This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2963416, IEEE Access

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2019.Doi Number

An Adaptive Kalman Filtering Approach to Sensing and Predicting Air Quality Index Values

Jibo Chen[1], Keyao Chen[2], Chen Ding[3], Guizhi Wang[3], Qi Liu[4], and Xiaodong Liu[5]

[1]Binjiang College, Nanjing University of Information Science & Technology, Wuxi 214105, China
[2]National Climate Center, China Meteorological Administration, Beijing 100081, China
[3]School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing 210044, China
[4]Shandong Beiming Medical Technology Ltd, Jinan 250000, China
[5]School of Computing, Edinburgh Napier University, 10 Colinton Road, Edinburgh, EH10 5DT, UK

Corresponding author: Keyao Chen (chenky@cma.gov.cn).

This work was supported by Major Program of the National Social Science Fund of China (Grant No. 17ZDA092).

ABSTRACT In recent years, Air Quality Index (AQI) have been widely used to describe the severity of haze and other air pollutions yet suffers from inefficiency and compatibility on real-time perception and prediction. In this paper, an Auto-Regressive (AR) prediction model based on sensed AQI values is proposed, where an adaptive Kalman Filtering (KF) approach is fitted to achieve efficient prediction of the AQI values. The AQI values were collected monthly from January 2018 to March 2019 using a WSN-based network, whereas daily AQI values started to be collected from October 1, 2018 to March 31, 2019. These data have been used for creation and evaluation purposes on the prediction model. According to the results, predicted values have shown high accuracy compared with the actual sensed values. In addition, when monthly AQI values were used, it has depicted higher accuracy compared to the daily ones depending on the experimental results. Therefore, the hybrid AR-KF model is accurate and effective in predicting haze weather, which has practical significance and potential value.

INDEX TERMS Real-time sensing and predicting, Kalman Filter, Air Quality Index, Simulation.

I. INTRODUCTION

In recent years, haze pollution has raised great concern in worldwide societies and scientific communities, due to its influencing living environment of human beings, even as potential impedance of the social progress from the world economic development perspective. The main causes of the haze pollution are the emission of exhaust gas from industrial production, smoke and dust from coal combustion, waste gas from vehicles, and dust from construction sites [1]-[2]. Different data retrieval methods have been used from historical monitoring results to WSN-based collection [3], as well as its optimization [4]-[5]. In this case, if the relative humidity is high and the air flow is relatively slow, the air will easily saturate and condense to form haze through the cooling of atmospheric radiation. Haze affects people's life in many aspects. First, it has direct harm to public health. For example, it causes rise of respiration, cardiovascular and cerebrovascular diseases. Meanwhile, it can lead to huge direct and indirect losses to the social economy. In severe haze pollutions, public and private transportation can be affected due to the reduction of visibility. Therefore, measuring, monitoring, and predicting air quality and haze pollutions become critical in order to achieve eventual reduction of haze risks in practical life.

Although the cause of air pollution is complex and stochastic, this does not mean that the prediction of the air pollution cannot be done. In recent years, many researchers and scholars have shown high concern in the analysis and the prediction of the Air Quality Index (AQI). Such work can be concluded into two groups including deterministic approaches and statistical approaches. The former approaches focus on the physical theory in atmosphere and meteorological processes with concern on high-volume historical data, so diffusion models of the atmospheric pollution were generally presented by using specific mathematical approaches. Chen *et al.* [6] simulated the $PM_{2.5}$ formation and emission based on Community Multi-scale Air Quality (CMAQ) model. Saide *et al.* [7] proposed a WRF–

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2963416, IEEE Access

Chem model with optimal parameters. On this basis, a forecasting system was developed in order to describe air quality and meteorological measurements. However, these proposed models usually require a large number of historical data in meteorological aspects, which are difficult to obtain for researchers in practice. Moreover, limited knowledge of pollutants evolution processes and experience of parameter selection would affect forecasting accuracy.

On the other hand, statistical approaches for prediction have been widely adopted recently due to their flexibility and simplicity. The well-used statistical models include AutoRegressive Integrated Moving Average (ARIMA), Grey Models (GM), Support Vector Regression (SVR), Multiple Linear Regression (MLR), Artificial Neural Networks (ANN), and other hybrid models. For instance, Yang *et al.* [8] presented the formation cause analysis of haze by time series methods, then a vector autoregressive model was constructed to predict daily haze increment. The results have showed good stability in short-term prediction. Carbajal *et al.* [9] introduced a fuzzy system to classify parameters, and then proposed an autoregressive model to predict the AQI based on the system. Combarro [10] employed the SVR method to determine the elements which had the greatest impact on the air quality in Oviedo city in Spain. Wu *et al.* [11] predicted the annual average concentration of $PM_{2.5}$ in the three different regions of China in 2020 based on the fractional order accumulation grey model called FGM(1, 1). The results showed that its forecasting performance was better than traditional grey model. Challoner *et al.* [12] used two different models to predict air quality indices. A personal exposure activity location model was used to predict the outdoor air quality of a specific building, while an artificial neural network model was used to predict the indoor air quality. The above models were combined to fit the relationship between indoor air and outdoor air of the building. Liu *et al.* [13] promoted a seq2seq model to predict air quality with historical air quality data and introduced n-step recurrent prediction to solve error problems. Bai *et al.* [14] established a W-BPNN model by using a wavelet technique and a backpropagation neural network (BPNN) to predict daily air pollutants concentration. The results showed that the prediction accuracy of hybrid model was better than that of BPNN model. Wu *et al.* [15]-[16] proposed an optimal hybrid model, which combined secondary decomposition, neural network and optimization algorithms to predict air quality index. All of the solutions above focused on the prediction of the AQI with limited introduction of their data sources. Instead, how to retrieve real-time data via a WSN-based network becomes challenging. However, due to the uncertainty and diffusion of air pollution, some individual statistical models tended to introduce biases for air quality prediction. Meanwhile, hybrid models obtained better forecasting results to some extent. Furthermore, a Kalman Filter (KF) approach can strengthen the ability of dealing with stochastic uncertainty combined with its state-space equation. Therefore, a hybrid model applying the KF approach to a statistical model is proposed in this study.

The remainder of this paper is organized as follows. Section 2 mainly presents related works using Kalman filtering models. Section 3 describes detailed data collecting, processing and modeling processes proposed in this paper. The experimental results of the Kalman filtering approach is depicted in Section 4. Section 5 discusses the Kalman filtering method for prediction. Finally, the last section makes a conclusion of this paper.

II. Related Works

In 1960, Kalman [17] proposed a state-space model into the filtering method and derived a set of recursive estimation equation called simple Kalman filters. With the popularization and improvement of the Kalman filtering model, it has been widely applied in different fields, such as hydrology, physics, mechanical control and economy.

A Seasonal Autoregressive Integrated Moving Average (SARIMA) with a Generalized Autoregressive Conditional Heteroscedasticity (GARCH) approach was proposed in order to predict traffic flow [18]. On this basis, an adaptive Kalman filter was used to realize the proposed model to improve the forecasting performance. Hua *et al.* [19] used a Weather Research and Forecast (WRF) model to compare the observed wind speed with the predicted wind speed, and then revised the predicted wind speed on the basis of the Kalman filter theory in order to reduce systematic and random errors. Finally, the forecasting accuracy has been well improved. An unscented Kalman Filter (UKF) approach with support vector regression (SVR) was adopted to conduct the short-term prediction of wind speed. Meanwhile, compared with four different models, the hybrid UKF-SVR model achieved better forecasting performance [20]. Lai *et al.* [21] proposed a Kalman Filtering algorithm to predict six kinds of different air pollutants compared with common forecasting models. The results demonstrated that the KF model could obtain the optimal prediction results. Galanis *et al.* [22] improved the prediction performance of regional weather by applying a nonlinear function to the classical Kalman filter algorithm. Chaabene and Ammar [23] proposed an autoregressive moving average model for medium-term forecasting based on a Kalman filter, yielding higher accuracy compared with the short-term weather forecasting. Kumar [24] introduced a Kalman filtering technique (KFT) to predict traffic flow with limited input data. The result proved the suitability of the presented prediction without enough data. Xing *et al.* [25] proposed a temperature model to construct a state of charge (SOC) estimation method. An Unscented Kalman Filter approach was used to deal with various uncertainties, such as environment variation, intercellular variation and modeling inaccuracy by adjusting the model parameters in each sampling step. Mastali *et al.* [26] employed an Extended Kalman filter (EKF) to predict the state of the batteries. On

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2963416, IEEE Access

this basis, a dual concept was introduced in the extended Kalman filter model in order to improve the forecasting accuracy. According to the results, the filters have kept small maximum errors indicating that the validity of Kalman filter. Soubdhan *et al.* [27] proposed the framework of a linear dynamic Kalman filter for predicting solar and photovoltaic production including probabilistic initialization, expectation maximization (EM) and auto regressive (AR) models. Two common Kalman filtering methods including EKF and UKF have been employed to fuse the pseudo-range, ranging information and location information for indoor localization, and the experiment results proved that the positioning performance of the nodes have improved obviously combined with Kalman Filter approach [28]. Rigatos and Siano [29] used a nonlinear Kalman filter to predict the default probability of financial companies and estimated the default risk by predicting the ratio of option to asset values.

Aimed at evaluating the level of air pollution, AQI has been chosen as an effective index to measure the comprehensive level of air quality, to which the Chinese environment minister has paid great attention in recent years. Moreover, the Kalman filter approach has been gradually applied to the fields of economy and finance, but seldomly used in the field of meteorology. In this paper, AQI data have been analyzed, evaluated and predicted by using the Kalman filter approach in order to accurately predict the air quality in the near future.

III. Data processing and model selection

In this section, three types of time-series data processing methods will be firstly analyzed and compared, i.e. Artificial Neural Networks (ANN), Wavelet Transform (WT) and Kalman Filter (KF). The ANN method [30]-[33] has been well applied in the field of image and voice processing, e.g. pattern detection and recognition. It has strong generalization and fault-tolerant features for the description of nonlinear systems. However, in regard to a large amount of system noises, an ANN model can fall into a local minimum value, resulting in serious prediction errors to a certain extent. The WT method [34]-[36] is a powerful tool for non-stationary signal processing. Due to the variation of signal characteristics in different scales, it is difficult for wavelet functions to be derived from a specific basis function, in order to achieve proper approximation on local signals in different scales. Therefore, reconstructed signals can lose the original time domain during de-noising processes. The Kalman filter method [37]-[38] updates and processes real-time data through an accurate mathematical model, which is convenient to be programed to realize the prediction efficiently. The state space model of the Kalman filter can estimate current time state by using the estimated values of previous time steps and the observed value of current time step, so the state estimation can achieve high accuracy and is therefore suitable for linear discrete finite-dimensional systems due to its strong ability of handling the stochastic uncertainty.

Considering the stationary characteristics of an AQI dataset, an autoregressive linear prediction model is firstly established in the following part of this section. Then, based on the recurrent relationship between the front and back terms of the AR model, the AQI data can be corrected by a Kalman filter approach. The historical AQI values collected from January 2014 to September 2018 via a WSN-based network were used as training data, and 182 data from October 2018 to March 2019 were used as test data to evaluate the prediction performance of the model. In order to compare the forecasting performance of different time scales including daily data and monthly mean data. Similarly, 48 AQI monthly data from January 2014 to December 2017 have been employed as training data to predict near future monthly mean values and the AQI data from January 2018 to March 2019 were correspondingly used as test data.

A. Brief Introduction to the Autoregressive Model

The autoregressive (AR) model [39] is a linear model which uses the linear combination of initial random variables to describe current random variables. As a popular linear regression model, it is used to fit stationary time series, which has been applied in the prediction of economics, informatics and natural phenomena in recent years.

Let a time series be $x(1), x(2), \cdots, x(t)$, and the predicted values of the $t+1$ time series has the following structure:

$$\begin{cases} x(t+1) = \varphi_1 x(t) + \varphi_2 x(t-1) + \cdots + \varphi_p x(t-p+1) + \varepsilon(t+1) \\ \varphi_p \neq 0 \\ E(\varepsilon(t)) = 0, Var(\varepsilon(t)) = \sigma_\varepsilon^2 \\ E(x(s)\varepsilon(t)) = 0, \forall s < t \end{cases} \quad (1)$$

where $1 \leq p \leq t$ The model is called a p-order AR model, denoted as AR (p), where φ represents model parameters, p represents the highest order number of the model $\varepsilon(t)$ is a zero-mean white noise random disturbance sequence. σ_ε^2 is the variance of $\varepsilon(t)$. The current random disturbance term is independent from the past sequence values. Generally, the model can be simplified as:

$$x(t+1) = \varphi_1(t) + \varphi_2 x(t-1) + \cdots + \varphi_p x(t-p+1) + \varepsilon(t). \quad (2)$$

B. Brief Introduction to the Kalman Filter

The Kalman Filter model (KF) was introduced to improve accuracy of AQI forecasting in this paper combined with its strong capacity of handling stochastic uncertainty. The Kalman filter approach as a statistical approach was proposed in 1960 for the first time. Then it has been applied in different fields especially in meteorological applications because of its good prediction performance. The Kalman filter can calculate the optimal estimation parameters by the

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2963416, IEEE Access

minimum mean square error (MSE) for many problems, which has high efficiency.

For general linear stochastic systems without an input parameter, the state space representation equation is given:

$$X_{t+1} = AX_t + W_t, \tag{3}$$

$$Z_{t+1} = BX_t + V_t, \tag{4}$$

where $t \geq 2$. Equation (3) is the state equation of the system. Equation (4) is the measurement equation of the system. X_{t+1} is an *n*-dimensional state vector at time $t+1$. A is the state transition matrix; W_t is the process noise vector of *p*-dimensional system. Z_{t+1} is an m-dimensional observation vector at time $t+1$. B is the predicted output transfer matrix. V_t is a q-dimensional observation noise vector. Let W_t and V_t be white noises, which are independent of each other and obey normal distribution. Q_t represents the covariance matrix of a process noise vector and R_t denotes the covariance matrix of an observation noise vector.

In this paper, the AR model is introduced into the state equation of the Kalman filter to simplify the corresponding processes.

Let $x_1(t) = x(t), x_2(t) = x(t-1), \cdots, x_p(t) = x(t-p+1)$, thus the AR model can be expressed as:

$$x_1(t+1) = \varphi_1 x_1(t) + \varphi_2 x_2(t) + \cdots + \varphi_p x_p(t) + \varepsilon(t+1). \tag{5}$$

According to (5),

$$x_2(t+1) = x_1(t), x_3(t+1) = x_2(t), \cdots, x_p(t+1) = x_p(t). \tag{6}$$

Therefore, the vector form of state equation of the AR-KF model can be written as follows:

$$\begin{bmatrix} x_1(t+1) \\ x_2(t+1) \\ x_3(t+1) \\ \vdots \\ x_p(t+1) \end{bmatrix} = \begin{bmatrix} \varphi_1(t) & \varphi_2(t) & \cdots & \varphi_{p-1}(t) & \varphi_p(t) \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ \vdots \\ x_p(t) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \varepsilon(t+1), \tag{7}$$

According to (2) to (7), the observation equation based on the Kalman filter can be obtained as follows:

$$Z(t+1) = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} \times \begin{bmatrix} x_1(t+1) & x_2(t+1) & \cdots & x_p(t+1) \end{bmatrix}^T, \tag{8}$$

$$\hat{X}_{t|t} = \hat{X}_{t|t-1} + G_t\left(Z_t - B_t \hat{X}_{t|t-1}\right), \tag{9}$$

$$G_t = P_{t,t-1} B_t^T \left(B_t P_{t,t-1} B_t^T + R_t\right)^{-1}, \tag{10}$$

$$P_{t,t-1} = A_{t,t-1} P_{t-1,t-1} A_{t,t-1}^T + \Gamma_{t,t-1} Q_{t-1} \Gamma^T{}_{t,t-1}, \tag{11}$$

$$P_{t,t} = \left(I - G_t B_t\right) P_{t,t-1}, \tag{12}$$

where $\hat{X}_{t|t-1}$ is the state estimation at time t under the condition of at time t-1. $\hat{X}_{t|t}$ is the optimal state estimation at time t after considering $\hat{X}_{t|t-1}$. Z_t represents the observation vector, G_t is the Kalman gain and P is the error covariance matrix.

This algorithm is implemented according to Table 1.

TABLE 1. The Adapted Kalman Filter Algorithm.

Input: X_{t-1}, Z_t

Output: X_t

1	$\bar{X}_t = A_t X_{t-1}$
2	$\bar{P}_t = A_t P_{t-1} A_t^T + Q_t$
3	$G_t = \bar{P}_t B_t^T (B_t \bar{P}_t B_t^T + R_t)^{-1}$
4	$X_t = \bar{X}_t + G_t (Z_t - B_t \bar{X}_t)$
5	$P_t = (I - G_t B_t) \bar{P}_t$
6	**return** X_t, P_t

IV. Empirical Analysis

According to the Ambient Air Quality Index (AQI) Technical Regulations (HJ 633-2012). the AQI can be classified into the following six grades:

TABLE 2. AQI Values and Air Quality Classification Table.

AQI value	Air quality level
0-50	Excellent
51-100	Good
101-150	Light pollution
151-200	Moderate pollution
201-300	Heavy pollution
>300	Serious pollution

This paper takes the historical data of the AQI as the reference data of haze concentration and combines the Kalman filter approach with autoregressive (AR) model to predict AQI values.

A. Data Source and Processing

In this paper, 1826 historical data of AQI concentrations have been employed from January 1, 2014 to December 30, 2018 in Nanjing. The data can be obtained from the website of Weather Post-report (http://www.tianqihoubao.com/).

The AQI data for establishing the model are from January 1, 2014 to December 31, 2018.The time series diagram of AQI data in Nanjing can be shown in Figure 1.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2963416, IEEE Access

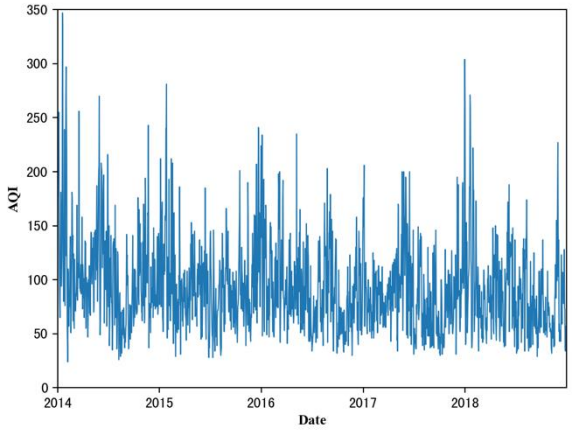


FIGURE 1. **AQI Daily Value Chart.**

In Fig. 1, it can be found that the fluctuation of AQI data in Nanjing is relatively stable. The range of AQI values is from 40 to 300. According to the applicable conditions of the AR model, stationary time-series data need to be used to train the proposed model. In order to test the stationarity of the time series data theoretically, the Augmented Dickey-Fuller (ADF) technique as a unit root method has been used. If there is no unit root in the time series data, the data are stationary; otherwise, it is a non-stationary series. Then in the former case, the p value can be calculated to be $p = 1.1157 \times 10^{-11}$, which is less than 0.05, i.e. no unit root in the sequence. Therefore, the daily AQI data is stationary which can be used as input data in the AR model.

Then, the order number of the AR model needs to be determined. Because the tailing and truncation of autocorrelation and partial autocorrelation can be performed depending on subjective operations. To improve the objectivity of this part, a Bayesian Information Criterion (BIC) approach is adopted to select the order p of the model by setting the upper and lower bounds and then traversing them one by one.

Define

$$BIC(p) = p\ln(N) - 2\ln(L), \tag{13}$$

where N represents the number of the samples, p denotes the order number of the model parameters, L represents the likelihood function.

When the BIC reaches the minimum value, the p value is chosen as the order of the optimal AR model under the criterion. According to programming experiments, the parameter has been optimally set to $p = 3$.

Therefore, the AR (3) model can be established as follows:

$$x_t = \varphi_0 + \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \varphi_3 x_{t-3}. \tag{14}$$

Next, the parameters of the AR model can be estimated by the least square method, which can be obtained by programming experiments, as shown in Table 3.

TABLE 3. **Parameter values of the model.**

φ_0	φ_1	φ_2	φ_3
93.4757	1.5511	-0.6928	0.1268

Then the three-order autoregressive model has been established by using 1734 AQI data from January 1, 2014 to September 30, 2018, and the model expression is shown below with 4 significant digits retained:

According to (3):

$$x_t = 93.4757 + 1.5511_1 x_{t-1} - 0.6928_2 x_{t-2} + 0.126 x_{t-3}, \tag{15}$$

where $t > 3$. This equation is the measurement equation of the Kalman filtering process. According to the (7), the time state transition matrix $\boldsymbol{A}$ of the Kalman filter is

$$\begin{bmatrix} 1.5511 & -0.6928 & 0.1268 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

1734 daily AQI samples from January 1, 2014 to September 30, 2018 in Nanjing were simulated by the Kalman filter and AR model, and the simulation results are shown in Figure 2. The abscissa represents the observation time and the ordinate indicates the AQI value. The red line represents the forecasting values by the hybrid KF-AR model, the blue line represents the AR forecasting values and the green line represents the true values.

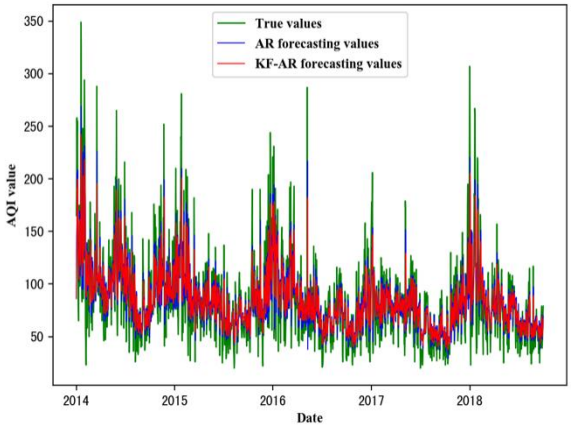


FIGURE 2. **AQI Daily Value Training Data Simulation Chart.**

As can be intuitively seen in Figure 2, the AQI values of December and January of each year from 2014 to 2018 are higher than that of other months. The AQI values of Nanjing are high in every winter and the trend is consistent with the regularly occurring haze in people's life, which shows the haze pollution will be more serious in winter than in other time; while the haze pollution concentration will reduce in summer to some extent, indicating that the

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2963416, IEEE Access

model is reasonable. In recent years, except for the sudden increase of the AQI value in the winter in 2018, the peak value curve of Nanjing in winter shows a decreasing trend. Such phenomenon should be closely related to the environmental air control taken by the Nanjing Municipal Government in recent years.

Since the fluctuation trend of the simulated value curve and the real value curve of the AQI daily values is intuitively consistent (Figure 2), 182 data of Nanjing from October 1, 2018 to March 31, 2019 have been used as test samples to fit into the model, in order to further analyze the prediction ability of the model through the test results. The simulation result is shown in Figure 3.

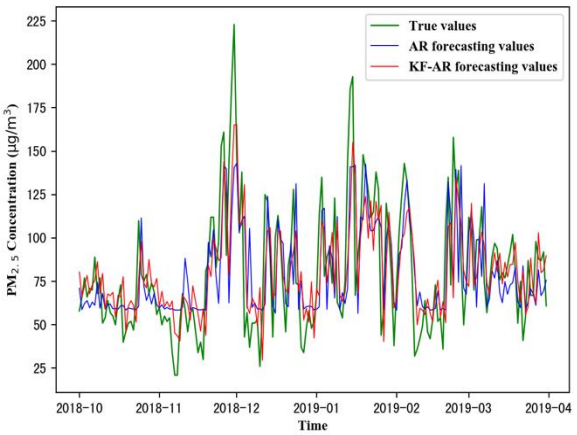


FIGURE 3. **AQI Daily Value Test Data Simulation Chart.**

The test results have shown that the curves of the test values by KF-AR model and the real values have good consistency (Figure 3), and both the two curves are in a downward trend. The errors from the hybrid model are mainly scattered in extreme values compared with the AR model, which are often caused by abnormal factors, such as temperature inversion, rain, wind, coals burning, and automobile exhaust. Specifically, the curve appears a bit smooth from the AR model at the beginning compared with the KF-AR model, while the KF-AR model fits well over the entire period. Therefore, it is difficult to fit well by the AR model in the whole periods. Compared with the real values and the predicted values, the root mean square error from October 1, 2018 to March 31, 2019 is 26.27 according to the KF-AR model, while the root mean square error of the AR model is 27.07, which is higher than that of the KF-AR model.

B. Modeling Process and Result Analysis of AQI Monthly Mean Sequence

Based on the previous analysis, the daily data are processed into monthly data to calculate the monthly AQI mean value series. The monthly mean values of 48 months from January 2014 to December 2017 have been collected to perform the model training experiment, whereas the monthly mean values of 15 months from January 2018 to March 2019 were selected to conduct the test. Based on the previous BIC criterion, the order of the AR model is set as $p = 4$. A four-order autoregressive model has been established for the 48 monthly AQI mean data from January 2014 to December 2017. According to the (2), it can be refined as follows with 4 significant digits being retained:

$$x_t = 0.4568_1 x_{t-1} - 0.0661_2 x_{t-2} - 0.14606 x_{t-3} + 0.194 x_{t-4}. \quad (16)$$

The state transition matrix $\boldsymbol{A}$ is determined:

$$\begin{bmatrix} 0.4568 & -0.0661 & -0.1460 & 0.1941 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

48 monthly mean value of the AQI data in Nanjing from January 2014 to December 2017 have been used as training data. The simulation results of the 48 samples in the first three years have been obtained, as shown in Figure 4.

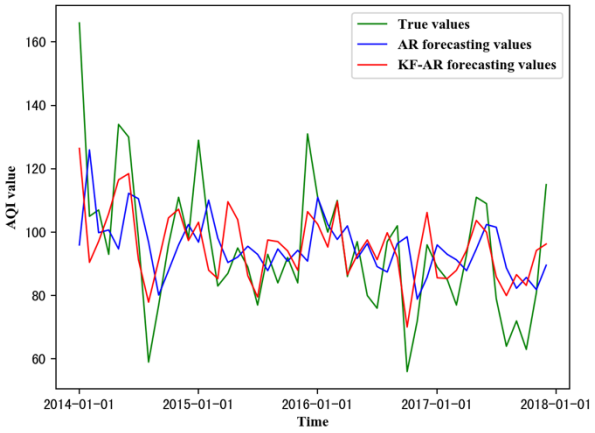


FIGURE 4. **AQI Monthly Mean Value Training Data Simulation Chart.**

As indicated in Figure 4, the monthly mean value series reach peak around the January of each year which shows the same trend as the daily value. The curve of real values has fluctuation since 2014, but generally depicts a slow downward trend and the air quality level is “Good”. The main reason is that the government has strengthened the pollution control and reduced emissions of air pollution in recent years.

Then, 15 monthly mean value data of AQI from January 2018 to March 2019 in Nanjing were used as test samples to continue fitting the series data with the model. The prediction ability of the model for the monthly mean series was further discussed according to the test results, as shown in Figure 5.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2963416, IEEE Access

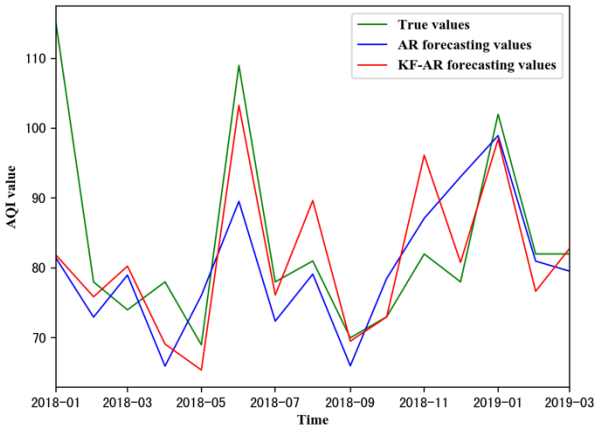


FIGURE 5. **AQI Monthly Mean Value Test Data Simulation Chart.**

According to Fig. 5, although there are some deviations in certain months, the trend of the real value curve of the AQI monthly mean data is consistent with the general trend of the predicted value curve compared with the AR model, which implies that the prediction effect is favorable. Specifically, in some time periods when AQI fluctuates greatly, the non-linear prediction performance based on the hybrid model outweighs the AR model. Also, in recent months, the prediction results of air quality show that the grade level is good. At the same time, the educational level of the society is increasing year by year, so the environmental protection awareness is gradually enhanced, and the air quality in the future is expected to be gradually improved. According to the difference between the real values and the test values, the root mean square error of two models are 10.37 and 11.84, respectively, which is less than the daily data prediction results. Moreover, the forecasting accuracy of the KF-AR model is less than that of the AR model. Table 4 gives the results of all experimental simulations, which are consistent with previous analysis. This phenomenon shows the forecasting performance of the hybrid model outperforms the AR model. Plus, the accuracy of monthly mean data prediction is better than the daily one.

TABLE 4. **Results of experimental simulation.**

RMSE	Daily Data	Monthly Mean Data
AR	28.07	11.84
KF-AR	26.27	10.37

V. Discussion

In this paper, an adaptive Kalman filtering model is introduced to improve the prediction accuracy of air quality based on the autoregressive model. The proposed methods can provide accurate data support for haze prevention and control. The results have shown that the Kalman filter based on the AR model can be well fitted into the AQI series data retrieved in Nanjing compared with individual model. Furthermore, this method can be extended to other fields for air quality prediction, such as $PM_{2.5}$, PM_{10}, SO_2, etc. It can also be combined with other prediction models, such as Support Vector Machine (SVM) and Artificial Neural Network (ANN) models to realize the hybrid prediction of the AQI in future.

VI. CONCLUSION

An adaptive Kalman filter approach based on the AR model has been proposed by training, testing and predicting the daily and monthly AQI series data in Nanjing. The paper finds that the model is more effective in predicting the monthly AQI series data collected via a WSN in Nanjing than in predicting daily AQI series data. Moreover, the forecasting performance of the hybrid KF-AR model is better than that of the individual AR model. Then a conclusion can therefore be drawn as follows:

(1) The prediction method based on the Kalman filter in this paper can predict the monthly mean value of AQI, which shows that the method depicts good prediction ability for AQI prediction and practical significance in the field of haze prediction.

(2) According to the training and test results, the AQI is decreasing in recent years. This not only means that haze prevention and control in Nanjing has achieved obvious effects but reflects the gradual improvement of people's environmental and ecological protection awareness as well.

(3) By observing the prediction curve of the AQI monthly mean series data, it is found that there is a certain delay error in the time. Therefore, the Kalman filtering model can be further improved on the AQI prediction in the future, by considering the integration of other regression methods to improve the Kalman filter and correct the delay error of the predicted values.

REFERENCES

[1] L. Miller, and X. Xu, "Ambient $PM_{2.5}$ human health effects—findings in China and research directions," *Atmosphere*, vol. 9, no. 11, pp. 424, Oct. 2018.

[2] Y. Hao, and Y. M. Liu, "The influential factors of urban $PM_{2.5}$ concentrations in China: a spatial econometric analysis," *J. Clean Prod.*, vol. 112, pp. 1443-1453, Jan. 2016.

[3] J. Chen, Y. Song, and G. Wang, "WSN-aided haze pollution governance: modelling public willingness based on structural equations," *Int. J. Sensor Networks*, vol. 29, no. 2, pp. 111-120, Feb. 2019.

[4] T. Qiu *et al.*, "Robustness optimization scheme with multi-population co-evolution for scale-free wireless sensor networks," *IEEE-ACM Trans. Netw.*, vol. 27, no. 3, pp. 1028-1042, Jun. 2019.

[5] T. Qiu, *et al.*, "TOSG: A topology optimization scheme with global-small-world for industrial heterogeneous Internet of Things," *IEEE Trans. Ind. Inform.*, vol. 15, no. 6, pp. 3174-3184, Jun. 2019.

[6] J. Chen, *et al.*, "Seasonal modeling of $PM_{2.5}$ in California's San Joaquin Valley," *Atmos. Environ.*, vol. 92, pp. 182 – 190, Aug. 2014.

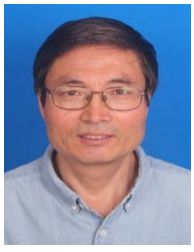
[7] P.E. Saide, *et al.*, "Forecasting urban PM_{10} and $PM_{2.5}$ pollution episodes in very stable nocturnal conditions and complex terrain using WRF–Chem CO tracer model," *Atmos. Environ.*, vol. 45, no. 16, pp. 2769-2780, May. 2011.

[8] X. P. Yang, et al., "A long-term prediction model of Beijing Haze episodes using time series analysis," *Comput. Intell. Neurosci.*, vol. 2016, pp. 1-7, Jul. 2016.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2963416, IEEE Access

[9] J. J. Carbajal, et al., "Assessment and prediction of air quality using fuzzy logic and autoregressive models". *Atmos. Environ.*, vol. 60, pp. 37-50, Dec. 2012.
[10] E. F. Combarro, "A SVM-based regression model to study the air quality at local scale in Oviedo urban area (Northern Spain): a case study," *Appl. Math. Comput.*, vol. 219, no. 17, pp. 8923–8937, May. 2013.
[11] L. Wu, and H. Zhao, "Using FGM (1, 1) model to predict the number of the lightly polluted day in Jing-Jin-Ji region of China," *Atmos. Pollut. Res.*, vol. 10, no. 2, pp. 552-555, Mar. 2019.
[12] A. Challoner, F. Pilla and L. Gill, "Prediction of indoor air exposure from outdoor air quality using an artificial neural network model for inner city commercial buildings," *Int. J. Environ. Res. Public Health*, vol. 12, no. 12, pp. 15233-15253, Dec. 2015.
[13] B. Liu *et al*., "A sequence-to-sequence air quality predictor based on the n-step recurrent prediction," *IEEE Access*, vol. 7, pp. 43331-43345, Mar. 2019.
[14] Y. Bai *et al.*, "Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions," *Atmos. Pollut. Res.*, vol. 7, no. 3, pp. 557-566, May 2016.
[15] Q. L. Wu, and H. X. Lin, "A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors," *Sci. Total Environ.*, vol. 683, pp. 808-821, Sep. 2019.
[16] Y. Zhan *et al*., "Spatiotemporal prediction of continuous daily $PM_{2.5}$ concentrations across China using a spatially explicit machine learning algorithm," *Atmos. Environ.*, vol. 155, pp. 129-139, Apr. 2017.
[17] R. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Engineering*., vol. 4, pp. 35-45, 1960.
[18] J. Guo, W. Huang and B. M. Williams, "Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification," *Transp. Res. Part C Emerg. Technol.*, vol. 43, pp. 50-64, Jun. 2014.
[19] S. Hua *et al.*, "Wind speed optimisation method of numerical prediction for wind farm based on Kalman filter method," *J. Eng.*, vol. 2017, no. 13, pp. 1146-1149, 2017.
[20] K. Chen, and J. Yu, "Short-term wind speed prediction using an unscented Kalman filter based state-space support vector regression approach," *Appl. Energy*, vol. 113, pp. 690-705, Jan. 2014.
[21] X. Lai, *et al*., "IoT Implementation of Kalman Filter to Improve Accuracy of Air Quality Monitoring and Prediction," *Applied Sciences*, vol. 9, no. 9, pp. 1831, May 2019.
[22] G. Galanis, P. Louka, P. Katsafados, I. Pytharoulis, and G. Kallos, "Applications of Kalman filters based on non-linear functions to numerical weather predictions," *Ann. Geophys*., vol. 24, pp. 2451-2460, 2006.
[23] M. Chaabene, and M. B. Ammar, "Neuro-fuzzy dynamic model with Kalman filter to forecast irradiance and temperature for solar energy systems," *Renew. Energy*., vol. 33, pp. 1435-1443, Jul. 2008.
[24] S.V. Kumar, "Traffic flow prediction using Kalman filtering technique," *Procedia Eng.*, vol. 187, pp. 582-587, 2017.
[25] Y. Xing *et al*., "State of charge estimation of lithium-ion batteries using the open-circuit voltage at various ambient temperature," *Appl. Energy*., vol. 113, no. 1, pp. 106-115, Jan. 2014.
[26] M. Mastali *et al.*, "Battery state of the charge estimation using Kalman filtering," *J. Power Sources.*, vol. 239, pp. 294-307, Oct. 2013.
[27] T. Soubdhan *et al*., "A Robust Forecasting Framework based on The Kalman Filtering Approach with a Twofold Parameter Tuning Procedure: Application to Solar and Photovoltaïc Prediction," *Sol. Energy*., vol. 131, pp. 246-259, Jun. 2016.
[28] T. Y. Zhou *et al.*, "Improved GNSS Cooperation Positioning Algorithm for Indoor Localization, CMC: Computers," *Comput., Mater. Con.*, vol. 56, no. 2, pp. 225-245, 2018.
[29] G. Rigatos, and P. Siano, "Forecasting of Power Corporations' Default Probability with Nonlinear Kalman Filtering," *IEEE Syst. J.*, vol. 12, pp. 1099-1107, Jun. 2018.
[30] D. Plonis *et al.*, "Predicting the frequency characteristics of hybrid meander systems using a feed-forward backpropagation network," *Electronics*, vol. 8, no. 1, pp. 85, Jan. 2019.
[31] E. Turajlic, A. Begović, and N. Škaljo, "Application of Artificial Neural Network for Image Noise Level Estimation in the SVD domain," *Electronics*, vol. 8, no. 2, pp. 163, Feb. 2019.
[32] C. S. Yuan *et al.*, "Fingerprint liveness detection from different fingerprint materials using convolutional neural network and principal component analysis," *CMC-Comput. Mat. Contin.*, vol. 53, no. 3, pp. 357-371, 2017.
[33] C. Anitescu *et al.*, "Artificial neural network methods for the solution of second order boundary value problems," *CMC-Comput. Mat. Contin.*, vol. *59*, pp. 345-359, 2019.
[34] N. F. Wang, D. X. Jiang, and W. G. Yang, "Dual-Tree complex wavelet transform and SVD-Based acceleration signals denoising and its application in fault features enhancement for wind turbine," *J. Vib. Eng. Technol.*, vol. 7, no. 4, pp. 311-320, Aug. 2019.
[35] Z. Yang, L. Ce, and L. Lian, "Electricity price forecasting by a hybrid model, combining wavelet transform, ARMA and kernel-based extreme learning machine methods," *Appl. Energy*, vol. 190, pp. 291-305, Mar. 2017.
[36] P. M. R. Bento, "A bat optimized neural network and wavelet transform approach for short-term price forecasting," *Appl. Energy*, vol. 210, pp. 88-97, Jan. 2018.
[37] Y. Zhang *et al.*, "A Novel Adaptive Kalman Filter With Colored Measurement Noise," *IEEE Access*, vol. 6, pp. 74568-74578, Nov. 2018.
[38] R. Ferrero, F. Gandino, and M. Hemmatpour, "Estimation of Displacement for Internet of Things Applications with Kalman Filter," *Electronics*, vol. 8, pp. 985, 2019.
[39] Y. Liu *et al.*, "A Novel Content Popularity Prediction Algorithm Based on Auto Regressive Model in Information-Centric IoT," *IEEE Access*, vol. 7, pp. 27555-27564, Feb. 2019



Jibo Chen is presently an Associate Professor at the School of Mathematics and Statistics, Nanjing University of Information Science and Technology, China. His research interests include structural equation modelling and other statistical modelling.

Keyao Chen is presently an engineer at National Climate Center, China Meteorological Administration, China. His research interests include decision making and technical support on Global Climate Change, Historical Climate Reconstruction and Climate Change.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2963416, IEEE Access

Chen Ding is presently a master graduate from the School of Mathematics and Statistics, Nanjing University of Information Science and Technology, China. His research interests include mathematical statistics, and its applications.

Guizhi Wang is presently a Professor at the School of Mathematics and Statistics, Nanjing University of Information Science and Technology, China. Her research interests include mathematical statistics, non-parametric statistical theory and its applications.

Qi Liu (M'11, SM'18) received the B.S. degree in Computer Science and Technology from Zhuzhou Institute of Technology, China in 2003, and M.S. and Ph.D. in Data Telecommunications and Networks from the University of Salford, UK in 2006 and 2010. His research interests include context awareness, data communication in MANET and WSN, and smart grid. His recent research work focuses on intelligent agriculture and meteorological observation systems based on WSN.

Xiaodong Liu (M'00, SM'17) received his PhD in Computer Science from De Montfort University and joined Napier in 1999. He is a Reader and is currently leading the Software Systems research group in the IIDI, Edinburgh Napier University. He was the director of Centre for Information & Software Systems. He is an active researcher in software engineering with internationally excellent reputation and leading expertise in contex-aware adaptive services, service evolution, mobile clouds, pervasive computing, software reuse, and green software engineering.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.