Development of Dissolved Oxygen Forecast Model Using Hybrid Machine Learning Algorithm with Hydro-Meteorological Variables

Abul Abrar Masrur Ahmed (masrur@outlook.com.au)
University of Southern Queensland https://orcid.org/0000-0002-7941-3902
M A I Chowdhury
Shahjalal University of Science and Technology School of Applied Science and Technology
Oli Ahmed
Leading University Department of Civil Engineering
Ambica Sutradhar
Leading University Department of Civil Engineering

Research Article

Keywords: Dissolved Oxygen, MARS, MODWT, Surma River, Bangladesh

Posted Date: December 9th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1100147/v1

Development of dissolved oxygen forecast model using hybrid machine learning algorithm with hydro-meteorological variables

A. A. Masrur Ahmed*:

School of Modern Sciences, Leading University, Sylhet-3112, Bangladesh.

And

School of Sciences, University of Southern Queensland, Springfield, QLD 4300, AUSTRALIA. Email: AbulAbrarMasrur.Ahmed@usq.edu.au; masrur@outlook.com.au

M A I Chowdhury:

Department of Civil and Environmental Engineering, Shahjalal University of Science and Technology, 3114, Bangladesh. Email: maiccee@gmail.com

Oli Ahmed:

School of Modern Sciences, Leading University, Sylhet-3112, Bangladesh.

Email: oliahmed3034@gmail.com

Ambica Sutradhar:

School of Modern Sciences, Leading University, Sylhet-3112, Bangladesh.

Email: ambicasutradhar@gmail.com

Corresponding Author Email: masrur@outlook.com.au

Abstract:

The ability to predict dissolved oxygen, which is a critical water quality (WQ) parameter, is critical for aquatic managers responsible for maintaining ecosystem health and the management of reservoirs affected by WQ. This paper reports forecasting dissolved oxygen (DO)

concentration using multivariate adaptive regression splines (MARS) of running river water using a set of water quality and hydro-meteorological variables. This study's key objectives were to assess input selection methods and five multi-resolution analyses as a data extraction approach. Moreover, the hybrid model is prepared by maximum overlap discrete wavelet transformation (MODWT) with the MARS model (*i.e.,* MODWT-MARS). The proposed model is further compared with numerous machine learning methods. The result shows that the hybrid algorithms (i.e., MODWT-MARS) outperformed the other models (r = 0.981, WI = 0.990, $RMAE$ = 2.47% and MAE = 0.089). This hybrid method may serve as the foundation for forecasting water quality variables with fewer predictor variables.

Keywords: Dissolved Oxygen, MARS, MODWT, Surma River, Bangladesh

Nomenclature

ACF	Auto Correlation Function
ANN	Artificial Neural Network
BF	Basis Functions
BNR	Bayesian Ridge Regression
BOD	Biological Oxygen Demand
COD	Chemical oxygen demand
CCF	Cross Correlation Function
CEEMDAN	Complete ensemble empirical mode decomposition with Adaptive Noise
CEEMDAN-BNR	Hybrid Model integrating the CEEMDAN algorithm with BNR
CEEMDAN-KNN	Hybrid Model integrating the CEEMDAN algorithm with KNN
CEEMDAN-KRR	Hybrid Model integrating the CEEMDAN algorithm with KRR
CEEMDAN-MARS	Hybrid Model integrating the CEEMDAN algorithm with MARS
CEEMDAN-RF	Hybrid Model integrating the CEEMDAN algorithm with RF
CEEMDAN-SVR	Hybrid Model integrating the CEEMDAN algorithm with SVR
EEMD	Ensemble empirical mode decomposition

EEMD-BNR	Hybrid Model integrating the EEMD algorithm with BNR
EEMD-KNN	Hybrid Model integrating the EEMD algorithm with KNN
EEMD-KRR	Hybrid Model integrating the EEMD algorithm with KRR
EEMD-MARS	Hybrid Model integrating the EEMD algorithm with MARS
EEMD-RF	Hybrid Model integrating the EEMD algorithm with RF
EEMD-SVR	Hybrid Model integrating the EEMD algorithm with SVR
EMD	Empirical Mode Decomposition
EMD-BNR	Hybrid Model integrating the EMD algorithm with BNR
EMD-KNN	Hybrid Model integrating the EMD algorithm with KNN
EMD-KRR	Hybrid Model integrating the EMD algorithm with KRR
EMD-MARS	Hybrid Model integrating the EMD algorithm with MARS
EMD-RF	Hybrid Model integrating the EMD algorithm with RF
EMD-SVR	Hybrid Model integrating the EMD algorithm with SVR
DWT	Discrete wavelet Transformation
DWT-BNR	Hybrid Model integrating the DWT algorithm with BNR
DWT -KNN	Hybrid Model integrating the DWT algorithm with KNN
DWT -KRR	Hybrid Model integrating the DWT algorithm with KRR
DWT -MARS	Hybrid Model integrating the DWT algorithm with MARS
DWT -RF	Hybrid Model integrating the DWT algorithm with RF
DWT -SVR	Hybrid Model integrating the DWT algorithm with SVR
ECDF	Empirical Cumulative Distribution Function
ELM	Extreme Learning Machine
FE	Forecasting Error
GCV	Generalised Cross-Validation
IMF	intrinsic mode functions
KNN	K- Nearest Neighbourhood
KRR	Kernel Ridge Regression
LM	Legates-McCabe’s Index
LSSVM	Least Square Support Vector Machine

MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MARS	Multivariate adaptive regression splines
MLP	Multi-Layer Perceptron
MODWT	Maximum Overlap Discrete Wavelet Transformation
MODWT -BNR	Hybrid Model integrating the MODWT algorithm with BNR
MODWT -KNN	Hybrid Model integrating the MODWT algorithm with KNN
MODWT -KRR	Hybrid Model integrating the MODDWT algorithm with KRR
MODWT -MARS	Hybrid Model integrating the MODWT algorithm with MARS
MODWT -RF	Hybrid Model integrating the MODWT algorithm with RF
MODWT -SVR	Hybrid Model integrating the MODWT algorithm with SVR
MRA	Multi-resolution Analysis
MSE	Mean Squared Error
NCA	Neighbourhood Component Analysis
NSE	Nash–Sutcliffe Efficiency
PACF	Partial Auto-Correlation Function
r	Correlation Coefficient
RBF	Radial Basis Function
RF	Random Forest
RMSE	Root-Mean-Square-Error
RRMSE	Relative Root-Mean-Square Error
SVR	Support Vector Regression
TDS	total dissolved solids
WQ	Water Quality

Introduction:

The deterioration of the quality of water sources throughout the world is considered a worldwide issue of importance. Because of the rapid rise of communities and the diversity of

their activities, this deterioration is speeding up, and it could constitute a serious threat to the aquatic environment and human health (Henderson et al., 2009; Hur and Cho, 2012; Mouri et al., 2011; Su et al., 2011). It is widely recognised that the degradation of the quality of water resources across the world is a global issue of critical importance. The rapid growth of communities and the diversity of their activities contribute to the acceleration of this deterioration, which may eventually pose a significant threat to the aquatic environment and human health. The decline of the quality of water bodies throughout the world is considered a worldwide issue of importance. Because of the rapid rise of populations and the diversity of their activities, this deterioration is speeding up. It has the potential to pose a significant hazard to the aquatic environment as well as public safety.

The dissolved oxygen (DO) in water is a critical water quality variable that is crucial for the proper functioning of the aquatic ecosystem (Ranković et al., 2010). DO demonstrate the water pollution in rivers (Heddam and Kisi, 2018; Mohan and Kumar, 2016) and the state of the river's ecosystems (Mellios et al., 2015; Ranković et al., 2010). The concentration of dissolved oxygen (DO) in aquatic systems refers to the metabolism of the aquatic systems, and it reflects the transient balance between the oxygen system and the metabolic activity. The concentration of DO is affected by a variety of parameters, including salinity, temperatures, and pressure (US-Geological-Survey, 2016). Researchers investigated the concentration and change of DO over the last decade since the dynamics of DO are nonlinear and many models are not relevant (Kisi et al., 2020).

There are various methods available for estimating the DO concentration, but most of them are time-consuming and expensive to use since they require numerous parameters that are not readily available in most cases (Suen and Eheart, 2003). More to the point, conventional data processing techniques are no longer appropriate for water quality modelling, which may be linked to the explanation that many parameters affecting water quality have a complicated nonlinear interaction with one another (Ahmed, 2017; Xiang et al., 2006). There are specific issues in developing a water quality model for tiny streams or rivers due to the lack of available

data, a lack of investment, and a large number of different inputs to consider. As a result, certain well-known water quality analysis models, such as the United States Environmental Protection Agency (USEPA): QUAL2E and QUAL2K, WASP6, require a great deal of information that is not always readily available (Ahmed, 2017). Moreover, these models are complex and sensitive and, therefore, tough to recognise.

Machine learning-based data-driven algorithms have become potentially popular in the field of water quality modelling (Ahmed, 2017; Ahmed and Shah, 2017a; Forough et al., 2019; Kuo et al., 2004; Tomic et al., 2018) and hydrological modelling (Ahmed et al., 2021d; Ahmed and Shah, 2017b; Ali et al., 2018; Deo and Şahin, 2015a; Deo and Şahin, 2015b; Prasad et al., 2018; Yaseen et al., 2016). Notably, they have been effectively demonstrated to forecast hydro-metrological variables, *e.g*., precipitation (Akbari Asanjan et al., 2018; Hamidi et al., 2015; Nguyen-Huy et al., 2017; Tripathi et al., 2006; Yang et al., 2018; Yoo and Cho, 2018), drought index (Deo and Şahin, 2015b), streamflow (Ahmed and Shah, 2017b; Deo and Sahin, 2016; Prasad et al., 2017), runoff (Hu et al., 2018; Kratzert et al., 2018), flood (Arto et al., 2019; Chen et al., 2014; Elsafi, 2014; Le et al., 2019), evapotranspiration (Ahmed et al., 2021a) and soil moisture (Ahmed et al., 2021b; Ahmed et al., 2021d). In addition, a number of artificial intelligence (AI)-based models for predicting and estimating DO concentrations have been developed, soft computing (Tao et al., 2019), artificial neural networks (ANNs), and hybrid ANN (Keshtegar et al., 2019; Zounemat-Kermani et al., 2019), fuzzy-based models (Heddam, 2017; Raheli et al., 2017), support vector machine (SVM) (Heddam and Kisi, 2018), extreme learning machine (ELM) (Heddam, 2016; Heddam and Kisi, 2017; Zhu and Heddam, 2020), and other potential approaches were applied for dissolved oxygen concentration modelling.

This study investigates the utilisation of multivariate adaptive regression splines (MARS) (Friedman, 1991) to describe DO dynamics' intrinsic nonlinear and multidisciplinary relationship. Like neural networks, no prior information on the form of the numerical function is required for MARS. MARS's benefits are that it can accomplish complex data by grouping related data collected, permitting it to understand easily (Zhang and Goh, 2016). Considering the positive attribute, the MARS model has been used in hydrology (Deo et al., 2017b; Heddam

and Kisi, 2018; Kisi and Parmar, 2016; Yin et al., 2018) and the energy sector (Al-Musaylh et al., 2019). Heddam and Kisi (2018) applied the least-square support vector machine (LSSVM), multivariate adaptive regression splines, and M5 model tree (M5T) for daily dissolved oxygen forecasting. The authors found the MARS model a substantial forecasting approach with a limited number of fo predictor variables. Therefore, incorporating the hybrid approaches and a potential feature selection algorithm may increase the result of forecasting. Nevertheless, the hybrid MARS models are yet to be executed in the study sites of Bangladesh.

Using multi-resolution analysis (MRA), a technique for extracting data features, the prediction performance can be enhanced significantly. Using the EMD, you can decompose a signal in accordance with the spirit of the Fourier series into a specific number of components. A coefficient representing Gaussian white noise with a unit variance is introduced sequentially to the time series in CEEMDAN-based decomposition to reduce the complexity and avoid the intricacy of the time series (Prasad et al., 2018). A coefficient denoting Gaussian white noise with covariance matrices is introduced sequentially to the time series in CEEMDAN-based decomposition to reduce the complexity and prevent the intricacy of the time series (Di et al., 2014). Previous studies have used CEEMDAN in forecasting soil moisture (Ahmed et al., 2021c; Prasad et al., 2018; Prasad et al., 2019) with an earlier version (*i.e.*, EEMD) used in forecasting streamflow (Seo and Kim, 2016) and rainfall (Beltrán-Castro et al., 2013; Jiao et al., 2016; Ouyang et al., 2016). Discrete wavelet transform (DWT) has been employed (Deo and Sahin, 2016; Deo et al., 2016; Nourani et al., 2014; Nourani et al., 2009) in different fields of hydrology. On the other hand, DWT has a limitation that prevents it from extracting all of the features of the predictors in its entirety. An enhanced discrete wavelet transform (DWT), such as the MODWT, can solve these problems (Cornish et al., 2006; Prasad et al., 2017; Rathinasamy et al., 2014). Al-Musaylh et al. (2020) successfully used MODWT to decompose the short-term electricity of Australia. The study incorporated the MODWT by separately splitting the data to training, testing, and validation to calculate the detailed approximation, as Quilty and Adamowski (2018) prescribed. The potential application of MODWT is further approved by Prasad et al. (2017) where they used MODWT to forecast streamflow. However,

neither the MODWT nor the DWT decomposition model has incorporated the MARS model in DO forecasting, as attempted in this study.

Uses a possible feature-selection technique in this investigation (i.e., NCA). As a result of the algorithm being slowed down by the extraneous and redundant features, the prediction model is less accurate (Arhami et al., 2013). Different feature selection methods have been utilised in predictive models (Ahmed et al., 2021c; Prasad et al., 2017; Prasad et al., 2019). The NCA method has been successfully applied by Ahmed et al. (2021c) to forecast surface soil moisture. The study demonstrates that the feature weight calculated by NCA was found successful in forecasting soil moisture and to the study by Ghimire et al. (2019b), where they applied NCA for solar radiation forecasting. Forecasting DO concentration with a machine learning method incorporated with the NCA feature selection method, and feature extraction methods would substantially increase forecasting performance.

The objectives of our study are (1) to build a machine learning predictive approach to forecast the DO concentration, incorporating MODWT (*i.e*., data decomposition) with NCA (*i.e*., a feature selection method) to produce a MARS-based forecasted model, (2) The MARS model was compared to other well-known machine learning approaches, such as Bayesian ridge regression (BNR), kernel ridge regression (KRR), support vector regression (SVR), and random forest (RF) models, to examine its forecasting capacity, (3) To explore the data decomposition ability of MODWT compared with other MRA methods (*i.e.,* EMD, DWT, EMD, and CEEMDAN).

2. Materials and methods

2.1 Theoretical frameworks of proposed models

2.1.1 Multivariate adaptive regression spline (MARS)

According to Friedman (1991), a non-parametric and nonlinear regression technique, the MARS, was utilised in this investigation. MARS uses numerous splines to build various nodes between these lines (Friedman, 1991). The underlying functional link between inputs and outputs is not assumed in the MARS model. The data in each spline is assigned using Basis

Functions (BF) in MARS models. It is possible to express the BF as a single equation between two knots. Two adjacent data domains converge at a knot, and the output is continuous. An adaptive regression algorithm is used (Heddam and Kisi, 2018). Using numerous lines, the MARS model depicts the piecewise relationship between the input and output variables. The over-fitting of training data is avoided by setting a predefined minimum number of observations between knots (Heddam and Kisi, 2018).

Let y be the target output, and a matrix of n input variables be the vector x = $(x_1, \dots, x_n)$. The data are then presumed to be created from an undisclosed "true" model. In the case of a straight answer, this will be as follows:

$$y = f(x_1, \dots., x_n) + \xi = f(x) + \xi$$

(1)

In which ξ is the distribution of the model error, and n is the number of training data points. By adding sufficient BFs, MARS approximates the f(.). For linear functions piecewise: max (*0, x-t*) where a knot exists at position *t* (Zhang and Goh 2013). The max (.) equation implies that only the positive portion of (.) is used; otherwise, a zero value will be given corresponding to:

$$\text{max (0, x-t)} = \begin{cases} x - t, & if\ x \geq t \\ 0, & otherwise \end{cases}$$

(2)

Thus, f (x) is constructed as a linear BF(x) combination:

$$f(\text{x}) = \beta_0 + \sum_{i=1}^{n} \beta_i\, BF(x)$$

(3)

The coefficients β are constants, calculated using the form of least-squares. Initially, f (x) is applied to input data in a forward-backwards stepwise process to determine the knot's position where the feature value varies (Deo et al., 2017b). A broad model is built at the end of the forwards' stage to over-fit the qualified input data. According to the generalised cross-validation, the model is optimised by deleting one last basis function from the model (GCV). GCV for a model is computed as follows for the training data with n observations:

$$\mathrm{GCV} = \frac{\frac{1}{n}\sum_{i=1}^{n}[y_i - f(x_i)]^2}{\left[1 - \frac{M + d \times (M-1)/2}{n}\right]^2}$$

(4)

Where M is the number of BF, d is the penalising parameter, n is the number of measurements, and $f(x_i)$ denotes the MARS model's expected values.

2.1.5 Maximal overlap discrete wavelet transforms (MODWT)

Distinctive wavelet transforms (DWTs) are modified by the maximal overlap discrete wavelet transform (MODWT) (Li et al., 2017). Ideally, time series analysis can be done using the MODWT's appealing qualities, which prevent missing data without subsampling. MODWT's ability to extract additional information is enhanced because the coefficients of decomposed components in each layer are identical to the original time series. Time-series data are broken down into high-pass and low-pass filters using MODWT, which handles two feature sets. Further, high-pass filters can be broken down into several information levels depending on the suitable time frame (He et al., 2017). Low-pass filters reflect the real-time-series signal pattern called an approximation. The signal $\varkappa_m$ is decomposed through wavelet low-pass ι_m and high-pass detail filters h_m and reconstructed by digital reconstruction filters complementing decomposition filters. This principle is described in the equations below:

$$\mathcal{X}_{m+1}(K) = \sum_p h_{p-2k}\mathcal{X}_m(P)$$

(5)

$$d_{m+1}(K) = \sum_p l_{p-2k}\mathcal{X}_m(P)$$

(6)

$$\mathcal{X}_m(K) = \sum_p h'_{p-2k}\mathcal{X}_{m+1}(P) + \sum_p l'_{p-2k}\mathcal{X}_{m+1}(P)$$

(7)

2.2 Comparing models

Breiman (2001) created an RF-based on a random forest (RF), which included methods for regression and classification. The bootstrap resampling procedure generates a new set of training data from the initial training sample set N, and then bootstrap-set random forests are

built using K decision trees. The RF model's full specifications may be read here (Ali et al., 2020).

A machine learning kernel method known as SVR (Support Vector Regression) can be used for various purposes, including forecasting time series. SVRs that use kernels can also learn the nonlinear trend of the training data. There are three SVR models to pick from, each with a different kernel (RBF, poly, and linear) (Yang et al., 2017).

With regularisation and the kernel technique, it is possible to reduce overfitting using the KRR (Kernel Ridge Regression) regression model (Saunders et al., 1998). The "kernel technique" can be used to generate a nonlinear form of ridge regression. Extending the general framework, kernel ridge regression allows nonlinear prediction. Linear, polynomial and Gaussian kernels are only some of the many options available for enhancing overall performance (You et al., 2018).

The Bayesian modelling approach uses hierarchical data (Huang and Abdel-Aty, 2010). Bayesian regression uses this regularisation parameter, which is easy to tailor to the data. The Gaussian maximum posterior estimate is discovered before the coefficient w and, with an accuracy of λ (-1), is treated as a random variable instead of a lambda.

K-Nearest Neighbors (KNN) algorithm is implemented using instance-based learning, which serves two purposes: 1) estimating the test data density function and 2) categorising the test data obtained from the test patterns (Shabani et al., 2020). Choosing the number of neighbours (k) is a crucial stage. This method's efficiency depends on selecting samples from the nearest reference database (or most similar). If k is significant, other points from other classes can be placed inside the desired range of possibilities (Wu et al., 2008).

This study incorporated four more decomposition methods and machine learning methods (i.e., DWT, EMD, EEMD, and CEEMDAN). Hyperspectral feature extraction is DWT-assisted, and the features are evaluated for their efficacy in discriminating between subtly different ground covers (Bruce et al., 2002). The theoretical explanation of the method is explained by other researchers (Agbinya, 1996; Fowler, 2005; Shensa, 1992). Most recently, Huang et al. (Huang et al., 1998) developed an empirical mode decomposition (EMD) method

for analysing the information contained in data derived from non-stationary and nonlinear systems. This algorithm decomposes the signal into a series of oscillatory functions that are "well-behaved," which are referred to as the intrinsic-mode functions in this context (IMFs). When used with the powerful adaptive EMD tool, it behaves as a dyadic filter bank (Flandrin et al., 2004). It is handy for filtering out noise in the measurement domains (Khaldi et al., 2008). Torres et al. (2011) implemented the CEEMDAN process to reduce the computational cost and retain the ability to eliminate mode mixing. The readers are requested to go through the previous studies (Ahmed et al., 2021c; Zhang et al., 2017; Zhou et al., 2019) for getting further information on CEEMDAN.

2.2 Study area and data

The Surma River, Bangladesh, provided daily water quality factors. Figure 01 depicts the Surma River monitoring stations. This river drains one of the heaviest runoffs in the Surma-Meghna system (Chowdhury and Ali, 2006). The Surma River originates in Assam's Cachar district, flows through Bangladesh's Sylhet and Sunamganj districts, joins the Meghna River near Bhairab Bazar Kishoreganj, and empties into the Bay of Bengal. Many studies found regarding water quality analysis (Ahmed, 2017; Ahmed and Shah, 2017a; Ahmed and Shah, 2017b), riverbank erosion (Islam and Hoque, 2014), stream flows (Ahmed and Shah, 2017b), and water level modelling (Biswas et al., 2009). The Surma River's Keane Bridge station provided the study's water quality variables between January 2017 and December 2019 obtained 15 cm to 20 cm below the surface.

The selection of prospective predictive factors is critical for predictive modelling. Various studies reveal that some variables predict DO better than others (Ahmed, 2017; Tomic et al., 2018). Ahmed (2017) used Biological Oxygen Demand (BOD) and Chemical oxygen demand (COD) for predicting the dissolved oxygen of the Surma River. Ay and Kisi (2012) observed that the temperature, pH, and electrical conductivity are highly influential over Fountain Creek, Colorado. However, Ranković (2010) claimed that pH and water temperature have a practical relation in DO prediction, whereas nitrates, chloride, and total phosphate have

poor connections. It is found that pH is a standard variable for predicting DO values using ANN, followed by temperature. However, along with pH and temperature, some authors used oxygen-containing (PO_4^{3-}, NO_3-N) variables or oxygen demanding variables (NH^{-4}N, COD, and BOD) (Wen et al., 2013). Turbidity (Iglesias et al., 2014) and total solid can be considered essential water quality parameters, as their high value indicates typically high values of other parameters associated with water quality. Humidity, water temperature, rainfall, total dissolved solids (TDS), pH, turbidity, and air temperature. The missing values were interpolated from two adjacent values. The fundamental statistics of the input variables are tabulated in Table 1.

2.3 Development of MODWT-MARS model

The multi-phase MODWT-MARS model and other benchmark models were created in Python using the sci-kit-learn machine learning platform (Pedregosa et al., 2011b). All simulations were performed on a machine with an Intel i7 processor running at 3.6GHz and 16 GB of RAM. Furthermore, a software platform such as "MATLAB2020" is employed for feature selection using neighbourhood component analysis (NCA). However, tools such as *matplotlib* (Barrett et al., 2005) and *seaborn* (Waskom et al., 2020) are employed to visualise the forecasted DO. Figure 2 depicts the workflow of the proposed MODWT-MARS model.

The wavelet transformation using MODWT was combined with the predictor variables filtered by the NCA approach to creating the MODWT-MARS model. Identifying the wavelet-scaling filter types and decomposition level is vital in creating a substantial wavelet transformation model. Because there is no one approach to choose the optimal filter, Al-Musaylh et al. (2020) used a trial and error strategy. Quilty and Adamowski (2018) discovered an issue in the forecast model inputs due to erroneous wavelet decomposition during the wavelet-based forecasting model. The inaccuracy can be traced back to the decomposition process's boundary conditions. They identified three problems: 1) improper use of future data, 2) unsuitable selection of decomposition levels and filters, and 3) incorrect division of validation and calibration data. The readers are encouraged to look up more information about the findings of Quilty and Adamowski (2018). The authors' concerns about the development

of MODWT and DWT decomposition were addressed in this study. After separating the DO variables to resolve more comprehensive information to create the MODWT-MARS model, Fig. 3 displays the time-series of the intrinsic mode functions (IMFs) and the residual components and decomposed components of MODWT.

There is no formula for verifying whether or not a model's valid predictors are present (Tiwari and Adamowski, 2013). Although the research describes three input selection strategies for picking the time series of lagged memories of DO and predictors for an optimum model, the literature does not specify which method should be used. The autocorrelation function (ACF), partial autocorrelation function (PACF), and cross-correlation function (CCF) approaches are the three types of approaches to consider. A substantial antecedent behaviour in terms of the lag of DO from the predictors was found in this study, utilising PACF as the predictor (Tiwari and Adamowski, 2013; Tiwari and Chatterjee, 2011). Fig. 4 demonstrates the PACF for DO time series showing the antecedent behaviour in terms of the lag of DO and decomposed components of DO using MODWT. It is clear from the figure that antecedent monthly delays are found significant.

The cross-correlation function determines which predictor's antecedent lag selects the input signal pattern and which pattern the predictor selects (Adamowski et al., 2012). The cross-correlation function is used to establish the statistical similarity between the predictors and the target variable. The cross-correlation function between the predictors and the DO for the River Surma is depicted in Figure 5a. Afterwards, a set of significant input combinations were determined by assessing r_{cross} of each predictor with DO. In this plot, a 95% confidence level of the statistically significant r_{cross} is shown in the blue line. It is found from the figure, the correlation of respective data with DO was found as highest for all stations at lag zero (r_{cross} ≈ 0.25 – 0.45). A similar procedure is maintained for the decomposed predictor variables. Fig 5b to 5f demonstrate the r_{cross} value between $\#d_1$ (DO) and $\#d_n$ (Predictors) and their respective residuals (n = 1 to 4). Fig. 5 shows that the r_{cross} value was ranged between 0.25 to 0.50 found more than 95% confidence level. The predictor data sets are normalised (Ahmed, 2017; Ali et al., 2019) between 0 and 1 to minimise one variable's overestimation.

$$DO_{norm} = \frac{DO - DO_{min}}{DO_{max} - DO_{min}} \quad (6)$$

In Eqs, (30), DO is the respective predictors, DO_{min} is the minimum value for the predictors, DO_{max} is the maximum value of the data and DO_{norm} is the normalised value of the data. After normalising the predictor variables, the data sets are partitioned 70% of the data sets as training, 15% of the data as testing, and the remaining 15% of the data sets are used for validation. The theoretical descriptions of the models have been elucidated in Section 2.

Python-based *Scikit-learn* (Pedregosa et al., 2011a) was used to build this study's SVR, RF, KRR, BNR, and KNN model. For SVR, the RBF (Radial Basis Function) was employed in developing the SVR model (Suykens et al., 2002). The RBF uses a faster function during training to examine non-linearities between the objective and predictor variables (Goyal et al., 2014; Lin, 2003; Maity et al., 2010). The tricky process of creating an accurate SVR model required identifying the 3D parameters (C, σ, and ε) (Hoang et al., 2014). This is why the NCA algorithm was used to select the parameters with the smallest weight value.

Alternatively, the MARS model adopted the Python-based Py-earth package (Rudy and Cherti, 2017). The two MARS models used are cubic or linear piecewise functions. This study used a piecewise cubic model because it provided a smoother response. Also, the generalised recursive partitioning regression was adopted since it can handle multiple preconditioners. A forward and backward selection was used for optimisation. The algorithm initially ran with a 'naïve' model that only contained the intercept term. By iteratively adding the reflected pairs of basis functions, the training MSE was reduced.

The accuracy of the hybrid MARS and other comparing models was constructed using piecewise cubic and linear regression functions, respectively. The best MARS model was selected using the lowest Generalised Cross-Validation (GCV) (Lin, 2003); the MODWT-MARS model yielded the lowest RMSE and the highest LM, demonstrating the most accurate predictions.

2.4 Model evaluation benchmarks

Several statistical score metrics were considered in the rigorous evaluation of the proposed model (i.e., MODWT-MARS) compared with the counterpart models. The commonly adopted model score metrics such as Pearson's Correlation Coefficient (r), root mean square error (RMSE; mg/l), mean absolute error (MAE; mg/l), Nash- Sutcliffe efficiency (NSE), Absolute Percentage Bias (APB; %), and Willmott's Index agreement (WI) (Krause et al., 2005; Legates and McCabe, 1999; Nash and Sutcliffe, 1970; Willmott et al., 2012) were used as the popular metrics used elsewhere (Ahmed et al., 2021c; Ghimire et al., 2019c). Due to the stations' geographic alterations, the percentage error measures, relative error values such as RRMSE, RMAE, and MAPE were considered. Owing to the inherent merits and weaknesses of the metrics, combining them is prudent (Sharma et al., 2019). Different sets of model evaluation metrics such as RMSE, MAE, and r^2 (coefficient of determination) (Chu et al., 2020); NSE, RMSE, MAE, and PERS (persistence index) (Tiwari and Chatterjee, 2010); Legates McCabe's Index (LM), Willmott's Index (WI), RRMSE, RMAE (Ali et al., 2019; Ghimire et al., 2019b; Yaseen et al., 2019) were selected for evaluating the model with numerous sets of variables. The correlation coefficient (*r*) provides information about the linear association between forecasted and observed DO data; therefore, it is limited in its capacity. However, *r* is considered oversensitive to extreme values (Willmott et al., 1985). Moreover, RMSE and MAE can provide appropriate information regarding the forecasting skill, whereby RMSE evaluates the robustness of the model related to high values but focuses on the deviation of the forecasted value from the observed (Deo et al., 2017a). Alternatively, MAEs are not a perfect replacement for RMSEs (Chai and Draxler, 2014). The Nash-Sutcliffe Efficiency (NSE) is a widely used model evaluation criteria for the hydrological models. NSE is a dimensionless metric and a scaled version of MSE, offering a better physical interpretation (Legates and McCabe, 2013). However, the NSE over-emphasises the higher values of outliers, and lower values are neglected (Legates and McCabe, 1999). Due to the standardisation of the observed and predicted means and variance, the robustness of *r* is limited. Willmott's Index (WI) was utilised to address this issue by considering the mean squared error ratio instead of the differences. The mathematical notations of the statistical metrics are as follows:

$$\text{MAE (mg/l)} = \frac{1}{N}\sum_{i=1}^{N}|DO_{for} - DO_{obs}| \quad (7)$$

$$\text{RMSE (mg/l)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(DO_{for} - DO_{obs})^2} \quad (8)$$

$$\text{NSE} = 1 - \left[1 - \frac{\sum_{i=1}^{N}(DO_{for})^2}{\sum_{i=1}^{N}(DO_{obs} - \overline{DO}_{for})^2}\right] \quad (9)$$

$$r = \left\{\frac{\sum_{i=1}^{N}(DO_{obs} - \overline{DO}_{obs})(DO_{for} - \overline{DO}_{for})}{\sqrt{\sum_{i=1}^{N}(DO_{obs} - \overline{DO}_{obs})^2 \sum_{i=1}^{N}(DO_{for} - \overline{DO}_{for})^2}}\right\}^2$$

(10)

$$\text{RRMSE (\%)} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(DO_{for} - DO_{obs})^2}}{\frac{1}{N}\sum_{i=1}^{N}(DO_{obs})} \text{ x } 100$$

(11)

$$\text{RMAE (\%)} = \frac{\frac{1}{N}\sum_{i=1}^{N}|DO_{for} - DO_{obs}|}{\frac{1}{N}\sum_{i=1}^{N}(DO_{obs})} \text{ x } 100$$

(12)

$$\text{MAPE (\%)} = \frac{1}{N}\left(\sum_{N}^{i=1}\left|\frac{(DO_{for} - DO_{obs})}{DO_{obs}}\right|\right) * 100$$

(13)

$$\text{WI} = 1 - \left[\frac{\sum_{i=1}^{N}(DO_{for} - DO_{obs})^2}{\sum_{i=1}^{N}(|DO_{for} - \overline{DO}_{obs}| + |DO_{obs} - \overline{DO}_{obs}|)^2}\right]$$

(14)

$$\text{APB (mg/l)} = \left[\frac{\sum_{i=1}^{N}(|DO_{for} - DO_{obs}|) * 100}{\sum_{i=1}^{N}|DO_{obs}|}\right]$$

(15)

$$LM = 1 - \left[\frac{\sum_{i=1}^{N}|DO_{for} - DO_{obs}|}{\sum_{i=1}^{N}||DO_{obs} - \overline{DO}_{obs}||}\right]$$

(16)

Where DO_{obs} and DO_{for} denote the observed and model-forecasted values from the i^{th} element; $\overline{DO}_{obs}$ and $\overline{DO}_{for}$ denote their average, respectively, and N represents the observation's number of the DO.

3.0 Results

In this research work, the appropriateness of the hybrid MODWT-MARS model compared against the CEEMDAN-MARS, EEMD-MARS, standalone MARS, and standalone SVR models for forecasting DO have been investigated. Though the mathematical metrics are so ambiguous that there is no way to evaluate the suitable alternative, it is reasonable to use multiple performance evaluation approaches. Compared with the other models, the hybrid and standalone models of BNR, KNN, KRR, and RF outstripped all decomposition methods. The performance of MODWT-MARS has revealed that the NCA algorithm helped choose the relevant features to assist the MARS in better emulating the future DO concentration. MODWT captured significant performance matrices (*i.e., r*, *NSE*, *WI*, *RMSE*, *MAE*); the MODWT-MARS model outperforms all the other developed models.

This study used the NCA algorithm to screen the appropriate predictor variables to use in the model. Table 2 provides the input combination for forecasting DO. The robustness of the NCA integrated BNR, KNN, KRR, MARS, and SVR model is provided in Tables A1–A6 in terms of statistical metrics would be found as supplementary materials. Tables show that each model's optimum standalone models were found between combinations from 19 to 29. For the case of the BNR model (Table 3), the standalone model (BNR_{28}) shows poor performance (r = 0.809, WI = 0.887, $RMAE$ = 7.55% and MAE = 0.275) comparing with the BNR-MODWT model (r = 0.977, WI = 0.987, $RMAE$ = 3.37% and MAE = 0.117). Moreover, the hybrid models showed improved performance ranging from 0.888 to 0.977 and 7.17% to 3.37% for r and $RMAE$ accordingly. The $MARS_{29}$ model was found as the optimum model (r = 0.824, WI = 0.895, $RMAE$ = 7.97% and MAE = 0.277) among all combinations of MARS model. The MODWT-MARS model was found as the highest performed model with substantial performance parameters (r = 0.981, WI = 0.990, $RMAE$ = 2.47% and MAE = 0.089) which is followed by CEEMDAN-MARS model (r = 0.949, WI = 0.971, $RMAE$ = 4.65% and MAE = 0.156). Mentionable that the highest model of SVR was found for CEEMDAN-SVR (r = 0.971, WI = 0.983, $RMAE$ = 3.36%) compared with the optimum standalone model (SVR_{20}). Mentionable that KRR, KNN, and RF model provides poor performance comparatively.

Further analysis through box plot showing the forecasted vs observed DO and absolute forecasting error of all hybrid models is illustrated in Fig. 6. The absolute forecasted error was determined as $|FE| = DO^{for} - DO^{obs}$. The box plot demonstrates the observed (DO^{obs}) data dispersal and forecasted (DO^{for}) DO from the proposed machine learning approaches and other comparing models. Fig. 6b, 6c, and 6e visualise the quartiles' data with distinctly larger outliers. The lower end of the plot lies between the lower quartile (25^{th} percentile) and the upper quartile (75^{th} percentile). The MODWT-MARS model shows an identical prediction compared with MODWT-SVR with higher outliers for the SVR model. A more in-depth inspection of the absolute forecast error (|FE|) from the hybrid MODWT-MARS model further strengthens the suitability of the hybrid MARS approach in predicting the DO of Surma River the narrowest distribution compared with other models. A significant percentage (98%) of the |FE| in the first error brackets ($0<|FE|<0.25$) for the MODWT-MARS model, while for MODWT-SVR, the percentage is 95%.

The empirical cumulative distribution function (ECDF) visualisation demonstrates the forecast error data's feature from the least to highest and perceives the full features circulated across the dataset. Fig. 7 represents the empirical CDF of all six models for objective models and comparing models. The hybrid MODWT-MARS model was seen as reasonably sound against other models. The MODWT-MARS generated errors significantly lower from 0 to 0.25mg/l. In the model-like KNN, KRR, and RF, the distribution of CDF was larger comparatively. The analysis also revealed that the standalone models showed a poor distribution, proving that MODWT-MARS was the most precise and responsive model.

To analyse the proposed MODWT-MARS model's further robustness, the models' forecasting performance was further assessed based on RRMSE and MAPE for all tested models, as shown in Fig. 8. From Fig. 8, the magnitude of RRMSE and MAPE for the objective model (MODWT-MARS) is significantly low, which clarifies the potential merits of the proposed model. The best RRMSE (3.6%) and MAPE (2.2%) were found for the MODWT-MARS model, which is followed by the MODWT-BNR model with moderate RRMSE (4.0%) and MAPE (3%). Besides, KNN, KRR, and RF models with MODWT showed RRMSE (11.5%

to 13.5%) and MAPE (9.5% to 12%) value, demonstrating poor performance. The analysis revealed that the hybrid MODWT-MARS model captures the future DO with higher accuracy.

Compared with standalone models using the Taylor diagram (Taylor, 2001), the proposed model performance analysis is an improved interpretation presented in Fig. 9. The Taylor diagram demonstrates that the MODWT-MARS model with the NCA algorithm is closer to the observation than the comparing models. The forecasted DO again illuminates the proposed model's better pertinency than the standalone and benchmark models. The benchmark models' performance with CEEMDAN and MODWT (*i.e.,* MODWT-SVR, CEEMDAN-SVR, and CEEMDAN-MARS) achieved closer proximity concerning the observed values. However, the proximity of observed DO for the MARS model with MODWT feature extraction is the closest.

The scatter plot of the forecasted and observed DO for the proposed MODWT-MARS model portrayed a detailed comparison of DO forecasting (Figure 10). The scatter plots comprise with the coefficient of determination (R^2) with goodness-of-fit between forecasted vs observed DO and a least-square fitting line and the corresponding equation; $DO_{for} = m * DO_{obs} + C$, where, '*m*' is referred to as the gradient, and *'C'* is denoted as the y-intercept. Figure 10 reveals that the proposed model displays significant performance with a more considerable R^2 value. The DO forecasting using a hybrid machine learning model (*i.e.,* MODWT-MARS) performed significantly better than the other models. The magnitudes registered from the hybrid MODWT-MARS model were the closest to unity, which, in pairs ($m|R^2$), are 0.978|0.976, followed by MODWT-SVR (0.939|0.965). Moreover, the CEEMDAN-SVR (0.699|0.795) and CEEMDAN-MARS (0.700|0.794) models provide a comparatively lower pair. Alternatively, y-intercepts [*ideal value = 0*] was found close to zero i.e., 0.084 for the proposed model. However, for the other models, the y-intercept deviated from the ideal value with more outliers.

To attain a different interpretation of the proposed MODWT-MARS model's accuracy, the time series plot is used to comprehend the proposed model's forecasting ability. Fig. 11 demonstrates the time series plot of forecasted and observed DO with MODWT-MARS

compared to MARS's standalone model. Results show that the proposed MODWT-MARS model is found close to the observed DO revealed a high predictive accuracy. After applying the NCA algorithm as a feature selection approach and MODWT as a feature extraction technique, the improvement of forecasted DO is improved.

Notably, five unique decomposition algorithms, EMD, EEMD, CEEMDAN, DWT, and MODWT, are incorporated to enhance the MARS-based predictive model. In terms of r, LM, and APB of DO forecasting, the MODWT effectively forecasts improvement. In the MARS model, *r* and *LM* values of using the MODWT model increased by ~19% and ~20% accordingly, and APB decreased by ~68%. Similarly, for the BNR model, MODWT feature-extraction skill increased *r* and *LM* values up to ~21% and ~59% accordingly, and APB is decreased by ~57%. Additionally, *r* and *LM* values for the MARS model with CEEMDAN are increased by ~15% and ~50%, respectively. Similarly, the inclusion of DWT, EMD, and EEMD also substantially improved the *r*, *LM,* and *APB* values.

3.1 Discussion

Another essential aspect of inefficiency is determining the optimal combination of input variables, which is one of the most critical factors in inefficiency and model configuration. According to the findings of this study, different input combinations have varying effects on the outcomes. Then, several input variables must be analysed, and the most appropriate collection of variables must be employed to optimise the products. Every model should have its ideal combination; yet, the most effective combination is rare throughout the various models. Al-Musaylh et al. (2019) used the hybrid MARS model in forecasting electricity demand with a good performance. This study demonstrated profound forecasting of Dissolved Oxygen (DO) concentration. Our findings have led to better forecasting than any algorithm evaluated in both the standalone and hybrid versions. We propose more studies to forecast DO using wet and dry season data sets and compare the results with the whole dataset's findings. Different pre-processing techniques could also enhance the projection accuracy of the MARS model. First, it is possible to implement a suitable feature selection approach such as NCA

(Ahmed et al., 2021c; Ghimire et al., 2019a) algorithm to pick the input variables that significantly impact the model. The feature weight calculated using neighbourhood component analysis (NCA) respective to predictor variables was added one by one based on the highest to lowest feature weight to improve the model performance. The optimum combination of input parameters was found significantly in the proposed hybrid MARS model.

4.0 Conclusion:

This research studies the hybrid machine learning model is incorporating neighbourhood component analysis (NCA) as a feature selection method and multivariate adaptive regression splines (MARS) as a predictive model for DO forecast of the River Surma. Five distinct data extraction approaches were utilised for developing the model. A new approach to DO forecast was created using a decedent lagged memory framework to explain the problem more appropriately and its consequences afterwards. The MODWT-MARS model is the objective modelling approach that provides the optimal performance among the benchmarked models. From this analysis, the following inference can be made.

1. The achieved results demonstrated that the NCA algorithm would be a helpful option for getting the predictor variables' substantial features. The model's performance metrics indicate that the NCA algorithm was a suitable tool for feature selection, as the NCA and MODWT optimised models showed better performance than standalone models.
2. The proposed hybrid MODWT-MARS model showed the best performance in forecasting dissolved oxygen concentration of the Surma River. The superior status of the MODWT-MARS model is endorsed with a low MAE (0.089) and high NS (0.990) value. The percentage increase in the correlation coefficient (r) values to 20% and LM index values to 19% with their respective standalone models. Precisely, the MODWT-MARS had the best performance considering r (0.981), WI (0.990), RMSE (0.121mg/l), and MAE (0.089mg/l) values.

3. Based on the conclusions, it is recognised that the hybrid MODWT-MARS model with the NCA feature algorithm shows superior forecasting of DO. The station's antecedent values of water quality parameters and hydro-meteorological variables embed the machine learning approach's future forecasting success. Therefore, this type of forecast is applied to DO forecasting for better water quality management.

Credit authorship contribution statement

A. A. Masrur Ahmed: Writing - original, Conceptualisation, Methodology, Software, Model development, and application. **M. A. I. Chowdhury:** Conceptualisation, Writing - review & editing. **Mumtaz Ali**: Writing: Review, **O Ahmed:** Data Collection & Writing. **A Sutradhar:** Data Collection & Writing.

Acknowledgements

The authors want to thank Leading University for allowing us to conduct laboratory testing of water parameters.

References

Adamowski, J., Fung Chan, H., Prasher, S.O., Ozga-Zielinski, B. and Sliusarieva, A., 2012. Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. Water Resources Research, 48(1).

Agbinya, J.I., 1996. Discrete wavelet transform techniques in speech processing, Proceedings of Digital Processing Applications (TENCON'96). IEEE, pp. 514-519.

Ahmed, A., Deo, R.C., Feng, Q., Ghahramani, A., Raj, N., Yin, Z. and Yang, L., 2021a. Hybrid deep learning method for a week-ahead evapotranspiration forecasting. Stochastic Environmental Research and Risk Assessment: 1-19.

Ahmed, A., Deo, R.C., Raj, N., Ghahramani, A., Feng, Q., Yin, Z. and Yang, L., 2021b. Deep Learning Forecasts of Soil Moisture: Convolutional Neural Network and Gated Recurrent Unit Models Coupled with Satellite-Derived MODIS, Observations and Synoptic-Scale Climate Index Data. Remote Sensing, 13(4): 554.

Ahmed, A.A.M., 2017. Prediction of dissolved oxygen in Surma River by biochemical oxygen demand and chemical oxygen demand using the artificial neural networks (ANNs). Journal of King Saud University - Engineering Sciences, 29(2): 151-158.

Ahmed, A.A.M., Deo, R.C., Raj, N., Ghahramani, A., Feng, Q., Yin, Z. and Yang, L., 2021c. Deep Learning Forecasts of Soil Moisture: Convolutional Neural Network and Gated Recurrent Unit Models Coupled with Satellite-Derived MODIS, Observations and Synoptic-Scale Climate Index Data. Remote Sensing, 13(4): 554.

Ahmed, A.A.M. and Shah, S.M.A., 2017a. Application of adaptive neuro-fuzzy inference system (ANFIS) to estimate the biochemical oxygen demand (BOD) of Surma River. Journal of King Saud University - Engineering Sciences, 29(3): 237-243.

Ahmed, A.M., Deo, R.C., Ghahramani, A., Raj, N., Feng, Q., Yin, Z. and Yang, L., 2021d. LSTM integrated with Boruta-random forest optimiser for soil moisture estimation under RCP4. 5 and RCP8. 5 global warming scenarios. Stochastic Environmental Research and Risk Assessment: 1-31.

Ahmed, A.M. and Shah, S.M.A., 2017b. Application of artificial neural networks to predict peak flow of Surma River in Sylhet Zone of Bangladesh. International Journal of Water, 11(4): 363-375.

Akbari Asanjan, A., Yang, T., Hsu, K., Sorooshian, S., Lin, J. and Peng, Q., 2018. Short-Term Precipitation Forecast Based on the PERSIANN System and LSTM Recurrent Neural Networks. Journal of Geophysical Research: Atmospheres, 123(22).

Al-Musaylh, M.S., Deo, R.C., Adamowski, J.F. and Li, Y., 2019. Short-term electricity demand forecasting using machine learning methods enriched with ground-based climate and ECMWF Reanalysis atmospheric predictors in southeast Queensland, Australia. Renewable and Sustainable Energy Reviews, 113.

Al-Musaylh, M.S., Deo, R.C. and Li, Y., 2020. Electrical energy demand forecasting model development and evaluation with maximum overlap discrete wavelet transform-online sequential extreme learning machines algorithms. Energies, 13(9): 2307.

Ali, M., Deo, R.C., Downs, N.J. and Maraseni, T., 2018. Multi-stage hybridised online sequential extreme learning machine integrated with Markov Chain Monte Carlo copula-Bat algorithm for rainfall forecasting. Atmospheric research, 213: 450-464.

Ali, M., Deo, R.C., Maraseni, T. and Downs, N.J., 2019. Improving SPI-derived drought forecasts incorporating synoptic-scale climate indices in multi-phase multivariate empirical mode decomposition model hybridised with simulated annealing and kernel ridge regression algorithms. Journal of Hydrology, 576: 164-184.

Ali, M., Deo, R.C., Xiang, Y., Li, Y. and Yaseen, ZM, 2020. Forecasting long-term precipitation for water resource management: a new multi-step data-intelligent modelling approach. Hydrological Sciences Journal, 65(16): 2693-2708.

Arhami, M., Kamali, N. and Rajabi, M.M., 2013. Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by Monte Carlo simulations. Environmental Science and Pollution Research, 20(7): 4777-4789.

Arto, I., Garcia-Muros, X., Cazcarro, I., Gonzalez-Eguino, M., Markandya, A. and Hazra, S., 2019. The socioeconomic future of deltas in a changing environment. Sci Total Environ, 648: 1284-1296.

Beltrán-Castro, J., Valencia-Aguirre, J., Orozco-Alzate, M., Castellanos-Domínguez, G. and Travieso-González, C.M., 2013. Rainfall forecasting based on ensemble empirical mode decomposition and neural networks, International Work-Conference on Artificial Neural Networks. Springer, pp. 471-480.

Biswas, R., Jayawardena, A. and Takeuchi, K., 2009. Prediction of water levels in the Surma River of Bangladesh by artificial neural network, Proceeding of 2009 Annual Conference.

Breiman, L., 2001. Random Forests. Machine Learning, 45: 5–32.

Bruce, L.M., Koger, C.H. and Li, J., 2002. Dimensionality reduction of hyperspectral data using discrete wavelet transform feature extraction. IEEE Transactions on geoscience and remote sensing, 40(10): 2331-2338.

Chai, T. and Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7(3): 1247-1250.

Chen, L., Ye, L., Singh, V., Zhou, J. and Guo, S., 2014. Determination of Input for Artificial Neural Networks for Flood Forecasting Using the Copula Entropy Method. Journal of Hydrologic Engineering, 19(11).

Chowdhury, R.K. and Ali, S.I.M., 2006. Investigation of Phosphate and Ammonia-Nitrogen concentrations at some selected locations of the Malnichara channel and the Surma river.

Chu, H., Wei, J. and Wu, W., 2020. Streamflow prediction using LASSO-FCM-DBN approach based on hydro-meteorological condition classification. Journal of Hydrology, 580.

Cornish, C.R., Bretherton, CS and Percival, D.B., 2006. Maximal overlap wavelet statistical analysis with application to atmospheric turbulence. Boundary-Layer Meteorology, 119(2): 339-374.

Deo, R.C., Downs, N., Parisi, A.V., Adamowski, J.F. and Quilty, J.M., 2017a. Very short-term reactive forecasting of the solar ultraviolet index using an extreme learning machine integrated with the solar zenith angle. Environ Res, 155: 141-166.

Deo, R.C., Kisi, O. and Singh, V.P., 2017b. Drought forecasting in eastern Australia using multivariate adaptive regression spline, least square support vector machine and M5Tree model. Atmospheric Research, 184: 149-175.

Deo, R.C. and Sahin, M., 2016. An extreme learning machine model for the simulation of monthly mean streamflow water level in eastern Queensland. Environ Monit Assess, 188(2): 90.

Deo, R.C. and Şahin, M., 2015a. Application of the Artificial Neural Network model for prediction of monthly Standardised Precipitation and Evapotranspiration Index using hydro-meteorological parameters and climate indices in eastern Australia. Atmospheric Research, 161-162: 65-81.

Deo, R.C. and Şahin, M., 2015b. Application of the extreme learning machine algorithm for the prediction of monthly Effective Drought Index in eastern Australia. Atmospheric Research, 153: 512-525.

Deo, R.C., Wen, X. and Qi, F., 2016. A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. Applied Energy, 168: 568-593.

Di, C., Yang, X. and Wang, X., 2014. A four-stage hybrid model for hydrological time series forecasting. PLoS One, 9(8): e104663.

Elsafi, S.H., 2014. Artificial Neural Networks (ANNs) for flood forecasting at Dongola Station in the River Nile, Sudan. Alexandria Engineering Journal, 53(3): 655-662.

Flandrin, P., Rilling, G. and Goncalves, P., 2004. Empirical mode decomposition as a filter bank. IEEE signal processing letters, 11(2): 112-114.

Forough, K.-T., Mousavi, S.-F., Khaledian, M., Yousefi-Falakdehi, O. and Norouzi-Masir, M., 2019. Prediction of Water Quality Index by Support Vector Machine: a Case Study in the Sefidrud Basin, Northern Iran. Water Resources, 46(1): 112-116.

Fowler, J.E., 2005. The redundant discrete wavelet transform and additive noise. IEEE Signal Processing Letters, 12(9): 629-632.

Friedman, J.H., 1991. Multivariate adaptive regression splines. The annals of statistics: 1-67.

Ghimire, Deo, Raj and Mi, 2019a. Deep Learning Neural Networks Trained with MODIS Satellite-Derived Predictors for Long-Term Global Solar Radiation Prediction. Energies, 12(12).

Ghimire, S., Deo, R.C., Downs, N.J. and Raj, N., 2019b. Global solar radiation prediction by ANN integrated with European Centre for medium range weather forecast fields in solar rich cities of Queensland Australia. Journal of Cleaner Production, 216: 288-310.

Ghimire, S., Deo, R.C., Raj, N. and Mi, J., 2019c. Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms. Applied Energy, 253.

Goyal, M.K., Bharti, B., Quilty, J., Adamowski, J. and Pandey, A., 2014. Modeling of daily pan evaporation in sub tropical climates using ANN, LS-SVR, Fuzzy Logic, and ANFIS. Expert systems with applications, 41(11): 5267-5276.

Hamidi, O., Poorolajal, J., Sadeghifar, M., Abbasi, H., Maryanaji, Z., Faridi, H.R. and Tapak, L., 2015. A comparative study of support vector machines and artificial neural

networks for predicting precipitation in Iran. Theoretical and applied climatology, 119(3-4): 723-731.
He, F., Zhang, Y., Liu, D., Dong, Y., Liu, C. and Wu, C., 2017. Mixed wavelet-based neural network model for cyber security situation prediction using MODWT and Hurst exponent analysis, International Conference on Network and System Security. Springer, pp. 99-111.
Heddam, S., 2016. Use of Optimally Pruned Extreme Learning Machine (OP-ELM) in Forecasting Dissolved Oxygen Concentration (DO) Several Hours in Advance: a Case Study from the Klamath River, Oregon, USA. Environmental Processes, 3(4): 909-937.
Heddam, S., 2017. Fuzzy neural network (EFuNN) for modelling dissolved oxygen concentration (DO), Intelligence Systems in Environmental Management: Theory and Applications. Springer, pp. 231-253.
Heddam, S. and Kisi, O., 2017. Extreme learning machines: a new approach for modeling dissolved oxygen (DO) concentration with and without water quality variables as predictors. Environmental Science and Pollution Research, 24(20): 16702-16724.
Heddam, S. and Kisi, O., 2018. Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree. Journal of Hydrology, 559: 499-509.
Henderson, R.K., Baker, A., Murphy, K., Hambly, A., Stuetz, R. and Khan, S., 2009. Fluorescence as a potential monitoring tool for recycled water systems: a review. Water research, 43(4): 863-881.
Hoang, N.-D., Pham, A.-D. and Cao, M.-T., 2014. A novel time series prediction approach based on a hybridisation of least squares support vector regression and swarm intelligence. Applied Computational Intelligence and Soft Computing, 2014.
Hu, C., Wu, Q., Li, H., Jian, S., Li, N. and Lou, Z., 2018. Deep Learning with a Long Short-Term Memory Networks Approach for Rainfall-Runoff Simulation. Water, 10(11).
Huang, H. and Abdel-Aty, M., 2010. Multilevel data and Bayesian analysis in traffic safety. Accident Analysis & Prevention, 42(6): 1556-1565.
Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.-C., Tung, C.C. and Liu, H.H., 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences, 454(1971): 903-995.
Hur, J. and Cho, J., 2012. Prediction of BOD, COD, and total nitrogen concentrations in a typical urban river using a fluorescence excitation-emission matrix with PARAFAC and UV absorption indices. Sensors, 12(1): 972-986.
Iglesias, C., Torres, J.M., Nieto, P.G., Fernández, J.A., Muñiz, C.D., Piñeiro, J. and Taboada, J., 2014. Turbidity prediction in a river basin by using artificial neural networks: a case study in northern Spain. Water resources management, 28(2): 319-331.
Islam, M.S. and Hoque, F., 2014. River Bank Erosion of the Surma River Due to Slope Failure. International Journal of Research and Innovations in Earth Science, 1(2): 54-58.
Jiao, G., Guo, T. and Ding, Y., 2016. A new hybrid forecasting approach applied to hydrological data: a case study on precipitation in Northwestern China. Water, 8(9): 367.
Keshtegar, B., Heddam, S. and Hosseinabadi, H., 2019. The employment of polynomial chaos expansion approach for modeling dissolved oxygen concentration in river. Environmental Earth Sciences, 78(1): 1-18.
Khaldi, K., Alouane, M.T.-H. and Boudraa, A.-O., 2008. A new EMD denoising approach dedicated to voiced speech signals, 2008 2nd International Conference on Signals, Circuits and Systems. IEEE, pp. 1-5.
Kisi, O., Alizamir, M. and Gorgij, A.D., 2020. Dissolved oxygen prediction using a new ensemble method. Environmental Science and Pollution Research: 1-15.

Kisi, O. and Ay, M., 2012. Comparison of ANN and ANFIS techniques in modeling dissolved oxygen, Proceedings of the Sixteenth International Water Technology Conference (IWTC 16), Istanbul, Turkey, pp. 7-10.

Kisi, O. and Parmar, K.S., 2016. Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution. Journal of Hydrology, 534: 104-112.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K. and Herrnegger, M., 2018. Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. Hydrology and Earth System Sciences, 22(11): 6005-6022.

Krause, P., Boyle, D. and Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Advances in geosciences, 5: 89-97.

Kuo, Y.M., Liu, C.W. and Lin, K.H., 2004. Evaluation of the ability of an artificial neural network model to assess the variation of groundwater quality in an area of blackfoot disease in Taiwan. Water Res, 38(1): 148-58.

Le, Ho, Lee and Jung, 2019. Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting. Water, 11(7).

Legates, D.R. and McCabe, G.J., 1999. Evaluating the use of "goodness-of-fit" Measures in hydrologic and hydroclimatic model validation. Water Resources Research, 35(1): 233-241.

Legates, D.R. and McCabe, G.J., 2013. A refined index of model performance: a rejoinder. International Journal of Climatology, 33(4): 1053-1056.

Li, M., Chen, W. and Zhang, T., 2017. Application of MODWT and log-normal distribution model for automatic epilepsy identification. Biocybernetics and Biomedical Engineering, 37(4): 679-689.

Lin, H., 2003. Hydropedology: Bridging disciplines, scales, and data. Vadose Zone Journal, 2(1): 1-11.

Maity, R., Bhagwat, PP and Bhatnagar, A., 2010. Potential of support vector regression for prediction of monthly streamflow using endogenous property. Hydrological Processes: An International Journal, 24(7): 917-923.

Mellios, N., Kofinas, D., Laspidou, C. and Papadimitriou, T., 2015. Mathematical modeling of trophic state and nutrient flows of Lake Karla using the PCLake model. Environmental Processes, 2(1): 85-100.

Mohan, S. and Kumar, K.P., 2016. Waste load allocation using machine scheduling: model application. Environmental Processes, 3(1): 139-151.

Mouri, G., Takizawa, S. and Oki, T., 2011. Spatial and temporal variation in nutrient parameters in stream water in a rural-urban catchment, Shikoku, Japan: Effects of land cover and human impact. Journal of Environmental Management, 92(7): 1837-1848.

Nash, J.E. and Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—A discussion of principles. Journal of hydrology, 10(3): 282-290.

Nguyen-Huy, T., Deo, R.C., An-Vo, D.-A., Mushtaq, S. and Khan, S., 2017. Copula-statistical precipitation forecasting model in Australia's agro-ecological zones. Agricultural Water Management, 191: 153-172.

Nourani, V., Baghanam, A.H., Adamowski, J. and Kisi, O., 2014. Applications of hybrid wavelet–Artificial Intelligence models in hydrology: A review. Journal of Hydrology, 514: 358-377.

Nourani, V., Komasi, M. and Mano, A., 2009. A multivariate ANN-wavelet approach for rainfall–runoff modeling. Water resources management, 23(14): 2877-2894.

Ouyang, Q., Lu, W., Xin, X., Zhang, Y., Cheng, W. and Yu, T., 2016. Monthly rainfall forecasting using EEMD-SVR based on phase-space reconstruction. Water resources management, 30(7): 2311-2325.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. and Dubourg, V., 2011a. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12: 2825-2830.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. and Dubourg, V., 2011b. Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct): 2825-2830.

Prasad, R., Deo, R.C., Li, Y. and Maraseni, T., 2017. Input selection and performance optimisation of ANN-based streamflow forecasts in the drought-prone Murray Darling Basin region using IIS and MODWT algorithm. Atmospheric Research, 197: 42-63.

Prasad, R., Deo, R.C., Li, Y. and Maraseni, T., 2018. Soil moisture forecasting by a hybrid machine learning technique: ELM integrated with ensemble empirical mode decomposition. Geoderma, 330: 136-161.

Prasad, R., Deo, R.C., Li, Y. and Maraseni, T., 2019. Weekly soil moisture forecasting with multivariate sequential, ensemble empirical mode decomposition and Boruta-random forest hybridiser algorithm approach. Catena, 177: 149-166.

Quilty, J. and Adamowski, J., 2018. Addressing the incorrect usage of wavelet-based hydrological and water resources forecasting models for real-world applications with best practices and a new forecasting framework. Journal of hydrology, 563: 336-353.

Raheli, B., Aalami, M.T., El-Shafie, A., Ghorbani, M.A. and Deo, R.C., 2017. Uncertainty assessment of the multilayer perceptron (MLP) neural network model with implementation of the novel hybrid MLP-FFA method for prediction of biochemical oxygen demand and dissolved oxygen: a case study of Langat River. Environmental Earth Sciences, 76(14): 1-16.

Ranković, V., Radulović, J., Radojević, I., Ostojić, A. and Čomić, L., 2010. Neural network modeling of dissolved oxygen in the Gruža reservoir, Serbia. Ecological Modelling, 221(8): 1239-1244.

Rathinasamy, M., Khosa, R., Adamowski, J., Ch, S., Partheepan, G., Anand, J. and Narsimlu, B., 2014. Wavelet-based multiscale performance analysis: An approach to assess and improve hydrological models. Water Resources Research, 50(12): 9721-9737.

Rudy, J. and Cherti, M., 2017. Py-earth: a python implementation of multivariate adaptive regression splines.

Saunders, C., Gammerman, A. and Vovk, V., 1998. Ridge regression learning algorithm in dual variables.

Seo, Y. and Kim, S., 2016. Hydrological Forecasting Using Hybrid Data-Driven Approach. American Journal of Applied Sciences, 13(8): 891-899.

Shabani, S., Samadianfard, S., Sattari, M.T., Mosavi, A., Shamshirband, S., Kmet, T. and Várkonyi-Kóczy, A.R., 2020. Modeling pan evaporation using gaussian process regression k-nearest neighbors random forest and support vector machines; Comparative analysis. Atmosphere, 11(1): 66.

Sharma, E., Deoa, R.C., Prasadb, R. and Parisia, A.V., 2019. A hybrid air quality early-warning framework: Hourly forecasting model with online sequential extreme learning machine and empirical mode decomposition algorithm.

Shensa, M.J., 1992. The discrete wavelet transform: wedding the a trous and Mallat algorithms. IEEE Transactions on signal processing, 40(10): 2464-2482.

Su, S., Li, D., Zhang, Q., Xiao, R., Huang, F. and Wu, J., 2011. Temporal trend and source apportionment of water pollution in different functional zones of Qiantang River, China. Water research, 45(4): 1781-1795.

Suen, J.-P. and Eheart, J.W., 2003. Evaluation of neural networks for modeling nitrate concentrations in rivers. Journal of water resources planning and management, 129(6): 505-510.

Suykens, J.A., De Brabanter, J., Lukas, L. and Vandewalle, J., 2002. Weighted least squares support vector machines: robustness and sparse approximation. Neurocomputing, 48(1-4): 85-105.

Tao, H., Bobaker, A.M., Ramal, MM, Yaseen, Z.M., Hossain, M.S. and Shahid, S., 2019. Determination of biochemical oxygen demand and dissolved oxygen for semi-arid river environment: application of soft computing models. Environmental Science and Pollution Research, 26(1): 923-937.

Taylor, K.E., 2001. Summarising multiple aspects of model performance in a single diagram. Journal of Geophysical Research: Atmospheres, 106(D7): 7183-7192.
Tiwari, M.K. and Adamowski, J., 2013. Urban water demand forecasting and uncertainty assessment using ensemble wavelet-bootstrap-neural network models. Water Resources Research, 49(10): 6486-6507.
Tiwari, M.K. and Chatterjee, C., 2010. Development of an accurate and reliable hourly flood forecasting model using wavelet–bootstrap–ANN (WBANN) hybrid approach. Journal of Hydrology, 394(3-4): 458-470.
Tiwari, M.K. and Chatterjee, C., 2011. A new wavelet–bootstrap–ANN hybrid model for daily discharge forecasting. Journal of Hydroinformatics, 13(3): 500-519.
Tomic, S.A., Antanasijevic, D., Ristic, M., Peric-Grujic, A. and Pocajt, V., 2018. A linear and nonlinear polynomial neural network modeling of dissolved oxygen content in surface water: Inter- and extrapolation performance with inputs' significance analysis. Sci Total Environ, 610-611: 1038-1046.
Torres, M.E., Colominas, M.A., Schlotthauer, G. and Flandrin, P., 2011. A complete ensemble empirical mode decomposition with adaptive noise, 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp. 4144-4147.
Tripathi, S., Srinivas, V.V. and Nanjundiah, R.S., 2006. Downscaling of precipitation for climate change scenarios: A support vector machine approach. Journal of Hydrology, 330(3-4): 621-640.
US-Geological-Survey, 2016. National water information system data available on the world wide web (USGS water data for the nation).
Wen, X., Fang, J., Diao, M. and Zhang, C., 2013. Artificial neural network modeling of dissolved oxygen in the Heihe River, Northwestern China. Environmental monitoring and assessment, 185(5): 4361-4371.
Willmott, C.J., Ackleson, S.G., Davis, R.E., Feddema, J.J., Klink, K.M., Legates, D.R., O'donnell, J. and Rowe, C.M., 1985. Statistics for the evaluation and comparison of models. Journal of Geophysical Research: Oceans, 90(C5): 8995-9005.
Willmott, C.J., Robeson, S.M. and Matsuura, K., 2012. A refined index of model performance. International Journal of Climatology, 32(13): 2088-2094.
Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B. and Philip, S.Y., 2008. Top 10 algorithms in data mining. Knowledge and information systems, 14(1): 1-37.
Xiang, S., Liu, Z. and Ma, L., 2006. Study of multivariate linear regression analysis model for ground water quality prediction. Guizhou Science, 24(1): 60-62.
Yang, L., Feng, Q., Yin, Z., Wen, X., Deo, R.C., Si, J. and Li, C., 2018. Application of multivariate recursive nesting bias correction, multiscale wavelet entropy and AI-based models to improve future precipitation projection in upstream of the Heihe River, Northwest China. Theoretical and Applied Climatology, 137(1-2): 323-339.
Yang, P., Xia, J., Zhang, Y. and Hong, S., 2017. Temporal and spatial variations of precipitation in Northwest China during 1960–2013. Atmospheric Research, 183: 283-295.
Yaseen, ZM, Jaafar, O., Deo, R.C., Kisi, O., Adamowski, J., Quilty, J. and El-Shafie, A., 2016. Streamflow forecasting using extreme learning machines: A case study in a semi-arid region in Iraq. Journal of Hydrology, 542: 603-614.
Yaseen, Z.M., Sulaiman, S.O., Deo, R.C. and Chau, K.-W., 2019. An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction. Journal of Hydrology, 569: 387-408.
Yin, Z., Feng, Q., Wen, X., Deo, R.C., Yang, L., Si, J. and He, Z., 2018. Design and evaluation of SVR, MARS and M5Tree models for 1, 2 and 3-day lead time forecasting of river flow data in a semiarid mountainous catchment. Stochastic Environmental Research and Risk Assessment, 32(9): 2457-2476.

Yoo, C. and Cho, E., 2018. Comparison of GCM Precipitation Predictions with Their RMSEs and Pattern Correlation Coefficients. Water, 10(1).

You, Y., Demmel, J., Hsieh, C.-J. and Vuduc, R., 2018. Accurate, fast and scalable kernel ridge regression on parallel and distributed systems, Proceedings of the 2018 International Conference on Supercomputing, pp. 307-317.

Zhang, W. and Goh, A.T., 2016. Multivariate adaptive regression splines and neural network models for prediction of pile drivability. Geoscience Frontiers, 7(1): 45-52.

Zhang, W., Qu, Z., Zhang, K., Mao, W., Ma, Y. and Fan, X., 2017. A combined model based on CEEMDAN and modified flower pollination algorithm for wind speed forecasting. Energy Conversion and Management, 136: 439-451.

Zhou, Y., Li, T., Shi, J. and Qian, Z., 2019. A CEEMDAN and XGBOOST-Based Approach to Forecast Crude Oil Prices. Complexity, 2019: 1-15.

Zhu, S. and Heddam, S., 2020. Prediction of dissolved oxygen in urban rivers at the Three Gorges Reservoir, China: extreme learning machines (ELM) versus artificial neural network (ANN). Water Quality Research Journal, 55(1): 106-118.

Zounemat-Kermani, M., Seo, Y., Kim, S., Ghorbani, M.A., Samadianfard, S., Naghshara, S., Kim, N.W. and Singh, V.P., 2019. Can decomposition approaches always enhance soft computing models? Predicting the dissolved oxygen concentration in the St. Johns River, Florida. Applied Sciences, 9(12): 2534.

Table 1 Basic Statistics, i.e., minimum (min), maximum (max), mean (M), standard deviation (SD), and coefficient of variation (CV) of the water quality variables in Surma River, Sylhet, Bangladesh

Variable	Acronyms	Unit	Min	Max	Mean	SD	CV (%)
Humidity	h	%	0.01	3.79	0.53	0.70	132
Water Temperature	w	^{0}C	0.18	4.0	1.53	1.05	69
Rainfall	r	mm	8.00	127	32.66	20.99	64
TDS	td	Mg/l	10.0	522	142.3	102.15	72
pH	p	-	5.70	8.25	6.92	0.55	8
Turbidity	tr	(NTU)	4.18	42.62	11.84	7.37	62
Air Temperature	a	^{0}C	12.30	33.30	27.10	4.93	20.00
DO	d	(mg/l)	1.90	17.30	5.40	2.45	45

Table 2 Different input combinations prepared by using the NCA feature selection algorithm. Numerical values after the variable indicate respective lag memories of the datasets.

No.	Different input combinations
1	h_3
2	h_3,tr_5
3	h_3,tr_5, h_5
4	h_3,tr_5, h_5, h_6
5	h_3,tr_5, h_5, h_6, tr_4
6	h_3,tr_5, h_5, h_6, tr_4, tr_7
7	h_3,tr_5, h_5, h_6, tr_4, tr_7, tr_3
8	h_3,tr_5, h_5, h_6, tr_4, tr_7, tr_3, h_4
9	h_3,tr_5, h_5, h_6, tr_4, tr_7, tr_3, h_4, tr_6
10	h_3,tr_5, h_5, h_6, tr_4, tr_7, tr_3, h_4, tr_6, td_4
11	h_3,tr_5, h_5, h_6, tr_4, tr_7, tr_3, h_4, tr_6, td_4, td_2
12	h_3,tr_5, h_5, h_6, tr_4, tr_7, tr_3, h_4, tr_6, td_4, td_2, td_1
13	h_3,tr_5, h_5, h_6, tr_4, tr_7, tr_3, h_4, tr_6, td_4, td_2, td_1, td_8
14	h_3,tr_5, h_5, h_6, tr_4, tr_7, tr_3, h_4, tr_6, td_4, td_2, td_1, td_8, p_4
15	h_3,tr_5, h_5, h_6, tr_4, tr_7, tr_3, h_4, tr_6, td_4, td_2, td_1, td_8, p_4,a_2
16	h_3,tr_5, h_5, h_6, tr_4, tr_7, tr_3, h_4, tr_6, td_4, td_2, td_1, td_8, p_4,a_2, td_3
17	h_3,tr_5, h_5, h_6, tr_4, tr_7, tr_3, h_4, tr_6, td_4, td_2, td_1, td_8 ,p_4,a_2, td_3, p_3
18	h_3,tr_5, h_5, h_6, tr_4, tr_7, tr_3, h_4, tr_6, td_4, td_2, td_1, td_8 ,p_4,a_2, td_3, p_3, p_5
19	h_3,tr_5, h_5, h_6, tr_4, tr_7, tr_3, h_4, tr_6, td_4, td_2, td_1, td_8 ,p_4,a_2, td_3, p_3, p_5, a_4
20	h_3,tr_5, h_5, h_6, tr_4, tr_7, tr_3, h_4, tr_6, td_4, td_2, td_1, td_8 ,p_4,a_2, td_3, p_3, p_5, a_4, a_1
21	h_3,tr_5, h_5, h_6, tr_4, tr_7, tr_3, h_4, tr_6, td_4, td_2, td_1, td_8 ,p_4,a_2, td_3, p_3, p_5, a_4, a_1, w_3
22	h_3,tr_5, h_5, h_6, tr_4, tr_7, tr_3, h_4, tr_6, td_4, td_2, td_1, td_8 ,p_4,a_2, td_3, p_3, p_5, a_4, a_1, w_3, a_3
23	h_3,tr_5, h_5, h_6, tr_4, tr_7, tr_3, h_4, tr_6, td_4, td_2, td_1, td_8 ,p_4,a_2, td_3, p_3, p_5, a_4, a_1, w_3, a_3, w_2
24	h_3,tr_5, h_5, h_6, tr_4, tr_7, tr_3, h_4, tr_6, td_4, td_2, td_1, td_8 ,p_4,a_2, td_3, p_3, p_5, a_4, a_1, w_3, a_3, w_2, r_{13}
25	h_3,tr_5, h_5, h_6, tr_4, tr_7, tr_3, h_4, tr_6, td_4, td_2, td_1, td_8 ,p_4,a_2, td_3, p_3, p_5, a_4, a_1, w_3, a_3, w_2, r_{13}, w_4
26	h_3,tr_5, h_5, h_6, tr_4, tr_7, tr_3, h_4, tr_6, td_4, td_2, td_1, td_8 ,p_4,a_2, td_3, p_3, p_5, a_4, a_1, w_3, a_3, w_2, r_{13}, w_4, r_1
27	h_3,tr_5, h_5, h_6, tr_4, tr_7, tr_3, h_4,tr_6,td_4,td_2,td_1,td_8,p_4,a_2, td_3, p_3, p_5, a_4, a_1, w_3, a_3, w_2, r_{13}, w_4, r_1, w_1
28	h_3,tr_5, h_5,h_6,tr_4,tr_7,tr_3 h_4, tr_6, td_4, td_2, td_1, td_8 ,p_4,a_2, td_3, p_3,p_5,a_4,a_1,w_3,a_3,w_2, r_{13}, w_4, r_1, w_1, d_1
29	h_3,tr_5,h_5,h_6,tr_4,tr_7,tr_3,h_4, tr_6, td_4, td_2, td_1, td_8 ,p_4,a_2, td_3, p_3,p_5,a_4,a_1,w_3,a_3, w_2,r_{13},w_4,r_1,w_1,d_1,w_5
30	h_3, tr_5, h_5, h_6,tr_4,tr_7, tr_3,h_4,tr_6,td_4,td_2,td_1,td_8,p_4,a_2,td_3,p_3,p_5,a_4,a_1,w_3,a_3,w_2,r_{13},w_4,r_1,w_1,d_1,w_5,p_1

Fig. 1. The study region showing the Keane Bridge station of Surma River, Sylhet, Bangladesh

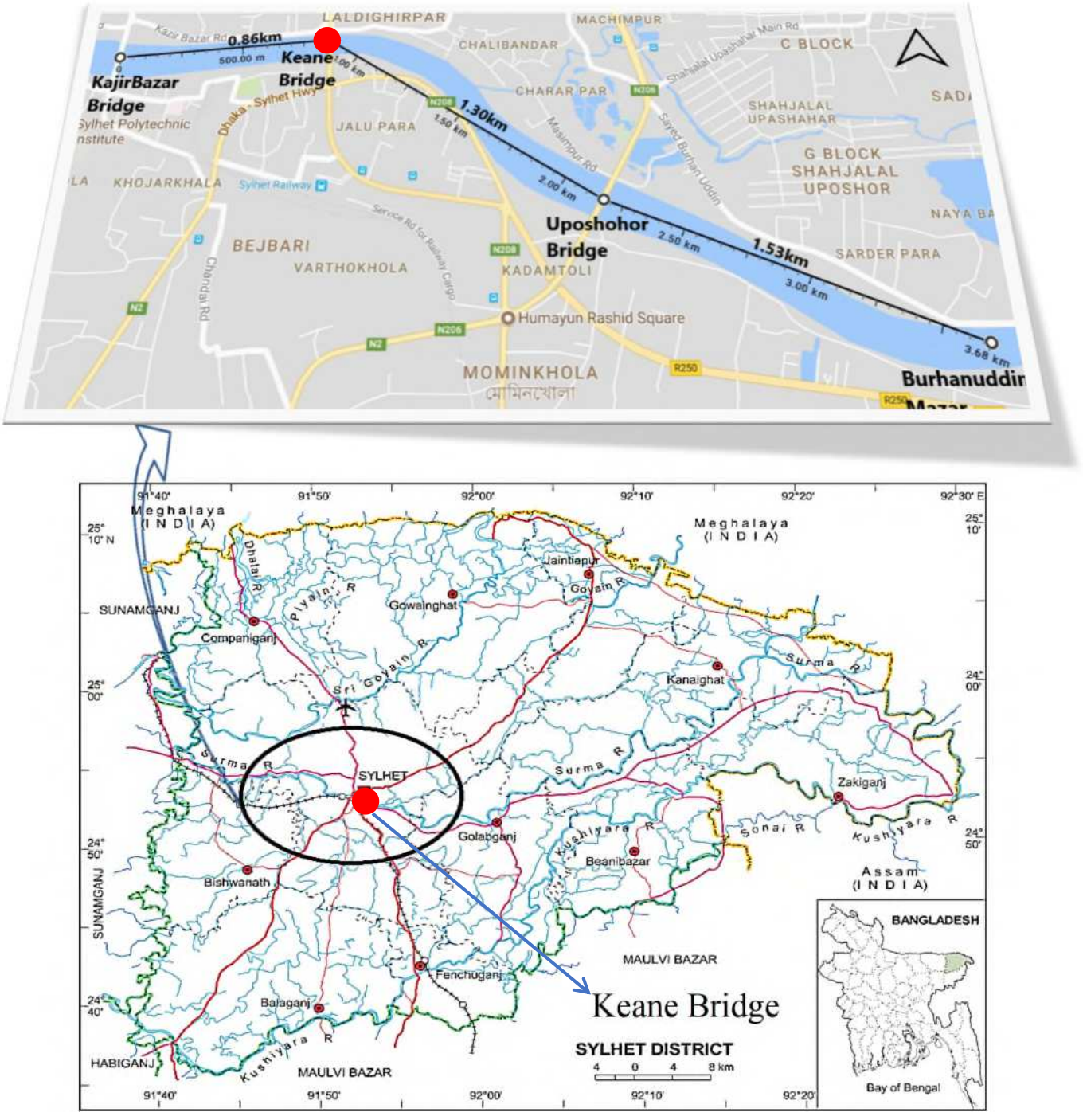
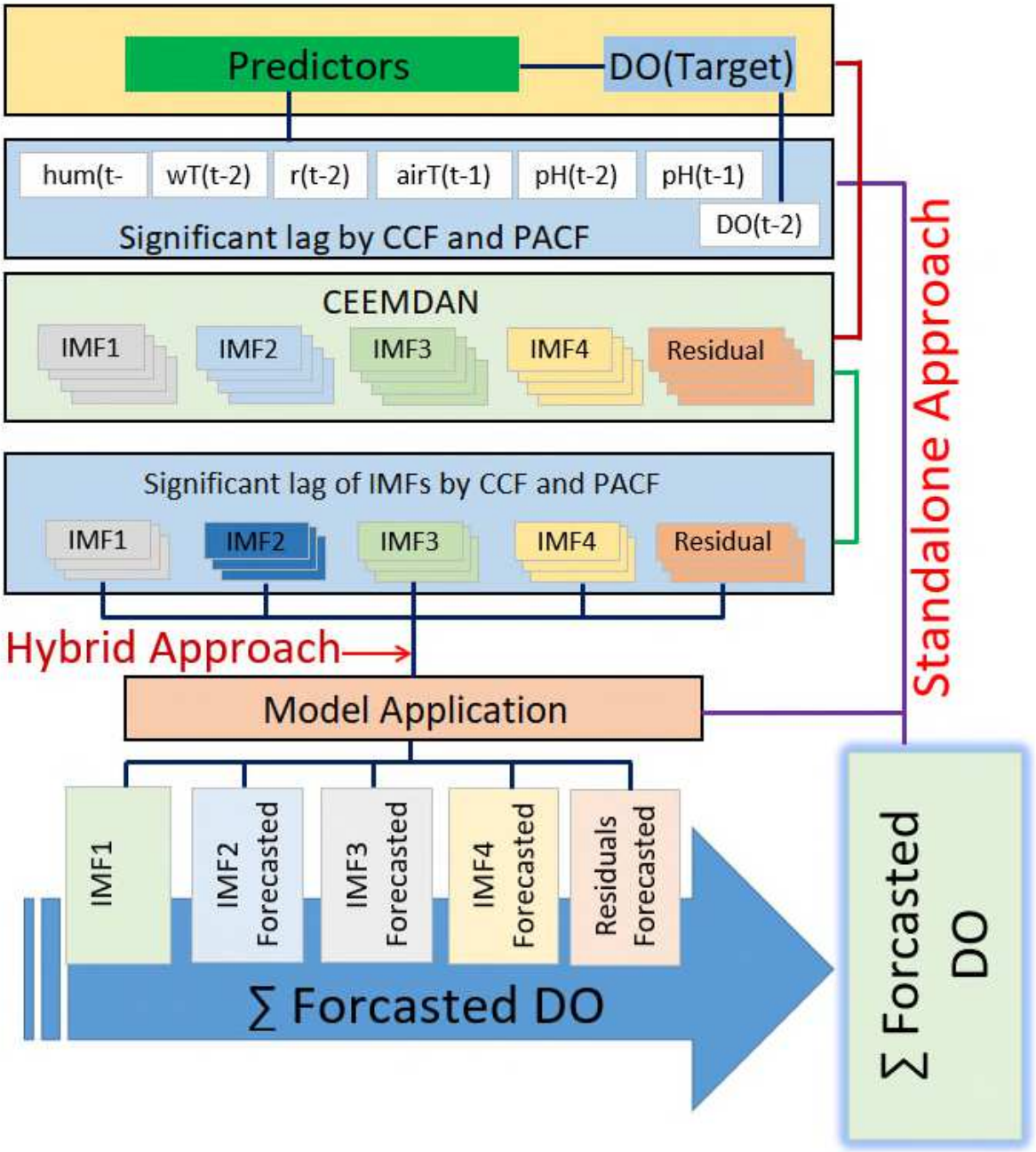


Fig. 2. Workflow detailing the steps in the model designing phase, as for the proposed hybrid CEEMDAN-MARS predictive models. Note: *IMF*= Intrinsic Mode Function, *CCF* = Cross-Correlation Functions, *PACF* = partial autocorrelation function, *CEEMDAN* = complete ensemble empirical mode decomposition with adaptive noise and *DO* = Dissolved Oxygen (mg/l)



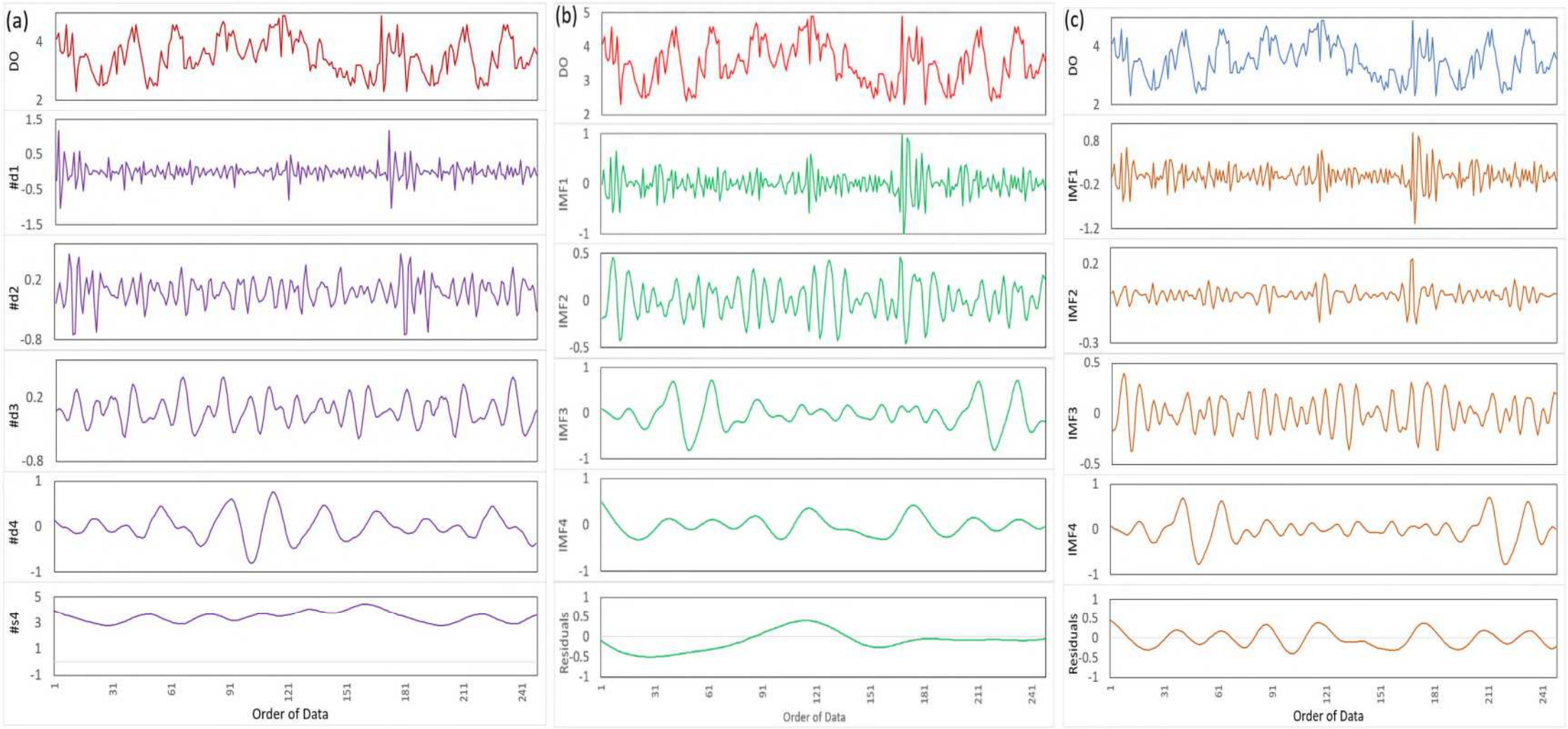


Fig.3. Time series of the a) maximum overlap discrete wavelet coefficient (MODWC) of Dissolved Oxygen using MODWT, and intrinsic mode functions (IMFs) and the residual components after decomposing the DO in the training period using b) CEEMDAN and c) EEMD. The time series of the actual DO is plotted at the top of the figure.

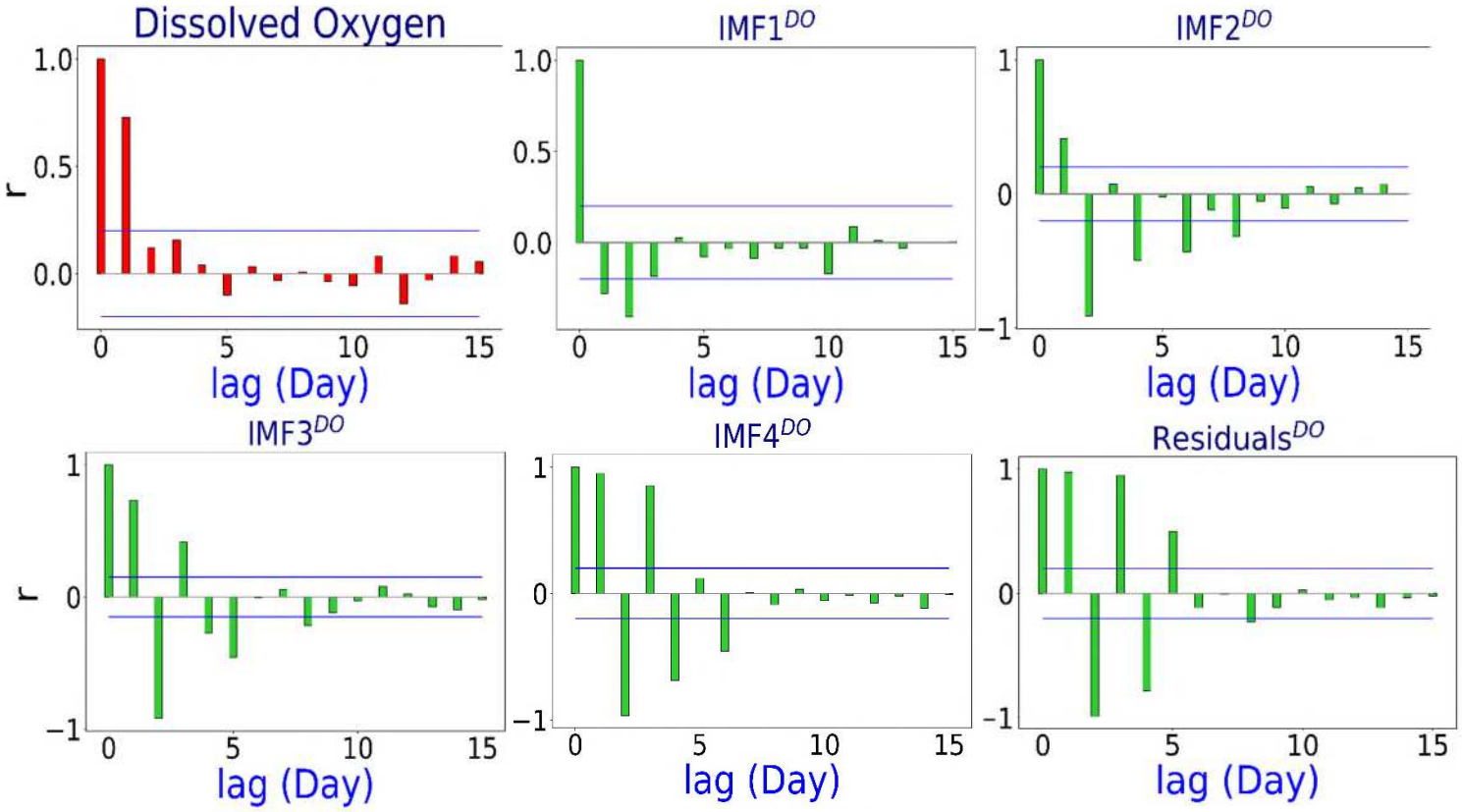


Fig. 4. Partial autocorrelation function (PACF) plot of the DO time series exploring the antecedent behaviour in terms of the lag of daily DO. The blue line in the figures indicates the ±95% confidence level.

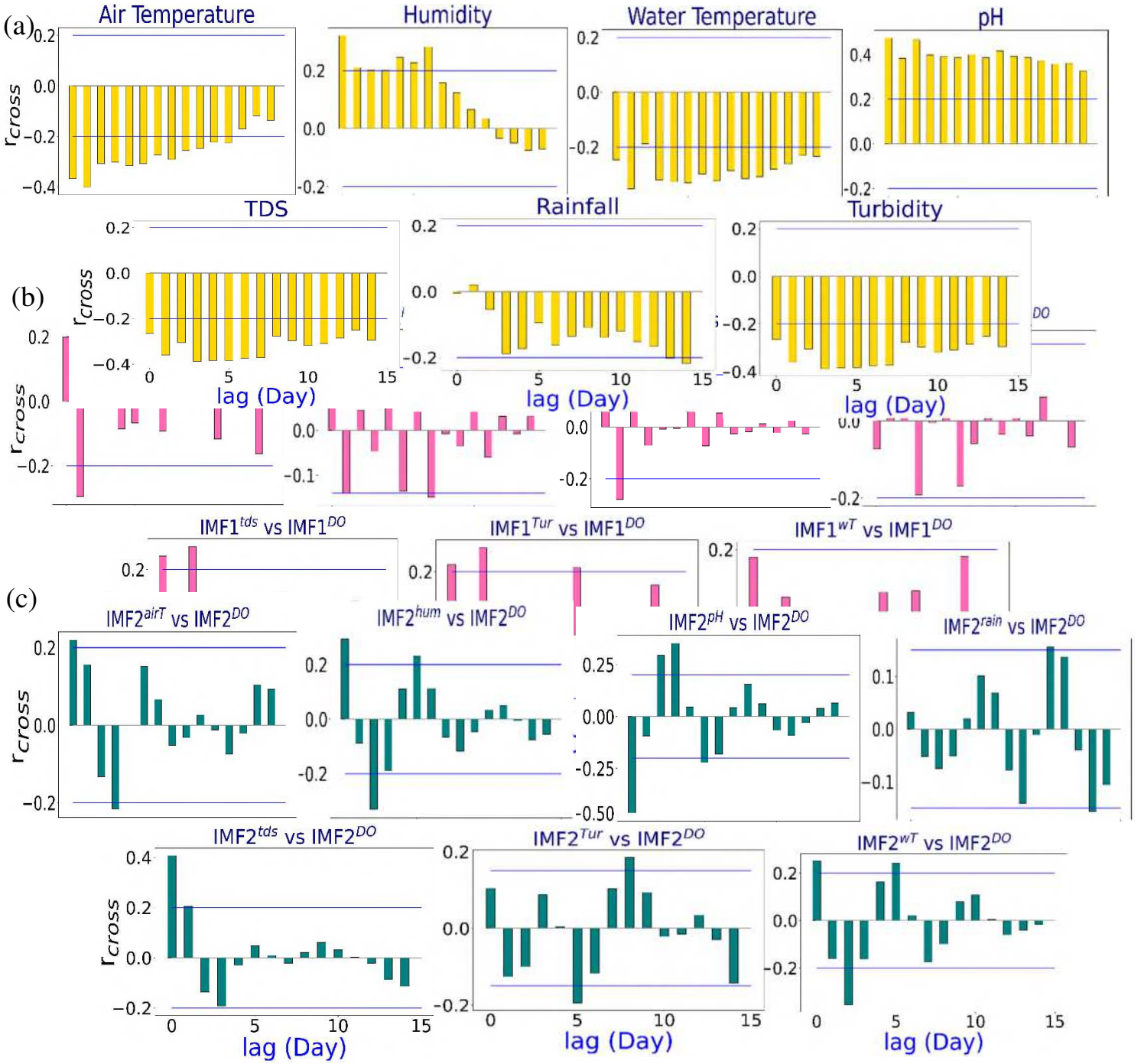

(a)
Air Temperature
Humidity
Water Temperature
pH
(b)
TDS
Rainfall
Turbidity
r_{cross}
lag (Day)
$IMF1^{tds}$ vs $IMF1^{DO}$
$IMF1^{Tur}$ vs $IMF1^{DO}$
$IMF1^{wT}$ vs $IMF1^{DO}$
(c)
$IMF2^{airT}$ vs $IMF2^{DO}$
$IMF2^{hum}$ vs $IMF2^{DO}$
$IMF2^{pH}$ vs $IMF2^{DO}$
$IMF2^{rain}$ vs $IMF2^{DO}$
$IMF2^{tds}$ vs $IMF2^{DO}$
$IMF2^{Tur}$ vs $IMF2^{DO}$
$IMF2^{wT}$ vs $IMF2^{DO}$

Fig. 5. An analysis of the statistically significant cross-correlation function plots of a) actual variables vs. DO, b) IMF1 of all variables vs. IMF1 of DO, c) IMF2 of all variables vs. IMF2 of DO, d) IMF3 of all variables vs. IMF3 of DO, e) IMF4 of all variables vs. IMF4 of DO, f) residuals of all variables vs. residuals of DO.

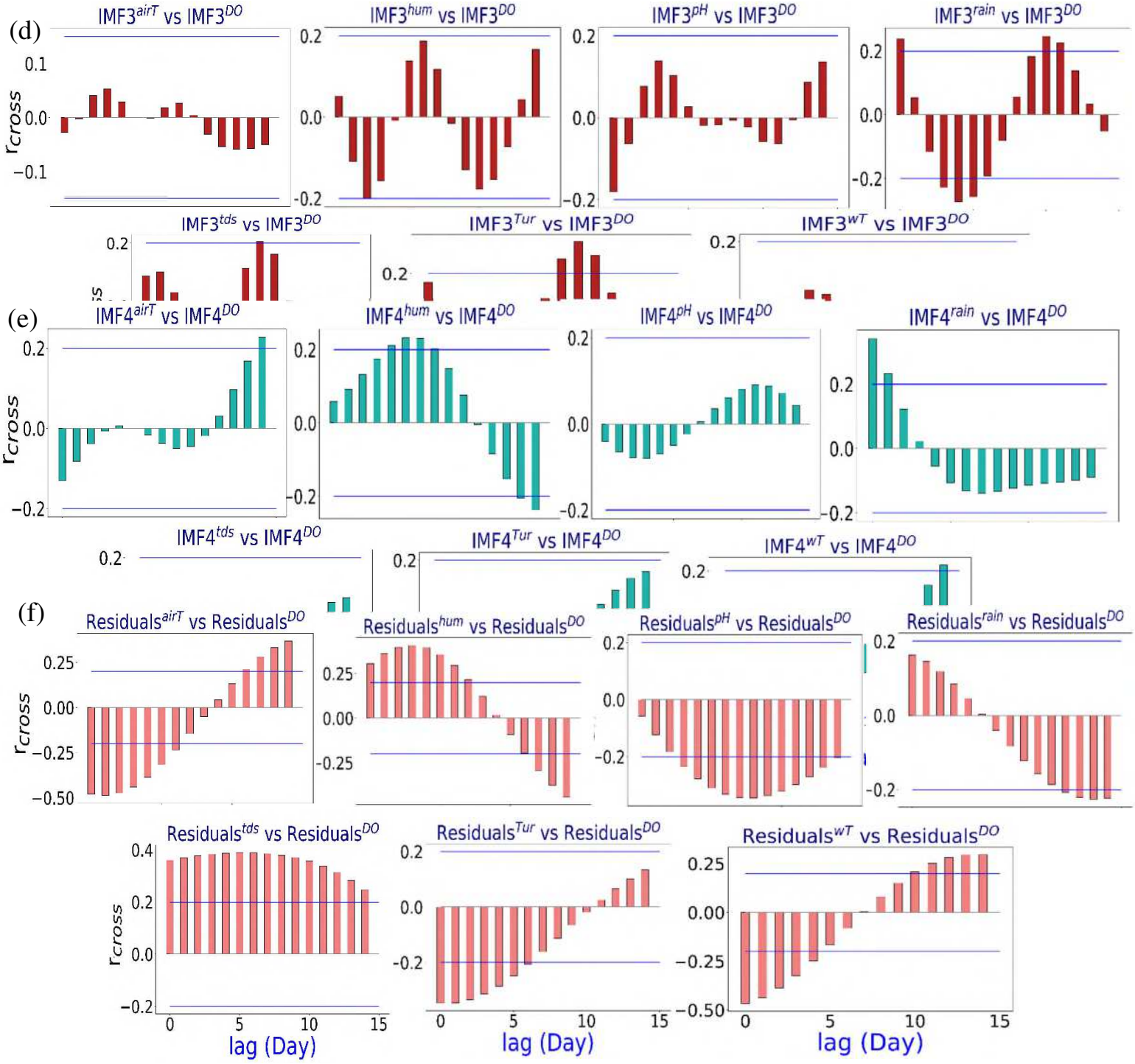


Fig.5. (Continued)

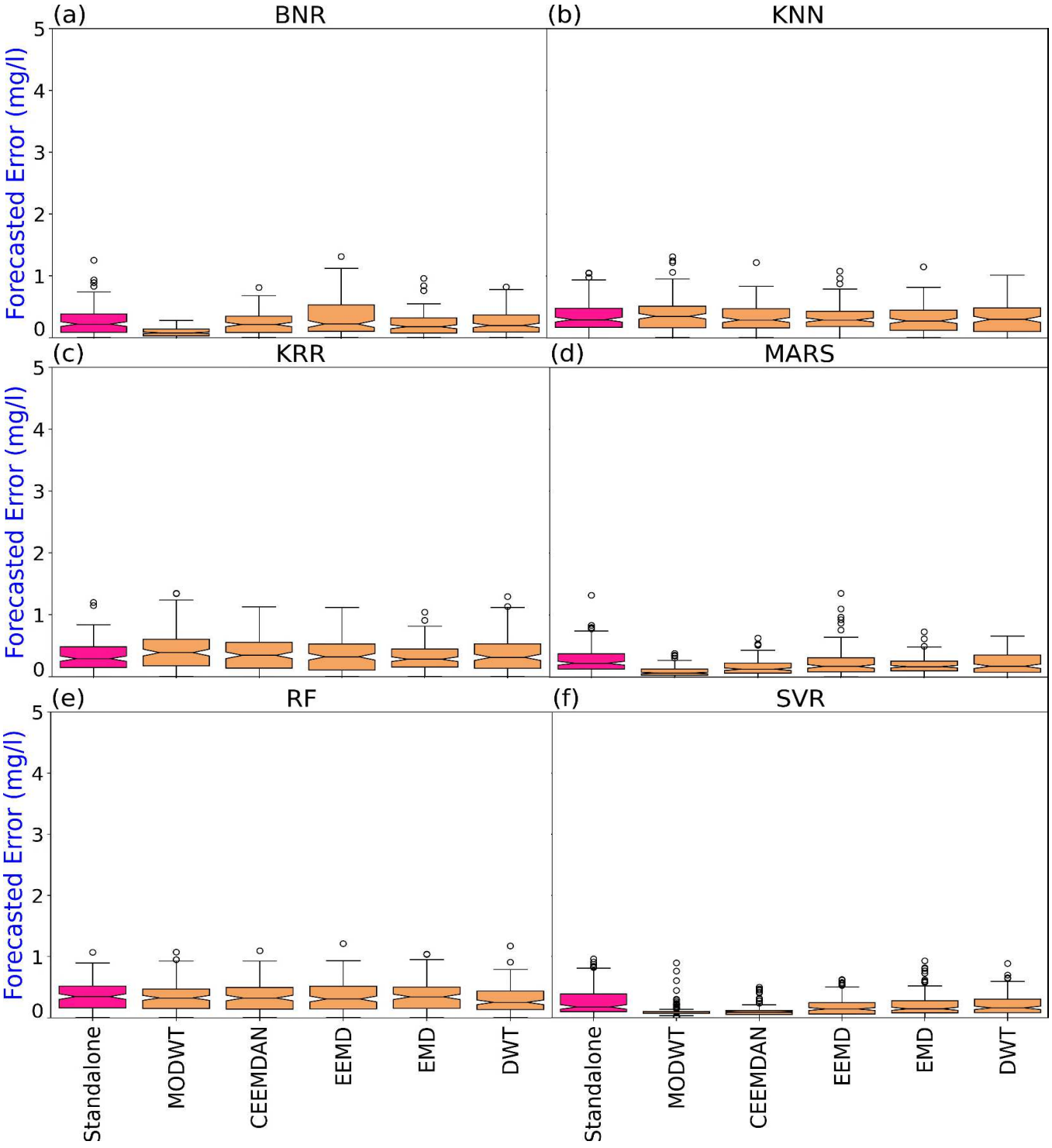


Fig. 6. Box plots of hybrid models (MODWT-MARS) and their respective standalone counterparts (i.e., MARS, BNR, KRR, KNN, RNN, and SVR) in forecasting DO in comparison with the observed DO of Surma River.

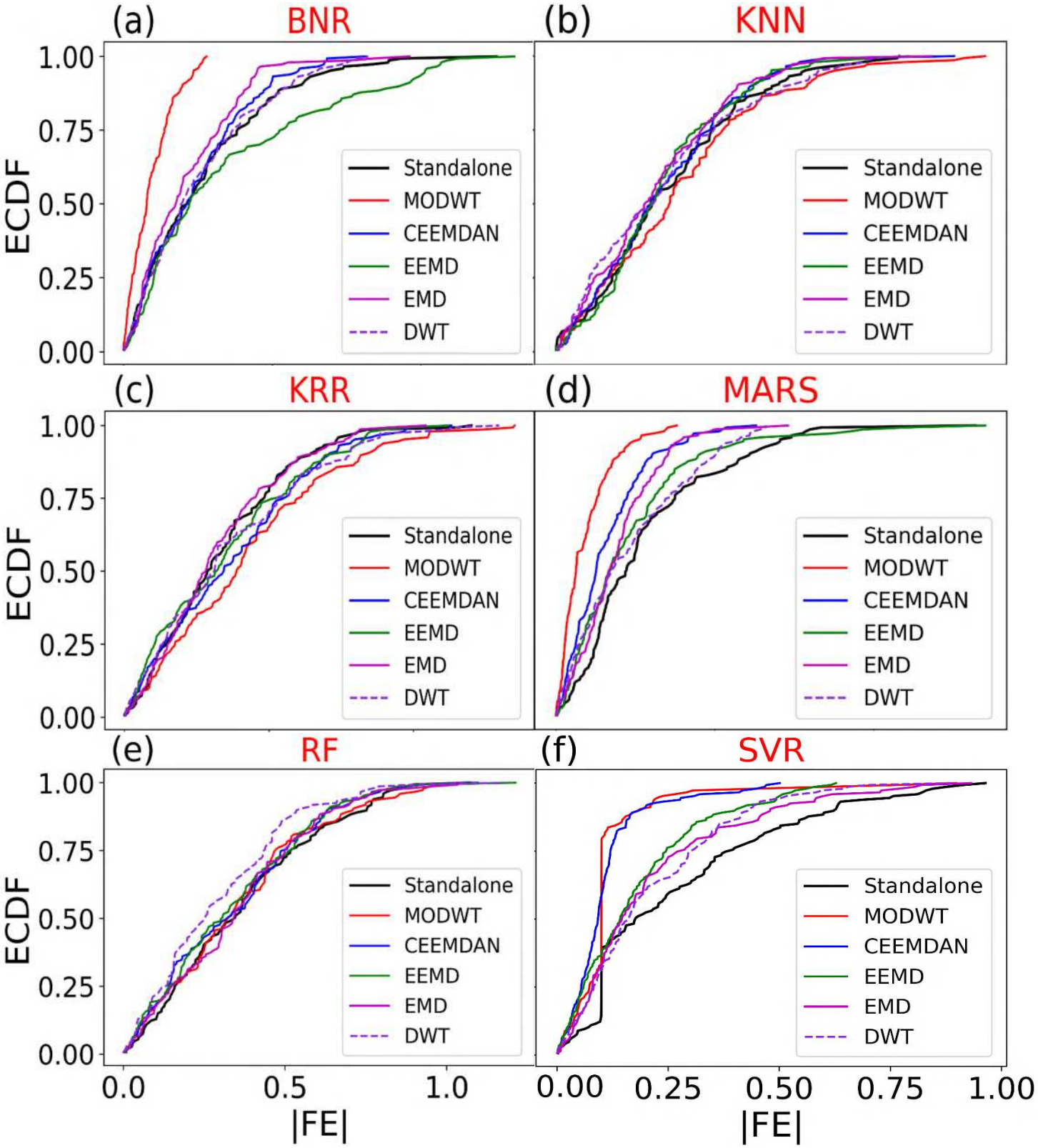

(a)
BNR
(b)
KNN
(c)
KRR
(d)
MARS
(e)
RF
(f)
SVR
ECDF
|FE|
Standalone
MODWT
CEEMDAN
EEMD
EMD
DWT
0.00
0.25
0.50
0.75
1.00
0.0
0.5
1.0

Fig. 7. Empirical Cumulative Distribution function (CDF) of forecasted error |FE| of DO generated by the proposed MODWT-MARS and comparing models.

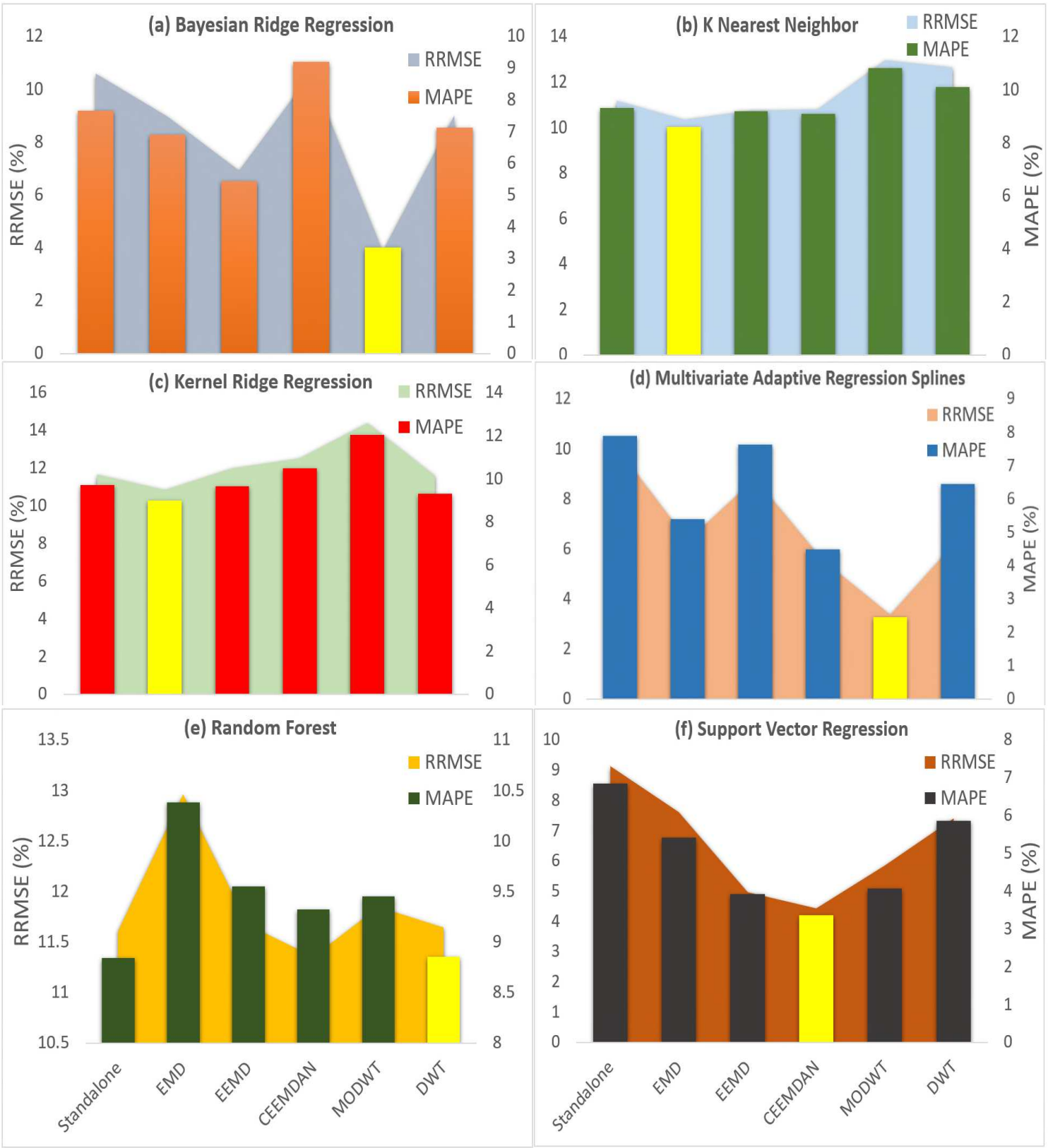

(a) Bayesian Ridge Regression
RRMSE
MAPE
RRMSE (%)
(b) K Nearest Neighbor
RRMSE
MAPE
MAPE (%)
(c) Kernel Ridge Regression
RRMSE
MAPE
RRMSE (%)
(d) Multivariate Adaptive Regression Splines
RRMSE
MAPE
MAPE (%)
(e) Random Forest
RRMSE
MAPE
RRMSE (%)
Standalone
EMD
EEMD
CEEMDAN
MODWT
DWT
(f) Support Vector Regression
RRMSE
MAPE
MAPE (%)
Standalone
EMD
EEMD
CEEMDAN
MODWT
DWT

Fig. 8. Comparison of the forecasting skill of proposed models in terms of RRMSE (%) and MAPE (%) in the testing period.

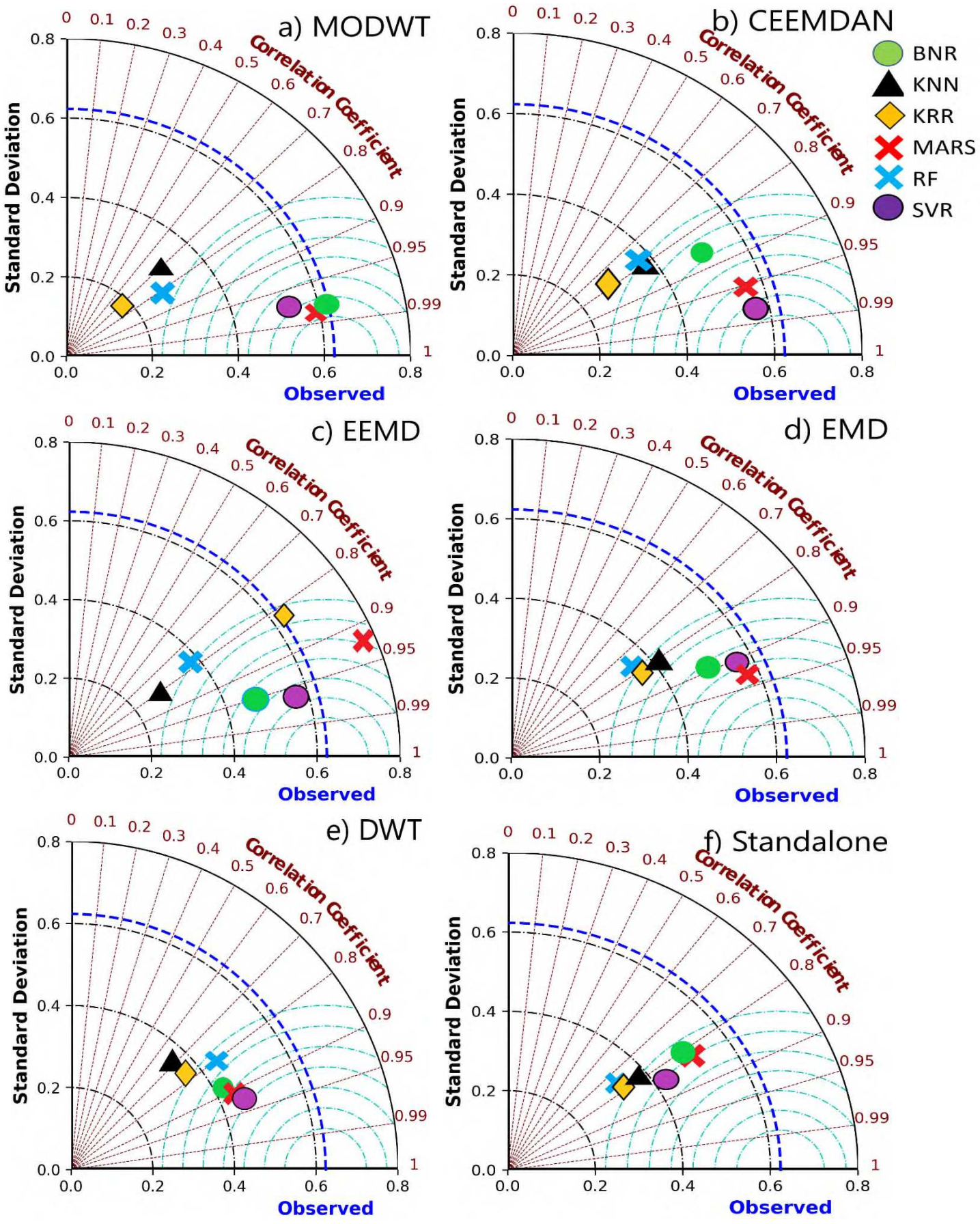

a) MODWT
b) CEEMDAN
c) EEMD
d) EMD
e) DWT
f) Standalone
BNR
KNN
KRR
MARS
RF
SVR
Correlation Coefficient
Standard Deviation
Observed
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.95
0.99
1
0.0
0.2
0.4
0.6
0.8

Fig. 9. Tylor diagram representing correlation coefficient together with the standard deviation difference for proposed hybrid models vs. benchmark models.

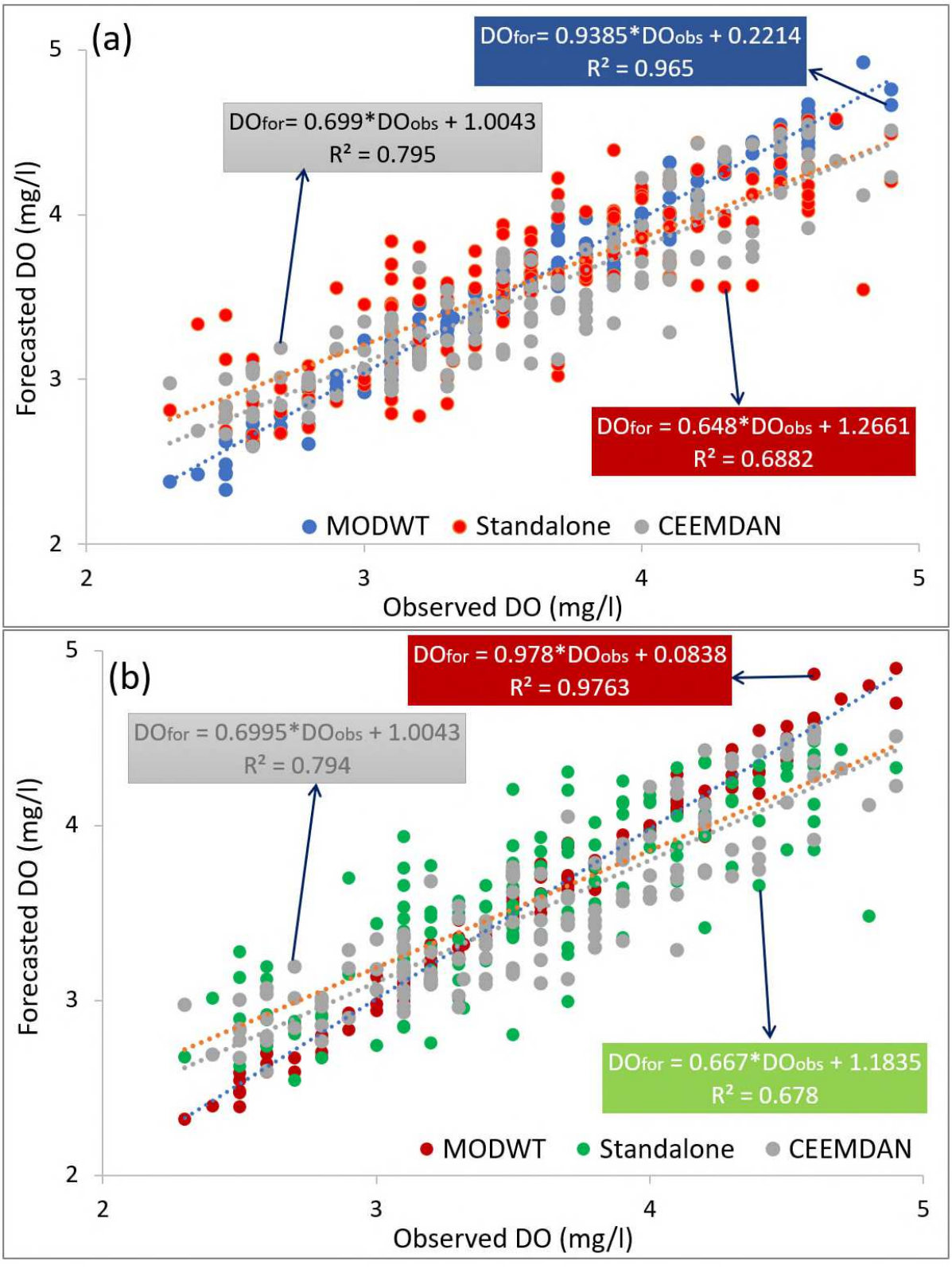


Fig.10. Scatter plot of forecasted vs observed DO, using a) Bayesian Ridge Regression and b) Multiple Adaptive Regression Splines model using MODWT and CEEMDAN decomposition. A least square regression line and coefficient of determination (R^2) with a linear fit equation are shown in each sub-panel.

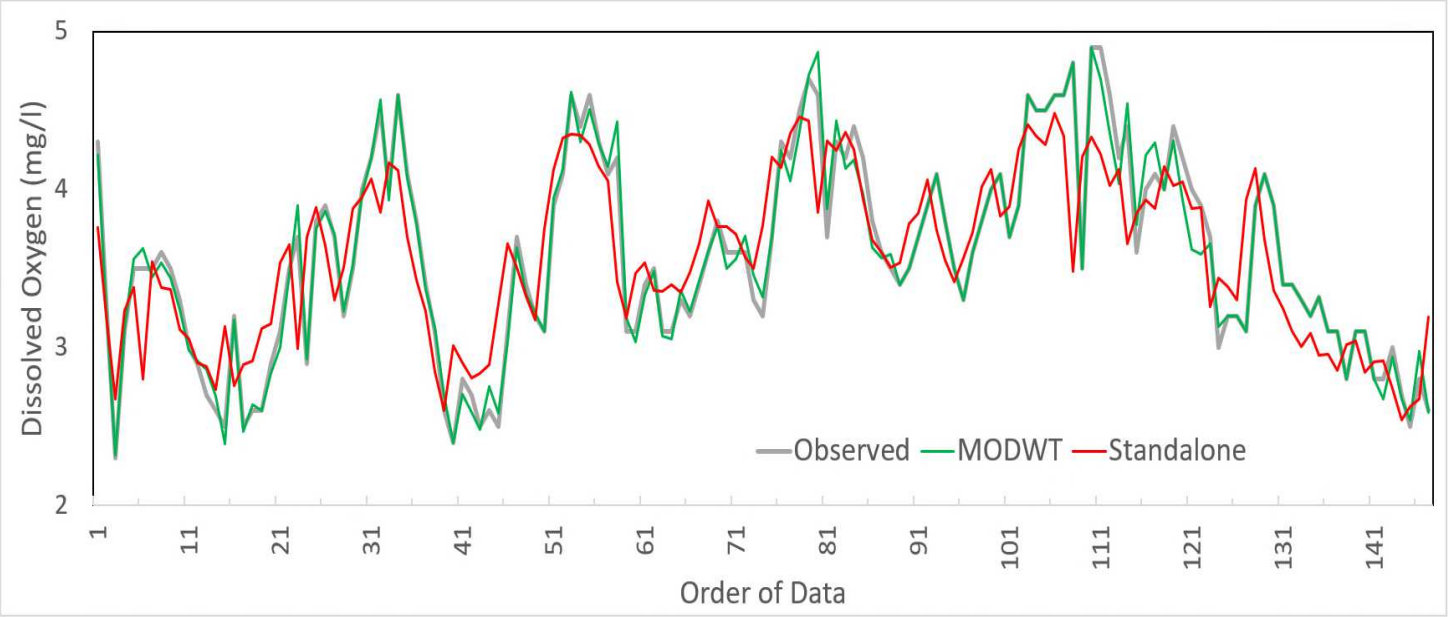


Fig. 11. Comparison between Forecasted DO and Observed DO during model testing using MODWT-MARS and Standalone MARS model.

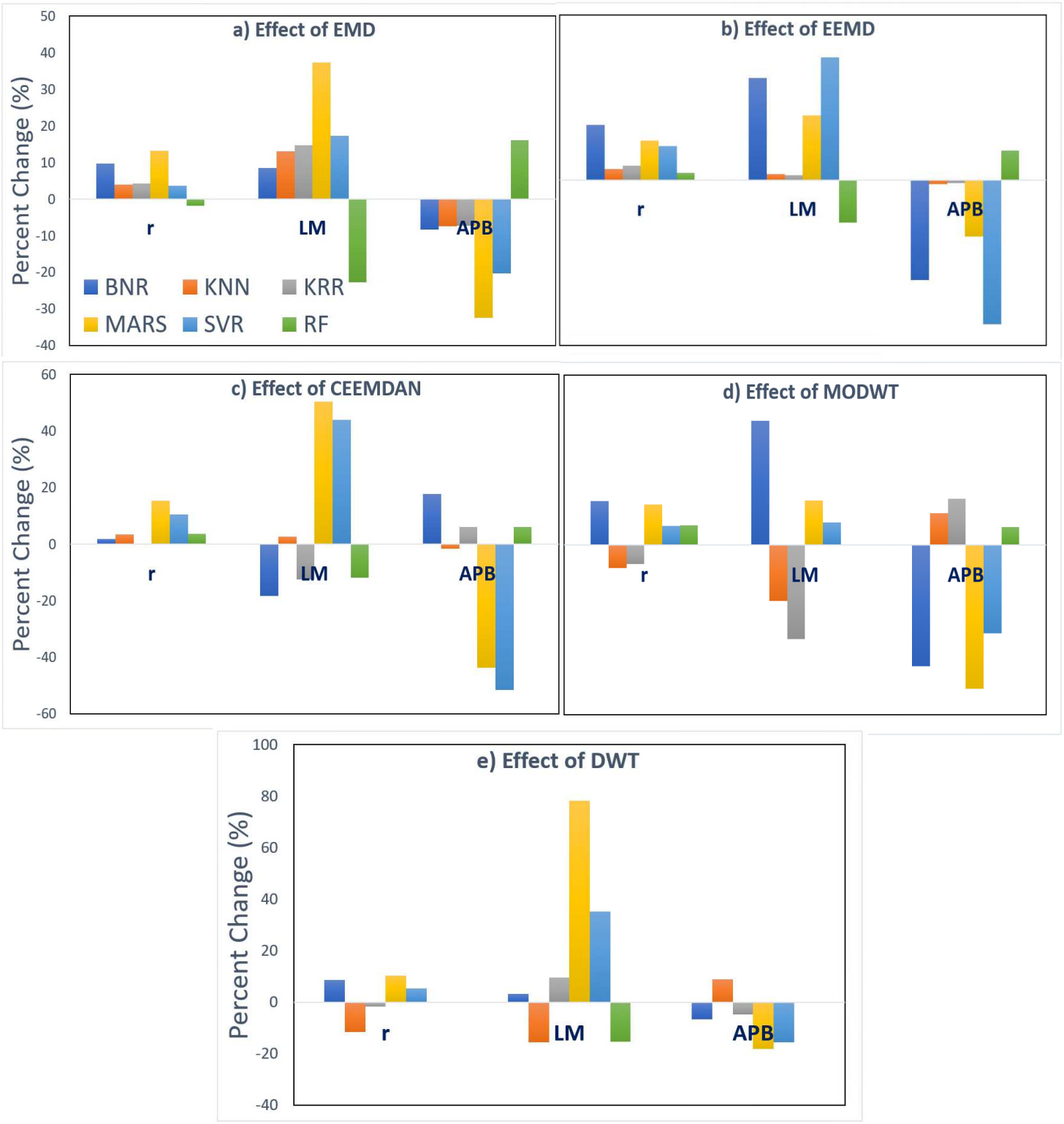


Fig. 12. Effect of (a) EMD, (b) EEMD, (c) CEEMDAN, (d) MODWT, and (E) DWT of the performance of six models based on *r*, *LM,* and *APB*.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- DOpaperMasrurSupplimentary.docx