, **10**, 243–252.
14. Wang, G. and Ma, J., A hybrid ensemble approach for enterprise credit risk assessment based on support vector machine. *Expert Syst. Appl.*, 2012, **39**, 5325–5331.
15. Chollet, F. and Allaire, J., *Deep Learning with R* Manning Publications Co, NY, USA, 2018.
16. Barredo, J. I., Major flood disasters in Europe: 1950–2005. *Nat. Hazards*, 2007, **42**, 125–148.
17. Sharma, A., Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 3 – a nonparametric probabilistic forecast model. *J. Hydrol.*, 2000, **239**, 249–258.
18. Salma, S., Rehman, S. and Shah, M. A., Rainfall trends in different climate zones of Pakistan. *Pak. J. Meteorol.*, 2012, **9**, 37–47.
19. Garai, S. *et al.*, An MRA based MLR model for forecasting Indian annual rainfall using large scale climate indices. *Int. J. Environ. Climate Change*, 2023, **13**, 137–150.
20. Archer, D. R. and Fowler, H. J., Using meteorological data to forecast seasonal runoff on the River Jhelum, Pakistan. *J. Hydrol.*, 2008, **361**, 10–23.
21. Garai, S. and Paul, R. K., Development of MCS based-ensemble models using CEEMDAN decomposition and machine intelligence. *Intell. Syst. Appl.*, 2023, **18**; <https://doi.org/10.1016/j.iswa.2023.200202>.
22. Liyew, C. M. and Melese, H. A., Machine learning techniques to predict daily rainfall amount. *J. Big Data*, 2021, **8**, 1–18.
23. Paul, R. K. and Garai, S., Performance comparison of wavelets-based machine learning technique for forecasting agricultural commodity prices. *Soft Comput.*, 2021, **25**, 12857–12873.
24. Liu, T., Luo, Z., Huang, J. and Yan, S., A comparative study of four kinds of adaptive decomposition algorithms and their applications. *Sensors (Switzerland)*, 2018, **18**, 1–51.
25. Huang, N. E. *et al.*, The empirical mode decomposition and the Hubert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc., London, Ser. A*, 1998, **454**, 903–995.
26. Torres, M. E. *et al.*, A complete ensemble empirical mode decomposition with adaptive noise. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 4144–4147.
27. Wu, Z. and Huang, N. E., Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Adv. Adapt. Data Anal.*, 2009, **1**, 1–41.
28. Yeh, J. R., Shieh, J. S. and Huang, N. E., Complementary ensemble empirical mode decomposition: a novel noise enhanced data analysis method. *Adv. Adapt. Data Anal.*, 2010, **2**, 135–156.
29. Breiman, L., Bagging predictors. *Risks*, 1996, **24**, 123–140.
30. Schapire, R. E., Freund, Y., Bartlett, P. and Lee, W. S., Boosting the margin: a new explanation for the effectiveness of voting methods. *Ann. Stat.*, 1998, **26**, 1651–1686.
31. Breiman, L., Random forests. *Machine Learning*, 2001, **45**, 5–32.
32. Moore, I. D., Grayson, R. B. and Ladson, A. R., Digital terrain modelling: a review of hydrological, geomorphological and biological applications. *Hydrol. Process.*, 1991, **5**, 3–30.
33. Hoerl, A. E. and Kennard, R. W., Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 1970, **12**, 55–67.

ACKNOWLEDGEMENT: We thank the anonymous reviewer for useful suggestions that helped improve the manuscript.

Received 9 May 2023; revised accepted 12 October 2023

doi: 10.18520/cs/v126/i3/360-367