mathematics

Article

Variational Bayesian Network with Information Interpretability Filtering for Air Quality Forecasting

Xue-Bo Jin [1,2], Zhong-Yao Wang [1,2], Wen-Tao Gong [1,2], Jian-Lei Kong [1,2,*], Yu-Ting Bai [1,2], Ting-Li Su [1,2], Hui-Jun Ma [1,2] and Prasun Chakrabarti [3]

[1] Artificial Intelligence College, Beijing Technology and Business University, Beijing 100048, China
[2] China Light Industry Key Laboratory of Industrial Internet and Big Data, Beijing Technology and Business University, Beijing 100048, China
[3] Department of Computer Science and Engineering, ITM SLS Baroda University, Vadodara 391510, India
* Correspondence: kongjianlei@btbu.edu.cn

Abstract: Air quality plays a vital role in people's health, and air quality forecasting can assist in decision making for government planning and sustainable development. In contrast, it is challenging to multi-step forecast accurately due to its complex and nonlinear caused by both temporal and spatial dimensions. Deep models, with their ability to model strong nonlinearities, have become the primary methods for air quality forecasting. However, because of the lack of mechanism-based analysis, uninterpretability forecasting makes decisions risky, especially when the government makes decisions. This paper proposes an interpretable variational Bayesian deep learning model with information self-screening for PM2.5 forecasting. Firstly, based on factors related to PM2.5 concentration, e.g., temperature, humidity, wind speed, spatial distribution, etc., an interpretable multivariate data screening structure for PM2.5 forecasting was established to catch as much helpful information as possible. Secondly, the self-screening layer was implanted in the deep learning network to optimize the selection of input variables. Further, following implantation of the screening layer, a variational Bayesian gated recurrent unit (GRU) network was constructed to overcome the complex distribution of PM2.5 and achieve accurate multi-step forecasting. The high accuracy of the proposed method is verified by PM2.5 data in Beijing, China, which provides an effective way, with multiple factors for PM2.5 forecasting determined using deep learning technology.

Keywords: multiple factors; time series forecasting; deep learning; interpretability; data filtering; variational Bayesian

MSC: 68T07

check for updates

Citation: Jin, X.-B.; Wang, Z.-Y.; Gong, W.-T.; Kong, J.-L.; Bai, Y.-T.; Su, T.-L.; Ma, H.-J.; Chakrabarti, P. Variational Bayesian Network with Information Interpretability Filtering for Air Quality Forecasting. *Mathematics* **2023**, *11*, 837. https://doi.org/10.3390/math11040837

Academic Editors: Ioannis E. Livieris, Panagiotis Pintelas and Stavros Stavroyiannis

Received: 2 January 2023
Revised: 31 January 2023
Accepted: 1 February 2023
Published: 7 February 2023

1. Introduction

Nowadays, people's living standards are improving with the gradual development of social and economic levels, and they are paying more and more attention to their well-being. It is well known that air quality has a significant impact on health. Therefore, time series forecasting technology for air quality has attracted wide attention. Among the many factors affecting air quality, PM2.5 is the most significant. PM2.5 is a mixture of particles with a diameter of less than or equal to 2.5 microns, including toxic and harmful substances such as elemental carbon, volatile organic compounds (VOC), sulfides, condensates, metal particles, etc. [1], with which it is easy to cause various respiratory and cardiovascular diseases and seriously affect people's health [2]. Accurately forecasting the concentration of PM2.5 can supply complete air quality forecasting information, provide a scientific and accurate theoretical basis for the prevention and control of air pollution, and make the government and the public understand air quality comprehensively.

With the development of sensing technology, historical PM2.5 concentration and other air pollutant data are convenient to obtain, which makes it possible to model the changing pattern of PM2.5. However, the formation mechanism and change process of PM2.5 data are very complex. They are also affected by seasonal and geographical conditions, including the temporal and spatial dimensions. PM2.5 is a kind of non-stationary time-series data with complex nonlinear and distribution characteristics that increases the difficulty of forecasting PM2.5 concentration accurately. The related research has always been the hotspot in the time series forecasting field.

There have been three methods for air quality forecasting, i.e., the statistical, machine learning, and deep learning models. Simple statistical models include time series models: the autoregressive (AR) models [3,4], the moving average (MA) models and the autoregressive moving average (ARMA) models [5], the machine learning models including artificial neural networks (ANN) [6,7], and random forest (RF) [8], etc. These methods are easy to implement and interpret. However, their small parameter scales and simple structures give them a low ability for feature representation and nonlinear fitting. These methods are only suitable for application to small and stable data sets and cannot solve the large data sets with strong nonlinearity or address the high complexity of the actual PM2.5 data. However, the prediction models may use the state space models and input–output representations [9,10], and the parameters of the models can be obtained by using the parameter identification methods [11,12] such as the least squares algorithms, the Newton algorithms [13–15] and the gradient search algorithms [16,17] and so on.

Deep learning methods have been widely used in air quality forecasting on the basis of their powerful modeling nonlinear ability for multivariable and multi-channel massive time series. In particular, recurrent neural networks (RNN) [18], long short-term memory networks (LSTM) [19], gated recurrent units (GRU) [20], etc., have become necessary ways to perform air quality forecasting.

On the other hand, mechanism-based analysis is still an essential way of air quality forecasting. For example, air pollution sources such as PM2.5 have temporal and spatial characteristics, and the pollution in adjacent areas is relatively similar. PM2.5 will also be affected by other air quality and meteorological factors. For example, excessive humidity can easily make PM2.5 dissipate more slowly. Therefore, for PM2.5, which has many influencing factors and complex changes, it is necessary to consider the impact of elements on the concentration change of PM2.5. Researchers found that mechanical pollution factors can help air quality forecasting accuracy. While the formation of its mechanism is very complex, the forecasting accuracy cannot be guaranteed because it only uses the mechanical method in the actual air quality forecasting. Therefore, in recent years, the deep network has become the primary method of air quality forecasting.

However, when training the deep network, it is incorrect to blindly use all the obtained data because a large number of data often have redundant information, which increases the training cost of the network and does not improve or even reduce the forecasting accuracy. Therefore, it is necessary to minimize useless information to improve the network's training efficiency and forecasting accuracy. Researchers have proposed information screening methods and applied them to air quality forecasting. The standard variable screening methods include the Granger causality analysis method [21], mutual information method [22], Spearman rank correlation coefficient [23], a data screening based on single Gauss [24], etc. These methods can quantitatively analyze the relationship between the factors in the multi-dimensional time series and eliminate the variables' useless, inconsistent and conflicting factors. However, these methods can only analyze the conflict and inconsistent relationship but do not consider the redundant relationship. This is because it is known that the redundant input data will make the deep neural network overfit in the modeling process, thus reducing the forecasting performance. Therefore, reducing redundant information remains an open research direction, particularly screening the input information of the deep network effectively.

In addition, the collected air quality data often has another problem: it contains a complex noise distribution. As we know, the complex noise distribution will mask the essential characteristics of the time series data, such as periodicity, seasonality, etc., and make it challenging to model, thus reducing its forecasting accuracy. Researchers have tended to add the denoising steps before forecasting, such as in the wavelet transform denoising method [25], empirical mode decomposition (EMD) method [26], etc. Conversely, for deep networks, wavelet transform and EMD, an adjustment in parameters is required; therefore, these methods cannot achieve an end-to-end forecasting network.

The main contributions of this paper are as follows.

(1) According to the air quality formation mechanism, this paper provides an interpretable information screening mechanism based on multivariable and multi-channel massive time series data. Compared with the existing methods [21–23], the interpretable information screening mechanism can mine the correlation and redundancy between multiple time series input variables that affect air quality, extract helpful information more effectively and eliminate information redundancies.

(2) The interpretable screening filter is embodied in the learning framework to build an end-to-end forecasting network. The screening filter learns the parameters from the input data through Bayesian hyperparametric optimization with multiple Gaussian peaks.

(3) The variational inference structure is introduced into the gated recurrent unit (GRU), following the interpretable information screening layer, to mine the spatial and temporal relation of PM2.5 and enhance the modeling ability of the network for nonlinearity and complex noise distribution.

This paper is organized as follows: Section 2 introduces the related research work for the air quality field, and Section 3 describes the data sets used. Section 4 introduces in detail the methods and forecasting models proposed. Section 5 establishes experiments and the analysis of the results. In Section 6, we give conclusions and suggest future related work.

2. Related Works

2.1. PM2.5 Forecasting Method Based on Traditional Methods

Traditional PM2.5 forecasting methods are mainly statistical models and machine learning methods. Because of their relatively simple structures, the statistical methods mostly consider the formation mechanism of PM2.5. They are widely used in the field of air quality forecasting. Liu et al. [27] combined ARIMA with numerical forecasting to forecast the daily and hourly PM2.5 concentration in Hong Kong. Zeng et al. [28] studied the relationship between PM2.5 and meteorological factors in Chengdu within 24 h. They used the generalized additive model to forecast the concentration of PM2.5. Although the traditional forecasting method based on statistics has a relatively good capability in forecasting PM2.5, it has limitations because the formation of PM2.5 is very complex.

Machine learning technology in air pollution forecasting is mainly based on historical data and was modeled nonlinearly, which is more in line with the nonlinearity of actual air pollution data, thus producing higher forecasting accuracy. Wang et al. [29] used an optimal network structure, based on the BP neural network, to forecast the concentration of PM2.5. Fang et al. [8] used a machine learning method to forecast the concentration of PM2.5 in Beijing, China, based on ground LiDAR and meteorological data. Chang et al. [30] used the self-organizing mapping method to extract the temporal and spatial characteristics of PM2.5 concentration and used the back-propagation neural network to make a forecast. Shahriar et al. [31] evaluated hybrid models (ARIMA, ANN, SVM, PCR, DT, and CatBoost) to forecast environmental PM2.5 concentration in many cities in Bangladesh. Among these models, CatBoost has the best performance. Carreno et al. [32] used machine learning techniques to forecast particulate matter levels based on meteorological and climatic features in Talca, Chile.

With the development of sensors and storage technology, there are more and more data about air quality. Although the traditional forecasting methods based on machine learning techniques have a relatively good effect on forecasting PM2.5, these methods have limitations and are more suitable for small data sets. These traditional forecasting methods of simple structure can no longer meet the complex modeling capabilities required in the context of big data, and it is easy to fall into overfitting, thus reducing the forecasting accuracy.

2.2. PM2.5 Forecasting Method Based on Deep Learning

A deep learning network has recently been widely used in air quality forecasting because of its robust modeling and learning ability. Sun et al. [33] proposed a PM2.5 concentration estimator based on a deep convolution neural network. Shi et al. [34] proposed using an improved integrated depth neural network method based on an attention mechanism to forecast the concentration of PM2.5. Mengfan et al. [35] proposed a new PM2.5 concentration hybrid forecasting model that combines a long short-term memory neural network (LSTM) and a specific convolutional neural network (CNN) with a 1 × 1 core size. Wang et al. [36] proposed a spatiotemporal convolution recursive long short-term memory (CR-LSTM) neural network model to forecast PM2.5 for long-term forecasting. Wang et al. [37] proposed a PM2.5 forecasting model that combines content and a bidirectional gated recurrent unit based on sense. These models do not consider the selection of input data for the mechanical characteristics of PM2.5 but improve the forecasting performance based on the improvement of the model. Prihatno et al. [38] proposed a single-dense layer bidirectional long short-term memory (BiLSTM) model to forecast the PM2.5 concentrations in the indoor environment by using time series data.

Although the deep learning method is widely used in the field of time series data for its strong modeling ability, there are limitations to multi-dimensional time series data. Many conflicting and inconsistent data often occur between multi-dimensional time series, such as air quality data. They will reduce the learning efficiency of deep learning and affect the understanding of the data characteristics of the model, thus reducing the accuracy of forecasting.

2.3. Multi-Factor PM2.5 Forecasting Method Based on Variable Screening

Given the problems in multi-dimensional air quality time series, many researchers have proposed various methods to consider the relationship between variables to reduce the dimensions of the input series. Zhu et al. [39] proposed an attention-based parallel network (Apnea) to forecast PM2.5 and used the maximum information coefficient (MIC) to conduct spatio-temporal correlation analysis, taking complete account of the linear and nonlinear relationship between the data of each monitoring station. Liu et al. [7] proposed a feature selection algorithm, based on pseudo F statistics, which obtains prime variables related to PM2.5 and then uses support vector regression to forecast PM2.5. Pak et al. [40] used mutual information (MI) to analyze the spatial–temporal correlation of air quality data, taking complete account of the whole region of China centered on the target monitoring station and historical air quality meteorological data. Cifuentes et al. [41] analyzed the impact of different forecasting variables based on the Spearman coefficient, principal component analysis (PCA), and meteorological data on air quality forecasting. Zhu et al. [42] quantified the relationship between different variables through the Pearson coefficient and considered the influence of different monitoring sites and seasons on PM2.5 forecasting.

Although the above methods consider the correlation between variables, these feature selection methods have limitations because they need to be calculated separately and then artificially screened. Moreover, these methods can only analyze the correlation between time series and cannot interpret their redundancy. Finally, the data can be input into the model for training. As such, the selection step is separated from the network and is only a data preprocessing process, which cannot realize an end-to-end forecasting network.

2.4. PM2.5 Forecasting Method for Noise Problems

Air quality time series often contain complex noise distribution, affecting the forecasting model's learning and modeling for the time series data [43]. Therefore, the data processing methods of noise reduction and denoising have been favored by researchers. Jin et al. [44] proposed a decomposition integration forecasting method based on wavelet denoising for PM2.5 data. Samal et al. [45] proposed a hybrid PM2.5 forecasting framework called a time convolution denoising automatic encoder (TCDA) network, which uses the denoising self-encoder to denoise PM2.5 data. Cai et al. [46] designed a denoising self-coding deep network based on LSTM to develop the accuracy of air pollutants forecasting models. Jin et al. [47] proposed a hybrid deep learning forecaster, using empirical mode decomposition (EMD) to decompose the data into components. They then used a deep network to forecast the PM2.5 in Beijing, China.

Almost all of the above methods are preprocessing for the deep networks, increasing the complexity of forecasting models and operation steps. Besides, these methods have some limitations, such as a lack of a strict mathematical foundation, a small scope of application, etc. Moreover, most of these methods need enough prior knowledge.

Based on the problems and advantages of the above research, this paper considers the correlation and redundancy between input data. It builds a self-learning optimization layer with an interpretable information screening mechanism to improve forecasting network interpretability and accuracy. Further, we create a Bayesian GRU network with variational inference, overcoming the problem that traditional deep learning struggles to fit the complex noise distribution and improving the model's multi-step forecasting accuracy.

3. Data Set and Spatial Correlation Analysis

3.1. Data Set

This paper uses the hourly PM2.5 and meteorological data in Beijing from January 2019 to December 2021. Each data set contains 26,280 data points. The meteorological data include temperature, wind direction, and humidity. The sampling frequency of all data is 1 h. The data are normalized as follows:

$$x' = \frac{x - \mu_x}{\sigma_x} \tag{1}$$

where x represents the input observation data, μ_x represents the mean value of the observation data, and σ_x represents the variance of the observation data. The role of Z-score standardization is to unify the data distribution of characteristics and reduce the impact of the characteristics of different distributions on the final results.

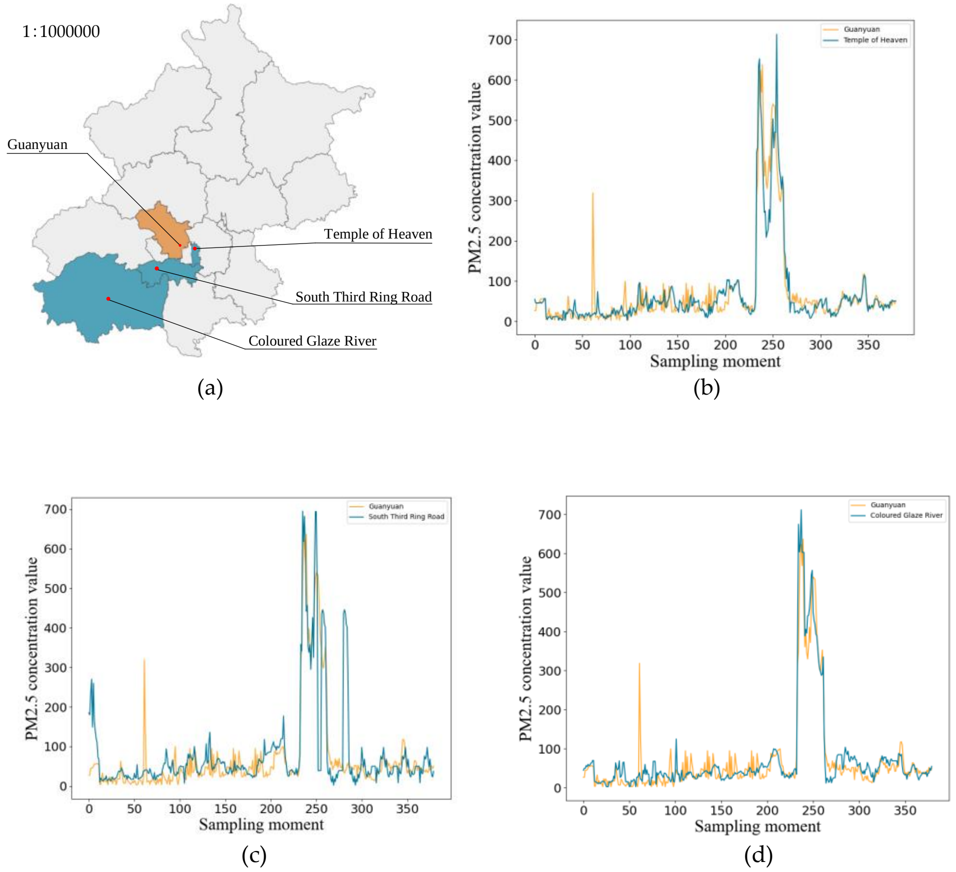
3.2. Spatial Correlation Analysis

The spatial correlation in the air quality between different areas in Beijing is shown and analyzed below.

Figure 1a shows the geographic location of four different areas in Beijing. The orange area represents "Guanyuan" in the Haidian District in Beijing. The blue areas represent the "Temple of Heaven" in the Dongcheng District in Beijing, the "South Third Ring Road" in the Fengtai District in Beijing, and the "Coloured Glaze River" in the Fangshan District in Beijing, respectively. We will analyze the spatial correlation in the air quality between the "Guanyuan" area and the other three areas.

Figure 1b shows 300 samplings from the "Guanyuan" area and "Temple of Heaven" area, and it can be seen that there is a solid spatial correlation in the air quality data between these two areas. It can also be seen from Figure 1a that those two areas are very near to each other. The high degree of coincidence proves a redundant relationship between them. It is bad practice to put this kind of data with solid redundancies into the neural network for training. For example, a large amount of data leads to an increase in training time; or, when training the neural network, it will lead to overfitting the data and poor forecasting results in practical application.

Figure 1c shows the spatial correlation in the air quality data of the "Guanyuan" area and the "South Third Ring Road" area. It can be seen that the similarity of the data curve between the "Guanyuan" area and the "South Third Ring Road" area is different from that of the "Guanyuan" area and the "Temple of Heaven" area, as shown in Figure 1b. Still, there is also a specific correlation between them. In addition, air pollutants are very vulnerable to the weather and so the wind blows air pollutants from one area to another relatively close location. Moreover, the “Guanyuan” area and the “South Third Ring Road” are somewhat near, and the other area will affect their air quality. There is a specific correlation between the data of the two regions, but the redundancy between them has been less extensively compared than the “Guanyuan” area and the “Temple of Heaven” area.



(a) (b) (c) (d)

Figure 1. Spatial Correlation between different areas in Beijing. (**a**) the geographic location of four different areas in Beijing. The scale is 1:1,000,000. (**b**) PM2.5 changes in Guanyuan, Haidian, and Temple of Heaven, Dongcheng. (**c**) PM2.5 changes in Guanyuan, Haidian, and South Third Ring Road, Fengtai. (**d**) PM2.5 changes in Beijing Guanyuan and Coloured Glaze River, Fangshan.

Figure 1d shows the spatial correlation in the air quality data of the “Guanyuan” area and the “Coloured Glaze River” area. It can be seen there is less similarity between the two regions when comparing Figure 1b,c. It can also be seen that the air pollutants in the two areas will not affect each other due to the distance between the two regions, and that the correlation is low. Therefore, when forecasting the air quality in the “Guanyuan” area, it is not necessary to send the “Coloured Glaze River” data into the neural network for training.

This paper will design a deep forecasting network based on information interpretability filtering. As the first layer of the network, the information interpretability filtering layer selects the m-dimensional variable with high correlation and low redundancy from the n-dimensional time series data. Then, these variables are input to train the deep forecasting model. The flow of this model is shown in Figure 2. We will detail the information interpretability filtering in Section 4 and the deep forecasting model in Section 5.

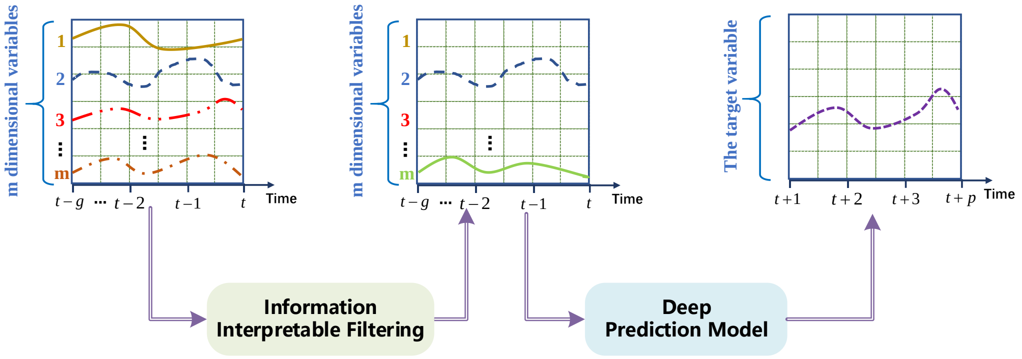

m dimensional variables
1
2
3
⋮
m
$t-g$ ⋯ $t-2$ $t-1$ t Time
m dimensional variables
1
2
3
⋮
m
$t-g$ ⋯ $t-2$ $t-1$ t Time
The target variable
$t+1$ $t+2$ $t+3$ $t+p$ Time
Information Interpretable Filtering
Deep Prediction Model

Figure 2. A deep forecasting model based on information interpretability filtering (m < n). The dotted lines with different colors in the figure represent different variables.

4. Information Interpretable Filtering

From a mechanism analysis, the relationship between meteorological elements and air quality contains a complex implicit nonlinear relationship [48–50]. For example, when the humidity is high, the water vapor content in the air is significant and tiny water particles surround the PM2.5 solid particles. Due to the increase in moisture content, the density and concentration of PM2.5 particles decrease, decreasing the PM2.5 value [51]. At the same time, rainfall has a significant impact on air quality. Rainwater will adsorb air pollutants and cause them to settle, thereby reducing the concentration of PM2.5 [52]. In addition, there will be a more significant impact between the regions. Due to the flow and diffusion of air, this method will be helpful for forecasting of air quality. However, the amount of information on these mechanisms is vast and complex, making accurate modeling difficult. Air quality forecasting modeling based on big data has recently received extensive attention.

The researchers found that redundant input data not only increase the training cost of the network but also reduce the forecasting accuracy due to overfitting. We know that the air quality data contained redundant information if the two regions were very close. Therefore, using a large amount of data cannot effectively improve forecasting performance. On the other hand, in deep learning networks, more input data will not make the network work more effectively.

The researchers began to consider the correlation between the data and used the method of data analysis to select the data. However, the lack of interpretability of classical data analysis methods makes the forecasting results lack analytical support. This paper will give an information filtering framework, with interpretability based on optimization strategies, by the following two steps:

Step 1: First, mutual information is used to calculate the relationship degree between the variables in the data set, including PM2.5 content or meteorological factor, e.g., temperature, humidity, and wind speed, in different regions.

The mutual information (MI) is calculated in formula (2):

$$I(x, y^i) = p\left(x, y^i\right) \log \frac{p(x, y^i)}{p(x)p(y^i)} \tag{2}$$

where x is set as the target variable, i.e., PM2.5 content; y^i is the variable to be selected; $p(x, y^i)$ is the random distribution of the two variables; $p(x)$ and $p(y^i)$ are marginal distributions.

When the mutual information value between x and y^i is more extensive, they have more related information. However, the data with a high correlation may also contain more redundant information, especially in air quality forecasting (as discussed in Section 3). Redundant information is unfavorable for neural network training. Therefore, we chose the most relevant variables but excluded the data with high redundancy. Consequently, we discussed the concept of "effective" mutual information.

Step 2: Adaptive information distance (AID) is proposed to select the input variables with high correlation but low redundancy for the deep network:

$$D(y^i) = \sqrt{(y^i - x)^T S^{-1} (y^i - x)} \tag{3}$$

where y^i is the selected variable based on the mutual information (2), S is the parameter that needs to be determined so that the selected variables can have high correlation and low redundancy, and x is the target variable.

In the information interpretability filtering framework, the parameter S of AID, which directly affects forecasting performance, is optimized to ensure the reliability and validity of variable screening via the Bayesian hyperparameter optimization method [53]. In this study, we use the root mean square error (RMSE) as the objective function for optimizing hyperparameters:

$$loss(w) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_t - \hat{x}_t)^2} \tag{4}$$

where $w = [S, x]$ is the hyperparameter that needs to be optimized for AID, T is the number of input samples, x_t is the target variable, $\hat{x}_t$ is the forecasting, and t is the time index. The set of hyperparameters w^* can be obtained by minimizing $loss(w)$:

$$w^* = \underset{w \in W}{\text{argmin}}\, loss(w) \tag{5}$$

where w^* is the optimal parameter determined by Bayesian hyperparameter optimization, w is a set of input parameters, and W is the parameter space of multi-dimensional parameters.

Information interpretability filtering consists of mutual information, AID, and Bayesian hyperparameter optimization. Among these, the mutual information methods selects variables with a high correlation with the target variable. AID sets variables with lower redundancy so that the variables with high correlation and low redundancy will be filtered out. Bayesian hyperparameter optimization learns the parameters required by AID according to the different input data in order to obtain correct input data, thereby improving the forecasting accuracy of the target variable. The computational flow chart of the interpretability filtering of information is shown in Figure 3.

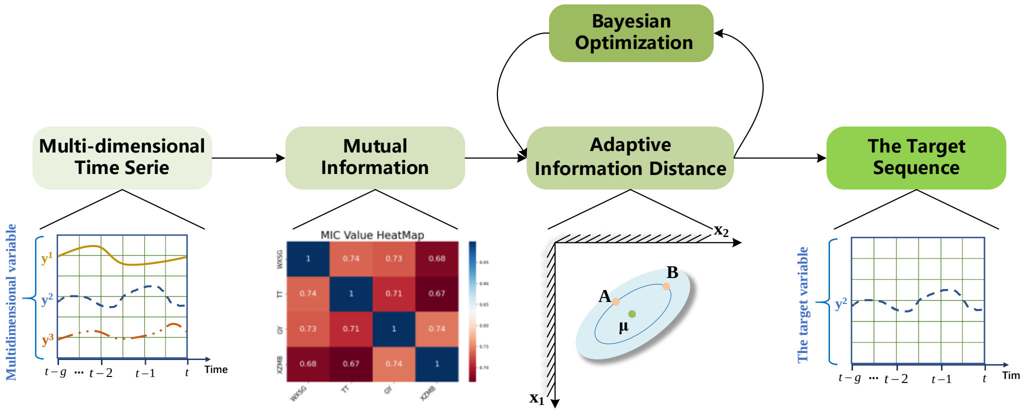


Figure 3. Information interpretability filtering. The dotted lines with different colors in the figure represent different variables.

5. Deep Forecasting Network

Here, we use variational Bayesian gated recurrent unit (VBGRU) in this forecasting model. VBGRU selects GRU as the primary network structure and uses variational methods to train weights with distributed characteristics. The network structure of VBGRU is shown in Figure 4.

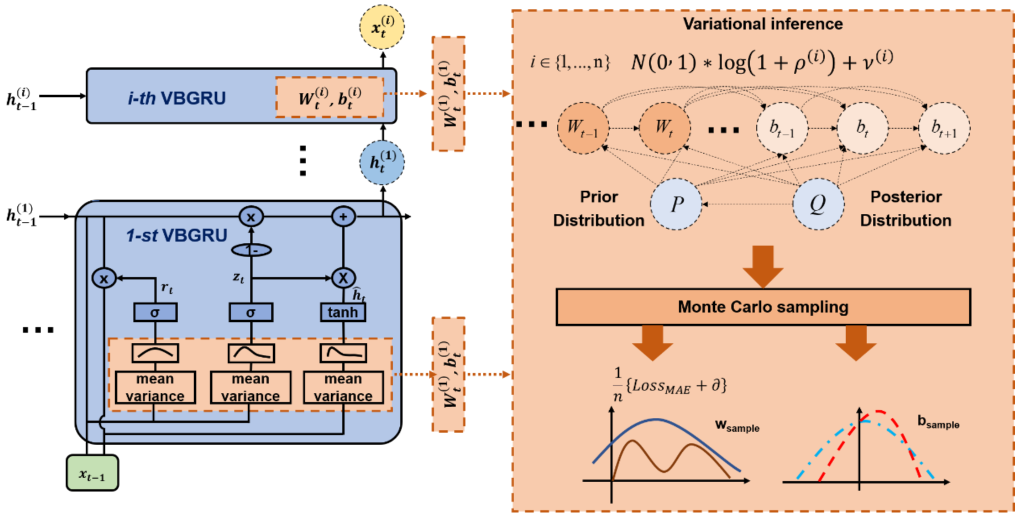


Figure 4. Deep forecasting network. In the lower right corner of the figure, the blue line represents the weight and bias generated by initialization, and the red line represents the new weight and bias sampled from the corresponding mean and variance through the Monte Carlo sampling module.

As shown in Figure 4, the weights and biases of VBGRU are transformed into distribution via the variational inference method. The specific process is: first, initialize the weight W and bias b of the VBGRU to the distribution of a specific mean and a specific variance. Then, the new weight W_{sample} and bias b_{sample} are sampled from the corresponding mean and variance through the Monte Carlo sampling module. The weight calculation method not only optimizes the performance indicators of the model but also learns the uncertainty of the network forecasting on a specific data point. In addition, VBGRU can obtain multiple model outputs by adding sampling points to calculate the uncertainty of the model at a specific point.

Let $W_{(n)}^{(i)}$ denote the nth sampling weight of the ith layer and $b_{(n)}^{(i)}$ denote the bias. In deep Bayesian networks, neither weight nor bias is a definite number, but results are obtained by sampling on a distribution; the distribution parameters $\rho^{(i)}$ and $\nu^{(i)}$ of weights and biases are obtained through training, and their relationship with weights $W_{(n)}^{(i)}$ and biases $b_{(n)}^{(i)}$ are shown in formulas (6) and (7).

$$W_{(n)}^{(i)} = N(0,1) * \log(1 + \rho^{(i)}) + \nu^{(i)} \tag{6}$$

$$b_{(n)}^{(i)} = N(0,1) * \log(1 + \rho^{(i)}) + \nu^{(i)} \tag{7}$$

Like ordinary deep models, deep Bayesian networks need to determine the optimization goal of the network model through the loss function. Additionally, the loss function needs to be guaranteed to be differentiable so that the network model parameters can be updated using the back-propagation algorithm. The deep Bayesian network uses the variational inference method to calculate the approximate distribution of the complex distribution of model parameters. The degree of approximation between the two is measured by the Kullback–Leibler divergence (KL divergence) in order to realize the differential operation of the loss function on its related variables. However, since the forecasting task is usually performed to forecast a specific value, only using the KL divergence as the loss error between the forecasting result and the network output will mean the model can only learn the distribution characteristics of the data. Therefore, the loss function of the deep Bayesian network is composed of two parts, the differentiable mean absolute error (MAE) and the KL divergence.

Therefore, the formula of the loss function of the deep Bayesian network is as follows:

$$Loss = Loss_{MAE} + \partial \cdot \left[\log\left(Q\left(\omega^{(n)} \middle| \theta \right) \right) - \log\left(P\left(\omega^{(n)} \right) \right) \right] \tag{8}$$

where ∂ represents an error weight parameter, generally set as the reciprocal of the number N of all training samples, namely $\partial = 1/N$; $P(\omega)$ is a manually set low-entropy prior distribution; $Q(\omega|\theta)$ is the posterior distribution of a given parameter.

Through the above analysis, we replace all the weights and biases inside the GRU with different distributions, initialize it to a standard normal distribution, and update the weight parameters of the network model through the Adam optimizer to obtain the best network parameters, that is, to obtain the optimal mean and variance of the weight distribution and bias distribution. Similarly, when using the trained model, the weight and bias distribution are sampled multiple times by sampling, and various sets of forecasting results are obtained. Finally, the multi-group forecasting results are averaged, which are the forecasted values output by the network.

6. Experiments

6.1. Experiment Setup and Evaluation Indicators

The learning and forecasting step of the network model is set to 24, i.e., 24 h of historical data are input into the model as a sample to forecast the next 24 h. The selection of the hyperparameter of the forecasting model is based on experience and multiple attempts. The batch size is set to 40, the number of training epochs is set to 50, and the learning rate is set to 0.001. The Adam optimization algorithm is used to make the model perform supervised learning.

We adopt three evaluation functions to evaluate the forecasting performance of the model, namely: root means square error (*RMSE*), mean square error (*MSE*), and mean absolute error (*MAE*). They are calculated by formulas (9)–(11) [24], respectively.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\hat{x}_t - x_t)^2} \tag{9}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (\hat{x}_t - x_t)^2 \tag{10}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |(\hat{x}_t - x_t)^2| \tag{11}$$

where n represents the total number of samples in the data set, x_t and $\overline{x}_t$ represent the actual value of PM2.5 and the average value of the actual value at time t, respectively; $\hat{x}_t$ and $\overline{\hat{x}}_t$ represent the forecasted value and average value of PM2.5 concentration obtained through experiments at time t, respectively. *RMSE*, *MSE*, and *MAE* indicate that the smaller the value is, the better the model's forecasting performance will be.

Analysis of variance is a statistical analysis method used to analyze the influence of categorical independent variables on numerically dependent variables. It can also be used to analyze the significance test of the difference between the mean values of two or more samples. We also used a one-way analysis of variance to validate the results. Variance analysis first acquires statistics on each factor in the target, divides the total change into test value and error value, and constructs statistic F. The formula is as follows:

$$F = \frac{MSTR}{MSE} \tag{12}$$

where $MSTR$ measures the variation among the $k \geq 2$ samples and MSE measures the variation within the samples. When the value of F is higher than the value of F-crit (the critical value of the F-test), it means that there is a significant difference in the level of different factors or the factors have a substantial impact on the results. The opposite means that there is no significant difference in the level of different factors, or that these factors have no substantial impact on the results.

6.2. Numerical Experiment and Analysis of PM2.5 in Different Regions

In this part, we analyze the correlation between PM2.5 in the Guanyuan (GY) area of Haidian District with PM2.5 in other different regions [39], including Temple of Heaven (TOH) in Dongcheng District, South Third Ring Road (STRR) in Fengtai District and Coloured Glaze River (CGR) in Fangshan District. Figure 5 shows a schematic diagram of the positions between different regions. The MI value and AID value between PM2.5 in different regions and PM2.5 in the GY region are shown in Table 1.

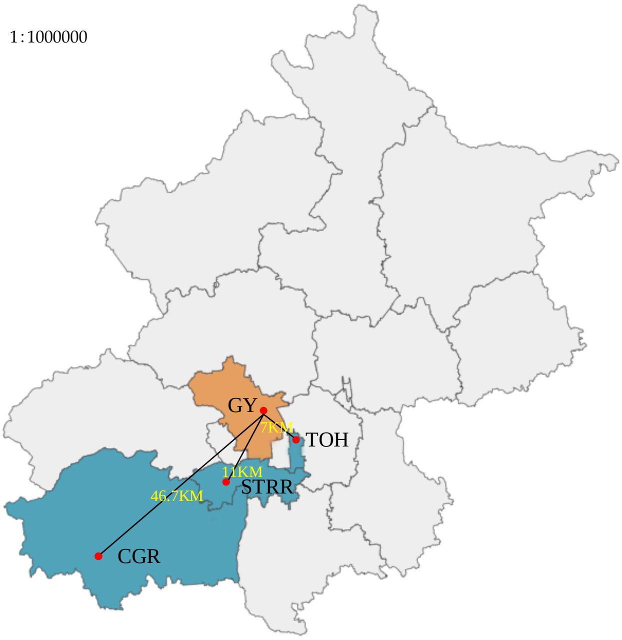


Figure 5. Distance between target area and other areas. The scale is 1:1,000,000.

Table 1. Comparison of MI value and AID value between PM2.5 of GY and PM2.5 of TOH, STRR, CGR.

	TOH	**STRR**	**CGR**
MI	0.71	0.56	0.43
AID	1.39	1.46	1.53

It can be seen from Table 1 that the MI values between PM2.5 in the GY region and the other three areas are TOH (0.71), STRR (0.56), and CGR (0.43) in order. Moreover, although the TOH region has a high correlation with the GY region, its AID is 1.39, indicating that there is a high degree of redundancy between the two. Its AID value is also significant, indicating less information about the relationship between the two, while the MI and AID values of STRR are 0.56 and 1.46, respectively, indicating a good correlation and low redundancy. In contrast, the correlation between the CGR and GY regions is very low. As such, we choose PM2.5 in the STRR region as an auxiliary variable to forecast PM2.5 in the GY region.

Figure 6 shows the comparison results of the three evaluation metrics. It is clear from the graph that the RMSE, MSE, and MAE of the combined STRR have the slightest error from the actual values.

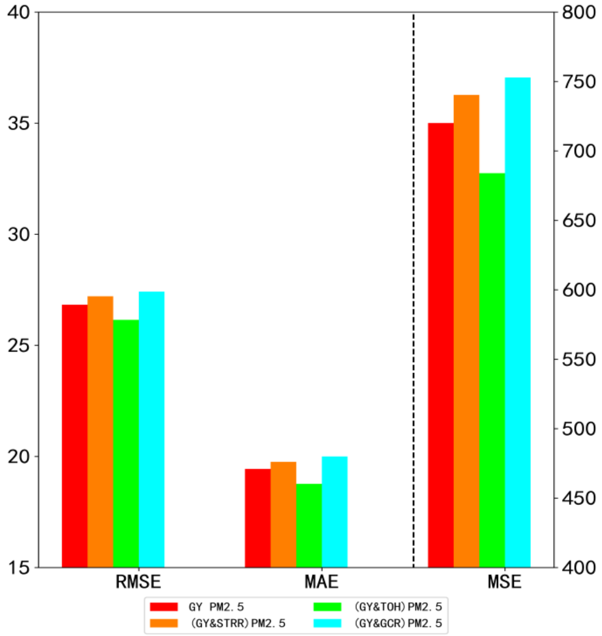


Figure 6. Two-coordinate error histogram of forecast results based on PM2.5 in different regions.

Table 2 shows that the RMSE, MSE, and MAE results of univariate forecasting using only PM2.5 in the GY region are 26.83, 720.09, and 19.44. When adding PM2.5 in different regions as auxiliary variables to forecast PM2.5 in the GY region, only the use of PM2.5 in the STRR region as an auxiliary variable improved the forecasting accuracy. The results were 26.15, 684.02, and 18.77, which are 2.5%, 5.0%, and 3.5%, respectively, lower than PM2.5 in the GY region, and the fluctuation of the results is also lower than that for PM2.5 in the GY region. Using PM2.5 in the other two regions as an auxiliary variable for forecasting did not improve the forecasting results but increased the forecasting error and reduced the model's forecasting performance. Besides, since the forecasting model in this experiment is the same, the training time for using different regions as the input data of the network is all around 46 s. The results shown in Table 2 verify the correctness of the analysis of the developments in Table 1 and verify the correctness of the relationship between region distance and forecasting performance described in Figures 1–3 in Section 3.

Table 2. Analysis of the forecasting results based on the historical data of PM2.5 in different regions as the input data of the network.

Input Data	RMSE	MSE	MAE	Training Time (s)
GY PM2.5 (only GY's PM2.5 data as input)	26.83 ± 0.0012	720.09 ± 0.0712	19.44 ± 0.007	44.25
(GY & TOH) PM2.5 (GY and TOH's PM2.5 data together as input)	27.21 ± 0.0054	740.42 ± 0.0616	19.76 ± 0.0032	48.44
(GY & STRR) PM2.5 (GY and STRR's PM2.5 data together as input)	26.15 ± 0.0011	684.02 ± 0.0702	18.77 ± 0.0015	44.81
(GY & CGR) PM2.5 (GY and CGR's PM2.5 data together as input)	27.42 ± 0.0089	752.90 ± 0.0811	20.00 ± 0.0033	46.35

We also used a one-way analysis of variance to validate the results. The result of the one-way analysis of variance is shown in Table 3.

Table 3. One-way analysis of variance on forecasting results based on the historical data of PM2.5 in different regions as the input data of the network. In this table, SS is the sum of squares and df is the degrees of freedom. The MS (mean squares) is the estimate of variance given by SS/df.

Source	SS	df	MS	F	F-Crit
Type	13,019.67	3	4339.89	10.42	2.61
Error	8,753,742.65	21,020	416.45		
Total	8,766,762.32	21,023			

Table 3 shows that the value of F (10.42) is higher than the F-crit (2.61) value, which indicates that using different regions as the network's input data significantly impacts the forecasting results.

6.3. Numerical Experiment and Analysis of Meteorological Factors

In this part, we analyze the correlation between the other three meteorological factors and PM2.5 in the adjacent area of the Haidian District, including temperature, wind direction, and humidity. The MI value and AID value between PM2.5 in the GY area and meteorological factors in Haidian District are shown in Table 4.

Table 4. Comparison of MI value and AID value between PM2.5 in the GY area and temperature, wind direction, and humidity in the Haidian District.

Variables	Temperature	Wind Direction	Humidity
MI	0.26	0.22	0.34
AID	1.70	1.64	1.69

It can be seen from Table 4 that the MI values between PM2.5 of GY area and meteorological factors in Haidian District are temperature (0.34), wind direction (0.26), and humidity (0.22), indicating that humidity has the highest correlation with PM2.5 in the GY area. Moreover, the AID of temperature is the highest, the AID of humidity is slightly lower than that of temperature, and the AID of wind direction is the lowest, indicating that both temperature and humidity have low redundancy with GY PM2.5. Therefore, we believe that using humidity as an auxiliary variable to forecast PM2.5 of GY will improve forecasting accuracy.

To verify our conclusion, the optimal results obtained in Table 4 are combined with each meteorological factor to forecast the future of PM2.5. The forecasting results are shown in Figure 7 and Table 4. Figure 7 shows the comparison results of the three evaluation indicators of (GY & STRR) PM2.5, combined with meteorological factors in different regions. It is clear from the figure that the RMSE, MSE, and MAE of the combined humidity have the slightest error from the actual values. From the results listed in Table 5, it can be seen that the forecasting errors obtained by adding two factors, temperature and wind direction, as auxiliary variables have improved to varying degrees; only the forecasting error of adding the humidity factor as an auxiliary variable has a significant decrease, and the RMSE has decreased from 26.15 to 25.44, a reduction of 2.7%; MSE decreased from 684.02 to 647.17, a decrease of 5.4%; MAE decreased from 18.77 to 17.95, a decrease of 4.4%. Besides, since the forecasting model in this experiment is the same, the training time using different regions as the input data of the network is all around 46 s, and there is no significant difference in the fluctuation of results.

Therefore, we used both PM2.5 and humidity in the STRR region as auxiliary variables to forecast PM2.5 concentrations in GY. At the same time, we also verified the validity of the method proposed in this study and proved that other air qualities and meteorological factors are closely related to the changes in PM2.5 concentration.

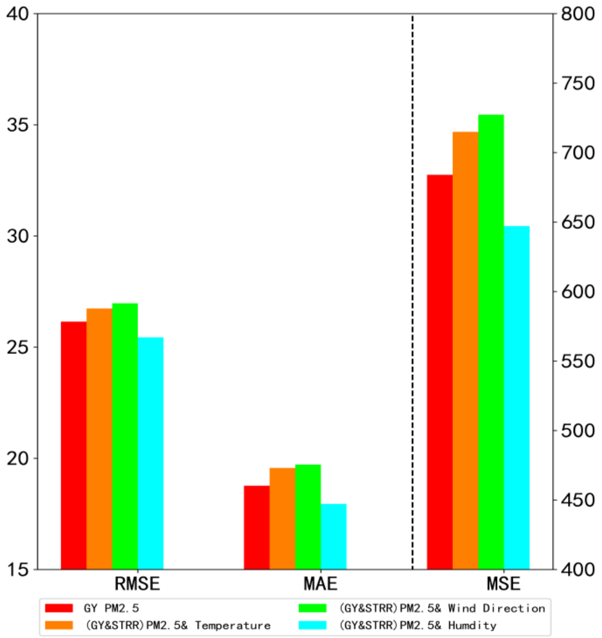


Figure 7. Two-coordinate error histogram of forecasting results based on different meteorological factors.

Table 5. Analysis of forecasting results of PM2.5 in space (GY & STRR) combined with historical data of meteorological factors in different regions as input data of the network.

Data	RMSE	MSE	MAE	Training Time (s)
(GY & STRR) PM2.5 (GY & STRR's PM2.5 data together as input)	26.15 ± 0.0011	684.02 ± 0.0702	18.77 ± 0.0015	44.81
(GY & STRR) PM2.5 & temperature (GY & STRR's PM2.5 data and temperature together as input)	26.74 ± 0.0015	714.94 ± 0.0757	19.57 ± 0.0022	45.62
(GY & STRR) PM2.5 & wind direction (GY & STRR's PM2.5 data and wind direction together as input)	26.97 ± 0.0033	727.31 ± 0.0815	19.72 ± 0.0063	47.91
(GY & STRR) PM2.5 & humidity (GY & STRR's PM2.5 data and humidity together as input)	25.44 ± 0.0021	647.17 ± 0.0603	17.95 ± 0.0027	45.21

We also used a one-way analysis of variance to validate the results. The result of the one-way analysis of variance is shown in Table 6.

Table 6. One-way analysis of variance on forecasting results based on the historical data of PM2.5 in space (GY & STRR), combined with historical data of meteorological factors in different regions as input data of the network. In this table, SS is the sum of squares and df is the degrees of freedom. The MS (mean squares) is the estimate of variance given by SS/df.

Source	SS	df	MS	F	F-Crit
Type	6934.63	3	2311.54	5.75	2.61
Error	8,445,274.19	21,020	401.77		
Total	8,452,208.81	21,023			

Table 6 shows that the value of F (5.75) is higher than the F-crit (2.61), which indicates that using different meteorological factors in different regions as the network's input data significantly impacts the forecasting results.

6.4. Interpretability Analysis

With the advent of big data, deep learning methods have successfully proven their superiority in processing big data. Still, for deep learning methods, it is not that the more data the model trained, the better the results obtained were. A large amount of data often has a lot of redundant information, which will not only increase the computational cost and reduce the convergence speed of the network but also affect the forecasting performance of the network.

This paper introduces AID based on MI to perform a nonlinear transformation on information and then change the metric value of MI. We aim to eliminate data with high redundancy while selecting highly correlated variables. Whether a different regional factor variable or a meteorological factor variable, this paper calculates the AID between each variable and PM2.5. On this basis, the information filtering method proposed in this paper not only improves the forecasting performance of the network but also has sufficient interpretability.

When analyzing the factors of different regions, we take the TOH region as an example. It can be seen from Table 1 that the MI values between PM2.5 in the TOH region and PM2.5 in the GY region are the highest at 0.71 compared with other regions. This shows that the PM2.5 in the TOH area has the highest correlation with the PM2.5 in the GY area. This is because this phenomenon occurs because the distance between the TOH and GY areas is the closest compared to other areas. Moreover, the AID values between PM2.5 in the TOH region and PM2.5 in the GY region are the lowest at 1.39 compared with other regions, indicating that there is significant redundancy between PM2.5 in the two regions. This is because PM2.5 particles in adjacent areas will undergo a process of dissipating with climatic conditions within a short time. Although there is a substantial time correlation between PM2.5 in adjacent areas, it also leads to a redundant relationship between PM2.5 in adjacent areas. When analyzing meteorological factors, we take humidity as an example. As shown in Table 3, the MI values between humidity and PM2.5 are the highest at 0.34 compared with other meteorological factors, indicating that humidity has the highest correlation with PM2.5. This is because when the humidity rises, the PM2.5 particles will absorb the water vapor in the air and settle after absorbing a certain amount of water vapor, thereby reducing the concentration of PM2.5. It can be seen that humidity will significantly impact the concentration of PM2.5. Besides, because PM2.5 has no inclusion relationship with humidity, there is no redundant information between the data. Therefore, it can also be seen from Table 3 that the AID values between humidity and PM2.5 are the second highest relative to other meteorological factors at 1.69.

This paper also used each variable as an auxiliary variable to forecast the future PM2.5, as shown in Tables 2 and 4. The results show that the variables selected by the information filtering module are critical for forecasting results. These key variables have a significant correlation and low redundancy with PM2.5. Their addition will significantly improve the forecasting performance of the network. It also shows that these variables play a vital role in promoting the network's ability to learn the changing pattern of PM2.5 better.

7. Conclusions and Future Work

In view of the problem that changes in weather data are affected by both temporal and spatial dimensions, as well as the challenge of redundant and conflicting information existing between multi-dimensional air quality data, it is difficult to forecast them accurately. In this paper, an interpretable variational Bayesian deep learning model with an information self-screening function is proposed to forecast the future PM2.5 concentration. The data self-screening layer's ability to screen variables with high correlation and low redundancy, and the anti-noise interference ability of the variational Bayesian gated cyclic unit, are fully utilized in this paper to improve the forecasting accuracy and robustness of future air quality. The validity of the method proposed is shown by the verification experiments of Beijing air quality data and meteorological data and by the careful consideration of RMSE, MSE, MAE, and other indicators.

In future research, we will try to use the method proposed in this paper on more air pollution data sets to test the applicability of this method. Additionally, we will continue to improve our network structure further to improve the overall performance of the forecasting method by means of some identification modeling methods [54–56], such as the multi-innovation theory [57,58] and the hierarchical principle [59–61].

Author Contributions: Conceptualization, X.-B.J.; methodology, Z.-Y.W.; software, Z.-Y.W.; validation, Z.-Y.W.; formal analysis, J.-L.K.; investigation, Y.-T.B.; resources, J.-L.K.; data curation, T.-L.S.; writing—original draft preparation, Z.-Y.W. and W.-T.G.; writing—review and editing, Z.-Y.W.; visualization, W.-T.G.; supervision, T.-L.S.; project administration, X.-B.J.; funding acquisition, H.-J.M. and P.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China, No. 62173007, 62006008, 61903009.

Data Availability Statement: Not applicable.

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

References

1. Jin, X.B.; Wang, Z.Y.; Kong, J.L.; Bai, Y.T.; Su, T.L.; Ma, H.J.; Chakrabarti, P. Deep Spatio-Temporal Graph Network with Self-Optimization for Air Quality Prediction. *Entropy* **2023**, *25*, 247. [CrossRef]
2. Menares, C.; Perez, P.; Parraguez, S. Forecasting PM2.5 levels in santiago de chile using deep learning neural networks. *Urban Clim.* **2021**, *38*, 100906. [CrossRef]
3. Li, J.; Li, X.; Wang, K. Atmospheric PM2.5 concentration prediction based on time series and interactive multiple model approach. *Adv. Meteorol.* **2019**, *2019*, 1279565. [CrossRef]
4. Lee, C.M.; Ko, C.N. Short-term load forecasting using lifting scheme and ARIMA models. *Expert Syst. Appl.* **2011**, *38*, 5902–5911. [CrossRef]
5. Ding, F.; Shi, Y.; Chen, T. Performance analysis of estimation algorithms of non-stationary ARMA processes. *IEEE Trans. Signal Process.* **2006**, *54*, 1041–1053. [CrossRef]
6. D'Amico, A.; Ciulla, G. An intelligent way to predict the building thermal needs: ANNs and optimization. *Expert Syst. Appl.* **2022**, *191*, 116293. [CrossRef]
7. Liu, W.; Liang, S.; Yu, Q. PM2.5 concentration prediction based on pseudo-F statistic feature selection algorithm and support vector regression. *IOP Conf. Ser. Earth Environ. Sci.* **2020**, *569*, 012039. [CrossRef]
8. Fang, Z.; Yang, H.; Li, C. Prediction of PM2.5 hourly concentrations in Beijing based on machine learning algorithm and ground-based LiDAR. *Arch. Environ. Prot.* **2021**, *47*, 98–107.
9. Li, M.H.; Liu, X.M. Maximum likelihood hierarchical least squares-based iterative identification for dual-rate stochastic systems. *Int. J. Adapt. Control Signal Process.* **2021**, *35*, 240–261. [CrossRef]
10. Li, M.H.; Liu, X.M. Iterative identification methods for a class of bilinear systems by using the particle filtering technique. *Int. J. Adapt. Control Signal Process.* **2021**, *35*, 2056–2074. [CrossRef]
11. Ding, F.; Chen, T. Combined parameter and output estimation of dual-rate systems using an auxiliary model. *Automatica* **2004**, *40*, 1739–1748. [CrossRef]
12. Ding, F. Coupled-least-squares identification for multivariable systems. *IET Control Theory Appl.* **2013**, *7*, 68–79. [CrossRef]
13. Xu, L. Separable Newton recursive estimation method through system responses based on dynamically discrete measurements with increasing data length. *Int. J. Control Autom. Syst.* **2022**, *20*, 432–443. [CrossRef]
14. Xu, L. Separable multi-innovation Newton iterative modeling algorithm for multi-frequency signals based on the sliding measurement window. *Circuits Syst. Signal Process.* **2022**, *41*, 805–830. [CrossRef]
15. Xu, L. Separable synchronous multi-innovation gradient-based iterative signal modeling from on-line measurements. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 6501313. [CrossRef]
16. Ding, F.; Liu, G.; Liu, X.P. Partially coupled stochastic gradient identification methods for non-uniformly sampled systems. *IEEE Trans. Autom. Control* **2010**, *55*, 1976–1981. [CrossRef]
17. Ding, F.; Chen, T. Parameter estimation of dual-rate stochastic systems by using an output error method. *IEEE Trans. Autom. Control* **2005**, *50*, 1436–1441. [CrossRef]
18. Connor, J.T.; Martin, R.D.; Atlas, L.E. Recurrent neural networks and robust time series prediction. *IEEE Trans. Neural Netw.* **1994**, *5*, 240–254. [CrossRef]
19. Tian, Y.; Zhang, K.; Li, J. LSTM-based traffic flow prediction with missing data. *Neurocomputing* **2018**, *318*, 297–305. [CrossRef]
20. Pan, C.; Tan, J.; Feng, D. Prediction intervals estimation of solar generation based on gated recurrent unit and kernel density estimation. *Neurocomputing* **2020**, *453*, 552–562. [CrossRef]
21. Wang, J.; Song, G. A Deep Spatial-temporal ensemble model for air quality prediction. *Neurocomputing* **2018**, *314*, 198–206. [CrossRef]

22. Wang, G.; Awad, O.I.; Liu, S. NOx emissions prediction based on mutual information and back propagation neural network using correlation quantitative analysis. *Energy* **2020**, *198*, 117286. [CrossRef]
23. Song, H.Y.; Park, S. An Analysis of correlation between personality and visiting place using spearman's rank correlation coefficient. *KSII Trans. Internet Inf. Syst. (TIIS)* **2020**, *14*, 1951–1966.
24. Jin, X.B.; Gong, W.T.; Kong, J.L.; Bai, Y.T.; Su, T.L. A variational Bayesian deep network with data self-screening layer for massive time-series data prediction. *Entropy* **2022**, *24*, 335. [CrossRef]
25. Wu, D.; Wang, X.; Wu, S. A hybrid method based on extreme learning machine and wavelet transform denoising for stock prediction. *Entropy* **2021**, *23*, 440. [CrossRef]
26. Ruiz-Aguilar, J.J.; Turias, I.; González-Enrique, J. A permutation entropy-based EMD–ANN forecasting ensemble approach for wind speed prediction. *Neural Comput. Appl.* **2021**, *33*, 2369–2391. [CrossRef]
27. Liu, T.; Alexis, K.H.; Lau, K.S. Time series forecasting of air quality based on regional numerical modeling in Hong Kong. *J. Geophys. Res. Atmos.* **2018**, *123*, 4175–4196. [CrossRef]
28. Zeng, Y.; Daniel, A.; Jaffe, X.Q. Prediction of potentially high PM2.5 concentrations in Chengdu, China. *Aerosol Air Qual. Res.* **2020**, *20*, 956–965. [CrossRef]
29. Wang, X.; Wang, B. Research on prediction of environmental aerosol and PM2.5 based on artificial neural network. *Neural Comput. Appl.* **2018**, *31*, 8217–8227. [CrossRef]
30. Chang, F.J.; Chang, L.C.; Kang, C.C. Explore spatio-temporal PM2.5 features in northern Taiwan using machine learning techniques. *Sci. Total Environ.* **2020**, *736*, 139656. [CrossRef]
31. Shahriar, S.A.; Kayes, I.; Hasan, K. Potential of ARIMA-ANN, ARIMA-SVM, DT and CatBoost for atmospheric PM2.5 forecasting in Bangladesh. *Atmosphere* **2021**, *12*, 100. [CrossRef]
32. Carreno, G.; Lopez-Cortes, X.A.; Marchant, C. Machine learning models to forecasting critical episodes of environmental pollution for PM2.5 and PM10 in Talca, Chile. *Mathematics* **2022**, *10*, 373. [CrossRef]
33. Sun, K.; Tang, L.; Qian, J.S. A deep learning-based pm2.5 concentration estimator. *Displays* **2021**, *69*, 102072. [CrossRef]
34. Shi, P.; Fang, X.; Ni, J. An Improved attention-based integrated deep neural network for PM2.5 concentration prediction. *Appl. Sci.* **2021**, *11*, 4001. [CrossRef]
35. Mengfan, T.; Siwei, L.; Lechao, D.; Senlin, H. Including the feature of appropriate adjacent sites improves the PM2.5 concentration prediction with long short-term memory neural network model. *Sustain. Cities Soc.* **2021**, *76*, 103427. [CrossRef]
36. Wang, W.; Mao, W.; Tong, X. A novel recursive model based on a convolutional long short-term memory neural network for air pollution prediction. *Remote Sens.* **2021**, *13*, 1284. [CrossRef]
37. Wang, B.; Kong, W.; Zhao, P. An air quality forecasting model based on improved convnet and RNN. *Soft Comput.* **2021**, *25*, 9209–9218. [CrossRef]
38. Prihatno, A.T.; Nurcahyanto, H.; Ahmed, M.F.; Rahman, M.H.; Alam, M.M.; Jang, Y.M. Forecasting PM2.5 concentration using a single-dense layer BiLSTM method. *Electronics* **2021**, *10*, 1808. [CrossRef]
39. Zhu, J.; Deng, F.; Zhao, J. Attention-based parallel networks (APNet) for PM2.5 spatiotemporal prediction. *Sci. Total Environ.* **2021**, *769*, 145082. [CrossRef]
40. Pak, U.; Ma, J.; Ryu, U. Deep learning-based PM2.5 prediction considering the spatiotemporal correlations: A case study of Beijing, China. *Sci. Total Environ.* **2020**, *699*, 133561. [CrossRef]
41. Cifuentes, F.; Gálvez, A.; González, C.M. Hourly ozone and PM2.5 prediction using meteorological data–alternatives for cities with limited pollutant information. *Aerosol Air Qual. Res.* **2021**, *21*, 200471. [CrossRef]
42. Zhu, M.; Xie, J. Investigation of nearby monitoring station for hourly PM2.5 forecasting using parallel multi-input 1D-CNN-biLSTM. *Expert Syst. Appl.* **2022**, *211*, 118707. [CrossRef]
43. Xing, G.; Zhao, E.; Zhang, C. A Decomposition-ensemble approach with denoising strategy for PM2.5 concentration forecasting. *Discret. Dyn. Nat. Soc.* **2021**, *2021*, 1–13. [CrossRef]
44. Jin, X.B.; Zhang, J.H.; Su, T.L. Modeling and analysis of data-driven systems through computational neuroscience wavelet-deep optimized model for nonlinear multicomponent data forecasting. *Comput. Intell. Neurosci.* **2021**, *2021*, 1–13. [CrossRef] [PubMed]
45. Samal, K.K.R.; Babu, K.S.; Das, S.K. Temporal convolutional denoising autoencoder network for air pollution prediction with missing values. *Urban Clim.* **2021**, *38*, 100872. [CrossRef]
46. Cai, J.; Dai, X.; Hong, L. An air quality prediction model based on a noise reduction self-coding deep network. *Math. Probl. Eng.* **2020**, *2020*, 3507197. [CrossRef]
47. Jin, X.-B.; Yang, N.-X.; Wang, X.-Y.; Bai, Y.-T.; Su, T.-L.; Kong, J.-L. Deep hybrid model based on EMD with classification by frequency characteristics for long-term air quality prediction. *Mathematics* **2020**, *8*, 214. [CrossRef]
48. Ding, F.; Ma, H.; Pan, J.; Yang, E.F. Hierarchical gradient- and least squares-based iterative algorithms for input nonlinear output-error systems using the key term separation. *J. Frankl. Inst.* **2021**, *358*, 5113–5135. [CrossRef]
49. Zhang, X. State estimation for bilinear systems through minimizing the covariance matrix of the state estimation errors. *Int. J. Adapt. Control Signal Process.* **2019**, *33*, 1157–1173. [CrossRef]
50. Chen, Z.; Chen, D.; Zhao, C. Influence of meteorological conditions on PM2.5 concentrations across China: A review of methodology and mechanism. *Environ. Int.* **2020**, *139*, 105558. [CrossRef]
51. Liao, T.; Wang, S.; Ai, J. Heavy pollution episodes, transport pathways and potential sources of PM2.5 during the winter of 2013 in Chengdu (China). *Sci. Total Environ.* **2017**, *584*, 1056–1065. [CrossRef] [PubMed]

52. Guo, L.C.; Zhang, Y.; Lin, H. The washout effects of rainfall on atmospheric particulate pollution in two Chinese cities. *Environ. Pollut.* **2016**, *215*, 195–202. [CrossRef] [PubMed]
53. Jin, X.B.; Zheng, W.Z.; Kong, J.L.; Wang, X.Y.; Bai, Y.T.; Su, T.L.; Lin, S. Deep-learning forecasting method for electric power load via attention-based encoder-decoder with Bayesian optimization. *Energies* **2021**, *14*, 1596. [CrossRef]
54. Xu, L.; Yang, E.F. Auxiliary model multiinnovation stochastic gradient parameter estimation methods for nonlinear sandwich systems. *Int. J. Robust Nonlinear Control* **2021**, *31*, 148–165. [CrossRef]
55. Wan, L.J. Decomposition- and gradient-based iterative identification algorithms for multivariable systems using the multi-innovation theory. *Circuits Syst. Signal Process.* **2019**, *38*, 2971–2991. [CrossRef]
56. Xu, L.; Song, G.L. A recursive parameter estimation algorithm for modeling signals with multi-frequencies. *Circuits Syst. Signal Process.* **2020**, *39*, 4198–4224. [CrossRef]
57. Ding, F.; Chen, T. Performance analysis of multi-innovation gradient type identification methods. *Automatica* **2007**, *43*, 1–14. [CrossRef]
58. Ding, F.; Liu, X.; Liu, G. Multiinnovation least squares identification for linear and pseudo-linear regression models. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **2010**, *40*, 767–778. [CrossRef]
59. Zhang, X. Hierarchical parameter and state estimation for bilinear systems. *Int. J. Syst. Sci.* **2020**, *1*, 275–290. [CrossRef]
60. Liu, S.Y.; Hayat, T. Hierarchical principle-based iterative parameter estimation algorithm for dual-frequency signals. *Circuits Syst. Signal Process.* **2019**, *38*, 3251–3268. [CrossRef]
61. Xu, L.; Hayat, T. Hierarchical recursive signal modeling for multi-frequency signals based on discrete measured data. *Int. J. Adapt. Control Signal Process.* **2021**, *35*, 676–693. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.