

A novel model for hourly PM_{2.5} concentration prediction based on CART and EELM

Zhigen Shang^{a*}, Tong Deng^b, Jianqiang He^a, Xiaohui Duan^a

^a *Department of Automation, Yancheng Institute of Technology, Yancheng 224051, China*

^b *The Wolfson Centre for Bulk Solids Handling Technology, Faculty of Engineering & Science, University of Greenwich, Kent, ME4 4TB, UK*

ABSTRACT

Hourly PM_{2.5} concentrations have multiple change patterns. For hourly PM_{2.5} concentration prediction, it is beneficial to split the whole dataset into several subsets with similar properties and to train a local prediction model for each subset. However, the methods based on local models need to solve the global-local duality. In this study, a novel prediction model based on classification and regression tree (CART) and ensemble extreme learning machine (EELM) methods is developed to split the dataset into subsets in a hierarchical fashion and build a prediction model for each leaf. Firstly, CART is used to split the dataset by constructing a shallow hierarchical regression tree. Then at each node of the tree, EELM models are built using the training samples of the node, and hidden neuron numbers are selected to minimize validation errors respectively on the leaves of a sub-tree that takes the node as the root. Finally, for each leaf of the tree, a global and several local EELMs on the path from the root to the leaf are compared, and the one with the smallest validation error on the leaf is chosen. The meteorological data of Yancheng urban area and the air pollutant concentration data from City Monitoring Centre are used to evaluate the method developed. The experimental results demonstrate that the method developed addresses the global-local duality, having better performance than global models including random forest (RF), ν -support vector regression (ν -SVR) and EELM, and other local models based on season and k -means clustering. The new model has improved the capability of treating multiple change patterns.

Keywords: PM_{2.5} concentration prediction; Local model; Classification and regression tree (CART); Extreme learning machine (ELM); Ensemble model

1. Introduction

Predicting concentrations of particulate matter in the air is important for control and reduction of airborne pollution. Particulate matter refers to small particles consisting of dust, dirt, soot, smoke, and liquid droplets suspended in the air. These particles vary widely in size (aerodynamic diameter). Particles with an aerodynamic diameter less than 2.5 μ m (PM_{2.5}) are known as fine particulates, which are seriously harmful to human health because of its absorption of toxic substances such as carcinogenic organic compounds and heavy metals (Thomaidis et al., 2003; Zhou et al., 2014). Exposure to high concentrations of PM_{2.5} has been linked to many diseases, such as diabetes, lung cancer, respiratory

*Corresponding author at: Department of Automation, Yancheng Institute of Technology, Yancheng 224051, China.
E-mail address: zgshang@ycit.edu.cn (Z. Shang).

32 and cardiovascular diseases (Requia et al., 2017). It is thought that prediction of PM_{2.5} concentrations is beneficial to
33 improve early warning procedures.

34 In recent years, a variety of models have been developed to predict PM_{2.5} concentrations (Niu et al., 2016; Wang et
35 al., 2017). These models are basically classified into two categories: deterministic and statistical models. Deterministic
36 models, called chemical transport models (CTMs) focus on understanding the underlying complex interactions among
37 meteorology, chemistry and emission. CTMs simulate the atmospheric chemistry and physics in the emission, transport
38 and transformation processes. Thus, CTMs require sufficient information of pollutant sources, reaction mechanisms and
39 chemical kinetics and so on (Sun et al., 2013). Insufficient information of pollutant sources and improper representation
40 of physicochemical processes limit their application in many places (Qin et al., 2014). Another disadvantage of CTM
41 approaches is high computational cost (Doraiswamy et al., 2010). On the other hand, statistical models aim to develop
42 the relationships between selected input variables and air pollutant concentrations using various regression models.
43 These models usually need a sufficient amount of historical data from monitoring stations. Compared to deterministic
44 models, statistical models have the advantage of easy, quick and economical implementation when given sufficient
45 historical data (Wang et al., 2017). Furthermore, statistical approaches are generally more competent in capturing the
46 underlying site-specific dependencies between air pollutant concentrations and selected variables. Therefore, statistical
47 models are less complex, computationally intensive but more accurate (Perez, 2012).

48 Statistical models mainly include linear and generalized linear regression, nonlinear regression, autoregressive
49 integrated moving average (ARIMA), hidden Markov model (HMM), random forest (RF), support vector regression
50 (SVR) and artificial neural network (ANN). Vlachogianni et al. (2011) adopted the linear regression model to predict
51 NO_x and PM₁₀ concentrations using NO, NO₂, CO, O₃ and PM_{2.5} concentrations. The generalized linear model was used
52 to predict PM₁₀ concentrations in urban areas (Garcia et al., 2016). Cobourn (2010) presented a PM_{2.5} prediction model
53 based on nonlinear regression and back-trajectory PM_{2.5} concentrations. However, the methods based on linear,
54 generalized linear and nonlinear regression tend to oversimplify the relationships between air pollutant concentrations
55 and predictor variables. Ni et al. (2017) designed an ARIMA time series model to explore the prediction of PM_{2.5} in the
56 short-term time series. However, the ARIMA model, being a linear model, cannot be well adapted to the nonlinear air
57 pollutant series (Niu et al., 2016). Dong et al. (2009) developed a method based on the hidden semi-Markov model to
58 predict PM_{2.5} concentration levels. Moreover, Sun et al. (2013) proposed an HMM with different emission distributions
59 to predict 24-hour-average PM_{2.5} concentrations in Northern California. The HMM model, however, suffers from
60 several inherent shortcomings, such as computationally expensive training and its sensitivity to initial condition
61 (Budalakoti et al., 2009). Random forest is a substantial modification of bagging that builds a large collection of
62 de-correlated trees, and then averages them (Breiman, 2001). Yu et al. (2016) used the random forest algorithm to

63 predict air quality for urban sensing systems. Support vector regression (SVR) is based on structural risk minimization,
64 and it has a unique global optimal solution. Kernel function applied in SVR maps the original space into a
65 high-dimensional feature space, where a linear regression model fits the data more appropriately. The SVR is widely
66 utilized to predict $PM_{2.5}$ and other pollutant concentrations (Lu and Wang, 2005; Niu et al., 2016; Xu et al., 2017).
67 However, several model parameters in SVR play a critical role in prediction performance. SVR requires considerable
68 computational cost to fine-tune the parameters. The ANN models have also gained popularity in their use for the
69 prediction of $PM_{2.5}$ concentrations due to the capability of detecting complex underlying nonlinear relationships
70 (McKendry, 2002; Qin et al., 2014; Voukantsis et al., 2011). However, the results of an ANN model are sensitive to the
71 selection of model parameters, and the model requires considerable expertise for fine-tuning the parameters.

72 Extreme learning machine (ELM), proposed by Huang et al. (2006), is an efficient learning algorithm for
73 single-hidden layer feedforward neural networks (SLFNs). In ELM, the parameters of the hidden layer are randomly
74 initialized, and the weights of the output layer are analytically computed by using Moore-Penrose generalized inverse.
75 Thus, ELM model can be run in an extremely low computational time. However, Moore-Penrose generalized inverse
76 leads ELM to suffer from the overfitting problem. Wang et al. (2017) utilized differential evolution (DE) to optimize the
77 parameters of the hidden layer for predicting air quality index. In addition to optimization algorithm, ensemble
78 algorithm can also be used to improve the performance of ELM (Zhou et al., 2002).

79 The selection of input variables for an ANN based prediction model is a critical issue, since irrelevant or noisy
80 variables may result to an unnecessarily complex model and poor generalization (Voukantsis et al., 2011). The linear
81 framework, including Pearson correlation coefficient, multiple linear regression and forward stepwise regression and so
82 on, has been widely used to select a reduced set of input variables (Cobourn, 2010; Díaz-Roles et al., 2008; Ni et al.,
83 2017; Russo et al., 2013). However, the linear framework has the limited capability of stimulating the complex
84 relationships between air pollutant concentrations and input variables. Voukantsis et al. (2011) employed a combination
85 of linear regression and ANN to select input variables for air pollutant prediction. To be completely free from the usual
86 linear framework, Poggi and Portier (2011) used the RF variable importance to determine input variables for PM_{10}
87 prediction. In the RF framework, the most widely used measure of importance of a given variable is the mean difference
88 of prediction errors of the trees (mean squared error (MSE) for regression) before and after the values of this variable
89 are randomly permuted in the out-of-bag (OOB) samples (Poggi and Portier, 2011). If a variable is important, a random
90 permutation will degrade the prediction, and the mean difference will be a large value.

91 Due to seasonal variations and diurnal variations, $PM_{2.5}$ concentrations have multiple change patterns (He et al.,
92 2017; Kassomenos et al., 2014; Liu et al., 2015). McKendry (2002) suggested that for air pollution prediction, hybrid
93 models and local multi-layer perceptions (MLPs) may be superior to a single global MLP. Hybrid models try to

94 construct a set of prediction models and combine them. [Díaz-Roles et al. \(2008\)](#) constructed a hybrid model,
95 aggregating the outputs of ARIMA and ANN to capture different patterns in PM₁₀ concentrations. The hybrid model
96 developed by [Perez \(2012\)](#) used nearest neighbour model as a correction which was applied to the results obtained from
97 the ANN model. Recently, hybrid models based on time series decomposition have been gradually developed, which
98 mainly consist of the following three steps: (1) decompose the output time series into several components; (2) train a
99 prediction model for each component; (3) aggregate the outputs from different component models. For air pollutant,
100 some hybrid prediction models apply wavelet transform as a tool to decompose output time series ([Bai et al., 2016](#);
101 [Osowski and Garanty, 2007](#)). Additionally empirical mode decomposition (EMD) and its variations also have been used
102 to decompose the original time series ([Niu et al., 2016](#); [Wang et al., 2017](#); [Xu et al., 2017](#); [Zhou et al., 2014](#); [Zhu et al.,](#)
103 [2017](#)).

104 Compared to hybrid models, local models split the input space into subspaces with similar properties and construct a
105 prediction model for each subspace. In this study, global models are referred as single models developed on the total
106 training data wherein air pollutant concentrations are from an air quality monitoring station, but local models are trained
107 on the subsets of the training data used for global models. The local models based on season divide the training set into
108 seasonal subsets, and train a model for individual season. [Perez and Gramsch \(2016\)](#) developed an ANN model to
109 predict hourly PM_{2.5} in Santiago de Chile for the season when high concentration episodes occur frequently, with the
110 training data from the same season for years 2010 and 2011. To improve air pollutant prediction, [Feng et al. \(2011\)](#) used
111 clustering method to divide the dataset into several clusters and built an ANN for individual cluster. However, the
112 output was ignored when using clustering algorithm to split the dataset, which was unreasonable. Moreover, [Bettenburg](#)
113 [et al. \(2012\)](#) found that balancing thinking locally and acting globally is important for local models. Normally global
114 models have the risk of underfitting due to multiple patterns, while local models tend to suffer from the overfitting
115 problem. Furthermore, a global model may be beneficial for learning some patterns. Consequently, air pollutant
116 prediction based on local models needs to address the global-local duality.

117 This study develops a framework based on CART and EELM to deal with the global-local duality. Through
118 constructing a shallow regression tree by using CART, the whole dataset is divided into subsets in a hierarchical manner.
119 For each node of the tree, the EELMs are trained using the samples belonging to the node, and hidden node numbers are
120 selected to minimize validation errors respectively on the leaves of a sub-tree that takes the node as the root. For each
121 leaf, there are a global and several local EELMs on the path from the root node to the leaf, and the EELM with the
122 smallest validation error on the leaf is selected. Before the implementation of the CART-EELM, input variables are
123 selected by using the RF model. The meteorological data of Yancheng urban area and the air pollutant concentration
124 data from City Monitoring Centre are used to evaluate the capability of the CART-EELM model in dealing with the

125 global-local duality.

126 The rest of this paper is organized as follows: CART and EELM are introduced, and CART-EELM is further
127 proposed in Section 2. In Section 3, our CART-EELM is evaluated with comparison to several other models. Section 4
128 draws the final conclusions.

129 2. Methodology

130 2.1. Selection of input variables using RF

131 In Poggi and Portier (2011), input variables were selected for PM₁₀ prediction through the analysis of the RF
132 variable importance. In this study, the RF model is also employed to provide variable importance ranking, but the
133 variable selection is performed based on the cross-validation error of the RF model. The selection of input variables
134 consists of the following steps:

135 Step1: Assess the OOB error of the RF model, compute the importance scores of input variable candidates, and rank
136 the candidates in a descending order of importance. To minimize sampling effects, we run the RF ten times on the
137 training set, and the importance score of each candidate is the mean of the scores observed from ten RF models.

138 Step 2: Invoke the most important k variables at the beginning, implement sequential introduction, and use the
139 cross-validation error of the RF model to evaluate the different combinations of input variable candidates. The
140 combination with the lowest error is chosen.

141 Step 3: Return to Step 1 until no further candidates can be rejected.

142 2.2. CART

143 CART, developed by Breiman et al. (1984), explores the structure of the training set and generates easily
144 understandable decision rules for regression or classification. The basic idea of the algorithm is to recursively partition
145 the input space into binary subsets where the output becomes successively more homogeneous.

146 Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^l$ be a set of training samples, where $\mathbf{x}_i \in \mathbf{R}^m$ is the i th input vector and $y_i \in \mathbf{R}$ is the
147 corresponding output. CART begins with the root node, which is associated with the entire input space \mathbf{R}^m . The root
148 node contains all the training samples. The next step is calculating the first split. For a regression problem, the split is to
149 minimize the expected sum variances for two resulting subsets:

$$\begin{aligned} & \min_{j,c} \frac{1}{l} \left(\sum_{k \in S_L} (y_k - \bar{y}_L)^2 + \sum_{k \in S_R} (y_k - \bar{y}_R)^2 \right) \\ & \text{s.t. } S_L = \{i | \mathbf{x}_{ij} \leq c, i = 1, \dots, l\}, \\ & \quad S_R = \{i | \mathbf{x}_{ij} > c, i = 1, \dots, l\}, \\ & \quad j \in \{1, \dots, m\} \end{aligned} \tag{1}$$

151 where S_L and S_R are the sets of training indices going to left child node and right child node, \bar{y}_L and \bar{y}_R are the mean

152 values of the outputs of samples in two subsets. The optimal j and c can be easily determined by discrete search over the
153 m input dimensions and l samples. The children of the root node are recursively split in the same manner until some
154 stop criterion is satisfied.

155 CART has low computational complexity because of its recursive computation. By moving from the root node to
156 the terminal node (leaf), each example is then assigned to a unique leaf where the predicted value is determined.
157 Normally, the mean value of the outputs of samples fallen into the leaf is chosen as the predicted value. So CART is
158 nonparametric and can find complex relationships between input and output variables. Therefore, CART also has the
159 advantage of discovering nonlinear structures and variables interactions in the training samples (Brezigar-Masten and
160 Masten, 2012).

161 Since the split in CART aims to minimize the diversity of outputs, the model is a natural fit for pattern segmentation.
162 In this study, the results of CART are used to replace k -means clustering, where CART is used to segment the change
163 patterns of hourly PM_{2.5} concentrations. The k -means clustering is a kind of unsupervised learning, while CART is a
164 form of supervised learning. So CART has the advantage of considering the output when segmenting patterns.

165 2.3. EELM

166 In ELM, the hidden layer parameters are randomly initialized. ELM is mathematically modelled by

$$167 \quad o_i = \mathbf{g}(\mathbf{x}_i)^T \boldsymbol{\mu}, \quad (2)$$

168 where $\mathbf{g}(\mathbf{x}_i) \in \mathbf{R}^p$ is the output vector of the hidden layer, $\boldsymbol{\mu} \in \mathbf{R}^p$ is the output weight vector, and p is the number
169 of the hidden neurons. The RBF nodes are used in this study, therefore Eq. (2) is rewritten as

$$170 \quad o_i = \sum_{j=1}^p \mu_j \exp(-\|\mathbf{x}_i - \mathbf{a}_j\|^2 / 2b_j^2), \quad (3)$$

171 where $\mathbf{a}_j \in \mathbf{R}^m$ and $b_j \in \mathbf{R}^+$ are the centre and impact factor of j th RBF node. ELM computes the output weight
172 vector $\boldsymbol{\mu} = \mathbf{H}^+ \mathbf{y}$ by Moore-Penrose generalized inverse, where $\mathbf{H} = [g(\mathbf{x}_1), \dots, g(\mathbf{x}_l)]^T$, and $\mathbf{y} = [y_1, \dots, y_l]^T$. Thus,
173 the ELM has an extremely low computational time. However, Moore-Penrose generalized inverse leads ELM to suffer
174 from the overfitting problem.

175 Ensemble algorithm is one of methods to improve the generalization performance of ELM. Several practical works
176 showed that the performance of a single ELM can be improved by using properly ensemble techniques, which develop a
177 population of ELM-based learners and then combine them to create improved results (Xue et al., 2014).

178 Bagging and boosting are typical representatives of ensemble methods. Khoshgoftaar et al. (2011) proved that
179 bagging generally outperforms boosting. Xue et al. (2014) proposed a novel ensemble of extreme learning machine
180 based on Genetic algorithms (GE-ELM), which outperforms both bagging and boosting. However, GE-ELM needs

181 considerably more training time. It is worth noting that in this study, any ensemble methods can be used. By considering
182 performance and computational time, bagging is chosen to ensemble ELMs. If it is assumed that all ELMs in EELM
183 have the same number of hidden neurons, there is only one parameter in the model, hidden neuron number, which needs
184 to be optimized when using EELM.

185 *2.4. CART-EELM model*

186 For prediction of hourly PM_{2.5} concentration, a novel method on training local models based on combination of
187 CART and EELM is developed, which aims at addressing the global-local duality and improving the prediction
188 accuracy. The algorithm developed includes the following steps:

189 Step 1: Construct the CART tree using the training set. The deeper branches in the tree may be affected by outliers.
190 Moreover, local models are trained at each non-root node of the tree in our algorithm. Thus, a shallow tree is
191 constructed to capture concentration change patterns and to ensure that each leaf has enough training samples for its
192 local model. To generate a shallow tree, a large value of the minimum number of samples in a leaf is set. Furthermore,
193 considering the fact that samples with the low-value outputs may take the most of the dataset, the maximum depth of
194 the tree is also set to prevent these samples from being excessively split. A CART tree is constructed using the total
195 training samples, and is then applied for splitting the validation samples. So each node has its own training and
196 validation samples.

197 Step 2: Train EELMs using the associated samples. Each node in the hierarchical tree trains EELMs using the
198 associated training samples of the node. At the root node, global models use the total training set, but utilize different
199 validation subsets from the leaves of the tree to respectively determine the model parameters. At each internal node,
200 local EELMs are trained with its own training samples, and the model parameters are chosen to minimize the validation
201 errors respectively on the leaves of a sub-tree that takes the node as the root. At each leaf node, a local EELM is
202 obtained with its own training and validation samples.

203 Step 3: Compare a global and local EELMs associated to the leaf. For each leaf of the tree, a global and several
204 local EELMs on the path from the root node to the leaf are compared, and the one with the minimum validation error on
205 the leaf is chosen.

206 Given a testing sample, it is assigned into a unique leaf where its prediction model is determined. The testing
207 procedure consists of two steps:

208 Step1: Assign the testing sample to a unique leaf using the splitting rules of the developed tree.

209 Step2: Make a predicted value for the testing sample using the prediction model chosen for the assigned leaf.

210 Local models at deep levels (e.g., at the leaves) may suffer from the overfitting problem. A global model trained on
211 the total dataset is hard to learn well all the patterns, but may be beneficial for capturing some patterns. Firstly, the

212 prediction model for a leaf is chosen based on the comparison of a global model and local models on the path from the
213 root node to the leaf, which can address the global-local duality. For the testing samples in a leaf (an input subspace),
214 training a prediction model in a larger input space and selecting the model parameter to minimize the validation error on
215 the leaf samples may be beneficial. The CART-EELM model aims to seek the best training set for each subspace (i.e.,
216 each leaf) based on the hierarchical regression tree. Secondly, at each non-leaf node, we choose different model
217 parameters for different leaves of a sub-tree that takes the node as the root node, since different leaves may have distinct
218 patterns. Finally, benefitting from high computational efficiency of EELM, the CART-EELM model has acceptable
219 computational cost.

220 3. Data and model implementation

221 3.1. Data

222 The experiments on a real-world air pollution dataset are made to prove the effectiveness of the proposed model.
223 The dataset is from Yancheng city, which is one of the 13 cities under the direct administration of Jiangsu Province,
224 China. Yancheng city spans between northern latitude $32^{\circ}34'-34^{\circ}28'$, eastern longitude $119^{\circ}27'-120^{\circ}54'$. As shown in
225 Fig. 1, the city is located in the middle of the northern plain of Jiangsu, and on the east of the city is the Yellow Sea.
226 Yancheng is fast growing urbanization. Increased energy consumption and the number of vehicles prompt the air
227 pollutant exposure levels especially for the fine particulates. Although clean energy and green technologies are
228 encouraged in Yancheng and neighbouring cities, this city still has experienced severe pollution events.



229

230

Fig. 1. Location of Yancheng Environment Monitoring Central Station

231 To monitor air pollution, networks of monitoring stations in major cities of China have been established. The hourly
 232 average concentrations of six air pollutants ($PM_{2.5}$, O_3 , PM_{10} , SO_2 , NO_2 , CO) are automatically recorded at monitoring
 233 stations. In Yancheng urban district, there are five monitoring stations: Wenfeng Middle School, Yancheng Power Plant,
 234 Administrative Committee of Development Zone, Baolong Square and City Monitoring Centre.

235 Since early 2013, the Environmental Protection Administration began to publish these real-time data to the public.
 236 Now the data from the station of City Monitoring Centre are used in this study, which is located in an area with the high
 237 population density. The air quality data encompasses July 1, 2015 to June 30, 2018. In addition, the meteorological data
 238 of Yancheng urban area from July 1, 2015 to June 30, 2018, which consists of surface wind speed (WS) and direction
 239 (WD), temperature (T), surface air relative humidity (RH) and rainfall (R), are also incorporated into the models. The
 240 data from July 1, 2015 to December 31, 2017 are used for training; the data from January 1, 2018 to June 30, 2018 are
 241 used for testing. The models involved are developed in Python 3.6 using Anaconda 3. The experiments are conducted
 242 on a computer with a Win10 64 bit OS running on Intel Core i5-7440HQ with 8 GB RAM.

243 Some values in the air quality and meteorological data are missing for the studied period. The missing values are
 244 interpolated by using cubic spline interpolation when the missing gap is less than 3 hours and there is no missing value
 245 in 3 hours before and after. Let the observed $PM_{2.5}$ value at time $t+1$, denoted as $PM_{2.5}(t+1)$, be the output of a sample.
 246 The corresponding input candidates include air pollutants at times $t-2$, $t-1$ and t , and the meteorological data at time t .
 247 Moreover, the meteorological data at time $t+1$ is also included as input candidates, since the predicted meteorological
 248 data for the next hour can be used in practical implementation. The -sine and cosine transformations are employed for
 249 the wind direction (Feng et al, 2015). Weekend indicator (1 if yes, 0 otherwise) for time $t+1$, abbreviated WEI($t+1$), is
 250 added due to alterations in traffic patterns and industrial behaviours at weekends. The corresponding input candidates
 251 are given in Table 1. After the samples with any missing value are deleted, we have 19627 training samples and 3898
 252 testing samples.

253 **Table 1:** The input variable candidates corresponding to the output $PM_{2.5}(t+1)$.

Input candidate names	Input candidate vector
XC_0, \dots, XC_{17}	$CO(t-2), CO(t-1), CO(t), NO_2(t-2), NO_2(t-1), NO_2(t),$ $O_3(t-2), O_3(t-1), O_3(t), SO_2(t-2), SO_2(t-1), SO_2(t),$ $PM_{10}(t-2), PM_{10}(t-1), PM_{10}(t), PM_{2.5}(t-2), PM_{2.5}(t-1), PM_{2.5}(t)$
XC_{18}, \dots, XC_{29}	$T(t), T(t+1), \cos(WD(t)), \cos(WD(t+1)), -\sin(WD(t)),$ $-\sin(WD(t+1)), WS(t), WS(t+1), R(t), R(t+1), H(t), H(t+1)$
XC_{30}	WEI ($t+1$)

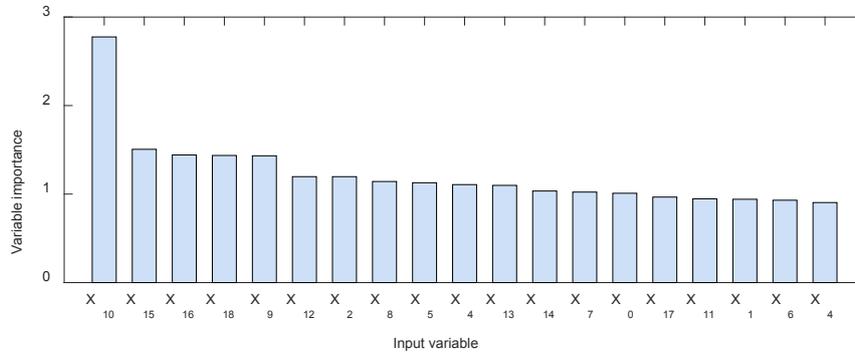
254 3.2. Selection of input variables using RF

255 We use 5-fold cross-validation error of the RF model to evaluate the different combinations of input variable
 256 candidates. The most important 15 variables are invoked in each iteration. In the first iteration, five candidates,
 257 including XC_{11} , XC_{15} , XC_{22} , XC_{28} and XC_{31} , are rejected. In the second iteration, another seven candidates (XC_1 , XC_2 ,

258 $XC_4, XC_{10}, XC_{14}, XC_{21}$ and XC_{27}) are rejected. No candidates can be rejected in the third iteration. The selected input
 259 variables are given in Table 2. Finally, we run the RF model to obtain the importance scores of the selected variables.
 260 As shown in Fig. 2, the variable X_{10} (the lagged 1-hour $PM_{2.5}$ concentration) is far more important than any other
 261 variable. The variable X_{15} (the lagged 1-hour wind speed) is the second important variable, which is followed by the
 262 variable X_{16} (the wind speed of the next hour). Let us remark that both XC_{26} and XC_{27} are rejected due to the small
 263 importance. Similarly in Poggi and Portier (2011), the daily total rainfall was not retained in the prediction model of the
 264 daily mean PM_{10} concentration.

265 **Table 2:** The selected input variables corresponding to the output $PM_{2.5}(t+1)$.

Input variable names	Input vector
X_0, \dots, X_{10}	$CO(t), NO_2(t-1), NO_2(t), O_3(t-2), O_3(t-1), O_3(t),$ $SO_2(t), PM_{10}(t-2), PM_{2.5}(t-2), PM_{2.5}(t-1), PM_{2.5}(t)$
X_{11}, \dots, X_{18}	$T(t), T(t+1), -\sin(WD(t)), -\sin(WD(t+1)),$ $WS(t), WS(t+1), H(t), H(t+1)$



267
268 **Fig. 2.** Importance scores of the selected input variables

269 **3.3. CART-EELM model**

270 **3.3.1. Splitting the dataset by using CART**

271 This study constructs a shallow tree to capture concentration change patterns and ensure that each leaf has
 272 enough samples for its local model. To train a shallow regression tree, the minimum number of samples in a leaf
 273 and the maximum depth of the tree are set to 1000 and 3, respectively. In Fig. 3, the results of splitting the training
 274 set by using CART are given. The serial number for the root node is 0, and all the nodes choose X_{10} (the lagged
 275 1-hour $PM_{2.5}$ concentration) to split its own dataset, since X_{10} contains the most information on next hour
 276 prediction (Lyu, et al, 2017). The persistence model is the simplest one, whose prediction for a given hour is the
 277 observed value of the previous hour.

278 The number of samples in nodes #3 and #4 reaches 10209, accounting for 52.02% of the total training set.
 279 Non-leaf nodes indicate the splitting rules. The value in each node is the mean value of the outputs of samples
 280 fallen into the node. There are significantly different MSE values for different leaves, which sharply increase as
 281 the serial numbers of the leaves increase.

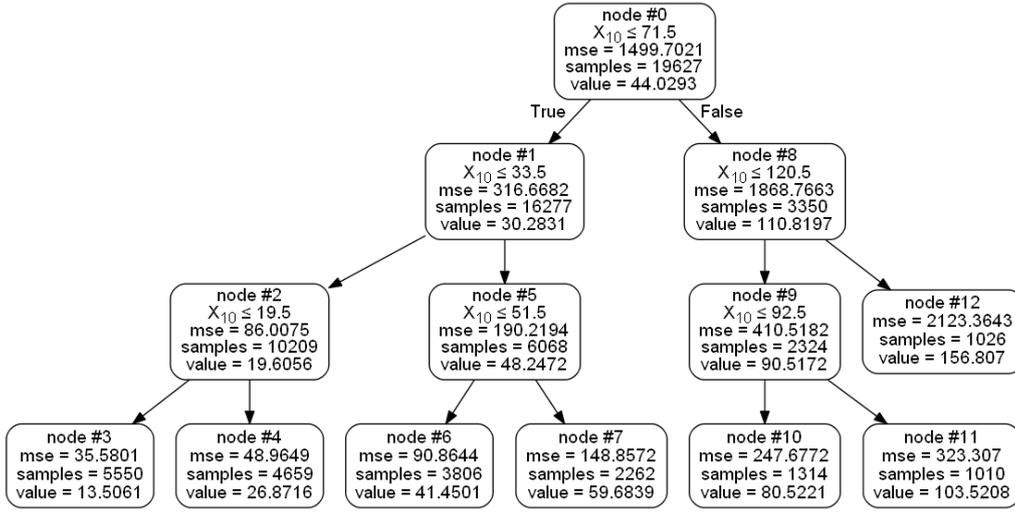


Fig. 3. Training results of the CART

3.3.2. Prediction model based on the CART-EELM

Each node in Fig. 3 has its own dataset. The hidden neuron numbers of EELM models are selected from $\{30, 45, \dots, 300\}$. To seek the optimal numbers, 5-fold cross-validation is implemented. For each non-leaf node of the CART tree, EELMs are trained using the samples belonging to the node, and the hidden neuron numbers are selected to minimize validation errors respectively on the leaves of a sub-tree that takes the node as the root node. Before training EELMs, the input variables are normalized within $[0, 1]$ by

$$\bar{x}_i^d = \frac{x_i^d - \min(x_i^d |_{i=1}^l)}{\max(x_i^d |_{i=1}^l) - \min(x_i^d |_{i=1}^l)}, \quad (4)$$

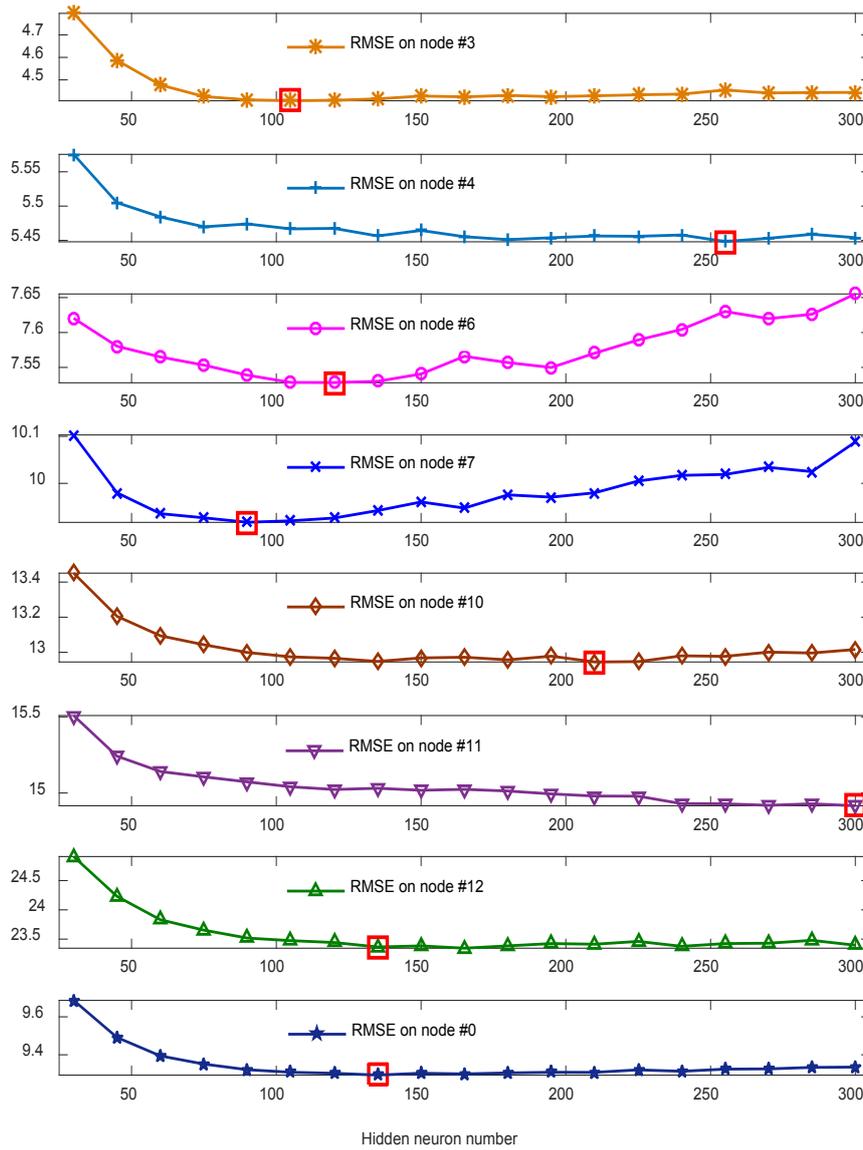
where l denotes the number of samples, p the number of numerical variables, x_i^d the origin value of the d th number variable, and \bar{x}_i^d the normalized value of the d th number variable.

The root node has 7 leaves. In order to show the selection procedure clearly, an example of training a global model at node #0 for node #3 is given. The total training set (the training set that belongs to node #0) is first divided into 5 separate subsets with approximately equal size. A series of 5 models are then trained, each using a different combination of 4 subsets. For each candidate of hidden neuron numbers, the validation error on node #3 is evaluated for each model, utilizing the subset of the data that are not used in training and belongs to node #3. Then the validation error is determined by averaging over all 5 validation subsets. The root mean square error (RMSE) is used as the error criterion, which is calculated by

$$\text{RMSE} = \sqrt{\frac{1}{l} \sum_{i=1}^l (y_i - \hat{y}_i)^2}, \quad (5)$$

where \hat{y}_i is the predicted value.

303 The first subplot in Fig. 4 demonstrates the average validation error on node #3 when using different
 304 candidates of hidden neurons numbers. The same selection procedure is implemented for other 6 leaves at node #0.
 305 The top first 7 subplots in Fig.3 show the validation errors of all global models at node #0. For the global models
 306 trained at node #0 for 7 leaf nodes, the optimal numbers of hidden nodes are 105, 255, 120, 90, 210, 300 and 135,
 307 respectively. It shows that different leaves may require different numbers of hidden neurons. The last subplot of
 308 Fig. 4 gives the validation errors of a global EELM on the total validation set. Its optimal model parameter is 135,
 309 which is resulted from balancing different leaf validation errors to minimize the total validation error.



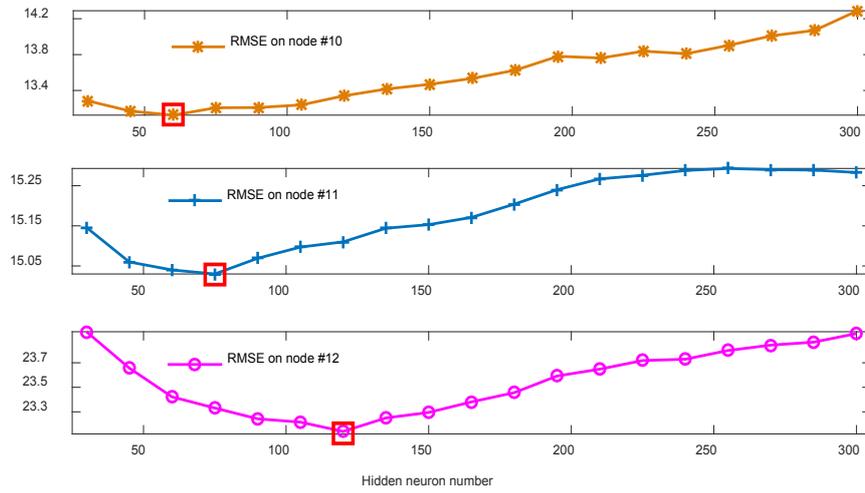
310

311

312

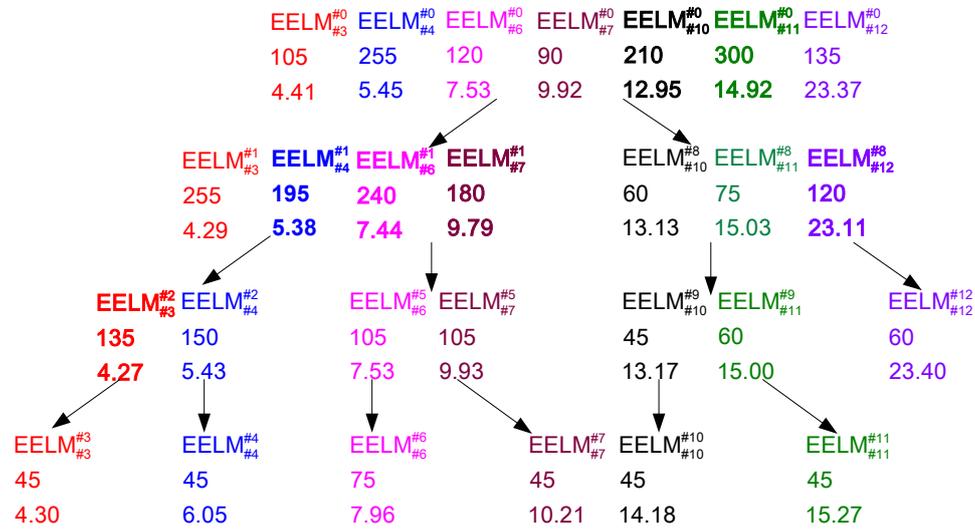
Fig. 4. Validation errors of the global models

313 Additional example of training local models is given at node #8 for nodes #10, #11 and #12. In Fig. 5, it shows
 314 the parameter selection results. For local models trained at node #8 for nodes #10, #11 and #12, the optimal
 315 numbers of hidden nodes are 60, 75 and 135, respectively.



316
317 **Fig. 5.** Validation errors of local models at node #8

318 The models at all the nodes are organized in a hierarchical fashion, as shown in Fig. 6. $EELM_{\#10}^{\#8}$, is shown as
319 an example to clarify the information provided in the node. $EELM_{\#10}^{\#8}$ represents the EELM model trained at node
320 #8 for node #10. The optimal parameter is selected as 60 with the RMSE error of 13.13. For node #10 (a leaf), a
321 global model and three local models, namely $EELM_{\#10}^{\#0}$, $EELM_{\#10}^{\#8}$, $EELM_{\#10}^{\#9}$ and $EELM_{\#10}^{\#10}$, are compared in
322 terms of the RMSE error, and finally $EELM_{\#10}^{\#0}$ is chosen. Consequently the prediction models for other 6 leaves
323 are obtained.



324
325 **Fig. 6.** EELM models at all nodes of the CART tree

326 For the input region of node #3, where X_{10} is not larger than $19.5\mu\text{g}\cdot\text{m}^{-3}$, the prediction model is trained at
327 node #2, where X_{10} is not larger than $33.5\mu\text{g}\cdot\text{m}^{-3}$. Other three leaves in the left branch of the root node choose
328 node #1, where X_{10} is not larger than $71.5\mu\text{g}\cdot\text{m}^{-3}$, to train the prediction models. For nodes #10 and #11, the
329 prediction models are trained at the root node. The prediction model of node #12 is trained at node #8, where X_{10}
330 is larger than $71.5\mu\text{g}/\text{m}^3$. The selection procedure provides the flexibility to address the global-local duality.

331 3.4. Global models

332 In order to evaluate the model effectiveness, the model has been compared with the persistence model and
333 global models, including RF, ν -SVR and EELM. The model is also compared with local models based on season
334 and clustering algorithm. In the RF model, it consists of 100 trees, and the model has a single parameter, i.e., the
335 minimum number of samples in a leaf, which is selected from $\{1,2,\dots,8\}$. The value of ε in SVR chosen as a
336 prior, is hard to determine. To overcome the difficulty of ε determination, the ν -SVR (Schölkopf et al., 2000) is
337 used, whereby ν controls the number of support vectors and training errors. In this study, Gaussian kernel function
338 is applied in the ν -SVR:

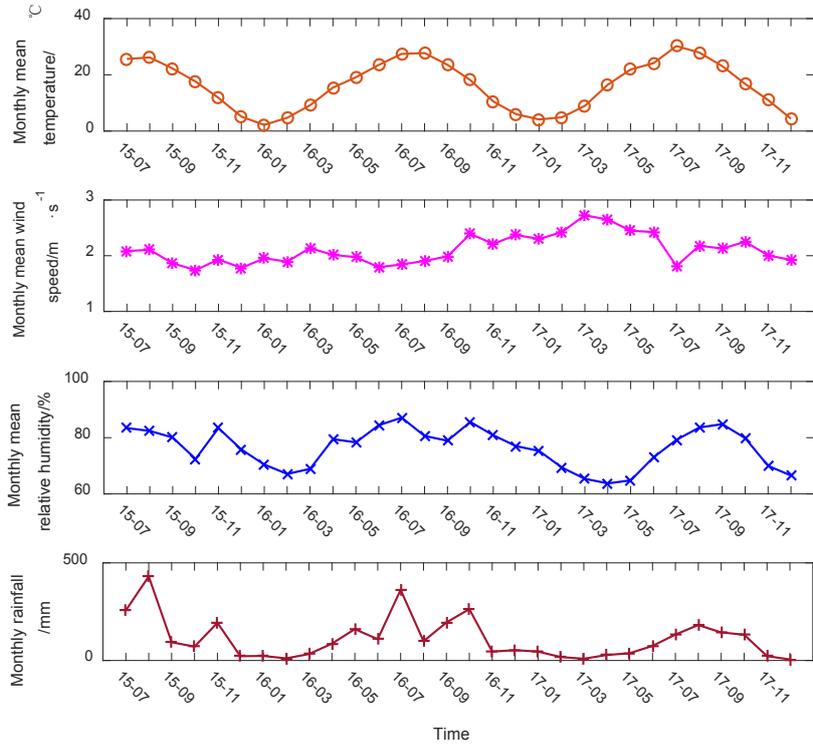
$$339 k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\gamma^2). \quad (6)$$

340 The ν -SVR model owns three parameters: ν , γ and the regularization parameter C , which are selected from
341 $\{0.3,0.4,\dots,0.7\}$, $\{0.1,0.15,\dots,0.45\}$ and $\{0.5,1,\dots,3\}$, respectively. Thus in the ν -SVR, there are 240 combinations
342 of three parameter values. The parameter selection based on 5-fold cross-validation is also applied to several
343 global models. The optimal parameter of the RF turns out to be 1. The optimized parameters of the ν -SVR are
344 selected as $\nu=0.4$, $\gamma=0.45$ and $C=1$. For the EELM, the optimized parameter is 135, which is indicated in the
345 last subplot in Fig. 4.

346 3.5. Local models based on season and k -means

347 The seasonal models split the training set into three subsets based on prevailing weather patterns that may
348 influence the $\text{PM}_{2.5}$ buildup. Fig. 7 presents the monthly mean temperatures, wind speeds and relative humidities as
349 well as the monthly rainfalls over the period between July 1, 2015 and December 31, 2017. The monthly mean
350 concentrations of $\text{PM}_{2.5}$ are indicated in Fig. 8. The $\text{PM}_{2.5}$ monthly mean concentration is negatively correlated to
351 the monthly mean temperature and relative humidity as well as the monthly rainfall. The correlation coefficients
352 are -0.88, -0.58 and -0.59, respectively.

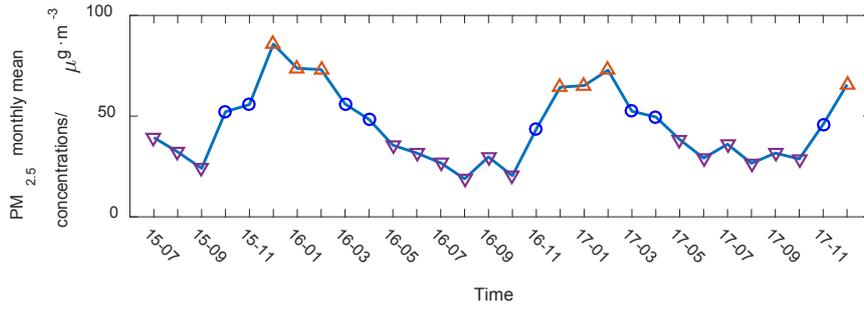
353 A winter season model is trained for the time period between December and February, in which the monthly
354 mean temperatures and the monthly rainfalls were generally low. As shown in Fig. 8, the monthly mean
355 concentrations of $\text{PM}_{2.5}$ exceeded $60 \mu\text{g}\cdot\text{m}^{-3}$ during this period. The data from March, April and November are
356 used for the model corresponding to middle $\text{PM}_{2.5}$ levels. It should be noted that in October 2015, high
357 concentration episodes occurred, and the data from this month are used for middle $\text{PM}_{2.5}$ levels. Finally, the third
358 model is trained for the time period between May and October. During this period, the monthly mean temperatures
359 and the monthly rainfalls were generally high, and the monthly mean concentrations of $\text{PM}_{2.5}$ were lower than 40
360 $\mu\text{g}\cdot\text{m}^{-3}$.



361

362

Fig. 7. Monthly mean concentrations from July, 2015 to December, 2017



363

364

Fig. 8. Monthly mean concentrations of PM_{2.5} from July, 2015 to December, 2017

365

366

367

The k -means clustering model aims to partition the l samples into k sets $\mathcal{S}=\{S_1, S_2, \dots, S_k\}$ in such a way that within each cluster the average dissimilarity of the samples from the cluster mean is minimized. Mathematically, the objective of the k -means clustering is to find

368

$$\arg \min_{\mathcal{S}} \sum_{i=1}^k \sum_{x \in S_i} \|x - \mu_i\|^2, \quad (7)$$

369

370

371

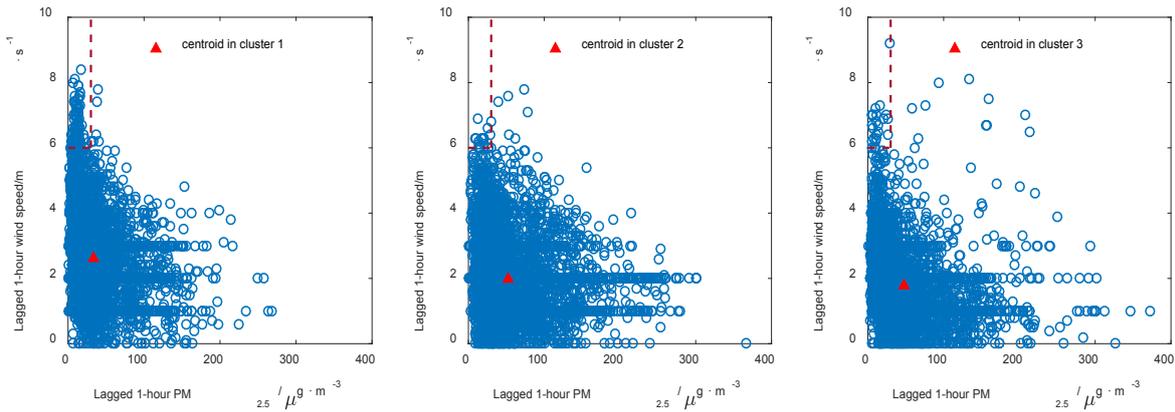
372

373

374

where μ_i is the mean of points in S_i . Unfortunately, such optimization by complete enumeration is feasible only for very small data sets. The strategy based on iterative greedy descent is the most popular method to obtain a good suboptimal partition. In general, there are three steps involved in the strategy. Namely, 1. initialize k cluster centroids; 2. assign each training example to its closest centroid; 3. recalculate each cluster centroid using the examples assigned to it in Step 2, and go back to Step 2 until convergence. The algorithm repeats Step 2 and Step 3 until the centroids do not change between two consecutive iterations.

375 The prediction method based on k -means divides the dataset into several clusters using k -means and builds a
 376 local EELM for individual cluster. The number of clusters is critical to the performance of the local models. The
 377 value of k can be determined by various methods. In this study, 5-fold cross-validation is used to choose the
 378 optimal k ranging from 2 to 7. Firstly, the total training set is partitioned into k clusters. Secondly in each cluster,
 379 we train a local EELM, whose hidden neuron number is selected by 5-fold cross-validation. Then for a certain k ,
 380 the validation error averaged over different clusters is obtained. Finally, the validation errors for different values
 381 of k are compared, and the one with the minimum validation error is chosen. In the study, the optimal number of
 382 clusters turns out to be 3.



383
 384 **Fig. 9.** Scatter plots between the lagged 1-hour PM_{2.5} concentration and the lagged 1-hour wind speed in three clusters

385 To extract the characteristics in three clusters, the scatter plots between the lagged 1-hour PM_{2.5} concentration
 386 and the lagged 1-hour wind speed in three clusters are given in Fig. 9, since these two variables are the two most
 387 important predictors. In Fig. 9, cluster 1 tends to include more samples that have low lagged 1-hour PM_{2.5} (less
 388 than 30 μg·m⁻³) but high lagged 1-hour wind speed (higher than 6 m·s⁻¹). In three clusters, the two most important
 389 elements of the centroids are (33.17, 2.61), (52.47, 1.98) and (47.17, 1.78), respectively. It may be unreasonable
 390 that the latter two are closely spaced.

391 3.6. Prediction results and discussion

392 The training times required by different models are given in Table 3. The RF model has the shortest training
 393 time, while the ν -SVR has the highest computational cost, since the ν -SVR needs to solve a quadratic
 394 programming problem for each combination of parameter values. The training time of the k -means-EELM is more
 395 than four times that of the EELM. Due to high computational efficiency of the EELM, the training times of three
 396 kinds of local models are significantly shorter than that of the ν -SVR.

397 The testing errors from different models are also given in Table 3. The testing set is split by the CART in Fig.
 398 3. The errors in seven input regions, corresponding to the seven leaves, are also included in Table 3. The
 399 persistence model services as a benchmark for other models. Except in the nodes #3 and #4, the prediction

400 accuracy of the RF is inferior to those of the ν -SVR and the EELM. Splitting the training set based on season
 401 increases the performance of the EELM, but the k -means-EELM is slightly inferior to the EELM in terms of
 402 accuracy on the total testing set.

403 **Table 3:** Testing results of different models.

Models	Training time/m	Testing RMSE/ $\mu\text{g}\cdot\text{m}^{-3}$	Testing RMSEs on the leaves/ $\mu\text{g}\cdot\text{m}^{-3}$						
			#3	#4	#6	#7	#10	#11	#12
Persistence	-	10.10	4.99	6.23	8.25	10.44	12.34	15.71	20.01
RF	23.08	9.81	4.61	6.07	8.02	9.58	11.53	14.63	20.87
ν -SVR	1201.12	9.23	4.80	6.11	7.94	9.49	10.78	13.52	18.25
EELM	38.03	9.22	4.82	6.10	7.92	9.57	10.97	13.30	18.10
Seasonal EELM	35.42	9.12	4.86	6.06	7.88	9.45	10.86	13.28	17.69
k -means-EELM	161.23	9.25	4.86	6.19	7.94	9.61	10.95	13.33	18.13
CART-EELM	87.05	8.96	4.63	6.04	7.86	9.35	10.82	13.20	17.02

404 For the results, the CART-EELM shows the most satisfactory accuracy on the total testing set. At leaves #4, #6
 405 and #7, the CART-EELM achieves the best prediction accuracy. At node #10, the prediction performance of
 406 CART-EELM is slightly worse than the ν -SVR. At node #12, the prediction error of the CART-EELM drops by
 407 5.97% and 3.79% compared with the EELM and the seasonal EELM, respectively. To further compare the
 408 CART-EELM with other models, the mean absolute error (MAE), the mean absolute percentage error (MAPE) and
 409 correlation coefficient the on the total testing set are given in Table 4. The MAE and the MAPE are calculated by

$$410 \quad \text{MAE} = \frac{1}{l} \sum_{i=1}^l |y_i - \hat{y}_i|, \quad (8)$$

$$411 \quad \text{MAPE} = \frac{1}{l} \sum_{i=1}^l \left| \frac{y_i - \hat{y}_i}{y_i} \right|. \quad (9)$$

412 The CART-EELM model has the lowest testing MAE and MAPE on the total testing set. It is found that the
 413 correlation coefficient has the lowest sensitivity. Among these statistical indicators, the RMSE is proposed as the
 414 key one for the description of the model skill (Thunis et al., 2012).

415 **Table 4:** Other statistical indicators of different models.

Models	Testing MAE/ $\mu\text{g}\cdot\text{m}^{-3}$	Testing MAPE/%	correlation coefficient
Persistence	6.43	16.85	0.97
RF	6.24	16.76	0.97
ν -SVR	6.04	16.78	0.98
EELM	5.99	16.46	0.98
Seasonal EELM	5.99	16.51	0.98
k -means-EELM	6.04	16.71	0.98
CART-EELM	5.86	16.01	0.98

416 According to the newly updated Ambient Air Quality Standard (GB3095-2012) and combined with the actual
 417 air pollution condition of three Chinese cities, the ambient air quality levels and the corresponding concentration
 418 ranges of $\text{PM}_{2.5}$ were developed (Xu et al., 2017). Table 5 presents the levels and the corresponding ranges. The
 419 correct estimations, the underestimates and the overestimates are counted for the EELM, the seasonal EELM and
 420 the CART-EELM. Compared with an overestimation, an underestimation may bring more harm to the public.
 421 Statistical results of three models are indicated in Table 6. The first row of the CART-EELM, is taken as an
 422 example to clarify the information provided in Table 6. There are 437 samples with the actual levels being I.

423 Among these, 279 samples are correctly estimated and the rest 158 samples are overestimated as level II. From
 424 Table 6, it is found that the CART-EELM has the most correctly estimated samples and the least underestimated
 425 samples.

426 **Table 5:** The ambient air quality levels and the corresponding concentration ranges

Levels				
Good (I)	Regular (II)	Bad (III)	Very bad (IV)	Extreme bad (V)
(0,15]	(15,35]	(35,75]	(75,120]	(120,+∞)

427 **Table 6:** Statistical results of the estimated levels from the EELM, the seasonal EELM and the CART-EELM

Actual levels	EELM					Seasonal EELM					CART- EELM				
	I	II	III	IV	V	I	II	III	IV	V	I	II	III	IV	V
I	280	157	0	0	0	280	157	0	0	0	279	158	0	0	0
II	77	1214	130	1	0	85	1212	124	1	0	70	1223	129	0	0
III	2	104	1068	62	0	1	104	1066	65	0	1	103	1071	61	0
IV	0	0	79	414	27	0	0	78	417	25	0	0	69	423	28
V	0	0	0	38	245	0	0	1	39	243	0	0	0	35	248

428 In this study, the CART-EELM model has better performances compared to the other models discussed. The
 429 global model trained on the total dataset is hard to fit well all the patterns, and the local models trained at the
 430 leaves have the overfitting risk. The current model presented aims to deal with the global-local duality for each
 431 leaf. The prediction model for a leaf is selected among a global and several local EELMs on the path from the root
 432 node to the leaf, in terms of the validation error on the leaf. The selection procedure provides the CART-EELM the
 433 flexibility to deal with the global-local duality.

434 The evaluation results also show that splitting the dataset based on season increases the prediction accuracy of
 435 the EELM model. Compared to the global model based on the EELM, local models based on k -means cannot
 436 improve the performance. In the case, the CART-EELM model shows better effectiveness of the $PM_{2.5}$
 437 concentration prediction.

438 4. Conclusions

439 The multiple change patterns of $PM_{2.5}$ concentrations increase the difficulty of hourly $PM_{2.5}$ prediction. As
 440 local models show great potential to improve the prediction accuracy, local prediction models based on the
 441 CART-EELM is proposed, which uses the CART to split the training set into subsets in the fashion of a
 442 hierarchical tree and EELMs are trained at each node with its own training samples. The prediction model for each
 443 leaf is then selected among a global model and local models on the path from the root node to the leaf in terms of
 444 the validation error on the leaf.

445 For the experimental results of evaluation, it can be concluded: (1) the method can address the global-local
 446 duality for the prediction model on each leaf; (2) the CART-EELM method has better performances than the global
 447 models, including RF, ν -SVR and EELM; (3) the CART-EELM method also show outperforms compared to the
 448 seasonal EELM and the k -means-EELM.

449 Acknowledgements

450 This research was supported in part by the research fund of Key Laboratory for Advanced Technology in
451 Environmental Protection of Jiangsu Province (Grant No. AE201121).

452 References

- 453 Bai, Y., Li, Y., Wang, X.X., Xie, J.J., Li, C., 2016. [Air pollutants concentrations forecasting using back propagation](#)
454 [neural network based on wavelet decomposition with meteorological conditions](#). *Atmos. Pollut. Res.* 7, 557–566.
- 455 Bettenburg, N., Nagappan, M., Hassan, A.E., 2012. [Think locally, act globally: Improving defect and effort prediction](#)
456 [models](#). *IEEE International Working Conference on Mining Software Repositories*, 60–69.
- 457 Breiman, L., Friedman, J.H., Olshen, R., Stone, C.J., 1984. [Classification and regression trees](#). *Biometrics* 40, 358.
- 458 Breiman, L., 2001. [Random forests](#). *Mach. Learn.* 45, 5–32.
- 459 Brezigar-Masten, A., Masten, I., 2012. [CART-based selection of bankruptcy predictors for the logit model](#). *Expert Syst.*
460 [Appl.](#) 39 (11), 10153–10159.
- 461 Budalakoti, S., Srivastava, A.N., Otey, M.E., 2009. [Anomaly detection and diagnosis algorithms for discrete symbol](#)
462 [sequences with applications to airline safety](#). *IEEE Trans. Syst. Man Cybern. C Appl. Rev.* 39 (1), 101–113.
- 463 Cobourn, W.G., (2010). [An enhanced PM_{2.5} air quality forecast model based on nonlinear regression and](#)
464 [back-trajectory concentrations](#). *Atmos. Environ.* 44 (25), 3015–3023.
- 465 Díaz-Robles, L.A., Ortega, J.C., Fu, J.S., Reed, G.D., Chow, J.C., Watson, J.G., et al., 2008. [A hybrid ARIMA and](#)
466 [artificial neural networks model to forecast particulate matter in urban areas: the case of Temuco, Chile](#). *Atmos.*
467 [Environ. 42, 8331–8340.](#)
- 468 Dong, M., Yang, D., Kuang, Y., He D., Erdal S., Kenski D., 2009. [PM_{2.5} concentration prediction using hidden](#)
469 [semi-Markov model-based times series data mining](#). *Expert Syst. Appl.* 36 (5), 9046–9055.
- 470 Doraiswamy, P., Hogrefe, C., Hao, W., Civerolo, K., Ku, J.Y., Sistla, G., 2010. [A retrospective comparison of](#)
471 [model-based forecasted PM_{2.5} concentrations with measurements](#). *J. Air & Waste Manage. Assoc.* 60, 1293–1308.
- 472 Feng, Y., Zhang, W., Sun, D., Zhang, L., 2011. [Ozone concentration forecast method based on genetic algorithm](#)
473 [optimized back propagation neural networks and support vector machine data classification](#). *Atmos. Environ.* 45,
474 1979–1985.
- 475 Feng, X., Li, Q., Zhu, Y.J., Hou, J.X., Jin, L.Y., Wang, J.J., 2015. [Artificial neural networks forecasting of PM_{2.5}](#)
476 [pollution using air mass trajectory based geographic model and wavelet transformation](#). *Atmos. Environ.* 107,
477 118–128.
- 478 Garcia, J.M., Teodoro, F., Cerdeira, R., Coelho, L.M.R., Kumar, P., Carvalho, M.G., 2016. [Developing a methodology](#)
479 [to predict PM10 concentrations in urban areas using generalized linear models](#), *Environ. Technol.*, 37(18),
480 2316–2325.
- 481 He, J.J., Gong, S.L., Yu, Y., Yu, L.J., Wu, L., Mao, H.J., et al., 2017. [Air pollution characteristics and their relation to](#)
482 [meteorological conditions during 2014–2015 in major Chinese cities](#). *Environ. Pollut.* 223, 484–496.
- 483 Huang, G.B., Zhu, Q.Y., Siew, C.K., 2006. [Extreme learning machine: theory and applications](#). *Neurocomputing* 70
484 (1-3), 489–501.
- 485 Kassomenos, P.A., Vardoulakis, S., Chaloulakou, A., Paschalidou, A.K., Grivas, G., Borge, R., et al., 2014. [Study of](#)
486 [PM₁₀ and PM_{2.5} levels in three European cities: Analysis of intra and inter urban variations](#). *Atmos. Environ.* 87,
487 153–163.

488 Khoshgoftaar, T.M., Hulse, J.V., Napolitano, A., 2011. Comparing Boosting and Bagging techniques with noisy and
489 imbalanced data, *IEEE Trans. Syst. Man Cybern. A Syst. Humans* 41 (3), 552–568.

490 Liu, Z.R., Hu, B., Wang L.L., Wu, F.K., Gao, W.K., Wang, Y.S., 2015. Seasonal and diurnal variation in particulate
491 matter (PM₁₀ and PM_{2.5}) at an urban site of Beijing: analyses from a 9-year study. *Environ. Sci. Pollut. Res.* 22 (1),
492 627–642.

493 Lu, W.Z., Wang, W.J., 2005. Potential assessment of the “support vector machine” method in forecasting ambient air
494 pollutant trends. *Chemosphere*, 59, 693–701.

495 Lyu, B.L., Zhang, Y.Z., Hu, Y.T., 2017. Improving PM_{2.5} air quality model forecasts in China using a bias-correction
496 framework. *Atmosphere* 8, 147.

497 McKendry, I.G., 2002. Evaluation of artificial neural networks for fine particulate pollution. *J. Air & Waste Manage.*
498 *Assoc.* 52 (9), 1096–1101.

499 Ni, X.Y., Huang, H., Du, W.P., 2017. Relevance analysis and short-term prediction of PM_{2.5} concentrations in Beijing
500 based on multi-source data. *Atmos. Environ.* 150, 146–161.

501 Niu, M.F., Wang, Y.F., Sun, S.L., Li, Y.W., 2016. A novel hybrid decomposition-and-ensemble model based on
502 CEEMD and GWO for short-term PM_{2.5} concentration forecasting. *Atmos. Environ.* 134, 168–180.

503 Osowski, S., Garanty, K., 2007. Forecasting of the daily meteorological pollution using wavelets and support vector
504 machine. *Eng. Appl. Artif. Intel.* 20, 745–755.

505 Perez, P., 2012. Combined model for PM₁₀ forecasting in a large city. *Atmos. Environ.* 60, 271–276.

506 Perez, P., Gramsch, E., 2016. Forecasting hourly PM_{2.5} in Santiago de Chile with emphasis on night episodes. *Atmos.*
507 *Environ.* 124, 22–27.

508 Poggi, J.M., Portier, B., 2011. PM₁₀ forecasting using clusterwise regression. *Atmos. Environ.* 45, 7005–7014.

509 Qin, S.S., Liu, F., Wang, J.Z., Sun, B.B., 2014. Analysis and forecasting of the particulate matter (PM) concentration
510 levels over four major cities of China using hybrid models. *Atmos. Environ.* 98, 665–675.

511 Requia, W.J., Adams M.D., Koutrakis, P., 2017. Association of PM_{2.5} with diabetes, asthma, and high blood pressure
512 incidence in Canada: A spatiotemporal analysis of the impacts of the energy generation and fuel sales. *Sci. Total*
513 *Environ.* 584-595, 1077–1083.

514 Russo, A., Raischel, F., Lind, P.G., 2013. Air quality prediction using optimal neural networks with stochastic variables.
515 *Atmos. Environ.* 79, 822–830.

516 Schölkopf, B., Smola, A.J., Williamson, R.C., Bartlett, P.L., 2000. New support vector algorithms. *Neural Comput.* 12
517 (5), 1207–1245.

518 Sun, W., Zhang, H., Palazoglu, A., Singh, A., Zhang, W.D., Liu, S.W., 2013. Prediction of 24-hour-average PM_{2.5}
519 concentrations using a hidden Markov model with different emission distributions in Northern California. *Sci. Total*
520 *Environ.* 443, 93–103.

521 Thomaidis, N.S., Bakeas, E.B., Siskos, P.A., 2003. Characterization of lead, cadmium, arsenic and nickel in PM_{2.5}
522 particles in the Athens atmosphere, Greece. *Chemosphere* 52 (6), 959–966.

523 Thunis, P., Pederzoli, A., Pernigotti, D., 2012. Performance criteria to evaluate air quality modeling applications. *Atmos.*
524 *Environ.* 59, 476–482.

525 Vlachogianni, A., Kassomenos, P., Karppinen, A., Karakitsios, S., Kukkonen, J., 2011. Evaluation of a multiple
526 regression model for the forecasting of the concentrations of NO_x and PM₁₀ in Athens and Helsinki. *Sci. Total*
527 *Environ.* 409 (8), 1559–1571.

528 Voukantsis, D., Karatzas, K., Kukkonen, J., Räsänen, T., Karppinen, A., Kolehmainen, M., 2011. Intercomparison of air

529 quality data using principal component analysis, and forecasting of PM_{10} and $PM_{2.5}$ concentrations using artificial
530 neural networks, in Thessaloniki and Helsinki. *Sci. Total Environ.* 409, 1266–1276.

531 Wang, D.Y., Wei, S., Luo H.Y., Yue, C.Q., Grunder, O., 2017. A novel hybrid model for air quality index forecasting
532 based on two-phase decomposition technique and modified extreme learning machine. *Sci. Total Environ.* 580,
533 719–733.

534 Xu, Y.Z., Yang, W.D., Wang, J.Z., 2017. Air quality early-warning system for cities in China. *Atmos. Environ.* 148,
535 239–257.

536 Xue, X.W., Yao, M., Wu, Z.H., Yang, J.H., 2014. Genetic ensemble of extreme learning machine. *Neurocomputing* 129,
537 175–184.

538 Yu, R.Y., Yang, Y., Yang, L.Y., Han, G.J., Move, O.A., 2016. RAQ-A random forest approach for predicting air
539 quality in urban sensing systems. *Sensors* 16 (1), 86.

540 Zhou, Z.H., Wu, J.S., Tang, W., 2002. Ensembling neural networks: Many could be better than all. *Artif. Intell.* 137,
541 239–263.

542 Zhou, Q.P., Jiang, H.Y., Wang, J.Z., Zhou, J.L., 2014. A hybrid model for $PM_{2.5}$ forecasting based on ensemble
543 empirical mode decomposition and a general regression neural network. *Sci. Total Environ.* 496, 264–274.

544 Zhu, S.L., Lian, X.Y., Liu, H.X., Hu, J.M., Wang, Y.Y., Che, J.X., 2017. Daily air quality index forecasting with hybrid
545 models: A case in China. *Environ. Pollut.* 231, 1232–1244.