Development of novel hybrid machine learning models for monthly thunderstorm frequency prediction over Bangladesh

Md. Abul Kalam Azad
Begum Rokeya University
A R M Towfiqul Islam (towfiq_dm@brur.ac.bd)
Begum Rokeya University https://orcid.org/0000-0001-5779-1382
Md. Siddiqur Rahman
Begum Rokeya University
Kurratul Ayen
Begum Rokeya University

Research Article

Keywords: Thunderstorm, Hybrid model, Ensemble empirical mode decomposition, Sensitivity analysis, Random Forest, Bangladesh

Posted Date: February 10th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-204328/v1

Version of Record: A version of this preprint was published at Natural Hazards on April 15th, 2021. See the published version at https://doi.org/10.1007/s11069-021-04722-9.

Development of novel hybrid machine learning models for monthly thunderstorm frequency prediction over Bangladesh

Md. Abul Kalam Azad[1]*, Abu Reza Md. Towfiqul Islam [1]*, Md. Siddiqur Rahman[1], Kurratul Ayen[1]

[1]*Department of Disaster Management, Begum Rokeya University, Rangpur 5400, Bangladesh*

[*]**Corresponding author: towfiq_dm@brur.ac.bd**; **suborno19@gmail.com**

ORCID: 0000-0001-5779-1382

Tel: +880-2-58616687

Fax: +880-2-58617946

Submission: January, 2021

Abstract

Accurate thunderstorm frequency (TSF) prediction is of great significance under climate extremes for reducing potential damages. However, TSF prediction has received little attention because a thunderstorm event is a combination of intricate and unique weather scenarios with high instability, making it difficult to predict. To close this gap, we proposed a novel hybrid machine learning model through hybridization of data pre-processing Ensemble Empirical Mode Decomposition (EEMD) with two state-of-arts models namely artificial neural network (EEMD-ANN), support vector machine (EEMD-SVM) for TSF prediction at three categories of yearly frequencies over Bangladesh. We were demarcated the yearly TSF datasets into three categories for the period 1981-2016 recorded at 28 sites; high (March-June), moderate (July-October), and low (November-February) TSF months. The performance of the proposed EEMD-ANN and EEMD-SVM hybrid models was compared

with classical ANN, SVM, Autoregressive Integrated Moving Average (ARIMA). EEMD-ANN and EEMD-SVM hybrid models showed 8.02%-22.48% higher performance precision in terms of root mean square error (RMSE) compared to other models at high, moderate and low-frequency categories. Eleven out of 21 input parameters were selected based on the Random Forest (RF) variable importance analysis. The sensitivity analysis results showed that each input parameter was positively contributed to building the best model of each category and thunderstorm days are the most contributing parameters influencing TSF prediction. The proposed hybrid models outperformed the conventional models where EEMD-ANN is the most skillful for high TSF prediction, and EEMD-SVM is for moderate and low TSF prediction. The findings indicate the potential of hybridization of EEMD with the conventional models for improving prediction precision. The hybrid model developed in this work can be adopted for TSF prediction in Bangladesh as well as different parts of the world.

Keywords: Thunderstorm; Hybrid model; Ensemble empirical mode decomposition; Sensitivity analysis, Random Forest, Bangladesh

1. Introduction

Thunderstorms are spectacular mesoscale phenomena that affect the environment and pose a severe threat to life, economy, agriculture, and infrastructures. A thunderstorm event results from a turbulent convective activity, which may bring about heavy rainfall, lightning, hail, tornadoes, and thunder (Islam et al., 2020). Thunderstorms occur in almost every region of the world because of meteorological instability and strong moisture convergence, which causes serious convections. It usually exists for less than an hour and typically has varying sizes ranging from a few kilometers to a few hundred kilometers (Saha and Quadir 2016). It is now a well-acknowledged fact that the climate system is getting warmer, which has implications for thunderstorm occurrences (Allen et al., 2014; Trenberth et al., 2007). Severe thunderstorms frequency is likely to increase in the 21st century due to the increasing convective instability (Rädler et al., 2019). Therefore, it is essential to predict the number of thunderstorm events that occur in a particular period under changing meteorological conditions in a given location. Predicting the number of thunderstorm phenomena could provide insights about future thunderstorm incidents under the climate change scenario.

Thunderstorm frequency (TSF) can be defined as the number of thunderstorm occurrences in a given location over a day, month, season, or annum. It is estimated that daily TSF is nearly 45,000 and annually 16 million

worldwide (Siddiqui and Rashid 2008). Many parts of South Asia experience higher TSF during the summer months (March-May) when high temperatures prevail at lower levels create a volatile atmosphere. Each year Bangladesh and its surroundings witness high TSF, especially during the pre-monsoon and early months of the monsoon season; however, thunderstorms occur in all seasons. Spatially, TSF is highest in the northeastern part and less in the southeastern and northwestern parts of Bangladesh (Mannan et al., 2016). Before 1981, the country endured thunderstorm strikes in about nine days in May, which later rose to 12 days. Besides, thunderstorms associated disasters cause severe damage to agricultural yields and infrastructures and lives on the ground and in aviation. Due to the exorbitant impact of thunderstorms on human life and the economy, the Government of Bangladesh declared it a natural disaster on 17 May 2016 (Wahiduzzaman et al., 2020). In contrast, thunderstorms bring crucial rainfall during the dry season, which benefits the country's crop production and cleans the air from dust, haze, and pollutants. A TSF prediction model can help prepare and design a more useful crop calendar adaptive to thunderstorm events. Besides, a TSF prediction model is essential for policymakers to adopt a mitigation plan for reducing the potential damages of thunderstorm casualties.

Thunderstorm prediction is a challenging task due to its small spatiotemporal extension, and the event is a combination of very complex and unique weather scenarios, which are highly unstable. Despite the challenges, many a researcher has attempted to predict thunderstorms worldwide, e.g., Jacovides and Yonetani (1990) in Cyprus; Mills and Colquhoun (1998) in Australia; Haklander and Delden (2003) in Netherland; Manzato (2007) in Italy; Zhen-hui et al. (2013) in China; Ali et al. (2011) in Malaysia; Litta et al. (2013) and Meher et al. (2019) in India; Collins and Tissot (2015) in the USA; Dowdy (2016) in the temperate and tropical regions; Osuri et al. (2017) in Indian monsoon region; Rädler et al. (2019) in Europe; Chen et al. (2020) in Taiwan; Kulikov et al. (2020) in Russia; Bouttier and Marchal (2020) in Western Europe; Islam et al. (2020) in Bangladesh. A variety of approaches have been taken in those studies. For example, Collins and Tissot (2015) used and compared an ANN and MLR model for thunderstorms prediction within 400 km^2 of South Texas; Rädler et al. (2019) used an ensemble of 14 regional climate models such as AR-CHaMo models, EURO-CORDEX model to assess the changes in the frequency of thunderstorm. Most of the studies have focused on Numerical Weather Prediction (NWP) modeling or forecasting of a single thunderstorm event on an hourly basis based on the convective indices. However, studies focused on predicting monthly TSF based on the convective indices and other

thunderstorm-related parameters are still scarce in the literature (Islam et al., 2020). In the present study, we have employed machine learning models including Artificial Neural Network (ANN), Support Vector Machine (SVM), incorporated with Ensemble Empirical Mode Decomposition (EEMD), and Auto-Regressive Integrated Moving Average (ARIMA) modeling to predict the monthly TSF over Bangladesh.

Among the machine learning models, ANN is a powerful model that can identify complex inherent nonlinear relationships between responses and predictors. Therefore, ANN models have drawn attention in the thunderstorm forecasting community (Manzato. 2007; Collins and Tissot, 2015; Litta et al., 2013). SVM is also a useful prediction technique that was used before in thunderstorm prediction (Qiu et al., 2010; Zhen-hui et al., 2013). The time series model like ARIMA is widely used because it can characterize nonlinear data; this model was also applied previously in thunderstorm prediction (Islam et al., 2020). Though these models are not always efficient enough to predict a target dataset accurately. Due to this reason, many researchers have developed techniques that adjoin several types of methods to obtain more accuracy in their prediction (Chen and Letchford. 2007; Gao and Stensrud. 2014; Solari et al., 2017; Suparta and Putro. 2018; Bouttier and Marchal. 2020; Kamangir et al., 2020). The hybrid EEMD integrated machine learning models have successfully applied in different fields of studies, e.g., runoff (Tan et al., 2018); streamflow forecasting (Zhang et al. 2015); rainfall forecasting (Johny et al. 2020) wind speed forecasting (Yu, 2020); groundwater level (Gong et al., 2018). However, TSF prediction has received little attention in the existing literature due to its complicated nature and unique weather feature with high instability, making it difficult to predict. Our work fills this research gap in literature. Therefore, a hybrid EEMD-ANN and EEMD-SVM models, the combination of an ensemble empirical mode decomposition (EEMD) with an ANN and SVM model, are proposed as effective methods to predict monthly TSF. In this study, widely used convective indices and thunderstorm-related variables were used as input parameters. The EEMD-ANN and EEMD-SVM prediction results were compared with three conventional prediction methods, e.g., ANN, SVM, and ARIMA, based on five performance evaluation metrics, i.e., Coefficient of determination (R^2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Index of Agreement (IA) along with the Taylor diagram. Even though machine learning models can solve prediction problems with reasonable accuracy, their predictive capability relies significantly on the input data quality. In such a case, sensitivity analysis can help identify which input

parameter is remarkable in building an ideal model. So, sensitivity analysis is employed in this study to improve the performance of the models.

This study's primary objective is to predict and evaluate the performance of the monthly TSF based on the convective and thunderstorm-related indices. Compare to other earlier studies, our work has two novel aspects. First, we develop hybrid machine learning models to predict monthly TSF at three frequency metrics over Bangladesh for the first time in literature. Second, this study identifies the most contributing parameters influencing TSF prediction and select optimal input parameters using Random Forest model. It is hoped that the novel hybrid model proposed in this work would able to address the challenge of complicated nature of thunderstorm event due to its high instability and randomness.

2. Study area and data sources

The site selected for this research is Bangladesh, a part of Southeast Asia, geographically located between 20° 34' to 26° 38' North latitude and 88° 01' to 92° 42' East longitude. Bangladesh is the biggest deltaic country in the world, occupying 147,570 sq. km area. The three vigorous rivers, Padma, Jamuna, and Meghna, and their tributaries encompass 80% of Bangladesh's floodplains, leaving out the hilly parts. The geographical features of this narrow flat lowland are very well suited for convection, as the moisture conveyed by the monsoon winds from the highly elevated regions and the Bay of Bengal causes the development of convection (Ahmed et al., 2017). Here, the monsoon is probably the controlling feature of climatic variability (Islam et al., 2020), portrayed by pelter-bearing breezes, humbly warm temperatures, and high moisture in the air. As an outcome, thunderstorms, floods, and tidal floods are regular incidents in this country. There are three prominent seasons observed in Bangladesh, which are premonsoon, monsoon, and post-monsoon.

In this study, monthly TSF and TSD data were collected from 28 stations (Fig. 1) of the Bangladesh Meteorological Department (BMD) ranged from 1981 to 2016. There are more meteorological stations in Bangladesh, but those stations do not have long term records of thunderstorms, and few stations have excessive missing data. Therefore, we have excluded those stations. Although some of the selected stations have few missing data, we have filled them by obtaining the nearest station's value. Table S1 contains the missing data information. Thunderstorm frequencies are recorded eight times per day in each station of BMD with a three-hour interval according to the World Meteorological Organization (WMO) standard hour. The number of TSF

observations recorded per day is regarded as the daily TSF. From the daily observations, monthly TSF is computed for each station. Description of the 28 meteorological stations of BMD and an overview of the annual TSF data are given in Table 1. The number of days with thunderstorm observations in a month is regarded as monthly TSD. Daily precipitation data was also collected from the same 28 stations of the BMD from 1981 to 2016. We then converted the daily precipitation into monthly average data and used them as predictors for the model building.

Single point data of Dew/Frost Point at 2 meters (DP), Relative Humidity at 2 meters (RH), Wind Speed range at 50 meters (WS50), and Earth Skin Temperature (ST) data were used as a predictor in building TSF prediction model. These parameters were obtained for the specific latitude and longitude of the selected 28 stations from the NASA Langley Research Center Atmospheric Science Data Center Surface Meteorological and Solar Energy (SSE) web portal supported by the NASA LaRC POWER Project (https://power.larc.nasa.gov/data-access-viewer/), which has a 0.5˚×0.5˚ gridded global dataset. Moreover, the monthly averaged CAPE, Convective Rain Rate (CRR), Convective Precipitation (CPRCP), K Index (KI), and Total Totals (TT) were obtained from Climate Data Store (CDS) of Copernicus Climate Change Service (https://cds-dev.copernicus-climate.eu/cdsapp#!/dataset/reanalysis-era5-single-levels-monthly-means?tab=overview). We have calculated the average of 28 stations for all the data and then used them as country average because this approach helps the future assumption of thunderstorm frequency within a large area. Table 2 corresponds to the descriptions of the datasets used in this study.

For convenience in predicting monthly TSF accurately, we have classified all the data in three categories of time-series, e.g., HTSF (High Thunderstorm Frequency; containing high-frequency months of March, April, May, June), MTSF (Moderate Thunderstorm Frequency; containing moderate-frequency months of July, August, September, October), and LTSF months (Low Thunderstorm Frequency; having low-frequency months of November, December, January, February). This classification helps differentiate the months with high, moderate, and low TSF, and thus, it reduces abrupt fluctuations in the time-series, which increases the prediction accuracy.

3. Methods used

3.1 Parameter Selection

The input parameters were selected based on a Random Forest (RF) relative importance technique performed in Salford Predictive Modeler 8.2. The RF algorithm is a popular and highly flexible supervised artificial intelligence applied to measure the importance of various contributing factors (Rahman et al., 2020). The RF method details can be found in the literature (Rahman and Islam, 2019; Salam and Islam, 2020). We have initially considered 21 input parameters and tested different combinations among them, but some of the parameters were not suitable enough to efficiently predict TSF (Fig. S1). We have excluded 10 of the initially considered parameters affecting the model performances and finalize 11 input parameters (Table 2) for model building. The excluded parameters were Lifted Index (LI), maximum temperature (MaxT), precipitable water (PRW), diurnal temperature (DT), specific humidity (SH), wind speed range at 10 meters (WS10), minimum temperature (MinT), V component of wind (VCW), U component of wind (UCW), and surface pressure (SP). CAPE, CPRCP, CRR, DP, KI, PRCP, RH, ST, TSD, TT, and WS50 were the selected parameters for constructing prediction models. These parameters also have a high correlation with TSF (Fig. S2). Among the selected input parameters, CAPE, KI, and TT are well known for their potentiality in predicting thunderstorm events (Vujovic et al., 2015; Islam et al., 2020).

3.2 Artificial Neural Network (ANN)

Artificial Neural Network (ANN) is one of the most employed techniques for modeling accurate predictions to solve complex and nonlinear problems (Phuong et al., 2017; Alizadeh et al., 2018, Pham et al., 2019). ANNs are data processing systems that exploit learning algorithms to imitate knowledge and save this knowledge in weighted connections, similar to a human brain (Pradhan and Lee, 2010; Boateng et al., 2019). An ANN has numerous processing components called neurons (Boateng et al., 2019). The data are processed by the neurons and then feed-forwarded to the subsequent layer. Corresponding links between layers connect these neurons. On each connecting link, there is a numeric weight. An ANN structure consists of three main layers, e.g., input layers, hidden layers, and output layers. The input layer contains the variables used for model construction; the hidden layers analyze the interconnection between the input and the output parameters based on algorithms, and the output layer represents the predicted variables.

Unlike statistical models, ANNs can automatically synthesize their weights to elevate their attitude. (Boussabaine, 1996). ANN is like a 'black box' which lacks self-explanation. As a result, both ANNs and

193 statistical approaches can be ensembled into a robust and potent methodological platform despite their
194 differences (Karlaftis and Vlahogianni, 2011). At first, an ANN model has to be trained with an acquainted
195 dataset called the 'training' dataset. The model will then 'learn' by synthesizing the neurons' numerical weights
196 regarding the errors between the predicted output values and the target output values through the training process.
197 After the training period, the neural network delivers a model that can predict a target value from a specificity
198 input value. This research has used a backpropagation based neural network regression approach to predict
199 thunderstorm frequency. Here, we have used 11 variables as input parameters and two layers in the hidden unit—
200 the 1st hidden layer composed of 4 neurons and the 2nd layer consisting of 2 neurons (Fig. 2). We have set the
201 learning rate to 0.1 while using the sigmoid function as the activation function. A neural network model can be
202 expressed in mathematical form as Eq.

203 $$y(x) = K\left(\sum_{j=1}^{n} w_j(p)\,.\,x_j(p) + c\right) \quad (1)$$

204 Where,

205 $x_j(p) =$ Input variable in discrete-time t

206 $y(x) =$ Predicted thunderstorm frequency

207 $n =$ Hidden neuron by trial

208 $w_j(p) =$ Weight that connects the i^{th} neuron in the input layer

209 $c =$ Neuronal bias

210 $K(.) =$ Hidden transfer function

211 **3.3 Support Vector Machine (SVM)**

212 SVM, one of the most successful forecasting methods in recent years, was first proposed by Vapnik (1995). It is
213 remarkably capable of handling small-sized datasets and nonlinear problems (Liu and Wang, 2016; Ghimire et
214 al., 2019). So, it has been widely applied in regression modeling. It is one of the most effective predicting tools
215 often used as an alternative approach to ANN. The SVM approximates structural risk minimization based on
216 statistical learning theory (Meng et al., 2019) rather than empirical risk minimization (Huang et., 2014). The
217 SVM description is avoided in this paper because many documents and books have described SVM theory in

detail (Vapnik, 1998; Carrier et al., 2013; Ch et al., 2013; Chiogna et al., 2018; Meng et al., 2019). SVM's basic idea is using the maximum margin algorithm (Pham et al., 2019), which searches for a hyperplane with the largest separating margin between the observed data. SVM can simplify an intricate problem by mapping the complicated nonlinear problem input factors into high-dimension space with kernel functions, transforming the complicated nonlinear problem into a linear problem. This process can find the optical function to fit the observations while avoiding overfitting to maintain the model generality. The useful and popular SVM kernel functions are linear, polynomial, sigmoid, Radial Basis Function (RBF), and so on. This research employs SVM as a regression technique that uses RBF:

$$L(m, m_j) = \exp\left(-\frac{\|m - m_j\|^2}{2\varphi^2}\right) \quad (2)$$

where φ represents the Gaussian noise level of standard deviation.

3.4 Autoregressive Integrated Moving Average (ARIMA)

ARIMA models are widely employed statistical prediction techniques because of their ability to handle nonstationary series efficiently. ARIMA modeling's basic idea is, here, the examined time series is linear and follows a particular normal distribution (Box and Jenkins, 1970). In a traditional ARIMA (p, d, q) model, p is autoregressive (AR), d is the number of differences from the actual time-series data to make it stationary, and q is moving average (MA). The standard equation for ARIMA models is as follows:

$$d_t = \sum_{i=1}^{p} f_i \, d_{t-1} + \sum_{j=1}^{q} \theta_j \, e_{t-j} + \tau_t \quad (7)$$

where d_t is the observed value at time *t*, f_i is the i^{th} number of autoregressive parameter, θ_j is the j^{th} number of moving average parameter and τ_t is the error at time *t*. In this study, the Box–Jenkins methodology is used to formulate the ARIMA (1,1,1) (1,0,0) models for fitting TSF. This methodology comprises model identification, parameter estimation, and testing residual and forecast. A detailed description of the ARIMA model can be found in the literature (Contreras et al., 2003; Shadab et al., 2020).

3.5 Ensemble Empirical Mode Decomposition (EEMD)

Based on Hilbert-Huang Transform (HHT), Huang et al. (1998) first proposed Empirical Mode Decomposition (EMD), which has been employed effectively throughout the decades. This is because of the following

advantages: 1. EMD is a highly efficient and adaptive method for nonlinear and non-stable signals (Chen et al., 2021). 2. HHT is fully adaptive by initially introducing the intrinsic mode functions (IMFs), which is unlike the Wavelet Transform or Fourier Transform that needs a pre-determined basis. The brief mathematical process of the EMD can be found in the literature (Zhou et al., 2014; Wang et al., 2015; Fan et al., 2020). However, few of these IMFs may contain dramatic oscillations of different scales called "mode mixing" (Chen et al., 2021). This inconvenience can make these IMFs lose their physical signification and make the EMD algorithm less robust (Chen et al., 2021). The Ensemble Empirical Mode Decomposition (EEMD) was subsequently proposed by Wu and Huang (2009) to vanquish these shortcomings, adding a Gaussian white noise into the raw data series. It enables EEMD to automatically attribute signals with different time scales to the precise reference scales. In consequence, the correlation between the resultant IMFs and the raw series significantly improved. The processes of EEMD are as follows:

1. Add the normally distributed Gaussian white noise $\omega(t)$ to the target series $f(t)$ to get a new signal $F(t)$: $F(t) = f(t) + \omega(t)$;

2. Decompose $F(t)$ using EMD method. Obtain IMFs $C_i(t)$ and the residual $r(t)$:

$$F(t) = \sum_{i=1}^{n} C_i(t) + r(t);$$

3. Adding different white noise sequence to the same raw series and repeat the above steps;

4. Since the mean value of Gaussian white noise is equal to zero, the IMFs obtained are integrated and averaged as the final result:

$$\overline{IMF} = \frac{1}{N} \sum_{m=1}^{N} C_{j,m}$$

where $C_{j,m}$ represents the jth IMFs from the mth time, N denotes the number of the added white noise sequences. Resolved by the above process, we have obtained six IMFs in total from the raw data series.

3.6 Hybrid Model building

In weather prediction, predicting thunderstorms is one of the most challenging tasks because of the implicit nonlinearity of thunderstorms' physics and dynamics (Litta et al., 2013). A problem with ANN, SVM, and other linear and nonlinear prediction models is that they cannot accurately handle nonstationary data. To solve the

nonstationary problem, we have built an EEMD-ANN and an EEMD-SVM hybrid model. In our cases, EEMD decomposes the monthly TSF data of HTSF, MTSF, and LTSF into six IMFs and one residual, presented in Fig. 3(a–c). In each category's case, we have selected the first three IMFs as predictors of the hybrid models with the other meteorological predictors. The value of IMF1, IMF2, and IMF3 are more relevant to the original series, and they are the most nonlinear components of their respective series. On the contrary, the value of IMF4, IMF5, and IMF6 is minimal. The residuals' value is not entirely relevant, so that their contributions are lesser to fit the model, indicating difficulty in predicting the TSF more accurately. Besides, the correlation analysis (Table 3) suggested that the first three IMFs are the essential variables in predicting monthly TSF. Therefore, using these sub-series in building the models might enhance the performances by giving useful information on several resolution levels. Fig. (4) demonstrates the methodological procedures of the EEMD-ANN and EEMD SVM prediction models. As seen in Fig. (4), the main steps of the presented EEMD-ANN or EEMD-SVM prediction can be summarized as follows:

Step 1: Decompose the original time series into a finite set of IMFs and a residue using EEMD.

Step 2: Eliminate the irrelevant or redundant IMFs and residue and select the IMFs with the highest frequency bands and a more significant correlation with the original series.

Step 3: Combine the selected IMF with other input parameters and then apply the ANN or SVM model to construct a prediction model for predicting TSF.

Step: 4 Obtain the predicted output by the models.

3.7 Model performance evaluation

The performance of each prediction model is evaluated using the following metrics. By letting δ_t represent the reference values, $\hat{\delta}_t$ represents the predicted values at time t, and $\bar{\delta}$ denotes the mean of the reference values.

Coefficient of determination (R^2)

R^2 is the Coefficient of determination, which is a number between 0 and +1. It measures the degree of alignment between two parameters; in our case, the reference data (δ_t) and the predicted data ($\hat{\delta}_t$). It quantifies how well future outcomes are likely to be predicted by the model. The Coefficient of determination is calculated according to the formula as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(\delta_t - \hat{\delta}_t\right)^2}{\sum_{i=1}^{N}\left(\delta_t - \bar{\delta}\right)^2}$$

where $\bar{\delta}$ represents the average of the reference values.

Root Mean Square Error (RMSE)

The second statistical metric $RMSE$ is defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\delta_i - \hat{\delta}_i)^2}$$

where N is the number of points in the test dataset.

Mean Absolute Error (MAE.)

The third statistical metric, namely, the MAE, can be defined as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\delta_i - \hat{\delta}_i|$$

According to Eq. (), the MAE is the average of the absolute error between δ_i and $\hat{\delta}_i$ $(i = 1,2,\cdots,n)$.

Mean Absolute Percentage Error (MAPE)

The fourth criterion is $MAPE$, which is used to compute the relative error between $|\delta_i - \hat{\delta}_i|$ and $|\delta_i|$ $(i = 1,2,\cdots,n)$, which is defined as

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|\delta_i - \hat{\delta}_i|}{|\delta_i|} \times 100\%$$

Index of Agreement (IA.)

The IA is used in this study and is defined as follows:

$$IA = 1 - \frac{\sum_{i=1}^{N}(\hat{\delta}_i - \delta_i)^2}{\sum_{i=1}^{N}(|\hat{\delta}_i - \bar{\delta}| + |\delta_i - \bar{\delta}|)^2}$$

The IA is a dimensionless metric used in comparing different models. The outcome value of IA is always between 0 and 1. For a "perfect" model, the R^2 and IA are equivalent to 1, and the MAE, $MAPE$, and $RMSE$ are equal to

0. The three commonly used metrics, i.e., MAE, $MAPE$, and $RMSE$, all quantify the differences between the predicted and observed concentrations. However, the $RMSE$ is more sensitive to extreme values, and the MAPE is sensitive to small values because of the power term. IA summarizes the similarity between the predicted and observed propensities.

Taylor Diagram

Another model performance evaluation technique, "Taylor diagram," is used in this study. This technique is widely used to compare model data and tracking changes in model performances. The mathematical theory to construct the diagram can be found in the literature (Taylor, 2001; Pakalidou and Karacosta, 2018). The diagram provides the degree of similarity between the observed point and the test point. The closest the test point to the observed point, the highest the accuracy of the model.

3.8 Sensitivity Analysis

In this section, the sensitivity analysis of the input parameters was measured for the best performing model of each season using the following equation:

$$sensitivity = \frac{R^2 - \bar{R}^2}{R^2} \times 100$$

where R^2 is the square of the correlation coefficient of the best prediction model and $\bar{R}^2$ is the square of the correlation coefficient of the predicted model when a specific parameter is excluded from the model.

4. Results

4.1 Comparative analysis

In this section, the prediction results are presented along with a comparison of model performance metrics among the predicted outputs from different models.

Fig. (5a) depicts the prediction results of HTSF using ANN, SVM, EEMD–ANN, EEMD-SVM, and ARIMA models of the testing dataset. In general, the hydrograph illustrates that all the models except for ARIMA have an excellent performance for simulating the monthly TSF. Moreover, Fig. (5b) shows the scatter plots of prediction by these models, which indicates that the EEMD-ANN and EEMD-SVM model have the best performance for predicting the monthly TSF. Here, for EEMD-ANN and EEMD-SVM, the

least square fitting line is slightly closest to the best possible 45-degree fitting line than ANN, SVM, and ARIMA. From Fig (5a) and Fig (5b), it is not clear which model performs better between EEMD-ANN and EEMD-SVM, as both are very close.

It is evident from Fig. (5c) that EEMD-ANN generated the highest value of CC and lowered centered RMS difference, while the other hybrid model EEMD-SVM was approximately equal in this regard. Comparing the models' standard deviation suggests that the EEMD-ANN and EEMD-SVM models were more in agreement and closer to observed values than the conventional ANN and SVM. It can also be seen in Fig. (5c) that the EEMD-ANN has the standard deviation relative to the observed, but the traditional ANN has a standard deviation less than the observed. It shows that the hybrid EEMD-ANN outperforms classic ANN. Here also, the ARIMA model has the furthest distance from the reference data.

It can be observed from Table 4 that the EEMD-SVM model has a decent performance, acquiring good R^2, training, and testing RMSE, MAE, MAPE, and IA values of 0.978, 1.364, 1.367, 1.095, 8.475, and 0.994, respectively. The ANN and SVM models have also acquired a good R^2, training RMSE, testing RMSE, MAE, MAPE, and IA of 0.973 and 0.964, 1.478, and 1.67, 1.562 and 1.777, 1.099 and 1.446, 6.049 and 9.481, 0.992 and 0.99, respectively. ARIMA model has the worst performance compared with the other models, as observed in Table 4. The EEMD-ANN model has acquired the best score in all the validation metrics, gaining the best R^2, training, and testing RMSE, MAE, MAPE, and IA values of 0.982, 1.177, 1.241, 0.917, 5.738, and 0.995, respectively. It increases the R^2 and IA by 1% and 0.3% and reduces the training RMSE, testing RMSE, MAE, and MAPE by 20.34%, 20.58%, 16.52%, and 0.31%, respectively, compared to the conventional ANN.

The hydrograph of the MTSF prediction results of the testing dataset is presented in Fig. (6a). Again, the hydrograph demonstrates that all the approaches except ARIMA have a decent performance in predicting the monthly MTSF. The scatterplot of the predicted outputs suggests that the EEMD-ANN and EEMD-SVM approaches have the closest least-square fitting line to the 45-degree line (Fig. 6b).

The Taylor diagram (Fig. 6c) suggests that the EEMD-ANN and EEMD-SVM have performed better than conventional ANN, SVM, and ARIMA. For the hybrid models, the correlation coefficient is >0.95, while

for the conventional models, it is <0.95. The lower centered RMS difference also confirms that the hybrid models are more in agreement than the conventional models. It is immensely challenging to find out the better model between the hybrid models as both the models have performed almost equally.

The validation metrics from Table 4 confirms that conventional ANN and SVM are remarkably inferior to the hybrid approaches; the R^2, training RMSE, testing RMSE, MAE, MAPE and IA are 0.839 and 0.881, 1.306 and 1.644, 1.45 and 1.184, 1.161 and 0.93, 8.485 and 6.852, 0.952 and 0.967 respectively. Moreover, the ARIMA model performs worse than the other approaches. Both the EEMD-ANN and EEMD-SVM have superior scores in predicting MTSF. The EEMD-ANN has gained a substantial R^2, training RMSE, testing RMSE, MAE, MAPE, and IA of 0.923, 1.271, 1.124, 0.889, 6.133, and 0.971, respectively, which increases the R^2 and IA by 10.012% and 1.995% and reduces the training RMSE by 2.68%, testing RMSE by 22.482%, MAE by 23.428% and MAPE by 2.352% compared to the conventional ANN. The EEMD-SVM has also acquired a significant R^2, training RMSE, testing RMSE, MAE, MAPE, and IA of 0.924, 1.412, 1.089, 0.909, 6.323, and 0.973, respectively, which improves the R^2 and IA by 4.881% and 0.62% and minimizes the training RMSE by 14.112%, testing RMSE by 8.024%, MAE by 2.258% and MAPE by 0.529% while comparing with the conventional SVM.

Fig. (7a) exhibits the predicted outcomes of the testing dataset of the LTSF using ANN, SVM, EEMD-ANN, EEMD-SVM, and ARIMA models. The hydrograph suggests that these models' predicted values, except for the ARIMA model, agree well with the observations of LTSF. However, several points exhibit a clear difference between the predicted and observed values. These differences often occur during the month of abrupt changes of TSF (e.g., 2012, 2014-2016). The scatterplot (Fig. 7b) indicates that the least-square fit for the EEMD-SVM prediction is closest to the perfect 45-degree fit, closely followed by the EEMD-ANN, SVM, and ANN approaches. However, there is a massive difference between the perfect fit and the least-square fit for the ARIMA model prediction.

EEMD-SVM approach has performed better than all the models (Table 4) in almost every validation metrics, which is coherent with the scatterplot (Fig. 7b) and the Taylor diagram (Fig. 7c) results. It is the best among the LTSF prediction models used in this study, gaining an impressive R^2 of 0. 0.886, training

RMSE of 0.291, testing RMSE of 0.398, MAE of 0.325, MAPE of 17.925, and IA of 0.966. However, the EEMD-ANN has a better performance in acquiring the lower training RMSE (0.23) than EEMD-SVM, but it was not consistent in gaining better R^2 value (0.864), testing RMSE (0.439), MAE (0.352), MAPE (21.929), and IA (0.956). The conventional ANN, SVM, and ARIMA models were remarkably inferior to the hybrid approaches predicting the LTSF. Compared to the traditional SVM, the EEMD-SVM raises the R^2, IA and lessens the training RMSE, testing RMSE, MAE, MAPE by 4.03%, 1.90% and 10.39%, 14.83%, 11.79%, 32.07%, which is remarkable.

4.2 Sensitivity assessment

The sensitivity analysis result of the best HTSF prediction model, EEMD-ANN, is shown in Fig. (8a). The rankings of the three most sensitive parameters are TSD, IMF1, and IMF2. Apart from these, all the parameters positively contribute to achieving better performance during the TSF prediction of the high-frequency months. TT, CPRCP, CAPE, and PRCP also played a pivotal role in constructing the EEMD-ANN model, followed by CRR, IMF3, WS50, RH, ST, DP, and KI (Fig. 8a). A similar result is observed in the three most sensitive parameters (Fig. 8b) for the EEMD-SVM model while predicting MTSF. In building this EEMD-SVM model, TT, CAPE, IMF3 are the next ranked sensitive parameters, respectively, which are followed by CPRCP, CRR, WS50, DP, PRCP, RH, KI, and ST. Fig. (8c) depicts the sensitivity analysis of the best LTSF prediction model, which is EEMD-SVM. TSD, IMF1, IMF3 are the top three sensitive parameters for predicting LTSF using the EEMD-SVM model. IMF2, CAPE, CPRCP, CRR, TT, PRCP also play a vital role in constructing the model for predicting LTSF. Although the parameters like DP, KI, RH, ST, and WS50 have low sensitivity value, they help achieve better prediction accuracy.

5. Discussion and conclusion

Due to increasing computational abilities, machine learning algorithms in modeling severe weather events are becoming progressively popular in current atmospheric studies. It is because of robust prospects from its use in operational prediction (McGovern et al., 2017; Czernecki et al., 2016; Kamangir et al., 2020) and the generating of severe weather events that add forthcoming changes in their frequencies (Allen et al., 2015; Lee et al., 2020). It is a fact that most current machine learning models have superiority over conventional statistical models

(Gagne et al., 2017). Furthermore, some machine learning models like random forest permit studying variable importance, making it likely to obtain a better insight into the factors influencing physics behind such studied processes.

In this research, we assessed the use of machine learning algorithms in modeling high, moderate, and low frequencies thunderstorm events on a monthly scale. This analysis was based on the observed TSF dataset, and convective parameters come from ERA5 reanalysis datasets. In the case of HTSF, the hybrid EEMD-ANN outperformed other models based on the evaluation criteria. For MTSF and LTSF, the hybrid EEMD-SVM model has superior performance than other standalone models. Theoretically, the three sub-series with various thunderstorm frequencies were used instead of thunderstorm events because it has physical meaning. TSF analysis can be used operationally to help human policy-making by lessening the cognitive associated with thunderstorm event identification.

Uncertainty increases in low TSF months (winter) because of the low SST and northeast wind flow from the BoB and lowers vapor flux availability. We anticipate that due to low surface temperature and soil moisture, the winter season (November to February) is the least favorable for forming TSF. However, this outcome is not surprising, and it can be underlined by analyzing the TSF pattern of three categories. This work proposed a prediction strategy for TSF prediction circumventing the probable precision reduction triggering from calibrating the decomposition method during implementing and accepting the application of operational research reported in several previous works (Napolitano et al., 2011; Zhang et al., 2015; Johny et al., 2020).

The outcomes obtained in this research indicate that EEMD can efficiently increase prediction accuracy, and the proposed EEMD-ANN model can achieve notable improvement over the conventional ANN method in the high, medium, and low TSF monthly time-series predictions. The EEMD-ANN is more successful in capturing the HTSF monthly, showing remarkable precision than the SVM-EEMD model. Our finding is similar to the other hydrological time-series studies (Wang et al., 2015). One probable reason for the improved performance of EEMD-ANN can be the method's capability to solve complex and nonlinear problems (Phuong et al., 2017).

In predicting MTSF, the Taylor diagram suggests EEMD-ANN is slightly more accurate than the EEMD-SVM. But the validation metrics suggest otherwise. We have selected the EEMD-SVM as the best model for MTSF prediction because the difference between the models is very narrow in the Taylor diagram, and it has performed

439 better in most of the validation metrics. Besides, the EEMD-SVM has gained a substantial improvement in
440 testing RMSE than EEMD-ANN, which indicates better model fitting for EEMD-SVM. The difference in the
441 Taylor diagram and validation metrics results is probably due to the diagram's algorithm based on root mean
442 square difference, standard deviation, and correlation coefficient. The standard deviation of the observed and
443 predicted datasets might create this result difference. Also, the EEMD-SVM is more accurate in predicting the
444 LTSF monthly. This is because SVM has been incredibly robust and efficient in nonlinear noise mixed data
445 (Devak et al., 2015). Besides, the potential of decomposition might be more prominent in predicting the TSF
446 dataset in the EEMD-ANN or EEMD-SVM model than the standalone model because the hybrid model can
447 overcome the shortcomings of the standalone model to produce a synergetic impact on prediction. The hybrid
448 EEMD-SVM can help avoid the overfitting or underfitting problem of the SVM model caused by the input
449 parameters' improper determination. This also implies that the EEMD tool is applicable for decomposing
450 monthly TSF time series and the idea of "decomposition and ensemble" is suitable. The findings in this work
451 agree well with those obtained from the studies (Hawinkel et al., 2015; Wang et al., 2015; Czernecki et al., 2016).
452 Previous studies have found that hybrid EEMD-ANN and EEMD-SVM models outperformed the classical
453 models, which apply original datasets in other fields of studies, e.g., runoff (Tan et al., 2018); streamflow
454 forecasting (Zhang et al. 2015); rainfall forecasting (Johny et al. 2020) wind speed forecasting (Yu, 2020);
455 groundwater level (Gong et al., 2018). The hybrid model is robust, theoretically justified, and more realistic
456 compared to other standalone models. It can be said that the proposed methodology can not only predict the
457 complicated thunderstorm frequency over Bangladesh rationally well, but it can also attain extreme climatic
458 events.
459 Topographical differences, wind regimes, and the inland distance far from the coastal and hilly regions may
460 differ sensitivity results in these categories. Based on the sensitivity assessment, TSD, IMF1, and IMF2
461 generated the highest score, similar to other thunderstorm-associated parameters in India by Umakanth et al.
462 (2020). TSD is very high in sensitivity analysis due to an enhanced number of TSF causing moist air circulated
463 from the Bay of Bengal (BoB). When passing the equatorial belt, the southeastern air masses go into the
464 southwest monsoon due to Ferrell's law, which brings a large amount of thunderstorm in the country. Generally,
465 the high TSD in May and June is observed in the northeast region, close to Cherapunji, where the cloud formation

is high, and hill ranges generate a tremendous amount of water vapor flex and precipitation. In recent times, more vigorous and more continuous moist is derived from the BoB because of elevated sea surface temperature. The high sea surface temperature triggered a rise in CAPE in most parts of Bangladesh (Wahiduzzaman et al., 2020; Sahu et al., 2020). These findings support the outcomes of Glazer et al. (2020), who assessed the variations in TSD in Bangladesh due to global climate change and revealed an increase in TSD in many regions of the country. An increase in CAPE and higher moisture content in the BoB may play a vital role in enhancing TSD. Thus, the sensitivity assessment gives a physical means to capture the non-overlapping TSD that would otherwise trigger the concern of multicollinearity. This is obvious from the improvement in the model performance for high TSF identification reported in earlier works (Siddiqui and Rashid, 2008; Gagne et al., 2017). Generally, the problem of over-prediction is a familiar matter for predicting a severe extreme event that can be lessened using current machine learning methods (Czernecki et al., 2019), particularly if various data sources are coupled. All the other convective parameters, e.g., TT, CPRCP, CRR, KI, and the meteorological parameters, e.g., PRCP, RH, ST, WS50, have positively contributed to the best model building. This is because all these input parameters are positively associated with a high correlation with thunderstorm occurrences.

Employing these hybrid ensemble models to predict monthly TSF is crucial for further studies. There are some advantages to the proposed hybrid models. First, the basic principle of the EEMD is elementary, can still give a thorough understanding of the monthly TSF time series dataset. Second, it is suitable and adequate to couple the EEMD with ANN, ARIMA, SVM to predict the nonstationary and nonlinear TSF. Third, the EEMD-ANN and EEMD-SVM models' prediction outcomes are more precise when applying the TSF time series decomposition. Fourth, the developed hybrid models do not need complex policy-making about each specific case study's obvious form. Thus, developing these hybrid prediction models by integrating EEMD may lead to more robust and better prediction outcomes. It may also be useful in further studies focused on extreme events prediction for various problems related to effective disaster management. The application of machine learning algorithms in a thunderstorm prediction brings with a new promise for forthcoming studies concerning both operational predictors and meteorological research that intend to examine observed and future variations in frequencies of severe extreme events (Yasen et al., 2017; Taszarek et al., 2019).

It is worth mentioning that the limitation of this research lies in two perspectives. First, although monthly TSF data were taken from 28 stations, future studies using datasets from different regions may be needed to strengthen these valid conclusions because the performance of data-driven models is data-based and case reports explicit. Second, the coupling preprocessing technique with a machine learning algorithm, a division of the training and testing datasets, and model selection criteria are a vital factor affecting the overall performance of the hybrid models. Thus, future works are solicited, which may shed much light on this concern. Our future work includes the case-study concept of generating a seasonal TSF forecasting on a continental scale that can provide deep insight into the severe weather event's current knowledge. Also, testing our hybrid model is to forecast thunderstorm frequency for other similar climatic regions globally. The ERA5 based parameters can be used in the RF model that is yet more reliable than any sole parameters used in operational models (Gagne et al., 2017). In addition to this, good tuning of generated machine learning is feasible if it is more robustly fitted on a considerable number of datasets or adding new parameters from satellite datasets. Hence, with computational interests in modeling tools, machine learning models play a pivotal role in examining thunderstorm events' climatological perspective and improving operational prediction.

Acknowledgement

We are grateful to the Department of Disaster Management, Begum Rokeya University, Rangpur for all sort of assistant provided during this study. Furthermore, we would like to thank the Bangladesh Meteorological Department (BMD) for providing required data for this research.

Conflict of interest

None

References

Ahmed, M.K., Alam, M.S., Yousuf, A.H.M., Islam, M.M., 2017. A long-term trend in precipitation of different spatial regions of Bangladesh and its teleconnections with El Nino/ southern oscillation and Indian Ocean dipole. Theor. Appl. Climatol. 129 (1–2), 473–486. https://doi.org/10.1007/s00704-016-1765-2

Ali, A.F., Johari, D., Ismail, N.F.N., Musirin, I., Hashim, N., 2011. Thunderstorm forecasting by using artificial neural network. 5th International Power Engineering and Optimization Conference, Shah Alam, Selangor, 2011, pp. 369-374. https://10.1109/PEOCO.2011.5970391

519 Alizadeh, M., Alizadeh, E., Asadollahpour Kotenaee, S., Shahabi, H., Beiranvand Pour, A., Panahi,M., et al.,
520 2018. Social vulnerability assessment using artificial neural network (ANN) model for earthquake hazard in
521 Tabriz city, Iran. Sustainability. 10, 3376. https://doi.org/10.3390/su10103376
522 Allen, J.T., Karoly, D.J., 2014. A climatology of Australian severe thunderstorm environments 1979–2011: inter-
523 annual variability and ENSO influence. Int. J Climatol. 34, 81–97. https://doi.org/10.1002/joc.3667
524 Allen, J.T., Tippett, M.K., Sobel, A.H., 2015. An empirical model relating US monthly hail occurrence to large–
525 scale meteorological environment. J. Adv. Model. Earth Syst. 7, 226–243.
526 https://doi.org/10.1002/2014MS000397.
527 Boateng, E.B., Pillay, M., Davis, P., 2019. Predicting the level of safety performance using an artificial neural
528 network. In: Ahram, T., Karwowski, W., Taiar, R. (Eds.), Human Systems Engineering and Design. Advances
529 in Intelligent Systems and Computing. 876, 705-710. https://doi.org/10.1007/978-3-030-02053-8_107
530 Boussabaine, A.H., 1996. The use of artificial neural networks in construction management: a review. Constr.
531 Manag. Econ. 14 (5), 427-436. https://doi.org/10.1080/014461996373296
532 Bouttier, F., Marchal, H., 2020. Probabilistic thunderstorm forecasting by blending multiple ensembles. Tellus
533 A: Dynamic Meteorology and Oceanography. 72(1), 1-19. https://doi.org/10.1080/16000870.2019.1696142
534 Box, G., Jenkins, G., 1970. Time series analysis: forecasting and control. Wiley, Hoboken.
535 Carrier, C., Kalra, A., Ahmad, S., 2013. Using paleo reconstructions to improve streamflow forecast lead time
536 in the western United States. J.A.W.R.A. J. Am. Water Resour. Assoc. 49 (6), 1351–1366.
537 https://doi.org/10.1111/jawr.12088
538 Ch, S., Anand, N., Panigrahi, B.K., Mathur, S., 2013. Streamflow forecasting by SVM with quantum behaved
539 particle swarm optimization. Neurocomputing. 101, 18–23. https://doi.org/10.1016/j.neucom.2012.07.017
540 Chen, I., Hong, J., Tsai, Y., Fong, C., 2020. Improving Afternoon Thunderstorm Prediction over Taiwan through
541 3DVar-Based Radar and Surface Data Assimilation. Weather and Forecasting. 35(6), 2603-2620.
542 https://doi.org/10.1175/WAF-D-20-0037.1
543 Chen, L., Letchford, C.W., 2007. Numerical simulation of extreme winds from thunderstorm downbursts.
544 Journal of Wind Engineering and Industrial Aerodynamics. 95 (9–11), 977-990.
545 https://doi.org/10.1016/j.jweia.2007.01.021

Chen, Y., Dong, Z., Wang, Y., Su, J., Han, Z., Zhou, D., Zhang, K., Zhao, Y., Bao, Y., 2021. Short-term wind speed predicting framework based on EEMD-GA-LSTM method under large scaled wind history. Energy Conversion and Management. 227, 113559. https://doi.org/10.1016/j.enconman.2020.113559

Chiogna, G., Marcolini, G., Liu, W., Pérez Ciria, T., Tuo, Y., 2018. Coupling hydrological modeling and support vector regression to model hydropeaking in alpine catchments. Science of The Total Environment. 633, 220–229. https://10.1016/j.scitotenv.2018.03.162

Collins, W., Tissot, P., 2015. An artificial neural network model to predict thunderstorms within 400km^2 South Texas domains. Meteorol. Appl. 22, 650–665. https://10.1002/met.1499

Contreras, J., Espinola, R., Nogales, F.J., Conejo, A.J., 2003. A.R.I.M.A. models to predict next-day electricity prices. IEEE Transactions on Power Systems. 18(3), 1014-1020. https://10.1109/TPWRS.2002.804943

Czernecki, B., Taszarek, M., Kolendowicz, L., Konarski, J., 2016. Relationship between human observations of thunderstorms and the PERUN lightning detection network in Poland. Atmos. Res. 167, 118–128. https://doi.org/10.1016/j.atmosres.2015.08.003.

Devak, M., Dhanya, C., Gosain, A., 2015. Dynamic coupling of support vector machine and K-nearest neighbour for downscaling daily rainfall. Journal of Hydrology, 525: 286-301. https://doi.org/10.1016/j.jhydrol.2015.03.051

Dowdy, A.J., 2016. Seasonal forecasting of lightning and thunderstorm activity in tropical and temperate regions of the world. Sci. Rep. 6, 20874. https://doi.org/10.1038/srep20874

Fan, X., Zhang, Y., Krehbiel, P.R., Zhang, Y., Zheng, D., Yao, W., Xu, L., Liu, H., Lyu, W., 2020. Application of ensemble empirical mode decomposition in low-frequency lightning electric field signal analysis and lightning location. IEEE Trans. Geosci. Remote Sens. 59, 86-100. https://doi.org/10.1109/TGRS.2020.2991724

Gagne, D.J., McGovern, A., Haupt, S.E., Sobash, R.A., Williams, J.K., Xue, M., 2017. Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. Weather Forecast. 32, 1819–1840. https://doi.org/10. 1175/WAF-D-17-0010.1

Gao, J., Stensrud, D.J., 2014. Some Observing System Simulation Experiments with a Hybrid 3DEnVAR System for Storm-Scale Radar Data Assimilation. Monthly Weather Review. 142(9), 3326-3346. https://doi.org/10.1175/MWR-D-14-00025.1

573 Ghimire, S., Deo, R. C., Downs, N. J., Raj, N., 2019. Global solar radiation prediction by ANN integrated with
574 European Centre for medium range weather forecast fields in solar rich cites of Queensland Australia. Journal
575 of Cleaner Production. 216, 288-310. https://10.1016/j.jclepro.2019.01.158
576 Glazer, R.; Torres–Alavez, J.A.; Coppola, E.; Das, S.; Ashfaq, M.; Sines, T. Projected changes to Severe
577 Thunderstorm environments as a result of 21st century warming from RegCM CORDEX–CORE simulations.
578 EGU Gen. Assem. 2020, 2020, 970
579 Gong, Y., Wang, Z., Xu, G., Zhang, Z., 2018. A Comparative Study of Groundwater Level Forecasting Using
580 Data-Driven Models Based on Ensemble Empirical Mode Decomposition, Water 2018, 10, 730;
581 doi:10.3390/w10060730
582 Haklander, A.J., Delden, A.V., 2003. Thunderstorm predictors and their forecast skill for the Netherlands.
583 Atmospheric Research. 67–68, 273-299. https://doi.org/10.1016/S0169-8095(03)00056-5
584 Hawinkel, P.; Swinnen, E.; Lhermitte, S.; Verbist, B.; Van Orshoven, J.; Muys, B., 2015. A time series
585 processing tool to extract climate-driven interannual vegetation dynamics using ensemble empirical mode
586 decomposition (EEMD). Remote Sens. Environ. 169, 375–389
587 Huang, N.E., Zheng S., Steven R. L., Manli C.W., Hsing H.S., Quanan Z., Nai-Chyuan Y., Chi C.T., Henry H.
588 L., 1998. The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time
589 Series Analysis. Proceedings: Mathematical, Physical and Engineering Sciences 454, no. 1971, 903-95.
590 https://www.jstor.org/stable/53161
591 Huang, S., Chang, J., Huang, Q., Chen, Y., 2014. Monthly streamflow prediction using modified EMD-based
592 support vector machine. Journal of Hydrology. 511, 764-775. https://doi.org/10.1016/j.jhydrol.2014.01.062
593 Islam, A.R.M.T., Nafiuzzaman, M., Rifat, J., Rahman, M.A., Chu, R., Li, M., 2020. Spatiotemporal variations
594 of thunderstorm frequency and its prediction over Bangladesh. Meteorol. Atmos. Phys.
595 https://doi.org/10.1007/s00703-019-00720-6
596 Jacovides, C.P., Yonetani, T., 1990. An evaluation of stability indices for thunderstorm prediction in greater
597 Cyprus. American Meteorological Society. 5 (4), 559-569. https://doi.org/10.1175/1520-
598 0434(1990)005<0559:AEOSIF>2.0.CO;2

599 Johny, K., Pai1, ML, Adarsh S., 2020. Adaptive EEMD-ANN hybrid model for Indian summer monsoon rainfall
600 forecasting, Theoretical and Applied Climatology https://doi.org/10.1007/s00704-020-03177-5
601 Kamangir, H, Collins, W., Tissot P, King, S.A., 2020. A deep-learning model to predict thunderstorms within
602 400 km2 South Texas domains, Meteorological Application, https://doi.org/10.1002/met.1905
603 Karlaftis, M.G., Vlahogianni, E.I., 2011. Statistical methods versus neural networks in transportation research:
604 differences, similarities and some insights. Transport. Res. C Emerg. Technol. 19 (3), 387-399.
605 https://doi.org/10.1016/j.trc.2010.10.004
606 Kulikov, M.Y., Belikovich, M.V., Skalyga, N.K., Shatalina, M.V., Dementyeva, S.O., Ryskin, V.G., Shvetsov,
607 A.A., Krasil'nikov, A.A., Serov, E.A., Feigin, A.M., 2020. Skills of Thunderstorm Prediction by Convective
608 Indices over a Metropolitan Area: Comparison of Microwave and Radiosonde Data. Remote Sens. 12, 604.
609 https://doi.org/10.3390/rs12040604
610 Lee, JG, Ki-Hong Min, KH, Park, H., Kim, Y., Chung, CY, Chang, E.C., 2020. Improvement of the Rapid-
611 Development Thunderstorm (RDT) Algorithm for Use with the GK2A Satellite, Asia-Pacific Journal of
612 Atmospheric Sciences https://doi.org/10.1007/s13143-020-00182-6
613 Litta, A.J., Idicula, S.M., Mohanty, U.C., 2013. Artificial Neural Network model in prediction of meteorological
614 parameters during premonsoon thunderstorms. International Journal of Atmospheric Sciences. 2013, 14.
615 https://doi.org/10.1155/2013/525383
616 Liu, Y., Wang, R., 2016. Study on network traffic forecast model of S.V.R. optimized by G.A.F.S.A. Chaos,
617 Solitons & Fractals 89, 153-159. https://doi.org/10.1016/j.chaos.2015.10.019
618 Mannan, M.A., Chowdhury, M.A.M., Karmakar, S., Ahmed, S., Rahman, A., Mondal, M.S.H., 2016. Prediction
619 of heavy rainfall in association with severe thunderstorm in Bangladesh during pre-monsoon season. The
620 Atmosphere, Bangladesh Meteorological Department, Dhaka, Bangladesh. 6(1), 64-76.
621 Manzato, A., 2007. Sounding-derived indices for neural network based short-term thunderstorm and rainfall
622 forecasts. Atmospheric Research. 83, 349-365. https://10.1016/j.atmosres.2005.10.021
623 McGovern, A., Elmore, K.L., Gagne, D.J., Haupt, S.E., Karstens, C.D., Lagerquist, R., Smith, T., Williams, J.K.,
624 2017. Using artificial intelligence to improve real-time decision-making for high-impact weather. Bull. Am.
625 Meteorol. Soc. 98, 2073–2090. https://doi.org/10.1175/BAMS-D-16-0123.1.

Meher, J.M., Das, L., 2019. Skill of CMIP5 climate models to reproduce the stability indices in identifying thunderstorms over the Gangetic West Bengal. Atmospheric Research. 225, 172-180. https://doi.org/10.1016/j.atmosres.2019.04.006

Meng, E., Huang, S., Huang, Q., Fang, W., Wu, L., Wang, L., 2019. A robust method for non-stationary streamflow prediction based on improved EMD-SVM model. Journal of Hydrology. 568, 462-478. https://doi.org/10.1016/j.jhydrol.2018.11.015

Mills, G.A., Colquhoun, J.R., 1998. Objective prediction of severe thunderstorm environments: Preliminary results linking a decision tree with an operational regional N.W.P. model. Weather and Forecasting. 13, 1078-1092. https://doi.org/10.1175/1520-0434(1998)013<1078:OPOSTE>2.0.CO;2

Napolitano G, Serinaldi F, See L., 2011. Impact of EMD decomposition and random initialisation of weights in ANN hindcasting of daily stream flow series: an empirical examination. J Hydrol 406:199–214

Osuri, K.K., Nadimpalli, R., Mohanty, U.C., Chen, U., Rajeevan, M., Niyogi, D., 2017. Improved prediction of severe thunderstorms over the Indian Monsoon region using high-resolution soil moisture and temperature initialization. Sci. Rep. 7, 41377. https://doi.org/10.1038/srep4 1377

Pakalidou, N., Karacosta, P., 2018. Study of very long-period extreme precipitation records in Thessaloniki, Greece. Atmospheric Research. 208, 106-115. https://doi.org/10.1016/j.atmosres.2017.07.029

Pham, B.T., Nguyen, M.D., Dao, D.V. et al., 2019. Development of artificial intelligence models for the prediction of Compression Coefficient of soil: An application of Monte Carlo sensitivity analysis. Sci. Total Environ. 679, 172–184. https://doi.org/10.1016/j.scitotenv.2019.05.061

Phuong, N.T.B., Duy, N.B., Nghiem, N.C., 2017. Remote sensing for monitoring surface water quality in the Vietnamese Mekong delta: the application for estimating chemical oxygen demand in river reaches in Binh Dai, Ben Tre. Vietnam Journal of Earth Sciences 39, 256–268. https://doi.org/10.15625/0866-7187/39/3/10270

Pradhan, B., Lee, S., 2010. Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling. Environ. Model Softw. 25, 747–759. https://doi.org/10.1016/j.envsoft.2009.10.016

651 Qiu, G., Wu, Z., Li, Z., Du, Q., 2010. Application of least square support vector machine for thunderstorm
652 prediction. 8th World Congress on Intelligent Control and Automation, Jinan, 2010, pp. 345-349,
653 https://10.1109/WCICA.2010.5555057
654 Rädler, A.T., Groenemeijer, P.H., Faust, E., Sausen, R., Púčik, T., 2019. Frequency of severe thunderstorms
655 across Europe expected to increase in the 21st century due to rising instability. Clim. Atmos. Sci. 2, 30.
656 https://doi.org/10.1038/s41612-019-0083-7
657 Rahman, M.S., Azad, M.A.K., Hasanuzzaman, M., Salam, R., Islam, A.R.M.T., Rahman, M.M., Hoque,
658 M.M.M., 2020. How air quality and COVID-19 transmission change under different lockdown scenarios? A
659 case from Dhaka city, Bangladesh. Sci. Total Environ. 143161. https://doi.org/10.1016/j.scitotenv.2020.143161
660 Rahman, M.S., Islam, A.R.M.T., 2019. Are precipitation concentration and intensity changing in Bangladesh
661 overtimes? Analysis of the possible causes of changes in precipitation systems. Sci. Total Environ. 690, 370–
662 387. https://doi.org/10.1016/j.scitotenv.2019.06.529
663 Saha, T.R., Quadir, D.A., 2016. Variability and trends of annual and seasonal thunderstorm frequency over
664 Bangladesh. Inter. J. Climatol. 36, 4651–4666. https://10.1002/joc.4663
665 Sahu, R.K.; Dadich, J.; Tyagi, B.; Visa, N.K.; Singh, J., 2020. Evaluating the impact of climate change in
666 threshold valuesofthermodynamicindicesduringpre–monsoonthunderstormseasonoverEasternIndia. Nat.
667 Hazards, 102, 1541–1569
668 Salam, R., Islam, A.R.M.T., 2020. Potential of R.T., Bagging and R.S. ensemble learning algorithms for
669 reference evapotranspiration prediction using climatic data-limited humid region in Bangladesh. J. Hydrol. 590,
670 125241. https://doi.org/10.1016/j.jhydrol.2020.125241
671 Shadab, A., Ahmad, S., Said, S., 2020. Spatial forecasting of solar radiation using A.R.I.M.A. model. Remote
672 Sensing Applications: Society and Environment. 20, 100427. https://doi.org/10.1016/j.rsase.2020.100427
673 Siddiqui, Z.A., Rashid, A., 2008. Thunderstorm frequency over Pakistan. Pak. J. Meteorol. 5, 39-63.
674 Solari, G., Rainisio, D. De Gaetano, P., 2017. Hybrid simulation of thunderstorm outflows and wind-excited
675 response of structures. Meccanica. 52, 3197–3220. https://doi.org/10.1007/s11012-017-0718-x

676 Suparta, W., Putro, W.S., 2018. Parametric studies of A.N.F.I.S. family capability for thunderstorm prediction.
677 Space Science and Communication for Sustainability. Springer, Singapore. https://doi.org/10.1007/978-981-10-
678 6574-3_2
679 Tan QF, Lei XH, Wang X et al., 2018. An adaptive middle and long-term runoff forecast model using EEMD-
680 ANN hybrid approach. J Hydrol 567:767–780
681 Taszarek, M., Allen, J., Púčik, T., Groenemeijer, P., Czernecki, B., Kolendowicz, L., Lagouvardos, K., Kotroni,
682 V., Schulz, W., 2019. A climatology of thunderstorms across Europe from a synthesis of multiple data sources.
683 J. Clim. 32 (6), 1813–1837. https://doi.org/10.1175/JCLI-D-18-0372.1.
684 Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res.
685 106, 7183–7192. https://doi.org/10.1029/2000JD900719
686 Trenberth, K.E., Jones, P.D., Ambenje, P. et al., 2007. Observations: Surface and atmospheric climate change.
687 Chapter 3 in S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller
688 (eds.), Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth
689 Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge,
690 United Kingdom and New York, NY, U.S.A.
691 Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer Verlag, New York, U.S.A.
692 https://doi.org/10.1007/978-1-4757-3264-1
693 Vapnik, V., 1998. Statistical learning theory, 1. Wiley, New York.
694 Vujović, D., Paskota, M., Todorović, N., Vučković, V., 2015. Evaluation of the stability indices for the
695 thunderstorm forecasting in the region of Belgrade, Serbia. Atmospheric Research, 161–162, 143-152.
696 https://doi.org/10.1016/j.atmosres.2015.04.005
697 Wahiduzzaman, M., Islam, A.R.M.T., Luo, J., Shahid S, Uddin, M.J., Shimul, S.M., Sattar, M.A., 2020. Trends
698 and Variabilities of Thunderstorm Days over Bangladesh on the ENSO and IOD Timescales. Atmosphere. 11,
699 1176. https://doi.org/10.3390/atmos11111176
700 Wang WC, Chau KW, Xu DM, Chen X.Y., 2015. Improving forecasting accuracy of annual runoff time series
701 using ARIMA based on EEMD decomposition. Wat Resour Manage 29:2655–2675

Wang, J., Xu, X., Meng, X., 2015. The modified ensemble empirical mode decomposition method and extraction of oceanic internal wave from synthetic aperture radar image. Journal of Shanghai Jiaotong University (Science). 20(2), 243–250. https://10.1007/s12204-015-1614-y

Wu, Z., Huang, N.E., 2009. Ensemble empirical mode decomposition: a noise-assisted data analysis method. Advances in Adaptive Data Analysis 1, 1–41. https://doi.org/10.1142/S1793536909000047

Yasen, M., Al-Jundi, R., Al-Madi, N., 2017. Optimized ANN-ABC for Thunderstorms Prediction, IEEE, 98-103, DOI: 10.1109/ICTCS.2017.37

Umakanth, N., Satyanarayana, G.C., Simon, B., Rao, M.C., Babu, N.R., 2020. Long-term analysis of thunderstorm-related parameters over Visakhapatnam and Machilipatnam, India, Acta Geophysica https://doi.org/10.1007/s11600-020-00431-2

Yu, M., 2020. Short-term wind speed forecasting based on random forest model combining ensemble empirical mode decomposition and improved harmony search algorithm, International Journal of Green Energy, 17:5, 332-348, DOI: 10.1080/15435075.2020.1731816

Zhang X, Peng Y, Zhang C, Wang, B., 2015. Are hybrid models integrated with data pre-processing techniques suitable for monthly streamflow forecasting? Some experiment evidences. J Hydrol 530:137–152

Zhen-hui, W., Yi, Z., Jia, Z., 2013. A preliminary study on thunderstorm forecast with LS-SVM method. Journal of Tropical Meteorology. 19(1), 104-108.

Zhou, Q., Jiang, H., Wang, J., Zhou, J., 2014. A hybrid model for PM 2.5 forecasting based on ensemble empirical mode decomposition and a general regression neural network. Science of The Total Environment. 496, 264–274. https://doi.org/10.1016/j.scitotenv.2014.07.051

Figures

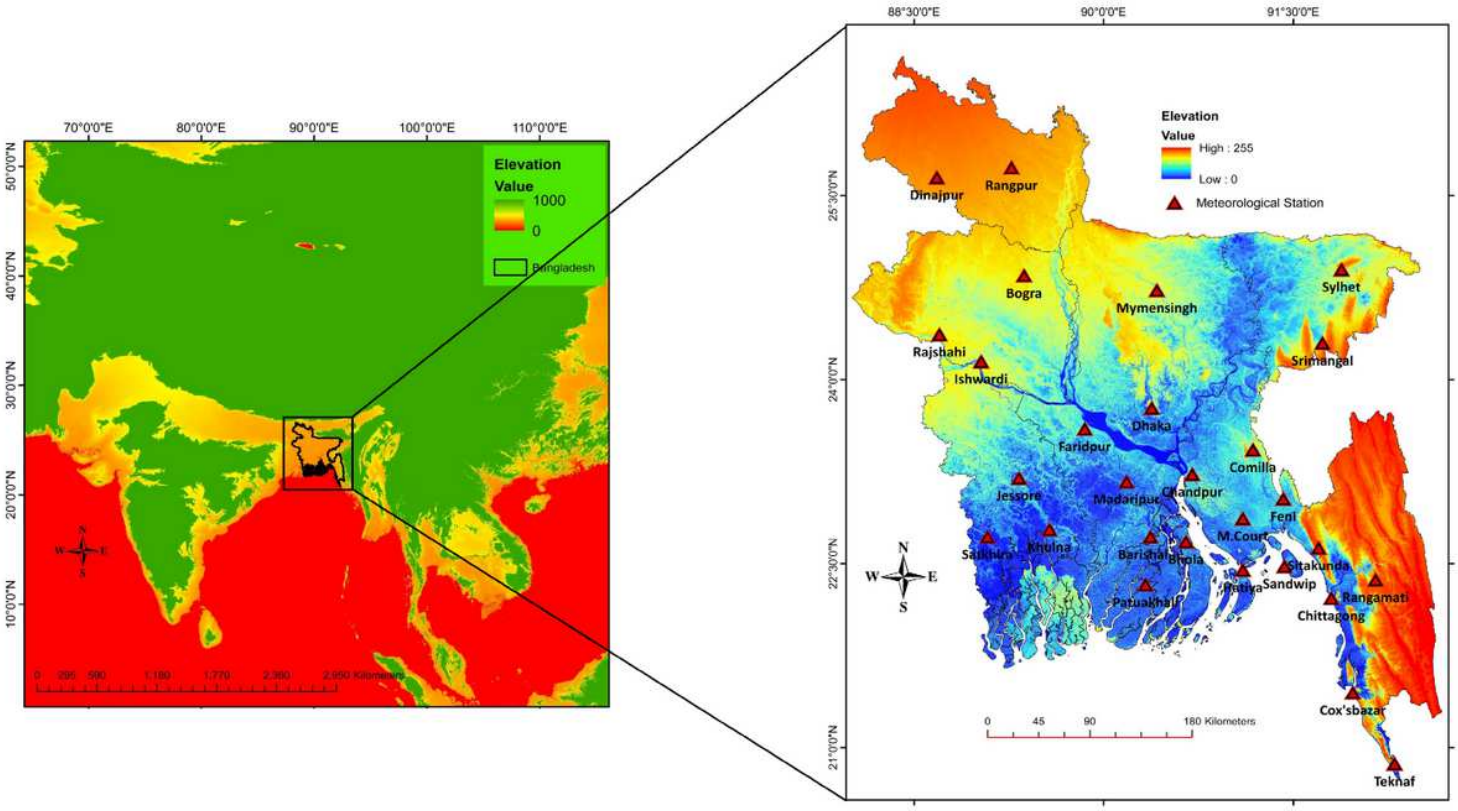

Figure 1

Geographical location of the study area, red delta signs represent the selected meteorological stations of BMD. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

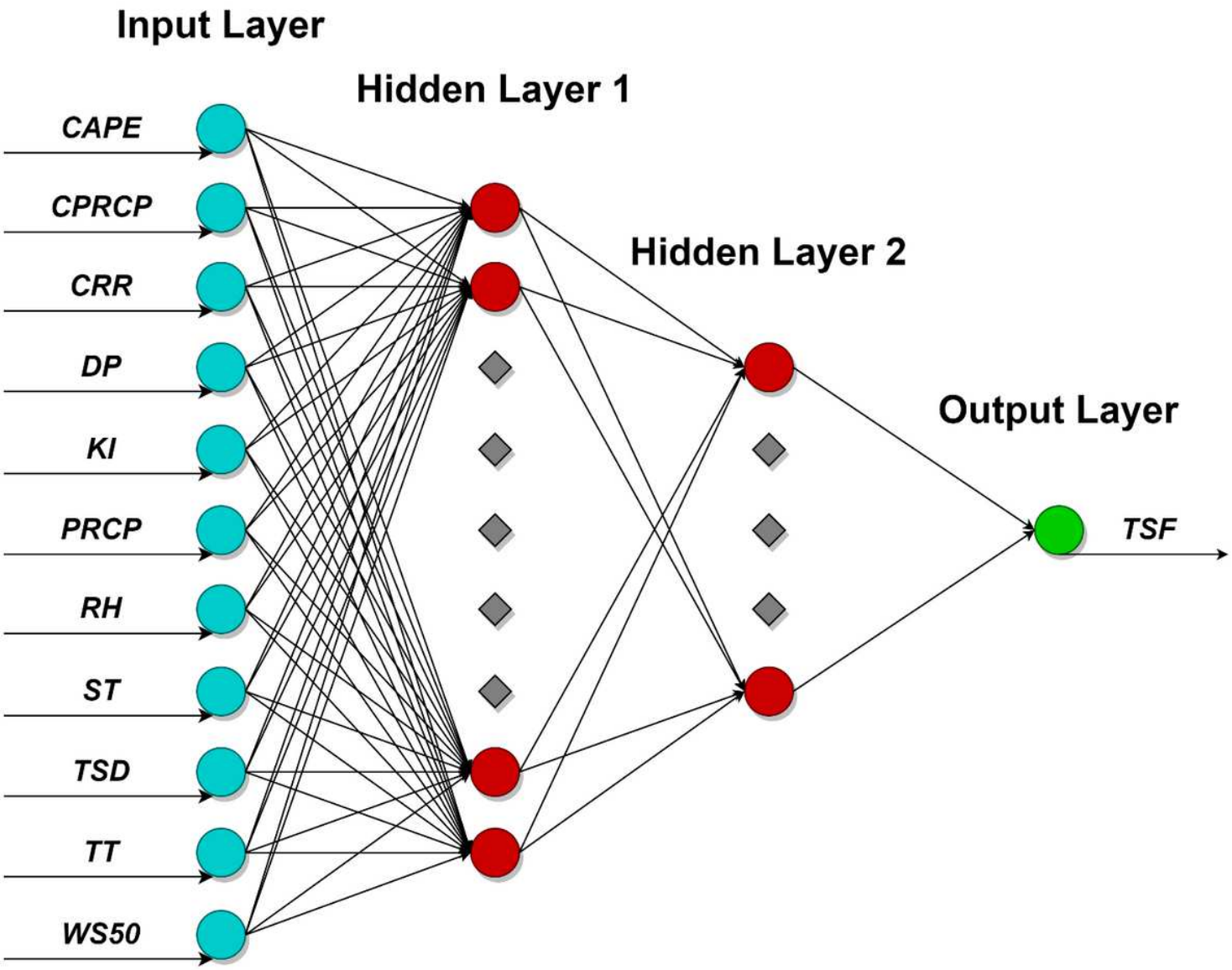

Figure 2

Architecture of the artificial neural network (ANN) model with an input layer, two hidden layers, and an output layer used for predicting monthly thunderstorm frequency.

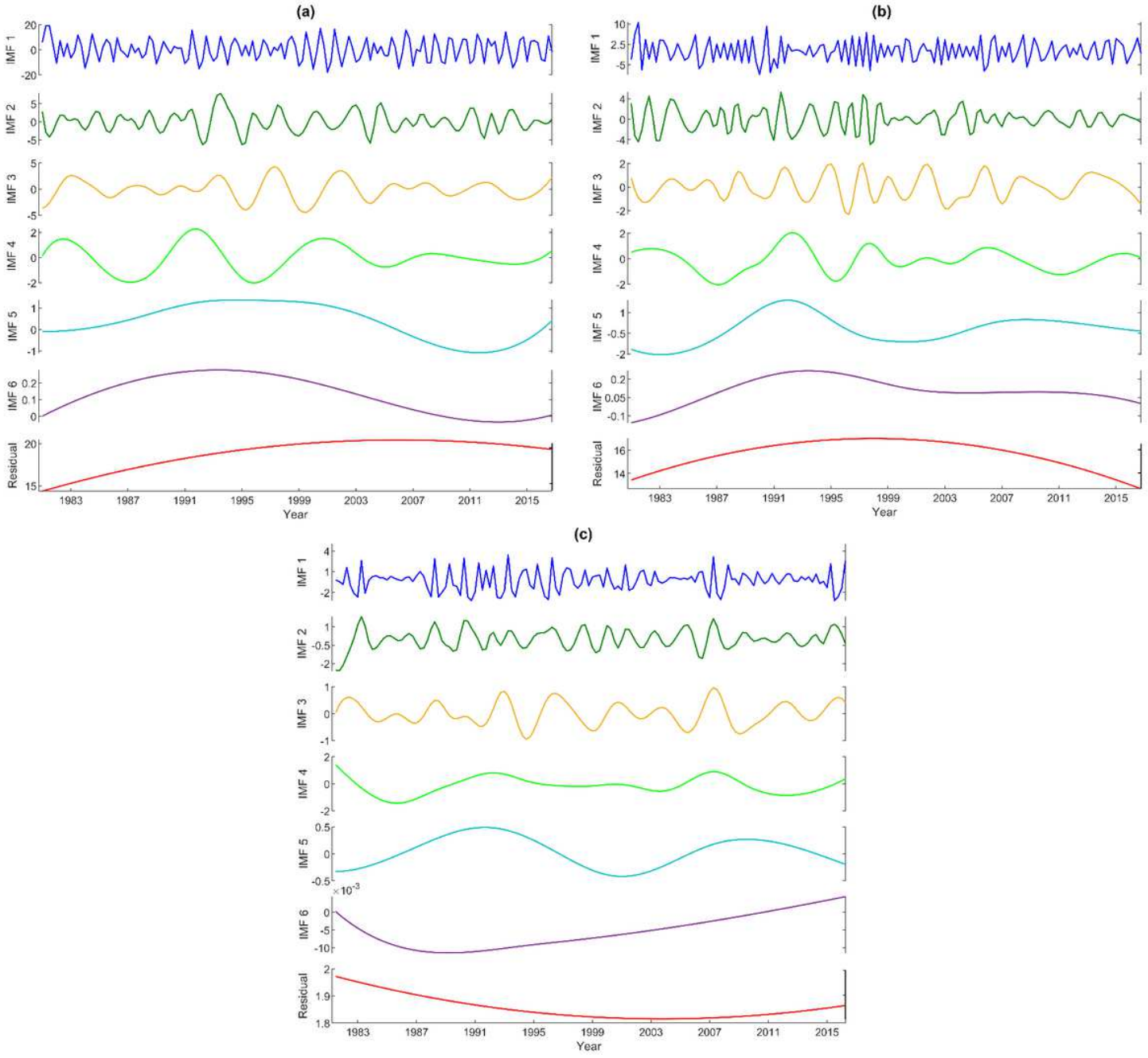

Figure 3

The decomposed sub-series of the original TSF data for high-frequency months (a), moderate-frequency months (b), and low-frequency months series (c) using EEMD.

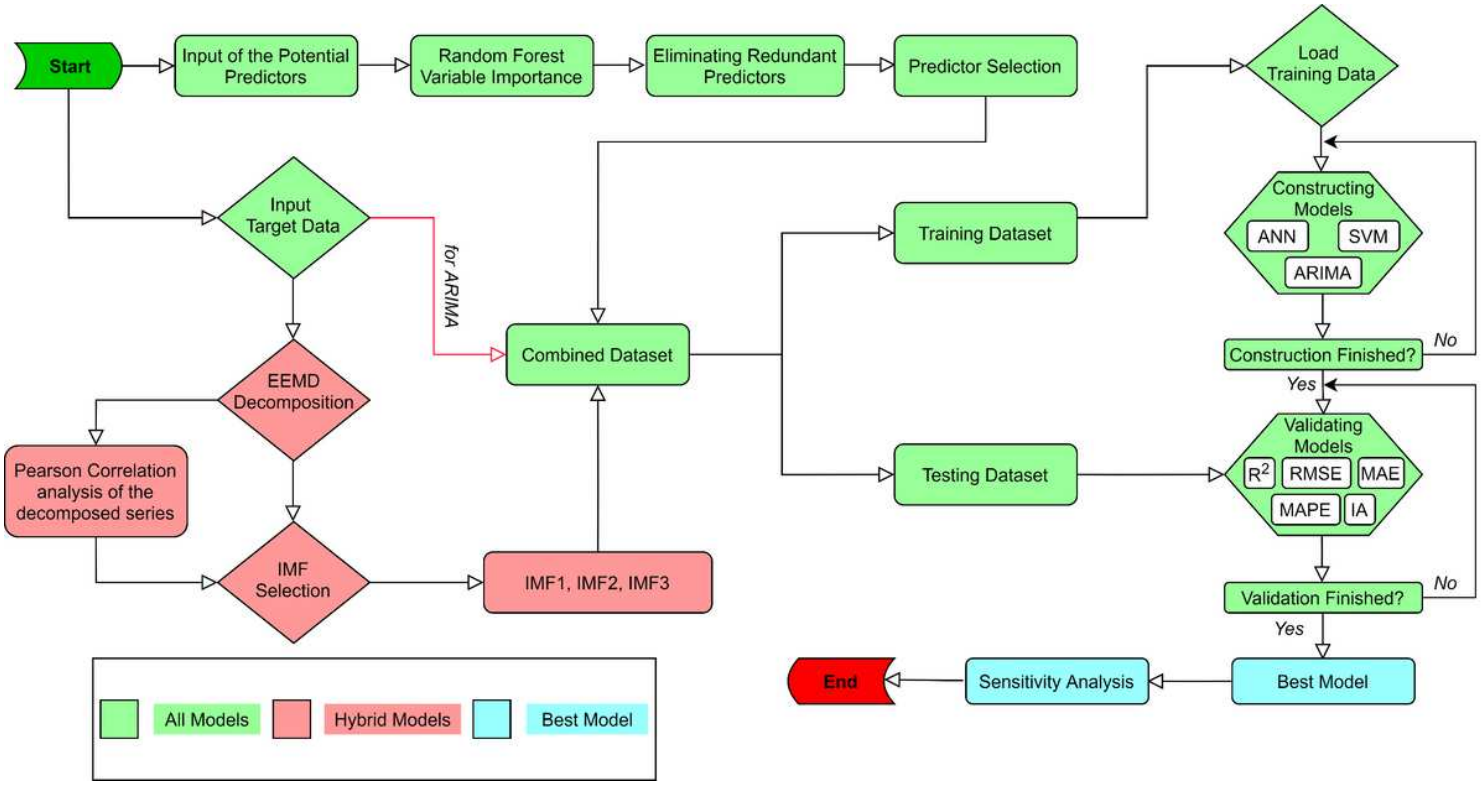

Figure 4

Flow diagram of the methodological processes.

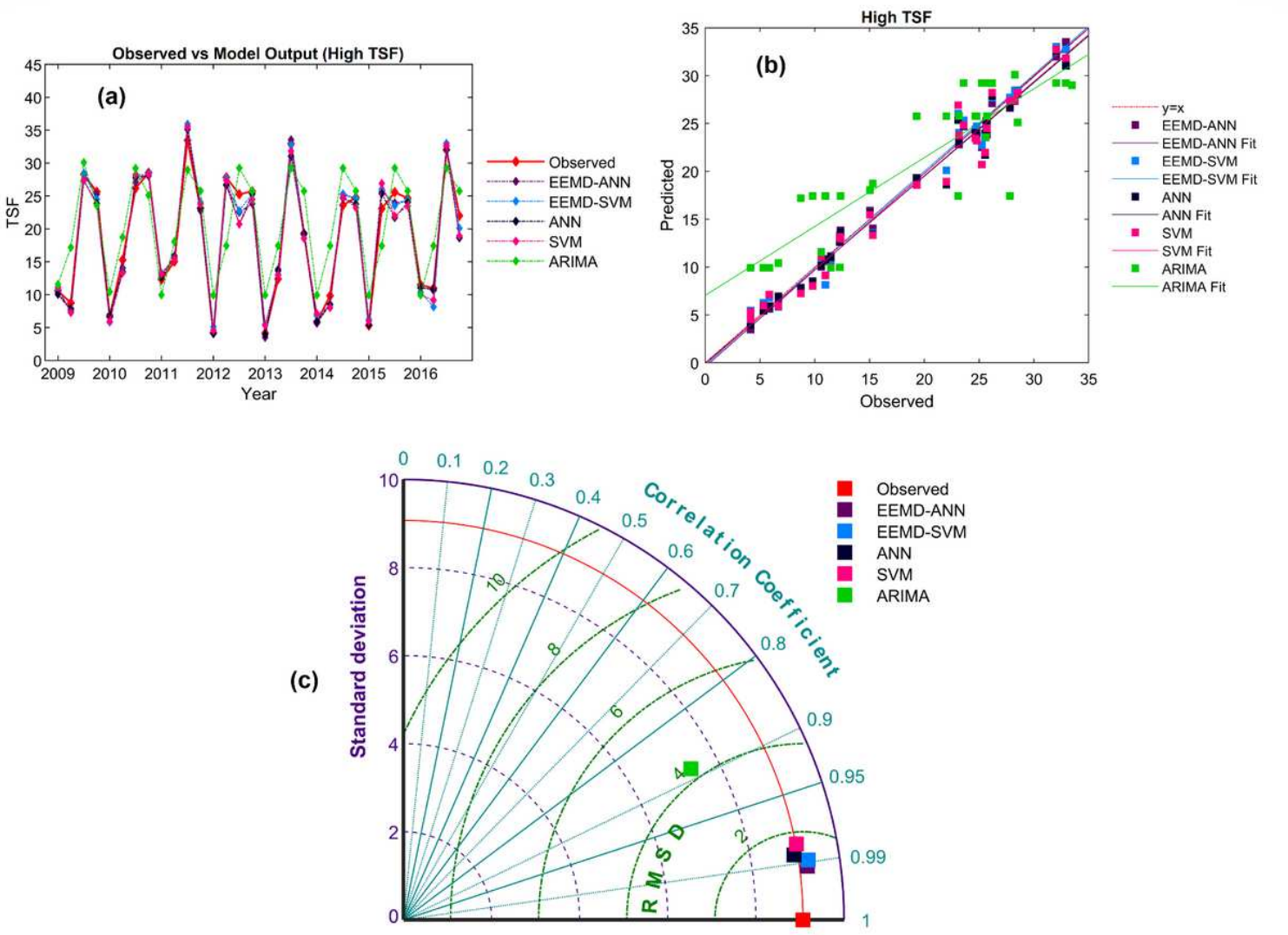

Figure 5

Observed and predicted output of the testing dataset of the high TSF series (a); scatterplot fitting of the prediction models (b); and Taylor Diagram of prediction by EEMD-ANN, EEMD-SVM, ANN, SVM, and ARIMA models (c). The deep cyan contours represent the Pearson correlation coefficient, green contours represent centered RMS error in the simulated field, and violet contours represent the Standard Deviation of the simulated pattern.

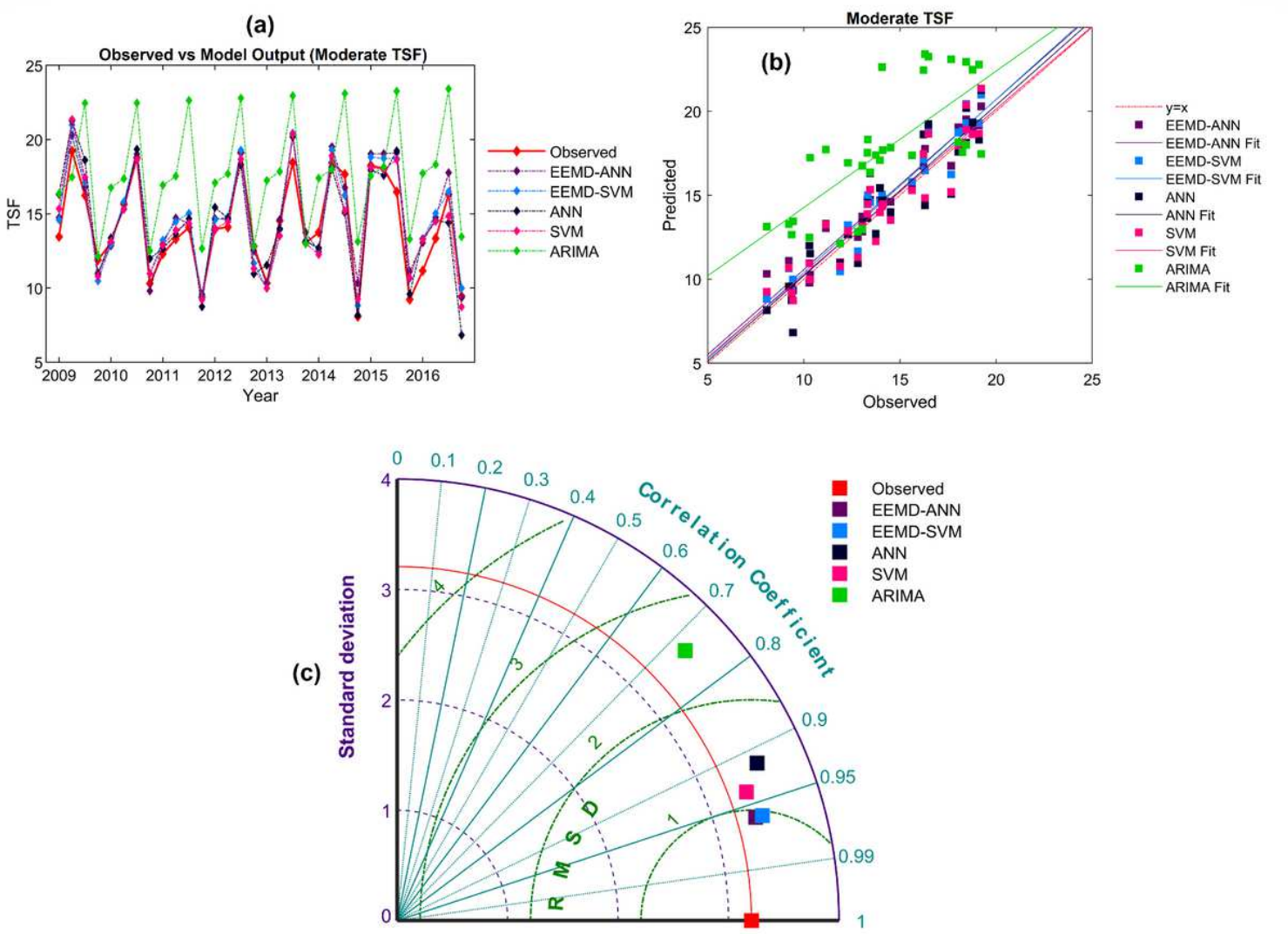

Figure 6

Observed and predicted output of the testing dataset of the moderate TSF series (a); scatterplot fitting of the prediction models (b); and Taylor Diagram of prediction by EEMD-ANN, EEMD-SVM, ANN, SVM, and ARIMA models (c). The deep cyan contours represent the Pearson correlation coefficient, green contours represent centered RMS error in the simulated field, and violet contours represent the Standard Deviation of the simulated pattern.

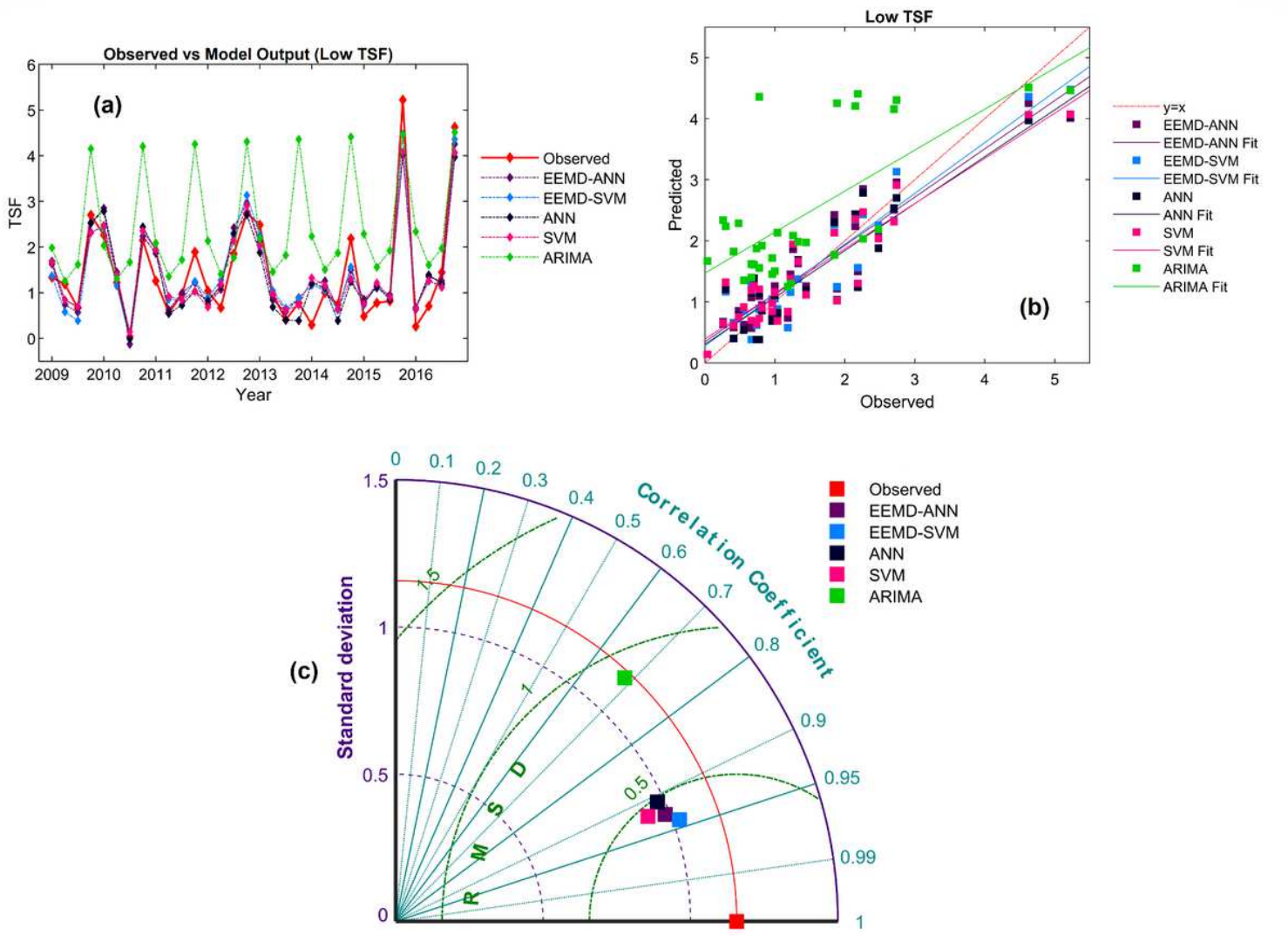

Figure 7

Observed and predicted output of the testing dataset of the low TSF series (a); scatterplot fitting of the prediction models (b); and Taylor Diagram of prediction by EEMD-ANN, EEMD-SVM, ANN, SVM, and ARIMA models (c). The deep cyan contours represent the Pearson correlation coefficient, green contours represent centered RMS error in the simulated field, and violet contours represent the Standard Deviation of the simulated pattern.

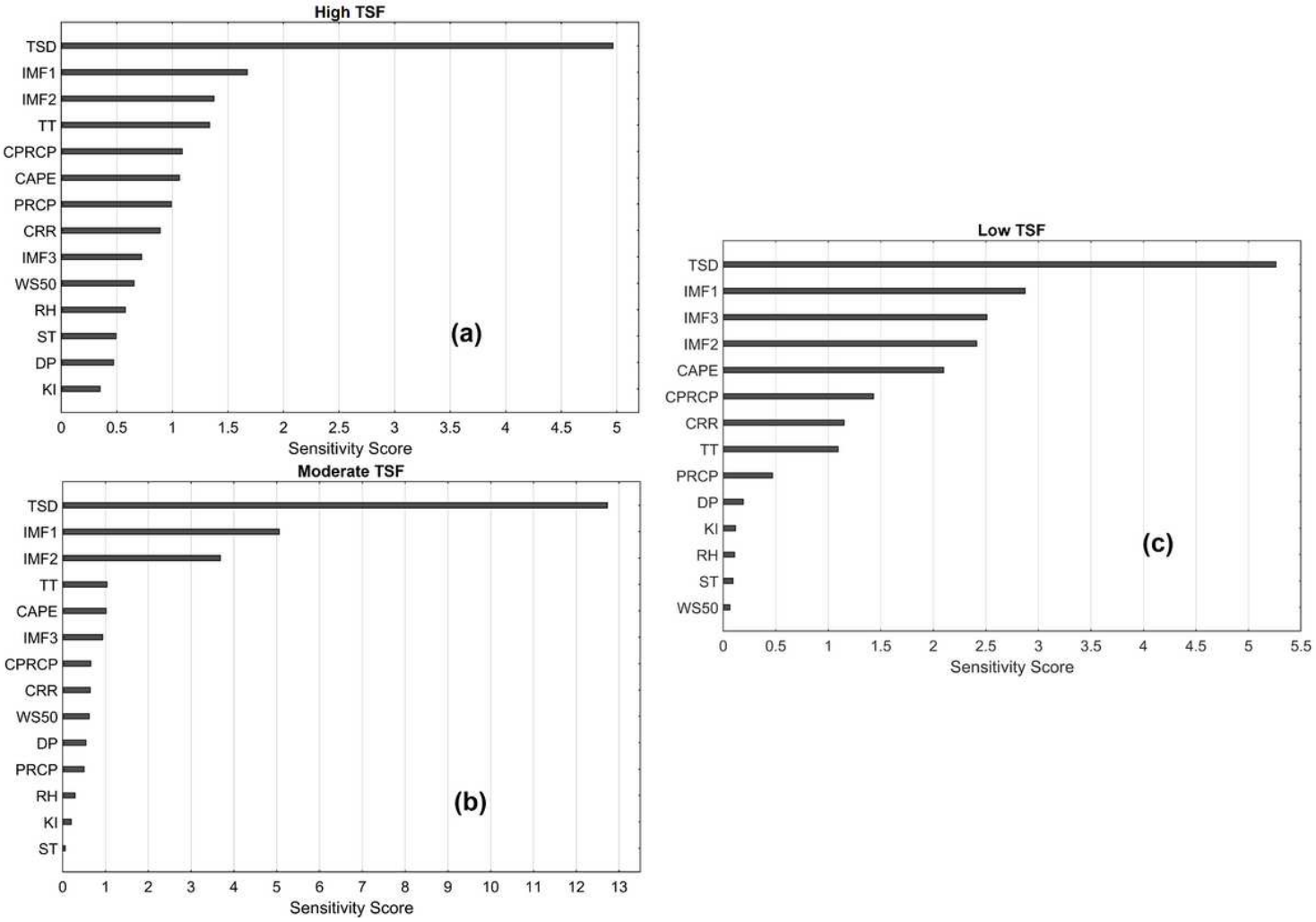

Figure 8

Sensitivity of the input parameters in building the best models in predicting high TSF (a), moderate TSF (b), and low TSF (c).