water

MDPI

Article

Monthly Streamflow Forecasting Using EEMD-Lasso-DBN Method Based on Multi-Scale Predictors Selection

Haibo Chu [1], Jiahua Wei [1,2,*] and Jun Qiu [1,2]

[1] State Key Laboratory of Hydroscience & Engineering, Tsinghua University, Beijing 100084, China; haibochu0613@163.com (H.C.); aeroengine@tsinghua.edu.cn (J.Q.)
[2] State Key Laboratory of Plateau Ecology and Agriculture, Qinghai University, Xining 810016, China
* Correspondence: weijiahua@tsinghua.edu.cn

Received: 3 October 2018; Accepted: 18 October 2018; Published: 22 October 2018

Abstract: For the inherent characteristics of a raw streamflow times series and the complicated relationship between multi-scale predictors and streamflow, monthly streamflow forecasting is very difficult. In this paper, an method was proposed integrating the ensemble empirical mode decomposition (EEMD), least absolute shrinkage and selection operator (Lasso) with deep belief networks (DBN) for forecasting monthly streamflow time series, which is EEMD-Lasso-DBN (ELD) method. To develop the ELD model, the raw streamflow time series was resolved into different elements, including intrinsic mode functions (IMFs) and residue series, using the EEMD technique. The predictors of each IMF element and residue were screened using the Lasso technique from a large number of candidate predictors, respectively. Then, the DBN models were built to simulate the complex relationship between the resolved elements and the selected predictors, respectively. The predicted results of the IMFs and residual series were assembled as an ensemble forecast for the raw streamflow time series and were compared with the other models. The monthly streamflow series from Tennessee, in the USA, were investigated using the ELD method. It was found that each IMF has different characteristics and physical meaning, corresponding to different predictors. The proposed ELD model can significantly improve the accuracy of monthly streamflow forecasting.

Keywords: streamflow forecasting; Lasso; deep belief networks; EEMD; Tennessee River Basin

1. Introduction

Monthly streamflow forecasting is essential for water resources' planning and management, such as dam construction, reservoir operation, and flood control [1,2]. However, streamflow forecasting is difficult because there are non-stationary characteristics of the raw monthly streamflow time series for intrinsic characteristics, a complex relationship between the impact predictors and monthly streamflow for an external environment, and multi-scale impact predictors, including atmospheric oscillations, sea surface temperatures, and precipitation. This influence also has the characteristics of uncertainty, temporal and spatial variation at different scales, and time-invariance, i.e., it may be different in different seasons [3,4]. Therefore, the selection of multi-scale predictors and forecasting methods are essential challenges that need to be solved in monthly streamflow forecasting.

The candidate predictors for streamflow forecasting has a wide range from local to global scales, and the selection of predictors has already been shown to be effective in improving the accuracy of streamflow forecasting. In previous studies, rainfall is often chosen as a predictor because it has a strong correlation with historical streamflow, however, the impact of other predictors on streamflow should not be ignored. Recently, some research has incorporated teleconnection predictors, such as

atmospheric oscillations and sea surface temperature (SST), into streamflow forecasting. Therefore, how to screen the predictors is a problem that needs to be further studied. Techniques, such as correlation and composite analysis, principal component analysis, and singular value decomposition (SVD) analysis, have been investigated to identify predictors for streamflow. Opitz-Stapleton et al. [5] used correlation and composite analysis to reveal the physical relationships between different climate factors and streamflow. Amigo et al. [6] used canonical correlation analysis to express deep insight into the link between different physicochemical properties and the concentration of metals in an estuary. Risko et al. [7] used singular value decomposition analysis to select the predictors of seasonal streamflow in West-Central Florida. Composite analysis requires limited data and is vulnerable to leveraging under the influence of a single large anomaly. SVD analysis needs evidence of coupling between the two datasets by the other methods. Lasso has been shown to be a useful technique in identifying significant predictors [8]. Previous studies attempted to identify substantial climatic factors (predictors) that drive streamflow variability in the raw time series. At different periodicity, climatic factors have different impacts on streamflow variability. In this paper, Lasso was used to select the predictors that significantly correlate with the different time series resolved from monthly streamflow representing the characteristics of a different periodicity.

Data preprocessing methods, including moving average (MA), wavelet analysis (WA), singular spectrum analysis (SSA), and principal component analysis (PCA), are useful tools for improving the forecasting accuracy by extracting more trends information and eliminating noise from monthly streamflow time series [9,10]. Wu et al. [11] investigated the performances of five models with SSA and MA on two real monthly streamflow series. Zhang et al. [12] developed six hybrid models with different combinations of wavelet analysis (WA), empirical mode decomposition (EMD), and singular spectrum analysis (SSA), and two modeling methods (i.e., the Artificial Neural Network (ANN) model and Autoregressive Integrated Moving Average (ARIMA) model). Mehr and Kahya [13] proposed a Pareto-optimal moving average multigene genetic programming (MA-MGGP) approach for single-station streamflow prediction as a parsimonious model. The EEMD, as an improved method of empirical mode decomposition (EMD), is proposed for the analysis of nonlinear and non-stationary time series, and it is useful to alleviate mode mixing issues by adding the noise to the raw data. EEMD resolves a nonlinear and complex data time series into several IMFs and a residue, and those IMFs and the residue reveal different characteristics [14–16].

The building of a model that simulates the complex relationship between the impact factors and streamflow is a significant problem that needs to be studied continuously. Classical forecasting models, such as auto-regression (AR), multiple linear regressions (MLR), and auto-regressive moving average (ARMA), are easy to set up, but cannot deal with non-linear problems, like streamflow forecasting, especially at monthly times scales [17–19]. Since the early 1990s, the ANN model has been devoted to the improvement of streamflow time series forecasting [20]. Recently, several works have been used to improve the forecasting abilities of the ANN model using EEMD. Wang et al. [21] coupled the ANN model with EEMD to predict the annual runoff. Nevertheless, ANN has weaknesses in determining the connection weight and the number of hidden layer nodes, and it also suffers from a slow convergence speed and is easy to trap in a local optimum. The ddeep belief network (DBN) is one of the most popular deep learning models, which has multiple hidden layers that can effectively and flexibly simulate the nonlinear and complex relationships compared with the ANN [22]. The EEMD and DBN should be coupled and adopted in the streamflow forecasting to enhance the accuracy.

The primary objective of this paper is to propose a new data-driven approach from predictors screening to streamflow forecasting, which integrates the Lasso, EEMD, and DBN. In this study, the EEMD method was applied to decompose the raw monthly streamflow time series into IMF elements, and one residual element. Then, the Lasso method was employed to select the predictors that significantly correlated with each extracted IMF element and the residual element of monthly streamflow. With the screening predictors as the models' input, the DBN model was used to predict each extracted IMF element and the residual element, respectively. The prediction results of the IMFs

and residual parts obtained by different DBN models were added up to assemble an output as the final prediction result.

2. Materials and Methods

The flowchart of the EEMD-DBN forecasting model based on Lasso is demonstrated in Figure 1.

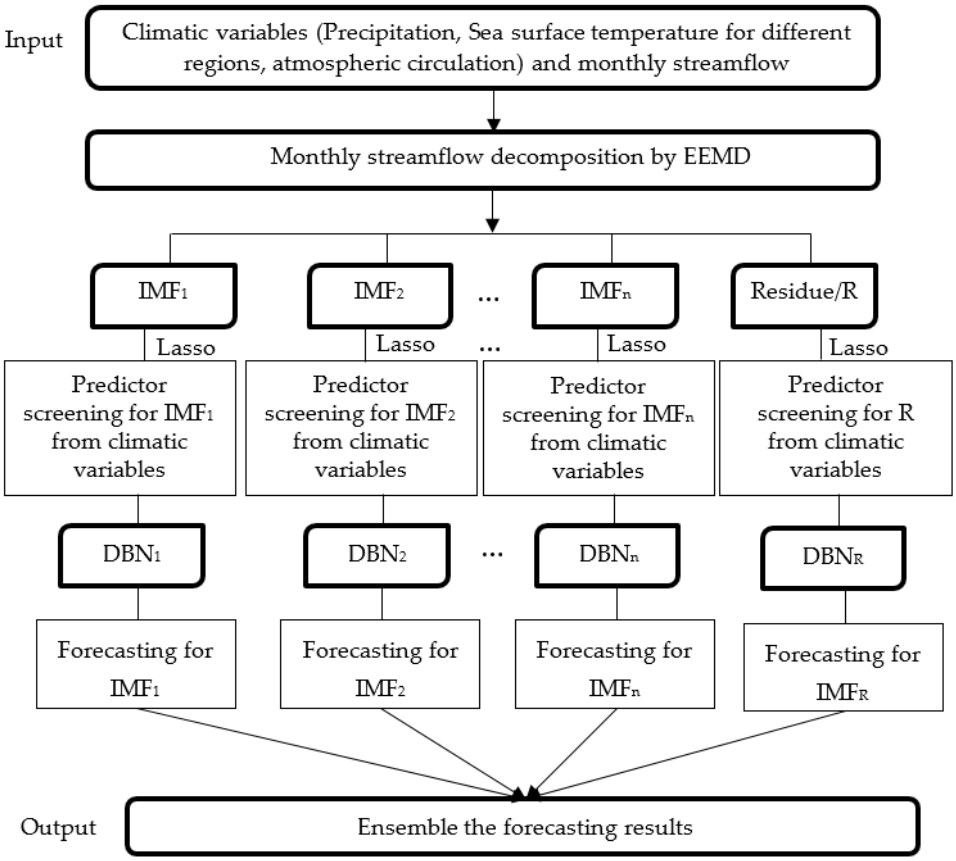

Figure 1. Flowchart of EEMD-Lasso-DBN for monthly streamflow forecasting.

From this figure, the presented EEMD-DBN based Lasso forecasting paradigm can be summarized as follows: Firstly, monthly streamflow data and candidate predictors, including precipitation, sea surface temperature for different regions, and atmospheric circulations, are collected. The EEMD method is applied to decompose the raw monthly streamflow time series, x(t), into m IMF elements, c_i(t), i = 1, 2 . . . , m, and one residual element, r(t). Secondly, the Lasso method is employed to select the predictors that significantly correlated with each extracted IMF element and the residual element of the monthly streamflow. Thirdly, the DBN models are developed to model the relationship between climate variables and IMFs and the residual element, respectively. Finally, the prediction results of the IMFs and residual parts obtained by different DBN models were added up to assemble an output as the final prediction result.

2.1. Ensemble Empirical Mode Decomposition (EEMD)

Ensemble empirical mode decomposition (EEMD) can be an effective tool to deal with non-linear and non-stationary time series, which is proposed in the improvement of the empirical mode decomposition (EMD) approach [23]. For the EMD approach, it is likely that there are similar oscillations in different modes or very different oscillations in a mode, so it cannot accurately reflect the characteristics of the raw data. To overcome this drawback, EEMD calculates the IMF elements by the mean of an ensemble of trials, while adding a white noise of a finite amplitude in each trial.

Two conditions should be satisfied when the EEMD approach is used: (1) The mean of the upper and lower envelopes must be equal to zero everywhere, and (2) the number of extreme data and the number of zero crossing must be equal or differ at most by one [24,25]. EEMD can decompose the time series into several IMFs, and an IMF means a simple oscillatory mode with a variable amplitude and frequency over time [26,27]. A detailed description of the process of extracting the IMF modes can be found in [28–30].

2.2. Least Absolute Shrinkage and Selection Operator (Lasso)

Least absolute shrinkage and selection operator, Lasso, is a popular statistical technique in conjunction with generalized linear models for predictor selection by shrinking some coefficients to zero. Lasso can constrain the regression coefficients by minimizing the residual sum of squares subject to the sum of the absolute values of the coefficient being less than a constant [31,32].

The coefficient estimates (β) are often calculated by the ordinary least squares method, which can be defined as:

$$L(\beta) = ||Y - X\beta|| \tag{1}$$

where $L(\beta)$ is the loss function.

To select the predictors, a penalty term has been added:

$$L(\beta) = ||Y - X\beta|| + \lambda\beta \tag{2}$$

where λ is a positive regularization parameter. The parameter, λ, controls the trade-off between the bias and parsimony of a fitted Lasso model; the higher the penalty parameter, λ, is, the higher the penalization on the predictors is, thus a sparser model will be screened with more coefficients shrank to zero [33].

When the Lasso model assumes N pairs of samples, then Equation (3) can be alternatively formulated as:

$$\beta = argmin\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2\right\} \tag{3}$$

$$\text{Subject to } \sum_{j=1}^{k}\beta_j \leq \lambda \tag{4}$$

Care should be taken to determine the values of λ, and it can effectively remove redundant parameters from the model, while accurately estimating the remaining significant parameters.

2.3. Deep Belief Networks (DBN)

Deep belief networks (DBN) consist of an unsupervised restricted Boltzmann machine (RBM) and a supervised network, as shown in Figure 2. Each RBM has a layer of input neurons and a single hidden layer with hidden-to-all-visible connections [34,35]. The DBN model can achieve a better performance with the initialization weight than that of the ANN model with random weights [36,37].

There is the pre-training part and fine-tuning part in the training process of the DBN model. Firstly, the parameters of the DBN model are initialized in the pre-training part, and the network is trained from layer to layer using the unsupervised learning method. When one layer of RBM is trained, the output of RBM can be transferred to the input layer of the next RBM. Secondly, the gradient descent method is used for the weights of the whole network. The fine-tuning process is similar to the training process of the back propagation algorithm [38].

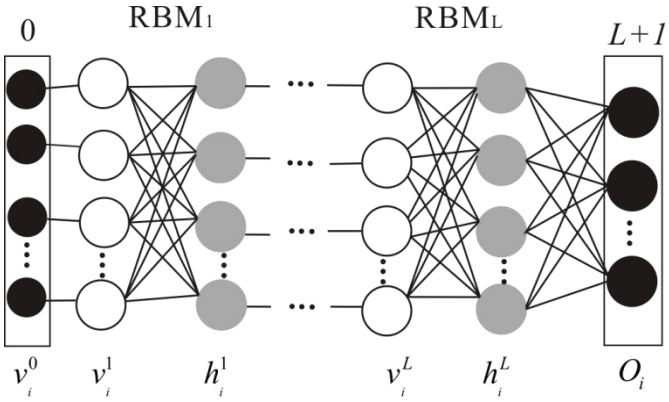

Figure 2. Structure of the Deep Belief Networks (DBN) model with three hidden layers.

There is one visible layer and one hidden layer in the RBM. There are no connections between neurons in the same layer, but full connections between neurons in different layers. The hidden and visible neurons are binary and stochastic, and v and h can be defined as the visible layer vector and the hidden layer vector, respectively. Then, the energy function of the RBM can be given by:

$$E(v,h|\theta) = -\sum_{i=1}^{m} a_i b_i - \sum_{j=1}^{n} b_j h_j - \sum_{i=1}^{m}\sum_{j=1}^{n} v_i w_{ij} h_j \quad (5)$$

where $\theta = (w_{ij}, a_i, b_j)$ is defined as the parameter of the RBM; w_{ij} represents the weights between visible neurons, i, and hidden neurons, j; a_i is the bias parameter of ith visible neurons; is the bias parameter of ith hidden neurons; and m and n are the numbers of visible and hidden neurons, respectively.

Hence, one of the important steps is to calculate the corresponding parameter,, and a detailed description of the process of calculating the parameters of the DBN model can be found in [39–41].

Different performance evaluation measures, including the root mean square error (RMSE), mean relative error (MAE), Nash-Sutcliffe coefficient (NS), and correlation coefficient (R^2) were employed to assess the results of the models.

3. Study Area and Data

3.1. Study Area

The Tennessee River was chosen as the study region (Figure 3), and it has an area of approximately 106,200 km^2. The river starts in eastern Tennessee and flows into the Ohio River. The basin is located in the temperate climate, with warm summers and mild winters. Annual mean precipitation is about 1400 mm, and it ranges from 1350 mm to 1450 mm from east to west. The greatest rainfall occurs in the winter and early spring, especially in March, while September and October are the driest months. The average annual temperature in the area is 13.9 °C, and it ranged from 11.1 °C to 14.4 °C across the area. The warmest months of the year are July and August, and the coldest months of the year are typically January and February. There is one gauging station 06010201 (35.48°, 87.82°) in the downstream of the Tennessee River. In the Tennessee River Basin, there are three main types of land use, with forests, pastureland, and cropland covering more than 48%, 20%, and 19% of the Tennessee River Basin, respectively. Developed land accounts for only about 8.6% of the area. The physiography of the Tennessee River is similarly diverse. There are four ecoregions, including the Blue Ridge, Ridge and Valley, South-west Appalachians, and Interior Plateau. The hydrological station is located in the Interior Plateau, which has well-developed karst terrain with thin soils and low-gradient streams with bedrock substrate covered by thin gravel deposits.

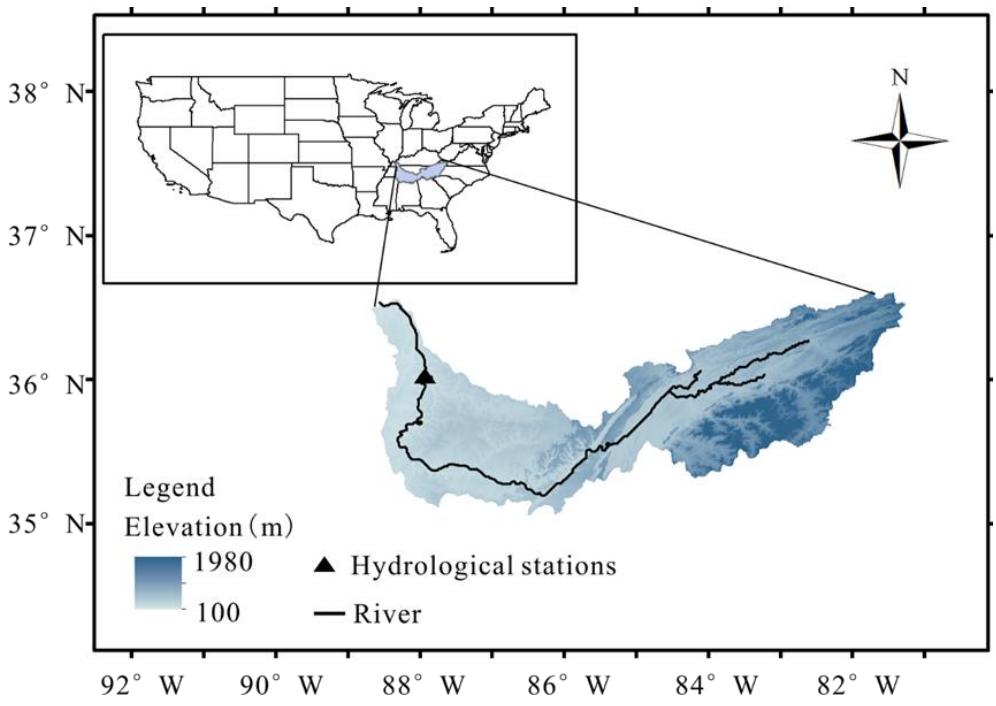

Figure 3. Location of the study area.

3.2. Data

The datasets used in this study are: (1) Natural streamflow: The unimpaired United States Geological Survey (USGS) stream discharge gauging stations in the Tennessee River watershed (Figure 3). For this study, monthly streamflow values for one gauging station were used from January 1950 through to December 2016; (2) precipitation: Monthly precipitation from 1950–2016 were obtained from the national climatic data center; and (3) large-scale climatic variability and sea surface temperature (SST): Sea surface temperature data from 1950–2016 were obtained from the Met Office Hadley Centre observation datasets. Large-scale climatic variability, including AO (Arctic Oscillation), PNA (Pacific-North American Pattern), PDO (Pacific Decadal Oscillation), NAO (North Atlantic Oscillation), SOI (Southern Oscillation Index), and AMO indices, can be obtained from National Oceanic and Atmospheric Administration (NOAA)(http://www.esrl.noaa.gov/psd/data/climateindi). The El Niño–Southern Oscillation (ENSO) index data were collected from the ERSSTv3b dataset. The model has been coded by MATLAB software. MATLAB function regress, newrb, svmtrain, and nnetfw is used to complete the analyses of MLR, RBFNN, SVR, and DBN models. The toolbox of MATLAB for Lasso and EEMD were also used.

4. Results and Discussion

In this study, the monthly streamflow time series of one unimpaired station from January 1950 through December 2016 were used (Figure 4). The inter-annual variability of the monthly streamflow time series of the station is large. For station 06010201, the minimum and maximum values in the entire flow data are in the range of 40–2517 m^3/s, and the average value and variance is 521 and 405 m^3/s, respectively. The Mann-Kendall test was used to study the trend characteristics of streamflow, and there is a statistics value of -0.414, which means that the streamflow in station 6010201 has a downward trend, but the trend is not significant.

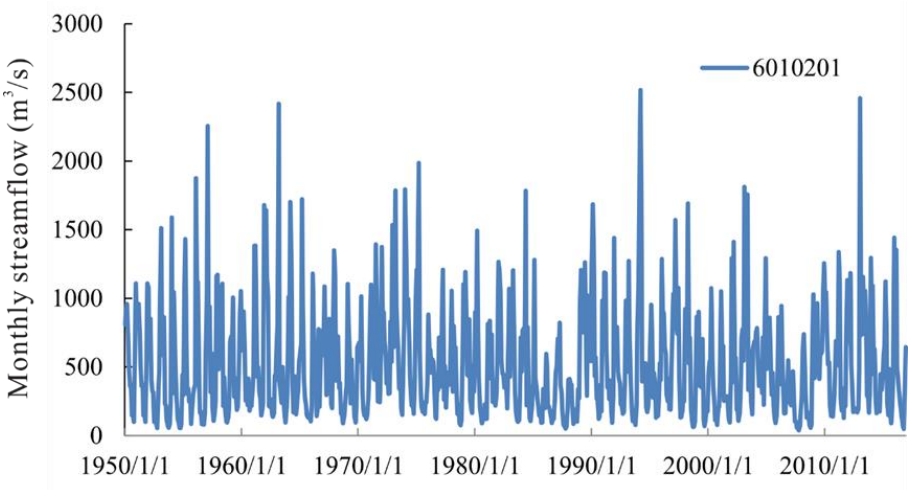

Figure 4. Monthly streamflow time series of the station in the study area.

4.1. Investigation of Main Elements Decomposition

The results of the EEMD are shown in Figure 5. It can be seen that the raw time series were resolved into eight IMFs and one residue, respectively. Each IMF has a different characteristic with a different variable amplitude and wave-number. The variable amplitude and wave-number represent the function of distance and their characteristic scales, so the IMFs are usually physically meaningful. IMF_1 has the characteristic of the maximum amplitude, highest frequency, and shortest wavelength. The other IMF elements have a decreasing trend in amplitude and frequency, and an increasing trend in wavelength. The residue part is a special element, which varies slowly around the long-term average.

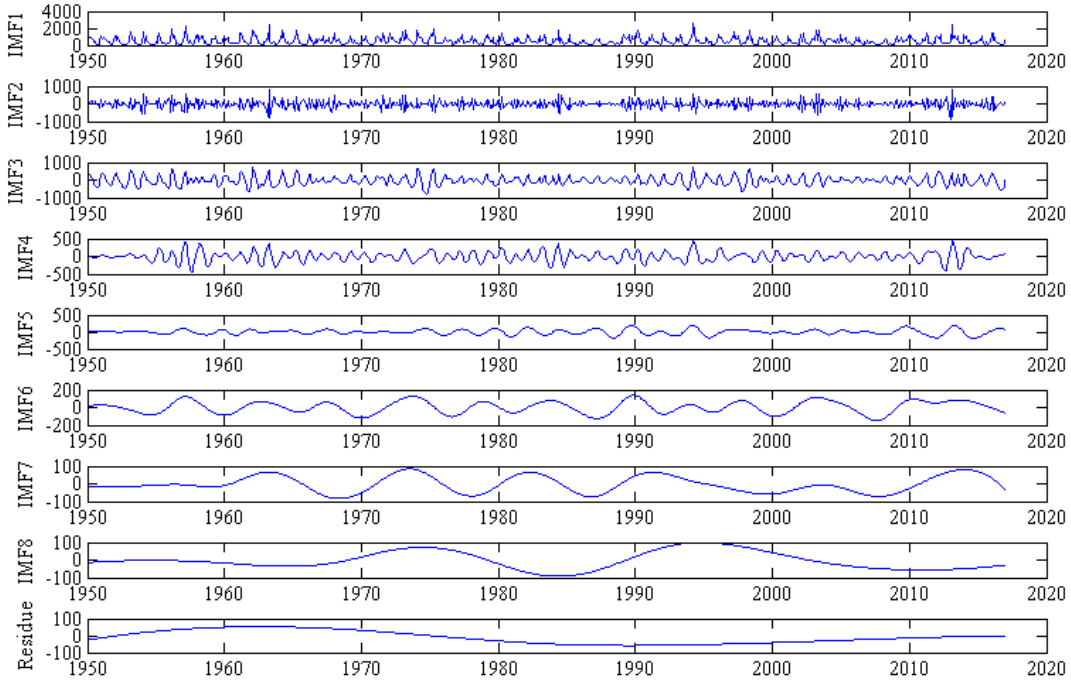

Figure 5. Resolved intrinsic mode functions (IMFs) of the streamflow data by Ensemble Empirical Mode Decomposition (EEMD).

The Lasso method was employed to screen the predictors that significantly correlated with each IMF and the residual element, respectively. The predictors of each IMF were different as shown in Table 1. Each IMF has different characteristics and physical meanings, corresponding to different

predictors. Most parts of IMF were affected by rainfall except for residue, and it indicated that rainfall was necessary for different oscillatory modes. IMF_1, IMF_4, and IMF_5 were only affected by rainfall, while IMF_2 and IMF_3 were both significantly related with the SST of different regions. The term, residue, correlated with large-scale atmospheric oscillations and SSTs except for rainfall. As the wavelength increased, the number of significantly correlated predictors also increased. The possible reason was that IMFs with long wavelengths have more complex physical meaning.

Table 1. The predictors of each IMF screened by Lasso.

Data	Predictors
IMF_1	rainfall
IMF_2	rainfall, SST3, SST4, SST6, SST7, SST10
IMF_3	rainfall, SST1, SST5, SST6, SST7, SST8, SST10
IMF_4	rainfall
IMF_5	rainfall
IMF_6	rainfall, AO, PNA, PDO, NAO, NINO3, SST3, SST6, SST7, SST10
IMF_7	rainfall, AO, PDO, NAO, SOI, NINO3, NINO3.4, SST1-4, SST6, SST10
IMF_8	rainfall, AO, PDO, NAO, SOI, NINO3, NINO4, NINO3.4, SST1-3, SST6-10
Residue	AO, PNA, PDO, SOI, NINO3, NINO4, NINO3.4, SST1-3, SST6-10

4.2. Investigation of Different Forecasting Methods

The multiple linear regression (MLR), radial basis function neural network (RBFNN), and support vector regression (SVR) models were also built and compared on their performance with the DBN model. Table 2 shows the forecasting performance evaluation of four models for station 06010201. During the calibration period, the order of the MAE and RMSE values were as follows: MLR (0.848–310) > RBFNN (0.741–278) > SVR (0.730–275) > DBN (0.637–251). During the validation period, the order of the MAE and RMSE values were as follows: MLR (1.093–308) > RBFNN (0.965–287) > SVR (0.952–286) > DBN (0.841–268). So, the order of the performance of the four models was: MLR < RBFNN < SVR < DBN. Compared with the three other models, the DBN model performs better in the calibration and validation period. The DBN models can discover the inherent features and hidden invariant structures in data from layer to layer. Therefore, the performance of the DBN models exhibits superiority and higher accuracy in monthly streamflow forecasting.

Table 2. Forecasting performance evaluation of four models for station 06010201.

	Model Calibration				Model Validation			
	MAE	**RMSE**	**NS**	**R^2**	**MAE**	**RMSE**	**NS**	**R^2**
MLR	0.848	310	0.420	0.629	1.09	308	0.400	0.594
RBFNN	0.741	278	0.534	0.742	0.965	287	0.479	0.627
SVR	0.730	275	0.542	0.745	0.952	286	0.483	0.649
DBN	0.637	251	0.618	0.790	0.841	268	0.546	0.723

4.3. Investigation of the Forecasting Models with EEMD

The MLR, RBFNN, SVR, and DBN models combined with EEMD were used to build the complex relationship between the impact factors at different scales and streamflow. The RBFNN model consists of one input layer, one hidden layer, and one output layer. There are two main parameters, including the hidden neurons and the spread of the radial basis function. A hidden layer with 120 neurons was used to develop the RBFNN model. The spread of the radial basis function was set at 0.9. The SVR model was developed based on the Vapnik-Chervonenkis dimension theory and the structure risk minimization principle, which has important parameters, including the regularization parameter, the width of the tube in loss function, and the spread. In this study, the regularization parameter (20), regression tube widths (0.001), and spread (0.8) were the optimal parameters.

The results are shown in Table 3 and Figure 6. The performances of the MLR, RBFNN, SVR, and DBN models during the calibration period were: As MAEs, 0.725, 0.699, 0.620, and 0.462; RMSEs, 274, 268, 247, and 202; NS, 0.546, 0.566, 0.632, and 0.754; and R^2, 0.747, 0.758, 0.799, and 0.873, respectively. During the model validation period, these models resulted in MAEs of 0.945, 0.913, 0.821, and 0.631; RMSEs of 285, 282, 264, and 221; NS of 0.485, 0.498, 0.560, and 0.691; and R^2 of 0.670, 0.710, 0.733, and 0.824, respectively. The results revealed that the forecasting models combined with EEMD can improve the accuracy of streamflow forecasting compared to the forecasting models without EEMD.

Table 3. Forecasting performance evaluation of four models with EEMD for station 06010201.

	Model Calibration				Model Validation			
	MAE	**RMSE**	**NS**	**R^2**	**MAE**	**RMSE**	**NS**	**R^2**
EEMD-MLR	0.725	274	0.546	0.747	0.945	285	0.485	0.670
EEMD-RBFNN	0.699	268	0.566	0.758	0.913	282	0.498	0.710
EEMD-SVR	0.620	247	0.632	0.799	0.821	264	0.560	0.733
EEMD-DBN	0.462	202	0.754	0.873	0.631	221	0.691	0.824

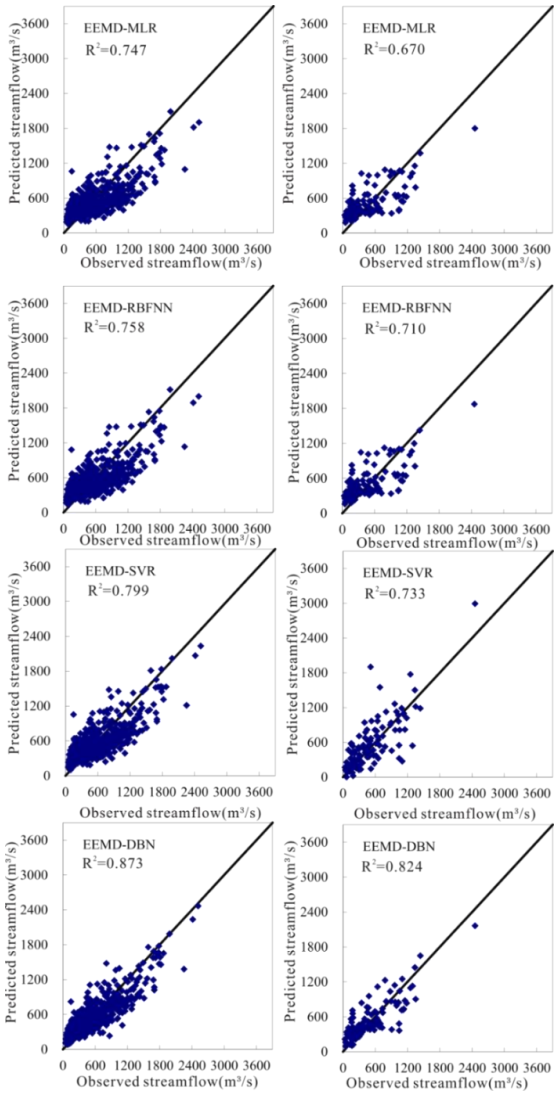

Figure 6. Predicted and observed streamflow of 06010201 station with EEMD during model calibration (**left** column) and validation (**right** column) using four different models.

4.4. Investigation of the Forecasting Models with EEMD and Lasso

Table 4 shows the forecasting performance evaluation of four models with and without EEMD and Lasso for station 06010201. The performances of the DBN, EEMD-DBN, and EEMD-Lasso-DBN models during the calibration period results in MAEs of 0.637, 0.462, and 0.245; RMSEs of 251, 202, and 158; NS of 0.618, 0.754, and 0.848; and R^2 of 0.790, 0.873, and 0.922, respectively. During the model validation period, these models resulted in MAEs of 0.841, 0.631, and 0.427; RMSEs of 268, 221, and 189; NS of 0.546, 0.691, and 0.775; and R^2 of 0.723, 0.824, and 0.885. It showed that the EEMD-Lasso-DBN model obtained a lower MAE and RMSE, and a higher NS and R^2 during both the model calibration and validation periods compared to DBN and EEMD-DBN. Tables 2–4 also indicated that the performances of the EEMD-Lasso-forecasting model (DBN/ANN/RBFNN/MLR) performed better than those of the EEMD-forecasting model, and each single forecasting model (DBN/ANN/RBFNN/MLR) performed worst. This analysis also indicated that EEMD and Lasso could assist in improving the performance of the forecasting models.

Table 4. Forecasting performance evaluation of four models with EEMD and Lasso for station 06010201.

	Model Calibration				**Model Validation**			
	MAE	**RMSE**	**NS**	**R^2**	**MAE**	**RMSE**	**NS**	**R^2**
EEMD-Lasso-MLR	0.674	261	0.586	0.770	0.884	277	0.515	0.712
EEMD-Lasso-RBFNN	0.544	225	0.694	0.838	0.728	243	0.627	0.766
EEMD-Lasso-SVR	0.413	187	0.788	0.892	0.580	210	0.722	0.822
EEMD-Lasso-DBN	0.245	158	0.848	0.922	0.427	189	0.775	0.885

Figures 7 and 8 illustrates the streamflow forecasts of 06010201 stations with and without EEMD and Lasso during model calibration and validation using four different models. The R^2 statistics of different models are given in Figures 7 and 8, and the EEMD-Lasso-DBN model has the highest R^2 in model calibration and the validation period with the value of 0.922 and 0.885, respectively, and the single MLR model has the smallest R^2 in model calibration and the validation period with a value of 0.629 and 0.594, respectively. The MLR model is often used to model linear relationships, but the relationship between the streamflow and climate variables has non-linear, non-stationary, and non-smoothness features, so the MLR model cannot be enough to deeply simulate the variability features in the relationship. From the Figures, it is obvious that the DBN models provided better performance than the other models in monthly streamflow forecasting, and EEMD-Lasso-DBN models seem to be more adequate than the single model for forecasting monthly streamflow.

In general, the complex hydrological time-series can be resolved into a simple sub-series by EEMD, and the characteristic can be seen more clearly at the daily, monthly, and annually periods than the raw data. From a large number of predictors, different predictors were selected for different scales by the Lasso method. Then, DBN models were developed for building the relationship between the selection predictors and different sub-series at different scales, respectively. The performance of the forecasting models has been improved compared to those of forecasting models that were built directly by raw data because the features of the decomposed elements can be extracted. Actually, monthly streamflow contains different frequency elements under the influence of multi-scale predictors. It is very difficult to clarify the internal mechanism of the phenomenon just by the raw time series when forecasting. Therefore, the EEMD-Lasso-DBN model performs better than the single model for forecasting monthly streamflow.

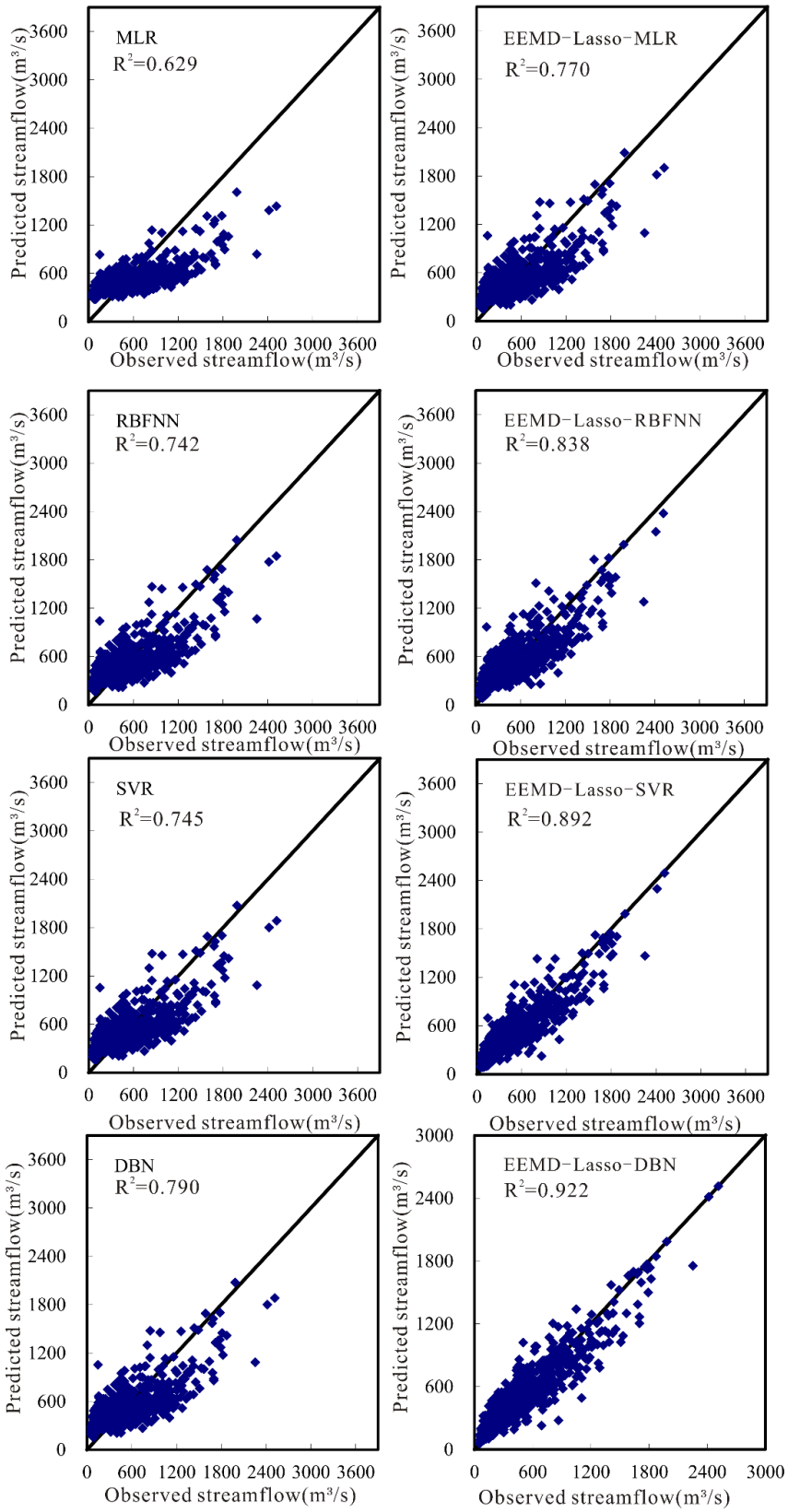

Figure 7. Predicted and observed streamflow of the 06010201 station with and without EEMD and Lasso during model calibration using four different models.

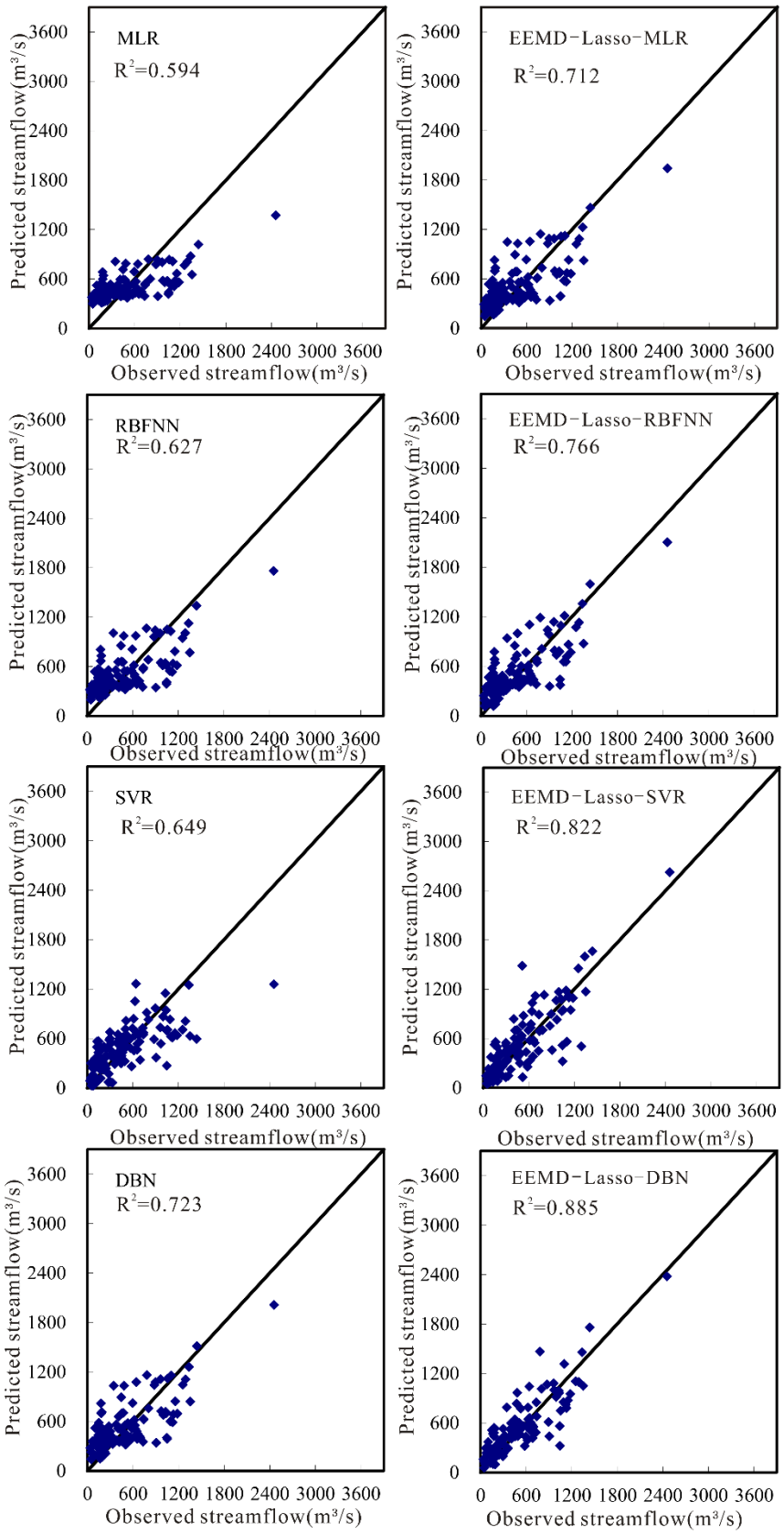

Figure 8. Predicted and observed streamflow of the 06010201 station with and without EEMD and Lasso during model validation using four different models.

5. Conclusions

This study aimed to develop a new approach based on EEMD, Lasso, and DBN for improving the accuracy of streamflow forecasting. In this paper, the EEMD method successfully resolved the raw monthly streamflow time series into eight IMF elements and one residual element. Each IMF and the residual may represent different possible physical meanings, which all have the respective characteristics with amplitude, frequency, and wavelength. Then, the Lasso method was employed to screen the predictors that significantly correlated with each IMF and the residual element, respectively. The predictors of each IMF were different; most parts of IMF were affected by rainfall expect for residue, and it indicated that rainfall was significant for different oscillatory modes. The predictor selection can detect the underlying relationship between historical data to give the highest accuracy.

Furthermore, DBN models were developed to simulate the relationship between each IMF and the corresponding predictors, respectively, so the tendencies of decomposed elements could be predicted. Finally, the prediction results of all the decomposed elements were assembled to produce an ensemble result for the raw monthly streamflow forecasting. Furthermore, four statistical measures (MAE, RMSE, NS, and R^2) were used to evaluate the forecasting performance of the different models. The comparison results suggested that the ELD method significantly produced more accurate forecasting than traditional forecasting models, and the single forecasting models without EEMD and Lasso perform worst among the developed models.

For future work, the new proposed approach can be used in the prediction of other time series. Furthermore, future improvement could be focused on predicting the decomposed elements by different approaches to get a higher accuracy.

Author Contributions: Formal analysis, J.W. and Q.J.; Methodology, H.C.; Validation, H.C.; Writing—original draft, H.C.; Writing—review & editing, J.W. and Q.J.

Funding: This research was funded by National key research and development project, grant number 2017YFC0403600; National Natural Science Foundation of China, grant number 51459003; the Chinese Ministry of Water Resources special funds for scientific research on public causes, grant number 201501028, the Science and Technology Projects State Grid Corporation of China, grant number 52283014000T.

Acknowledgments: Comments and suggestions from anonymous reviewers, the Associate Editor, and the Editor are greatly appreciated.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Carrier, C.; Kalra, A.; Ahmad, S. Using paleo reconstructions to improve streamflow forecast lead time in the western United States. *J. Am. Water Resour. Assoc.* **2013**, *49*, 1351–1366. [CrossRef]
2. Anghileri, D.; Voisin, N.; Castelletti, A.; Pianosi, F.; Nijssen, B.; Lettenmaier, D.P. Value of long-term streamflow forecasts to reservoir operations for water supply in snow-dominated river catchments. *Water Resour. Res.* **2016**, *52*, 4209–4225. [CrossRef]
3. Zhang, Y.; Lian, J.; Liu, F. An improved filtering method based on EEMD and wavelet-threshold for modal parameter identification of hydraulic structure. *Mech. Syst. Signal Process.* **2016**, *68*, 316–329. [CrossRef]
4. Wood, A.W.; Hopson, T.; Newman, A.; Brekke, L.; Arnold, J.; Clark, M. Quantifying streamflow forecast skill elasticity to initial condition and climate prediction skill. *J. Hydrometeorol.* **2016**, *17*, 651–668. [CrossRef]
5. Opitz-Stapleton, S.; Gangopadhyay, S.; Rajagopalan, B. Generating streamflow forecasts for the Yakima River Basin using large-scale climate predictors. *J. Hydrol.* **2007**, *341*, 131–143. [CrossRef]
6. Amigo, J.M.; Gredilla, A.; de Vallejuelo, S.F.O.; de Diego, A.; Madariaga, J.M. Study of parameters affecting the behaviour of trace elements in a polluted estuary. Canonical correlation analysis as a tool in environmental impact assessment. *Chemom. Intell. Lab.* **2012**, *119*, 1–10. [CrossRef]
7. Risko, S.L.; Martinez, C.J. Forecasts of seasonal streamflow in West-Central Florida using multiple climate predictors. *J. Hydrol.* **2014**, *519*, 1130–1140. [CrossRef]
8. McNeish, D.M. Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivar. Behav. Res.* **2015**, *50*, 471–484. [CrossRef] [PubMed]

9. Liu, Z.; Sun, W.; Zeng, J. A new short-term load forecasting method of power system based on EEMD and SS-PSO. *Neural Comput. Appl.* **2014**, *24*, 973–983. [CrossRef]
10. Kasiviswanathan, K.S.; He, J.; Sudheer, K.P.; Tay, J.H. Potential application of wavelet neural network ensemble to forecast streamflow for flood management. *J. Hydrol.* **2016**, *536*, 161–173. [CrossRef]
11. Wu, C.L.; Chau, K.W.; Li, Y.S. Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques. *Water Resour. Res.* **2009**, *45*, 8. [CrossRef]
12. Zhang, X.; Peng, Y.; Zhang, C.; Wang, B. Are hybrid models integrated with data preprocessing techniques suitable for monthly streamflow forecasting? Some experiment evidences. *J. Hydrol.* **2015**, *530*, 137–152. [CrossRef]
13. Mehr, A.D.; Kahya, E. A Pareto-optimal moving average multigene genetic programming model for daily streamflow prediction. *J. Hydrol.* **2017**, *549*, 603–615. [CrossRef]
14. Lee, H.S. Estimation of extreme sea levels along the Bangladesh coast due to storm surge and sea level rise using EEMD and EVA. *J. Geophys. Res. Ocean.* **2013**, *118*, 4273–4285. [CrossRef]
15. Fu, W.; Zhou, J.; Zhang, Y.; Zhu, W.; Xue, X.; Xu, Y. A state tendency measurement for a hydro-turbine generating unit based on aggregated EEMD and SVR. *Meas. Sci. Technol.* **2015**, *26*, 125008. [CrossRef]
16. Zhang, H.; Singh, V.P.; Wang, B.; Yu, Y. CEREF: A hybrid data-driven model for forecasting annual streamflow from a socio-hydrological system. *J. Hydrol.* **2016**, *540*, 246–256. [CrossRef]
17. Wang, W.; Van Gelder, P.H.; Vrijling, J.K.; Ma, J. Forecasting daily streamflow using hybrid ANN models. *J. Hydrol.* **2006**, *324*, 383–399. [CrossRef]
18. Tootle, G.A.; Singh, A.K.; Piechota, T.C.; Farnham, I. Long lead-time forecasting of US streamflow using partial least squares regression. *J. Hydrol. Eng.* **2007**, *12*, 442–451. [CrossRef]
19. Ausati, S.; Amanollahi, J. Assessing the accuracy of ANFIS, EEMD-GRNN, PCR, and MLR models in predicting PM 2.5. *Atmos. Environ.* **2016**, *142*, 465–474. [CrossRef]
20. Solomatine, D.P.; Xue, Y. M5 model trees and neural networks: Application to flood forecasting in the upper reach of the Huai River in China. *J. Hydrol. Eng.* **2004**, *9*, 491–501. [CrossRef]
21. Wang, W.C.; Chau, K.W.; Qiu, L.; Chen, Y.B. Improving forecasting accuracy of medium and long-term runoff using artificial neural network based on EEMD decomposition. *Environ. Res.* **2015**, *139*, 46–54. [CrossRef] [PubMed]
22. Shao, H.; Jiang, H.; Wang, F.; Wang, Y. Rolling bearing fault diagnosis using adaptive deep belief network with dual-tree complex wavelet packet. *ISA Trans.* **2017**, *69*, 187–201. [CrossRef] [PubMed]
23. Huang, N.E.; Wu, Z. A review on Hilbert-Huang transform: Method and its applications to geophysical studies. *Rev. Geophys.* **2008**, *46*, 2. [CrossRef]
24. Hawinkel, P.; Swinnen, E.; Lhermitte, S.; Verbist, B.; Van Orshoven, J.; Muys, B. A time series processing tool to extract climate-driven interannual vegetation dynamics using ensemble empirical mode decomposition (EEMD). *Remote Sens. Environ.* **2015**, *169*, 375–389. [CrossRef]
25. Ouyang, Q.; Lu, W.; Xin, X.; Zhang, Y.; Cheng, W.; Yu, T. Monthly rainfall forecasting using EEMD-SVR based on phase-space reconstruction. *Water Resour. Manag.* **2016**, *30*, 2311–2325. [CrossRef]
26. Barge, J.T.; Sharif, H.O. An ensemble empirical mode decomposition, self-organizing map, and linear genetic programming approach for forecasting river streamflow. *Water* **2016**, *8*, 2016247. [CrossRef]
27. Zang, H.; Fan, L.; Guo, M.; Wei, Z.; Sun, G.; Zhang, L. Short-Term Wind Power Interval Forecasting Based on an EEMD-RT-RVM Model. *Adv. Meteorol.* **2016**, *2016*, 1–10. [CrossRef]
28. Wang, W.C.; Xu, D.M.; Chau, K.W.; Chen, S. Improved annual rainfall-runoff forecasting using PSO–SVM model based on EEMD. *J. Hydroinform.* **2013**, *15*, 1377–1390. [CrossRef]
29. Peng, T.; Zhou, J.; Zhang, C.; Fu, W. Streamflow forecasting using empirical wavelet transform and artificial neural networks. *Water.* **2017**, *9*, 406. [CrossRef]
30. Guo, Z.; Zhao, W.; Lu, H.; Wang, J. Multi-step forecasting for wind speed using a modified EMD-based artificial neural network model. *Renew. Energy* **2012**, *37*, 241–249. [CrossRef]
31. Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429. [CrossRef]
32. Nardi, Y.; Rinaldo, A. Autoregressive process modeling via the lasso procedure. *J. Multivar. Anal.* **2011**, *102*, 528–549. [CrossRef]
33. Kwon, S.; Lee, S.; Na, O. Tuning parameter selection for the adaptive Lasso in the autoregressive model. *J. Korean Stat. Soc.* **2017**, *46*, 285–297. [CrossRef]

34. Chen, H.; Wang, J.; Tang, B.; Xiao, K.; Li, J. An integrated approach to planetary gearbox fault diagnosis using deep belief networks. *Meas. Sci. Technol.* **2016**, *28*, 025010. [CrossRef]
35. Zhang, R.; Shen, F.; Zhao, J. A model with fuzzy granulation and deep belief networks for exchange rate forecasting. In Proceedings of the 2014 International Joint Conference on Neural Networks (IJCNN), Beijing, China, 6–11 July 2014; pp. 366–373.
36. Chen, J.; Jin, Q.; Chao, J. Design of deep belief networks for short-term prediction of drought index using data in the Huaihe river basin. *Math. Probl. Eng.* **2012**, *2012*, 1–16. [CrossRef]
37. Dedinec, A.; Filiposka, S.; Dedinec, A.; Kocarev, L. Deep belief network based electricity load forecasting: An analysis of Macedonian case. *Energy* **2016**, *115*, 1688–1700. [CrossRef]
38. Chen, Z.; Li, W. Multisensor Feature Fusion for Bearing Fault Diagnosis Using Sparse Autoencoder and Deep Belief Network. *IEEE Trans. Instrum. Meas.* **2017**, *66*, 1693–1702. [CrossRef]
39. Zhong, P.; Gong, Z.; Li, S.; Schönlieb, C.B. Learning to Diversify Deep Belief Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote.* **2017**, *55*, 3516–3530. [CrossRef]
40. Agana, N.A.; Homaifar, A. EMD-Based Predictive Deep Belief Network for Time Series Prediction: An Application to Drought Forecasting. *Hydrology* **2018**, *5*, 18. [CrossRef]
41. Yang, H.; Hu, B.; Pan, X.; Yan, S.; Feng, Y.; Zhang, X.; Yin, L.; Hu, C. Deep belief network-based drug identification using near infrared spectroscopy. *J. Innov. Opt. Health Sci.* **2017**, *10*, 1630011. [CrossRef]