A novel PM2.5 concentrations probability density prediction model combines the least absolute shrinkage and selection operator with quantile regression

Shaomei Yang1 · Haoyue Wu1

Received: 20 February 2022 / Accepted: 2 June 2022 / Published online: 11 June 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract
PM2.5 has a significant negative impact on human health and atmospheric quality, and accurate prediction of its concentration is necessary. When using common point prediction models for PM2.5 concentration prediction, the influence of various uncertainties on PM2.5 concentrations makes the prediction results suffer from poor accuracy. To address this issue, this paper proposes the quantile regression neural network (QRNN) model based on the least absolute shrinkage and selection operator (LASSO), combined with kernel density estimation (KDE) for probabilistic density prediction of PM2.5 concentrations. The model uses LASSO regression to select the influential factors, and then the quartiles of daily PM2.5 concentrations calculated by the QRNN model are imported into the KDE model to obtain the probability density predictions of PM2.5 concentrations. In the paper, empirical analyses are carried out with the cities of Beijing and Jinan in China as well as six other datasets, and the prediction performance of the model is assessed by using evaluation criteria in both point prediction and interval prediction. The simulation reveals that the predictive performance of the LASSO-QRNN-KDE model is well, and the model is not only effective in filtering high-dimensional data, but also has a higher accuracy compared to common research models. In addition, the model is able to describe the uncertainty of PM2.5 concentration fluctuations and carry more information on the variation of PM2.5 concentrations, which can provide a novel and excellent PM2.5 concentration prediction tool for relevant policy makers.

Keywords PM2.5 concentration prediction · Probability density prediction · The least absolute shrinkage and selection operator · Quantile regression · Air pollution · Hybrid models

Introduction
PM2.5 refers to the particulate matter in the ambient air with an aerodynamic equivalent diameter less than or equal to 2.5 microns, which has a small particle size and can be used as a carrier of toxic substances. These fine particles can be inhaled into the human body through the respiratory system and adhere to the upper and lower respiratory tracts and lung lobes, causing damage to the nervous system, cardiovascular and cerebrovascular systems, and respiratory systems. People who have been exposed to fine particulate pollution for a long time are more likely to face cardiovascular and respiratory diseases (Geng et al. 2015; Lee et al. 2017; Mukhopadhyay and Sahu 2018). At the same time, PM2.5 pollution has a negative impact on the normal development of agriculture, industry, and cultural tourism. It will also affect traffic order and cause major economic losses. By predicting the concentration of PM2.5 and accurately forecasting it, it is possible to issue early warnings for heavily polluted air in order to take targeted protective measures. Therefore, in order to maintain the order of production activities, reduce public health risks, and provide meaningful reference for relevant policy makers, the prediction of PM2.5 concentration is of great significance.

Currently, prediction models for air pollutant concentrations such as PM2.5 can be divided into three main categories: deterministic models, classical statistical models, and machine learning models (Jiang and Qiao
Deterministic methods mainly include weather research forecasting model (WRF) (Cao et al. 2018), community multiscale air quality model (CMAQ) (Yang et al. 2019), and coupled WRF-CMAQ meteorological-chemical model (Syvakov et al. 2016). The above methods use relevant meteorological data and pollutant source data to simulate the process of pollutants from emission to accumulation to dispersion and transfer, so as to predict and analyze the concentration of pollutants. According to Lightstone et al. (2017), such methods are more complex and do not have a significant advantage in prediction accuracy over other models. At the level of statistical models, autoregressive integrated moving average model (ARIMA) and multiple linear regression model (MLR) are more used among scholars in the problem of PM2.5 concentration prediction (Zhang et al. 2018). However, the concentration of PM2.5 is affected by a variety of variable factors and has strong nonlinearity and complexity. Statistical models are not satisfactory in processing nonlinear time series data. Machine learning has the advantage of covering knowledge of probability theory, statistics, approximate theory, and knowledge of complex algorithms, and can simulate the way of human learning. Machine learning has superior predictive performance and is widely used in air pollutant concentration prediction. At the same time, the hybrid model is also widely used in the prediction of the concentration of air pollutants such as PM2.5. Huang and Kuo (2018) combined convolutional neural network (CNN) and long short-term memory (LSTM) and applied it to the PM2.5 prediction system. This method effectively improved the prediction accuracy of the LSTM model. Huang et al. (2021) used empirical mode decomposition (EMD) to decompose the PM2.5 concentration series. They input the multiple stationary sub-sequences and meteorological features obtained after decomposition into the constructed gated recurrent unit neural network (GRU) in order to train and predict them. The results showed that the accuracy of the EMD-GRU model is much higher than that of the single GRU model. Li et al. (2017) introduced the co-integration theory in order to solve the problem of the lack of evaluation of the possible correlation between different variables in the current research. The author used the flower pollination algorithm (FPA) to optimize the parameters of the support vector machine (SVM) model, and made high-precision predictions of the PM2.5 and PM10 concentrations in two important cities in Yunnan Province, China.

PM2.5 concentrations are influenced by a variety of factors, related not only to pollutant emissions, but also to meteorological factors. Common pollutant factors include PM10, NO₂, CO, and O₃. (Masiol et al. 2013; Xiao et al. 2018), while meteorological factors include wind speed, precipitation, temperature, and humidity (Liao et al. 2017; Wang et al. 2019; Xu et al. 2019). In the big data environment, related data sources are extensive and the characteristics are different, making the prediction of PM2.5 concentration more complicated. Reviewing the existing literature related to the prediction of PM2.5 and other air pollutant concentrations, it is found that most of the prediction methods obtain deterministic point prediction results, which can hardly reflect the uncertainty of their fluctuations. To solve the above problems, a probability density prediction method for PM2.5 concentration can be used to overcome the shortcomings of the traditional point prediction and interval prediction methods. In this way, the future PM2.5 concentration values and the concentration fluctuation intervals under the corresponding confidence intervals can be obtained, and the probability of occurrence of each point within the interval can also be visualized. With the probability density function of the predicted PM2.5 concentration, more detailed and effective information can be obtained than the traditional prediction methods. It is extremely helpful for the accurate prediction of PM2.5 concentration.

However, due to the complexity and variability of factors affecting PM2.5 concentration, the probability density function that accurately portrays the uncertainty of PM2.5 concentration is difficult to obtain. Therefore, in order to predict PM2.5 concentration more accurately, valuable information needs to be extracted from the wide range of influencing factors first, so as to filter out the influencing factors that have significant explanatory significance on PM2.5 concentration. The traditional classical method of variable selection is mainly represented by the subset selection method. This method compared 2ᵖ − 1 non-empty subsets of all p dimensional variables with a certain criterion to find the optimal subset. Here, the most representative selection criteria are mainly the Akaike information criterion (AIC) based on information theory proposed by Hirotsugu Akaike in 1974 (Hoerl and Kennard 2000), the BIC criterion based on Bayes’ method proposed by Schwarz (1978), and the cross-validation criterion. But for the information age where high-dimensional data are proliferating in various fields, the complexity and high cost of computation make the traditional variable selection methods no longer applicable to these high-dimensional data structures. Therefore, statisticians have started to focus on variable selection methods for high-dimensional data. In the field of variable selection for high-dimensional data, Tibshirani’s 1996 proposal of the LASSO algorithm (least absolute shrinkage and selection operator) under the ordinary linear model was of epoch-making significance (Tibshirani 1996). It was inspired by two algorithms, ridge Regression (Hoerl and Kennard 2000) and nonnegative garrote (Breiman 1995).

LASSO was essentially a coefficient compression method that compressed the coefficients of non-significantly influential variables to zero, achieving both variable selection and
parameter estimation at the same time. This method can be seen as an improvement of ridge regression, but it is milder in compressing the coefficients of significant variables with significant effects, thus ensuring the accuracy of parameter estimation. In a previous study, Koen et al. (2020) used a LASSO regression model to select variables closely related to four common chronic diseases in the Netherlands to predict disease prevalence. The results showed that the use of the LASSO algorithm significantly improved the predictive performance of the model. Jasleen and Mamta (2021) used a correlation-based adaptive LASSO algorithm to identify potentially important influences while studying pollutant concentrations and meteorological factors in Delhi and its surrounding cities. Evaluation of the model revealed better performance of this approach in extracting subset of features. In an empirical study, Li et al. (2017) found that real surplus management (REM) was strongly selected as a key crisis predictor by the LASSO algorithm, developing an updated distress forecasting model with surplus management considerations for the Chinese market. In addition, the method has been applied to the variable selection process of quantile regression models.

Quantile regression was first proposed by Koenker and Bassett (1978). Due to its unique explanatory power, it has become one of the key tools for data analysis and is widely used in various data analysis fields (Koenker 2015). Quantile regression can give a more comprehensive picture of the conditional distribution of the response variables. Instead of analyzing only the conditional expectation (mean) of the explanatory variable, the method can also analyze how the explanatory variables affect the median, quantile, etc. of the response variable. Kang et al. (2021) used quantile regression to study the impact of macroprudential policies on the level of bank financing of Chinese firms, and pointed out that the intensity of the impact of macroprudential policies was different at different quantile points. Salari et al. (2021) combined the idea of quantile regression in the study of the impact of multiple factors on the ecological footprint of emerging countries, and discussed the significant impact of different quantile factors. Sun et al. (2021) used quantile autoregressive models to analyze return and volatility series on multiple time scales using the crude oil market as an example. It was found that the autoregressive coefficients vary with quantile. The above study further demonstrates the advantages of quantile regression. Namely, the method can explain the effect of each input variable on the response variable at different quartiles and obtain a more detailed relationship between the input and response variables.

The approach of combining the LASSO algorithm with quantile regression has been applied to the power, financial, and medical sectors. He et al. (2019) predicted electricity consumption based on LASSO-quantile regression for Guangdong Province, China, and California, USA. Simulation results showed that the method provided better performance for electricity consumption forecasting. Cao et al. (2021) studied the optimal portfolio problem based on the LASSO-quantile regression method and showed that the results obtained by this method were close to the given value at risk under certain canonical conditions, demonstrating the robustness and effectiveness of this method. Taha et al. (2018) reduced the data dimensionality by LASSO and used quantile regression method to study and analyze the prostate cancer dataset. The results validated the superiority of this combined method.

There are no studies that have applied LASSO-quantile regression to air pollutant concentration prediction, to the best of our knowledge. In this paper, the LASSO-QRNN-KDE model is proposed to combine LASSO quantile regression with kernel density estimation. Firstly, this paper will apply the LASSO variable selection method to extract the factors that have significant effects on PM2.5 concentrations from the high-dimensional external influences data to obtain a lower dimensional data set. Subsequently, combine with the quantile regression method, a quantile regression model will be established to obtain the conditional quantiles of PM2.5 concentrations at different quantiles. Finally, the kernel density estimation method was employed to obtain the probability density prediction curves of PM2.5 concentrations on any consecutive day in the future month.

The main innovations and contributions of this paper can be divided into the following two aspects:

1. The LASSO-QRNN-KDE model proposed in the paper reduces the uncertainty of the prediction by downscaling the influencing factors of PM2.5 concentration and selecting the factors that have a significant influence on it. Moreover, it is able to obtain different probability distributions of PM2.5 concentrations at different quantile points and further obtain probability density curves. Compared with the common point prediction, the probability density prediction expands the prediction range and provides a more detailed and informative explanation for the prediction of PM2.5 concentrations.

2. This paper demonstrates the superiority of the method through experiments on two real datasets, Beijing and Jinan, China, and six other datasets. Three different scenarios are investigated and compared in this paper: without considering external factors, considering external factors without variable selection, and using the LASSO algorithm to filter external factors. The results show that the proposed LASSO-QR-KDE model has superior performance and performs well in point prediction as well as interval prediction. Moreover, when comparing five models commonly used to predict PM2.5 concentrations, namely ARIMA, BP, RBF,
CNN-LSTM, and EMD-GRU, the LASSO-QRNN-KDE model was found to achieve higher accuracy.

The rest of the paper is organized as follows: the second part describes the theoretical basis and methodology involved in the model proposed in the paper. The third part presents the constructed neural network model and its evaluation criteria. The fourth part gives an empirical analysis and discussion of the two cases and shows how the data are handled. Finally, the conclusions of the paper are given in the fifth part.

## Methodology

### LASSO method

The LASSO method adds a penalty term to optimize the traditional objective function to produce a sparse solution. The method avoids the problem of overfitting caused by excessive explanatory variables (He et al. 2021). It proceed by extracting the key variables from the high-dimensional variables. The method is effectively solving the problem of multicollinearity among variables and improving the explanatory accuracy of the model, which is highly evaluated and widely used. In addition, compared with ridge regression, the LASSO method can directly compress the regression coefficients of redundant predictor variables to zero, obtaining a more streamlined set of predictor variables. Over-compression of important regression coefficients can also be avoided.

Given a matrix of independent variables \(X = (x_1, x_2, \ldots, x_n)\), where \(X_j = (x_{j1}, x_{j2}, \ldots, x_{jm_j})^T\), \(j = 1, 2, \ldots, n\), the dependent variable \(Y = (y_1, y_2, \ldots, y_m)\). It is assumed that the data have been standardized. If \(\sum_{j=1}^{n} x_{ij} = 0\) and \(\sum_{i=1}^{m} x_{ij}^2 = 1\), then the linear model between \(Y\) and \(X\) is established as Eq. (1).

\[
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon, \tag{1}
\]

where \(\beta_0\) is the constant term, \(\beta_i\) is the coefficient of each variable, and \(\epsilon\) is the random disturbance term.

Suppose \(\beta = (\beta_1, \beta_2, \ldots, \beta_n)^T\), then the LASSO estimation of \(\beta\) is as in Eq. (2).

\[
\hat{\beta}_{\text{LASSO}} = \arg\min_{\beta} \sum_{i=1}^{m} (y_i - \beta^T x_i)^2, \quad \text{s.t. } \sum_{j=1}^{n} |\beta_j| \leq \lambda, \tag{2}
\]

where the penalty parameter is \(\lambda\) and \(\lambda \geq 0\). If \(\hat{\beta}^0\) is the least squares estimation of \(\beta\), then \(\lambda_0 = \sum_{j=1}^{n} |\hat{\beta}^0_j|\). When \(\lambda < \lambda_0\), the absolute value of the LASSO estimation of the regression coefficient \(\hat{\beta}_j\) in the model is smaller than the absolute value of \(\hat{\beta}^0\). As \(\lambda\) decreases gradually, the LASSO estimation of certain \(\hat{\beta}_j\) decreases or even goes to zero. The variable can be eliminated if its estimation is zero, which indicates that it has little association with the dependent variable \(Y\), and thus, the variable selection function can be achieved. The commonly used estimation and testing methods regarding the penalty parameter \(\lambda\) are the cross-validation criterion, the generalized validation method, and the AIC (Hoerl and Kennard 2000).

For the solution of the LASSO problem, the least angle regression (LARS) significantly enhances the computational efficiency of LASSO. Suppose \(A = \{j_1, j_2, \ldots, j_k\} \subseteq \{1, 2, \ldots, n\}\) is the indicator set, and define the matrix of \(m \times k\) as \(X_A = (\lambda_{j_1} x_{j_1}, \lambda_{j_2} x_{j_2}, \ldots, \lambda_{j_k} x_{j_k})\), where the value of \(\lambda\) can be determined by Eq. (3).

\[
\lambda_{jl} = \begin{cases} 1, & x_{jl} \text{ is positively correlated with } Y \\ 0, & x_{jl} \text{ is not correlated with } Y \\ -1, & x_{jl} \text{ is negatively correlated with } Y \\ \end{cases} \tag{3}
\]

Subsequently, suppose that \(G_A = X_A^T X_A\) and \(A_A = (I_A G_A^{-1} I_A)^{-1/2} u_A = X_A w_A\) is the angular component vector in \(X_A w_A = A_A G_A^{-1} A_A\), where \(I_A\) is a K-dimensional vector, and all elements are equal to 1. It can be shown that the angular component vector \(u_A\) is a unit vector and \(X_A u_A = A_A I_A\).

Establish the LARS estimation as \(\tilde{\mu} = X \hat{\beta}\). Assuming that the current estimate is \(\hat{\mu}_A\), \(\tilde{\hat{\mu}} = X' (Y - u_A)\) indicates that the respective variable is currently correlated with \(Y\). From the previous assumptions, the variables in \(A\) are maximally correlated with \(Y\), and the relationship can be expressed by Eq. (4).

\[
\begin{cases} \tilde{c}_j = \tilde{\hat{c}}, j \in A \\ c_j < \tilde{\hat{c}}, j \notin A \end{cases} \tag{4}
\]

where \(\tilde{\hat{c}} = \max_j \{c_j\}\).

(1) Calculate the value of \(\hat{c}\).  
(2) Calculate \(X_A A_A u_A\) by taking \(\lambda_j = \text{sign} \{c_j\} \{j \in A\}\).  
(3) Calculate \(\alpha = X' u_A\), with \(X\) being the full matrix of independent variables.  
(4) Calculate \(\hat{\mu}_{\alpha} = u_A + \hat{\gamma} u_A\), where \(\hat{\gamma} = \min_{j \notin A} \{\frac{\tilde{c}_j - c_j}{\lambda_{\alpha - u_A}} + \frac{c_j + c_j}{\lambda_{\alpha + u_A}}\}\) \(\min^+\) means that only the smallest positive value in the set is selected for calculation.  
(5) Set \(\hat{\mu}_A = u_{\alpha^+}\), and then returns to step (1) to continue the loop until all the independent variables are used for the operation.
Quantile regression

Quantile regression is a linear model that describes the relationship between the explanatory variable \( X \) and the conditional quantile of the response variable \( Y \). The formula is given in Eq. (5):

\[
Q(\tau | X) = a_0(\tau) + a_1(\tau)X_1 + a_2(\tau)X_2 + \cdots + a_n(\tau)X_n = X^T \alpha(\tau),
\]

(5)

where \( Q(\tau | X) \) is the estimate of the response variable \( Y \) corresponding to the explanatory variable \( X = [X_1, X_2, \cdots, X_n] \) under the \( \tau \) th quantile condition. \( \alpha(\tau) = [\alpha_0(\tau), \alpha_1(\tau), \cdots, \alpha_n(\tau)]^T \) is the model parameter associated with the quantile \( \tau \), which is estimated and solved by Eq. (6).

\[
\min_{\tau} \sum_{i=1}^{N} \rho_\tau(Y_i - X_i \alpha(\tau)) = \min \left[ \sum_{i=1}^{N} [Y_i - X_i \alpha(\tau)]^2 + \sum_{i=1}^{N} \rho_\tau(Y_i - X_i \alpha(\tau)) \right].
\]

(6)

where \( N \) is the number of samples, and \( \rho_\tau(\bullet) \) is the loss function of the model, which is expressed as in Eq. (7)

\[
\rho_\tau(u) = \begin{cases} \tau u, & u \geq 0 \\ (1 - \tau)u, & u < 0 \end{cases}
\]

(7)

After the obtained estimate of \( \alpha(\tau) \), the estimate of the response variable at the \( \tau \) th quantile can be solved according to Eq. (5).

Kernel density estimation

The solution of the density function of a random variable can be solved by parametric estimation methods and non-parametric estimation methods. However, the parametric estimation method requires prior estimation of the density function form of the random variable, which often leads to a large discrepancy between the actual density function and the estimated density function.

The kernel density estimation (KDE) method can overcome the drawbacks of the parameter estimation methods. KDE is used in probability theory to estimate unknown density functions and was proposed by Rosenblatt (1956) and Parzen (1962). The method enables to employ a set of random variables from the same unknown distribution function to estimate its density function. KDE has no requirement to know the prior distribution of the data in question or to make any assumptions about the data in question. The method requires only input variables, kernel density estimation function, and optimal bandwidth to perform the predictions.

\[ f(x) \] is the density function of the random variable \( x \), \( f(x) = F'(x) \). According to the principle of KDE, a simple estimate of \( f(x) \) can be obtained as in Eq. (8).

\[
f(x) = \frac{F(x + k) - F(x - k)}{2k},
\]

(8)

where \( k \) is a non-negative series of constants, and \( F(x) \) is the sample distribution function of the random variable \( x \). \( x_1, x_2, \cdots, x_n \) are \( n \) sample points of an independent identically distributed one-dimensional continuous overall sample. Its probability density function is \( f(x) \), and the kernel density estimate is as in Eq. (9).

\[
\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),
\]

(9)

where \( K(\bullet) \) is the kernel function, which is essentially the weight function. \( h \) is called the bandwidth, and its choice determines the smoothness and accuracy of the estimated density function.

LASSO-QRNN-KDE model

Quantile regression model based on LASSO

The precision of PM2.5 concentration prediction is influenced by various external factors, and the impact of different factors on the prediction results varies in degrees. According to the description of quantile regression and the LASSO method in the previous section, the quantile regression model is constructed as Eq. (10).

\[
Y_t = r(\tau) + \sum_{c=1}^{p} \alpha_c(\tau)Y_{t-c} + \sum_{d=1}^{m} \left( \sum_{j=1}^{q} \beta_{jd}(\tau)X_{d-j} \right) + \epsilon_t,
\]

(10)

where \( Y_t \) is the prediction series and \( X_i \) is the sequence of influencing factors. \( m \) indicates the number of influencing factors and \( c \) indicates the maximum number of lags of the prediction series, \( c = 1, 2, \cdots, p. \) \( \tau \) is the quantile, \( \tau \in (0, 1) \). The regression coefficient vector for the prediction target of the \( \tau \) th quantile is \( \alpha_{\tau} \), and the regression coefficient vector for the \( d \)th quantile is \( \beta_{\tau,d} \), \( d = 1, 2, \cdots, m. \) \( q \) indicates the maximum number of lags of the influence factor. \( r(\tau) \) is the intercept term and \( \epsilon_t \) is the error term. The predicted value of PM2.5 concentration \( y_t \) at the \( \tau \) th quantile can be further obtained, which is calculated as Eq. (11)

\[
Q_{\tau}(\tau | Y_t, X_t) = r(\tau) + \sum_{c=1}^{p} \alpha_c(\tau)Y_{t-c} + \sum_{d=1}^{m} \left( \sum_{j=1}^{q} \beta_{jd}(\tau)X_{d-j} \right)
\]

(11)

According to Eqs. (10) and (11), it can be observed that there is heterogeneity in \( r(\tau), \alpha_{\tau} \), and \( \beta_{\tau,d} \), which indicates the discrepancy in the importance of the studied...
samples. This reflects the impact of the selected influence factors on the PM2.5 concentration prediction results under different quantile points. The parameters $\alpha_i(\tau)$ and $\beta_j^d(\tau)$ in Eq. (11) can be solved by Eq. (12).

$$\min_{\alpha_i} \sum_{t=1}^{n} \rho_{\tau}\left( Y_t - r(\tau) - \sum_{c=1}^{p} \alpha_c(\tau)Y_{t-c} - \sum_{d=1}^{m} \left( \sum_{j=1}^{q} \beta_j^d(\tau)X_{t-d}^d \right) \right).$$

(12)

where $\rho_{\tau}(u)$ is the check function with the following Eq. (13).

$$\rho_{\tau}(u) = (\tau - I(u < 0))u$$

(13)

$I(u < 0)$ is called the indicator function and $u < 0$ is the conditional relation. When $u < 0$ is true, $I(u < 0) = 1$, and vice versa $I(u < 0) = 1$. To improve the speed of operation, the interior point algorithm is applied to solve the equation (Portnoy and Koenker 1997).

Since the influence of multiple factors on PM2.5 concentration is considered, this paper will apply the LASSO method to filter a large number of influence factors. The penalty parameter $L1$ is added to the original objective function, and Eq. (14) is obtained as follows.

$$\min_{\alpha} \sum_{t=1}^{n} \rho_{\tau}\left( Y_t - r(\tau) - \sum_{c=1}^{p} \alpha_c(\tau)Y_{t-c} - \sum_{d=1}^{m} \left( \sum_{j=1}^{q} \beta_j^d(\tau)X_{t-d}^d \right) \right) + \lambda \left( \|\alpha(\tau)\|_1 + \|\beta(\tau)\|_1 \right),$$

(14)

where $\lambda$ is the penalty parameter and its penalty enhances as the value taken increases. $\|\alpha(\tau)\|_1$ and $\|\beta(\tau)\|_1$ are the 1-norm. $\|\alpha(\tau)\|_1$ is the loss function that measures the effectiveness of the quantile regression model to fit the data. $\|\beta(\tau)\|_1$ is the penalty function, which can be used to achieve variable selection by constraining the insignificant coefficients to zero. To facilitate the solution of the LASSO-QRNN model, Eq. (14) can be transformed into Eq. (15).

$$\min_{\alpha, \beta} \sum_{t=1}^{n} \rho_{\tau}\left( Y_t - r(\tau) - \sum_{c=1}^{p} \alpha_c(\tau)Y_{t-c} - \sum_{d=1}^{m} \left( \sum_{j=1}^{q} \beta_j^d(\tau)X_{t-d}^d \right) \right),$$

s.t. $\|\alpha(\tau)\|_1 + \|\beta(\tau)\|_1 < \delta,$

(15)

where $\delta$ is the constraint parameter corresponding to $\lambda$. The degree of constraint on the coefficients increases as the value of $\delta$ is taken to decrease. The LASSO-QRNN model is used to demonstrate the distribution of PM2.5 concentrations at different quartiles, and to filter out the influencing factors that are significant to its interpretation. In this way, the effect of reducing the data dimensionality and optimizing the prediction model can be achieved.

### Probability density prediction of PM2.5 concentration

From the above method, the conditional quantile of PM2.5 concentration at different quantile points can be obtained, and each quantile is the prediction result of PM2.5 concentration at different quantile points. The mean square error performance of the Epanechnikov kernel function is optimal, and its efficiency loss is small. In this paper, Epanechnikov kernel function is used to predict the PM2.5 concentration at each quantile as its input value, and the prediction result of the probability density of PM2.5 concentration is then obtained. The expression of the Epanechnikov kernel function is as follows.

$$k(x) = \frac{3}{4} (M - x^2)M(|x| \leq 1),$$

(16)

where $M(\bullet)$ is an indicator function. When $|x| \leq 1$ is true $M(\bullet)$ takes the value of 1, and vice versa $M(\bullet)$ takes the value of 0.

### LASSO-QRNN-KDE neural network model

Neural networks employ nonlinear kernel functions that can provide a better way to deal with complex nonlinear objects. In the paper, a quantile regression neural network is used to predict the quantile corresponding to random variables at different quantile levels based on the single hidden layer neural network model proposed by Taylor (Madsen and Nielsen 1993). The hyperbolic tangent Sigmoid function is used as the hidden layer basis function of the neural network in solving the problem utilizing quantile regression neural network. First, a stable network structure is established between the input and response variables to achieve the prediction of different quartiles of the predicted object, and then the probability density can be estimated by using the kernel density estimation of nonparametric estimation. The expression of the quantile regression neural network model used in the paper is as follows.

$$Q_{\gamma}(\theta|X) = f[x, u(\theta), v(\theta)] = \sum_{j=1}^{r} \left\{ \frac{2v_j(\theta)}{1 + e^{-2 \sum_{i=1}^{r} u_{ij}(\theta)X_i}} - v_j(\theta) \right\},$$

(17)

where $\theta$ is the quantile, $u(\theta) = \{u_{ij}(\theta)\}$ is the weight matrix to be estimated between the input layer and the implied layer, and $V(\theta) = \{v_j(\theta)\}$ is the connection weight vector between the implied layer and the output layer. To achieve the parameter estimation of Eq. (17), which can be achieved by solving the optimization objective function Eq. (18).
To prevent the neural network model from falling into the process-fitting state, the corresponding penalty term is added to the objective function. The new objective function is as follows

$$E_{\theta} = \frac{1}{N} \sum_{i=1}^{N} \rho_{\theta} \{ Y_i - f[X, U(\theta), V(\theta)] \}$$  \hspace{1cm} (18)

where $\lambda_1$ and $\lambda_2$ are the penalty parameters of the model. By determining the optimal penalty parameters, the model can be effectively prevented from over-fitting to the empirical data, which can reduce the prediction error and improve the prediction accuracy. The optimal estimate of $U(\theta)$ and $V(\theta)$ can be obtained by optimizing Eq. (19) after getting the value of $\min_{U(\theta), V(\theta)} E_{\theta}(U(\theta), V(\theta))$. Bringing the estimated values into Eq. (17), the conditional quantile estimation function of the response variable can be obtained. The structural flow chart of the model is shown in Fig. 1.

**Evaluation criteria**

For testing the prediction performance of the method in the paper, a comprehensive and reasonable analysis and evaluation of the prediction results is required. In this paper, mean absolute percentage error, while RMSE is more sensitive to outlier data and can highlight error values with large effects, and the expressions are as follows.

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{A_i - F_i}{A_i} \right|,$$  \hspace{1cm} (20)

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (A_i - F_i)^2}{n}},$$  \hspace{1cm} (21)

where $n$ indicates the number of days of PM2.5 concentration being used for prediction, $A_i$ is the actual value, and $F_i$ is the predicted value. The closer the values of MAPE and RMSE to zero indicate the better model performance.

PICP is able to define the probability of the actual value falling in the prediction interval, and its expression is as follows.

$$\text{PICP} = \frac{1}{n} \sum_{i=1}^{n} c_i,$$  \hspace{1cm} (22)

where $c_i$ reflects the ability of the actual value to fall into the prediction interval at the confidence level.

$$c_i = \begin{cases} 0, & A_i \in [L_i, U_i] \\ 1, & A_i \not\in [L_i, U_i] \end{cases},$$  \hspace{1cm} (23)
where \( L_i \) and \( U_i \) are the lower and upper bounds of the \( i \) th prediction interval. In general, the value of PICP should be greater than or equal to the predetermined confidence level, and a larger PICP value indicates a higher confidence level of the model.

PINAW can reflect the sharpness of the probability density prediction, and a high sharpness probability density prediction is able to carry the higher amount of information. The expression for PINAW is as follows.

\[
PINAW = \frac{1}{nR} \sum_{i=1}^{n} (U_i - L_i),
\]

where \( R \) is the range of the actual value of the predicted series. The smaller the value of PINAW, the narrower the prediction interval, which means that the prediction interval can carry more information and has higher accuracy.

PINAD reflects the deviation between the actual value and the forecast interval, and its expression is as follows.

\[
PINDA = \frac{1}{nR} \sum_{i=1}^{n} D_i,
\]

where \( D_i \) is the deviation from the prediction interval for the \( i \) th prediction point, and is defined by the following equation.

\[
D_i = \begin{cases} 
L_i - A_i, & A_i < L_i \\
0, & L_i < A_i < U_i \\
A_i - U_i, & A_i > U_i 
\end{cases}
\]

PINAD calculates the cumulative deviation of the points where the actual values fall outside the prediction interval from the bounds of the prediction interval. When the confidence level of the prediction interval is satisfied, the smaller the PINAD, the better the performance of the interval prediction model.

**Empirical analysis**

**Case 1: Beijing, China**

Beijing is located in the north of China and has 12 national automatic air quality monitoring sites. In the twenty-first century, Beijing has developed rapidly and made great achievements in industrial construction as well as in the construction of transport networks. However, along with the economic boom, the air quality in Beijing has also been affected, and it is far from the national standard and the expectation of the people, especially the foggy weather makes the whole society pay extra attention to the fine particulate matter PM2.5 pollution which is more difficult to prevent and control. In recent years, Beijing has vigorously promoted clean energy and strengthened environmental supervision, enabling pollutant concentrations to drop significantly. However, the first quarter still has a high PM2.5 concentration, as shown in Fig. 2. In order to continuously promote the environmental construction, the prediction of PM2.5 is essential.

In this case, the data were selected from January 1, 2018, to January 31, 2019, for Beijing, China. The data were obtained from the National Meteorological Data Center of China (http://data.cma.cn/), and all data were normalized in advance. From the experience of previous studies, the...
concentration of PM2.5 is influenced not only by air quality but also by meteorological factors. In this paper, five air quality indicators and five meteorological factor indicators were selected as influencing factors to investigate the implications on PM2.5 concentration prediction.

According to the previous Eq. (11), the sequence of PM2.5 concentrations to be predicted is \((y_1, y_2, \cdots, y_n)\), where \(n\) is the number of samples to be predicted. The sequence of influencing factors is \(X_i\) and \(X_i = (X_i^1, X_i^2, \cdots, X_i^m)\), \(m\) is taken as 10 in this paper. \(X_i^d\) is the \(d\)th influencing factor, and \(X_i^d = (x_{i,1}^d, x_{i,2}^d, \cdots, x_{i,4}^d)\).

In this paper, \(X_i^1\) indicates daily PM10 concentration (\(\mu g/m^3\)), \(X_i^2\) indicates daily SO2 concentration (\(\mu g/m^3\)), \(X_i^3\) indicates daily CO concentration (mg/m3), \(X_i^4\) indicates daily NO2 concentration (\(\mu g/m^3\)), and \(X_i^5\) indicates daily O3 concentration (\(\mu g/m^3\)); \(X_i^6\) indicates daily average temperature (0.1°C), \(X_i^7\) indicates the daily relative humidity (%), \(X_i^8\) indicates the daily average wind speed (0.1 m/s), \(X_i^9\) indicates daily CO concentration (mg/m3), and \(X_i^{10}\) indicates daily average temperature (0.1°C), \(X_i^1\) indicates daily average wind speed (0.1 m/s), and \(X_i^2\) indicates daily average temperature (0.1°C).

Table 1: The statement of influencing factors of PM2.5 concentration in Beijing

| Name (codename)                  | Lags                  |
|----------------------------------|-----------------------|
| Daily PM2.5 concentration (\(Y_i\)) | \(y_{i,1}\) * \(y_{i,2}\) \(y_{i,3}\) \(y_{i,4}\) |
| Daily PM10 concentration (\(X_i^1\)) | \(x_{i,1}^1\) \(x_{i,2}^1\) \(x_{i,3}^1\) \(x_{i,4}^1\) |
| Daily SO2 concentration (\(X_i^2\)) | \(x_{i,1}^2\) * \(x_{i,2}^2\) \(x_{i,3}^2\) \(x_{i,4}^2\) |
| Daily CO concentration (\(X_i^3\)) | \(x_{i,1}^3\) * \(x_{i,2}^3\) \(x_{i,3}^3\) * \(x_{i,4}^3\) |
| Daily NO2 concentration (\(X_i^4\)) | \(x_{i,1}^4\) \(x_{i,2}^4\) \(x_{i,3}^4\) \(x_{i,4}^4\) |
| Daily O3 concentration (\(X_i^5\)) | \(x_{i,1}^5\) \(x_{i,2}^5\) \(x_{i,3}^5\) \(x_{i,4}^5\) |
| Daily average temperature (\(X_i^6\)) | \(x_{i,1}^6\) * \(x_{i,2}^6\) * \(x_{i,3}^6\) * \(x_{i,4}^6\) |
| Daily relative humidity (\(X_i^7\)) | \(x_{i,1}^7\) \(x_{i,2}^7\) \(x_{i,3}^7\) \(x_{i,4}^7\) |
| Daily average wind speed (\(X_i^8\)) | \(x_{i,1}^8\) \(x_{i,2}^8\) \(x_{i,3}^8\) \(x_{i,4}^8\) |
| The wind direction of daily maximum wind speed (\(X_i^9\)) | \(x_{i,1}^9\) * \(x_{i,2}^9\) \(x_{i,3}^9\) \(x_{i,4}^9\) |

Table 2: Structure of neural network in three situations (Beijing)

| Situation (codename) | Input layer–hidden layer–output layer |
|----------------------|--------------------------------------|
| Without considering external factors (A) | 7–3–1 |
| Considering external factors without variable selection (B) | 53–3–1 |
| LASSO-QRNN-KDE (C) | 10–3–1 |

The LASSO method was applied to the training data for variable filtering. From Appendix 2 (Table 11), it can be seen that Lasso proceeds to the 38th step with the smallest cp value, and then the optimal number of steps for Lasso regression is 38. Eight factors with non-zero coefficients were retained, as shown in the part marked with an asterisk in Table 1. The selected factors are those that have significant explanations for PM2.5 concentrations in Beijing. The above output shows that PM2.5 concentration in Beijing is influenced by its historical concentration, SO2 concentration, CO concentration, average temperature, wind speed, and wind direction.

Three situations are considered in the paper to make the empirical analysis more objective and to verify the validity of the LASSO-QRNN-KDE method for probability density prediction. They are probability density prediction without considering external factors, probability density prediction with considering external factors but without variable selection, and probability density prediction with applying LASSO for variable selection. After training, the neural network structures for the three cases are shown in Table 2. The number of times the neural network was trained was set to 1000, and the penalty parameters \(\lambda_1\) and \(\lambda_2\) in Eq. (19) were set to 0.5. The quartiles range from 0.05 to 0.95 with an interval of 0.05. With the trained neural network model, the continuous quartiles for each day in January 2019 can be obtained. The obtained PM2.5 concentration quantile is then used as an input variable and imported into the Epanechnikov kernel function to obtain the probability density curve of PM2.5 concentration for each day in January 2019.

Figures 3, 4, and 5 show the probability density curves of PM2.5 concentrations in Beijing for 8 days in January under three scenarios in turn. The horizontal coordinate value corresponding to the red dashed line in the figure indicates the actual PM2.5 concentration in Beijing on that day, while the solid blue line is the probability density curve and the highest point is the mode of the predicted value. The comparison of the three plots shows that the highest probability point (highest point) of the probability density curve often deviates from the actual value when no external factors are considered. On days...
1, 17, and 25 in Fig. 3, the actual values appear at the tail of the probability density curve, and the predicted values deviate significantly from the actual values. With external factors taken into account but not selected, the prediction situation improved. According to Fig. 4, the prediction value that deviates significantly from the actual value is significantly improved, and only on day 13 the actual value is at the tail of the probability density curve. In the case of using LASSO to select the variables, according to Fig. 5, the actual values are basically near the highest probability point of the probability density curve, and there is no case that the actual values are at the tail of the probability density curve. It can be seen that the LASSO-QRNN-KDE model has a good performance in predicting
the probability density of PM2.5 concentration, and the predicted values have a higher probability of being close to the actual values.

Appendix 3 (Table 13) gives the predicted values and relative errors of PM2.5 concentrations at the highest point of the probability density curve (mode) for January 2019 in Beijing under three situations. Combining with Fig. 6, it can be seen that the maximum relative errors in the three situations are 15.64%, 13.29%, and 11.42%, and the minimum relative errors are 0.88%, 0.99%, and 0.08%. The performance of the LASSO-QRNN-KDE model is significantly better than the other two situations.

According to the results in Table 3, in terms of point predictions, the best prediction results are obtained using the LASSO-QRNN-KDE model, where the MAPE is 2.40% and the RMSE is 1.63. In addition, the probability density

![Probability density curve of Beijing with LASSO-QRNN-KDE model](image)

**Fig. 5** Probability density curve of Beijing with LASSO-QRNN-KDE model

**Table 3** Prediction errors for three situations in Beijing

| Evaluation indicators | Without considering external factors (a) | Considering external factors without variable selection (b) | LASSO-QRNN-KDE (C) |
|-----------------------|-----------------------------------------|----------------------------------------------------------|-------------------|
| Point prediction      | MAPE 6.04%                               | 3.41%                                                    | 2.40%             |
|                       | RMSE 4.11                                | 2.74                                                     | 1.63              |
| Interval prediction   | PICP 94.81%                              | 96.08%                                                   | 98.34%            |
|                       | PINAW 29.76%                             | 33.52%                                                   | 22.07%            |
|                       | PINAD 3.37%                              | 3.34%                                                    | 1.85%             |
prediction considering external factors but without variable selection also performs better than the probability density prediction without considering external factors. In terms of interval prediction, it is evident that the prediction with external factors has a qualified PICP, and the PICP increases by 2.26% after selecting the influencing factors by LASSO, which indicates that the proposed model has more reliable prediction intervals. In addition, the proposed model has the smallest PINAW and PINAD, with a decrease of 11.45% and 1.49% respectively compared to the model without the influence filter, which means that the model has a higher sharpness and the prediction interval has a smaller deviation from the actual value.

In the prediction of PM2.5 concentrations with quantile regression neural networks, the prediction results using the LASSO-QRNN-KDE model outperformed the prediction results without considering external factors and without variable selection. This indicates that the probability density prediction of PM2.5 concentration should adequately consider the influence of external environment, and the optimal prediction model can be obtained by filtering multiple external factors.

**Case 2: Jinan, China**

Jinan is located in East China and has built 8 national automatic air environment quality monitoring points. Jinan used to be a heavy industrial production base, and heavy industrial enterprises have made outstanding contributions to the economic development of Jinan. However, the development of heavy industries has brought pollution in various aspects such as air and water. In addition, the high topography around Jinan makes its air less mobile and dust tends to accumulate over the city, which has led to Jinan being the city with the poor air quality in China. According to Fig. 7, it can be seen that its annual PM2.5 concentration has a U-shaped distribution. Accurate prediction of PM2.5 concentrations is essential to promote the local environment and to reduce public health risks.

In this case, data from January 1, 2020, to January 31, 2021, were selected from the National Meteorological Data Center of China (http://data.cma.cn/) for Jinan, China, and all data were pre-normalized. Consistent with case 1, in this case m is taken as 10, c and q as 4, and the confidence level is 95%. The statistical descriptions and correlation analysis of the data sets are shown in Appendix 1 (Tables 8 and 10). The data from January 1, 2020, to December 31, 2020, in Jinan were treated as training data, and the data from January 1, 2021, to January 31, 2021, were treated as test data using the rolling prediction method. From Appendix 2 (Table 12), it can be seen that the $c_p$ value at step 32 is the smallest, and the optimal number of steps for Lasso regression is obtained as 32. Ten factors were selected, as shown in the variables marked with an asterisk in Table 4. The results show that the PM2.5 concentration in Jinan is influenced by its historical concentration, PM10 concentration, CO concentration, $NO_2$ concentration, $O_3$ concentration, as well as wind speed and wind direction.

The same as case 1, case 2 also considered three cases. After training, the neural network structures for the three cases are shown in Table 5.

Figures 8, 9, and 10 show the probability density curves of PM2.5 concentrations in Jinan for eight days in January under three scenarios in turn. The highest probability point
Table 4 The statement of influencing factors of PM2.5 concentration in Jinan

| Name (codename)                      | Lags                      |
|--------------------------------------|---------------------------|
| Daily PM2.5 concentration ($Y_t$)    | $Y_{t-1}$, $Y_{t-2}$, $Y_{t-3}$, $Y_{t-4}$ |
| Daily PM10 concentration ($X_t^1$)  | $X_{t-1}^1$, $X_{t-2}^1$, $X_{t-3}^1$, $X_{t-4}^1$ |
| Daily SO2 concentration ($X_t^2$)   | $X_{t-1}^2$, $X_{t-2}^2$, $X_{t-3}^2$, $X_{t-4}^2$ |
| Daily CO concentration ($X_t^3$)    | $X_{t-1}^3$, $X_{t-2}^3$, $X_{t-3}^3$, $X_{t-4}^3$ |
| Daily NO2 concentration ($X_t^4$)   | $X_{t-1}^4$, $X_{t-2}^4$, $X_{t-3}^4$, $X_{t-4}^4$ |
| Daily O3 concentration ($X_t^5$)    | $X_{t-1}^5$, $X_{t-2}^5$, $X_{t-3}^5$, $X_{t-4}^5$ |
| Daily average temperature ($X_t^6$) | $X_{t-1}^6$, $X_{t-2}^6$, $X_{t-3}^6$, $X_{t-4}^6$ |
| Daily average air pressure ($X_t^7$) | $X_{t-1}^7$, $X_{t-2}^7$, $X_{t-3}^7$, $X_{t-4}^7$ |
| Daily average wind speed ($X_t^8$)  | $X_{t-1}^8$, $X_{t-2}^8$, $X_{t-3}^8$, $X_{t-4}^8$ |
| The wind direction of daily maximum wind speed ($X_t^{10}$) | $X_{t-1}^{10}$, $X_{t-2}^{10}$, $X_{t-3}^{10}$, $X_{t-4}^{10}$ |

of the probability density curve deviates from the actual value more often when the prediction is made without considering external factors, and the true value often appears at the end of the probability density curve. When external factors are taken into account, the prediction results are much better, especially when the LASSO method is used to select the variables and then make the prediction, the actual value basically appears at the highest point of the probability density curve. In this case, the LASSO-QRNN-KDE model also has a good performance in predicting the probability density of PM2.5 concentration.

Appendix 3 (Table 14) presents the predicted values of PM2.5 concentrations at the highest point of the probability density curve (mode) and the relative errors for Jinan in January 2021 under three situations. Figure 11 shows that the maximum relative errors in the three cases are 14.77%, 11.46%, and 6.52%, and the minimum relative errors are 1.02%, 1.53%, and 0.05%. The overall performance using the LASSO-QRNN-KDE model is significantly improved over the other two situations.

The experimental results in Table 6 demonstrate that the LASSO-QRNN-KDE model outperforms the other three cases, from both the point prediction perspective and the interval prediction perspective. When applying LASSO to variable screening, point prediction results were optimal, where MAPE is 2.36% and RMSE is 2.07. In addition, the PICP is greater than 95% for all three situations, with the model presented in the paper having a high PICP of 99.03%, which means that the LASSO-QRNN-KDE model performs well for interval coverage. Similarly, after screening for impact factors with LASSO, PINAW, and PINAD are 17.14% and 1.18% respectively, which are the best performers in the three situations. Similar to case 1, probability density prediction of
PM2.5 concentration using LASSO-QRNN-KDE model can get more reliable results.

**Comparative experiments**

In order to validate the accuracy and robustness of the LASSO-QRNN-KDE model, several comparison models and datasets from other regions are introduced below. Experiments are conducted to compare the performance of 5 other commonly used forecasting models with the model proposed in this paper, and to verify the capability of the proposed model to generalize using seven additional datasets.

**Comparisons with other models**

To demonstrate the superiority of the LASSO-QRNN-KDE method for PM2.5 concentration prediction, a variety of traditional prediction models and novel research models are...
introduced in this section for comparison with the proposed method. In this section, the prediction results are compared using the same data set. Five comparison models are presented below.

1. ARIMA (autoregressive integrated moving average model), which achieves time series forecasting by fitting the observed time series into a parametric model (Ruby-Figueroa et al. 2016);
2. BP (back propagation neural network), which continuously adjusts the weights and thresholds of the network through training, and thus learns the mapping relationships in the input and output data (Zhou and Qu 2017);
3. RBF (Radial Basis Function), similar to BP, but with good ability to approximate arbitrary nonlinear functions and to express the intrinsic unresolved regularities of the system (He et al. 2013);
4. CNN-LSTM (convolutional neural networks-long short-term memory), which uses CNN for feature extraction, and then inputs the extracted feature values into the LSTM structure (Huang and Kuo 2018).
5. EMD-GRU (empirical mode decomposition-gate recurrent unit), which uses EMD to decompose PM2.5 concentration sequences, and then the decomposed multiple smooth subsequences and meteorological features are sequentially input into the GRU neural network for training and prediction (Huang et al. 2021).

The prediction results of each model for PM2.5 concentrations in Beijing are shown in Fig. 12 and for Jinan City in Fig. 13. The prediction errors of the two cases are compared in Fig. 14. In addition, the daily forecast data for each model are shown in Appendix 4 along with the daily forecast relative errors.

Based on the experimental results of each model, the following observations are derived:

1. Compared with the traditional prediction model ARIMA and the classical machine learning methods BP and RBF, the evaluation indexes of the other three models are better, which reflects that the networks designed specifically for dealing with time series problems outperform the typical shallow machine learning methods in terms of time series prediction. Furthermore, for the task of PM2.5 concentration prediction, the prediction performance of the combined model is better than that of the single model.
2. In both cases, the RMSE and MAPE of the LASSO-QRNN-KDE model are taken to be the minimum values. Taking the data of the Beijing case as an example, the RMSE and MAPE of the proposed model decreased by 2.14 and 2.78%, respectively, compared to the better-performing EMD-GRU model.

Validation of the model with other data sets

To validate the generalization capability of the LASSO-QRNN-KDE model, six additional datasets are added to this section, and the statistical descriptions of these datasets are presented in Appendix 1 (Table 8). The prediction results
for each dataset are evaluated with the evaluation criteria proposed in the previous section, and the results are presented in Table 7.

(1) The prediction results from the six additional datasets show that the LASSO-QRNN-KDE model has strong robustness. The point predictions of the model
performed well after selecting the influencing factors for each region by LASSO method. Except for the XA dataset, the MAPE for all regions are below 3%. The RMSE of each dataset can basically be maintained around 2, indicating that the model proposed in the paper has a high accuracy in point prediction. In addition, the best point prediction performance and the worst point prediction performance were obtained for the AL dataset and the XA dataset. The MAPE and RMSE of the XA dataset increased by 2.24% and 1.82 compared to the AL dataset, and it can be reasonably inferred that the XA dataset is more volatile, resulting in poorer prediction performance.

(2) Based on the performance of interval prediction, the PICP of all six data sets is greater than the confidence level, which indicates that the model proposed in the paper has a high accuracy in point prediction. In addition, the best point prediction performance and the worst point prediction performance were obtained for the AL dataset and the XA dataset. The MAPE and RMSE of the XA dataset increased by 2.24% and 1.82 compared to the AL dataset, and it can be reasonably inferred that the XA dataset is more volatile, resulting in poorer prediction performance.

Table 7 Prediction results for each dataset

| Datasets | Cities       | MAPE(%) | RMSE  | PICP  | PINAW | PINAD  |
|----------|--------------|---------|-------|-------|-------|--------|
| SIZ      | Shijiazhuang | 2.33    | 2.12  | 95.00 | 20.33 | 2.11   |
| ZZ       | Zhengzhou    | 2.91    | 2.28  | 98.31 | 21.16 | 1.29   |
| XA       | Xian         | 4.02    | 2.84  | 96.48 | 25.04 | 3.04   |
| AL       | Alashan League | 1.78   | 1.02  | 96.18 | 22.94 | 2.62   |
| SZS      | Shizuishan   | 3.2%    | 1.46  | 97.50 | 23.48 | 1.46   |
| HEB      | Harbin       | 2.64%   | 2.53  | 98.09 | 24.40 | 1.03   |

Discussion

In this paper, a new hybrid model is proposed to accurately predict the short-term variability of PM2.5 concentrations. Based on previous studies, LASSO is used for the selection of factors influencing PM2.5 concentrations, which greatly improves the prediction accuracy of the model. This is similar to the results of the study by Koen et al. (2020). The LASSO approach does play a larger role in improving the accuracy of the model interpretation. In addition, the QRNN model is able to obtain the overall distribution of the response variables corresponding to the explanatory variables, which gives more information than the widely used mean regression (Sun et al. 2021). Based on the experimental results above, it can be seen that the model proposed in the paper can effectively select sequences that are significant to the prediction target and can obtain the probability distribution of the predicted values compared to the comparison model.

The two case studies and the six other datasets analyzed in this paper are from more polluted cities in China, for which PM2.5 prediction is more urgent. As the causes of high levels of PM2.5 vary from region to region, the LASSO method is used to select the influencing factors to obtain
more accurate predictions, as demonstrated in the comparison between case 1 and case 2. After the selection of influencing factors, the performance of the model improved significantly in all aspects.

The model presented in this paper performs well in predicting data sets from highly polluted areas, with better prediction performance for less fluctuating data sets. In future research, extension of the model to predict PM2.5 concentrations at lower levels and improvement of the model’s prediction performance on more volatile data sets could be considered. In addition, better selection methods for impact factors are also a future research direction.

Conclusion

In this paper, the LASSO-QR-KDE model is proposed to predict the probability density of PM2.5 concentrations. Comparative tests were constructed for analysis based on data sets from Beijing, Jinan, and six other regions in China. The model prediction performance is evaluated by using point predictors MAPE and RMSE and interval predictors PICP, PINAW, and PINAD, respectively, and compared with the control model to draw the following conclusions.

(1) The model proposed in the paper is applicable to the prediction of PM2.5 concentrations and shows high accuracy and robustness in prediction. In the comparison tests with other models and several data sets, the LASSO-QRNN-KDE model obtained the smallest MAPE and RMSE, which could be maintained at around 2% and 2. On the one hand, the LASSO method is able to reduce the dimensionality of the influencing factor data set without affecting the prediction performance and is suitable for prediction of PM2.5 concentrations with multiple influencing factors. On the other hand, the KDE model can overcome the shortcomings of parameter estimation and can fit the distribution according to the characteristics of PM2.5 concentration data, further reducing the prediction error. In addition, the LASSO-QRNN-KDE model can provide decision makers with more accurate prediction information by obtaining narrower prediction intervals and high interval coverage when making interval predictions. In tests on several datasets, the prediction intervals obtained deviated from the true values by a small margin, almost below 3%.

(2) When analyzing the datasets from Beijing and Jinan, China, different influencing factors were screened out, due to the variability of the regions. It is also because the model in this paper is able to identify the variability that it performs well in predicting PM2.5 concentrations. The model outperformed Beijing in predicting PM2.5 concentrations in Jinan for all indicators, due to the fact that the Jinan data is more stable than the Beijing data. This is also reflected in additional tests on six other datasets, with less volatile datasets having better predictions, such as the AL dataset.

(3) When the LASSO-QRNN-KDE model is used for prediction, the true value of the prediction point is likely to occur near the highest point of the probability density curve, which is the mode position. The model is able to obtain a probability density curve of PM2.5 concentrations for each day of the month, producing richer predictions than the single point prediction approach and the interval prediction approach. By making differentiated predictions for individual areas with high PM2.5 concentrations, the model can provide meaningful reference for relevant policy makers to effectively reduce and control air pollution and reduce public health risks.

Appendix 1

| Table 8 | Statistical description of data sets |
|---------|-------------------------------------|
| Cities  | Date                  | Size | Min | Max | Mean | Standard deviation | Variance |
| Beijing | 2018.01.01–2018.12.31 | 365  | 6   | 364 | 59.11 | 51.65               | 2677.11  |
| Jinan   | 2020.01.01–2020.12.31 | 365  | 2   | 282 | 63.01 | 41.04               | 1689.50  |
| Shijiazhuang | 2019.06.01–2021.05.31 | 730  | 17  | 326 | 77.27 | 52.15               | 2727.18  |
| Zhengzhou | 2019.06.01–2020.07.31 | 334  | 15  | 359 | 63.79 | 49.30               | 2437.86  |
| Xian    | 2020.04.01–2021.07.31 | 487  | 9   | 493 | 78.21 | 76.54               | 5873.71  |
| Alashan League | 2017.01.01–2018.07.31 | 577  | 12  | 315 | 35.22 | 23.27               | 542.86   |
| Shizuishan | 2015.01.01–2016.12.31 | 731  | 12  | 258 | 46.15 | 28.32               | 804.18   |
| Harbin  | 2019.01.01–2019.12.31 | 365  | 6   | 460 | 56.60 | 63.14               | 3997.66  |
Table 9  The results of the correlation analysis (Beijing)

| Variable | $Y_t$ | $X_1^t$ | $X_2^t$ | $X_3^t$ | $X_4^t$ | $X_5^t$ | $X_6^t$ | $X_7^t$ | $X_8^t$ | $X_9^t$ | $X_{10}^t$ |
|----------|-------|---------|---------|---------|---------|---------|---------|---------|---------|---------|----------|
| $Y_t$    | 1     |         |         |         |         |         |         |         |         |         |          |
| $X_1^t$  | 0.872*** | 1       |         |         |         |         |         |         |         |         |          |
| $X_2^t$  | 0.637*** | 0.592*** | 1       |         |         |         |         |         |         |         |          |
| $X_3^t$  | 0.850*** | 0.662*** | 0.572*** | 1       |         |         |         |         |         |         |          |
| $X_4^t$  | 0.755*** | 0.686*** | 0.594*** | 0.801*** | 1       |         |         |         |         |         |          |
| $X_5^t$  | -0.142*  | 0.023   | -0.255***| -0.269***| -0.320** | 1       |         |         |         |         |          |
| $X_6^t$  | -0.265***| -0.101  | -0.523***| -0.376***| -0.368***| 0.749***| 1       |         |         |         |          |
| $X_7^t$  | 0.327*** | 0.066   | -0.160***| 0.332*** | 0.150**  | 0.023   | 0.261** | 1       |         |         |          |
| $X_8^t$  | 0.165*** | -0.005  | 0.378*** | 0.272*** | 0.288*** | -0.726***| -0.896***| -0.210***| 1       |         |          |
| $X_9^t$  | -0.264***| -0.103  | -0.078  | -0.306***| -0.428***| 0.062   | -0.087  | -0.507***| -0.007 | 1       |          |
| $X_{10}^t$| 0.075    | 0.128   | -0.042  | -0.024  | 0.072   | 0.195** | 0.132*  | 0.01    | -0.161***| -0.049 | 1        |

Table 10  The results of the correlation analysis (Jinan)

| Variable | $Y_t$ | $X_1^t$ | $X_2^t$ | $X_3^t$ | $X_4^t$ | $X_5^t$ | $X_6^t$ | $X_7^t$ | $X_8^t$ | $X_9^t$ | $X_{10}^t$ |
|----------|-------|---------|---------|---------|---------|---------|---------|---------|---------|---------|----------|
| $Y_t$    | 1     |         |         |         |         |         |         |         |         |         |          |
| $X_1^t$  | 0.916*** | 1       |         |         |         |         |         |         |         |         |          |
| $X_2^t$  | 0.642*** | 0.603***| 1       |         |         |         |         |         |         |         |          |
| $X_3^t$  | 0.865*** | 0.770***| 0.616***| 1       |         |         |         |         |         |         |          |
| $X_4^t$  | 0.631*** | 0.607***| 0.547***| 0.707***| 1       |         |         |         |         |         |          |
| $X_5^t$  | -0.315***| -0.232***| -0.371***| -0.443***| -0.379***| 1       |         |         |         |         |          |
| $X_6^t$  | -0.443***| -0.367***| -0.693***| -0.480***| -0.464***| 0.749***| 1       |         |         |         |          |
| $X_7^t$  | 0.151*** | -0.031  | -0.346***| 0.251*** | 0.041   | -0.214***| 0.168***| 1       |         |         |          |
| $X_8^t$  | 0.361*** | 0.301***| 0.607***| 0.411***| 0.505***| -0.643***| -0.885***| -0.212***| 1       |         |          |
| $X_9^t$  | -0.152***| -0.033  | -0.007  | -0.288***| -0.447***| 0.119*  | 0.098   | -0.382***| -0.168***| 1       |          |
| $X_{10}^t$| 0.072   | 0.099   | -0.032  | -0.041  | -0.037  | 0.111   | 0.013   | -0.155***| -0.041  | 0.063   | 1        |
### Appendix 2

Table 11 LASSO regression results of Beijing data

| Step | Df | Rss  | Cp   | Step | Df | Rss  | Cp   | Step | Df | Rss  | Cp   |
|------|----|------|------|------|----|------|------|------|----|------|------|
| 0    | 1  | 4.76 | 419.17 | 19  | 16 | 2.24 | 70.00 | 38  | 31 | 1.83 | 37.67 |
| 1    | 2  | 4.18 | 333.85 | 20  | 17 | 2.24 | 71.06 | 39  | 32 | 1.83 | 39.35 |
| 2    | 3  | 3.33 | 208.06 | 21  | 18 | 2.23 | 71.49 | 40  | 33 | 1.82 | 39.81 |
| 3    | 4  | 3.00 | 160.10 | 22  | 19 | 2.14 | 60.01 | 41  | 34 | 1.80 | 39.78 |
| 4    | 5  | 2.94 | 152.46 | 23  | 20 | 2.11 | 57.53 | 42  | 35 | 1.79 | 39.09 |
| 5    | 6  | 2.74 | 125.38 | 24  | 21 | 2.10 | 59.03 | 43  | 36 | 1.78 | 39.93 |
| 6    | 7  | 2.68 | 118.01 | 25  | 22 | 2.07 | 55.93 | 44  | 37 | 1.77 | 41.10 |
| 7    | 8  | 2.68 | 118.95 | 26  | 21 | 2.07 | 53.53 | 45  | 38 | 1.76 | 41.31 |
| 8    | 9  | 2.65 | 117.39 | 27  | 22 | 2.05 | 52.10 | 46  | 39 | 1.75 | 41.62 |
| 9    | 10 | 2.65 | 118.72 | 28  | 23 | 2.04 | 53.22 | 47  | 40 | 1.74 | 42.04 |
| 10   | 9  | 2.60 | 110.15 | 29  | 24 | 2.04 | 55.05 | 48  | 41 | 1.71 | 40.00 |
| 11   | 10 | 2.57 | 106.98 | 30  | 25 | 2.03 | 56.14 | 49  | 42 | 1.71 | 41.79 |
| 12   | 11 | 2.47 | 94.52  | 31  | 26 | 1.97 | 48.49 | 50  | 41 | 1.70 | 38.48 |
| 13   | 12 | 2.38 | 82.09  | 32  | 27 | 1.92 | 43.32 | 51  | 42 | 1.70 | 40.39 |
| 14   | 11 | 2.35 | 75.60  | 33  | 28 | 1.91 | 43.94 | 52  | 43 | 1.70 | 42.19 |
| 15   | 12 | 2.34 | 76.18  | 34  | 29 | 1.89 | 43.30 | 53  | 44 | 1.70 | 37.67 |
| 16   | 13 | 2.32 | 75.94  | 35  | 30 | 1.85 | 38.86 | 54  | 45 | 1.83 | 39.35 |
| 17   | 14 | 2.28 | 70.92  | 36  | 31 | 1.84 | 38.96 | 55  | 46 | 1.82 | 39.81 |
| 18   | 15 | 2.26 | 70.80  | 37  | 32 | 1.83 | 40.20 | 56  | 47 | 1.80 | 39.78 |

Table 12 LASSO regression results of Jinan data

| Step | Df | Rss  | Cp   | Step | Df | Rss  | Cp   | Step | Df | Rss  | Cp   |
|------|----|------|------|------|----|------|------|------|----|------|------|
| 0    | 1  | 6.82 | 465.33 | 19  | 18 | 2.88 | 57.72 | 38  | 33 | 2.37 | 30.58 |
| 1    | 2  | 4.36 | 192.16 | 20  | 19 | 2.82 | 53.35 | 39  | 34 | 2.35 | 30.61 |
| 2    | 3  | 3.70 | 120.33 | 21  | 20 | 2.78 | 51.48 | 40  | 35 | 2.34 | 31.34 |
| 3    | 4  | 3.53 | 102.75 | 22  | 21 | 2.76 | 51.16 | 41  | 36 | 2.34 | 33.23 |
| 4    | 5  | 3.49 | 100.29 | 23  | 22 | 2.75 | 51.11 | 42  | 37 | 2.33 | 35.12 |
| 5    | 6  | 3.35 | 86.65  | 24  | 23 | 2.74 | 52.50 | 43  | 38 | 2.33 | 36.81 |
| 6    | 7  | 3.23 | 75.10  | 25  | 24 | 2.72 | 52.41 | 44  | 39 | 2.33 | 38.60 |
| 7    | 8  | 3.10 | 62.32  | 26  | 25 | 2.68 | 49.27 | 45  | 40 | 2.31 | 38.85 |
| 8    | 9  | 3.06 | 60.18  | 27  | 26 | 2.55 | 37.63 | 46  | 41 | 2.31 | 39.86 |
| 9    | 10 | 3.05 | 61.65  | 28  | 27 | 2.52 | 36.02 | 47  | 40 | 2.30 | 36.98 |
| 10   | 11 | 3.02 | 59.27  | 29  | 28 | 2.52 | 37.68 | 48  | 41 | 2.30 | 38.86 |
| 11   | 12 | 2.99 | 58.89  | 30  | 29 | 2.50 | 37.87 | 49  | 42 | 2.29 | 40.63 |
| 12   | 13 | 2.98 | 59.81  | 31  | 30 | 2.47 | 36.09 | 50  | 43 | 2.29 | 42.17 |
| 13   | 14 | 2.96 | 59.14  | 32  | 31 | 2.40 | 30.20 | 51  | 44 | 2.28 | 43.33 |
| 14   | 15 | 2.91 | 55.42  | 33  | 32 | 2.39 | 31.45 | 52  | 45 | 2.28 | 45.00 |
| 15   | 16 | 2.90 | 56.63  | 34  | 33 | 2.39 | 33.28 | 38  | 33 | 2.37 | 30.58 |
| 16   | 17 | 2.89 | 57.62  | 35  | 34 | 2.37 | 33.14 | 39  | 34 | 2.35 | 30.61 |
| 17   | 16 | 2.88 | 54.48  | 36  | 33 | 2.37 | 31.10 | 40  | 35 | 2.34 | 31.34 |
| 18   | 17 | 2.88 | 55.97  | 37  | 34 | 2.37 | 32.69 | 41  | 36 | 2.34 | 33.23 |
## Appendix 3

### Table 13  Prediction results for three situations of Beijing data

| Time/day | Actual value (μg/m³) | A | B | C |
|----------|----------------------|---|---|---|
|          |                      | Mode (μg/m³) | RE (%) | Mode (μg/m³) | RE (%) | Mode (μg/m³) | RE (%) |
| 1        | 28                   | 23.62        | 15.64 | 24.28        | 13.29 | 24.80        | 11.42 |
| 2        | 57                   | 48.83        | 14.33 | 50.36        | 11.65 | 52.79        | 7.39  |
| 3        | 123                  | 119.63       | 2.74  | 119.00       | 3.25  | 120.08       | 2.37  |
| 4        | 18                   | 17.20        | 4.43  | 19.07        | 5.95  | 17.32        | 3.78  |
| 5        | 17                   | 15.76        | 7.31  | 17.40        | 2.34  | 16.68        | 1.88  |
| 6        | 64                   | 70.25        | 9.76  | 67.67        | 5.74  | 62.66        | 2.09  |
| 7        | 34                   | 35.17        | 3.43  | 33.64        | 1.07  | 34.03        | 0.08  |
| 8        | 10                   | 9.50         | 5.01  | 10.12        | 1.23  | 9.71         | 2.92  |
| 9        | 41                   | 40.21        | 1.92  | 41.43        | 1.04  | 41.14        | 0.35  |
| 10       | 75                   | 72.85        | 2.87  | 77.27        | 3.03  | 77.35        | 3.13  |
| 11       | 103                  | 108.57       | 5.41  | 100.38       | 2.54  | 105.29       | 2.22  |
| 12       | 217                  | 218.91       | 0.88  | 221.36       | 2.01  | 214.63       | 1.09  |
| 13       | 128                  | 113.51       | 11.32 | 118.49       | 7.43  | 129.29       | 1.01  |
| 14       | 104                  | 100.32       | 3.54  | 101.32       | 2.58  | 108.19       | 4.03  |
| 15       | 9                    | 8.46         | 6.03  | 9.43         | 4.77  | 8.76         | 2.72  |
| 16       | 38                   | 39.85        | 4.87  | 37.46        | 1.43  | 36.69        | 3.45  |
| 17       | 45                   | 47.58        | 5.73  | 44.41        | 1.32  | 44.69        | 0.69  |
| 18       | 64                   | 72.95        | 13.98 | 62.35        | 2.58  | 63.18        | 1.28  |
| 19       | 11                   | 12.05        | 9.51  | 11.24        | 2.14  | 10.83        | 1.54  |
| 20       | 9                    | 9.35         | 3.88  | 8.91         | 0.99  | 9.05         | 0.60  |
| 21       | 24                   | 24.63        | 2.63  | 23.39        | 2.55  | 23.17        | 3.47  |
| 22       | 28                   | 30.06        | 7.34  | 28.58        | 2.08  | 27.71        | 1.02  |
| 23       | 23                   | 21.15        | 8.05  | 22.71        | 1.26  | 23.30        | 1.32  |
| 24       | 58                   | 60.45        | 4.22  | 60.00        | 3.45  | 56.91        | 1.88  |
| 25       | 10                   | 10.58        | 5.82  | 10.77        | 7.69  | 9.54         | 4.61  |
| 26       | 25                   | 24.04        | 3.86  | 25.36        | 1.43  | 24.77        | 0.93  |
| 27       | 49                   | 50.97        | 4.02  | 49.98        | 2.00  | 49.37        | 0.76  |
| 28       | 35                   | 36.97        | 5.63  | 33.91        | 3.11  | 35.46        | 1.31  |
| 29       | 106                  | 104.61       | 1.31  | 104.03       | 1.86  | 107.64       | 1.55  |
| 30       | 28                   | 29.31        | 4.69  | 27.50        | 1.78  | 28.57        | 2.02  |
| 31       | 13                   | 12.07        | 7.18  | 13.40        | 3.07  | 13.20        | 1.57  |
Table 14  Prediction results for three situations of Jinan data

| Time/day | Actual value (µg/m³) | A Mode (µg/m³) | RE (%) | B Mode (µg/m³) | RE (%) | C Mode (µg/m³) | RE (%) |
|----------|----------------------|----------------|--------|----------------|--------|----------------|--------|
| 1        | 46                   | 39.21          | 14.77  | 41.25          | 10.33  | 43.00          | 6.52   |
| 2        | 65                   | 70.17          | 7.96   | 72.45          | 11.46  | 68.66          | 5.63   |
| 3        | 142                  | 129.76         | 8.62   | 128.33         | 9.63   | 136.65         | 3.77   |
| 4        | 133                  | 141.55         | 6.43   | 139.28         | 4.72   | 135.78         | 2.09   |
| 5        | 28                   | 30.62          | 9.34   | 26.36          | 5.87   | 27.34          | 2.34   |
| 6        | 37                   | 39.18          | 5.88   | 35.93          | 2.88   | 35.86          | 3.08   |
| 7        | 16                   | 15.16          | 5.24   | 16.39          | 2.43   | 16.23          | 1.45   |
| 8        | 29                   | 28.48          | 1.79   | 28.50          | 1.71   | 29.67          | 2.32   |
| 9        | 30                   | 32.60          | 8.66   | 28.10          | 6.32   | 31.14          | 3.81   |
| 10       | 32                   | 33.20          | 3.74   | 33.73          | 5.42   | 30.59          | 4.41   |
| 11       | 38                   | 37.13          | 2.29   | 38.58          | 1.53   | 39.77          | 4.67   |
| 12       | 58                   | 53.57          | 7.63   | 61.17          | 5.46   | 56.20          | 3.11   |
| 13       | 70                   | 66.42          | 5.12   | 72.88          | 4.11   | 69.67          | 0.47   |
| 14       | 93                   | 98.04          | 5.42   | 87.36          | 6.06   | 93.05          | 0.05   |
| 15       | 95                   | 90.27          | 4.98   | 100.12         | 5.39   | 93.97          | 1.08   |
| 16       | 46                   | 47.68          | 3.65   | 44.42          | 3.43   | 44.92          | 2.34   |
| 17       | 49                   | 47.90          | 2.25   | 50.89          | 3.85   | 48.33          | 1.36   |
| 18       | 43                   | 41.53          | 3.41   | 44.24          | 2.88   | 44.00          | 2.32   |
| 19       | 56                   | 59.48          | 6.22   | 53.89          | 3.76   | 57.10          | 1.97   |
| 20       | 93                   | 97.93          | 5.3    | 88.47          | 4.87   | 91.55          | 1.56   |
| 21       | 148                  | 138.97         | 6.1    | 141.50         | 4.39   | 144.54         | 2.34   |
| 22       | 212                  | 217.94         | 2.8    | 220.67         | 4.09   | 216.43         | 2.09   |
| 23       | 107                  | 101.91         | 4.76   | 103.05         | 3.69   | 104.51         | 2.33   |
| 24       | 139                  | 132.01         | 5.03   | 141.61         | 1.88   | 140.08         | 0.78   |
| 25       | 168                  | 162.72         | 3.14   | 163.68         | 2.57   | 170.39         | 1.42   |
| 26       | 135                  | 133.62         | 1.02   | 130.37         | 3.43   | 133.27         | 1.28   |
| 27       | 77                   | 79.07          | 2.69   | 79.15          | 2.79   | 78.86          | 2.42   |
| 28       | 34                   | 32.35          | 4.84   | 35.52          | 4.46   | 32.69          | 3.84   |
| 29       | 44                   | 45.31          | 2.98   | 42.01          | 4.52   | 44.09          | 0.21   |
| 30       | 64                   | 61.27          | 4.26   | 61.73          | 3.54   | 63.66          | 0.53   |
| 31       | 83                   | 87.17          | 5.03   | 85.71          | 3.26   | 84.23          | 1.48   |
## Appendix 4

### Table 15 Predicted values under various models of Beijing data (µg/m²)

| Time/day | Actual value | LASSO-QRNN-KDE | ARIMA | BP | RBF | CNN-LSTM | EMD-GRU |
|----------|--------------|----------------|-------|----|-----|----------|---------|
| 1        | 24.80        | 31.66          | 22.15 | 30.08 | 26.57 | 29.58    | 24.80   |
| 2        | 52.79        | 75.88          | 48.76 | 58.70 | 49.06 | 58.42    | 52.79   |
| 3        | 120.08       | 97.35          | 109.97| 107.47| 104.16| 131.36   | 120.08  |
| 4        | 17.32        | 21.32          | 14.88 | 20.40 | 18.90 | 17.30    | 17.32   |
| 5        | 17.32        | 21.04          | 19.62 | 15.56 | 18.27 | 16.35    | 17.32   |
| 6        | 62.66        | 53.18          | 52.24 | 59.35 | 59.56 | 65.24    | 62.66   |
| 7        | 34.03        | 41.57          | 38.51 | 40.83 | 38.47 | 31.69    | 34.03   |
| 8        | 9.71         | 12.92          | 8.99  | 8.66  | 10.36 | 9.66     | 9.71    |
| 9        | 41.14        | 36.59          | 43.71 | 34.10 | 45.35 | 38.58    | 41.14   |
| 10       | 77.35        | 70.12          | 68.60 | 82.55 | 71.96 | 78.59    | 77.35   |
| 11       | 105.29       | 70.62          | 85.16 | 81.75 | 109.08| 95.95    | 105.29  |
| 12       | 214.63       | 155.48         | 175.70| 259.99| 225.22| 204.08   | 214.63  |
| 13       | 129.29       | 160.92         | 155.42| 134.66| 150.35| 137.06   | 129.29  |
| 14       | 108.19       | 122.76         | 127.12| 95.41 | 88.65 | 107.02   | 108.19  |
| 15       | 8.76         | 7.43           | 10.42 | 7.97  | 8.19  | 9.45     | 8.76    |
| 16       | 36.69        | 47.78          | 30.97 | 43.21 | 33.74 | 35.42    | 36.69   |
| 17       | 45.31        | 35.75          | 53.01 | 54.40 | 42.92 | 41.12    | 45.31   |
| 18       | 63.18        | 52.06          | 79.40 | 55.17 | 56.25 | 61.60    | 63.18   |
| 19       | 10.83        | 8.86           | 12.37 | 9.83  | 11.52 | 10.26    | 10.83   |
| 20       | 9.05         | 11.12          | 7.93  | 8.24  | 8.50  | 9.42     | 9.05    |
| 21       | 23.17        | 27.51          | 28.76 | 27.50 | 22.04 | 22.41    | 23.17   |
| 22       | 27.71        | 24.72          | 22.25 | 31.82 | 24.50 | 25.91    | 27.71   |
| 23       | 23.30        | 24.15          | 27.58 | 25.93 | 24.51 | 25.61    | 23.30   |
| 24       | 56.91        | 62.00          | 51.74 | 44.56 | 51.85 | 52.41    | 56.91   |
| 25       | 9.54         | 6.97           | 10.76 | 11.64 | 11.10 | 9.14     | 9.54    |
| 26       | 24.77        | 28.46          | 20.14 | 21.17 | 22.60 | 24.42    | 24.77   |
| 27       | 49.37        | 55.32          | 37.00 | 54.63 | 54.08 | 49.73    | 49.37   |
| 28       | 35.46        | 26.02          | 42.02 | 38.84 | 37.89 | 37.04    | 35.46   |
| 29       | 107.64       | 144.94         | 85.87 | 99.27 | 100.62| 116.27   | 107.64  |
| 30       | 28.57        | 33.98          | 21.60 | 24.02 | 23.36 | 26.26    | 28.57   |
| 31       | 13.20        | 9.65           | 15.53 | 11.72 | 11.36 | 13.55    | 13.20   |
Table 16 Relative errors under various models of Beijing data (%)

| Time/day | LASSO-QRNN-KDE | ARIMA | BP   | RBF  | CNN-LSTM | EMD-GRU |
|----------|----------------|-------|------|------|----------|---------|
| 1        | 11.42          | 13.07 | 20.88| 7.42 | 5.11     | 5.63    |
| 2        | 7.39           | 33.12 | 14.46| 2.98 | 13.93    | 2.49    |
| 3        | 2.37           | 20.85 | 10.59| 12.63| 15.32    | 6.80    |
| 4        | 3.78           | 18.45 | 17.32| 13.31| 4.99     | 3.89    |
| 5        | 1.88           | 23.77 | 15.44| 8.48 | 7.46     | 3.80    |
| 6        | 2.09           | 16.90 | 18.38| 7.26 | 6.93     | 1.94    |
| 7        | 0.08           | 22.25 | 13.25| 20.09| 13.14    | 6.80    |
| 8        | 2.92           | 29.24 | 10.09| 13.42| 3.55     | 3.38    |
| 9        | 0.35           | 10.76 | 6.60 | 16.84| 10.62    | 5.91    |
| 10       | 3.13           | 6.51  | 8.54 | 10.06| 4.06     | 4.78    |
| 11       | 2.22           | 31.44 | 17.32| 20.63| 5.90     | 6.84    |
| 12       | 1.09           | 28.35 | 19.03| 19.81| 3.79     | 5.95    |
| 13       | 1.01           | 25.72 | 21.42| 5.20 | 17.46    | 7.08    |
| 14       | 4.03           | 18.04 | 22.23| 8.26 | 14.76    | 2.91    |
| 15       | 2.72           | 17.40 | 15.76| 11.45| 9.01     | 5.02    |
| 16       | 3.45           | 25.73 | 18.50| 13.70| 11.20    | 6.80    |
| 17       | 0.69           | 20.56 | 17.79| 20.88| 4.62     | 8.63    |
| 18       | 1.28           | 18.66 | 24.07| 13.79| 12.11    | 3.75    |
| 19       | 1.54           | 19.42 | 12.43| 10.66| 4.69     | 6.75    |
| 20       | 0.60           | 23.58 | 11.86| 8.48 | 5.53     | 4.66    |
| 21       | 3.47           | 14.62 | 19.84| 14.57| 8.18     | 6.61    |
| 22       | 1.02           | 11.73 | 20.53| 13.64| 12.49    | 7.45    |
| 23       | 1.32           | 5.02  | 19.93| 12.75| 6.59     | 11.36   |
| 24       | 1.88           | 6.89  | 10.80| 23.18| 10.60    | 9.64    |
| 25       | 4.61           | 30.33 | 7.63 | 16.37| 10.98    | 8.63    |
| 26       | 0.93           | 13.82 | 19.45| 15.32| 9.61     | 2.33    |
| 27       | 0.76           | 12.90 | 24.48| 11.48| 10.37    | 1.49    |
| 28       | 1.31           | 25.67 | 20.05| 10.97| 8.25     | 5.83    |
| 29       | 1.55           | 36.74 | 18.99| 6.35 | 5.07     | 9.69    |
| 30       | 2.02           | 21.36 | 22.85| 14.22| 16.59    | 6.21    |
| 31       | 1.57           | 25.79 | 19.47| 9.81 | 12.64    | 4.23    |
Table 17: Predicted values under various models of Jinan data (μg/m³)

| Time/day | Actual value | LASSO-QRNN-KDE | ARIMA | BP | RBF | CNN-LSTM | EMD-GRU |
|----------|--------------|----------------|-------|----|-----|----------|--------|
| 1        | 46           | 43.00          | 36.65 | 38.90 | 49.37 | 47.60    | 40.88  |
| 2        | 65           | 68.66          | 52.15 | 57.93 | 61.16 | 61.17    | 60.94  |
| 3        | 142          | 136.65         | 159.07 | 124.32 | 154.55 | 161.06   | 130.78 |
| 4        | 133          | 135.78         | 112.45 | 111.48 | 148.07 | 146.57   | 138.88 |
| 5        | 28           | 27.34          | 24.25 | 31.82 | 24.63 | 24.79    | 30.06  |
| 6        | 37           | 35.86          | 30.47 | 45.90 | 40.46 | 34.17    | 39.12  |
| 7        | 16           | 16.23          | 18.90 | 13.24 | 17.37 | 15.09    | 16.78  |
| 8        | 29           | 29.67          | 30.28 | 33.81 | 30.13 | 30.40    | 28.10  |
| 9        | 30           | 31.14          | 30.46 | 33.26 | 25.81 | 26.23    | 30.65  |
| 10       | 32           | 30.59          | 34.81 | 27.29 | 27.93 | 29.09    | 31.28  |
| 11       | 38           | 39.77          | 42.14 | 30.66 | 33.62 | 35.20    | 39.42  |
| 12       | 58           | 56.20          | 49.18 | 50.19 | 51.67 | 61.47    | 59.93  |
| 13       | 70           | 69.67          | 83.12 | 78.88 | 74.07 | 74.29    | 68.56  |
| 14       | 93           | 93.05          | 112.21 | 104.38 | 85.14 | 86.26    | 87.89  |
| 15       | 95           | 93.97          | 123.91 | 108.10 | 88.16 | 88.51    | 99.05  |
| 16       | 46           | 44.92          | 57.08 | 51.20 | 51.79 | 44.04    | 42.37  |
| 17       | 49           | 48.33          | 59.23 | 43.63 | 44.36 | 51.76    | 43.84  |
| 18       | 43           | 44.00          | 34.77 | 48.04 | 38.67 | 45.67    | 44.00  |
| 19       | 56           | 57.10          | 46.13 | 62.96 | 50.47 | 58.42    | 54.84  |
| 20       | 93           | 91.55          | 75.11 | 80.31 | 84.65 | 98.33    | 89.80  |
| 21       | 148          | 144.54         | 175.25 | 166.57 | 131.22 | 138.75   | 138.90 |
| 22       | 212          | 216.43         | 246.75 | 246.13 | 234.75 | 220.88   | 205.09 |
| 23       | 107          | 104.51         | 139.12 | 91.24 | 94.71 | 113.23   | 110.57 |
| 24       | 139          | 140.08         | 174.56 | 155.64 | 160.00 | 158.11   | 145.70 |
| 25       | 168          | 170.39         | 134.22 | 147.49 | 185.34 | 158.73   | 175.41 |
| 26       | 135          | 133.27         | 108.04 | 119.41 | 153.13 | 128.74   | 142.24 |
| 27       | 77           | 78.86          | 90.75 | 85.26 | 86.36 | 81.71    | 74.77  |
| 28       | 34           | 32.69          | 40.29 | 30.67 | 37.19 | 32.09    | 35.71  |
| 29       | 44           | 44.09          | 52.79 | 39.60 | 49.02 | 41.10    | 45.02  |
| 30       | 64           | 63.66          | 50.85 | 69.27 | 72.86 | 66.39    | 60.81  |
| 31       | 83           | 84.23          | 98.54 | 89.08 | 95.97 | 87.17    | 78.01  |
Table 18 Relative errors under various models of Jinan data (%)

| Time/day | LASSO-QRNN-KDE | ARIMA | BP | RBF | CNN-LSTM | EMD-GRU |
|----------|----------------|-------|----|-----|----------|---------|
| 1        | 6.52           | 20.32 | 15.44 | 7.33 | 3.47     | 11.14   |
| 2        | 5.63           | 19.77 | 10.87 | 5.91 | 5.89     | 6.25    |
| 3        | 3.77           | 12.02 | 12.45 | 8.84 | 13.42    | 7.90    |
| 4        | 2.09           | 15.45 | 16.18 | 11.33 | 10.20    | 4.42    |
| 5        | 2.34           | 13.38 | 13.64 | 12.05 | 11.45    | 7.34    |
| 6        | 3.08           | 17.64 | 24.06 | 9.36 | 7.66     | 5.72    |
| 7        | 1.45           | 18.15 | 17.23 | 8.57 | 5.71     | 4.85    |
| 8        | 2.32           | 4.43  | 16.59 | 3.88 | 4.83     | 3.09    |
| 9        | 3.81           | 1.54  | 10.88 | 13.97 | 12.56    | 2.16    |
| 10       | 4.41           | 8.77  | 14.71 | 12.72 | 9.10     | 2.25    |
| 11       | 4.67           | 10.90 | 19.32 | 11.53 | 7.36     | 3.74    |
| 12       | 3.11           | 15.20 | 13.46 | 10.91 | 5.99     | 3.33    |
| 13       | 0.47           | 18.74 | 12.69 | 5.82 | 6.13     | 2.06    |
| 14       | 0.05           | 20.66 | 12.24 | 8.45 | 7.25     | 5.50    |
| 15       | 1.08           | 30.43 | 13.79 | 7.20 | 6.83     | 4.26    |
| 16       | 2.34           | 24.09 | 11.30 | 12.58 | 4.27     | 7.89    |
| 17       | 1.36           | 20.88 | 10.96 | 9.46 | 5.63     | 10.53   |
| 18       | 2.32           | 19.15 | 11.73 | 10.08 | 6.20     | 2.32    |
| 19       | 1.97           | 17.63 | 12.42 | 9.87 | 4.33     | 2.08    |
| 20       | 1.56           | 19.24 | 13.64 | 8.98 | 5.73     | 3.44    |
| 21       | 2.34           | 18.41 | 12.55 | 11.34 | 6.25     | 6.15    |
| 22       | 2.09           | 16.39 | 16.10 | 10.73 | 4.19     | 3.26    |
| 23       | 2.33           | 30.02 | 14.73 | 11.49 | 5.82     | 3.34    |
| 24       | 0.78           | 25.58 | 11.97 | 15.11 | 13.75    | 4.82    |
| 25       | 1.42           | 20.11 | 12.21 | 10.32 | 5.52     | 4.41    |
| 26       | 1.28           | 19.97 | 11.55 | 13.43 | 4.64     | 5.36    |
| 27       | 2.42           | 17.86 | 10.73 | 12.15 | 6.12     | 2.90    |
| 28       | 3.84           | 18.49 | 9.79  | 9.39  | 5.62     | 5.04    |
| 29       | 0.21           | 19.98 | 10.01 | 11.42 | 6.60     | 2.31    |
| 30       | 0.53           | 20.54 | 8.24  | 13.84 | 3.74     | 4.99    |
| 31       | 1.48           | 18.72 | 7.32  | 15.63 | 5.03     | 6.01    |

Author contribution Shaomei Yang: methodology, resources, and writing — review and editing. Haoyue Wu: conceptualization, software, validation, investigation, data curation, and writing — original draft.

Data availability The datasets analysed during the current study are available in the National Meteorological Data Center of China repository, http://data.cma.cn/.

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Competing interest The authors declare no competing interests.

References

Breiman L (1995) Better subset regression using the nonnegative garrote. Technometrics 37(4):373–384
Cao Q, Lian S, Chen SC et al (2018) WRF modeling of PM2.5 remediation by SALSCS and its clean air flow over Beijing terrain. Sci Total Environ 626(JUN.1):134
Cao MN, Tian P, Li GR (2021) Study on Lasso penalized quantile regression of optimal portfolio. J Syst Sci Math Sci 1–17
Geng G, Zhang Q, Martin RV et al (2015) Estimating long-term PM2.5 concentrations in China using satellite-based aerosol optical depth and a chemical transport model. Remote Sens Environ 166:262–270
He YY, Xu QF, Yang SL et al (2013) A power load probability density forecasting method based on RBF neural network quantile regression. Zhongguo Dianji Gongcheng Xuebao/Proc Chin Soc Electr Eng 33(1):93–98
He YY, Yang Q, Wang S et al (2019) Electricity consumption probability density forecasting method based on LASSO-quantile regression neural network. Appl Energy 233–234(JAN.1):565–575
He P, Lan W, Ding Y (2021) Is the Chinese stock market predictable?—an evidence based on the combination LASSO-logistic model. Stat Res 38(05):82–96

Hoerl A, Kennard R (2000) Taylor & Francis Online: Ridge regression: biased estimation for nonorthogonal problems - Technometrics - Volume 12, Issue 1. Technometrics 42(1):7

Huang CJ, Kuo PH (2018) A deep CNN-LSTM model for particulate matter (PM2.5) Forecasting in Smart Cities. Sensors 18(7):2220

Huang G, Li X, Zhang B et al (2021) PM2.5 concentration forecasting at surface monitoring sites using GRU neural network based on empirical mode decomposition. Sci Total Environ 768:144516

Jasleen KS, Mamta M (2021) An efficient correlation based adaptive LASSO regression method for air quality index prediction. Earth Sci Informat (prepublish)

Jiang F, Qiao YQ (2021) PM2.5 Concentration prediction based on sample entropy and improved extreme learning machine. Statist Decis 37(03):166–171

Kang Q, Wu J, Chen M, Jeon BN (2021) Do macroprudential policies affect the bank financing of firms in China? Evidence from a quantile regression approach. J Int Money Financ 115

Koenker R (2015) Quantile regression. J Econ Perspect 15(4):143–156

Koen F, Boshuizen HC, Nielen MMJ et al (2020) Mapping chronic disease prevalence based on medication use and socio-demographic variables: An application of LASSO in healthcare in the Netherlands

Koenker R, Bassett GW (1978) Regression quantiles[J]. Econometrica 46(1):211–244

Lee D, Mukhopadhyay S, Rushworth A et al (2017) A rigorous statistical framework for spatio-temporal pollution prediction and estimation of its long-term impact on health. Biostatistics 18(2):370–385

Li C, Lou C, Luo D, Xing K (2021) Chinese corporate distress prediction using LASSO: the role of earnings management. Int Rev Financ Anal (prepublish)

Liao T, Wang S, Ai J et al (2017) Heavy pollution episodes, transport pathways and potential sources of PM2.5 during the winter of 2013 in Chengdu (China). Sci Total Environ 584–585(apr.15):1056–1065

Lightstone SD, Moshary F, Gross B (2017) Comparing CMAQ forecasts with a neural network forecast model for PM2.5 in New York. Atmosphere 8(12):161

Li W, Kong D, Wu J (2017) A new hybrid model FPA-SVM considering cointegration for estimation of pollution concentration forecasting: a case study of Kunming and Yuxi, China. Comput Intell Neurosci 2017:1–11

Madsen K, Nielsen HB (1993) A finite smoothing algorithm for linear $\Sigma_1$ $\Sigma$ estimation. SIAM J Optim 3(2):223–235

Masiol M, Formenton G, Pasqualetto A et al (2013) Seasonal trends and spatial variations of PM10-bounded polycyclic aromatic hydrocarbons in Veneto Region, Northeast Italy. Atmos Environ 79(nov.):811–821

Mukhopadhyay S, Sahu SK (2018) A Bayesian spatiotemporal model to estimate long-term exposure to outdoor air pollution at coarser administrative geographies in England and Wales. J R Stat Soc Ser A Stat Soc 181(2):465–486

Parzen E (1962) On estimation of probability density function and model[J]. Ann Math Stat 33(3):1065–1076

Portnoy S, Koenker R (1997) The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. Stat Sci 12(4)

Rosenblatt M (1956) Remarks on some nonparametric estimates of a density function[J]. Annals of Mathematical Statistics 27(3):832–837

Ruby-Figueroa R, Saavedra J, Bahamonde N et al (2016) Permeate flux prediction in the ultrafiltration of fruit juices by ARIMA models. J Memb Sci S0376773816307207

Salari TE, Roumiani A, Kazemzadeh E (2021) Globalization, renewable energy consumption, and agricultural production impacts on ecological footprint in emerging countries: using quantile regression approach. Environ Sci Pollut Res Int

Schwarz GE (1978) Estimating the dimension of a model. Ann Stat 6(2)

Sheather SJ, Jones MC (1991) A reliable data-based bandwidth selection method for kernel density estimation. J R Stat Soc Ser B: Methodol 53(3):683–690

Shen Y, Wu W, Jie C (2018) Wind power forecasting using multi-objective evolutionary algorithms for wavelet neural network-optimized prediction intervals. Appl Sci 8(2):185

Sun J, Zhao X, Xu C (2021) Crude oil market autocorrelation: evidence from multiscale quantile regression analysis. Energy Econ 98

Syarakov D, Prodanova M, Georgieva E et al (2016) Simulation of European air quality by WRF–CMAQ models using AQMEII-2 infrastructure. J Comput Appl Math 232–245

Taha A, Rahim A, Habashah M, Intisar IA (2018) Bayesian variable selection and coefficient estimation in heteroscedastic linear regression model. J Appl Stat 45(14)

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B: Statistical Methodol 53(1):267–288

Wang H, Li J, Peng Y et al (2019) The impacts of the meteorology features on PM$_{(2.5)}$ levels during a severe haze episode in central-east China. Atmos Environ 197(JAN):177–189

Xiao L, Lang Y, Christakos G (2018) High-resolution spatiotemporal mapping of PM2.5 concentrations at mainland China using a combined BME-GWR technique. Atmos Environ 173(jan.):295–305

Xu H, Bechle MJ, Wang M et al (2019) National PM$_{(2.5)}$ and NO$_{(2)}$ exposure models for China based on land use regression, satellite measurements, and universal kriging. Sci Total Environ 655(MAR.10):423–433

Yang X, Wu Q, Zhao R et al (2019) New method for evaluating winter air quality: PM2.5 assessment using community multi-scale air quality modeling (CMAQ) in Xi’an. Atmos Environ 211(AUG.):18–28

Zhang L, Lin J, Qiu R et al (2018) Trend analysis and forecast of PM2.5 in Fuzhou, China using the ARIMA model. Ecol Indic 85(DEC.):702–710

Zhou K, Qu Z (2017) Application of BP neural network optimized by genetic simulated annealing algorithm to prediction of air quality index in Lanzhou. 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA). IEEE

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.