*mathematics*

MDPI

*Article*

# ResInformer: Residual Transformer-Based Artificial Time-Series Forecasting Model for PM2.5 Concentration in Three Major Chinese Cities

**Mohammed A. A. Al-qaness [1,*], Abdelghani Dahou [2], Ahmed A. Ewees [3], Laith Abualigah [4,5,6], Jianzhu Huai [7], Mohamed Abd Elaziz [8,9,10], Ahmed M. Helmi [11,12]**

1 College of Physics and Electronic Information Engineering, Zhejiang Normal University, Jinhua 321004, China
2 Mathematics and Computer Science Department, University of Ahmed DRAIA, Adrar 01000, Algeria
3 Department of Computer, Damietta University, Damietta 34517, Egypt
4 Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman 19328, Jordan
5 Faculty of Information Technology, Middle East University, Amman 11831, Jordan
6 Applied Science Research Center, Applied Science Private University, Amman 11931, Jordan
7 State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China
8 Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt
9 Artificial Intelligence Research Center (AIRC), Ajman University, Ajman 346, United Arab Emirates
10 Department of Electrical and Computer Engineering, Lebanese American University, Byblos 13-5053, Lebanon
11 Department of Computer Engineering, College of Engineering and Information Technology, Buraydah Private Colleges, Buraydah 51418, Saudi Arabia
12 Department of Computer and Systems Engineering, Faculty of Engineering, Zagazig University, Zagazig 44519, Egypt
* Correspondence: alqaness@zjnu.edu.cn



**Citation:** Al-qaness, M.A.A.; Dahou, A.; Ewees, A.A.; Abualigah, L.; Huai, J.; Abd Elaziz, M.; Helmi, A.M. ResInformer: Residual Transformer-Based Artificial Time-Series Forecasting Model for PM2.5 Concentration in Three Major Chinese Cities. *Mathematics* **2023**, *11*, 476. https://doi.org/10.3390/math11020476

Academic Editors: Chaman Verma, Maria Simona Raboaca, Zoltán Illés and Alexander B. Medvinsky

Received: 28 November 2022
Revised: 28 December 2022
Accepted: 11 January 2023
Published: 16 January 2023



**Abstract:** Many Chinese cities have severe air pollution due to the rapid development of the Chinese economy, urbanization, and industrialization. Particulate matter (PM2.5) is a significant component of air pollutants. It is related to cardiopulmonary and other systemic diseases because of its ability to penetrate the human respiratory system. Forecasting air PM2.5 is a critical task that helps governments and local authorities to make necessary plans and actions. Thus, in the current study, we develop a new deep learning approach to forecast the concentration of PM2.5 in three major cities in China, Beijing, Shijiazhuang, and Wuhan. The developed model is based on the Informer architecture, where the attention distillation block is improved with a residual block-inspired structure from efficient networks, and we named the model ResInformer. We use air quality index datasets that cover 98 months collected from 1 January 2014 to 17 February 2022 to train and test the model. We also test the proposed model for 20 months. The evaluation outcomes show that the ResInformer and ResInformerStack perform better than the original model and yield better forecasting results. This study's methodology is easily adapted for similar efforts of fast computational modeling.

**Keywords:** air pollution; PM2.5; deep learning; time series; forecasting

**MSC:** 68T07

## 1. Introduction

Air quality has attracted public attention in recent years [1,2], with a particular focus on pollution, specifically PM2.5 and PM10. As we know, PM2.5 and PM10 can worsen air pollution and influence public health [3,4]. Hence, it is essential to precisely forecast PM2.5 and PM10 to design proper deterrents and management criteria for enhancing air quality. Researchers have used a variety of techniques to determine the concentration of

atmospheric pollutants. Because of the significant adverse effects of air pollution on human health, research into air quality and pollution is critical for public health [5]. Aside from the well-known public health concerns created by air pollution, there are also illnesses related to air pollution throughout the world [6,7].

Air pollution is the greatest issue in many parts of the world due to development and industry [8,9]. Air pollution affects human health and has significant environmental and ecological consequences. The most severe concern in most world cities is airborne fine particulate matter (PM) with an aerodynamic diameter of PM2.5, or fewer [10]. The inhaled PM is absorbed by the respiratory tract, depending on its size. A recent World Health Organization (WHO) study found a link between PM intensity and 7 million worldwide fatalities per year. These fatalities were linked to stroke (33%), ischemic heart disease (IHD) (36%), ALRI (8%), lung cancer (LC) (6%), and COPD (17%). PM2.5 has the most severe documented effects on human health. Almost 87 percent of the world's population lives in areas where pollution levels exceed the WHO recommendations. In medium- and low-income nations, almost 90% of the population is exposed to hazardous concentrations of pollutants [11,12].

Forecasting is separated into two types: deterministic and probabilistic [13]. The outcomes of deterministic methods can yield a point forecasting result. The general public can act immediately based on the obtained results. However, deterministic forecasting methods are only sometimes accurate. The population must estimate the forecasting inaccuracy to reduce air pollution, which is impractical. Probabilistic forecasting methods can assess uncertainty within the error of deterministic forecasting and provide a prediction interval. The prediction interval includes the actual PM2.5 values with a high confidence level. With the probabilistic forecasting data, people can more readily create schedules to reduce air pollution [14].

*Main Objectives and Contributions*

Many Chinese cities have had severe air pollution in recent years due to the rapid development of the economy, industry, the vast number of vehicles and their pollution, extensive coal consumption, and vehicular exhaust [15,16]. In addition, metal emissions from anthropogenic sources have increased dramatically [17]. In the literature, many studies have been implemented in China to study the relation between PM2.5 and several known diseases, such as lung cancer [18], stillbirth [19], oral clefts [20], mouth, hand, and foot disease [21], preterm birth [22], and others [23,24].

China has created robust action plans for reducing air pollution, in which PM2.5 has received the highest priority due to its impacts on air quality [25]. To this end, building an intelligence-based time-series forecasting tool for air quality is necessary. Accordingly, this paper studies the forecasting of PM2.5 in three major cities in China: Beijing, Wuhan, and Shijiazhuang. The air quality forecasting and analysis of the mentioned cities have been addressed by many studies using different tools. For example, ref. [25] compared the air quality between Beijing and Los Angeles in terms of PM2.5. They found that the concentration of PM2.5 in Beijing has decreased in recent years. The authors of [26] studied the impacts of PM2.5 on the acute exacerbation of chronic obstructive pulmonary disease (AECOPD) using the records of the concentration of PM2.5 for Bejing for an extended period (2010–2019). They employed the Kolmogorov–Zurbenko filter method on the collected datasets using short-term and long-term scenarios. They found that the concentration of PM2.5 had a positive association with the AECOPD hospitalization risk. Yang et al. [27] suggested a forecasting tool using CNN, LSTM, and a combined CNN–LSTM for Beijing's PM2.5 concentrations. Zhang et al. [2] used the grey multivariable convolution model to predict the concentration of PM10 and PM2.5 in Shijiazhuang.

This paper presents a novel air pollution forecasting-based deep learning method. In the proposed method, an improved attention distillation operation is proposed to replace the original operation placed after the multi-head ProbSparse self-attention block in the Informer [28] model to boost the model performance and extract more relevant and meaningful features from the input sequences in the encoder block. Inspired by the inverted residual block in the neural networks such as MobileNetV3 [29], we replace the attention distillation in the Informer model with a residual-based structure. Furthermore, we implement two versions of the Informer model using the residual structure instead of the canonical convolution block, and we name these ResInformer and ResInformerStack. We compare the proposed version with the original informer model on different PM2.5 forecasting datasets. In short, our main contributions can be summarized as follows:

- Inspired by the Informer [28] model, we develop a new time-series forecasting model as a transformer-based scheme that shows significant prediction performance.
- We replace the self-attention distillation operation of the traditional Informer with a residual block for capturing dominating attention on the encoder side.
- We use new versions of the Informer and the developed ResInformer called InformerStack and ResInformerStack using the residual structure instead of the canonical convolution block. The four models are extensively compared using different performance indicators.
- We evaluate the mentioned models using PM2.5 datasets collected from three major Chinese cities, Beijing, Shijiazhuang, and Wuhan.

This paper is organized into the following sections. Section 2 presents the related works that use deep learning methods for air pollution forecasting. Section 3 shows the proposed deep learning method for air pollution forecasting. In Section 4, the experiments and results are detailed. Finally, the conclusions and future work directions are provided in Section 5.

## 2. Literature Review

Prior studies on the green economy can be divided into national, regional, industrial, and city [30,31].

Because the Community Radiative Transfer Model is oriented to the GOCART scheme [32], a novel LiDAR assimilation technique for the MOSAIC system was established in [32], based on the Community Radiative Transfer Model and 3DVAR methods, in order to enhance the prediction of PM2.5 3D distribution. The suggested model produced better results than the conventional models in the literature. Despite much research on PM2.5, the problem of gradient disappearance and the representative samples of wavelet orders and layers still needs to be solved. A new model based on a wavelet transform-stacked autoencoder was suggested in [33]. The conclusion was that such a unique approach may assist in improving the accuracy of the PM2.5 forecast.

An integrated approach of gated recurrent unit deep net based on observable mode decomposition (EMD-GRU) was proposed in [34] for forecasting the PM2.5 concentration. This approach first decomposed the PM2.5 concentration sequence using a decomposition model. It then fed the various stationary subsequences formed after the decomposition. The forecast results of the instance presented in this study demonstrated that the EMD-GRU model decreased the RMSE and SMAPE compared to the GRU model.

To examine the effects of COVID-19 on fine particulate matter (PM2.5) levels, a conditional variational autoencoder system was created in [35] based on deep learning to detect PM2.5 anomalies in Chinese cities during the COVID-19 outbreak. The authors demonstrated that the timing of the variation in the cities with uncontrolled PM2.5 anomalies corresponded to the WHO's responses to COVID-19. A Long Short-Term Memory-based deep neural network was used in [36] to anticipate the optimum PM2.5 concentrations. Compared to traditional multilayer neural networks, deep learning approaches have considerable advantages.

A unique annual nonlinear grey paradigm was initially constructed in [37] to solve such concerns by merging the climate sensitivity factor, the traditional Weibull Bernoulli grey model, and the cultural algorithm, which depicted the periodicity and nonlinearity of the source data concurrently. The suggested model provided early warning data to policymakers in order for them to build PM2.5 mitigation plans. Based on Sentinel–5Ps and ground-station observed data, researchers evaluated the shift in atmospheric pollutants over a partial to total lockdown time in 2020 [38]. The results showed that the average tropospheric NO2 level reduced significantly in 2020 due to the lockdown compared to the previous year.

Further research examined the reasons for these unexpected incidents [39]. The PM2.5 data were collected from surveillance systems installed as part of the Airbox project in Taichung. Short-term forecast reaction capability may increase significantly in the future. Ref. [40] provided a cellular automata (CA) system based on a multiple regression analysis model and many schematics to evaluate the formation and dispersion of PM2.5. The percent of the forecast errors in this experiment was almost smaller than 20 $g/m^3$.

The research in [41] investigated reliable PM2.5 prediction, which may help to reduce or avoid detrimental outcomes. The suggested method was unique because it used a genetic algorithm (GA) and an encoder–decoder (E–D) model to forecast PM2.5. The suggested strategy enhanced the accuracy by at least 13.7 percent by merging the GA-based feature extraction technique and the E–D model. Using various clustering algorithms, the authors in [42] demonstrated that it was feasible to correctly estimate acceptable PM2.5 concentrations with minimum computing time. The data were collected using an Internet of Things infrastructure comprised of Airbox devices for PM2.5 monitoring. A final comparison study was performed for several clustering algorithms regarding efficiency and computing time.

Another study proposed a novel picture-based predictor of PM2.5 concentration, which used photographs taken with mobile phones or cameras to estimate PM2.5 concentration in real time [43]. Naturalness statistics models were developed on entropy characteristics in spatial and transform domains using a large body of images recorded under favorable weather, i.e., poor PM2.5 concentration. In terms of the prediction accuracy and deployment efficiency, adequate experimental findings demonstrated the proposed model's advantage over current relevant state-of-the-art forecasts.

Another paper suggested a deep neural system PM2.5 prediction model that captured the temporal variations of ground-level PM2.5 by combining remote sensing atmospheric aerosol depth information from the Himawari-8 satellite with traditional meteorological information [44]. The suggested PM2.5 model considerably increased the accuracy of the PM2.5 estimate and offered a new viewpoint for PM2.5 monitoring by utilizing an end-to-end classification technique.

## 3. Proposed Deep Learning Model

Recent research has witnessed the development of deep learning models for long sequence time-series forecasting (LSTF) problems, such as Informer [28]. The Informer model is transformer-based architecture known for its superiority in forecasting long sequences compared to other models such as ARIMA, DeepAR, CNN, and LSTM. The Informer models are widely used in time-series forecasting, relying on sparse attention transformers to extract long-term temporal dependencies. In this section, we introduce an improved architecture of the Informer by replacing the self-attention distillation operation with a residual block to capture dominating attention on the encoder side. The residual block is a generalization of the residual connection inspired by the inverted residual block in the MobileNetV3 [29] architecture. The proposed architecture of the ResInformer is shown in Figure 1.

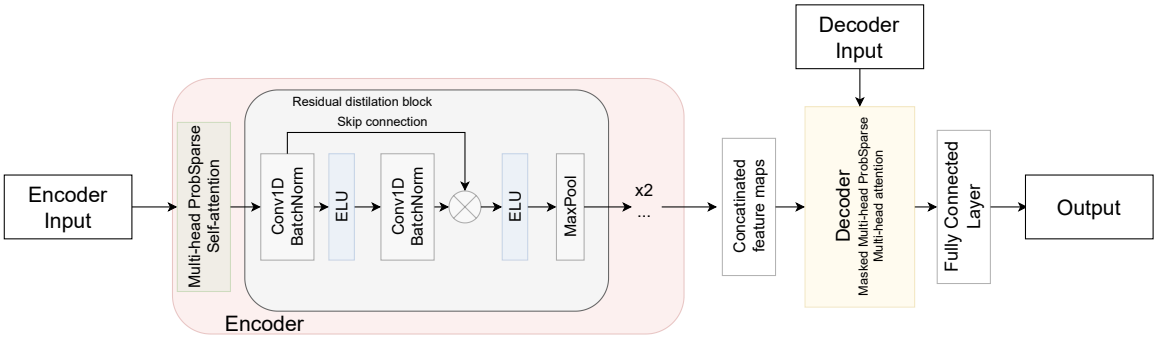



**Figure 1.** The proposed ResInformer architecture.

The main components of the Informer rely on the following parts: the self-attention mechanism, ProbSparse, which lowers the time complexity, the self-attention distilling operation, which reduces overheads (process longer inputs), and the batch sequence prediction, which is based on the generative style decoder. The ProbSparse self-attention mechanism was proposed to tackle the time and memory complexity in attention-based models when dealing with long input sequences. The ProbSparse attention is based on the KL-divergence to create a sparse attention version that can achieve $O(TlogT)$ complexity compared to the self-attention mechanism in a vanilla Transformer model. Compared to the canonical attention [45], the ProbSparse attention only selects a subset $u$ of dominant queries based on their variance (the largest variance is selected) over all keys. The sparse query matrix defined as $\bar{Q} \in \mathbb{R}^{L_Q \times d}$ replaces the old query in the ProbSparse attention, which consists only of the subset of Top-$u$ queries, and $d$ represents the corresponding input dimension. The ProbSparse attention is defined in Equation (1),

$$ATT(\bar{Q}, K, V) = Softmax(\frac{\bar{Q}K^t}{\sqrt{d}})V. \quad (1)$$

As shown in Figure 1, to perform information distillation, the Informer uses the multi-head ProbSparse Self-Attention Blocks followed by the self-attention distillation operation in a pyramid structure to distill the redundant combinations of value $V$. The operation sequences proposed to replace the actual distillation process are listed in Algorithm 1. The original Informer starts with a ProbSparse self-attention applied on the hidden representation $h_i$ from the $i^{th}$ block generating the $Q, K$, and $V$. Later, the self-attention distillation operation is applied using the dilated convolution (Conv1d) layers inspired by [46] and defined as in Equation (2),

$$X^t_{j+1} = MaxPool(ELU(Conv1D([X^t_j]_{AB}))), \quad (2)$$

where $j$ represents the $j^{th}$, $[\cdot]_{AB}$ represents the operations of the ProbSparse self-attention, and Conv1D, MaxPool, and ELU are the 1D convolution operation with the kernel size $K = 3$, the max-pooling operation to downsample $X^t$ to half with a stride equal to two at each block, and the ELU activation function, respectively [47].

In the proposed ResInformer, the self-attention distillation operation is proposed as listed in Algorithm 1. The model takes as input the hidden representation $h_i$ from the $i^{th}$ block and produces the $h_{i+1}$ representation for the $i + 1^{th}$ block.

The batch normalization is used to normalize the inputs used in Transformer architectures across the $C$ number of channels of the generated output in a 1D array of length $L$ from the Conv-1D layer over $N$ batches.

**Algorithm 1** ResInformer ProbSparse Self-Attention Block.

**Require:** $h_i$
**Ensure:** $h_{i+1}$
$residual \leftarrow h_i$
$h_{i+1} \leftarrow ATT(h_i, h_i)$
$h_{i+1} \leftarrow ELU(BatchNorm(Conv1D(h_{i+1})))$
$h_{i+1} \leftarrow MaxPool(ELU(BatchNorm(Conv1D(h_{i+1})) + residual))$

After concatenating the representation of the output resulting from the encoder block in a final hidden representation $h_L$, the decoder block is applied to generate the output sequence $y_L$, which is built based on the standard decoder architecture from [45]. The decoder block comprises a multi-head self-attention mechanism and a fully connected layer to generate the output sequence. The decoder uses a generative style to predict the next time step $y_{t+1}$, based on the previous time step $y_t$ and the currently hidden representation $h_t$. Concerning the loss function, we used the mean square error (MSE) loss function as in Equation (3),

$$loss_{mse}(y, \hat{y}) = \frac{1}{M} \sum_{t=1}^{n} \|y^{(t)} - \hat{y}^{(t)}\|_2^2, \tag{3}$$

where $\hat{y} \in \mathbb{R}^{M \times n}$ is the predicted output, and $Y$ is the ground truth.

Table 1 lists the main characteristics of the models used in our study.

**Table 1.** The models' characteristics.

| Characteristics | Informer | InformerStack | ResFormer | ResFormerStack |
|---|---|---|---|---|
| Transformer-based | ✓ | ✓ | ✓ | ✓ |
| ProbSparse self-attention | ✓ | ✓ | ✓ | ✓ |
| Self-attention distilling | ✓ | ✓ | ✗ | ✗ |
| Generative style decoder | ✓ | ✓ | ✓ | ✓ |
| Encoders/decoders | 1/1 | 2/1 | 1/1 | 2/1 |
| Residual distilation block | ✗ | ✗ | ✓ | ✓ |

## 4. Study Areas

Beijing, the capital of the People's Republic of China, is situated at the north tip of the North China Plain, with mountains to its north and west, having about 21,886,000 residents in a total area of 16,410 km$^2$, with a population density of 1333 person/km. Its urban population makes it the second-largest Chinese city.

It is also the nation's cultural, educational, and political center. Its climate is humid continental, influenced by the monsoon, with hot humid summers and cold dry winters. The city used to have poor air quality, especially in winter, caused by several factors, including car exhaust emissions, manufacturing in surrounding regions, and coal burning.

Shijiazhuang, the capital city of Hebei province, is situated at the west edge of the North China Plain and the east foot of the Taihang Mountains, about 266 km southwest of Beijing. Its climate is semi-arid continental, influenced by the monsoon, with hot humid summers and cold dry winters. It has a total area of 14,530 km$^2$, with an estimated resident population of 11,204,700 in 2021. The population density of Shijiazhuang is 771 person/km. It is a major industrial city in North China, with a state-level industrial zone. Its poor air quality ranked as high for air pollution in the nation and the world.

Wuhan, the capital city of Hubei Province, is situated at the confluence of the Yangtze River and the Han river, having a total area of 8483 km$^2$, with an estimated resident population of 13,648,900 in 2021. The population density of Wuhan is 1608 person/km. Its urban area consists of three towns, Wuchang, Hankou, and Hanyang, with 25% covered by water. Its climate is humid subtropical, with abundant rainfall in summer and four distinct seasons. It has a population of more than 11 million, making it the most populous city in

Central China. It is also an industrial hub in Central China, with three state-level industrial zones and over 350 research institutes. Figure 2 shows the geographical position of the three cities.

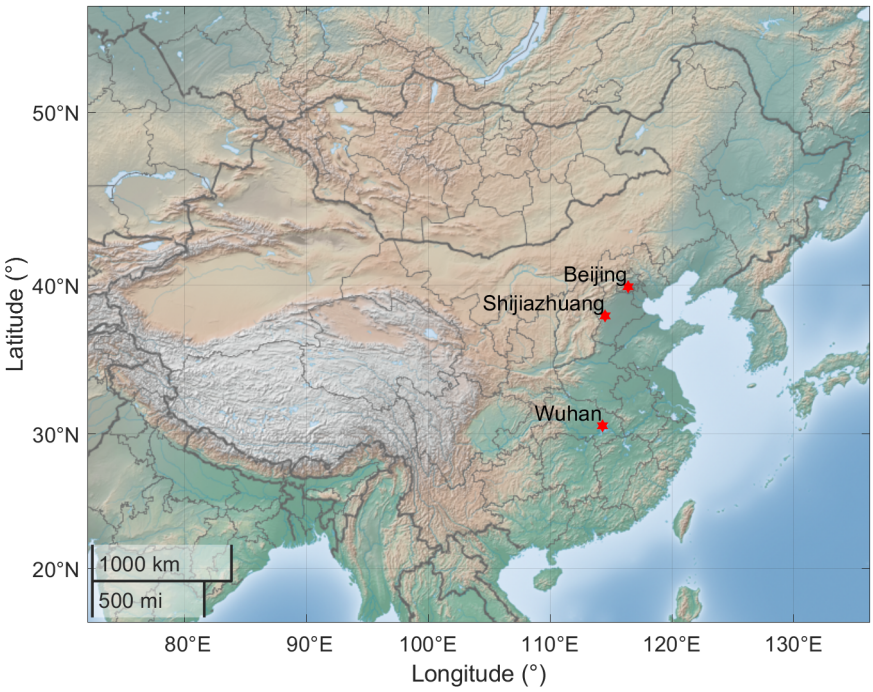



**Figure 2.** The study areas.

## 5. Experiments

### *5.1. Experimental Setup*

The ADAM [48] optimizer with an initial learning rate of $1E-4$ was used to train the models using a batch size of 16 for 50 epochs. The training process was stopped early within ten epochs. All the experiments were repeated ten times, implemented in PyTorch, and conducted on a single NVIDIA GTX1080 GPU with 8 GB of RAM. The ResInformer contained two encoder layers and one decoder layer with eight attention heads. The same configurations were set for the Informer, InformerStack, and ResInformer, with prediction windows equal to one day based on the datasets' time stamps. The InformerStack stacked the encoder layers with the following order 3,2,1. All models used an input sequence length of the encoder equal to 96 and a start token (label length) for the decoder equal to 48. The encoder and decoder input sizes were equal to seven.

Multiple evaluation metrics were used to evaluate the performance of the models, including the mean square error (MSE), the mean absolute error (MAE), the root mean square error (RMSE), the mean squared prediction error (MSPE), the mean absolute percentage error (MAPE), and the coefficient of determination ($R^2$). The following equations define the evaluation metrics used in the experiments.

$$MSE = \frac{1}{n}\sum_{t=1}^{n} \|y^{(t)} - \hat{y}^{(t)}\|_2^2; \tag{4}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n} \|y^{(t)} - \hat{y}^{(t)}\|_1; \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} \|y^{(t)} - \hat{y}^{(t)}\|_2^2}; \tag{6}$$

$$MSPE = \frac{1}{n}\sum_{t=1}^{n} \|y^{(t)} - \hat{y}^{(t)}\|_1 / y^{(t)}; \tag{7}$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n} \|y^{(t)} - \hat{y}^{(t)}\|_1 / \hat{y}^{(t)}; \tag{8}$$

$$R^2 = 1 - \frac{\sum_{t=1}^{n} \|y^{(t)} - \hat{y}^{(t)}\|_2^2}{\sum_{t=1}^{n} \|y^{(t)} - \bar{y}\|_2^2}, \tag{9}$$

where $\bar{y}$ is the mean of the ground truth values $y$, while $\hat{y}$ represents the predicted values.

### 5.2. Dataset Description

We obtained the air quality index datasets from the Worldwide COVID-19 Air Quality datasets that are publicly available online at "Air Quality Open Data Platform (https://aqicn.org/data-platform/covid19/, accessed on 1 March 2022)". As mentioned above, we selected three major Chinese cities, Beijing, Shijiazhuang, and Wuhan. The data were divided into three parts: training, validation, and testing sets with a ratio of 70:10:20. Concerning data preprocessing, all datasets were prepared with data standardization as defined in Equation (10) to boost the models' performance and to reduce the variance in the data. For the Beijing data, the dates ranged from 1 January 2014 to 17 February 2022. For the Wuhan data, the dates ranged from 1 January 2014 to 26 October 2021. For the Shijiazhuang data, the dates ranged from 21 February 2014 to 9 February 2022.

$$\tilde{x} = \frac{x - mean}{std}, \tag{10}$$

where *mean* is the mean of the training samples, and *std* is the standard deviation of the training samples.

### 5.3. Results

We compared the proposed ResInformer and its variant ResInformerStack to the aforementioned models, Informer and InformerStack.

Table 2 illustrates the evaluation results for the four models using five evaluation indicators, MAE, MSE, $R^2$, RMSE, MAPE, and MSPE.

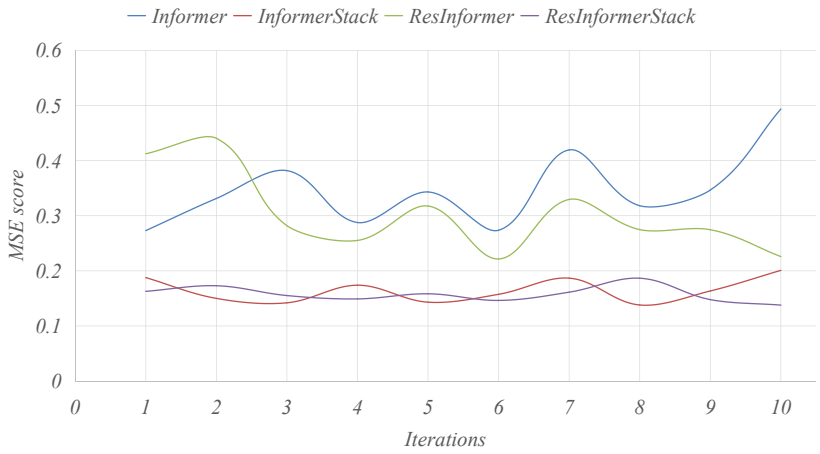
In the case of Beijing, it is clear that the ResInformerStack obtained the best values in terms of all performance indicators. The InformerStack model recorded the second rank in terms of the MSE, MAE, $R^2$, and RMSE. The Informer obtained the second rank in terms of the MAPE and MSPE.

In the case of Wuhan, the ResInformerStack model recorded the best values in terms of the MSE, $R^2$, RMSE, MAPE, and MSPE. In terms of the MAE, the best result was obtained by the InformerStack model. Moreover, the InformerStack recorded the second rank in several indicators, such as the MSE, $R^2$, RMSE, and MSPE. The Informer model recorded the second rank in terms of the MAPE.

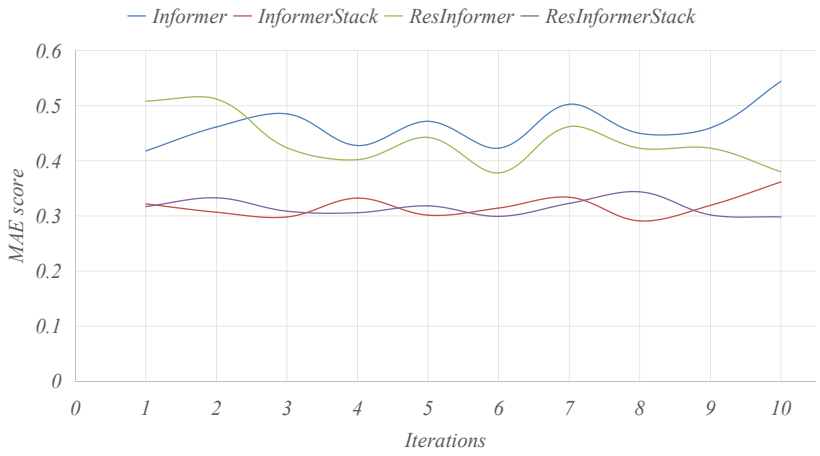
For the Shijiazhuang data, the ResInformer obtained the best MSE value, followed by the Informer, ResInformerStack, and the ResInformer, respectively. The Informer model obtained the best MAE value, followed by the ResInformerStack, InformerStack, and the ResInformer. In terms of the $R^2$, the best results were obtained by the Informer model, followed by the InformerStack, ResInformer, and the ResInformerStack, respectively. The ResInformer recorded the best RMSE value, followed by the Informer, ResInformerStck, and the InformerStack, respectively. Moreover, the Informer model obtained the first rank in terms of the MSPE, followed by the ResInformerStack, ResInformer, and the InformerStack, respectively.

For further analysis, Figures 3–5 show the prediction results of all the compared methods on the testing sets in terms of the MAE, MSE, and $R^2$, for Beijing, Wuhan, and Shijiazhuang, respectively. Here also, the ResInformerStack obtained the best results, except for

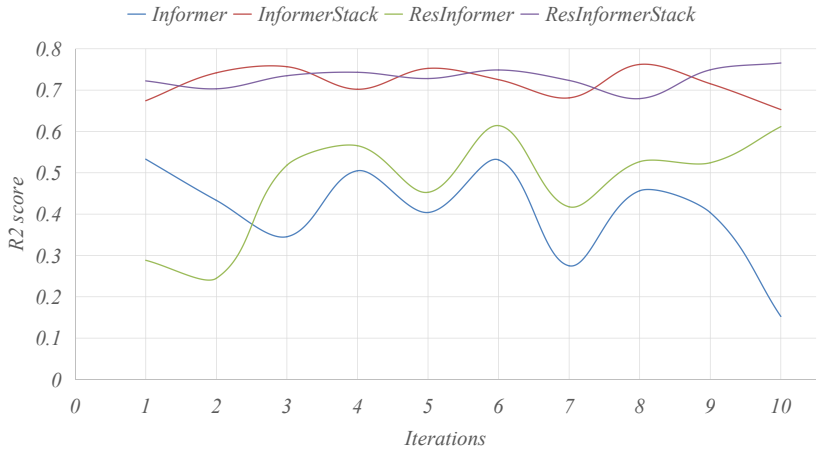
the data of Shijiazhuang city, in which the Informer model obtained the best results in terms of the MAE and $R^2$, whereas the ResInformer obtained the best MSE value.




(**a**)

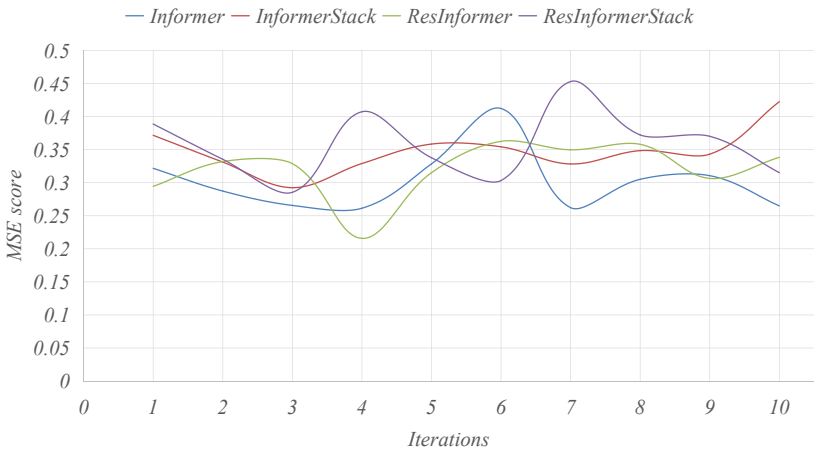



(**b**)



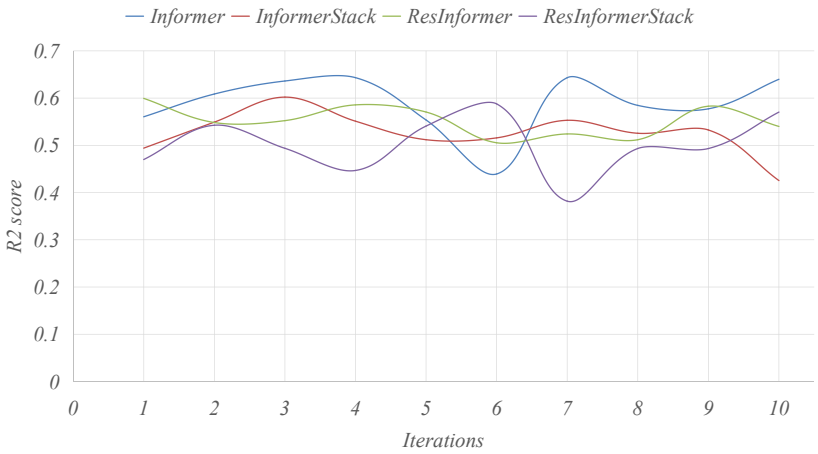

(**c**)

**Figure 3.** Wuhan PM2.5 prediction scores. (**a**) MSE score for testing data predictions; (**b**) MAE score for testing data predictions; (**c**) $R^2$ score for testing data predictions.



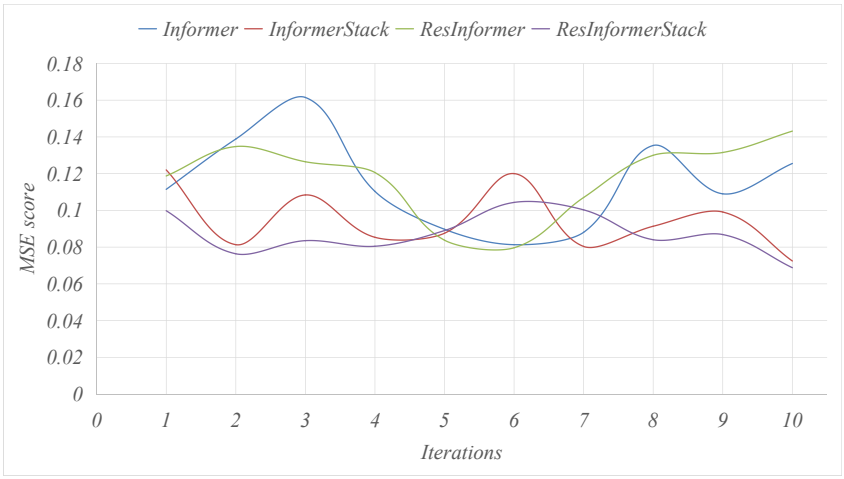

(**a**)


(**b**)



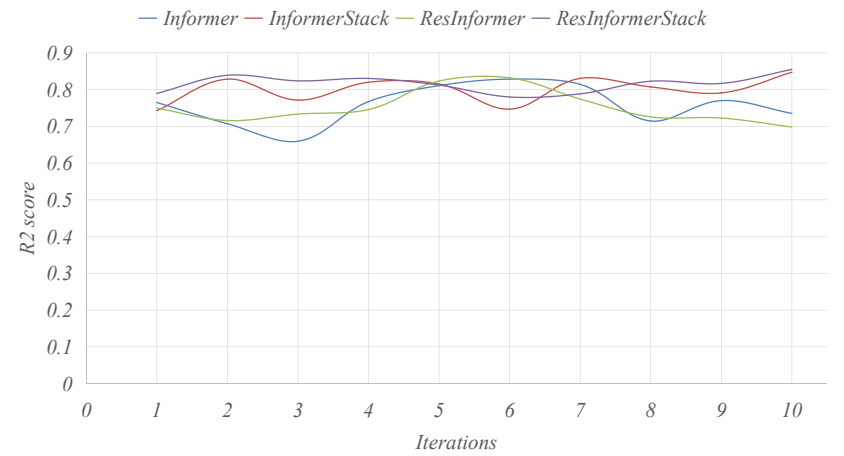

(**c**)

**Figure 4.** Shijiazhuang PM2.5 prediction scores. (**a**) MSE score for testing data predictions; (**b**) MAE score for testing data predictions; (**c**) $R^2$ score for testing data predictions.




(**a**)


(**b**)




(**c**)

**Figure 5.** Beijing PM2.5 prediction scores. (**a**) MSE score for testing data predictions; (**b**) MAE score for testing data predictions; (**c**) $R^2$ score for testing data predictions.

Additionally, Figures 6–8 display the predicted outputs compared to the real data records for the compared methods using the datasets of the three cities, Beijing, Wuhan, and Shijiazhuang, respectively. We selected 20% of the last records based on the time from each dataset for the test prediction. From these figures, we can see that the ResInformerStack

obtained the nearest value to the real record (ground truth) on the Beijing and Shijiazhuang datasets. In terms of the Wuhan dataset, it can be noticed that the Informer model performed similarly to ResInformerStack, where at some time stamps, the Informer model was relatively accurate. As noticed in the Beijing dataset charts, the InformerStack did not perform well on the February to March segment, whereas the other models performed better on this segment.

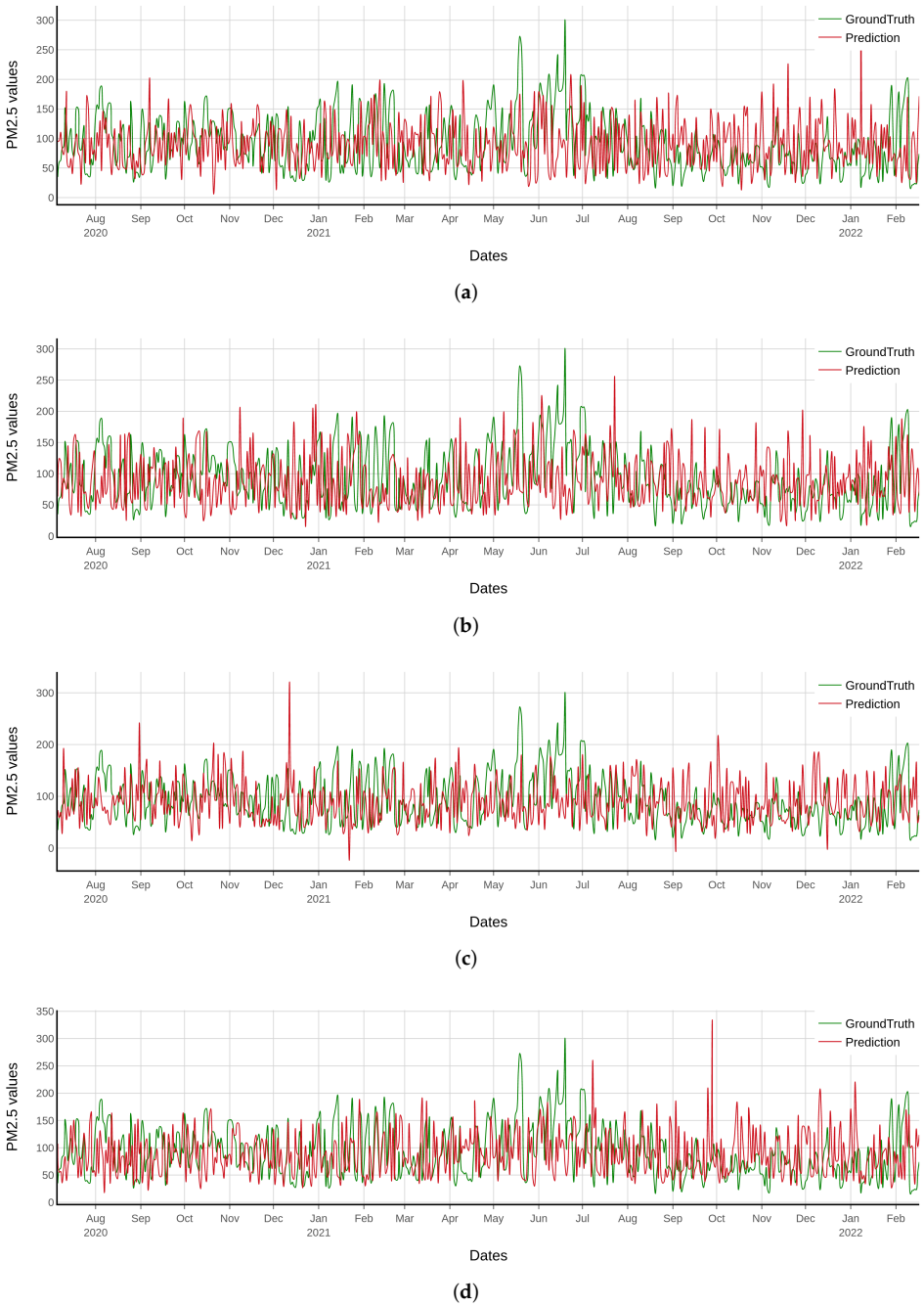



**Figure 6.** Beijing PM25 predictions. (**a**) PM2.5 predictions on the testing set using Informer; (**b**) PM2.5 predictions on the testing set using InformerStack; (**c**) PM2.5 predictions on the testing set using ResInformer; (**d**) PM2.5 predictions on the testing set using ResInformerStack.

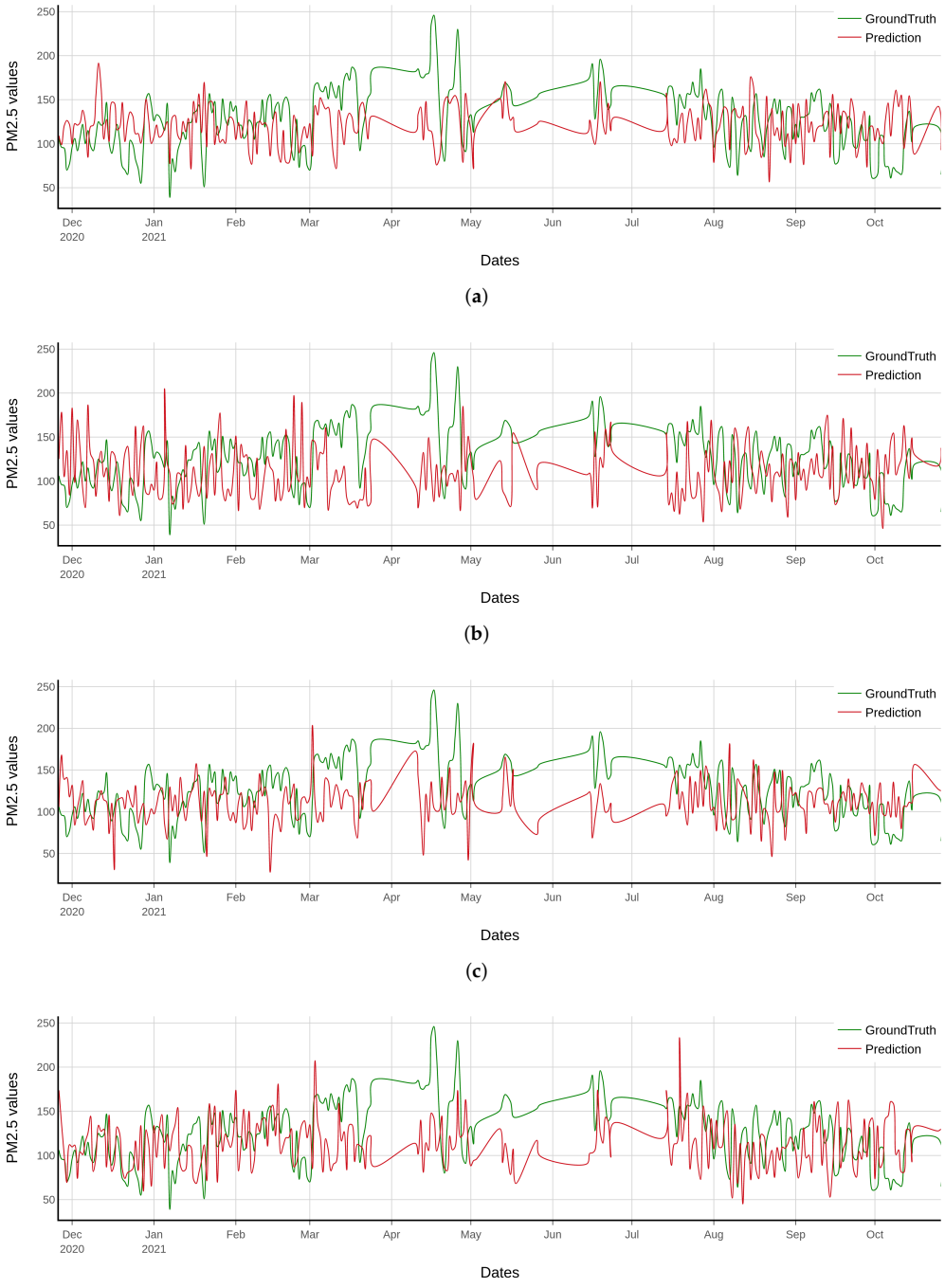
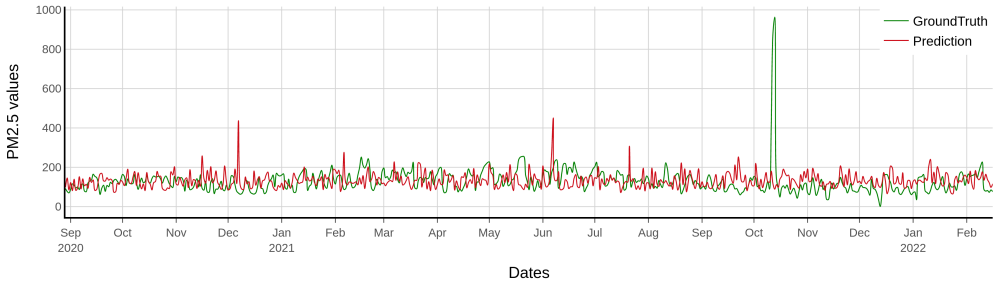
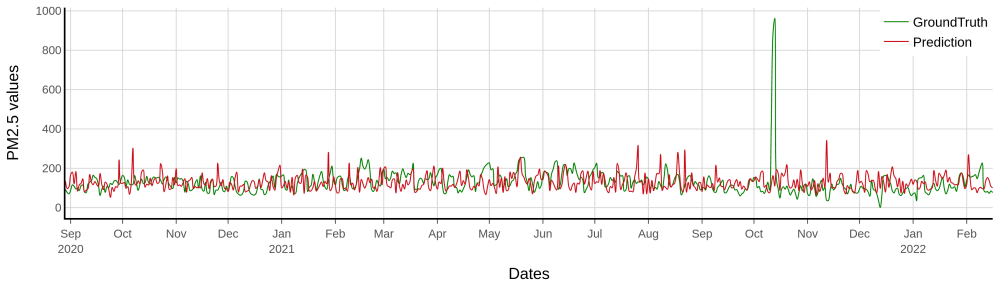


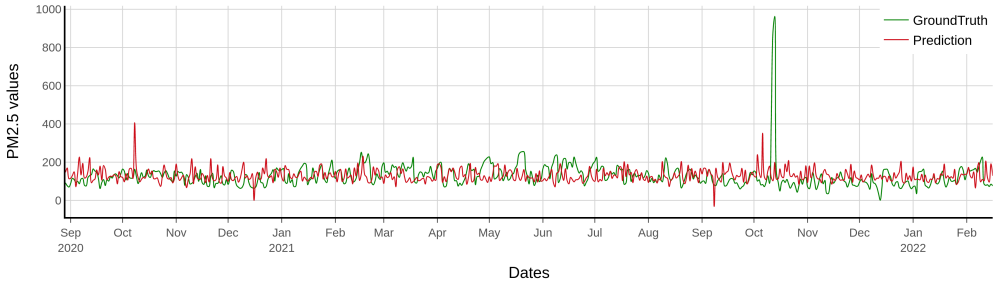

**Figure 7.** Wuhan PM2.5 predictions. (**a**) PM2.5 predictions on the testing set using Informer; (**b**) PM2.5 predictions on the testing set using InformerStack; (**c**) PM2.5 predictions on the testing set using ResInformer; (**d**) PM2.5 predictions on the testing set using ResInformerStack.



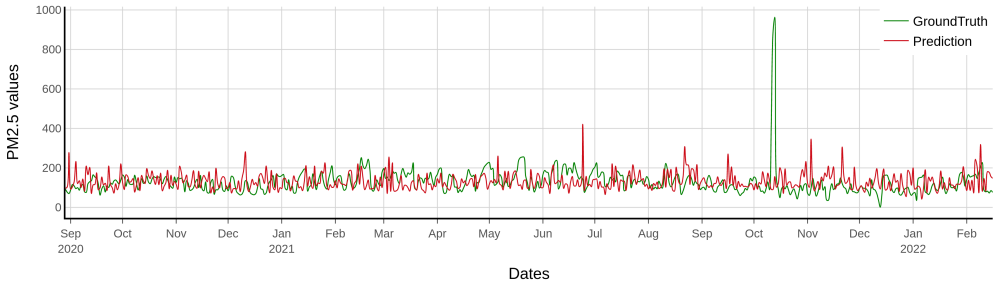

(**a**)




(**b**)




(**c**)




(**d**)

**Figure 8.** Shijiazhuang PM2.5 predictions. (**a**) PM2.5 predictions on the testing set using Informer; (**b**) PM2.5 predictions on the testing set using InformerStack; (**c**) PM2.5 predictions on the testing set using ResInformer; (**d**) PM2.5 predictions on the testing set using ResInformerStack.

**Table 2.** The prediction results of PM2.5.

| City | Model | MSE | MAE | $R^2$ | RMSE | MAPE | MSPE |
|---|---|---|---|---|---|---|---|
| Beijing | Informer | 0.0813 | 0.2159 | 0.8285 | 0.2852 | 0.8064 | 9.79 |
| | InformerStack | 0.0725 | 0.2012 | 0.8472 | 0.2692 | 0.8921 | 16.30 |
| | ResInformer | 0.0796 | 0.2130 | 0.8320 | 0.2822 | 0.8565 | 11.96 |
| | ResInformerStack | **0.0688** | **0.1964** | **0.8549** | **0.2623** | **0.7561** | **7.73** |
| Wuhan | Informer | 0.2730 | 0.4180 | 0.5329 | 0.5225 | 1.44 | 29.94 |
| | InformerStack | 0.1380 | **0.2911** | 0.7621 | 0.3716 | 1.49 | 26.94 |
| | ResInformer | 0.2215 | 0.3782 | 0.6142 | 0.4706 | 1.54 | 32.72 |
| | ResInformerStack | **0.1378** | 0.2982 | **0.7656** | **0.3712** | **1.39** | **21.19** |
| Shijiazhuang | Informer | 0.2614 | **0.2890** | **0.6433** | 0.5112 | **1.79** | **137.91** |
| | InformerStack | 0.2925 | 0.3081 | 0.6020 | 0.5408 | 2.25 | 263.97 |
| | ResInformer | **0.2158** | 0.3138 | 0.5857 | **0.4646** | 2.00 | 212.94 |
| | ResInformerStack | 0.2855 | 0.3055 | 0.4937 | 0.5343 | 1.91 | 185.91 |

## 6. Conclusions

China has implemented strict policies to tackle the problem of air pollution. Therefore, developing time-series forecasting tools for air quality in China is necessary. Thus, we selected three major cities in China, Beijing, Wuhan, and Shijiazhuang, to forecast particulate matter (PM2.5) concentrations using public datasets. The developed deep learning model was built based on the well-known Informer model, where the attention distillation block was boosted with a residual block-inspired structure from efficient networks. The developed structure was named ResInformer. With extensive evaluation experiments, we concluded that the developed ResInformer and its variant ResInformerStack performed better than the original Informer and its variant InformerStack in many cases.
However, the proposed model faced some limitations in the context of the encoder–decoder architectures, such as the large set of training parameters that needed to be set carefully before the training, which could affect the network size and complexity. In addition, the network speed and accuracy convergence should be further studied in depth, which will be considered in our future work, in order to balance the network performance and the time and resources cost. In future work, we will enhance the prediction accuracy of the ResInformer further so that it can be employed for other time-series forecasting applications.

**Author Contributions:** Conceptualization, M.A.A.A.-q. and A.D.; methodology, M.A.A.A.-q. and A.D.; software, A.D. and A.A.E.; validation, M.A.E., L.A. and J.H.; formal analysis, J.H.; investigation, M.A.A.A.-q.; resources, M.A.A.A.-q.; data curation, M.A.A.A.-q. and L.A.; writing—original draft preparation, M.A.A.A.-q., A.D., A.A.E. and A.M.H.; writing—review and editing, L.A., J.H. and M.A.E.; visualization, A.D. and J.H.; supervision, M.A.A.A.-q.; project administration, M.A.A.A.-q.; funding acquisition, M.A.A.A.-q. All authors have read and agreed to the submitted version of the manuscript.

**Funding:** This work was supported by the Scientific Research Center at Buraydah Private Colleges under the research project # BPC-SRC/2022-010.

**Data Availability Statement:** Available online at https://aqicn.org/data-platform/covid19/, (accessed on 1 March 2022).

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

## References

1. Barthwal, A.; Acharya, D.; Lohani, D. Prediction and analysis of particulate matter (PM2.5 and PM10) concentrations using machine learning techniques. *J. Ambient. Intell. Humaniz. Comput.* **2021**, 1–16. [CrossRef]
2. Zhang, Z.; Wu, L.; Chen, Y. Forecasting PM2.5 and PM10 concentrations using GMCN (1, N) model with the similar meteorological condition: Case of Shijiazhuang in China. *Ecol. Indic.* **2020**, *119*, 106871. [CrossRef]
3. Wu, J.; Li, T.; Zhang, C.; Cheng, Q.; Shen, H. Hourly PM 2.5 Concentration Monitoring With Spatiotemporal Continuity by the Fusion of Satellite and Station Observations. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 8019–8032. [CrossRef]

4. Xiao, F.; Yang, M.; Fan, H.; Fan, G.; Al-Qaness, M.A. An improved deep learning model for predicting daily PM2.5 concentration. *Sci. Rep.* **2020**, *10*, 20988. [CrossRef] [PubMed]
5. Al-Qaness, M.A.; Fan, H.; Ewees, A.A.; Yousri, D.; Abd Elaziz, M. Improved ANFIS model for forecasting Wuhan City air quality and analysis COVID-19 lockdown impacts on air quality. *Environ. Res.* **2021**, *194*, 110607. [CrossRef]
6. Şahin, C.B.; Dinler, Ö.B.; Abualigah, L. Prediction of software vulnerability based deep symbiotic genetic algorithms: Phenotyping of dominant-features. *Appl. Intell.* **2021**, *51*, 8271–8287. [CrossRef]
7. Danandeh Mehr, A.; Rikhtehgar Ghiasi, A.; Yaseen, Z.M.; Sorman, A.U.; Abualigah, L. A novel intelligent deep learning predictive model for meteorological drought forecasting. *J. Ambient. Intell. Humaniz. Comput.* **2022**, 1–15. [CrossRef]
8. Goudarzi, G.; Hopke, P.K.; Yazdani, M. Forecasting PM2.5 concentration using artificial neural network and its health effects in Ahvaz, Iran. *Chemosphere* **2021**, *283*, 131285. [CrossRef]
9. Jamei, M.; Karbasi, M.; Mosharaf-Dehkordi, M.; Olumegbon, I.A.; Abualigah, L.; Said, Z.; Asadi, A. Estimating the density of hybrid nanofluids for thermal energy application: Application of non-parametric and evolutionary polynomial regression data-intelligent techniques. *Measurement* **2021**, *189*, 110524. [CrossRef]
10. Barbera, E.; Curro, C.; Valenti, G. A hyperbolic model for the effects of urbanization on air pollution. *Appl. Math. Model.* **2010**, *34*, 2192–2202. [CrossRef]
11. Manojkumar, N.; Srimuruganandam, B. Health effects of particulate matter in major Indian cities. *Int. J. Environ. Health Res.* **2021**, *31*, 258–270. [CrossRef] [PubMed]
12. Yang, M.C.; Chen, M.C. Composite Neural Network: Theory and Application to PM2.5 Prediction. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 1311–1323. [CrossRef]
13. Liu, H.; Duan, Z.; Chen, C. A hybrid multi-resolution multi-objective ensemble model and its application for forecasting of daily PM2.5 concentrations. *Inf. Sci.* **2020**, *516*, 266–292. [CrossRef]
14. He, J.; Christakos, G.; Jankowski, P. Comparative Performance of the LUR, ANN, and BME Techniques in the Multiscale Spatiotemporal Mapping of PM 2.5 Concentrations in North China. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 1734–1747. [CrossRef]
15. Duan, W.; Wang, X.; Cheng, S.; Wang, R. Regional collaboration to simultaneously mitigate PM2.5 and O3 pollution in Beijing-Tianjin-Hebei and the surrounding area: Multi-model synthesis from multiple data sources. *Sci. Total. Environ.* **2022**, *820*, 153309. [CrossRef]
16. Hu, S.; Zhao, G.; Tan, T.; Li, C.; Zong, T.; Xu, N.; Zhu, W.; Hu, M. Current challenges of improving visibility due to increasing nitrate fraction in PM2.5 during the haze days in Beijing, China. *Environ. Pollut.* **2021**, *290*, 118032. [CrossRef]
17. Yang, X.; Zheng, M.; Liu, Y.; Yan, C.; Liu, J.; Liu, J.; Cheng, Y. Exploring sources and health risks of metals in Beijing PM2.5: Insights from long-term online measurements. *Sci. Total Environ.* **2022**, *814*, 151954. [CrossRef]
18. Pan, J.; Xue, Y.; Li, S.; Wang, L.; Mei, J.; Ni, D.; Jiang, J.; Zhang, M.; Yi, S.; Zhang, R.; et al. PM2.5 induces the distant metastasis of lung adenocarcinoma via promoting the stem cell properties of cancer cells. *Environ. Pollut.* **2022**, *296*, 118718. [CrossRef]
19. Yang, S.; Tan, Y.; Mei, H.; Wang, F.; Li, N.; Zhao, J.; Zhang, Y.; Qian, Z.; Chang, J.J.; Syberg, K.M.; et al. Ambient air pollution the risk of stillbirth: A prospective birth cohort study in Wuhan, China. *Int. J. Hyg. Environ. Health* **2018**, *221*, 502–509. [CrossRef]
20. Zhao, J.; Zhang, B.; Yang, S.; Mei, H.; Qian, Z.; Liang, S.; Zhang, Y.; Hu, K.; Tan, Y.; Xian, H.; et al. Maternal exposure to ambient air pollutant and risk of oral clefts in Wuhan, China. *Environ. Pollut.* **2018**, *238*, 624–630. [CrossRef]
21. Yang, Z.; Hao, J.; Huang, S.; Yang, W.; Zhu, Z.; Tian, L.; Lu, Y.; Xiang, H.; Liu, S. Acute effects of air pollution on the incidence of hand, foot, and mouth disease in Wuhan, China. *Atmos. Environ.* **2020**, *225*, 117358. [CrossRef]
22. Qian, Z.; Liang, S.; Yang, S.; Trevathan, E.; Huang, Z.; Yang, R.; Wang, J.; Hu, K.; Zhang, Y.; Vaughn, M.; et al. Ambient air pollution and preterm birth: A prospective birth cohort study in Wuhan, China. *Int. J. Hyg. Environ. Health* **2016**, *219*, 195–203. [CrossRef] [PubMed]
23. Wang, Q.; Zhu, H.; Xu, H.; Lu, K.; Ban, J.; Ma, R.; Li, T. The spatiotemporal trends of PM2.5-and O3-related disease burden coincident with the reduction in air pollution in China between 2005 and 2017. *Resour. Conserv. Recycl.* **2022**, *176*, 105918. [CrossRef]
24. Gao, X.; Koutrakis, P.; Coull, B.; Lin, X.; Vokonas, P.; Schwartz, J.; Baccarelli, A.A. Short-term exposure to PM2.5 components and renal health: Findings from the Veterans Affairs Normative Aging Study. *J. Hazard. Mater.* **2021**, *420*, 126557. [CrossRef] [PubMed]
25. Shao, M.; Wang, W.; Yuan, B.; Parrish, D.D.; Li, X.; Lu, K.; Wu, L.; Wang, X.; Mo, Z.; Yang, S.; et al. Quantifying the role of PM2.5 dropping in variations of ground-level ozone: Inter-comparison between Beijing and Los Angeles. *Sci. Total Environ.* **2021**, *788*, 147712. [CrossRef]
26. Lyu, B.; Cai, Y.; Sun, Z.; Li, J.; Liang, L. Evaluating temporally decomposed associations between PM2.5 and hospitalisation risks of AECOPD: A case study in Beijing from 2010 to 2019. *Atmos. Pollut. Res.* **2022**, *13*, 101356. [CrossRef]
27. Yang, J.; Yan, R.; Nong, M.; Liao, J.; Li, F.; Sun, W. PM2.5 concentrations forecasting in Beijing through deep learning with different inputs, model structures and forecast time. *Atmos. Pollut. Res.* **2021**, *12*, 101168. [CrossRef]
28. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 2–9 February 2021; Volume 35, pp. 11106–11115.

29. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
30. Chang, J.H.; Tseng, C.Y. Analysis of correlation between secondary PM2.5 and factory pollution sources by using ANN and the correlation coefficient. *IEEE Access* **2017**, *5*, 22812–22822. [CrossRef]
31. Guo, C.; Liu, G.; Lyu, L.; Chen, C.H. An unsupervised PM2.5 estimation method with different spatio-temporal resolutions based on KIDW-TCGRU. *IEEE Access* **2020**, *8*, 190263–190276. [CrossRef]
32. Cheng, X.; Liu, Y.; Xu, X.; You, W.; Zang, Z.; Gao, L.; Chen, Y.; Su, D.; Yan, P. Lidar data assimilation method based on CRTM and WRF-Chem models and its application in PM2.5 forecasts in Beijing. *Sci. Total Environ.* **2019**, *682*, 541–552. [CrossRef]
33. Qiao, W.; Tian, W.; Tian, Y.; Yang, Q.; Wang, Y.; Zhang, J. The forecasting of PM2.5 using a hybrid model based on wavelet transform and an improved deep learning algorithm. *IEEE Access* **2019**, *7*, 142814–142825. [CrossRef]
34. Huang, G.; Li, X.; Zhang, B.; Ren, J. PM2.5 concentration forecasting at surface monitoring sites using GRU neural network based on empirical mode decomposition. *Sci. Total Environ.* **2021**, *768*, 144516. [CrossRef] [PubMed]
35. Zhao, Y.; Wang, L.; Huang, T.; Tao, S.; Liu, J.; Gao, H.; Luo, J.; Huang, Y.; Liu, X.; Chen, K.; et al. Unsupervised PM2.5 anomalies in China induced by the COVID-19 epidemic. *Sci. Total Environ.* **2021**, *795*, 148807. [CrossRef] [PubMed]
36. Menares, C.; Perez, P.; Parraguez, S.; Fleming, Z.L. Forecasting PM2.5 levels in Santiago de Chile using deep learning neural networks. *Urban Clim.* **2021**, *38*, 100906. [CrossRef]
37. Zhou, W.; Wu, X.; Ding, S.; Ji, X.; Pan, W. Predictions and mitigation strategies of PM2.5 concentration in the Yangtze River Delta of China based on a novel nonlinear seasonal grey model. *Environ. Pollut.* **2021**, *276*, 116614. [CrossRef]
38. Bar, S.; Parida, B.R.; Mandal, S.P.; Pandey, A.C.; Kumar, N.; Mishra, B. Impacts of COVID-19 lockdown on NO2 and PM2.5 levels in major urban cities of Europe and USA. *Cities* **2021**, *117*, 103308. [CrossRef]
39. Liou, N.C.; Luo, C.H.; Mahajan, S.; Chen, L.J. Why Is Short-Time PM2.5 Forecast Difficult? The Effects of Sudden Events. *IEEE Access* **2019**, *8*, 12662–12674. [CrossRef]
40. Deng, F.; Ma, L.; Gao, X.; Chen, J. The MR-CA models for analysis of pollution sources and prediction of PM 2.5. *IEEE Trans. Syst. Man, Cybern. Syst.* **2017**, *49*, 814–820. [CrossRef]
41. Nguyen, M.H.; Le Nguyen, P.; Nguyen, K.; Nguyen, T.H.; Ji, Y. PM2.5 Prediction Using Genetic Algorithm-Based Feature Selection and Encoder-Decoder Model. *IEEE Access* **2021**, *9*, 57338–57350. [CrossRef]
42. Mahajan, S.; Liu, H.M.; Tsai, T.C.; Chen, L.J. Improving the accuracy and efficiency of PM2.5 forecast service using cluster-based hybrid neural network model. *IEEE Access* **2018**, *6*, 19193–19204. [CrossRef]
43. Gu, K.; Qiao, J.; Li, X. Highly efficient picture-based prediction of PM2.5 concentration. *IEEE Trans. Ind. Electron.* **2018**, *66*, 3176–3184. [CrossRef]
44. Sun, Y.; Zeng, Q.; Geng, B.; Lin, X.; Sude, B.; Chen, L. Deep learning architecture for estimating hourly ground-level PM 2.5 using satellite remote sensing. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1343–1347. [CrossRef]
45. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
46. Yu, F.; Koltun, V.; Funkhouser, T. Dilated residual networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 472–480.
47. Clevert, D.A.; Unterthiner, T.; Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). *arXiv* **2015**, arXiv:1511.07289.
48. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.