Article

Prediction of Air Pollution Concentration Based on mRMR and Echo State Network

Xinghan Xu ^{1,*} and Weijie Ren ^{2,*} ¹ Department of Environmental Engineering, Kyoto University, Kyoto 615-8540, Japan² Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China

* Correspondence: xu.xinghan.57u@st.kyoto-u.ac.jp (X.X.); renweijie@mail.dlut.edu.cn (W.R.); Tel.: +86-0411-84706002-2607 (W.R.)

Received: 15 March 2019; Accepted: 28 April 2019; Published: 1 May 2019

Abstract: Air pollution has become a global environmental problem, because it has a great adverse impact on human health and the climate. One way to explore this problem is to monitor and predict air quality index in an economical way. Accurate monitoring and prediction of air quality index (AQI), e.g., PM_{2.5} concentration, is a challenging task. In order to accurately predict the PM_{2.5} time series, we propose a supplementary leaky integrator echo state network (SLI-ESN) in this paper. It adds the historical state term of the historical moment to the calculation of leaky integrator reservoir, which improves the influence of historical evolution state on the current state. Considering the redundancy and correlation between multivariable time series, minimum redundancy maximum relevance (mRMR) feature selection method is introduced to reduce redundant and irrelevant information, and increase computation speed. A variety of evaluation indicators are used to assess the overall performance of the proposed method. The effectiveness of the proposed model is verified by the experiment of Beijing PM_{2.5} time series prediction. The comparison of learning time also shows the efficiency of the algorithm.

Keywords: air pollution concentration; prediction; feature selection; echo state network

1. Introduction

With the rapid advancement of urbanization and industrialization, air quality has deteriorated severely, which has negatively affected the quality of the living environment and even hindered economic growth in some areas [1]. In particular, the inhalable particles produced by industrial pollution have small particle size, large diffusion area and strong activity, thus they can enter the human body through the respiratory tract, which has adverse effects on human health. Therefore, the prediction of air pollutants plays a crucial role in the early warning and control of environmental pollution [2]. Therefore, modeling and forecasting the air quality index (e.g., PM_{2.5} concentration) has become an effective way to prevent and control air pollution, and it also provides a scientific basis for the development of effective measures [3]. The implementation of this idea can effectively reduce the health hazard of air pollution, thus achieving early warning and rational planning [4].

For a long time, many scholars have conducted in-depth research on air pollution. At the same time, a variety of predictive models have been proposed, such as autoregressive integrated moving average model [5], support vector machine [5], multiple linear regression model [6], neural networks [7,8], and so on [9]. All of them have been applied to predict air pollution concentration. Oprea et al. [10] applied artificial neural network and adaptive neuro-fuzzy inference system to predict PM_{2.5} concentration. Deng et al. [11] proposed heterogeneous space-time artificial neural networks to deal with spatial heterogeneity, which has been applied to predict the concentration of

fine particles in Beijing–Tianjin–Hebei. Ong et al. [12] proposed a deep recurrent neural network for time series prediction, which can accurately predict the $PM_{2.5}$ concentration. In order to solve the air pollution monitoring and prevention in Kunming and Yuxi, China, Li et al. [13] studied the method called co-integration flower-pollination-algorithm support vector machine for the prediction of PM_{10} concentration time series. In order to study the attenuation effect of haze on solar radiation scattering, Yao et al. [14] proposed a new SVM-based method that increases the accuracy of global solar radiation models by increasing variables of daily global solar radiation, sunshine hours, temperature, relative humidity, and air quality index. Reid et al. [15] selected an optimal prediction model based on 10-fold cross-validation to estimate $PM_{2.5}$ concentrations during wildfires in Northern California in 2008, and reliably predicted major wildfire events. As can be seen, different types of predictive models can successfully solve different problems of air pollution time series prediction.

Because the causes of air pollution are very complicated, analyzing the main pollutants and influence variables of air quality index will lay the foundation for the establishment of predictive models. Although the methods described earlier are effective, they do not analyze the validity of the input variables, so the models are likely to contain irrelevant or redundant information. In order to build an accurate predictive model, Sun et al. [16] applied Pearson correlation coefficient to analyze the relationship between $PM_{2.5}$ and other variables and selected the appropriate input variables according to the correlation order. Zhang et al. [17] investigated cross-correlations between $PM_{2.5}$ and four meteorological factors based on multifractal detrended cross-correlation analysis method, which reveals the impact of meteorological variables on $PM_{2.5}$ concentration. Zhu et al. [18] proposed a novel graphical causality analysis approach and analyzed the impact of meteorological and traffic variables on air quality indexes. Chen et al. [19] applied the convergent cross-mapping method to analyze the causal relationship between meteorological factors and $PM_{2.5}$ concentration in Beijing–Tianjin–Hebei region, and obtained quantitative causality analysis results. Therefore, for complex air pollution problems, we also need to choose the appropriate correlation or causality analysis method to analyze the impact variables, thus achieving high-precision detection and forecasting.

At present, air pollution has become one of the major environmental problems. However, the causes of air pollution are very complicated. Taking the $PM_{2.5}$ concentration as an example, its concentration is not only affected by the air pollutants such as NO_2 , CO, O_3 , and SO_2 , but also by meteorological variables such as temperature, pressure, humidity, wind speed and wind direction. How to choose effective information from a variety of variables for prediction is an important research topic. However, in the previous studies, the correlation between influencing factors and $PM_{2.5}$ was not considered in the establishment of most predictive models. In order to solve the above problem, this paper considers using the minimum redundancy maximum relevance (mRMR) [20] feature selection method to select the appropriate input variables, which can select the most relevant information and reduce redundant information. Furthermore, considering the chaotic characteristics of variables, we use phase space reconstruction to extract evolutionary information of relevant variables. At last, the new input variables are transferred to the supplemental leaky integrator echo state network (SLI-ESN) for prediction. The improved model not only enhances the feature extraction and memory ability of the reservoir for multivariate time series, but also improves the influence of historical evolution state on the current state. In practical applications, accurate prediction is conducive to monitoring air quality and making reasonable and scientific decision-making on air pollution prevention. In order to verify the validity of the proposed method, we select the dataset of air quality index and meteorological time series to predict $PM_{2.5}$ concentration in Beijing, China.

The rest of this paper are arranged as follows: Section 2 describes the preliminary knowledge of this paper, including feature selection method and echo state network; Section 3 introduces in detail the four aspects of the feature selection method, phase space reconstruction, $PM_{2.5}$ prediction model and algorithm flow; In Section 4, we analyzed the experimental results of the time series prediction of $PM_{2.5}$ concentration in Beijing, China. Section 5 gives conclusions and illustrates the challenges that will still be faced in the future.

2. Preliminaries

2.1. Feature Selection Method

Time series that affect air pollution are high-dimensional data, which not only contains rich information, but also has irrelevant or redundant factors. These irrelevant and redundant factors reduce the prediction accuracy and efficiency of the model. Thus, analyzing the relationship between variables and selecting valuable input variables are important for prediction.

Feature selection is the most typical data preprocessing method [21]. It consists of four parts: generation process, evaluation function, stop criterion and verification process. Currently, Common feature selection algorithms include random forest (RF), correlation feature selection (CFS), fast correlation-based filter (FCBF), mutual information (MI) [22], information gain (IG) [23], regularization models, relief-based algorithms, and genetic algorithm. However, these feature selection algorithms usually ignore the redundancy relationship between features. Random forests have over-fitting on noise regression problems and do not give continuous output. The CFS and FCBF feature selection methods are slow to calculate and cannot handle large-scale data efficiently. MI and IG do not require the type of data distribution but have high computational complexity for high-dimensional data. The regularization model has a better selection effect on high-dimensional data with a size much larger than the number of samples, but it does not perform satisfactorily in low-dimensional data selection. For the prediction of PM_{2.5} concentration with high noise, how to select appropriate feature selection method is very important. In this paper, the minimum redundancy maximum relevance (mRMR) [20] is used for feature selection. After this, a set of purest features can be obtained, and the redundancy features are removed while guaranteeing the maximum relevance.

2.2. Echo State Network

The echo state network (ESN) proposed by Jaeger et al. [24] is a new type of recurrent neural network. It consists of an input layer, a reservoir and an output layer. The structure of ESN is shown in Figure 1. W_{in} is the input weight matrix that connects the inputs and the reservoir, W is the internal connection weight matrix of the reservoir, W_{out} is the output weight matrix that connects the reservoir and the output, and W_{back} is the output-to-reservoir feedback connection weight matrix, which is usually set as a zero vector.

Figure 1. The basic structure of echo state network.

Assuming that the system obtains a time series $\mathbf{U}(n) = [u_1(n), u_2(n), \dots, u_K(n)]^T$ at time n , which is the input of the reservoir at this time. The output of the reservoir will be $\mathbf{y}_{in}(n) = \mathbf{U}(n + \lambda)$, and the prediction of different time steps is realized according to the adjustment of λ . The reservoir receives two inputs. The one is $\mathbf{U}(n)$ from the input layer and the other is $x(n - 1)$ from the previous state of the reservoir. The matrixes W and W_{in} are randomly generated and remain unchanged during the

training. The weight W is a large-scale sparse matrix, in which non-zero elements indicate activated neurons in the reservoir. The update formula of current state of reservoir is as follows:

$$\mathbf{x}(n+1) = \tanh(\mathbf{W}_{in} \times [1; \mathbf{U}(n+1)] + \mathbf{W} \times \mathbf{x}(n)) \quad (1)$$

where $\mathbf{x}(n) \in \mathbf{R}^{L \times 1}$ is the state of the reservoir at time n , and its initial state is a zero vector. The operator $\tanh(\cdot)$ is the hyperbolic tangent activation function. With the infinite growth of time, the dependence of the current state affected by the initial state of reservoir gradually decreases or even disappears [25]. The information in reservoir is transferred to the output layer with a linear connection, and the output of the network is:

$$\mathbf{y}(n) = \mathbf{W}_{out} \mathbf{x}(n) \quad (2)$$

Afterwards, Jaeger et al. [26] improved the ESN and proposed the leaky integrator ESN (LI-ESN) by introducing leaky integrator neurons into the reservoir. LI-ESN is a variant of ESN, but its reservoir contains the leak integrator neurons. The update formula of the state of reservoir in LI-ESN is:

$$\mathbf{x}(n+1) = a \times \tanh(\mathbf{W}_{in} \times [1; \mathbf{U}(n+1)] + \mathbf{W} \times \mathbf{x}(n)) + (1-a) \times \mathbf{x}(n) \quad (3)$$

where $a \in [0, 1]$ is the leaking rate. The leaky neuron in reservoir has a leaky integration of its activation, partially remembers its previous activation, and retains the effect of the previous moment on the current state.

3. PM_{2.5} Time Series Prediction Model

In order to accurately predict the concentration of PM_{2.5}, this paper proposes a PM_{2.5} time series prediction model. To handle high-dimensional input variables, the mRMR method is utilized to select optimal subset. Then, the phase space is reconstructed from the subset to obtain evolution information of time series. Finally, we predict PM_{2.5} time series based on supplemental leaky integrator echo state network (SLI-ESN) model, which is proposed in this paper.

3.1. mRMR Feature Selection Method

In this paper, mRMR method [20] is used for feature selection. Its basic idea is to maximize the correlation between input features and output, and to minimize the correlation between input features. The goal is to find the most representative feature subset. The evaluation of correlation between features is performed by mutual information method, and the appropriate feature subset is selected from original feature set. The maximum relevance and minimum redundancy of mRMR are calculated as follows:

$$\max D(S, C) = \frac{1}{|S|} \sum_{f_i \in S} I(f_i; C) \quad (4)$$

$$\min R(S) = \frac{1}{|S|^2} \sum_{f_i, f_k \in S} I(f_i; f_k) \quad (5)$$

where f_i and f_k represent the i th and k th feature in set S , respectively. C denotes the target output. $I(f_i; C)$ denotes the correlation function of the i th feature and output C . $I(f_i; f_k)$ represents the correlation function of the i th and k th feature. Mutual information is defined as follows:

$$I(x; y) = \iint P(x, y) \log \frac{P(x, y)}{P(x)P(y)} dx dy \quad (6)$$

Therefore, the expression of mRMR is as follows:

$$\max_S \left[\frac{1}{|S|} \sum_{f_i \in S} I(f_i; C) - \frac{1}{|S|^2} \sum_{f_i, f_k \in S} I(f_i; f_k) \right] \quad (7)$$

Applying the forward selection algorithm to solve the objective function (7), we can get the sort result of the input features. Finally, the optimal input feature subset can be obtained by cross-validation.

3.2. Phase Space Reconstruction

Phase space reconstruction is a very important step in time series analysis and prediction. In order to extract useful information from time series, Takens [27] proposed embedding theorem of phase space reconstruction. In the actual calculation, since the numerical differentiation is sensitive to the error. Therefore, phase space reconstruction of time series generally adopts coordinate delay method. The essence is to construct m -dimensional vector by different delay times of one-dimensional time series. Its calculation formula is as follows:

$$\mathbf{U}_i(n) = [u_i(n), u_i(n + \tau_i), \dots, u_i(n + (m_i - 1)\tau_i)] \quad (8)$$

where $\{u_i(n) | i = 1, 2, \dots, K\}$ are input time series, m_i is the embedding dimension, and τ_i is the delay time. The reconstructed phase space contains all the evolution information of original system.

The embedding theorem shows that for one-dimensional time series of infinitely long, noise-free m -dimensional chaotic attractors, an m -dimensional phase space can be found, as long as the embedding dimension satisfies $m \geq 2d + 1$, where d is dimension of the dynamic system [28]. However, the existing time series are finite-length sequences with noise. The embedding dimension and delay time cannot be arbitrarily selected, otherwise the phase space quality of the reconstruction will be seriously affected.

3.3. Supplementary Leaky Integrator Echo State Network

The echo state network, including the model proposed in this paper, has the ability to memorize historical information. However, due to the unpredictability of chaotic time series, it is difficult to accurately retain the information before the current state of the reservoir for a long time. Moreover, due to the complexity of PM_{2.5} time series, its current concentration information may be more dependent on the most recent state. Inspired by the leaked integrator reservoir, the update of the reservoir state is strongly dependent on the historical state. In this paper, considering the impact of previous state on current state, the state update formula of reservoir neurons is improved, and SLI-ESN model is proposed. The state update formula of reservoir is as follows:

$$\mathbf{x}(n+1) = a \times \tanh(\mathbf{W}_{in} \times [1; \mathbf{U}(n+1)] + \mathbf{W} \times \mathbf{x}(n)) + b \times \mathbf{x}(n) + (1 - a - b) \times \mathbf{x}(n-1) \quad (9)$$

As with LI-ESN, the meaning of the attenuation parameter a has not changed, and its satisfy $0 < a < 1$. The parameter b is the supplement factor, and it also need to satisfy $0 < b < 1$. More importantly, like the side length of a triangle, each coefficient must be greater than 0, and satisfy $a + b < 1$. Because the current state of reservoir is more dependent on the state of neighboring moment, this paper considers influence of the first two historical moments. When $n \geq 2$, supplement $(1 - a - b) \times \mathbf{x}(n-1)$ is established. Set $\mathbf{W}_{back} = 0$, the output layer is calculated as follows:

$$\mathbf{y}(n) = \mathbf{W}_{out} \mathbf{x}(n) \quad (10)$$

The calculation from reservoir to output layer is solved by ridge regression [29]. It is a biased estimation regression method for collinear data analysis. It overcomes ill-posed problem in least squares solution and prevents over-fitting problem. The expression is as follows:

$$W_{out} = (X^T X + kI)^{-1} X^T Y \quad (11)$$

where k is regularization parameter, I is the identity matrix, $x(n)$ and $y(n)$ are column vectors of X and Y , respectively.

3.4. Algorithm Flow

In order to accurately predict the PM_{2.5} time series, this paper proposes the prediction structure based on feature extraction and improved echo state network model. The process is summarized as follows.

1. Feature selection: select the optimal subset from original dataset based on mRMR feature selection method.
2. Phase space reconstruction: reconstruct phase space of the selected optimal subset based on Takens' theorem and form a new set of input features.
3. Data division: divide training set and testing set according to a certain proportion.
4. Model training: train SLI-ESN model using ridge regression algorithm on training set.
5. Prediction: predict PM_{2.5} time series using SLI-ESN model on testing set.

In order to briefly explain prediction process of the proposed model, the schematic diagram of the method is shown in Figure 2.

Figure 2. Schematic diagram of time series prediction based on supplementary leaky integrator echo state network (SLI-ESN). mRMR: minimum redundancy maximum relevance.

4. Results and Discussion

In order to prove the validity and practicability of the proposed method, the SLI-ESN model was applied to predict the actual observed PM_{2.5} time series of Beijing. At the same time, this paper also carried out comparative experiments of ESN, LI-ESN [26], extreme learning machine (ELM) [30], hierarchical ELM (H-ELM) [31] and stacked auto-encoder (SAE) [32]. The experimental environment is Windows 7 system, MATLAB 2016a programming software. The computer memory is 6 GB, clocked at 3.50 GHz, and Intel-i3 CPU.

Firstly, the AQI dataset and evaluation indicators are explained. Secondly, the feature selection experiment based on mRMR is introduced in detail, and the optimal subset is obtained. Finally, the

reconstructed features of optimal subset is used for training the SLI-ESN model, and the validity of the proposed method in PM_{2.5} time series prediction is verified by predictive indicators.

4.1. Data Description

This paper selects 8759 samples of hourly air pollution data from January to December 2016 in Haidian District, Beijing. The dataset is from the US Embassy (Harvard University Geographic Analysis Center Dataverse), including the average concentration of PM_{2.5}, PM₁₀, NO₂, CO, O₃, and SO₂ per hour, and hourly temperature (T), pressure (P), humidity (H), wind speed (WS) and wind direction (WD).

Because PM_{2.5} is a particulate matter that can enter the lungs, it has a great impact on human health and the quality of atmospheric environment. Therefore, predicting PM_{2.5} concentration can not only make an estimate of environmental quality, but also have a greater impact on better environmental governance and human health. This paper uses five predictive indicators to evaluate the prediction results of PM_{2.5} time series, namely root mean square error (RMSE), normalized root mean square error (NRMSE), mean absolute error (MAE), symmetric mean absolute percentage error (SMAPE) and Pearson correlation coefficient (R).

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{t=1}^N [\hat{y}(t) - y(t)]^2} \quad (12)$$

$$\text{NRMSE} = \frac{1}{\hat{y}_{\max} - \hat{y}_{\min}} \sqrt{\frac{1}{N} \sum_{t=1}^N [\hat{y}(t) - y(t)]^2} \quad (13)$$

$$\text{MAE} = \frac{1}{N} \sum_{t=1}^N |\hat{y}(t) - y(t)| \quad (14)$$

$$\text{SMAPE} = \frac{1}{N} \sum_{t=1}^N \frac{|y(t) - \hat{y}(t)|}{(|\hat{y}(t)| + |y(t)|)/2} \quad (15)$$

$$R = \frac{\sum_{t=1}^N (y(t) - \bar{y})(\hat{y}(t) - \bar{\hat{y}})}{\sqrt{\sum_{t=1}^N (y(t) - \bar{y})^2 \sum_{t=1}^N (\hat{y}(t) - \bar{\hat{y}})^2}} \quad (16)$$

where N is the number of samples, $\hat{y}(t)$ is the predicted output, and $y(t)$ is the target value.

In the above evaluation indicators, the smaller the values of the estimated indicators of RMSE, NRMSE, MAE, and SMAPE are, the better the prediction results of the model are. $R = 1$ indicates that $\hat{y}(t)$ and $y(t)$ are linear correlation, and $R = 0$ indicates there is no correlation. When $R \in (0, 1)$, it indicates that there is a correlation, and the larger value is, the stronger the linear correlation is.

4.2. Data Processing

The optimal subset selection of original data was performed using mRMR. In this paper, the first 75% of the dataset is used for training set, and the last 25% is used for testing set. The optimal subset is selected based on training set, and each model parameter in the simulation is obtained only from the training set.

Firstly, PM_{2.5} time series is chosen as reference variable. Other 10-dimensional variables are used as comparison variables. The data was quantitatively analyzed using mRMR to obtain sorting results: PM_{2.5}, CO, WS, PM₁₀, H, WD, SO₂, NO₂, P, T, and O₃. According to the influence of different factors on PM_{2.5} concentration, the correlation between different variables and PM_{2.5} is shown in Figure 3.

Figure 3. Correlation between different variables and PM_{2.5}.

According to the mRMR method, the irrelevant and redundant variables in the original dataset are reduced. The selection result of obtained optimal subset is shown in Figure 4.

Figure 4. Results of optimal subset selection.

According to the prediction results in Figure 4, when the predictor is with 5 dimensions, the prediction error is the smallest, namely 9.199. Therefore, the optimal subset is PM_{2.5}, CO, WS, PM₁₀, and H. The phase space reconstruction is then performed on the selected optimal subset.

The delay time τ and embedding dimension m calculated by C-C method [33] are shown in Table 1. The 5 variables in the optimal subset are expressed as bold in the table. As shown in Table 1, the delay time for the obtained optimal subset is [8,8,6,4,4], and the embedding dimension is [2,2,2,4,4] for PM_{2.5}, PM₁₀, CO, H and WS respectively.

Table 1. Phase space reconstruction parameters of variables. Abbreviations: hourly temperature (T), pressure (P), humidity (H), wind speed (WS), and wind direction (WD).

Variables	PM _{2.5}	PM ₁₀	NO ₂	CO	O ₃	SO ₂	T	P	H	WS	WD
τ	8	8	4	6	4	6	4	12	4	4	6
m	2	2	3	2	4	2	6	2	4	4	3

4.3. Experimental Results and Analysis

Phase space reconstruction is performed on the optimal subset to obtain a 14-dimensional reconstructed time series, which are used for inputs of prediction model. For the LI-ESN model, the reasonable range of the leaking rate a is (0, 1). For the SLI-ESN model, we use the cross-validation

to select the two parameters a and b . According to experience, the feasible and effective range of the supplementary factor b is mainly within $(0, 0.1)$. In this paper, ESN, LI-ESN, ELM, H-ELM and SAE are selected as comparison methods. The specific one-step (1 hour) prediction results are shown in Table 2.

Table 2. Comparison of one-step (1 hour) prediction results. Abbreviations: root mean square error (RMSE), normalized root mean square error (NRMSE), mean absolute error (MAE), symmetric mean absolute percentage error (SMAPE), Pearson correlation coefficient (R), echo state network (ESN), leaky integrator ESN (LI-ESN), extreme learning machine (ELM), hierarchical ELM (H-ELM), stacked auto-encoder (SAE), supplementary leaky integrator echo state network (SLI-ESN).

Methods	RMSE	NRMSE	MAE	SMAPE	R
ESN	10.2020	0.0160	6.6948	0.1053	0.9936
LI-ESN	9.7063	0.0152	6.2322	0.1001	0.9943
ELM	11.7330	0.0184	7.1902	0.1082	0.9914
H-ELM	14.1520	0.0222	8.0575	0.1102	0.9876
SAE	32.1700	0.0505	20.1840	0.2764	0.9448
SLI-ESN	9.3953	0.0147	5.8447	0.0894	0.9945

It can be seen from Table 2 that the proposed method achieves better prediction results in one-step (1 hour) prediction. The one-step (1 hour) prediction result of $PM_{2.5}$ concentration is shown in Figure 5. And Figure 6 plots the fit of predicted value to actual data. It can be seen from the figures that the prediction has a good linear relationship with actual value. SLI-ESN performs satisfactorily at peaks and undulating moments, which mainly depends on the full utilization of the historical state of the reservoir and the effective information obtained by phase space reconstruction.

Figure 5. $PM_{2.5}$ concentration one-step (1 hour) prediction curve.

Figure 6. PM_{2.5} concentration one-step (1 hour) predicted output fit curve.

At the same time, the simulation results of the five-step (5 hours) prediction are given in Figure 7. The prediction curve can better track original input, and the medium-term prediction effect is also good. Table 3 gives the five-step (5 hours) prediction results. The ten-step (10 hours) prediction result of SLI-ESN is shown in Figure 8. As seen in Figure 8, in some of the peaks, the prediction curve is still able to roughly fit the fluctuation trend of the original data, which is precisely because SLI-ESN makes full use of the role of historical information. Its comparison with other algorithms is shown in Table 4. The proposed algorithm achieves the optimal value on all four parameter indicators, except for SMAPE, which fully demonstrates the effectiveness of SLI-ESN in long-term prediction.

Figure 7. PM_{2.5} concentration five-step (5 hours) prediction curve.

Table 3. Comparison of five-step (5 hours) prediction results.

Methods	RMSE	NRMSE	MAE	SMAPE	R
ESN	43.1243	0.0677	29.3879	0.3777	0.8907
LI-ESN	41.1574	0.0646	27.6611	0.3577	0.9027
ELM	45.7617	0.0718	29.7892	0.3726	0.8678
H-ELM	49.9159	0.0783	32.1977	0.3974	0.8368
SAE	51.9847	0.0816	34.5487	0.4039	0.8394
SLI-ESN	37.6874	0.0591	25.5871	0.3392	0.9108

Figure 8. PM_{2.5} concentration ten-step (10 hours) prediction curve.**Table 4.** Comparison of ten-step (10 hours) prediction results.

Methods	RMSE	NRMSE	MAE	SMAPE	R
ESN	72.3723	0.2299	50.5928	0.5620	0.6858
LI-ESN	70.5700	0.2944	49.3628	0.5569	0.6895
ELM	71.6619	0.2429	47.1085	0.5385	0.6582
H-ELM	66.7932	0.2760	46.1529	0.5003	0.7053
SAE	71.7758	0.3631	49.7844	0.5617	0.6623
SLI-ESN	65.7108	0.1966	46.3633	0.5443	0.7314

Longitudinal comparison between Tables 2–4 shows that the longer the prediction time is, the larger the error will be, which is consistent with the basic characteristics of the chaotic time series. Comparing Tables 2–4, it can be found that the recurrent neural networks in three kinds of networks (SAE as the representative of deep learning, ELM and H-ELM as the representative of feedforward neural network and ESN as the representative of recurrent neural network) have more satisfactory prediction performance. This exactly demonstrates the validity of the structure of the reservoir in time series prediction.

In order to further illustrate the performance of the proposed method, the running time of all comparison methods are shown in Table 5. The results show that the training time based on the ESN

and ELM model is much smaller than the deep learning model. This is mainly because the training process of deep learning consumes a lot of time, and this time-consuming process is not required by the other neural networks. Moreover, SLI-ESN can complete training and testing within an acceptable time frame. It indicates that SLI-ESN can achieve good prediction results in both prediction accuracy and time consuming.

Table 5. Comparison of running time on six methods (second).

Methods	One-step (1 hour)		Five-step (5 hours)		Ten-step (10 hours)	
	Training Time	Testing Time	Training Time	Testing Time	Training Time	Testing Time
ESN	0.1145	0.0238	0.1138	0.0213	0.1281	0.0226
LI-ESN	0.1139	0.0225	0.1648	0.0307	0.5086	0.0418
ELM	0.0624	0.0312	0.0624	0.0312	0.0624	0.0312
H-ELM	0.4292	0.1039	0.4894	0.1160	0.1748	0.0696
SAE	143.1321	0.0292	139.4019	0.0268	145.4450	0.0816
SLI-ESN	3.1061	0.1730	3.3148	0.1669	4.3705	0.2193

5. Conclusions

PM_{2.5} is the main source of air pollution, and predicting PM_{2.5} concentration is of great significance for protecting the environment. In order to improve prediction accuracy and reliability, it is important to preprocess the data to eliminate irrelevant and redundant variables before prediction. In this paper, the mRMR is used to screen original dataset to obtain optimal subset. Phase space reconstruction is performed on the optimal subset. And the reconstructed data is used as new input time series of SLI-ESN model for prediction. Experiments show the validity of the SLI-ESN model, which has high prediction accuracy in medium and long term projects, good generalization performance and good application prospects.

Although this paper has achieved the desired results, there are still some issues that need to be addressed in future work. First of all, long-term predictions are not satisfactory. We want to extend the model to longer prediction interval, such as one day, one week or one month. In addition, optimal subset selection and model optimization take a lot of time. In the future, we expect to simultaneously implement input variable selection and model optimization based on an optimization algorithm. Optimization objects include, but are not limited to, input variables, model structure, and model parameters.

Author Contributions: Conceptualization, X.X. and W.R.; methodology, X.X. and W.R.; data curation, X.X.; writing—original draft preparation, X.X. and W.R.; writing—review and editing, X.X. and W.R.

Funding: This research was funded by the National Natural Science Foundation of China (61773087).

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

References

1. Liu, T.; Lau, A.K.H.; Sandbrink, K.; Fung, J.C.H. Time Series Forecasting of Air Quality Based On Regional Numerical Modeling in Hong Kong. *J. Geophys. Res. Atmos.* **2018**, *123*, 4175–4196. [[CrossRef](#)]
2. Cai, S.; Wang, Y.; Zhao, B.; Wang, S.; Chang, X.; Hao, J. The impact of the “air pollution prevention and control action plan” on PM_{2.5} concentrations in Jing-Jin-Ji region during 2012–2020. *Sci. Total Environ.* **2017**, *580*, 197–209. [[CrossRef](#)]
3. Li, L.; Zhang, J.H.; Qiu, W.Y.; Wang, J.; Fang, Y. An Ensemble Spatiotemporal Model for Predicting PM_{2.5} Concentrations. *Int. J. Environ. Res. Public Health* **2017**, *14*, 549. [[CrossRef](#)]
4. Han, W.; Tong, L.; Chen, Y.; Li, R.; Yan, B.; Liu, X. Estimation of High-Resolution Daily Ground-Level PM_{2.5} Concentration in Beijing 2013–2017 Using 1 km MAIAC AOT Data. *Appl. Sci.* **2018**, *8*, 2624. [[CrossRef](#)]
5. Wang, P.; Zhang, H.; Qin, Z.; Zhang, G. A novel hybrid-Garch model based on ARIMA and SVM for PM_{2.5} concentrations forecasting. *Atmos. Pollut. Res.* **2017**, *8*, 850–860. [[CrossRef](#)]

6. 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. [\[CrossRef\]](#)
7. Huang, C.-J.; Kuo, P.-H. A Deep CNN-LSTM Model for Particulate Matter (PM_{2.5}) Forecasting in Smart Cities. *Sensors* **2018**, *18*, 2220. [\[CrossRef\]](#)
8. Qiao, J.; Cai, J.; Han, H.; Cai, J. Predicting PM_{2.5} Concentrations at a Regional Background Station Using Second Order Self-Organizing Fuzzy Neural Network. *Atmosphere* **2017**, *8*, 10. [\[CrossRef\]](#)
9. Rybarczyk, Y.; Zalakeviciute, R. Machine Learning Approaches for Outdoor Air Quality Modelling: A Systematic Review. *Appl. Sci.* **2018**, *8*, 2570. [\[CrossRef\]](#)
10. Oprea, M.; Mihalache, S.F.; Popescu, M. Computational intelligence-based PM_{2.5} air pollution forecasting. *Int. J. Comput. Commun. Control* **2017**, *12*, 365–380. [\[CrossRef\]](#)
11. Deng, M.; Yang, W.; Liu, Q.; Jin, R.; Xu, F.; Zhang, Y. Heterogeneous Space–Time Artificial Neural Networks for Space–Time Series Prediction. *Trans. GIS* **2018**, *22*, 183–201. [\[CrossRef\]](#)
12. Ong, B.T.; Sugiura, K.; Zettsu, K. Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting PM_{2.5}. *Neural Comput. Appl.* **2016**, *27*, 1553–1566. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Li, W.; Kong, D.; Wu, J. A New Hybrid Model FPA-SVM Considering Cointegration for Particulate Matter Concentration Forecasting: A Case Study of Kunming and Yuxi, China. *Comput. Intel. Neurosci.* **2017**, *2017*, 2843651. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Yao, W.; Zhang, C.; Hao, H.; Wang, X.; Li, X. A support vector machine approach to estimate global solar radiation with the influence of fog and haze. *Renew. Energy* **2018**, *128*, 155–162. [\[CrossRef\]](#)
15. Reid, C.E.; Jerrett, M.; Petersen, M.L.; Pfister, G.G.; Morefield, P.E.; Tager, I.B.; Raffuse, S.E.; Balmes, J.R. Spatiotemporal prediction of fine particulate matter during the 2008 northern California wildfires using machine learning. *Environ. Sci. Technol.* **2015**, *49*, 3887–3896. [\[CrossRef\]](#)
16. Sun, W.; Sun, J. Daily PM_{2.5} concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm. *J. Environ. Manag.* **2017**, *188*, 144–152. [\[CrossRef\]](#)
17. Zhang, C.; Ni, Z.; Ni, L. Multifractal detrended cross-correlation analysis between PM_{2.5} and meteorological factors. *Physica A* **2015**, *438*, 114–123. [\[CrossRef\]](#)
18. Zhu, J.Y.; Zhang, C.; Zhang, H.; Zhi, S.; Li, V.O.; Han, J.; Zheng, Y. pg-causality: Identifying spatiotemporal causal pathways for air pollutants with urban big data. *IEEE Trans. Big Data* **2018**, *4*, 571–585. [\[CrossRef\]](#)
19. Chen, Z.; Xie, X.; Cai, J.; Chen, D.; Gao, B.; He, B.; Cheng, N.; Xu, B. Understanding meteorological influences on PM_{2.5} concentrations across China: a temporal and spatial perspective. *Atmos. Chem. Phys.* **2018**, *18*, 5343–5358. [\[CrossRef\]](#)
20. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [\[CrossRef\]](#)
21. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
22. Brown, G.; Pocock, A.; Zhao, M.J.; Luján, M. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *J. Mach. Learn. Res.* **2012**, *13*, 27–66.
23. Uğuz, H. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowledge-Based Syst.* **2011**, *24*, 1024–1032. [\[CrossRef\]](#)
24. Jaeger, H.; Haas, H. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science* **2004**, *304*, 78–80. [\[CrossRef\]](#)
25. Ozturk, M.C.; Xu, D.; Principe, J.C. Analysis and design of echo state networks. *Neural Comput.* **2007**, *19*, 111–138. [\[CrossRef\]](#)
26. Jaeger, H.; Lukoševičius, M.; Popovici, D.; Siewert, U. Optimization and applications of echo state networks with leaky-integrator neurons. *Neural Netw.* **2007**, *20*, 335–352. [\[CrossRef\]](#)
27. Takens, F. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence*; Springer: Berlin, Heidelberg, 1981; pp. 366–381.
28. Han, M.; Ren, W.; Xu, M.; Qiu, T. Nonuniform state space reconstruction for multivariate chaotic time series. *IEEE T. Cybern.* **2019**, *49*, 1885–1895. [\[CrossRef\]](#)
29. Løkse, S.; Bianchi, F.M.; Jenssen, R. Training echo state networks with regularization through dimensionality reduction. *Cogn. Comput.* **2017**, *9*, 364–378. [\[CrossRef\]](#)

30. Huang, G.B.; Zhou, H.; Ding, X.; Zhang, R. Extreme learning machine for regression and multiclass classification. *IEEE Trans. Syst. Man Cybern. Part B-Cybern.* **2012**, *42*, 513–529. [[CrossRef](#)]
31. Tang, J.; Deng, C.; Huang, G.B. Extreme learning machine for multilayer perceptron. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *27*, 809–821. [[CrossRef](#)]
32. Chen, Y.; Lin, Z.; Zhao, X.; Wang, G.; Gu, Y. Deep learning-based classification of hyperspectral data. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2014**, *7*, 2094–2107. [[CrossRef](#)]
33. Kim, H.; Eykholt, R.; Salas, J.D. Nonlinear dynamics, delay times, and embedding windows. *Physica D* **1999**, *127*, 48–60. [[CrossRef](#)]

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).