A decomposition based Long short term memory model (LSTM) for reservoir inflow forecasting

Kshitij Tandon[1], Subhamoy Sen[2*]

[1] *PhD student, i4S Laboratory, Indian Institute of Technology Mandi, Mandi, India*
[2] *Assistant Professor, i4S Laboratory, Indian Institute of Technology Mandi, Mandi, India*

[*] *Email: subhamoy@iitmandi.ac.in*

ABSTRACT

Accurate forecasting of the reservoir inflow is crucial for operations and management of water resources. Due to the nonlinearity and nonstationarity of the real hydrological data, an empirical mode decomposition based long short term memory (EMD-LSTM) model is proposed in this paper for daily reservoir inflow forecasting upto 10 days lead time. The accuracy and performance of the model is analysed using the mean absolute error (MAE), root mean square error (RMSE) and Nash-Sutcliffe efficiency (NSE). The performance of the proposed EMD-LSTM model is compared with artificial neural network (ANN) and long short term memory (LSTM) model for 3 days, 7 days and 10 days ahead lead time respectively. Daily inflow data from 2013-2022 of the Bhakra reservoir located on river Sutlej in Himachal Pradesh, India is used to demonstrate the proposed model. The overall results of the model were highly encouraging in terms of having Nash-Sutcliffe efficiency of upto 0.94 in validation stage for 10 days ahead forecast as compared to the ANN and LSTM models. Therefore, the model can provide useful information when the models are used for decision making and can ensure safe operations of reservoir systems.

Keywords: Reservoir Inflow, Artificial neural network, Long short term memory, Empirical mode decomposition

1. INTRODUCTION

The forecast of reservoir inflow is essential for the optimal functioning and management of the reservoirs. The accuracy of the forecast plays a major role in reservoir operation, power generation, flood control and for optimal distribution of water resources in a basin. The real hydrological data shows nonlinearity and nonstationarity due to the consequences of human actions and natural factors posing a challenge in forecasting the inflow. To deal with such challenges, various forecasting methods have been proposed in the past [Ramaswamy and Saleh(2020), Bai et al.(2016)Bai, Chen, Xie, and Li] adopting various methodologies ranging from first principal-based to data-based time series modelling or more recently the machine learning(ML) algorithms.

Data-driven time series model such as auto-regressive moving average (ARMA). [Mohammadi et al.(2006)Mohammadi, Eslami, and Kahawita] have been proposed to forecast floods with a forecast lead-time of up to three days. These models, however, do not consider the nonlinearity and/or non-stationarity in the data. To deal with such problems, AI prediction models through ML approaches have been widely used. ANN models have been extensively applied for reservoir inflow forecasting as they can handle the non-linearity associated with the reservoir inflow data[Wu et al.(2005)Wu, Han, Annambhotla, and Bryant]. In these ML-based predictor models, difficulties do originate from the selection process of the model parameters since their impacts on the prediction precision is mostly significant. To handle time series horological data, Recurrent neural networks (RNN) have been widely used especially when the model has to be estimated using a long sequence data. RNN, however, poses the problem of vanishing gradients and the performance of RNN in time series prediction does not improved much while the required computational demand is at times substantial. Long short-term memory network (LSTM), an improvised form of RNN, solves the problem of RNN in training long sequences and vanishing gradients by applying constant error flow within special memory cells of the LSTM networks. Many studies have successfully applied LSTM models for groundwater level prediction and daily runoff prediction [Zhang et al.(2018)Zhang, Zhu, Zhang, Ye, and Yang, Zuo et al.(2020)Zuo, Luo, Wang, Lian, and He]

The runoff process is a complex time series with multi-scale laws. The above mentioned prediction methods, however, do not take into consideration the random nature, trend and periodicity of the real hydrological data, thereby renders the AI prediction models sometimes incompetent. To extract such characteristics from the data, extensive research on decomposition-based deep learning models have been employed in reservoir inflow forecasting [Tang et al.(2015)Tang, Wang, He, and Wang]. Various signal processing methods such as wavelet analysis and empirical mode decomposition (EMD) [Bai et al.(2016)Bai, Chen, Xie, and Li, Kisi et al.(2014)Kisi, Latifoğlu, and Latifoğlu] have evolved over time which couples signal decomposition with the time series analysis. For performance enhancement of a model, several decomposition methods, such as Fourier transform, empirical mode decomposition, wavelet transforms have been coupled with a neural network based regression model[Maheswaran and Khosa(2013)] in this attempt.

In this paper, EMD is coupled with LSTM to enhance the accuracy of the forecast by decomposing the inflow time series into several intrinsic mode functions (IMFs). These IMFs are then provided as an input to the LSTM model. This model is further compared with the ANN and LSTM model to check for the performance and accuracy.

2. METHODOLOGY

2.1. *Artificial Neural Network*

ANN model is a supervised machine learning algorithm wherein the underlying correlation between training data (as input) and pertinent target data or labels (as output) are modelled using several layers of interconnected neurons. The model represents a complex and nonlinear process that enables identifying the suitable labels to the input once trained extensively using a rich archive of training data. In this study, the inflow along with its two lagged values are provided as input to the model and the inflow at the desired lead time (3 days, 7 days, 10 days) are considered as output. The input data is normalized between 0 to 1 in order to improve the simulation performance. The model consists of an input layer with three nodes, two hidden layer with 32 nodes and one output layer with one node. Relu activation function is used for the ANN model. Following, the network architecture is defined through rigorous trial and error process prior to training it using a part of the available time series data.

2.2. *Long short term memory model (LSTM)*

The traditional or vanilla Recurrent neural network (RNN) is a class of neural networks which uses recursive approaches to model sequential data. The output of the network at any time step t is dependent not only on the input at time step t but also on the recursive inputs before time step t. However, time series forecasting problems demands long term dependencies for which the RNN is often rendered incapable due to the much mentioned problem of vanishing and exploding gradients. Hochreiter et al [Hochreiter and Schmidhuber(1997)] proposed the Long short term memory model (LSTM) equipped with memory blocks to deal with vanishing gradients by memorizing the network parameters for long duration. This article employs this approach for modelling the time series. The model consists of one hidden layer with twelve nodes. The inflow with its two lagged value is provided as an input sequence to the model and Relu activation function is employed in the model.

2.3. *Empirical mode decomposition (EMD)*

Empirical mode decomposition proposed by Huang et al. 1998 is a signal decomposition technique developed especially for the non-linear and non-stationary time series data. The essence of EMD is to decompose the time series into finite number of intrinsic mode functions (IMFs) and a residual. The major advantage of EMD technique is that the decomposed components are the time series data that oscillate about zero, therefore can be effectively analysed. Also, the decomposed IMFs and the residue can be reassembled to realize the original time series without any information loss. By principle, IMFs should satisfy the following two criteria,

1. The number of zeros and extremes of the entire IMF data set should be same, and
2. The mean value of the envelope defined by local maxima and local minima is zero

The original sequence is decomposed into intrinsic mode functions (IMFs) and a residual as follows:

$$x_k = \sum_{j=1}^{n} c_j(t) + r_n(t) \tag{1}$$

where, $x(t)$ is the original time series, $\sum_{j=1}^{n} c_j(t)$ are the IMFs and $r_n(t)$ is the residual.

In the proposed model, EMD is combined with LSTM wherein instead of the raw data, the highly correlated IMFs are assumed as input with a presumption that these IMFs store the majority of the information required for the prediction and the noise or random nature in the data can be judiciously kept out of the modelling by removing the corresponding IMFs. The flow chart describing the proposed methodology is shown in the Figure. 1.

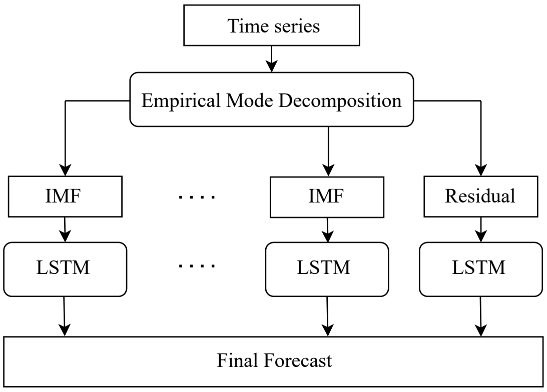

Figure 1. Flowchart describing the proposed EMD-LSTM model

2.4. *Model Performance Evaluation*

The time series model once trained needs to be validated for its prediction ability which is undertaken through several error and statistical indices, discussed in the following.

2.4.1. Nash-Sutcliffe efficiency (NSE) Nash-Sutcliffe efficiency (NSE) estimates the goodness of fit. It is evaluated by the following equation:

$$NSE = 1 - \frac{\sum_{i=1}^{N}(y_i^{act} - y_i^{pred})^2}{\sum_{i=1}^{N}(y_i^{act} - y_{act}^{avg})^2} \tag{2}$$

where, y_i^{pred}, y_i^{act} are the predicted and actual values of inflow, N is the number of samples, y_{act}^{avg} is the mean value of measured sample.

2.4.2. Mean absolute error (MAE) The mean absolute error depicts the average of the absolute error between the actual observed value y_i^{act} and the predicted value y_i^{pred}. It is given as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i^{act} - y_i^{pred}| \tag{3}$$

2.4.3. Root mean square error (RMSE) The Root mean square error (RMSE) measures the prediction performance of a model. The smaller the RMSE, the better is the accuracy of the prediction model. RMSE is evaluated as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i^{act} - y_i^{pred})^2} \tag{4}$$

3. STUDY AREA AND DATA DESCRIPTION

The Bhakra Nangal dam is the second highest concrete gravity dam in Asia located across the river Satluj in the state of Himachal Pradesh, India. The upstream catchment area of the dam is 56,980 km^2. The Rainfall in the catchment area varies over the basin with an annual average of 875 mm. The reservoir, Govindsagar has a gross capacity of 9876 m^3 with live storage capacity of 7814 m^3 above dead storage level of 445.62 m. The area covered by the reservoir is 168.35 km^2 when full. The total runoff is 16,441 m^3 for a mean year. River Satluj, which originates from the Himalayas, forms a major source of water in the region. Due to the climate change and rainfall pattern change, the magnitude of the streamflow is of concern for water management. The river flows are high from June to September because of the monsoon rains and melting of snow. Eventually, accurate forecasting of the inflow for such huge reservoir is imperative for the reservoir operation, flood control and also ensures the structural safety of the dam at the same time.

The real time observed daily inflow for the period 2013-2022 is available on the website of Bhakra Beas Management Board (BBMB). An univariate time series forecasting is performed to forecast the inflow for different lead-times (3, 5 and 7 days) with an objective of flood forecasting, reservoir safety and health assessment. The daily inflow data is split into training (calibration) and testing (validation) with 70:30 ratio each respectively as shown in Figure. 2

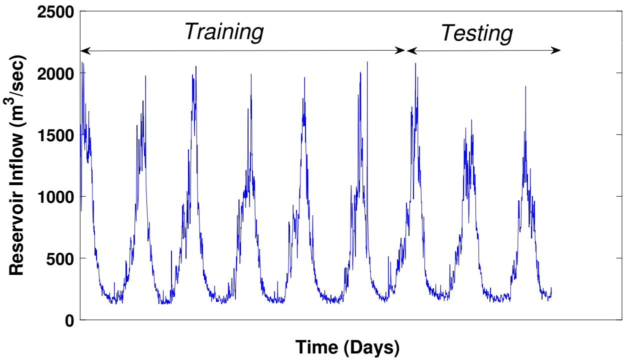

Figure 2. Reservoir Inflow Data of Bhakra reservoir from 2013 to 2022

4. RESULTS AND DISCUSSION

The time series data of the reservoir inflow is decomposed into nine IMF components and one residue in the order of decreasing frequency by using EMD. The last component is the residue which represents the general trend of the time series. All the extracted IMF components are not shown here for brevity. In order to explore the physical phenomenon of the extracted IMF components, a cross correlation study is carried out between each IMF component and the original time series data. As shown in Table I, IMFs 2, 3, 4, 5 and 6 have shown a significant positive correlation with the original data.

IMF	Cross- correlation coefficient
1	0.0285
2	0.1731
3	0.1691
4	0.4301
5	0.8199
6	0.1129
7	0.0829
8	0.0484
9	0.0757
Residual	0.0742

Table I. Correlation of IMFs with original data

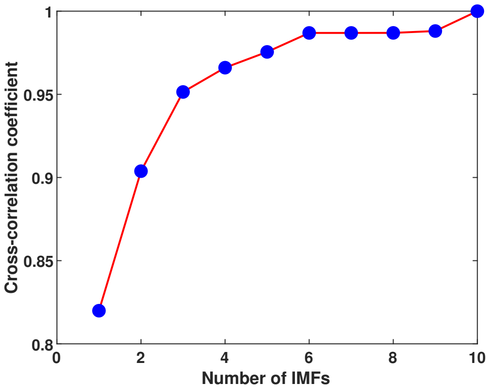

Figure 3. Variation in cross-correlation coefficient with the number of IMFs

In order to reduce the impact of noise (or random nature present in the data) on forecast and to maintain economy in computation, selection of the vital IMF components is essential. The selection is done on the basis of maximum positive correlation of the concerned IMF with the original time series data. Figure. 3 depicts the variation of correlation coefficients with the number of IMFs considered. The correlation increases with the increase in number of IMFs (arranged in descending correlation coefficient from Table I). The curve shows an increasing correlation up to six IMF's and becomes constant thereafter allowing identification of an elbow to decide the practical numbers of IMFs to be considered. Accordingly, six IMFs are considered for our proposed EMD-LSTM model. Figure. 4 shows the graphical representation of the six IMFs employed for EMD-LSTM model.

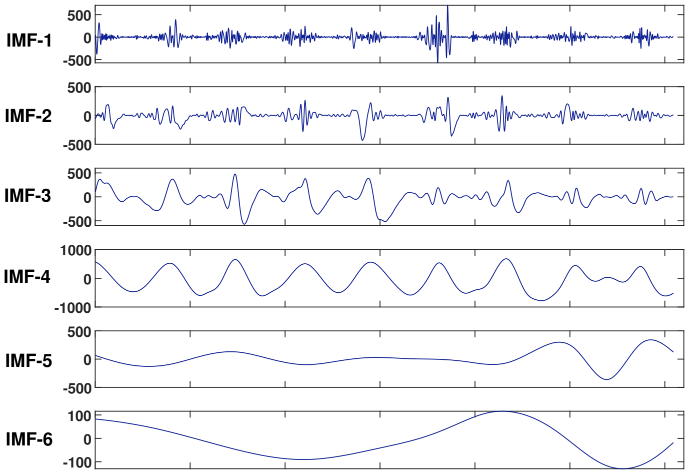

Figure 4. The intrinsic mode functions (IMFs) employed for prediction in EMD-LSTM model.

Figures 5 6 7 shows the forecast plots along with the scatter plots of the observed inflow and predicted inflow from the three models (ANN, LSTM and EMD-LSTM) for three lead times (3, 7 and 10 days). The predictive performance of the model decreases as the lead time increases. The comparison results reveal that the EMD-LSTM model provides relatively better performance and illustrates the good fit of the model against the observation than the LSTM and ANN model. The ANN model despite captures the underline physics of the system from the observed data for most of

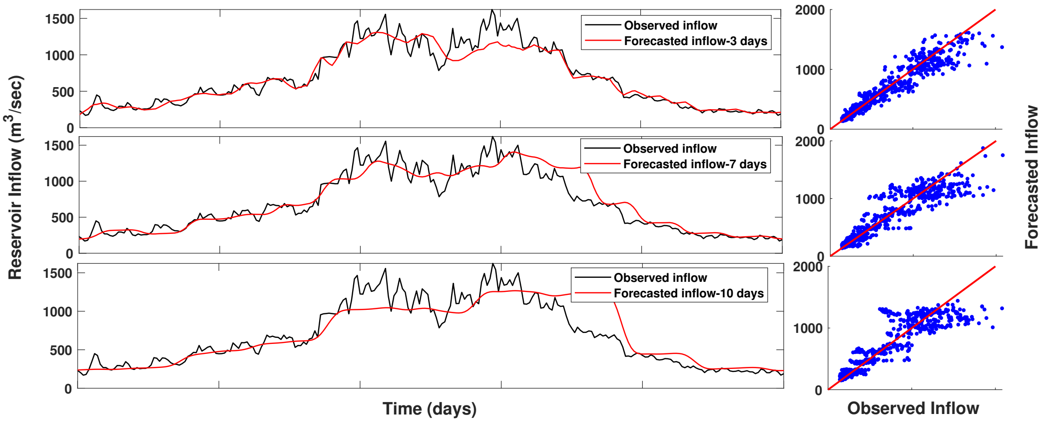

Figure 5. Reservoir inflow forecast for 3 days,7 days and 10 days ahead respectively by ANN model

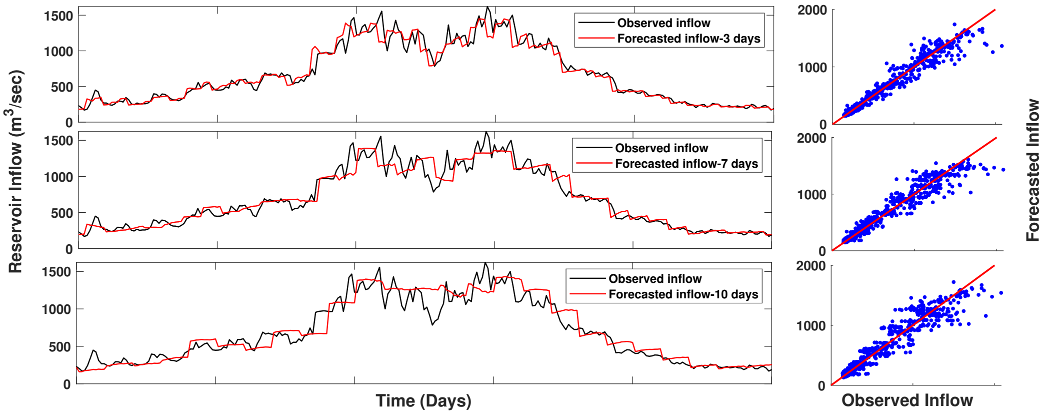

Figure 6. Reservoir inflow forecast for 3 days,7 days and 10 days ahead respectively by LSTM model

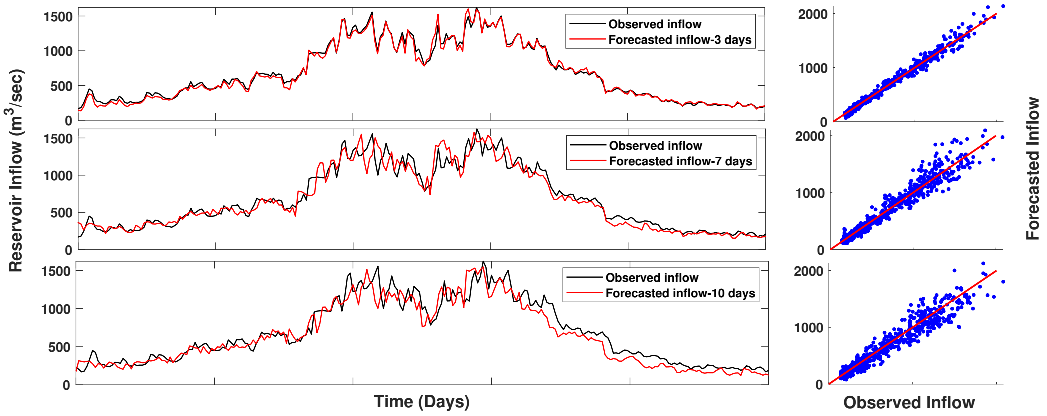

Figure 7. Reservoir inflow forecast for 3 days,7 days and 10 days ahead respectively by EMD-LSTM model

the parts yet it does fails to replicate the complete physical process for all data ranges. Eventually, ANN model fails to capture the variation for higher lead times. Overall, the scatter plot illustrates that the EMD-LSTM model can predict the inflow much better than its alternatives.

Table II depicts three performance indices (NSE, RMSE and MAE) for the performance evaluation of the three competing models for inflow prediction.The NSE values range from 0.84 to 0.92 for ANN model, 0.93 to 0.96 for the LSTM model and 0.94 to 0.98 for EMD-LSTM model for 3 days , 7 days and 10 days lead times respectively. The RMSE increased significantly with the increase in the lead time for all candidate models. Nevertheless, EMD-LSTM yielded quite low RMSE compared to its alternatives (Figure. 8 and Table II). The MAE results further signify that the proposed model predicts more accurately and the other alternatives have more errors.

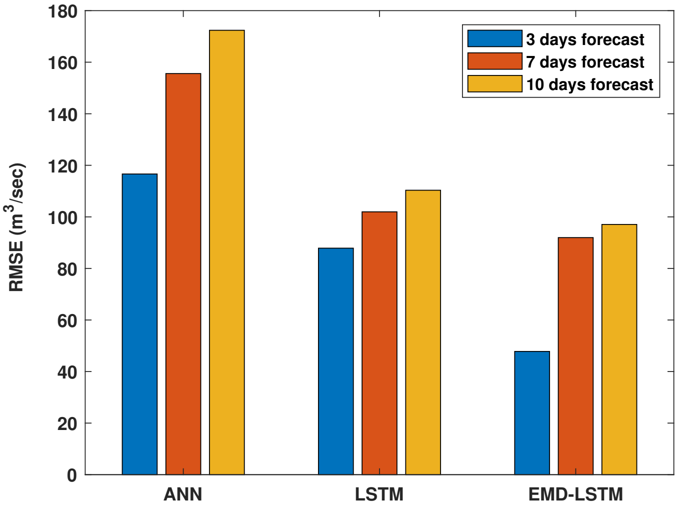

Figure 8. Comparison of root mean square error for ANN, LSTM and EMD-LSTM model respectively

	NSE			RMSE (m^3/sec)			MAE %		
Lead time (days)	3	7	10	3	7	10	3	7	10
ANN	0.92	0.87	0.84	116.63	155.59	172.38	68.75	93.94	106.57
LSTM	0.96	0.94	0.93	87.86	101.96	110.33	45.62	59.71	66.58
EMD-LSTM	0.98	0.95	0.94	47.80	91.96	97.06	31.98	59.18	73.21

Table II. Model Performance evaluation for 3 days, 7 days and 10 days ahead lead time respectively

5. CONCLUSIONS

By coupling EMD and LSTM, the proposed model significantly enhances the forecasting accuracy and prediction capacity. It deals effectively with the non stationarity and noise existing in the original time series. For small lead time, all the models perform notably well in terms of higher accuracy and less error. The proposed model performs significantly better among other models for higher lead time. ANN model is unable to capture the trend of original time series for higher lead time. The LSTM model is applicable in model fitting and forecasting. Moreover, EMD is convenient to reduce the noise and extracts original trend and periodicity of the original time series leading to better prediction results.

REFERENCES

[Ramaswamy and Saleh(2020)] V Ramaswamy and F Saleh. Ensemble based forecasting and optimization framework to optimize releases from water supply reservoirs for flood control. *Water Resources Management*, 34(3):989–1004, 2020.

[Bai et al.(2016)Bai, Chen, Xie, and Li] Yun Bai, Zhiqiang Chen, Jingjing Xie, and Chuan Li. Daily reservoir inflow forecasting using multiscale deep feature learning with hybrid models. *Journal of hydrology*, 532:193–206, 2016.

[Mohammadi et al.(2006)Mohammadi, Eslami, and Kahawita] Kourosh Mohammadi, HR Eslami, and Rene Kahawita. Parameter estimation of an arma model for river flow forecasting using goal programming. *Journal of Hydrology*, 331(1-2):293–299, 2006.

[Wu et al.(2005)Wu, Han, Annambhotla, and Bryant] Jy S Wu, Jun Han, Shastri Annambhotla, and Scott Bryant. Artificial neural networks for forecasting watershed runoff and stream flows. *Journal of hydrologic engineering*, 10(3):216–222, 2005.

[Zhang et al.(2018)Zhang, Zhu, Zhang, Ye, and Yang] Jianfeng Zhang, Yan Zhu, Xiaoping Zhang, Ming Ye, and Jinzhong Yang. Developing a long short-term memory (lstm) based model for predicting water table depth in agricultural areas. *Journal of hydrology*, 561:918–929, 2018.

[Zuo et al.(2020)Zuo, Luo, Wang, Lian, and He] Ganggang Zuo, Jungang Luo, Ni Wang, Yani Lian, and Xinxin He. Decomposition ensemble model based on variational mode decomposition and long short-term memory for streamflow forecasting. *Journal of Hydrology*, 585:124776, 2020.

[Tang et al.(2015)Tang, Wang, He, and Wang] Ling Tang, Shuai Wang, Kaijian He, and Shouyang Wang. A novel mode-characteristic-based decomposition ensemble model for nuclear energy consumption forecasting. *Annals of Operations Research*, 234(1):111–132, 2015.

[Kisi et al.(2014)Kisi, Latifoğlu, and Latifoğlu] Ozgur Kisi, Levent Latifoğlu, and Fatma Latifoğlu. Investigation of empirical mode decomposition in forecasting of hydrological time series. *Water resources management*, 28(12):4045–4057, 2014.

[Maheswaran and Khosa(2013)] R Maheswaran and Rakesh Khosa. Wavelets-based non-linear model for real-time daily flow forecasting in krishna river. *Journal of Hydroinformatics*, 15 (3):1022–1041, 2013.

[Hochreiter and Schmidhuber(1997)] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.