Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 5515103, 16 pages
https://doi.org/10.1155/2021/5515103

Research Article

Longer Time Span Air Pollution Prediction: The Attention and Autoencoder Hybrid Learning Model

Xin-Yu Tu[1], **Bo Zhang**[1,2,3], **Yu-Peng Jin**[1], **Guo-Jian Zou**[4], **Jian-Guo Pan**[1] and **Mao-Zhen Li**[1,5]

[1]*College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai 200234, China*
[2]*Shanghai Engineering Research Center of Intelligent Education and Bigdata, Shanghai Normal University, Shanghai 200234, China*
[3]*Institute of Artificial Intelligence on Education, Shanghai Normal University, Shanghai 200234, China*
[4]*College of Transportation Engineering, Tongji University, Shanghai 200082, China*
[5]*Department of Electronic and Electrical Engineering, Brunel University London, Uxbridge, UB8 3PH, UK*

Correspondence should be addressed to Bo Zhang; zhangbo@shnu.edu.cn and Jian-Guo Pan; panjg@shnu.edu.cn

Received 15 January 2021; Revised 21 March 2021; Accepted 25 May 2021; Published 2 June 2021

Academic Editor: Ana C. Teodoro

Air pollution has become a critical issue in human's life. Predicting the changing trends of air pollutants would be of great help for public health and natural environments. Current methods focus on the prediction accuracy and retain the forecasting time span within 12 hours. Shorter time span decreases the practicability of these perditions, even with higher accuracy. This study proposes an attention and autoencoder (A&A) hybrid learning approach to obtain a longer period of air pollution changing trends while holding the same high accuracy. Since pollutant concentration forecast highly relates to time changing, quite different from normal prediction problems like autotranslation, we integrate "time decay factor" into the traditional attention mechanism. The time decay factor can alleviate the impact of the value observed from a longer time before while increasing the impact of the value from a closer time point. We also utilize the hidden states in the decoder to build connection between history values and current ones. Thus, the proposed model can extract the changing trend of a longer history time span while coping with abrupt changes within a shorter time span. A set of experiments demonstrate that the A&A learning approach can obtain the changing trend of air pollutants, like PM2.5, during a longer time span of 12, 24, or even 48 hours. The approach is also tested under different pollutant concentrations and different periods and the results validate its robustness and generality.

1. Introduction

Air pollution has become increasingly severe with the gradual maturity of global industrialization. Studies have pointed out that air pollution might be one of the key factors leading to noncommunicable diseases and shorter life expectancy [1]. Countries around the world have tackled this problem with various measures and people have tended to change their activities in certain ways to cope with the declining air quality. In the fight against this havoc, forecasting air quality would be the essential and effective way to help human beings.

Thanks to the progress in sensor and recording devices, voluminous information about atmospheres and pollutant concentrations has been gathered during these years. These data have in turn boosted studies on identifying influences of pollutants (e.g., PM2.5, PM10, and AOI) on public and societies [2]. Among them, foreseeing the changing trend of air quality is one the most challenging ones. Facing the colossal raw data, prediction approaches naturally deploy statistics-based methods. For example, ARMA (Autoregressive Moving Average) [3] uses one-dimensional data to predict air pollutions. Although multiple variables are added, this kind of method still follows the way of linear

models and thus can hardly be applied to the long-time span sequential prediction, a kind of nonlinear problem.

Nonlinear models can exert their expertise at extracting deep features from gigantic data, especially when the raw data are more precise. For example, neural networks (NNs) like RNN, LSTM [4, 5], and GRU [6, 7] have been used for various nonlinear problems, including image processing, natural language processing, and automatic driving [8]. Bidirectional LSTM methods have been used to predict air contamination [9]. Current studies on air quality perdition devoted to improving the accuracy while retaining the prediction period within 12 hours; some of them even limited the time within 1 hour to seek the utmost accuracy. Despite the high accuracy, short-time prediction has barely any practical values, given that fighting air pollution needs longer time to prepare.

In this paper, we design a novel attention and autoencoder (A&A) learning model to predict the changing of air pollutants longer than 12 hours. This approach is composed of an Encoder and Decoder structure and LSTM models that contain the attention mechanism. The cooperation between these models is expected to exert a longer-span sequential prediction.

Targeting the temporal characteristic of the pollutant changing trend, this study specifically incorporates the attention mechanism to emphasize the "time" feature through the following techniques:

(1) Add a time factor in the attention mechanism

(2) Consider the decoder hidden states when outputting prediction results

(3) Use the window algorithm to keep the concerned history data stable

The remainder of the paper is organized as follows. Section 2 classifies and reviews related works. Section 3 describes the architecture of the A&A learning approach. Section 4 presents the experimental process. Section 5 draws certain conclusions and points to directions for further work. The block diagram of our research is depicted in Figure 1, and detailed research steps can be found in Section 4. The brief steps are listed as follows:

(1) First, after confirming the specific city and on which year we are going to investigate, we get two datasets of air pollution and atmosphere

(2) Then, we need to preprocess the data

(a) Using mean value interpolation to fill in the missing data

(b) Adjusting both datasets to the same record frequency

(c) Merging two datasets, splitting the train/test datasets, and segmenting the dataset into multiple batches

(3) After that, processed data will be sent to A&A learning for prediction

(4) Finally, we evaluate the predicted value through evaluation metrics and get the result

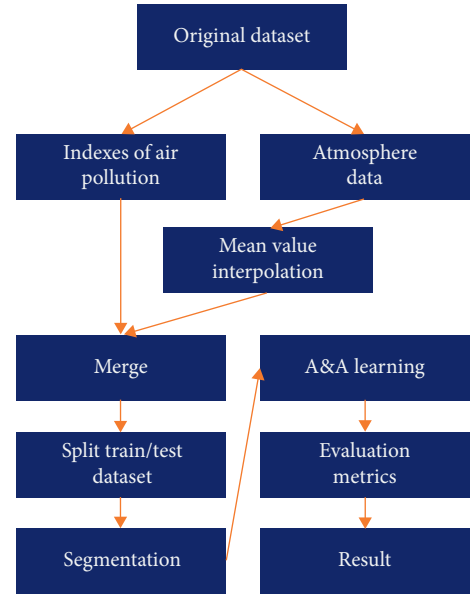

Figure 1: Brief introduction of A&A learning.

2. Related Work

2.1. General Classification. Previous approaches to forecasting air pollution can be classified into theory-based methods and statistics-based ones:

(1) Theory-based methods focus on chemical or physical factors that could affect the variation tendencies of pollutants and construct numerical models to simulate these dynamic processes.

(2) Statistics-based methods analyze a sea of data about past air qualities and atmospheres (e.g., PM2.5, PM10, and AOI) and extract potential relationships among them to forecast the variation tendencies of pollutants.

2.2. Theory-Based Methods. Two theory-based air quality models have been widely accepted around the industry: the community multiscale air quality (CMAQ) [10] modeling system and the comprehensive air quality model with extensions (CAMx) [11]. Both were developed based on the assumption of "one atmosphere." They involved conversion and influences among different types of air contaminants and simulated scales of multipollutants [12].

CMAQ targets a majority of air contaminants and estimates the overall air quality covering multiple regions. Its universal applicability, however, comes with certain innate deficiencies, like errors caused by manually set parameters and perturbed mass conservation induced by inconsistent meteorological fields [13].

CAMx is a publicly available photochemical grid model that was developed and coded during the late 1990s by modern and modular coding practices [14]. It was proved to be effective in simulating various atmospheres and different scales of particulate pollutants covering cities or even

crossing continents. Yet problems still exist in accuracy and comprehensiveness of compound pollutants simulation. For instance, the particular matter for comprehensive air quality model with extensions (PMCAMx) predicts even higher values of O3 and PM than CMAQ does [15]. Although theory-based models will work efficiently when they are fed with accurate data, the mega-computation and inevitable gentle errors caused by the models do influence their performance. Plus, it is hard to collect abundant data for them in practice.

2.3. Statistics-Based Methods. Meanwhile, the acceleration of computing power has brought statistics-based machine-learning models back to life. Machine-learning approaches have been widely used for classification and regression, thanks to their ability to extract potential relationships among a vast number of features, as well as thinking like a human and reacting reasonably without a predesigned program. They are traditionally divided into three categories: supervised, unsupervised, and semisupervised study, based on different amounts of labels used during the training period [16].

Some researchers have adopted certain machine-learning techniques to predict air pollutions, like SVM algorithms [17, 18] and the semiexperimental regression model [19], and obtained satisfying results on small-scale data. However, they can barely forecast long-period trends.

Artificial NNs, a branch of machine learning, have been proved to be promising for time-dependent forecast, for example, RNN [20], LSTM [21], and GRU [22]. An artificial NNs approach is to construct a model by simulating human's NNs. With multiple nerve cells depicting the profound connections among data, nonlinear activations, and the backpropagation algorithm [23], artificial NNs could classify or predict things more practically and be apt at solving nonlinear, random, and irregular problems.

2.4. NN-Based Methods. In the field of air quality prediction, it is believed that accuracy is partially inversely proportional to time span. This conclusion has kept most researchers who deployed statistics-based methods focusing on the accuracy of prediction while shortening the limit of time to one hour [24] or one day (with only one value representing the pollution level each day) [25], which is far from meeting practical needs. Our A&A learning approach, however, puts real-life activities first by extending the forecast time to 12 or 24 hours, or potentially even 48 hours. This approach extracts the nonlinear spatial and temporal features of the changing of air pollutant concentration from previous data. Its main purpose is not to push the accuracy to the climax but to predict the tendency of the air pollution indexes within a longer period while retaining a higher accuracy.

Numerous barriers could impede obtaining accurate air quality forecasting. First and foremost, air pollutant concentration is influenced by certain human activities, like letting off fireworks or driving cars, and by meteorological factors such as wind speed and sea level pressure [25, 26]. To tackle this complex situation, the proposed A&A approach deploys LSTM and GRU that can transform numeric air pollution information into characteristic vectors and clarify deeper regulations between these features by capturing the temporal information of fifteen dimensions of data. These two methods can contribute to the prediction of PM2.5 distribution for a longer time span.

Prediction for a longer time span in this paper is a typical sequence-to-sequence problem [27]: the observation period versus the prediction period which are not equal. This separates the proposed model into an encoder and a decoder; thus, the model can conduct feature extracting and forecasting separately. The core of both is LSTM. The encoder, fed with original data, is to finish the excavating task, compress the distribution of air pollutant features during the observed period into a cell state, and send it to the decoder. Once the decoder gets the cell state, prediction can be initialized based on it. Thus, the inequality between the two time spans disappears.

Another critical issue is that pure RNN, LSTM, and GRU models could cause the vanishing gradient problem. Due to the "tanh" activation frequently reused in these models, gradients drop dramatically along with time. That means previous information could vanish during a long sequence prediction. To solve this problem, we adopt an attention-theory-based [28] algorithm to quantify to what extend the previous pollutant information can influence current prediction. Thus, the condensation of PM2.5 can be forecasted by the matrix output from the concatenated encoder and decoder, and the model's ability to seek long-term regularity among the historical data can be significantly improved.

To sum up, our research emphasizes the following critical tasks:

(1) Extract correlated features by a sequence-to-sequence framework and transform them into vectors

(2) Quantify the influences between the observed period and the forecasting period

(3) Develop a time-related attention algorithm to extract the regularity of historical data and quantify its weights

3. Architecture of the A&A Learning Approach

3.1. Overview of A&A Learning. The proposed A&A learning approach uses the improved attention mechanism to build the encoder-decoder structure with the LSTM model. This architecture can tackle particular problems within the NNs while strengthening the extracting ability.

First, the encoder-decoder structure can avoid the conflict between overfitting and linear prediction. We have conducted multiexperiments that used pure LSTM to forecast air quality. Although it could output the regularity of the variation of PM2.5, the fixed vector size drove its performance unstable. As the training epochs getting bigger, the error was becoming significantly high. If we enlarged the vector size, the overfitting problem would be more severe; and if we did it conversely, the result would tend to be more linear, hindering the deeper seeking of information about the air quality. To alleviate the conflict, we adopt the

sequence-to-sequence structure to retain more historical data for later prediction.

Second, the improved attention mechanism can facilitate the effect of "time" in the prediction results. The combination of LSTM and sequence-to-sequence models is not enough, and the improved attention mechanism can help the approach to obtain results based on different periods of time.

Our approach intends to highlight the constraint of "time" through the following three ways:

3.1.1. Multiply a Time Factor according to the Time Gap between the Previous Data and the Current Ones. Although the traditional attention mechanism is powerful, it evenly allocates the impact of each observed point on the predicted point. In the air quality prediction task, however, the impact of pollutants on the air quality at the predicted time changes with the time span between the observed time and the predicted time; that is, the longer the time gap between the observed time and the predicted time, the less the impact. Therefore, certain time-related decay factors and numerical scoring mechanisms must be incorporated into the attention mechanism of this study.

Actually, the improved attention mechanism adds a time factor to the attention queries. The time factor is calculated by the difference between the observed time and the predicted time and helps to modify the hidden states.

3.1.2. Take the Decoder Sequence into the Attention Queries' Calculation. The traditional attention mechanism only considers the encoder's hidden state and current hidden states; thus, the current predicted values cannot impact later prediction, which is against common sense. Therefore, the improved attention mechanism involved the hidden state of the decoder to push the prediction spreading forward. Use a fixed window size to simulate the time step.

When adding the hidden state of the decoder to the attention values, the stable size of the attention values will be boosted along with the prediction proceeding. Given the decaying impact, a window is "framed" on the sequence of the hidden states. The windowing algorithm abandons the hidden state at the longest time point when a new hidden state from the decoder joins; thus, the total amount of the attention values can keep stable.

To make the procedure clear, we have made a block diagram of the A&A learning approach which can be found in Figure 2 and there are 6 steps to predict the value.

Step 1. First, a batch of features with fifteen dimensions will be sent to train the encoder.

Step 2. Then, the cell state of encoder and the last row of fifteen features will be transferred to decoder through route 3 and all hidden states will be recorded and sent to the improved attention mechanism through route 2.

Step 3. After using the last row of fifteen features as the input and the cell state transferred from encoder, the decoder will give out current hidden state, which will be sent to the attention mechanism through route 4.

Step 4. Next, we will multiply time factor to emphasize the importance of hidden states with a closer time span, add the hidden state of one step before produced by decoder, and drop the most ancient one. After that, hidden states with stable size will be sent to attention mechanism through route 5.

Step 5. Additionally, we calculate the result by using current hidden state and hidden states with stable size and sent to decoder through route 6.

Step 6. Finally, after a full connect layer, the predicted value will be sent through route 7 and the current hidden state will be sent to the improved attention mechanism, and it will also be used to update the hidden states through step 4 before the next round of forecast.

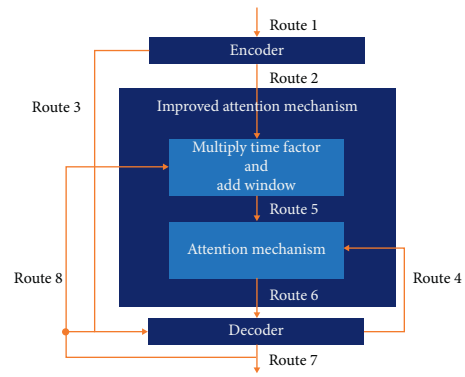

Figure 2: Block diagram of the A&A learning approach.

Also, we have prepared the overall structure of the A-A learning approach in Figure 3 to illustrate this progress in detail.

3.2. Recurrent NNs. Artificial NNs can be divided into feedforward NNs and recurrent NNs [29]. Backpropagation NNs, the typical feedforward NNs, contain no links among the neurons of the same layers, which means that no interactions between historical and current data exist. This could lead to the model's insensitivity to the variation of air pollutant concentration.

Recurrent NNs, however, can connect previous and current data. Each node in the network is a computing unit that incorporates both the hidden state from the previous units and the current input data and outputs a relevant result. This mechanism can help the model to retrospect previous information during prediction.

Suppose the input series is denoted as $X = \{x_1, x_2, x_3, \ldots, x_{t-1}, x_t\}$, the hidden state, containing the condensed historical data, as $S = s_1, s_2, s_3 \ldots, s_{t-1}, s_t\}$, and the output of each time step as $O = \{o_1, o_2, o_3, \ldots, o_{t-1}, o_t\}$. Recursion of the recurrent NN along with the time of the sequence can be calculated as

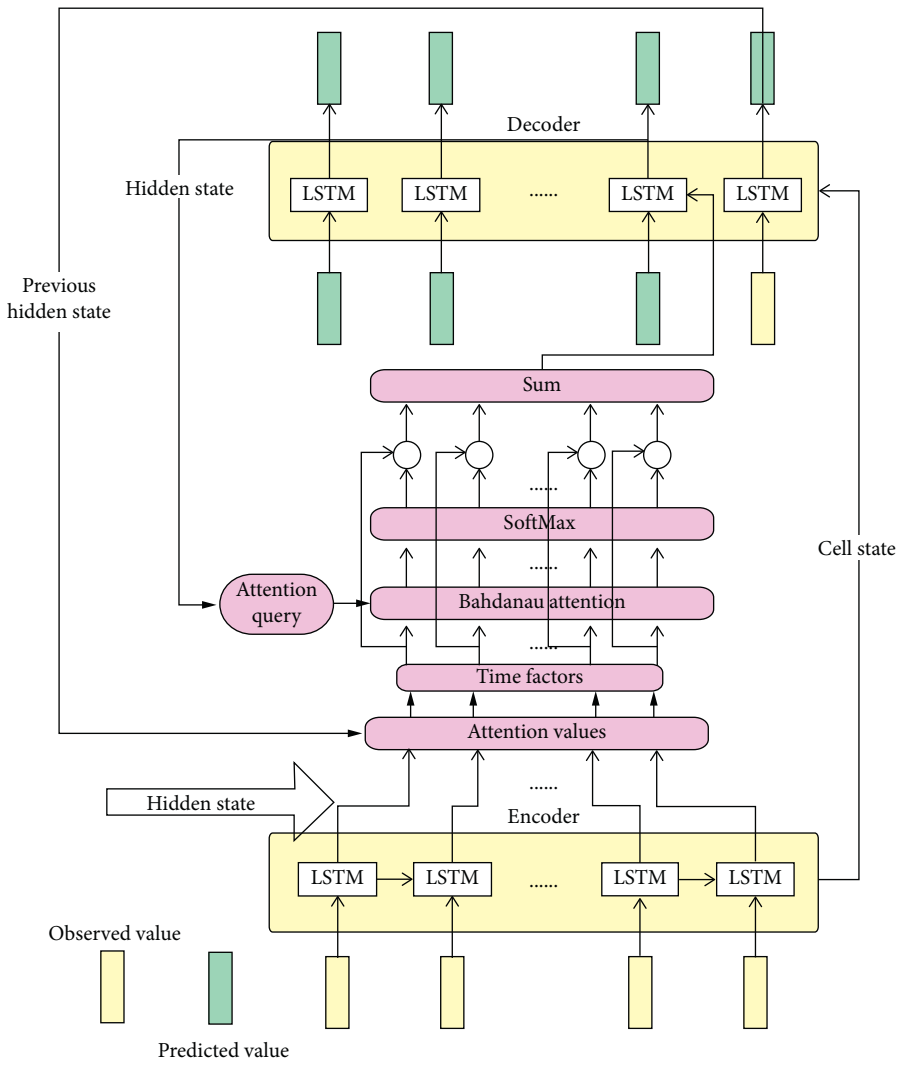

Figure 3: Overview of the A&A learning approach.

$$s_t = f\left(U \cdot x_t + W \cdot s_{t-1}\right),$$
$$o_t = softmax\left(Vs_t\right), \tag{1}$$

where "f" represents the activation functions, among which "tanh" and "relu"' are the frequently used; "U, V, and W" are the shared weights of the RNN. They represent the critical feature that can enable the RNN to capture relationships among diverse units and can also cause gradient vanishing after a long-sequenced calculation.

3.3. Long Short-Term Memory (LSTM). LSTM was invented to solve the gradient vanishing problem by Hochreiter, S and Schmidhuber, J in 1997. It updates and eliminates the information of hidden states and cell states of the units. The extra state, the cell state, stretches throughout the LSTM units, enabling a longer memory of historical information on the basic RNN. Thus, LSTM holds the ability to process longer sequenced data.

Suppose the input series is denoted as $X = \{x_1, x_2, x_3, \ldots, x_{t-1}, x_t\}$, the hidden state, containing the condensed historical data, as $S = \{s_1, s_2, s_3, \ldots, s_{t-1}, s_t\}$, the cell state on behalf of the current state of LSTM as $C = \{c_1, c_2, c_3, \ldots, c_{t-1}, c_t\}$, and the output of each time step as $O = \{o_1, o_2, o_3 \ldots, o_{t-1}, o_t\}$. Recursion of the RNN along with the time of the sequence is calculated as

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right),$$
$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right),$$
$$C_t^{'} = \sigma\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right),$$
$$C_t = f_t * C_t + i_t * C_t^{'}, \tag{2}$$
$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right),$$
$$h_t = h_t * f\left(C_t\right),$$

where $[a, b]$ means concatenating matrixes a and b; "$W_?$" and "$b_?$" represent the outcome weight and the bias of various equations, respectively; the subscripts "i," "f," "o," and "c" indicate that the parameters are related to the input gate, the forget gate, the output gate, and the cell state, respectively; $C_t^{'}$, the supplementary cell state, represents the

alternative information for cell states update; "σ" is the sigmoid function; and "f" is the tanh activation.

First, the forget gate drops the reluctant part of the cell state to simulate the loss of the historical data with time stepping forward according to both the previous hidden state and the current input. Second, the input gate controls updating the information for the cell state. Third, after updating the cell state according to the previous cell state C_{t-1} and the supplementary cell state $C_t^{'}$, the output gate chooses the information for the hidden state that will affect the calculation of the following LSTM cells.

Although LSTM is apt at solving long sequence problems, compared to RNNs, its attention in this study needs to be paid to the saltation and high-value part of the data, because air pollution concentration does not change dramatically in a high frequency.

3.4. Gated Recurrent Unit. A Gated Recurrent Unit (GRU) is a simplified edition of LSTM and has been well applied. It uses the update gate (z_t) that is transferred from the forget, the input and the update gate in LSTM, and the reset gate $(r)_t$ and merges the hidden state with the output.

The merged state, the hidden state, is updated according to both the previous hidden state h_{t-1} and the supplementary state $h_{t-1}^{'}$ and their weights depend on the result of the update gate.

$$\begin{aligned} r_t &= \sigma\left(W_r \cdot \left[h_{t-1}, x_t\right]\right), \\ z_t &= \sigma\left(W_z \cdot \left[h_{t-1}, x_t\right]\right), \\ h_t^{'} &= f\left(W \cdot \left[r_t \cdot h_{t-1}, x_t\right]\right), \\ h_t &= \left(1 - z_t\right) \cdot h_{t-1} + z_t \cdot h_t^{'}, \end{aligned} \tag{3}$$

where "W_r,""W_z," and W represent the weights for the diverse steps of calculation.

The GRU holds the same capability of producing excellent results with that of LSTM [30].

3.5. Sequence-to-Sequence Structure. As mentioned before, prediction for air quality requires alignment between the historical data and the result matrix; the sequence-to-sequence structure, also called encoder-decoder structure, can well solve this problem by using two RNN models. This structure separates the prediction process into two parts based on the observed period and the forecasting period. During the observed period, one RNN model is used to extract information from historical data of each time step and condense it into a hidden state. This hidden state is then transmitted to the other RNN model used in the forecasting period, initializing the following prediction with the latest preserved data. The current simulated value is calculated after a full-connected network. Thus, the current predicted value is taken as the input of the next perdition and the procedure is iterated until the final sequence is obtained. This state-sharing strategy balances the nonaligned time spans between the input data and the output data, allowing a longer sequence of data to be analyzed.

In this study, we replace the RNN models with LSTM networks for better extracting results and nonlinear simulating ability.

Although it is claimed that LSTM can help to relieve the vanishing gradient caused by the reuse of the "sigmoid" function [31], the conflict between overfitting and losing historical information generated by the fixed vector size of the cell state still exists. When the size of the vector grows, the overfitting problem becomes severer, while retraining the overfitting leads to a shortage of historical information for the following prediction. Extra measures should be deployed to avoid this.

3.6. Attention Mechanism. The attention mechanism initially developed for image processing has been proved to be powerful in multiple fields of deep learning in recent years. Vaswani et al. [32] indicated that the attention mechanism performed outstandingly in sequential tasks, like translation of NLP (Natural Language Processing). Other studies [33, 34] demonstrated that combing the sequence-to-sequence structure and the attention mechanism could obtain effective results in air quality prediction.

Targeting the LSTM's dilemma between overfitting and previous data losing, the attention mechanism designates an energy function and an alignment vector to quantify the similarity between the previous hidden state of decoder h_{t-1}^d and all the previous encoder hidden states H_E at time "t."

Assume the time span of the observed period in the sequence-to-sequence structure is denoted as O and the predicted period as P, we can get the preserved encoder hidden state series of the encoder $H_E = \{h_1^e, h_2^e, h_3^e, \ldots, h_{O-1}^e, h_O^e\}$, the hidden state of the decoder $H_D = \{h_1^d, h_2^d, h_3^d, \ldots, h_{P-1}^d, h_P^e\}$, the alignment vector $A = \{a_1, a_2, a_3, \ldots, a_{O-1}, a_O\}$, and the output of the decoder $O_D = \{o_1^d, o_2^d, o_3^d, \ldots, o_{P-1}^d, o_P^e\}$. The variant "$p_t$" is a pollutant vector representing the output of the attention mechanism. It quantifies the contributions of all the information to the current output of forecasting through a weighted summation. The weights are calculated through the energy function and the SoftMax function [35]. Therefore, the parameters of LSTM of the decoder with the attention mechanism at time "t" are updated by the following equations:

$$\begin{aligned} p_t &= \sum_{t=1}^{O} a_t h_t^e, \\ O_t^d &= F\left(h_t^d, p_t\right), \end{aligned} \tag{4}$$

where F represents a full-connection layer.

The energy function $E = \{e_1, e_2, e_3, \ldots, e_{O-1}, e_O\}$ represents the correlation between the current hidden state and the previous one. It is also needed in the calculation of alignment vector at time "t" (a_t).

Multiple energy functions can be used by the attention mechanism. Among them, the Bahdanau attention [36] and cosine similarity attention are the most related ones for this study.

Bahdanau attention is as follows:

$$e_j = v_a^T f\left(W_a h_j^e + U_a h_{t-1}^d\right). \tag{5}$$

Cosine similarity is as follows:

$$\begin{aligned} e_j &= \frac{h_j^e \cdot h_{t-1}^d}{\left(h_j^e * h_d^{t-1}\right)}, \\ a_j &= \frac{\exp\left(e_j\right)}{\sum_{k=1}^{O} \exp\left(e_k\right)}, \end{aligned} \tag{6}$$

where "j" and "k" represent the encoder time steps, ranging from 1 to the time span of the observed period O.

Thanks to the attention mechanism and the encoder-decoder structure, the proposed approach can retrospect abundant information when simulating the current concentration of PM2.5. The windowed attention queries also ensure that the approach pays closer attention to the time span between the current hidden state and the previous ones without altering the core task of the traditional attention mechanism-fixed alignment. In addition, the time decay factor contributes an extra part to the high concentration on time of the A&A learning approach. The difference can be found in Figures 4 and 5, which depicts a typical attention-mechanism-based algorithm and the improved attention mechanism-based one, respectively.

3.7. The A&A Learning. As the A&A learning approach incorporates the time factor, a time decay factor "D" is added to the calculation of the energy function in the improved attention mechanism. Additionally, the retrospective hidden states include not only the hidden states of the encoder but also those of the decoder. Thanks to the window whose time span matches the length of the observed period O, the size of the retrospective hidden states keeps stable, and the information for both the observed period and the predicted period can be analyzed and extracted.

The A&A learning approach forecasts the concentration of the specific air pollution index PM2.5 based on the following equations.

For LSTM in the sequence-to-sequence structure,

$$\begin{aligned} f_t &= \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right), \\ i_t &= \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right), \\ C'_t &= \sigma\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right), \\ C_t &= f_t * C_t + i_t * C_t^{'}, \\ o_t &= \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right), \\ h_t &= h_t * f\left(C_t\right). \end{aligned} \tag{7}$$

For the improved attention mechanism (the Bahdanau attention is taken as an example).

Assume that the retrospective hidden states are $H_R = \{h_1^r, h_2^r, h_3^r, \ldots, h_{O-1}^r, h_O^r\}$, and at time "t", H_R is composed of the hidden states of the encoder whose time step ranges from "t-1" to "O" ("O" refers to the length of the observed period) and the decoder from time step "1" to "$t-2$"; thus, the series of all related hidden states used in the calculation can be denoted as $H_R = \{h_{t-1}^e, h_{t-2}^e, \ldots, h_{O-1}^e, h_O^e, h_1^d, h_2^d, \ldots, h_{t-3}^d h_{t-2}^d\}$. Before sending the hidden states H_R into the improved attention model, the approach should multiply the time decay factor $D = \{d_1, d_2, \ldots, d_{O-1}, d_O\}$ to relieve the influence of those data that are highly correlated with the decoder's hidden state but hold a long-time span from it. Thus, the pollutant vector p_t is updated by the following equation:

$$p_t = \sum_{t=1}^{O} a_t d_t h_t^r. \tag{8}$$

The forecasted value of time "t" is formed with the pollutant vector and the current hidden state of the decoder.

$$O_t^d = F\left(h_d^t, c_t, p_t\right). \tag{9}$$

The energy function and the alignment vector remain unchanged.

$$\begin{aligned} e_j &= v_a^T f\left(W_a h_d^{t-1} + U_a h_c^j\right), \\ a_j &= \frac{\exp\left(e_j\right)}{\sum_{k=1}^{O} \exp\left(e_k\right)}. \end{aligned} \tag{10}$$

To sum up, the A&A learning approach inherits the strength of the attention mechanism to extract more historical information and puts emphasis on time with the help of the windowed hidden states and the time decay factor.

4. Experiments

To evaluate the performance of the A&A learning approach, we conducted certain experiments on the datasets of real atmosphere and pollutant data.

4.1. Experiment Setup and Data Collection. All experiments were conducted on a PC Server with an AMD Ryzen 7 3700X 8-Core CPU of 3.6 GHz, an 8G GeForce RTX 2080 SUPER GPU, and a 32 GB memory.

We utilized two datasets for the experiments: the atmosphere dataset with four dimensions of features (e.g., air temperature and dew point temperature), originated from the U.S. National Climate Data Center, and the air pollutant dataset with fifteen dimensions of features (e.g., PM2.5, PM10, and AOI), gleaned from the China Meteorological Bureau. Both datasets contained data from 2017 to 2018.

We used a crawling tool to collect data every single or three hours. Since this process might be interrupted by unexpected events, like hardware or power failures, a tiny missing rate, lower than 5%, was allowed in our datasets. And we adopted the mean imputation method to fill up the missing data, as air quality usually does not change dramatically in a short period [37].

4.2. Feature Selection. We conducted a correlation analysis on the air quality indexes to achieve two goals. First, the irrelevant features were identified and eliminated for the sake of the convergence of the model. Second, a specific index that held strong relations with all other indexes was

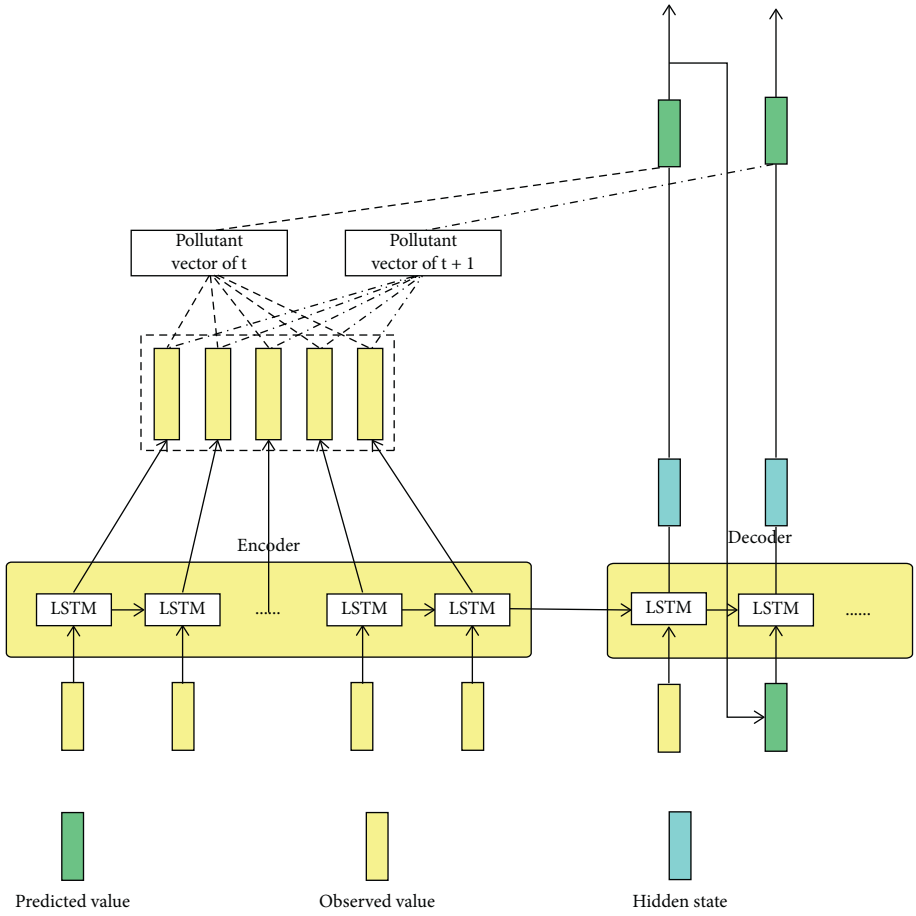

Figure 4: A typical attention-mechanism-based algorithm.

determined and chosen as the object for later experiments. This is because encompassing multi-indexes into the model would incur higher errors, and one specific index related to all others is enough to well reflect the performance of the model.

The correlation result of the indexes in 2018 (the most recent year in our datasets) is shown in Figure 6.

The results indicated that features related to ozone were redundant, since no others well correlated with them, and the index PM2.5 could be the specific index due to its high relevance with others. Thus, the task of the proposed A&A learning approach in the following experiments was to predict the concentration of PM2.5 for future twelve or twenty-four hours. A typical PM2.5 concentration variation of Shanghai ranging from 2018.01.01 to 2018.02.28 is shown in Figure 7.

4.3. Experiment Design. We compared the performances of the pure Decoder-Encoder model, the traditional attention-mechanism-based Decoder-Encoder model, and the pure LSTM model with the proposed A&A learning approach.

To validate the robustness of our model, we took two cities with different degrees of contamination as the forecasting targets: Shanghai with an average concentration of PM2.5 of 35.97 and a standard deviation of 28.85 and Wulumuqi with an average concentration of 56.35 and a standard deviation of 57.82.

We conducted sensitivity tests on the hyperparameter of node size of LSTM (or GRU) to prove that our model does not heavily depend on specific parameter combinations. We also run the model over various time spans to validate its predictive power.

4.4. Evaluation Metrics. The error between the predicted value of a model and the observed value can be utilized to indicate the performance. To quantify the errors, we adopted certain indicators.

The mean absolute error (MAE) is the frequently used indicator to reflect the numeric gap between the ground truth and the predicted value. As the concentration of the pollutant could be lower than 20 or higher than 200, the mean absolute percentage error (MAPE) was chosen to

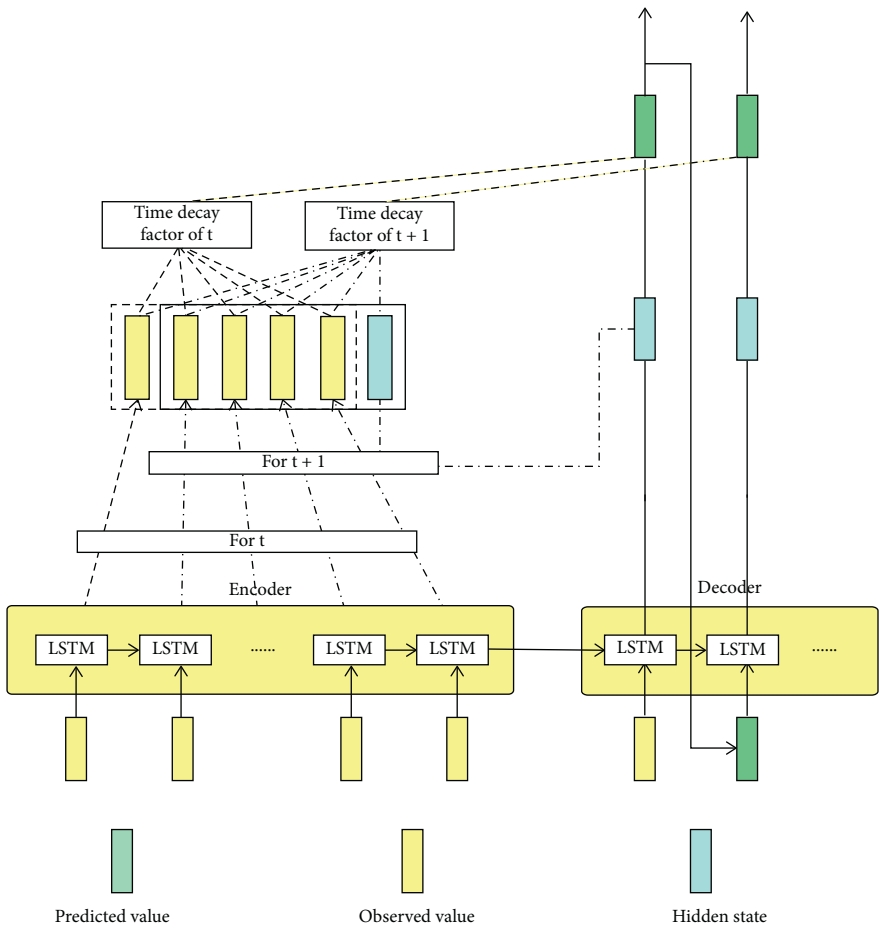

Figure 5: The improved attention mechanism-based algorithm.

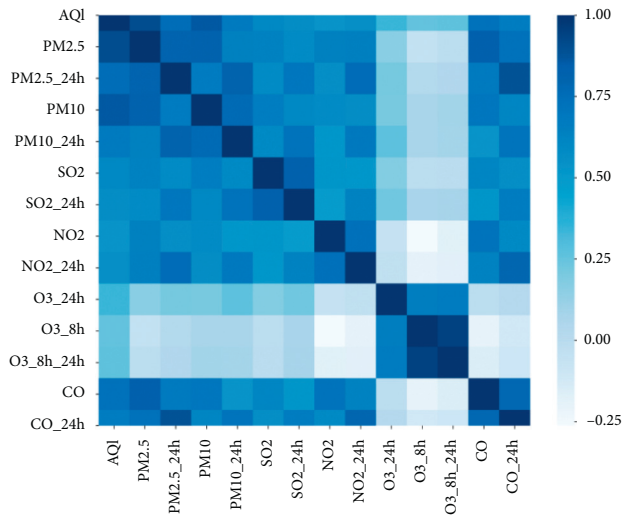

Figure 6: Correlation result of multiple indexes.

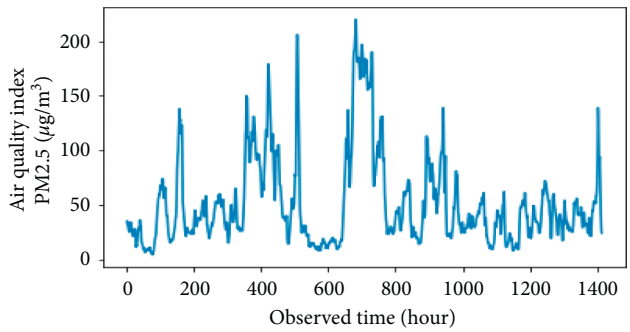

Figure 7: A typical PM2.5 concentration variation of Shanghai ranging from 2018.01.01 to 2018.02.28.

depict the errors. Finally, a correlation analysis was conducted on the predicted values and the real values in order to quantify model's ability to extract tendency.

$$\begin{aligned} \mathrm{MAE} &= \frac{1}{n}\sum_{i=1}^{n}\left|\left(y_i - y_t^{'}\right)\right|, \\ \mathrm{MAPE} &= \frac{\sum_{i=1}^{n}\left|y_i - y_t^{'}\right| * 100}{(y * n)}, \\ \mathrm{COR} &= \frac{\mathrm{Cov}\left(Y, Y'\right)}{\sqrt{\delta Y} * \sqrt{\delta Y'}}, \end{aligned} \tag{11}$$

where "y_i" represents the predicted values, "y_i" the observed values, and "n" the time span of the predicted period.

4.5. Data Preprocessing. Data preprocessing is one of the most critical factors that could influence the training of a model. The amount, reliability, and appropriateness of the dataset and even the interpretation of the data could affect the ultimate results. The atmospheric dataset needed to be split into multiple parts with the same observed period and the predicted period and then be fed into the minibatch training and the encoder-decoder structure. During the training period, the segmentation is illustrated in Figure 8, with time span of observed period being 3, predicted period 2, and step of the window 3. The green blocks of this illustration are values we can observe and the orange part is values we are going to predict.

In the experiments, we separated each dataset into an observation part (denoted as "T_O") and a prediction part ("T_p"). Then, a typical data unit could then be formed with shape of [$T_O + T_P$, features], where "features" represent the number of factors used in the prediction. The time step T_S, representing the gap between two adjacent data units, was determined, and the data units could be concatenated with shape of [batch size, $T_O + T_P$, features] by the minibatch algorithm during the training period.

We then adopted Z-score normalization [38] to eliminate the effects of different dimensions, thus accelerating the model's convergence. The normalization is described as

$$x' = \frac{x - \mu}{\sigma}, \tag{12}$$

where "x" stands for the transformed data and x for the untransformed; μ and σ represent the mean and the standard values of this feature, respectively.

The training set contained the data of the whole year of 2017 and the data of 2018 before the unstable predicting period. The time span of the tests was set as a constant number of 360 hours.

5. Results and Discussions

5.1. Comparison Results. The results of the general comparison between the A&A learning approach, the LSTM, the GRU, the pure Decoder-Encoder model, and the traditional Attention Mechanism-based Decoder-Encoder model on datasets of Shanghai (SH) and Wulumuqi (WLMQ) from 2017.01.01 to 2018.12.31 will be shown in this section.

The first experiment, conducted on the data of Shanghai, set the training epochs as 50, the time step of data separation as 48, the node size as 64, the time span of the observation as 48 hours, the prediction time as 24 hours, and the split rate of training as 0.7. The results are listed in Table 1. The segmentation of [23, 72] stands for time span of 72 hours in total and the predicted time span of 24 hours, which also indicates that the observed time span takes up to 48 hours.

The second experiment changed the node size to 32 and kept other parameters the same. The results are listed in Table 2. To test our model's generalization, we conducted another experiment on the dataset of Wulumuqi, with a node size of 64, and the results are listed in Table 3. The fourth experiment changed the node size to 32 and kept the same parameters with those in the fourth one and the results are listed in Table 4.

To clarify the performance of LSTM, GRU, SEQ2SEQ, Attention, and A&A learning, a sort of line chart depicting the predicted result of the first and third experiments has been shown in Figures 9–18. Compared with A&A learning, other models have a common weakness and periodicity and it is hard for them to seize the variation trend. Thus, we can draw the conclusion that the A&A learning performs better than any other models on both Shanghai and Wulumuqi datasets in all four experiments. As it is depicted above, the accuracy of the proposed model increased alongside the node size. This phenomenon reflects the A&A learning's requirement of a larger node size to restore more previous information to get an accurate result, and among these five models, it is clear to notice that the COR of A&A learning rank much higher than any other models, which has also been depicted in Figures 13 and 18 with high consistency between the predicted value and the real one. Meanwhile, when we compared the result of all five models, the stable performance of A&A learning under different pollute condition demonstrated the generalization and effectiveness of our model.

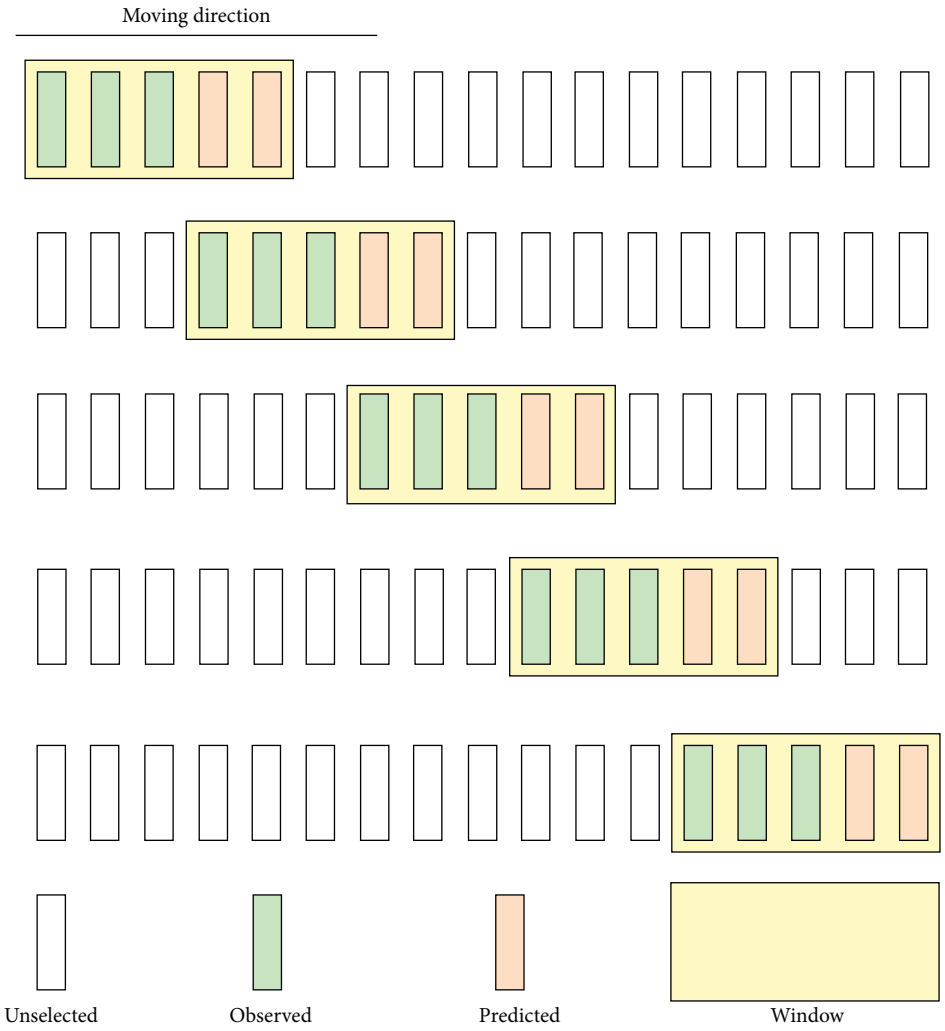

Figure 8: Data segmentation.

Table 1: Experiment on the dataset of Shanghai with time span of 48, node size of 64, and data segmentation of [24, 72].

Model	MAE	MAPE	COR
LSTM	15.56	97.32	0.27
GRU	15.38	103.51	0.21
SEQ2SEQ	13.21	83.27	0.19
Attention	14.53	89.41	0.21
A&A learning	8.97	56.95	0.63

Table 2: Experiment on the dataset of Shanghai with time span of 48, node size of 32, and data segmentation of [24, 72].

Model	MAE	MAPE	COR
LSTM	11.36	73.02	0.28
GRU	11.38	69.20	0.30
SEQ2SEQ	12.69	76.01	0.26
Attention	11.21	70.45	0.35
A&A learning	10.93	61.72	0.37

Table 3: Experiment on the dataset of Wulumuqi with time span of 48, node size of 64, and data segmentation of [24, 72].

Model	MAE	MAPE	COR
LSTM	13.01	97.32	0.31
GRU	16.97	127.54	0.40
SEQ2SEQ	14.83	97.16	0.41
Attention	14.68	105.19	0.32
A&A learning	11.21	80.31	0.55

Table 4: Experiment on the dataset of Wulumuqi with time span of 48, node size of 32, and data segmentation of [24, 72].

Model	MAE	MAPE	COR
LSTM	12.57	90.21	0.33
GRU	11.79	77.83	0.35
SEQ2SEQ	15.02	111.47	0.32
Attention	13.91	99.75	0.21
A&A learning	10.41	67.20	0.35

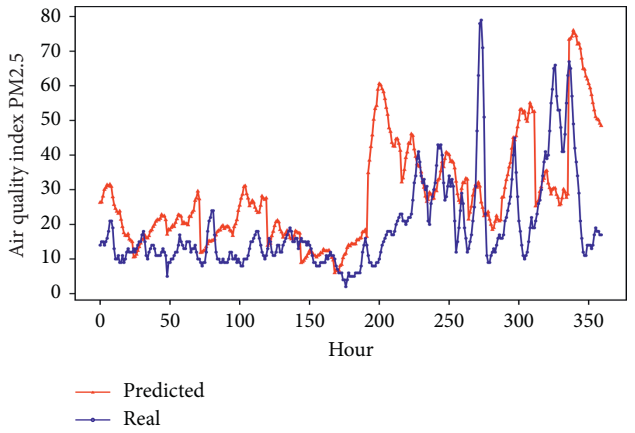

Figure 9: LSTM result on Shanghai dataset with node size of 64.

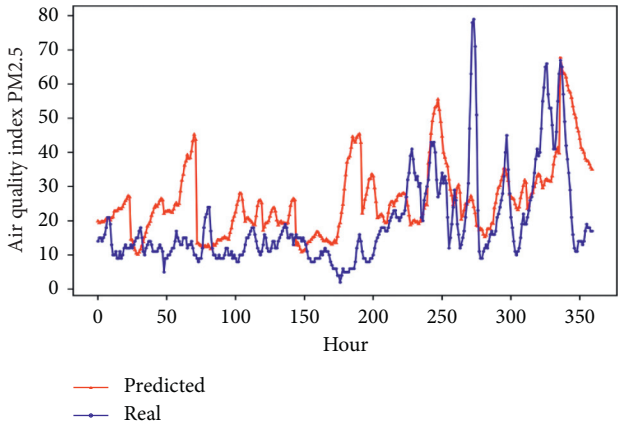

Figure 10: GRU result on Shanghai dataset with node size of 64.

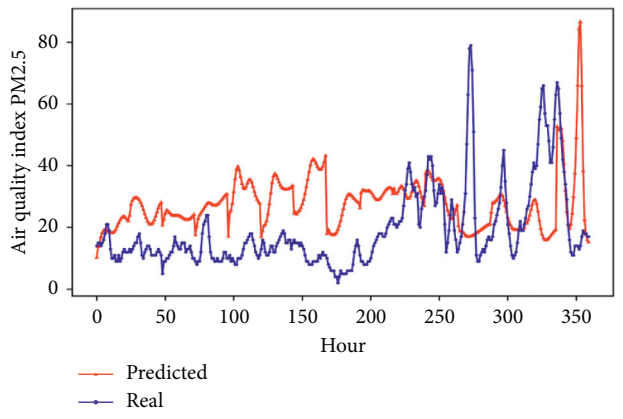

Figure 11: SEQ2SEQ result on Shanghai dataset with node size of 64.

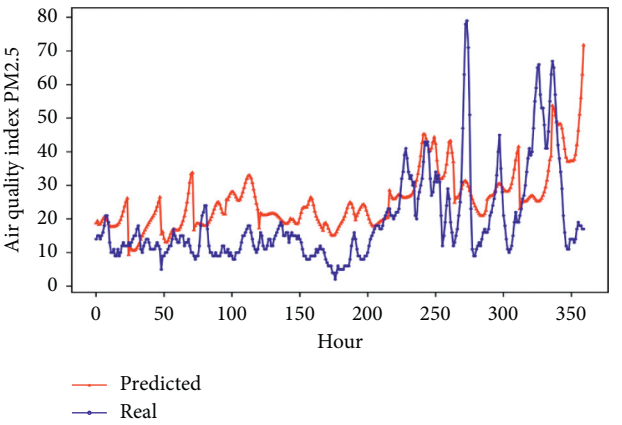

Figure 12: Attention result on Shanghai dataset with node size of 64.

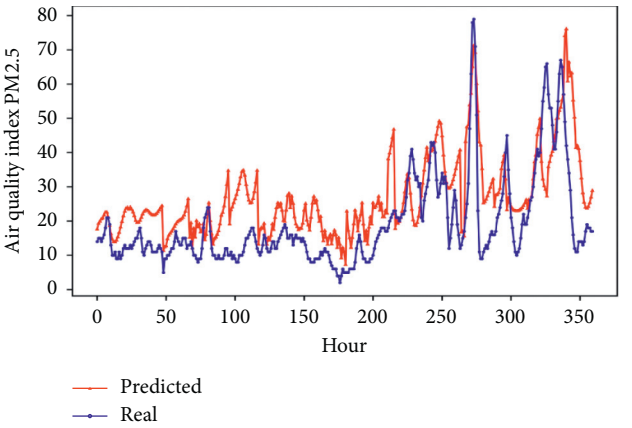

Figure 13: A&A learning result on Shanghai dataset with node size of 64.

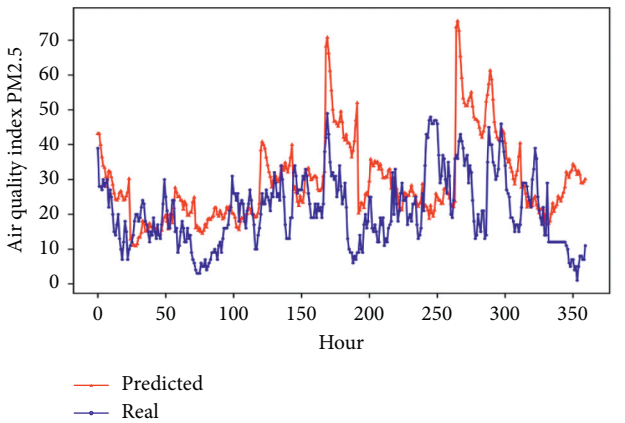

Figure 14: LSTM result on Wulumuqi dataset with node size of 64.

To investigate the ultimate forecasting ability of these modes, we then conducted a series of experiments with the data segmentation of [11, 34] and [48, 144], where the parameters were not specifically adjusted to favor our model. With various predicted spans, we will figure out how long the A&A learning can predict. The results are listed in Table 5 (in the column of Dataset, W represents Wulumuqi and S Shanghai, and for the evaluation metrics like MAE, MAPE, and COR, the former score represents the result with LSTM's node size of 32 and the latter one represents that of 128.)

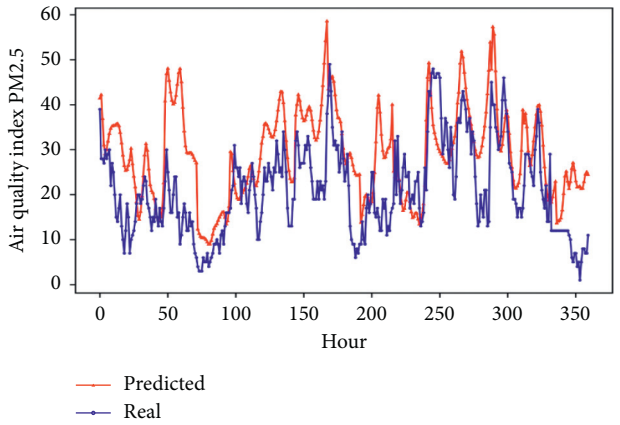

Figure 15: GRU result on Wulumuqi dataset with node size of 64.

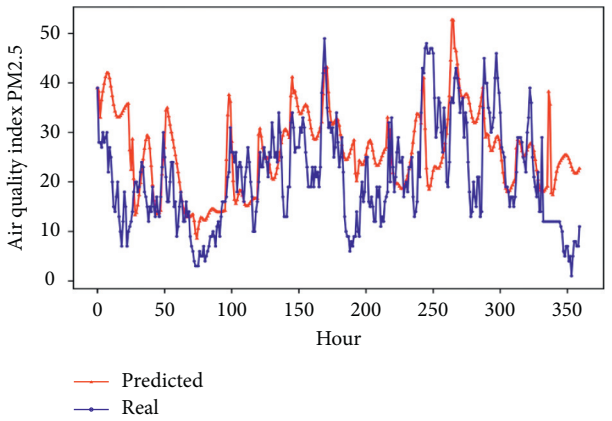

Figure 16: SEQ2SEQ result on Wulumuqi dataset with node size of 64.

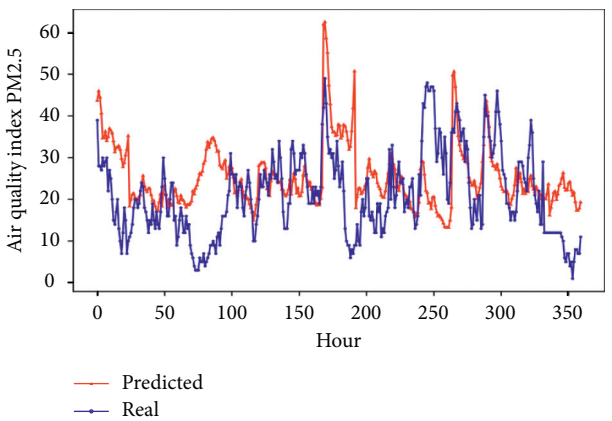

Figure 17: Attention result on Wulumuqi dataset with node size of 64.

When the time spans varied from [23] and [72] to [11, 34] and [48, 144], the results remained broadly the same. With a node size of 128, A&A learning maintains its well performance. However, with a node size of 32, A&A learning still ranked the top place, which implied that higher

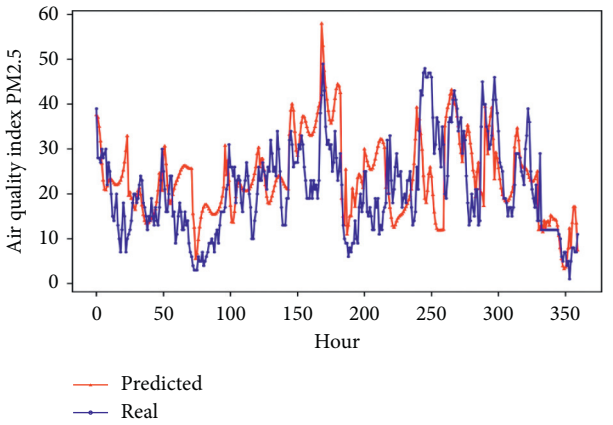

Figure 18: A&A learning result on Wulumuqi dataset with node size of 64.

prediction accuracy of the A&A learning depends on a specific node size for its inner model, and the longer the time span for prediction is, the larger the node size is required.

In Table 5, although the A&A learning approach did not get the first place in all the results, its results only have a gentle difference from the first one. As we take the predicted value as the input of the following value which is going to be predicted, the accumulative errors will grow unavoidably. Through this way, our prime target is to fetch the variation tendency which can be reflected by the COR result, and thus, it is possible for us to forecast the variation trend rather than focusing on a specific value in one moment.

Though LSTM and GRU performed outstandingly, their simple network structures made their result fluctuate around a certain line which has been illustrated in Figures 9 and 10. But if the training epochs were enlarged, the simple network would hamper them from further extracting information, and the evaluation indicators of LSTM and GRU would rise dramatically. This result inspired us to pretrain the A&A learning to limit the variation scope of PM2.5 for better performance and it will be one of our future concerns.

5.2. Limitations. Despite higher errors, the tendency of PM2.5 was captured by the A&A learning approach. Unlike previous studies, the approach focuses on the changing tendency of PM2.5 within the following 24 or even 48 hours. We prefer to forecast for a longer time span with an acceptable sacrifice of accuracy because it is more valuable for real-life activities.

Nevertheless, the accuracy of the prediction would be improved if the following three limitations were lifted.

First, the data of the two datasets were not gleaned at the same frequency: the atmosphere dataset was recorded every three hours, while the pollution dataset was recorded every single hour, so we used the interpolation algorithm to fill up the missing data, which probably contributed to the errors.

Second, errors could have been accumulated when the input features, features of PM2.5 in these experiments, were produced. The original data we collected about PM2.5 were not enough for predicting for 12, 24, or even 48 hours. All

Table 5: Results for various data segmentation.

Model	Dataset/data segmentation	MAE	MAPE	COR
LSTM	W/[12, 48]	**12.32**/14.37	62.74/69.47	0.34/0.35
GRU	W/[12, 48]	13.19/15.37	78.34/74.21	0.33/0.32
SEQ2SEQ	W/[12, 48]	15.83/17.74	70.96/68.42	0.43/0.39
Attention	W/[12, 48]	16.84/17.83	75.53/86.13	0.27/0.31
A&A learning	W/[12, 48]	12.53/**13.74**	**57.27/56.73**	**0.45/0.41**
LSTM	W/[48, 144]	17.32/16.57	63.25/74.46	**0.45**/0.42
GRU	W/[48, 144]	16.37/15.74	64.71/77.43	0.37/0.36
SEQ2SEQ	W/[48, 144]	19.52/18.23	76.47/82.54	0.39/0.37
Attention	W/[48, 144]	18.23/17.94	**59.85**/54.32	0.42/0.44
A&A learning	W/[48, 144]	**12.73/10.35**	65.07/**51.04**	**0.45/0.49**
LSTM	S/[12, 48]	12.42/11.91	53.29/51.17	0.47/0.56
GRU	S/[12, 48]	13.29/12.78	61.79/63.54	0.41/0.39
SEQ2SEQ	S/[12, 48]	11.89/12.31	66.92/57.97	0.33/0.41
Attention	S/[12, 48]	11.66/10.01	54.68/51.78	0.42/0.45
A&A learning	S/[12, 48]	**10.21/9.97**	**49.39/45.47**	**0.55/0.61**
LSTM	S/[48, 144]	15.34/14.41	**55.82**/54.91	0.42/0.53
GRU	S/[48, 144]	16.27/15.13	63.90/62.60	0.51/0.32
SEQ2SEQ	S/[48, 144]	19.27/17.31	66.06/63.34	0.41/0.49
Attention	S/[48, 144]	22.77/19.99	77.69/68.05	0.39/0.53
A&A learning	S/[48, 144]	**13.17/12.71**	59.31/**53.21**	**0.50/0.61**

the input features were calculated based on the previously predicted ones, through which the errors could easily accumulate and greatly damage the correlation of the predicted values and the real ones. Although we deployed windowed time steps and the decoder hidden state that contained the attention mechanism to adjust the accumulative errors, the real atmosphere and pollution status could be better simulated in an alternative way, for example, incorporating related theories into the model.

Third, as the goal of the A&A learning approach is to balance errors and forecast time spans, a specified loss function should have been deployed to fulfill this purpose, which would have helped to smooth the convergence of the model.

Table 6: Results for various models on dataset Shanghai with segmentation of [24, 72]

Method	MAE
SVM	35.5
HMM	36.3
SVR	34.3
Random forest	36.8
XGBoost	32.1
BP	30.0
MLP	31.5
RNN	27.4
Bi-LSTM [40]	23.7
RLSTM	22.9
EDSModel [39]	15.5
A&A learning	**10.2**

6. Discussion

Prediction for indexes of air pollution such as PM2.5 is of utmost importance for a healthy living atmosphere. The A&A learning has attached adequate importance to time factors and we reformed the input structure of the attention mechanism. In comparison with other attention-based methods, the performance of A&A learning on the dataset of Shanghai proves better than all methods [39] when the segmentation is [23] and [72] and the result is shown in Table 6.

On the other hand, the practical value of models for air pollution prediction task can be hindered by the limited forecast time span. For previous study mainly focusing on accuracy [41–43], the time span was restricted to one hour and the result was bound to be accurate since it is hard for air pollution values to change dramatically during one hour without specific activity of human beings. Thus, the correlation coefficient result of the predict and real value will remain high even if we let the current value as the predicted one. In this way, our research focuses on prolonging the time span and, at the same time, controlling the error under an acceptable level. The result of A&A learning shown in Section 5 was better than other commonly used models and we hope that more researches can appear in this field.

7. Conclusions

In this paper, we propose a novel A&A learning approach to predict air pollution for a longer period. As air pollution prediction holds a strong relation with the factor of "time," we add a time factor to simulate the "decay" effect of time on prediction. We had also included the decoder hidden states that can obtain the variation trend of the history and predicted data. Moreover, to make the total number of hidden states stable, we proposed a window method. As it is shown in experiments, the A&A learning approach can predict the air quality for a longer period while being more alert to recent changes and retain a good accuracy. Compared with

other models, the A&A learning performs in a quite stable and robust way under different circumstances, although it must rely on adjusted parameters to keep higher accuracy.

The following conclusions were drawn:

(1) A&A learning can prolong the time span of air pollution forecast to over 48 hours.

(2) A&A learning performs well in both datasets with a high and low level of pollution.

(3) Air pollution forecast problem is a problem of time series prediction and more emphasis on the models should be laid on time factors.

For future study, we are convinced that A&A learning can support longer air pollution prediction with the combination of the traditional method, for example, using CAMx to impose restrictions on accumulative errors of input features.

Meanwhile, this method will contribute to a longer prediction in stock market or other time-related tasks [44] and we are convinced that it would help those issues laying great emphasis on variation tendency to get a prediction with longer time span and limit the accumulative errors within acceptable levels. Also, when we use the traditional attention mechanism, the performance of A&A learning will be better than the traditional machine-learning method such as KNN and NNs on other problems [45].

Future study will focus more on practical values of NNs and thus prolong the prediction time span would be of great help. We hold the view that, for future study, more features will be considered for higher accuracy and we intend to add factors of human activities, like fireworks and traffic restrictions on private cars, to the A&A learning approach and improve the prediction accuracy or even identify the major contributor to air pollutions. Further, given the accumulated prediction results, the A&A learning has room to improve. For example, if the more accurate results of professional models, like weather forecasting models, were deployed as the input of the decoder, the accuracy of the A&A learning would be improved.

Data Availability

The data used in the study can be found at https://github.com/SkyTu/The-Attention-and-Autoencoder-Hybrid-Learning-Model. All related data can be found at http://www.cma.gov.cn/ and http://eosweb.larc.nasa.gov/cgi-bin/sse/grid.cgi?email=zhenhuawan@gmail.com.

Conflicts of Interest

The authors declare that they have no known conflicts of interest or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (61802258, 61572326, and 61702333), the Natural Science Foundation of Shanghai (18ZR1428300), and the Shanghai Committee of Science and Technology (17070502800).

References

[1] K. Balakrishnan, S. Dey, T. Gupta et al., "The impact of air pollution on deaths disease burden and life expectancy across the states of India: the global burden of disease study 2017," *The Lancet Planetary Health*, vol. 3, no. 1, pp. e26–e39, 2019.

[2] L. Li, J. Qian, C.-Q. Ou, Y.-X. Zhou, C. Guo, and Y. Guo, "Spatial and temporal analysis of air pollution index and its timescale-dependent relationship with meteorological factors in Guangzhou China 2001-2011," *Environmental Pollution*, vol. 190, pp. 75–81, 2014.

[3] D. C. Carslaw and K. Ropkins, "Openair-an R package for air quality data analysis," *Environmental Modelling & Software*, vol. 27, pp. 52–61, 2011.

[4] F. A. Gers, S. Jürgen, and F. Cummins, "Learning to forget: continual prediction with LSTM," pp. 850–855, 1999.

[5] T. C. Bui, V. D. Le, and S. K. Cha, "A deep learning approach for forecasting air pollution in South Korea using LSTM," 2018, https://arxiv.org/abs/1804.07891.

[6] R. Dey and M. Fathi, "Gate-variants of gated recurrent unit (GRU) neural networks," in *Proceedings of the 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, Boston, MA, USA, August 2017.

[7] J. Becerra-Rico, "Airborne particle pollution predictive model using Gated Recurrent Unit (GRU) deep neural networks," *Earth Science Informatics*, vol. 13, no. 3, pp. 821–834, 2020.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, June 2016.

[9] B. Zhang, H. Zhang, G. Zhao, and J. Lian, "Constructing a PM2.5 concentration prediction model by combining auto-encoder with Bi-LSTM neural networks," *Environmental Modelling & Software*, vol. 124, Article ID 104600, 2020.

[10] C. Zhang, L. Di, Z. Sun, L. Lin, E. G. Yu, and J. Gaigalas, "Exploring cloud-based web processing service: a case study on the implementation of CMAQ as a service," *Environmental Modelling & Software*, vol. 113, pp. 29–41, 2019.

[11] S. Ma, X. Zhang, C. Gao et al., "Multimodel simulations of a springtime dust storm over northeastern China: implications of an evaluation of four commonly used air quality models (CMAQ v5.2.1 CAMx v6.50 CHIMERE v2017r4 and WRF-Chem v3.9.1)," *Geoscientific Model Development*, vol. 12, no. 11, pp. 4603–4625, 2019.

[12] ENVIRON International Corporation, *User's Guide to the Comprehensive Air Quality Model with Extensions Version 5.40*, ENVIRON International Corporation, Novato CA, 2014.

[13] Y. Hu, M. Talat Odman, and A. G. Russell, "Mass conservation in the community multiscale air quality model," *Atmospheric Environment*, vol. 40, no. 7, pp. 1199–1204, 2006.

[14] H. Xiao, Q. Wu, X. Yang, L. Wang, and H. Cheng, "Numerical study of the effects of initial conditions and emissions on PM2.5 concentration simulations with CAMx v6.1: a Xi'an case study," *Geoscientific Model Development*, vol. 14, no. 1, pp. 223–238, 2021.

[15] Y. Zhang, B. Pun, S.-Y. Wu, K. Vijayaraghavan, and C. Seigneur, "Application and evaluation of two air quality models for particulate matter for a southeastern U.S. Episode," *Journal of the Air & Waste Management Association*, vol. 54, no. 12, pp. 1478–1493, 2004.

[16] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, Cambridge, MA, USA, 2012.
[17] W. C. Leong, R. O. Kelani, and Z. Ahmad, "Prediction of air pollution index (API) using support vector machine (SVM)," *Journal of Environmental Chemical Engineering*, vol. 8, no. 3, Article ID 103208, 2020.
[18] J. Fan, L. Wu, X. Ma, H. Zhou, and F. Zhang, "Hybrid support vector machines with heuristic algorithms for prediction of daily diffuse solar radiation in air-polluted regions," *Renewable Energy*, vol. 145, pp. 2034–2045, 2020.
[19] C. Wen, S. Liu, X. Yao et al., "A novel spatiotemporal convolutional long short-term neural network for air pollution prediction," *Science of the Total Environment*, vol. 654, pp. 1091–1099, 2019.
[20] H. Chang-Hoi, I. Park, H. R. Oh et al., "Development of a PM2 5 prediction model using a recurrent neural network algorithm for the Seoul metropolitan area Republic of Korea," *Atmospheric Environment*, vol. 245, 2020.
[21] C.-J. Huang and P.-H. Kuo, "A deep CNN-LSTM model for particulate matter (PM2.5) forecasting in smart cities," *Sensors*, vol. 18, no. 7, p. 2220, 2018.
[22] T. Gangopadhyay et al., "Temporal attention and stacked LSTMs for multivariate time series prediction," 2018, https://arxiv.org/abs/1809.04206.
[23] Y. Chen, J. M. An, and A. Yanhan, "A novel prediction model of PM2.5 mass concentration based on back propagation neural network algorithm," *Journal of Intelligent and Fuzzy Systems*, vol. 37, no. 7, pp. 1–9, 2019.
[24] V. Athira, P. Geetha, R. Vinayakumar, K. P. Soman et al., "Deepairnet: applying recurrent networks for air quality prediction," *Procedia Computer Science*, vol. 132, pp. 1394–1403, 2018.
[25] X. Feng, Q. Li, Y. Zhu, J. Hou, L. Jin, and J. Wang, "Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation," *Atmospheric Environment*, vol. 107, pp. 118–128, 2015.
[26] Y. Zhang, "Dynamic effect analysis of meteorological conditions on air pollution: a case study from Beijing," *Science of the Total Environment*, vol. 684, pp. 178–185, 2019.
[27] I. Sutskever, V. Oriol, and V. Quoc, "Sequence to sequence learning with neural networks," 2014, https://arxiv.org/abs/1409.3215.
[28] D. R. Liu, S. J. Lee, Y. Huang, and C. J. Chiu, "Air pollution forecasting based on attention-based LSTM neural network and ensemble learning," *Expert Systems*, vol. 37, Article ID e12511, 2020.
[29] R. Hecht-Nielsen, *'Theory of the Backpropagation Neural network*, Academic Press, Cambridge, MA, USA, 1992.
[30] I. H. Fong, "Predicting concentration levels of air pollutants by transfer learning and recurrent neural network," *Knowledge-Based Systems*, vol. 192, no. 15, pp. 105622.1–105622.10, 2020.
[31] X. Yin, J. A. N. Goudriaan, E. A. Lantinga, J. A. N. Vos, and H. J. Spiertz, "A flexible sigmoid function of determinate growth," *Annals of Botany*, vol. 91, no. 3, pp. 361–371, 2003.
[32] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," 2017, https://arxiv.org/abs/1706.03762.
[33] M. Cai, Y. Yin, and M. Xie, "Prediction of hourly air pollutant concentrations near urban arterials using artificial neural network approach," *Transportation Research Part D: Transport and Environment*, vol. 14, no. 1, pp. 32–41, 2009.
[34] B. Liu, S. Yan, J. Li et al., "A sequence-to-sequence air quality predictor based on the n-step recurrent prediction," *IEEE Access*, vol. 7, pp. 43331–43345, 2019.
[35] R. A. Dunne and N. A. Campbell, "On the pairing of the softmax activation and cross-entropy penalty functions and the derivation of the softmax activation function," in *Proceedings of the 8th Aust. Conf. on the Neural Networks*, Melbourne, Australia, February 1997.
[36] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," 2015, https://arxiv.org/abs/1506.07503.
[37] X. F. Li, M. J. Zhang, S. J. Wang, A. F. Zhao, and Q. Ma, "Variation characteristics and influencing factors of air pollution index in China," *Huan Jing Ke Xue = Huanjing Kexue*, vol. 33, no. 6, p. 1936, 2012.
[38] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition*, vol. 38, no. 12, pp. 2270–2285, 2005.
[39] B. Zhang, G Zou, D Qin, Y Lu, Y Jin, and H Wang, "A novel Encoder-Decoder model based on read-first LSTM for air pollutant prediction," *The Science of the Total Environment*, vol. 765, no. 3, Article ID 144507, 2021.
[40] B. Zhang, "Constructing a PM2.5 concentration prediction model by combining auto-encoder with Bi-LSTM neural networks," *Environmental Modelling and Software*, vol. 124, Article ID 104600, 2019.
[41] Y. Bai, Y. Li, B. Zeng, C. Li, and J. Zhang, "Hourly PM2.5 concentration forecast using stacked autoencoder model with emphasis on seasonality," *Journal of Cleaner Production*, vol. 224, pp. 739–750, 2019.
[42] J. Zhu, P. Wu, H. Chen, L. Zhou, and Z. Tao, "A hybrid forecasting approach to air quality time series based on endpoint condition and combined forecasting model," *International Journal of Environmental Research and Public Health*, vol. 15, no. 9, p. 1941, 2018.
[43] M. Đurić and D. Vujović, "Short-term forecasting of air pollution index in Belgrade Serbia," *Meteorological Applications*, vol. 27, p. 5, 2020.
[44] A. Glowacz, "Fault diagnosis of electric impact drills using thermal imaging," *Measurement*, vol. 171, Article ID 108815, 2021.
[45] Y. Zhang, B. Chen, G. Pan, and Y. Zhao, "A novel hybrid model based on VMD-WT and PCA-BP-RBF neural network for short-term wind speed forecasting," *Energy Conversion and Management*, vol. 195, pp. 180–197, 2019.
[46] J. Chen, J. Lu, J. C. Avise, J. A. DaMassa, M. J. Kleeman, and A. P. Kaduwela, "Seasonal modeling of PM2.5 in California's san joaquin valley," *Atmospheric Environment*, vol. 92, pp. 182–190, 2014.