applied sciences

MDPI

Article

A Machine Learning-Based Ensemble Framework for Forecasting $PM_{2.5}$ Concentrations in Puli, Taiwan

Peng-Yeng Yin [1,*], Alex Yaning Yen [2], Shou-En Chao [3], Rong-Fuh Day [3] and Bir Bhanu [4]

[1] Department of Computer Science and Information Engineering, China University of Technology, Taipei 11695, Taiwan
[2] Center for the Conservation of Cultural Heritage, China University of Technology, Taipei 11695, Taiwan; alexyen@cute.edu.tw
[3] Department of Information Management, National Chi Nan University, Nantou 54561, Taiwan; shouen.chao@gmail.com (S.-E.C.); rfday@ncnu.edu.tw (R.-F.D.)
[4] Visualization and Intelligent Systems Laboratory, University of California, Riverside, CA 92521, USA; bhanu@ee.ucr.edu
* Correspondence: pengyengyin@gmail.com

Abstract: Forecasting of $PM_{2.5}$ concentration is a global concern. Evidence has shown that the ambient $PM_{2.5}$ concentrations are harmful to human health, climate change, plant species mortality, etc. $PM_{2.5}$ concentrations are caused by natural and anthropogenic activities, and it is challenging to predict them due to many uncertain factors. Current research has focused on developing a new model while overlooking the fact that every single model for $PM_{2.5}$ prediction has its own strengths and weaknesses. This paper proposes an ensemble framework which combines four diverse learning models for $PM_{2.5}$ forecasting in Puli, Taiwan. It explores the synergy between parametric and non-parametric learning, and short-term and long-term learning. The feature set covers periodic, meteorological, and autoregression variables which are selected by a spiral validation process. The experimental dataset, spanning from 1 January 2008 to 31 December 2019, from Puli Township in central Taiwan, is used in this study. The experimental results show the proposed multi-model framework can synergize the advantages of the embedded models and obtain an improved forecasting result. Further, the benefit obtained by blending short-term learning with long-term learning is validated, in surpassing the performance obtained by using just single type of learning. Our multi-model framework compares favorably with deep-learning models on Puli dataset. It also shows high adaptivity, such that our multi-model framework is comparable to the leading methods for $PM_{2.5}$ forecasting in Delhi, India.

Keywords: $PM_{2.5}$; short-term learning; long-term learning; multi-model framework; forecast

Citation: Yin, P.-Y.; Yen, A.Y.; Chao, S.-E.; Day, R.-F.; Bhanu, B. A Machine Learning-Based Ensemble Framework for Forecasting $PM_{2.5}$ Concentrations in Puli, Taiwan. *Appl. Sci.* **2022**, *12*, 2484. https://doi.org/10.3390/app12052484

Academic Editor: José Carlos Magalhães Pires

Received: 26 January 2022
Accepted: 24 February 2022
Published: 27 February 2022

1. Introduction

The urbanization of human history has inevitably increased the scale and region of the industrial and food production. The immense amount of particulate matter with aerodynamic diameter $\leq$2.5 μm ($PM_{2.5}$) generated during the production drifts into the air and even infiltrates into our houses [1]. Other anthropogenic pollution sources include vehicle exhaust, burning activities, coal, and gasoline combustion, to name a few [2]. Research studies have shown that ambient $PM_{2.5}$ concentrations greatly contribute to human respiratory diseases [3,4] and cancers [5]. The World Health Organization (WHO) reported that ambient $PM_{2.5}$ concentrations caused an estimated 3 million premature deaths worldwide in 2012 [6].

The concentration of $PM_{2.5}$ is hard to forecast due to anthropogenic activities, natural landscape profiles, and uncertain weather conditions. Some anthropogenic activities and natural scenarios have periodic characteristics. These periodic characteristics could be long-term (from a season to a year) or short-term (from a couple of hours to several days).

For long-term examples, the power plants in Taiwan are used to raise their production during summer in response to the high energy demand, and the monsoon usually brings northerly serious air pollutants to Taiwan's west coast in every winter. As short-term examples, the Taiwanese burn a lot of incense and joss papers on some religious ceremony days, many people gather at night for barbecue activities on the Mid-Autumn Festival Day, and anticyclones usually cause the ambient pollutants to last for several days.

The diverse methodologies adopted in the literature for $PM_{2.5}$ forecasting range from regression and autoregression, machine learning, meta-evolution, receptor model, to hybrid methods. These methods can be classified as parametric or non-parametric ones. The parametric methods require explanatory variables in addition to the $PM_{2.5}$ series itself, while the non-parametric methods need only the $PM_{2.5}$ series. Parametric methods are effective when the selected explanatory variables are informative to describe the $PM_{2.5}$ concentrations. The widely used explanatory variables include wind speed and direction, precipitation, temperature, relative humidity, atmospheric pressure, land use, traffic amount, road types, satellite images, etc. Non-parametric methods analyze the temporal trends or series components in the historical $PM_{2.5}$ data and use them as input for predicting the next $PM_{2.5}$ concentration. It is worth noting that hybrid methods are currently more popular because each type of methodology has its own strengths and weaknesses and a hybrid of multiple models is likely to complement one another. Several works have shown that hybrid methods can boost the performance obtained by using a single method [7–9].

This paper proposes an ensemble framework that combines four diverse prediction models. Both ensemble and hybrid methods combine multiple models to enhance the overall performance, however, there is still a difference between them. In the ensemble method, the adopted models work independently with the same input data, and a learning scheme (such as probabilistic, weighting, or voting mechanism) is used to integrate the outputs from the multiple models. While in the hybrid method, the embedded models usually work cooperatively in a one-way fashion. The output of a preceding model is used as the input of the next model. So there is no need to integrate the results of respective models.

Our ensemble framework combines four diverse models, namely, the cluster linear regression, Fourier series descriptor, short-term multi-layer perceptron, and long-term multi-layer perceptron. Our ensemble framework includes parametric and non-parametric models, short-term and long-term learning schemes in a single framework. To the best of our knowledge, this paper is the first attempt to develop an ensemble of such diverse models and learning schemes for $PM_{2.5}$ forecasting. In addition to the framework design, the feature selection is deliberate. Four types of features, including periodic variables, meteorological variables, short-term meteorological variables, and short-term autoregression variables, are considered. A spiral validation process [10,11] is devised to reduce the size of the feature set by retaining the most effective variables. We choose a central Taiwan township named Puli as our studied area. Our experimental results with the $PM_{2.5}$ and meteorology datasets for the twelve year period from 2008 to 2019 show that the proposed multi-model framework can synergize the advantages of the embedded models and obtain an improved forecasting result, indicating that the design of our ensemble framework is promising. We also show that the performance obtained by using the proposed multi-model framework surpasses that obtained by using either short-term learning ensemble or long-term learning ensemble, revealing the importance of training with both short-term and long-term data in order to make the forecasting system adaptive to varying pollution events and weather conditions.

The remainder of this paper is organized as follows. Section 2 reviews the relevant research on $PM_{2.5}$ forecasting and presents the contributions of this paper. Section 3 elucidates the proposed multi-model ensemble framework. Section 4 presents the experimental results and comparative performance. Finally, Section 5 concludes this work.

2. Related Work and Contributions

2.1. Related Work

We have performed a comprehensive review of recent work on $PM_{2.5}$ forecasting. Current forecasting approaches can be classified into the following five categories.

(1) ***Regression or autoregression.*** *First*, regression (either linear or nonlinear) techniques stem from a set of explanatory variables (such as meteorological metrics, traffic conditions, road types, etc.) and use them to estimate the response variable (i.e., $PM_{2.5}$ concentrations) by training on cross-sectional data. Regression techniques are useful for analyzing the relationship between explanatory variables and the response variable. However, the selection of effective explanatory variables depends on the local meteorological patterns and it is critical to the prediction accuracy. Multivariate linear regression models for the forecasting of concentrations of NOx and PM_{10} in Athens and Helsinki have been intensively evaluated by [12] using NO, NO_2, CO, O_3, and $PM_{2.5}$ as the explanatory variables. Nonlinear regression models have been deployed in [13] to forecast the $PM_{2.5}$ air quality and detect unhealthy $PM_{2.5}$ event in the Louisville, Kentucky metropolitan area. Separate nonlinear regression models were presented by [14] for primary $PM_{2.5}$, $PM_{2.5}$ sulfate ion, and $PM_{2.5}$ nitrate ion. The explanatory variables of the nonlinear regression models included facility emissions rates in tons per year and the distance between the single emissions source and receptor. The multiple nonlinear regressions were found useful in predicting Beijing's daily $PM_{2.5}$ concentration with one nonlinear regression applied to each season of meteorological conditions [15]. The graphically and temporally weighted regression has been attempted in [16] to find the relationship between the surface $PM_{2.5}$ concentration and the satellite-derived aerosol optical depth (AOD) data. Dynamic multiple linear equations have been used in [17] to model the hourly $PM_{2.5}$ concentration in relation to meteorological characteristics. The authors showed that the prediction accuracy of the dynamic system could surpass to that obtained by nonlinear models. A spatiotemporal land use regression method was applied to forecast the $PM_{2.5}$ concentration in finely-grained spatial grids [18]. The result showed that the $PM_{2.5}$ concentration level in primary schools is significantly different to the mean $PM_{2.5}$ level of the corresponding city.

Second, autoregression methods, such as autoregressive integrated moving average model (ARIMA) and generalized autoregressive conditional heteroscedasticity (GARCH), require no explanatory variables but focus on discovering the temporal trends contained in the $PM_{2.5}$ time series. ARIMA was adopted in [19] to explore the short-term time series and estimate the mean daily $PM_{2.5}$ concentration. A GARCH model was used in [8] to capture linear and nonlinear panel information for $PM_{2.5}$ concentrations forecasting.

(2) ***Machine learning.*** Machine learning approaches are the main stream in the literature of $PM_{2.5}$ forecasting. The approaches, such as artificial neural networks (ANNs), fuzzy systems, and support vector machine (SVM), are able to learn the relationship function between the explanatory variables and the responsive variable ($PM_{2.5}$ concentrations) based on training data. The advantage of machine learning is that there is no prior assumption for the form of the relationship function. In other words, a black-box learning is implicitly performed without the need to anticipate the analytic form of the relationship. General regression neural networks (GRNN) have been applied in [20] to predict all decomposed $PM_{2.5}$ components obtained by the ensemble empirical mode decomposition (EEMD) method. The EEMD method has also been used in [21] to decompose the $PM_{2.5}$ historical data and the least square support vector machine (LSSVM) is employed to predict all reconstructed components and integrate them to produce the final forecasting result. A specific type of ANN named multiple layer perceptron (MLP) is adopted in [22] to learn the relationship between $PM_{2.5}$ concentration and a set of meteorological factors with satellite-derived AOD data. An ensemble model was established in [23] to combine $PM_{2.5}$ estimates from three machine learning methods, namely, neural network, random forest, and gradient boosting. An improved deep learning model for predicting daily $PM_{2.5}$ concentration in the Beijing–Tianjin–Hebei area has been proposed in [24]. The improved deep learning model considers spatiotemporal correlations among surrounding monitoring sites with spatial

site-density information. Firstly, an MLP is applied to generate weighted $PM_{2.5}$ time series data based on air pollution concentration, wind condition, and the distance between the focused site and its neighboring sites. Second, both the original and weighted $PM_{2.5}$ time series are fed into a long short-term memory (LSTM) to extract spatiotemporal features. Finally, these spatiotemporal features and meteorological data are input to another MLP to predict the daily $PM_{2.5}$ concentration. Vast amounts of meteorological and pollutant data have been dealt with by [25] based on deep learning techniques. A deep learning network consisting of a convolutional neural network (CNN) and an LSTM is deployed. The CNN extracts meteorological and pollutant features from 14 sites in Shanghai and the LSTM is used to model time dependence of pollutants. The experimental results showed that the proposed combined model outperforms classic deep learning models.

(3) ***Meta-evolution.*** Evolutionary algorithms are a family of nature-inspired optimization algorithms based on evolution in biology. Among them, genetic algorithm (GA), particle swarm optimization (PSO), ant cuckoo search algorithm (CSA) are notable ones. Meta-evolution is a framework where the evolutionary algorithms are applied to fine tune the hyperparameters of a machine learning model to improve its performance. Meta-evolution frameworks have been developed to predict $PM_{2.5}$ concentrations. For example, GA [26], PSO [27], and CSA [28] have been applied to optimize the parameter settings of artificial neural networks and support vector machines to estimate the $PM_{2.5}$ mass by using meteorological variables as inputs.

(4) ***Receptor model***. In this model, the relationship between the pollution and its particulate sources is estimated according to the pollutants' conservation on chemical mass balance (CMB) or the chemical transportation by the profiles of particulate mass sources along the dispersion route. The receptor model is commonly used for $PM_{2.5}$ dispersion prediction of line or point sources such as roadways or burning sites. The dispersion of $PM_{2.5}$ along urban highway in India has been analyzed in [29]. The vehicle types, sizes, and ages are considered in the emission mass balance and the molecular settling velocity and meteorological conditions are incorporated into the line source dispersion prediction model. The impact of Taiwan's barbecue festival on air quality is studied in [30], which monitored the $PM_{2.5}$ mass before, in, and after, the barbecue festival and compared the contribution from different chemical species. Another type of receptor model is based on statistical analysis, such as positive matrix factorization (PMF) [31]. PMF justifies the number of factors for finding the best-fit model to interpret the solution. PMF is used to analyze the contribution of various source apportionments in a mixture of $PM_{2.5}$ samples.

(5) ***Hybrid approaches.*** In order to obtain better prediction performance, hybrid methods combining multiple types of prediction approaches are proposed. For instance, receptor model was used in [23] to simulate the spatiotemporal transportation of $PM_{2.5}$ mass, and a back propagation neural network was applied to calibrate the simulations with meteorological and land use data. A hybrid GARCH model combining ARIMA and a SVM has been proposed in [8] for $PM_{2.5}$ concentrations forecasting. A multi-model fusion method for $PM_{2.5}$ prediction is proposed in [9]. The backpropagation neural network is used as the fusion model to integrate the decisions from multiple regression methods. The prediction accuracy of the fusion method has been shown to be superior to that obtained from a single model. Kumar et al. [32] proposed a hybrid machine-learning method to predict the $PM_{2.5}$ concentration in Delhi on an hourly basis. The method firstly uses an extra-trees regression to learn the correlation between the $PM_{2.5}$ and meteorological time series. Then the AdaBoost is applied to boost the performance by assigning stronger weight to misclassified samples.

2.2. Research Trends and Contributions of This Paper

Depending on the applications, the time unit for $PM_{2.5}$ forecasting varies significantly. For short-term forecasting, the $PM_{2.5}$ concentration is estimated for the next immediate hour or in the next 6 h, 12 h, or 24 h [8,9,17,19,22,27,29]. For middle-term forecast, the daily and weekly $PM_{2.5}$ concentration are predicted in [7,15,16,18–21,23,26,28,30]. Rela-

tively few works [18,23] have contemplated the long-term prediction such as seasonal and yearly $PM_{2.5}$.

For the country of the studied area, most existing works chose the cities in China as their investigation fields because these cities are among those which have the most serious $PM_{2.5}$ pollution in the world. Some other studies have been conducted in the rest of Asia (e.g., Taiwan, Iran, and India) and America (USA, Chile).

From our literature review, we observe three research trends as follows. (1) The regression techniques and the machine learning (including ANN) models constitute the two main classes of approaches used in the literature for $PM_{2.5}$ forecasting. (2) Both parametric (which requires explanatory variables) and non-parametric (which requires no explanatory variables) models are widely adopted in the literature, and there are no significant performance advantages of one over the other. Various phenomena reveal that the relationship between $PM_{2.5}$ concentrations and the explanatory variables is too complex to be described by a parametric model. The $PM_{2.5}$ time series itself manifests some salient trends and the usage of non-parametric model is desired to realize the temporal trends. (3) Different lengths of training time span have been considered in the literature. It ranges from a week [8], a month [28,29], to multiple years [7,15]. Nevertheless, none of the approaches in the literature has fused both the short-term and long-term learning. This is the promising direction that will be explored in this paper. (4) An increasing number of recent work has developed hybrid approaches which combine multiple models in a single framework to exploit the strengths and weaknesses of individual models. In particular, regression has been hybridized with ANN [9] and SVM [8].

The characteristics of existing literature can be compared according to the following three aspects: forecasting approaches, time unit of forecasting, and the country of studied area, as listed in Table 1.

Table 1. Comparison of recent works on $PM_{2.5}$ forecasting.

	Regression or Autoregression	Artificial Neural Network (ANN)	Machine Learning (ANN Excluded)	Meta-Evolution	Receptor Model	Time Units of Forecasting	Country of Studied Area
Zhu and Fan, 2015				•		day	China
Tsai et al., 2015					•	day	Taiwan
Yin et al., 2016	•					day	China
Ausati and Amanollahi, 2016		•				day	Iran
Di et al., 2016 *		•			•	day	USA
Zhang et al., 2016				•		hour	China
Ni et al., 2017	•					hour/day	China
Niu et al., 2017			•			day	China
Wang et al., 2017 *	•		•			hour	China
Sun and Sun, 2017				•		day	China
Mao et al., 2017		•				hour	China
Dhyni et al., 2017					•	hour	India
Guo et al., 2017	•					day	China
Moisan et al., 2018	•					hour	Chile
Zhang et al., 2019 *	•	•				hour	China
Di et al., 2019			•			day/season/year	USA
Qin et al., 2019		•				hour	China
Zhang et al., 2020	•					week/season/year	China
Xiao et al., 2020		•				day	China
Kumar et al., 2020 *	•		•			hour	India

* hybrid approaches.

In light of these trends, this paper proposes (1) a multi-model ensemble framework for the next 24-h $PM_{2.5}$ concentration forecast in Puli, Taiwan. The multiple models complement one another from different learning perspectives and they collaborate to maximize the overall performance. (2) Four regression and machine learning models are deployed in our framework. Different learning characteristics, namely, parametric and non-parametric, short-term and long-term, are exploited in the proposed framework. (3) Our experiments, conducted with the historical Puli, Taiwan dataset, show promising results.

3. Proposed Methods

3.1. Framework Architecture

We develop a multi-model framework for dealing with the next 24-h $PM_{2.5}$ forecasting problem. Figure 1 shows the architecture of our framework. The response variable is the $PM_{2.5}$ concentration to be forecasted and the explanatory variables are relevant periodic and meteorological factors which are selected based on a spiral validation method. The properties and trends of the response variable are learned through both long-term and short-term data. The long-term training set consists of nine-year hourly data for the response variable and the explanatory variables, while the short-term training set consists of the dataset for the same variables but with the data records available only for the immediate past week of the test day. Our system intends to not only capture the long-term relationship between the response variable and the explanatory variables but also reveal the recent $PM_{2.5}$ trends that have incurred by the emerging climate patterns or pollutant sources. The long-term relationship is learned by the cluster linear regression and a multi-layer perceptron, and the short-term trends are identified by a Fourier series descriptor and a multi-layer perceptron. A random forest classifier is applied to estimate the most probable value class of the $PM_{2.5}$ which then guides the multi-model strategy method to integrate the forecasts from all long-term and short-term predictors to produce the final forecast. The feature selection strategy and each of our learning models are elucidated in the following subsections.

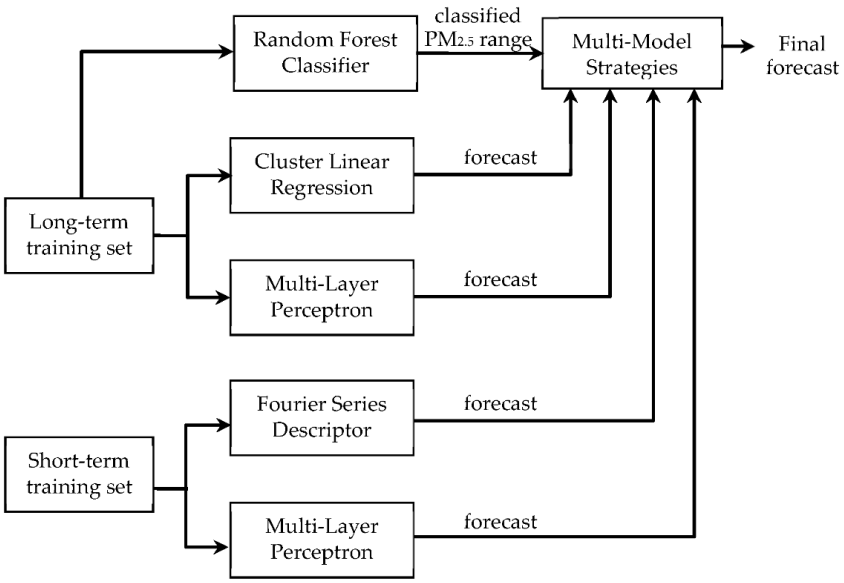


Figure 1. Architecture of our multi-model framework.

3.2. Feature Selection

The selection of effective features for forecasting $PM_{2.5}$ concentration depends on spatiotemporal meteorological patterns and local geographical terrains. This phenomenon manifests in the discrepancy among the chosen features in the literature for studies conducted at different places [7,15,16,20,29]. It resembles the selection of appropriate functional components in the software design life cycle (SDLC) of a software project. In light of this,

we adopt the *spiral model* broadly used in the SDLC field to determine the features which will be employed in our forecasting framework. The spiral model [10,11,33,34] as shown in Figure 2 repeats spiral iterations of four phases: planning, design, conduct, and evaluation. In the planning phase, system and unit specifications are acquired. Design phase starts with the software design according to the specifications. In the construct phase, the code implementation of the system prototype is fulfilled. In the evaluation phase, the users' feedback with the prototype is collected for system evaluation and risk analysis. Then the software development process enters into the next spiral iteration to enhance the feedback evaluation and resolve the risk until the system evaluation and incurred risk are acceptable.

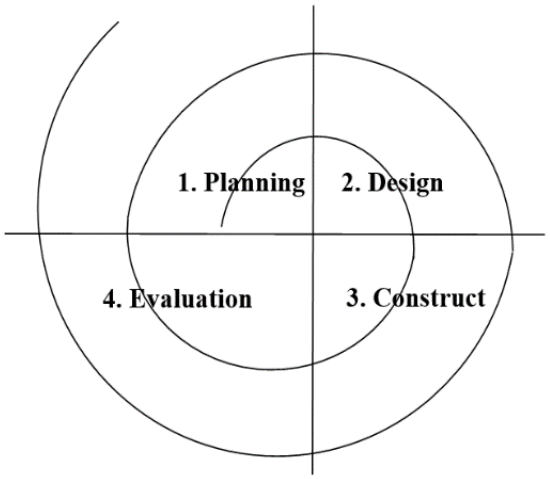


Figure 2. The spiral model of software development life cycle.

The planning and design in the SDLC depend on the system requirement and unit specifications which are not complete enough in the early SDLC phases, so the spiral model allows the software development to incrementally converge to the ideal system. Analogously, the design of effective features for forecasting $PM_{2.5}$ depends on the geographical places where the investigation is conducted because the formation of $PM_{2.5}$ concentrations are highly related to local anthropogenic and meteorological patterns. Therefore, we adapt the spiral model to a feasible feature selection model for prediction of $PM_{2.5}$ concentration. We start with an initial set of diverse features which are determined by the literature review and our own invention, such as the temporal periodical features and compound features. Then, in each spiral iteration, we respectively remove each feature from the current set to conduct the production of respective new system prototypes with the same training set. The resulting risk (prediction RMSE as will be defined) of each system prototype is evaluated from system simulation over the validation set. Based on the evaluation, the feature whose test removal has resulted in the remaining feature set having the minimum risk in terms of RMSE with the validation set is actually removed from the current feature set. Then the spiral feature validation process enters into the next spiral iteration until the current feature set contains only one feature. Finally, the best set of features can be determined by comparing the risk variations of generated prototypes along the spiral iterations. In the following, we describe our feature engineering process in detail.

Table 2 shows the initial set of features with which we start the spiral validation process for feature selection. The initial set contains 20 features determined by our own invention (the periodic features and short-term compound features as will be noted) and the literature review (the meteorological and autoregression features). These features are classified in four categories. (1) *Periodic variables*. We preliminarily examined our dataset and found a commonly observed situation in Puli Township that the daily low $PM_{2.5}$ appears in early afternoon, and the daily high $PM_{2.5}$ is usually observed around midnight. Figure 3a shows a typical example of daily periods observed within 1 February 2011 and 8 February 2011. We anticipate that this salient daily period would be a useful

indicator for adjusting the prediction for $PM_{2.5}$ concentration. Moreover, the $PM_{2.5}$ time series in our multi-year dataset also has a yearly periodic trend. Figure 3b shows the $PM_{2.5}$ yearly periodic trend within 2008 and 2016. The low monthly $PM_{2.5}$ is seen in every summer, and the high monthly $PM_{2.5}$ in a year commonly appears in winter. With the observations from daily and yearly trends of $PM_{2.5}$ variations, we propose periodic variables based on sine and cosine values of daily and yearly ordinal hour to capture these trends. (2) *Meteorological variables*. The meteorological feature variables are commonly used in the $PM_{2.5}$ forecasting literature. We include in the initial feature set the broadly used variables as follows: temperature (Temp), relative humidity (RH), precipitation (Prep), wind speed (WS), and direction sine (sin_w) and cosine (cos_w). (3) *Short-term history meteorological variables*. In addition to the meteorological condition within the current hour, the $PM_{2.5}$ concentration is strongly related to its recent weather status. For example, a constantly blowing strong wind for a couple of hours would significantly mitigate the $PM_{2.5}$ concentrations. Therefore, we record short-term history meteorological variables observed in a time window of prior six hours. To explore more effective features, we propose new compound feature variables as follows. As the wind speed (WS) and wind direction sine (sin_w) and cosine (cos_w) have cohesive meaning as a whole, some important information may be missing if these features are used independently. We propose the compound features *sinw*_WS and *cosw*_WS by calculating the sum of product of WS and sin_w and of WS and cos_w, respectively. Moreover, a complex feature combining *sinw*_WS and *cosw*_WS is proposed. The feature ST_WB, defined as the root of the sum of square *sinw*_WS and square *cosw*_WS, is found to have higher prediction capability than *sinw*_WS and *cosw*_WS, as will be noted. (4) *Short-term history autoregression variables*. The $PM_{2.5}$ series has strong autocorrelation characteristic. Figure 4 shows the autocorrelation of the $PM_{2.5}$ as a function of the time lag in hours. It is seen that the autocorrelation has a daily periodic trend and decays with the time lag length. Thus, we devise three autoregression variables within 48-h time lag.

Table 2. Initial feature variables and their removal order in the spiral validation process.

Initial Variables	Variable Descriptions	Remova Order	Finally Retained
	(1) Periodic variables		
sin_d	Sine of ordinal hour in a day	15	√
cos_d	Cosine of ordinal hour in a day	16	√
sin_y	Sine of ordinal hour in a year	3	
cos_y	Cosine of ordinal hour in a year	6	
	(2) Meteorological variables		
Temp	Temperature	14	√
RH	Relative humidity	8	
Prep	Precipitation	7	
WS	Wind speed	10	√
sin_w	Sine of wind direction	2	
cos_w	Cosine of wind direction	5	
	(3) Short-term history meteorological variables		
ST_Temp	Mean temperature in prior six hours	17	√
ST_RH	Mean relative humidity in prior six hours	19	√
ST_Prep	Mean precipitation in prior six hours	9	√
ST_WS	Mean wind speed in prior six hours	13	√
*sinw*_WS	Sum of product of *sin_w* and WS in prior six hours	1	
*cosw*_WS	Sum of product of *cos_w* and WS in prior six hours	4	
ST_WB	Rooted sum of square *sinw*_WS and square *cosw*_WS	18	√
	(4) Short-term history autoregression variables		
L_$PM_{2.5}$	Last hour $PM_{2.5}$ in the preceding day	20	√
D1_$PM_{2.5}$	Mean hourly $PM_{2.5}$ in the preceding day	11	√
D2_$PM_{2.5}$	Mean hourly $PM_{2.5}$ in the day 24 h ahead	12	√

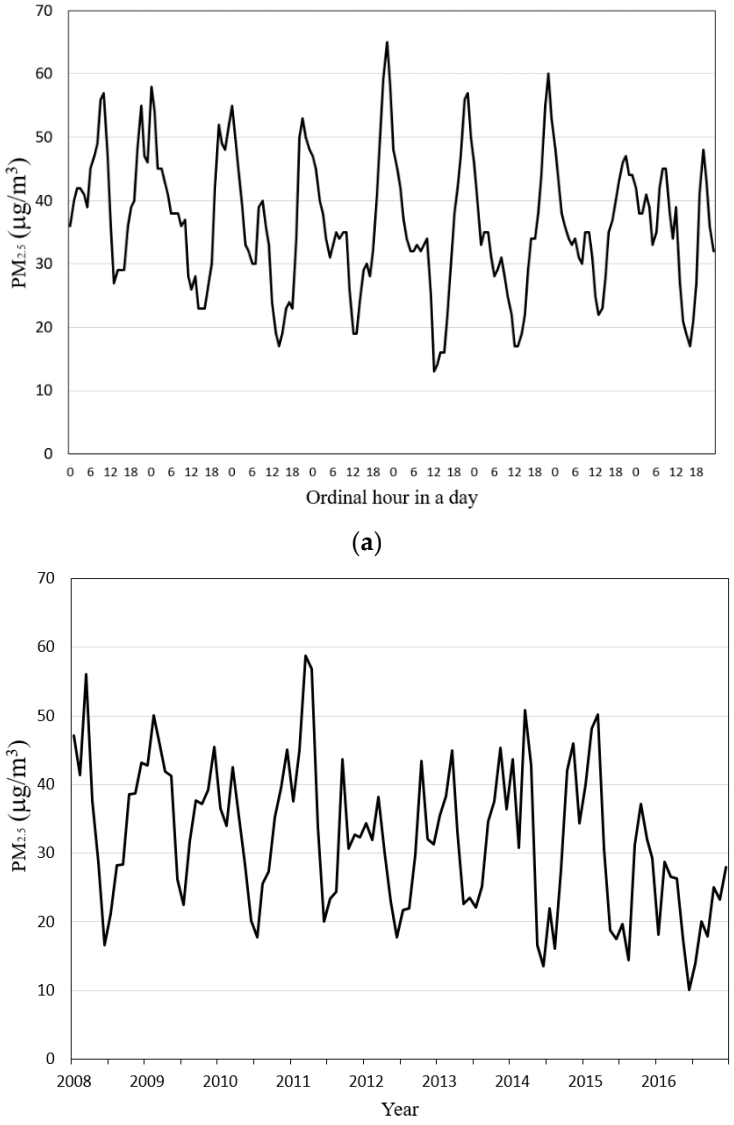


Figure 3. (**a**) The daily and (**b**) yearly periodic trends of $PM_{2.5}$ time series.

After we have prepared the initial feature set of 20 candidate features, the spiral validation process is applied to assist us to asymptotically converge to an ideal set of effective features. As previously noted, in each spiral iteration, we respectively remove each feature from the current set to conduct the production of respective new system prototypes with the same training set. The resulting risk of each system prototype is evaluated from system simulation over the validation set. The training set covers eight-year data for Puli Township from 2008 to 2015, while the validation set contains the entire year 2016 data. Based on the evaluation, the feature whose test removal has resulted in the remaining feature set having the minimum risk in terms of RMSE with the validation set is actually removed from the current feature set. Then the spiral feature validation process enters into the next spiral iteration until the current feature set contains only one feature. Finally, the best set of features can be determined by comparing the risk variations of generated prototypes along the spiral iterations. To illustrate, Figure 5 shows the variations of the risk in terms of RMSE obtained by the retained feature set in each spiral iteration by applying the spiral validation process. The order of removed features in the sequential spiral iterations is shown in Table 2. We observe that the RMSE stays at a relatively low level as the number of retained features in the feature set decreases from 12 to 7. We

finally determine to retain 12 features (i.e., removing 8 features) as indicated in Table 2 because then the final feature set contains some members from each feature category and it potentially has better generalization capability than other feature selections. Note that, our spiral validation process is a feature selection heuristic to asymptotically obtain the near-optimal feature set from 20 initial potential features. For an exhaustive search of the optimal feature set, we require to estimate in total $\binom{20}{1} + \binom{20}{2} + \ldots + \binom{20}{20}$ system prototypes which is computationally prohibitive.

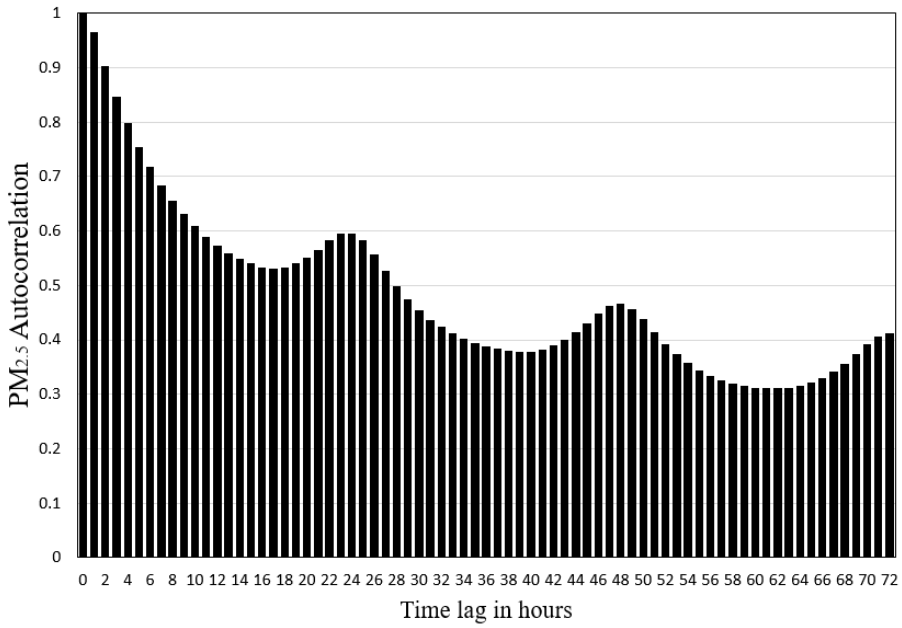


Figure 4. The autocorrelation characteristic of $PM_{2.5}$ series.

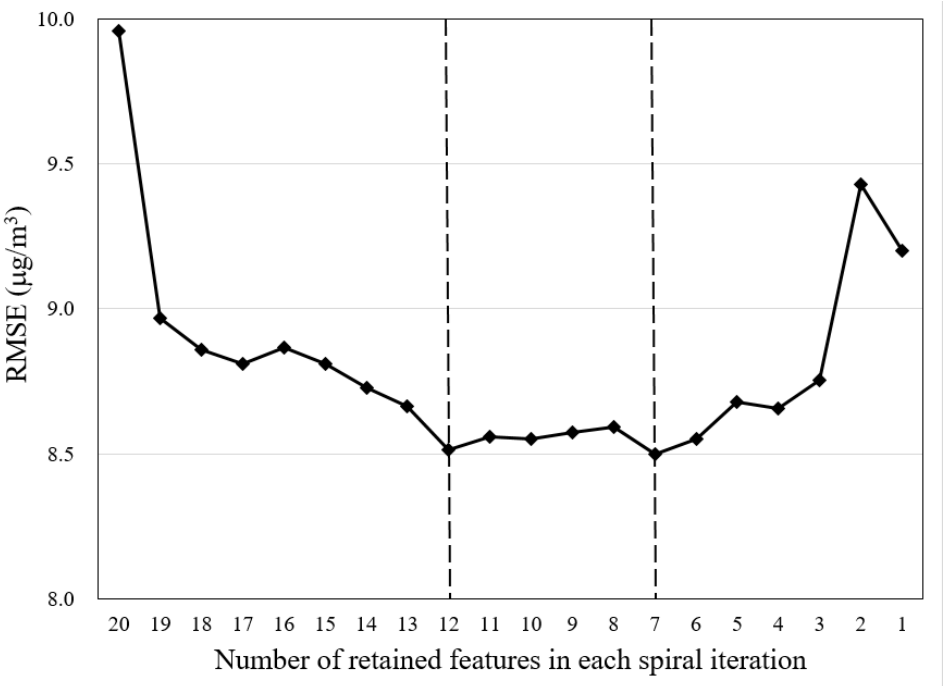


Figure 5. Variations of forecasting RMSE with the retained feature set in each spiral iteration.

Our multi-model framework has two learning tasks, (1) the relationship between the meteorological variables and the $PM_{2.5}$ concentration, and (2) the trends embodied in the

$PM_{2.5}$ series. For the first learning task, we devise the cluster linear regression and the multi-layer perceptron. The learning scheme for the second task is two-fold. One is through the feature design where we have selected periodic variables and autoregression variables as previously noted. The other is accomplished by the Fourier series descriptor which learns the main components in the $PM_{2.5}$ series. The two learning tasks are conducted in both short-term and long-term manners, as described in the following sections.

3.3. Long-Term Learning

3.3.1. Cluster Linear Regression

To investigate the relationship between the $PM_{2.5}$ concentration and the explanatory variables, some existing works deployed various forms of linear regression, such as multiple linear regression [20] and dynamic multiple regression [17]. We devise another form of linear regression, namely, the cluster linear regression, which partitions the feature space into several clusters and interprets the relationship between the features and the response variable for each cluster by linear regression. The clustering process is conducted by a decision tree with six features. The features are selected by reference to the maximal information gain principle where the feature reducing the most regression RMSE (and thus with better interpretation capability) is selected next in the decision tree. To avoid selecting redundant features, the collinearity between the next feature and the features already selected is tested. In details, the clustering starts with the initial feature set I, consisting of all 20 features listed in Table 2 and an empty set J of features for clustering. For each feature variable x contained in I, we find the optimal threshold to partition x values into two subgroups and interpret all the plots (x, $PM_{2.5}$) in each subgroup by linear regression. The optimal value of the threshold is the one which minimizes the sum of the RMSE for the two subgroups. After all features in I have been examined, the best feature x^* with the minimum sum of the RMSE is selected. We then test the Pearson collinearity between x^* and any variables already contained in J. If their absolute Pearson correlation is all no greater than 0.7, we update the two feature sets by $I = I - \{x^*\}$ and $J = J \cup \{x^*\}$. Otherwise, we only perform $I = I - \{x^*\}$ and test the next best feature x^* with the minimum sum of the RMSE. The process is iterated until J contains six features resulting in 64 clusters, i.e., the decision tree has six levels. As the amounts of our nine-year long-term data are huge, the maximal level of the decision tree is set to six by trading off the regression interpretability and the computational efficiency.

3.3.2. Multilayer Perceptron

Multilayer perceptron (MLP) is an artificial neural network in which in addition to the input layer and the output layer, one or more non-linear hidden layers can be deployed in the network. In our long-term learning, the MLP is adopted to learn the non-linear relationship between long-term meteorological data and $PM_{2.5}$ variations. Consequently, the input layer consists of neurons corresponding to the 12 features which are determined by the spiral validation process as noted in Section 3.2. The 12 features are listed in the last column of Table 2 as the finally retained features. The MLP is constructed with one hidden layer which contains 12 neurons. The output layer has only one neuron which receives the values from the last hidden layer and transforms them to the final output as the $PM_{2.5}$ forecasting.

3.4. Short-Term Learning

We contemplate the short-term trends of $PM_{2.5}$ concentration emerge via two pathways. One is formed by emerging pollution sources, such as dust-storms, burning of agricultural wastes, incense burning in a religious ceremony, or barbecue activities in festivals. The other one is incurred by particular weather patterns, such as stagnant wind or upper-level anticyclone, which cause the pollutants hard to dissipate. Both trends are temporary, usually disappear within one week. Hence, we use the data observed in the immediate past week as the training set for our short-term predictors. Two predictors are proposed for revealing

each type of the short-term trends. First, Fourier series is employed to approximate the short-term pollution trend because Fourier series is a non-parametric descriptor which only focuses on the response variable (i.e., $PM_{2.5}$ concentration). Second, the MLP is adopted to learn the short-term relationship between the changes of meteorological patterns and the variations of $PM_{2.5}$ concentration. We articulate the two short-term predictors in the following.

3.4.1. Fourier Series Descriptor

Fourier series can be used to represent a given function, say $f(t)$, in the form of the Fourier polynomial as follows.

$$f(t) = \frac{1}{2}a_0 + a_1 \cos t + b_1 \sin t + a_2 \cos 2t + b_2 \sin 2t + \ldots + a_n \cos nt + b_n \sin nt + \ldots \quad (1)$$

Since the principal information of $f(t)$ is reserved in lower degree terms, we can approximate $f(t)$ by the Fourier polynomial of degree n when n is sufficiently large. The value of coefficients a_i and b_i leading to the optimal approximation can be determined by finding the orthogonal projection of $f(t)$ onto the space spanned by the Fourier polynomial of degree n,

$$\begin{aligned} proj_{Fourier} f(t) &= \left(f(t), \tfrac{1}{\sqrt{2\pi}}\right)\tfrac{1}{\sqrt{2\pi}} \\ &+ \left(f(t), \tfrac{1}{\sqrt{\pi}}\cos t\right)\tfrac{1}{\sqrt{\pi}}\cos t + \left(f(t), \tfrac{1}{\sqrt{\pi}}\sin t\right)\tfrac{1}{\sqrt{\pi}}\sin t \\ &+ \ldots \\ &+ \left(f(t), \tfrac{1}{\sqrt{\pi}}\cos nt\right)\tfrac{1}{\sqrt{\pi}}\cos nt + \left(f(t), \tfrac{1}{\sqrt{\pi}}\sin nt\right)\tfrac{1}{\sqrt{\pi}}\sin nt \end{aligned} \quad (2)$$

To describe the short-term pollution trend, the hourly $PM_{2.5}$ data, $g(t)$, acquired in the immediate past week are converted into the target function in the domain $[-\pi, \pi]$ as follows.

$$f(t) = g\left(\frac{t+\pi}{2\pi} \times 7 \times 24\right), t \in [-\pi, \pi], \quad (3)$$

To Illustrate, Figure 6 shows the $PM_{2.5}$ series (during 13 October to 19 October 2019) approximation by using the Fourier polynomial of degree n equivalent to 50, 60, and 70, respectively. It can be observed that the Fourier polynomials with n = 50 and 60 give a better approximation than that with n = 70. To determine the best value of n, we have explored several values for the Fourier polynomial degree and found that n = 60 provides the best result as will be shown by the experiments in Section 4.2.

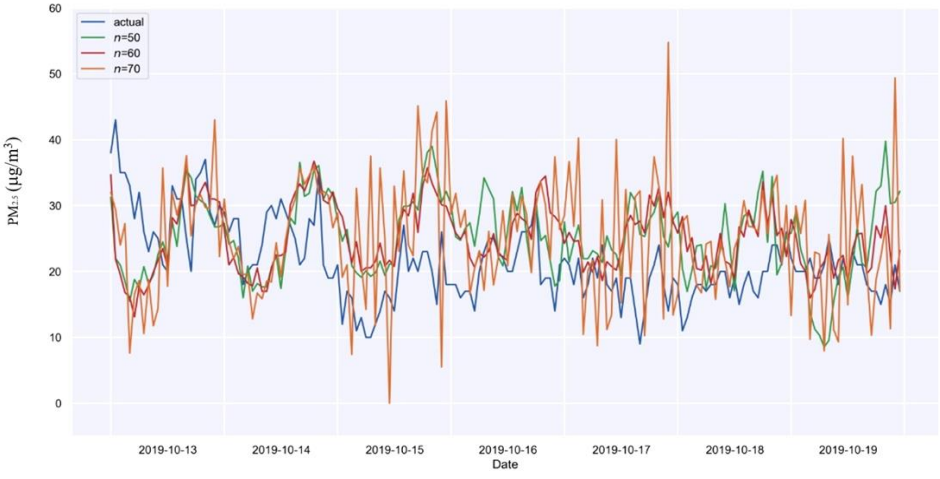


Figure 6. Illustration of $PM_{2.5}$ series approximation by using the Fourier polynomial of degree n equivalent to 50, 60, and 70, respectively.

3.4.2. Multilayer Perceptron

As noted, MLP has been deployed in our long-term learning scheme. Again, we adopt MLP in our short-term learning scheme. In contrast to the long-term MLP, which learns the main relationship between variables with the consensus over nine years, the short-term MLP aims to learn the recent emerging weather patterns that make the $PM_{2.5}$ variations unusual in long-term history. Hence, the short-term MLP is trained with the feature dataset for the immediate past week of the test day.

3.5. Multi-Model Integration Strategies

So far, we have proposed four models for $PM_{2.5}$ forecasting. Let the forecasts made by the four models be denoted by $\hat{z}_1$, $\hat{z}_2$, $\hat{z}_3$, and $\hat{z}_4$, respectively. To combine their forecasts to produce an improved one, five multi-model integration strategies and a lower bound are presented as follows.

- Averaged strategy. This strategy simply determines the final forecast as the mean of the four individual forecasts as calculated by $\hat{z} = \frac{1}{4}\sum_{i=1}^{4}\hat{z}_i$.
- Weighted strategy. The strategy considers the final forecast as the sum of weighted forecasts made by individual models. Let the prediction RMSE error estimated for the ith model during its training phase be e_i. The normalized reciprocal error is adopted as the weight for the model, i.e., $w_i = e_i^{-1} / \sum_{j=1}^{4} e_j^{-1}$. The final forecast is then calculated by $\hat{z} = \sum_{i=1}^{4} w_i \hat{z}_i$.

The rest of the strategies work with the classified value range of $PM_{2.5}$. We use the long-term training set to produce the distributions of $PM_{2.5}$ concentrations. The 10-percentile and the 90-percentile are used to separate the $PM_{2.5}$ concentrations into low, middle, and high ranges. A random forest classifier is trained with the long-term dataset with the labeled ranges. During the test process, the random forest classifier receives the inputs and estimates the most probable value range of the $PM_{2.5}$. Let the estimate made by the random forest classifier be denoted as C_{rfc}, which is L, M, or H, if the classification result is labeled as low, middle, or high $PM_{2.5}$ range. We propose three advanced multi-model integration strategies as follows.

- Max_Avg_Min strategy. This strategy determines the final forecast according to the classified value range of the $PM_{2.5}$. If the test instance is classified as in the high/low range, the maximal/minimal forecast value made by individual models is output as the final forecast. If it is classified in the middle range, the strategy outputs the same forecast value as that made by the Averaged strategy. In other words, the final prediction determined by the Max_Avg_Min strategy can be calculated as follows.

$$\hat{z} = \begin{cases} \max(\hat{z}_1, \hat{z}_2, \hat{z}_3, \hat{z}_4) & \text{if } C_{\text{rfc}} = \text{'H'} \\ \sum_{i=1}^{4} \hat{z}_i / 4 & \text{if } C_{\text{rfc}} = \text{'M'} \\ \min(\hat{z}_1, \hat{z}_2, \hat{z}_3, \hat{z}_4) & \text{if } C_{\text{rfc}} = \text{'L'} \end{cases} \tag{4}$$

- Max_Wgt_Min strategy. This strategy resembles the Max_Avg_Min strategy by assigning the maximal or minimal forecast value made by individual models as the final forecast if the classified range is high or low. However, if the test instance is classified

as in the middle range, the strategy outputs the same forecast value as that made by the weighted strategy, i.e.,

$$\hat{z} = \begin{cases} \max(\hat{z}_1, \hat{z}_2, \hat{z}_3, \hat{z}_4) & \text{if } C_{\text{rfc}} = \text{'H'} \\ \sum_{i=1}^{4} w_i \hat{z}_i & \text{if } C_{\text{rfc}} = \text{'M'} \\ \min(\hat{z}_1, \hat{z}_2, \hat{z}_3, \hat{z}_4) & \text{if } C_{\text{rfc}} = \text{'L'} \end{cases} \quad (5)$$

- Adpt_Wgt strategy. This strategy adopts the adaptive weighting scheme to produce the final forecast. In precise terms, the set of the individual forecast which falls in the classified $PM_{2.5}$ range is identified. The final forecast is determined by calculating the sum of the weighted forecasts which are contained in the identified set. So the Adpt_Wgt strategy will adapt to the models which are validated by the random forest. The Adpt_Wgt strategy can be realized by the following formula.

$$\hat{z} = \begin{cases} \sum w_i \hat{z}_i / \sum w_i & \text{if } C_{\text{rfc}} = \text{'H'} \text{ and } \forall \hat{z}_i \in \text{H} \\ \sum w_i \hat{z}_i / \sum w_i & \text{if } C_{\text{rfc}} = \text{'M'} \text{ and } \forall \hat{z}_i \in \text{M} \\ \sum w_i \hat{z}_i / \sum w_i & \text{if } C_{\text{rfc}} = \text{'L'} \text{ and } \forall \hat{z}_i \in \text{L} \end{cases} \quad (6)$$

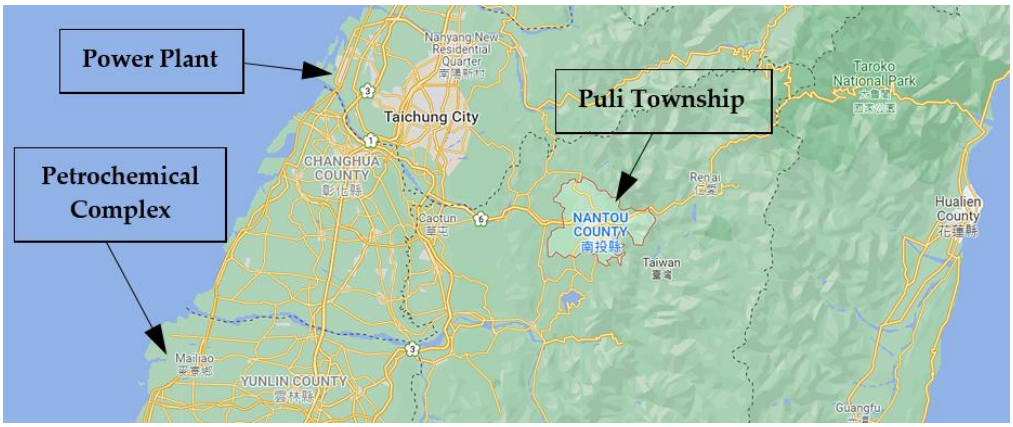
- Lower bound. To realize how well our multi-model strategies work for combining multiple forecasts, a lower bound for the forecasting error is calculated for comparison. The lower bound is the best forecasting root mean square error (RMSE) or mean average error (MAE) that could be possibly obtained by selecting a model for each prediction. That is, for each instance in the test set, the best of the four model forecasts which is nearest to the actual $PM_{2.5}$ is manually selected. After the best forecasts for all test instances have been selected, their RMSE and MAE are calculated and designated as the lower bounds. It is noted that here the lower bound is only referring to the optimal performance by model selection. It is not intended to indicate the global lower bound for any forms of multi-model hybridization.

4. Proposed Experimental Results and Comparative Performance

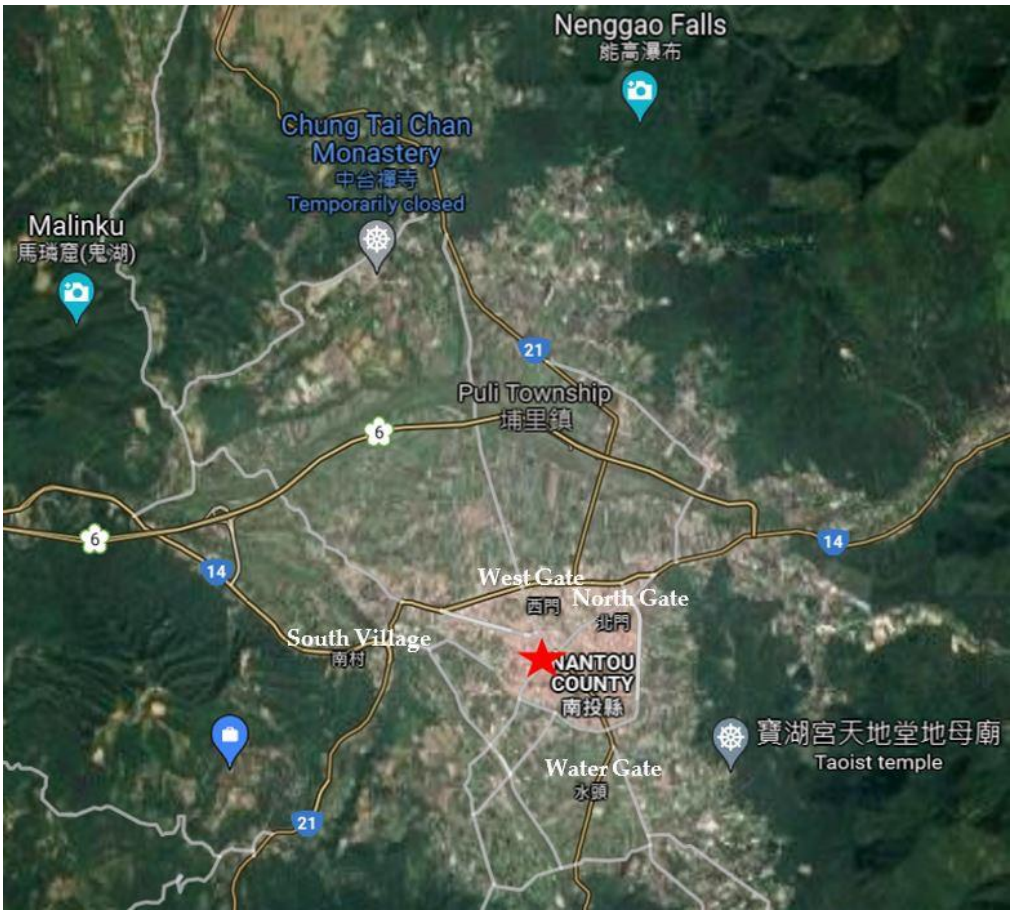
4.1. Dataset Description and Forecast Performance Measures

Our studied field is Puli Township located in Central Taiwan as shown in Figure 7a. There is a metropolitan (Taichung City) and several power plants and petrochemistry complexes to the west of Puli. Puli is a mountain basin with river outlets to the west part. To make our research practical and verifiable to the public, we chose the dataset of real $PM_{2.5}$ concentration and meteorological data values to perform comparative analysis of the predictive models, According to the $PM_{2.5}$ dataset maintained by Taiwan Environmental Protection Administration (EPA), Puli is notorious for its high air pollution rank in the list of all EPA supersites [35]. Therefore, we chose the hourly $PM_{2.5}$ dataset available at Puli supersite (https://airtw.epa.gov.tw/CHT/Query/His_Data.aspx, accessed on 16 December 2021). Figure 7b shows the basin geography of Puli Township and the location of the EPA supersite. The supersite is located in the central Puli downtown which is the most populated area with many shops, restaurants, and temples nearby. In addition to local sources, the air pollution from the western metropolitan drifts through the river valley into Puli basin. The EPA supersite applies the beta attenuation monitoring (BAM) technique for $PM_{2.5}$ measurement. The BAM employs the energy absorption of beta radiation by suspended particles extracted from the air flow. The attenuation caused by suspended particles is exponentially dependent on the particle mass in the sample. For the features to be used as the explanatory variables in our parametric models, we obtained the hourly meteorological dataset from the Taiwan Central Weather Bureau (http://e-service.cwb.gov.tw/HistoryDataQuery/, accessed on 16 December 2021), which tallies all the raw features we need, namely, temperature, relative humidity, wind speed and direction, and precipitation. The time span of the $PM_{2.5}$ and the meteorological datasets is between 2008 and 2019. The data for the early nine years (from 1 January 2008 to 31

December 2016) are designated as the long-term training set. The remaining three years (from 1 January 2017 to 31 December 2019) are used as the test set for evaluating the performance of competing models. The seven days prior to each test day are used as the short-term training set for the corresponding test.



(**a**)



(**b**)

Figure 7. Location of the studied field. (**a**) Puli Township and western plausible pollution sources. (**b**) Basin geography of Puli Township. The EPA supersite marked by a red star is located in the central Puli downtown.

To evaluate the forecast accuracy, we adopt three performance measures, namely, the root mean square error (RMSE), the mean absolute error (MAE), and the mean absolute

percentage error (MAPE), which have been broadly used in the literature and they are defined as follows.

$$\text{RMSE} = \left(\frac{1}{T} \sum_{t=1}^{T} (\hat{y}_t - y_t)^2 \right)^{\frac{1}{2}}, \tag{7}$$

$$\text{MAE} = \frac{1}{T} \sum_{t=1}^{T} |\hat{y}_t - y_t|, \tag{8}$$

$$\text{MAPE} = \frac{1}{T} \sum_{t=1}^{T} |\hat{y}_t - y_t| / y_t, \tag{9}$$

where y_t and $\hat{y}_t$ are the observed $PM_{2.5}$ value at time t and the corresponding predicted value, and T is the number of observed $PM_{2.5}$ records in the evaluation period.

The computation platform for our experiments is a personal computer with a 3.6 GHz CPU and 32 GB RAM, the programs were codified using Python 3.6.8 and the Scikit-learn package.

4.2. Performance of Single Models

Before we evaluate the proposed multi-model ensemble framework, the parameters of the individual models, cluster linear regression (CLR), long-term multilayer perceptron (LMLP), Fourier series descriptor (FSD), and short-term multilayer perceptron (SMLP), are tuned in advance to serve as a baseline for comparison. As previously described in Section 3.3.1, the CLR has been well constructed by a decision tree partition. The main parameters of the MLP are the number of hidden layers and the number of neurons in each hidden layer. We first constructed various one-hidden-layer MLPs with the number of neurons equivalent to 8, 10, 12, denoted by MLP(8), MLP(10), and MLP(12), respectively. The 3-year (2017–2019) test set forecasting performance is listed in Table 3 where the best result is shown in boldface. It is seen that MLP(12) outperforms the other counterparts. To test if the MLP performance improves with the number of hidden layers, the three-hidden-layer MLP with 12 neurons in each hidden layer, denoted by MLP(12, 12, 12), is constructed for comparison. We observe that increasing the number of hidden layers in MLP does not show the performance improvement in forecasting. This observation is consistent with the results from the literature [36], which has shown that an MLP with only one hidden layer can be satisfactorily used in different fields of engineering. In this paper, both the long-term and short-term MLPs are implemented with one-hidden-layer with 12 neurons. For the FSD, we have to determine the best degree n of the Fourier polynomial. The 3-year test set forecasting performance of the FSD, with n equivalent to 30, 40, 50, 60, 70, 80, and 90, is shown in Table 4. We see that there is no clear trend about the optimal value of n. Consider the tradeoff between the forecasting performance and computational efficiency, we chose n = 60 in our FSD setting. Finally, Table 5 tabulates the RMSE, MAE, and MAPE test performance obtained by each individual model for all days in the three-year test set. It is seen that the LMLP is the best among the four models for the three-year test set, CLR ranks at the second place, followed by FSD and SMLP.

Table 3. Three-year forecasting performance obtained by various MLPs.

	RMSE	MAE	MAPE
MLP(8)	8.63	6.36	0.42
MLP(10)	7.84	5.74	0.40
MLP(12)	**7.82**	**5.71**	**0.38**
MLP(12, 12, 12)	8.14	5.84	0.40

Table 4. Three-year forecasting performance obtained by various FSDs.

	RMSE	MAE	MAPE
FSD(30)	10.19	7.69	0.61
FSD(40)	11.89	9.23	0.73
FSD(50)	10.14	7.74	0.62
FSD(60)	**9.59**	**7.35**	**0.60**
FSD(70)	12.95	9.66	0.75
FSD(80)	9.59	7.36	0.59
FSD(90)	12.65	9.50	0.73

Table 5. Three-year forecasting performance obtained by various single models.

	RMSE	MAE	MAPE
CLR	8.94	6.29	0.43
LMLP	**7.82**	**5.71**	**0.38**
FSD	9.59	7.35	0.60
SMLP	10.73	8.04	0.60
Mean	9.27	6.85	0.50

4.3. Performance of Short-Term and Long-Term Learning Ensembles

This section compares the performance of short-term and long-term learning ensembles. The short-term ensemble is reduced from our multi-model framework by only activating the short-term models, namely, FSD and SMLP. The final forecast made by the short-term ensemble is calculated by applying the weighted strategy. Similarly, the long-term ensemble received the forecasts from CLR and LMLP only and combined them by the weighted strategy. The RMSE, MAE, and MAPE performance of short-term and long-term learning ensembles are shown in Table 6. It is seen that the forecasting performance obtained by the long-term ensemble is around the mean performance, which is achievable by the two embedded models (i.e., CLR and LMLP), and so is the performance of short-term ensemble achievable by FSD and SMLP. The implications are that the ensemble combining the forecasting results by two long-term learning models cannot create additional merits in performance improvement. Similar situation applies for the short-term ensemble. However, the short-term models have the potential to promote the performance of the long-term models although the short-term models are outperformed by the long-term models. This observation is validated by the superior performance of various multi-model strategies as shown in Table 7. The multi-model strategies enable the short-term models to compensate the test cases at which the long-term models performs worse. It is also worthy to note that the forecasting capability of the long-term ensemble may stay effective for some years. In our study, it remains effective in the three test years. However, the maximal number of straight years the long-term ensemble is applicable after a training process still needs further verification.

Table 6. Three-year forecasting performance obtained by short-term and long-term learning ensembles.

	RMSE	MAE	MAPE
Long-term ensemble	7.86	5.67	0.38
Short-term ensemble	10.42	7.83	0.58

Table 7. Three-year forecasting performance obtained by various multi-model strategies and deep-learning models.

Models	RMSE	MAE	MAPE	R^2
Averaged	8.02	5.92	0.44	0.54
Weighted	7.70	5.64	0.41	0.57
Max_Avg_Min	8.01	5.91	0.44	0.54
Max_Wgt_Min	**7.69**	**5.63**	**0.41**	**0.57**
Adpt_Wgt	8.31	6.20	0.51	0.50
Mean	7.95	5.86	0.44	0.54
Lower Bound	4.73	2.92	0.21	0.84
LSTM	8.09	5.50	0.40	0.59
CNN	9.20	6.97	0.58	0.39

4.4. Performance of Various Multi-Model Strategies

This section presents the comparative performances of our multi-model strategies. The state-of-the-art deep learning techniques, in particular, a convolutional neural network (CNN) and a long short-term memory (LSTM) network are constructed for a comparison with our multi-model strategies. The input layer of the CNN contained the same features used in our ensemble models. Five kernels are used to generate the feature maps, which were fed into a fully connected network to learn the next day 24-h $PM_{2.5}$ forecasts. The LSTM learns the autoregression relations in the original $PM_{2.5}$ time series. We applied the additive decomposition method to decompose the $PM_{2.5}$ time series into the trend series, cycle series, seasonal series, and residue series. The LSTM takes the trend, cycle, and seasonal series as model input, and uses the residue series for adjusting the model.

Table 7 shows the RMSE, MAE, MAPE, and R^2 of the forecast for all days in the three test years by applying various multi-model strategies and deep-learning models. We observe the following implications. (1) The excellent performance of the lower bound implies that the four individual models do have complementary features, which are essential to build an effective multi-model framework. (2) For our proposed multi-model strategies, the Max_Wgt_Min is the best strategy among all. The weighted strategy ranks at the second place, followed by Max_Avg_Min, Averaged, and Adpt_Wgt. (3) Both Max_Wgt_Min and weighted strategies are able to further improve the performance of the best individual model, namely, the LMLP as shown in Table 5. The remaining strategies Max_Avg_Min, Averaged, and Adpt_Wgt are inferior to LMLP, but are significantly better than CLR, FSD, and SMLP. The mean performance of all multi-model strategies overcomes the mean performance of all individual models, indicating the benefit offered by the ensemble. (4) As for the deep-learning models, LSTM significantly outperforms CNN. However, LSTM is better than Adpt_Wgt only out of our five ensemble strategies. CNN is the worst among all compared models. It is well known that CNN prevails in learning spatial information of multiple $PM_{2.5}$ sensors as revealed in [24,25]. However, this study does not consider spatial information and only uses one $PM_{2.5}$ supersite station as the forecasting reference. (5) Although our multi-model strategies can improve the performance obtained by a single model, there is still a gap to the optimal lower bound obtained by manual model selection. It is a promising direction for future research to develop a more intelligent collaborative learning strategy (e.g., deep reinforcement learning) to meet the high prediction accuracy that is currently only attainable by manual model selection.

4.5. Comparative Performances on Delhi Dataset

To test the applicability of our ensemble framework on the dataset acquired in other countries, we chose the dataset used in [32] which applies time series analysis and regression to forecast the hourly $PM_{2.5}$ concentration in the R.K. Puram area in Delhi, India. The time span of the dataset is from 1 January 2018 to 30 November 2019, and it is separated to a training set and a test set by respectively using 80% and 20% of the entire dataset. The features contained in the dataset include solar radiance, air pressure, temperature,

wind speed, wind direction and $PM_{2.5}$ time series. Kumar et al. [32] conducted a comprehensive comparison on several state-of-the-art machine learning techniques. We use the same experimental settings and apply our ensemble framework on Delhi dataset. Table 8 shows the comparative performances of various methods on Delhi dataset. We can see that all of the single models used in our framework, if they are executed as a stand-alone predictor, are inferior to all the compared models in Kumar et al. [32]. Our long-term or short-term ensemble can improve their elementary models, but still cannot be fairly comparable to the leading group. With our multi-model ensemble strategies, the forecasting performance obtained by using weighted, Max_Wgt_Min, and Adpt_Wgt is comparable to the leading methods in Kumar et al. [32]. The best ensemble, weighted strategy, can surpass all but one method, the ET + AdaBoost, in Kumar et al. [32]. This result indicates that our ensembles adapt very well to the Delhi dataset. It is worthy to note that the formation and transportation of $PM_{2.5}$ concentration are dependent on the geographical terrains and local anthropogenic activities, so the set of effective meteorological variables for forecasting $PM_{2.5}$ could vary with the studied location, as revealed by many previous researchers [7,15,16,20,29]. The features used in our ensemble models for Puli dataset are different to those used in [32]. Consider this situation, it is a promising result to note that our ensemble framework can achieve the best performance of almost all compared models for Delhi dataset.

Table 8. Comparative performances of various methods on Delhi dataset.

Sources	Models	RMSE	MAE	MAPE
Single models	CLR	28.23	19.73	0.41
	LMLP	29.43	21.45	0.41
	FSD	40.51	25.38	0.48
	SMLP	**29.32**	**19.12**	**0.40**
Ensembles	Long-term ensemble	27.16	19.17	0.38
	Short-term ensemble	27.13	18.82	0.38
	Averaged	26.75	17.81	0.34
	Weighted	**25.26**	**16.93**	**0.32**
	Max_Avg_Min	26.95	18.17	0.35
	Max_Wgt_Min	25.52	17.34	0.33
	Adpt_Wgt	25.36	17.38	0.34
	Lower Bound	16.58	9.30	0.16
Kumar et al. (2020)	Decision trees (DT)	38.13	22.18	–
	Random forest (RF)	25.83	15.21	–
	Extra trees (ET)	25.37	15.04	–
	DT + AdaBoost	25.40	14.46	–
	RF + AdaBoost	25.30	14.99	–
	ET + AdaBoost	**25.11**	**14.79**	–
	LSTM	28.97	16.66	–

5. Conclusions

In this paper, we proposed a multi-model framework for $PM_{2.5}$ forecasting. The framework combines four diverse learning models, namely, cluster linear regression (CLR), long-term multi-layer perceptron (LMLP), Fourier series descriptor (FSD), and the short-term multi-layer perceptron (SMLP). The feature set fed into the multi-model framework is selected by a spiral validation process which evaluates the risk for eliminating every feature. We explore the collaborations between parametric and non-parametric learning, short-term and long-term learning. Our experiments with Puli dataset spanning 1 January 2008 to 31 December 2019, show that the proposed multi-model framework can synergize the advantages of the embedded models and can obtain an improved forecasting result. We also show that the performance obtained by a mixture of short-term and long-term learning can surpass that obtained by applying just a single type of learning. A promising direction

for future research is the application of evolutionary instance selection [37] to appropriately train the forecasting system under the environmental scenarios of current testing.

Author Contributions: Conceptualization, P.-Y.Y., A.Y.Y. and R.-F.D.; methodology, P.-Y.Y. and A.Y.Y.; software, S.-E.C.; validation, S.-E.C.; writing—original draft preparation, P.-Y.Y. and B.B.; writing—review and editing, P.-Y.Y. and B.B.; visualization, S.-E.C.; funding acquisition, P.-Y.Y. and R.-F.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Technology of ROC, under grant MOST 110-2410-H-163-001-MY2, grant MOST 107-2410-H-260-015-MY3, grant MOST 107-2420-H-260-002-HS3, and the Environmental Protection Administration of ROC, under grant EPA-107-FA12-03-A150.

Data Availability Statement: The raw data used in this study can be access at https://airtw.epa.gov.tw/CHT/Query/His_Data.aspx and http://e-service.cwb.gov.tw/HistoryDataQuery/, accessed on 16 December 2021.

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

References

1. Lee, W.C.; Shen, L.; Catalano, P.J.; Mickley, L.J.; Koutrakis, P. Effects of future temperature change on $PM_{2.5}$ infiltration in the Greater Boston area. *Atmos. Environ.* **2017**, *150*, 98–105. [CrossRef]
2. Liang, C.S.; Duan, F.K.; He, K.B.; Ma, Y.L. Review on recent progress in observations, source identifications and countermeasures of $PM_{2.5}$. *Environ. Int.* **2016**, *86*, 150–170. [CrossRef] [PubMed]
3. Hwang, S.L.; Lin, Y.C.; Guo, S.E.; Chi, M.C.; Chou, C.T.; Lin, C.M. Emergency room visits for respiratory diseases associated with ambient fine particulate matter in Taiwan in 2012: A population-based study. *Atmos. Pollut. Res.* **2017**, *8*, 465–473. [CrossRef]
4. Song, C.; He, J.; Wu, L.; Jin, T.; Chen, X.; Li, R.; Ren, P.; Zhang, L.; Mao, H. Health burden attributable to ambient $PM_{2.5}$ in China. *Environ. Pollut.* **2017**, *223*, 575–586. [CrossRef] [PubMed]
5. Chen, Y.C.; Chiang, H.C.; Hsu, C.Y.; Yang, T.T.; Lin, T.Y.; Chen, M.J.; Chen, N.T.; Wu, Y.S. Ambient $PM_{2.5}$-bound polycyclic aromatic hydrocarbons (PAHs) in Changhua County, Central Taiwan: Seasonal variation, source apportionment and cancer risk assessment. *Environ. Pollut.* **2016**, *218*, 372–382. [CrossRef] [PubMed]
6. WHO Media Centre. Ambient (Outdoor) Air Quality and Health. 2016. Available online: http://www.who.int/mediacentre/factsheets/fs313/en/ (accessed on 16 December 2021).
7. Di, Q.; Koutrakis, P.; Schwartz, J. A hybrid prediction model for $PM_{2.5}$ mass and components using a chemical transport model and land use regression. *Atmos. Environ.* **2016**, *131*, 390–399. [CrossRef]
8. Wang, P.; Zhang, H.; Qin, Z.; Zhang, G. A novel hybrid-Garch model based on ARIMA and SVM for $PM_{2.5}$ concentrations forecasting. *Atmos. Pollut. Res.* **2017**, *8*, 850–860. [CrossRef]
9. Zhang, B.; Li, X.; Zhao, Y.; Li, Y.; Wang, X. Air quality $PM_{2.5}$ prediction based on multi-model fusion. In Proceedings of the 2019 Chinese Control and Decision Conference (CCDC), Nanchang, China, 3–5 June 2019.
10. Pew, R.W.; Mavor, A.S. (Eds.) *Human-System Integration in The System Development Process: A New Look*; National Academy Press: Washington, DC, USA, 2007.
11. Shylesh, S. *A Study of Software Development Life Cycle Process Models*; Elsevier SSRN: Amsterdam, The Netherlands, 2017.
12. Vlachogianni, A.; Kassomenos, P.; Karppinen, A.; Karakitsios, S.; Kukkonen, J. Evaluation of a multiple regression model for the forecasting of the concentrations of NOx and PM10 in Athens and Helsinki. *Sci. Total Environ.* **2011**, *409*, 1559–1571. [CrossRef]
13. Cobourn, W.G. An enhanced $PM_{2.5}$ air quality forecast model based on nonlinear regression and back-trajectory concentrations. *Atmos. Environ.* **2010**, *44*, 3015–3023. [CrossRef]
14. Baker, K.R.; Foley, K.M. A nonlinear regression model estimating single source concentrations of primary and secondarily formed $PM_{2.5}$. *Atmos. Environ.* **2011**, *45*, 3758–3767. [CrossRef]
15. Yin, Q.; Wang, J.; Hu, M.; Wong, H. Estimation of daily $PM_{2.5}$ concentration and its relationship with meteorological conditions in Beijing. *J. Environ. Sci.* **2016**, *48*, 161–168. [CrossRef]
16. Guo, Y.; Tang, Q.; Gong, D.Y.; Zhang, Z. Estimation ground-level $PM_{2.5}$ concentrations in Beijing using a satellite-based geographically and temporally weighted regression model. *Remote Sens. Environ.* **2017**, *198*, 140–149. [CrossRef]
17. Moisan, S.; Herrera, R.; Clements, A. A dynamic multiple equation approach for forecasting $PM_{2.5}$ pollution in Santiago, Chile. *Int. J. Forecast.* **2018**, *34*, 566–581. [CrossRef]
18. Zhang, T.; Liu, P.; Sun, X.; Zhang, C.; Wang, M.; Xu, J.; Pu, S.; Huang, L. Application of an advanced spatiotemporal model for $PM_{2.5}$ prediction in Jiangsu Province, China. *Chemosphere* **2020**, *246*, 125563. [CrossRef] [PubMed]
19. Ni, X.Y.; Huang, H.; Du, W.P. Relevance analysis and short-term prediction of $PM_{2.5}$ concentrations in Beijing based on multi-source data. *Atmos. Environ.* **2017**, *150*, 146–161. [CrossRef]
20. Ausati, S.; Amanollahi, J. Assessing the accuracy of ANFIS, EEMD-GRNN, PCR, and MLR models in predicting $PM_{2.5}$. *Atmos. Environ.* **2016**, *142*, 465–474. [CrossRef]

21. Niu, M.; Gan, K.; Sun, S.; Li, F. Application of decomposition-ensemble learning paradigm with phase space reconstruction for day-ahead $PM_{2.5}$ concentration forecasting. *J. Environ. Manag.* **2017**, *196*, 110–118. [CrossRef]
22. Mao, X.; Shen, T.; Feng, X. Prediction of hourly ground level $PM_{2.5}$ concentrations 3 days in advance using neural networks with satellite data in eastern China. *Atmos. Pollut. Res.* **2017**, *8*, 1005–1015. [CrossRef]
23. Di, Q.; Amini, H.; Shi, L.; Kloog, I.; Silvern, R.; Kelly, J.; Sabath, M.B.; Choirat, C.; Koutrakis, P.; Lyapusting, A.; et al. An ensemble-based model of $PM_{2.5}$ concentration across the contiguous United States with high spatiotemporal resolution. *Environ. Int.* **2019**, *130*, 104909. [CrossRef]
24. Xiao, F.; Yang, M.; Fan, H.; Fan, G.; Al-Qaness, M.A.A. An improved deep learning model for predicting daily $PM_{2.5}$ concentration. *Sci. Rep.* **2020**, *10*, 20988. [CrossRef]
25. Qin, D.; Yun, J.; Zou, G.; Yong, R.; Zhao, Q.; Zhang, B. A novel combined prediction scheme based on CNN and LSTM for urban PM25 concentration. *IEEE Access* **2019**, *7*, 20050–20059. [CrossRef]
26. Zhu, H.; Fan, L. $PM_{2.5}$ forecasting based on artificial neural network and genetic algorithm. *Int. J. Simul. Syst. Sci. Technol.* **2015**, *16*, 10.1–10.5.
27. Zhang, C.J.; Dai, L.J.; Ma, L.M. Rolling forecasting model of $PM_{2.5}$ concentration based on support vector machine and particle swarm optimization. In Proceedings of the International Symposium on Optoelectronic Technology and Application 2016, Beijing, China, 9–11 May 2016; p. 101561I. [CrossRef]
28. Sun, W.; Sun, J. Daily $PM_{2.5}$ concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm. *J. Environ. Manag.* **2017**, *188*, 144–152. [CrossRef]
29. Dhyani, R.; Sharma, N.; Maity, A.K. Prediction of $PM_{2.5}$ along urban highway corridor under mixed traffic conditions using CALINE4 model. *J. Environ. Manag.* **2017**, *198*, 24–32. [CrossRef] [PubMed]
30. Tsai, Y.I.; Sopajaree, K.; Kuo, S.C.; Yu, S.P. Potential $PM_{2.5}$ impacts of festival related burning and other inputs on air quality in an urban area of southern Taiwan. *Sci. Total Environ.* **2015**, *527–528*, 65–79. [CrossRef] [PubMed]
31. Reff, A.; Eberly, S.I.; Bhave, P.V. Receptor modeling of ambient particulate matter data using positive matrix factorization: Review of existing methods. *J. Air Waste Manag. Assoc.* **2007**, *57*, 146–154. [CrossRef]
32. Kumar, S.; Mishra, S.; Singh, S.K. A machine learning-based model to estimate $PM_{2.5}$ concentration levels in Delhi's atmosphere. *Heliyon* **2020**, *6*, e05618. [CrossRef]
33. Boehm, B.W. A spiral model of software development and enhancement. *IEEE Comput.* **1988**, *21*, 61–72. [CrossRef]
34. Boehm, B.W. *Spiral Development: Experience, Principles, and Refinements*; Special Report; CMU/SEI-2000-SR-008; Software Engineering Institute: Pittsburgh, PA, USA, 2000.
35. Hsu, C.H.; Cheng, F.Y. Classification of weather patterns to study the influence of meteorological characteristics on $PM_{2.5}$ concentrations in Yunlin County, Taiwan. *Atmos. Environ.* **2016**, *144*, 397–408. [CrossRef]
36. Govindaraju, R.S. ASCE Task Committee on Application of Artificial Neural Networks in Hydrology 2000. Artificial neural networks in hydrology. II: Hydrology applications. *J. Hydrol. Eng.* **2000**, *5*, 124–137.
37. Derrac, J.; García, S.; Herrera, F. A survey on evolutionary instance selection and generation. *Int. J. Appl. Metaheuristic Comput.* **2010**, *1*, 60–92. [CrossRef]