RF-KDE-QSR Model for Estimating the Scale of Epidemics

Chuwei Liu, Jianping Huang, Siyu Chen, Jiaqi He, Shikang Du, Nan Yin, Chao Zhang, and Danfeng Wang

***Abstract*—Infectious diseases are posing an increasingly serious threat to human society. It is urgent to make a rapid estimate of the scale of outbreaks when the disease information is still unclear in the early stages of the outbreak, so as to buy time for a timely response to infectious diseases and provide reference for the allocation of medical resources and the formulation of control measures. Based on this, this study took the concentrated outbreak of COVID-19 in various cities in China as an example, collected 22 meteorological, social-ecological and population mobility indicators, and established a random forest-kernel density estimation-quantile stepwise regression (RF-KDE-QSR) model to make a preliminary estimate of the daily outbreak scale in cities. The RF model was used for preliminary estimation, and the KDE-QSR model was used for residual correction to correct the prediction results. The evaluation of the prediction accuracy proved the effectiveness of the prediction model. When the RF model was used alone, the R-squared (R^2) was 0.82 and the corrected R^2 was 0.90. The KDE-QSR model effectively improved the prediction accuracy of the model.**

***Index Terms*—Concentrated outbreak, COVID-19, epidemic, random forest (RF), kernel density estimation (KDE), quantile stepwise regression (QSR)**

I. Introduction

An epidemic is defined as a sudden increase in the number of disease cases in a specific area or population within a short period of time, exceeding the normally expected level [1], [2]. Since the 21st century, the accelerated urbanization process and the overuse of the natural environment have increased the possibility of disease outbreaks becoming epidemics [3], [4]. In recent years, infectious diseases have become more and more frequent, and this trend is likely to continue and worsen in the future. The occurrence and spread of epidemics must be observed and prevented to ensure public health safety. It is particularly important to find the key natural and social factors that affect the occurrence of infectious diseases and build models to quickly estimate the scale of the disease in the early stages of the disease.

As the most serious epidemic so far in the 21st century, the novel coronavirus pneumonia emerged at the end of 2019 and rapidly spread worldwide. COVID-19 has a high incidence rate and a long duration. It has had a significant negative impact on human health, social and economic growth, and other aspects [5]. China is among the nations with the most effective epidemic control measures. However, the spread of COVID-19 continues to have a serious impact on the economic growth and public life of China [6], [7]. Studies have shown that, like other infectious diseases, climate and environmental factors [8], [9], [10], socioeconomic factors [11], [12], and population migration and aggregation [13], [14] all have a significant impact on the scale of COVID-19. This makes it possible to use the above factors to estimate the scale of COVID-19.

This study aims to establish a model taking the outbreak of COVID-19 in China as an example, using available or predictable indicators such as meteorological indicators, socioeconomic indicators, and population migration indicators, to quickly estimate the scale of the infectious disease in the city in the early stages of the infectious disease when there is still a lack of disease-related data, and provide a reference for the control of infectious diseases and the normalization of infectious disease prevention. We use the random forest (RF) model, kernel density estimation (KDE) method, and quantile stepwise regression (QSR) model to construct a model framework for estimating the scale of concentrated infectious disease outbreaks with socioeconomic indicator data, meteorological data, and population mobility data. The specific contributions are as follows:

1) The RF-KDE-QSR model is constructed to quickly estimate the scale of the outbreak in the early stage of an infectious disease outbreak, when there is a lack of virus detection and detailed epidemiological information, cure, death, etc. This helps to quickly respond to sudden infectious diseases and also provides support for the normalization of infectious disease prevention;
2) A residual correction method combining KDE and QSR is developed to effectively improve the prediction accuracy of RF.

The rest of this article is organized as follows. Section II introduces related work and background. Section III describes

Received 19 June 2024; revised 8 October 2024 and 29 October 2024; accepted 18 November 2024. Date of publication 5 December 2024; date of current version 2 June 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 423B2506 and Grant 42175106, in part by China Meteorological Administration under Grant U2242209, in part by the Major Project of Guangzhou National Laboratory under Grant GZNL2024A01004, in part by Gansu Province Intellectual Property Program (Oriented Organization) Project under Grant 22ZSCQD02, and in part by the Outstanding Postgraduate "Innovation Star" Project of the Department of Education of Gansu Province, China under Grant 2023CXZX-103. *(Corresponding author: Jianping Huang.)*

Chuwei Liu, Jianping Huang, Siyu Chen, Chao Zhang, and Danfeng Wang are with the Collaborative Innovation Center for Western Ecological Safety, College of Atmospheric Sciences, Lanzhou University, Lanzhou 730000, China (e-mail: liuchw19@lzu.edu.cn; hjp@lzu.edu.cn; chensiyu@lzu.edu.cn).

Jiaqi He, Shikang Du, and Nan Yin are with the College of Earth and Environmental Science, Lanzhou University, Lanzhou 730000, China.

This article has supplementary downloadable material available at https://doi.org/10.1109/TCSS.2024.3507733, provided by the authors.

Digital Object Identifier 10.1109/TCSS.2024.3507733

the proposed framework and data in detail. Section IV provides experimental results and performance analysis, and discusses them. Section V summarizes this study, and Section VI discusses the generalization and application potential of the model as well as its limitations.

II. Related Work and Background

Many prediction methods are applied to predict the scale of epidemics. Traditional time series models are relatively simple and easy to operate models. They capture the patterns and structures of data changes over time, understand and summarize past behaviors, and predict the future [15], such as the autoregressive integrated moving average (ARIMA) model [16] and the Markov chain model [17]. They are widely used in infectious disease prediction. However, the original traditional time series models rely on historical data, and they cannot capture external factors such as public health interventions, climate change, and social behavior changes that may play an important role in the spread of infectious diseases, nor can they capture complex nonlinear relationships [18].

System dynamics modeling is more commonly used in infectious disease prediction. This type of model simulates and predicts the spread of infectious diseases by simulating the behaviors of the components of the system and the complex interactions between their feedback relationships and steady-state equilibrium [19], [20]. The more common models include the susceptible-infectious-removed (SIR) model [21] and its extensions, such as the susceptible-exposed-infectious-recovered (SEIR) model [22], [23], [24], [25] which describes the transmission process more complexly, and the susceptible-infectious-susceptible (SIS) model [26], which is used to describe the transmission process of infectious diseases without the formation of lasting immunity. This type of model provides an in-depth understanding of the transmission mechanism and has good scalability. It can incorporate external environmental factors such as climate, social and human behaviors [27], [28] into the model to describe the process of epidemic spread in more detail. However, this type of model requires more detailed data to accurate estimation of parameters such as infection rate and recovery rate, which depends on virus detection, epidemiological information, and expert expertise. This limits the application of the model.

Machine learning is widely used in infectious disease prediction. They do not require complex assumptions and can well capture nonlinear relationships in data, obtain the laws and trends of infectious disease transmission, and thus make predictions [29]. Many machine learning models have achieved good results in predicting the incidence of COVID-19. For example, artificial neural networks (ANN) and long short-term memory (LSTM) are used to predict the spread of COVID-19 at the global and national scales [30], [31]. RF and decision tree algorithms are used to predict the number of cases [32]. Support vector machines (SVM) are also used to predict the number of cases and deaths of COVID-19 [33]. These models are more flexible and less dependent on early-stage outbreak data, allowing for the consideration of various external factors that influence the spread of infectious diseases. Some existing studies focus on the national or provincial scale. Our study focuses on a more detailed city scale, taking into account the differences in external factors of each city. We considered the different development levels and urban functional positioning of each city, and introduced indicators such as per capita road area, urban passenger volume, and the proportion of science and education expenditure, which are still relatively lacking in existing forecasts. Our prediction indicators did not include disease-related data, which may lose prediction accuracy, but it enables to quickly estimate the scale of the disease in the early stage of the infectious disease when there is a lack of disease data. In addition, we constructed a KDE-QSR model to capture the error characteristics of the RF prediction results and correct the prediction results.

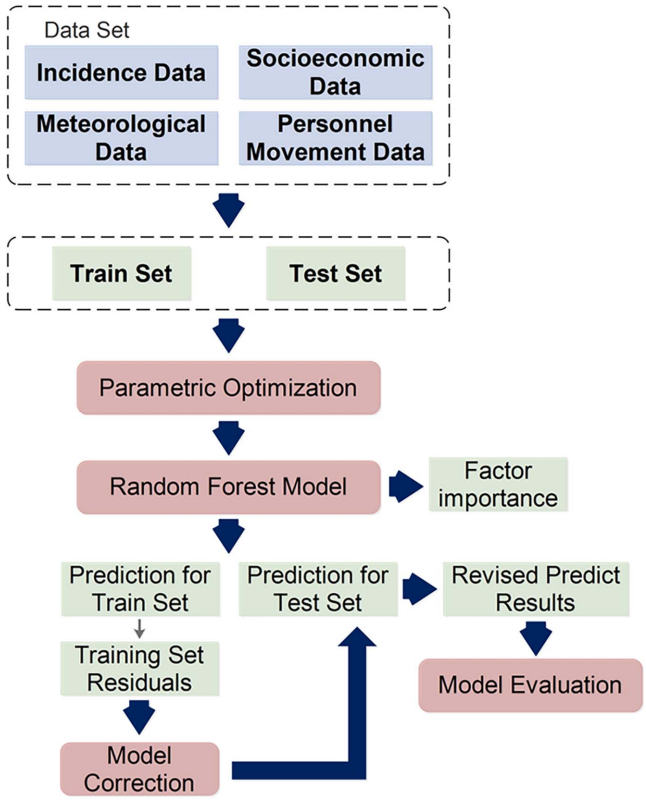


Fig. 1. Study framework.

III. Materials and Methods

Fig. 1 shows the research framework. The following is an explanation of the framework.

A. Data for Study

This study used the daily incidence of COVID-19 in cities across China as the response variable. These data were obtained from the websites of the health commissions of provinces in China. See Fig. R1 (see the supplementary information) for time distribution of the total number of cases in each province. Due to the different methods of data disclosure in each province, to standardize the counting, only confirmed cases were counted. Cities in China mainly experienced concentrated outbreaks and

public harm, rather than new cases appearing every day. Consequently, this study focuses on the daily incidence of concentrated outbreaks. In this study, a concentrated outbreak is defined as an event that lasts more than 3 days and has a maximum daily case count of more than five. If no new COVID-19 cases were reported for three consecutive days, the incident was considered to be over. The current study compiled all concentrated outbreak events in China from the beginning of January 2020 through the end of May 2022. After eliminating missing data on the response and predictor variables, we retained a total of 4991 records, involving 179 cities.

The influence of both natural and social factors on the COVID-19 spread was evaluated (Table I). Daily maximum temperature (TEM_Max; unit: °C), daily average temperature (TEM_Avg; unit: °C), daily minimum temperature (TEM_Min; unit: °C), daily average relative humidity (RH_Avg; unit: %), average daily precipitation (PRE_Day; unit: mm), daily average wind speed (WIN_Avg; unit: m/s), daily average pressure (PRS_Avg, unit: hPa) and daily sunshine hours (SH, unit: h) were obtained from the Chinese Academy of Sciences Resources and environmental science data platform (URL: https://www.resdc.cn/data.aspx?-DATAID=230). Daily absolute humidity (AH, unit: g/m^3) was calculated from TEM_Avg to RH_Avg by the following formula [34]:

$$\mathrm{AH} = 2.1674 \times \mathrm{RH\_Avg} \times \frac{6.112 \times e^{\left(\frac{17.67 \times \mathrm{TEM\_Avg}}{243.5 + \mathrm{TEM\_Avg}}\right)}}{273.15 + \mathrm{TEM\_Avg}}. \quad (1)$$

Details of meteorological stations are shown in Table R1.

NO_2 concentration data was collected from the China National Environmental Monitoring Center (http://www.cnemc.cn/sssj/). Correspondingly, the average NO_2 drop value (AD_NO_2; unit: μg/m^3), which is the difference between the average NO_2 concentration value of the week before the event and the NO_2 concentration value during the event, was applied here. Motor vehicle emissions are considered to be the main source of pollutants such as NO_2 in the air, so this indicator is used to characterize the intensity of urban blockade and control [35]. Urban district population (Pop; unit: 10 000 persons), per capita gross regional product (GRP_PC; unit: yuan), the total annual number of passengers transported by buses and trolley buses (PTB; unit: 10 000 person-times), population density (PD; unit: person/square kilometer) and road surface area per capita (RSA_PC; unit: m^2) were collected from "China City Statistical Yearbook" (2021 and 2022) and "China Urban Construction Statistical Yearbook" (2020, 2021, and 2022). If the data for the current year are missing, the adjacent year was used as a substitute. The proportion of expenditure for science and technology and expenditure for education in local general public budget expenditure (PESEG; unit: %) was calculated by the following formula:

$$\mathrm{PESEG} = \frac{\mathrm{EST} + \mathrm{EE}}{\mathrm{GPBE}} \times 100\% \quad (2)$$

where EST is the expenditure for science and technology, EE is the expenditure for education and GPBE is the local general public budget expenditure. The Urbanization rate (UR; unit: %) was calculated by the following formula:

$$\mathrm{UR} = \frac{\mathrm{UP} + \mathrm{UTP}}{\mathrm{UDP} + \mathrm{UDTP}} \times 100\% \quad (3)$$

where UP refers to the urban population, UTP refers to urban temporary population, UDP refers to the urban district population, UDTP refers to urban district temporary population. The EST, EE, and GPBE were taken from "China City Statistical Yearbook" (2021 and 2022), and the UP, UTP, UDP, and UDTP were taken from "China Urban Construction Statistical Yearbook" (2020, 2021, and 2022). The proportion of the population over 65 years old (PO65; unit: %) is provided by the seventh population census. There were also indicators reflecting population mobility. The average immigration index for the preceding week (AII_P) and the daily average immigration index for the previous day (AII) during the outbreak are provided by the Baidu migration data platform (https://qianxi.baidu.com/#/). Here, we consider the different infection intensities of different strains and divide the outbreaks into three categories according to the time of appearance of the main strains in China (Stage, see Table I). The two distinguishing time points are the end of

TABLE I
PREDICTOR FACTORS

Name	Factor	Unit
TEM_Max	Daily maximum temperature	°C
TEM_Avg	Daily average temperature	°C
TEM_Min	Daily minimum temperature	°C
RH_Avg	Daily average relative humidity	%
PRE_Day	Average daily precipitation	mm
WIN_Avg	Daily average wind speed	m/s
PRS_Avg	Daily average pressure	hPa
SH	Daily sunshine hours	h
AH	Daily absolute humidity	g/m^3
AD_NO_2	Average NO_2 drop value	μg/m^3
Pop	Population	10 000 persons
PO65	Proportion of the population over 65 years old	%
GRP_PC	Per capita gross regional product	yuan
PTB	Total annual volume of passengers transported by buses and trolley buses	10 000 person-times
PD	Population density	Person/square kilometer
PRA_PC	Road surface area per capita	m^2
PESEG	Proportion of expenditure for science and technology and expenditure for education in local general public budget expenditure	%
UR	Urbanization rate	%
AII_P	Average immigration index for the previous week	
AII	Average immigration index for the previous day	
Stage	The stage of the outbreak	
Days_n	The day the disease occurred in this event	

May 2021 (the first case of Delta strain in China was discovered in Guangzhou) [36] and December 2021 (the first case of Omicron strain in China was discovered in Tianjin) [37]. In addition, we also include the day the disease occurred in this event (Days_n) as a predictor variable. See Fig. R2 for a display of some data in the dataset.

B. RF Module

In this study, the factors in Table I were used as predictor variables, the daily incidence scale was used as the response variable, and the training set and testing set were divided in a ratio of 75:25 to construct the RF model. RF is a machine learning algorithm based on decision tree ensemble [38], which is widely used in regression and classification tasks. The RF model has advantages over other supervised learning models. It avoids overfitting samples and provides an average-based solution. It also has good tolerance for outliers and noise [39] and is suitable for nonlinear and high-dimensional complex regression problems [40].

C. KDE-QSR Residual Correction Model

Fig. 2 shows the process of KDE-QSR residual correction. It is assumed that the prediction error of the testing set and the prediction residual of the training set conform to the same distribution characteristics. Considering that the RF model may not have completely learned the information of the predictor variables, the additional information contained in the residuals is used to revise the prediction results and improve the model's prediction ability. Here, we establish QSR for the predictor variables and the training set residuals and apply the established model to the residual correction of the test set. Quantile regression does not require assumptions about the data distribution, and is more robust to outliers and extreme values than the classic linear regression model [41], [42]. Quantile regression aims to estimate the specific quantile of the dependent variable under given conditions, and its objective function is the weighted absolute deviation (Eτ) [43], [44], [45]

$$E_\tau = \sum_{i=1}^{n} \rho_\tau(y_i - \hat{y}_i) \tag{4}$$

where y_i and $\hat{y}_i$ are the observed and predicted values. ρ_τ is known as the pinball loss function [46], and is given by

$$\rho_\tau(u) = \begin{cases} \tau u & u \geq 0 \\ (\tau - 1)u, & u < 0 \end{cases} \tag{5}$$

among them, τ is the target quantile, and u is the difference between the observed value and the predicted value of the dependent variable.

We use the Akaike information criterion (AIC) criterion and the stepwise regression method to find a suitable response variable to build a model. Using the stepwise regression method, the model is built by gradually adding or removing predictor variables to achieve the goal of minimizing information loss. Among them, AIC is a commonly used model selection criterion used to evaluate the goodness of fit of a statistical model on a given data set, taking into account the complexity and goodness of fit of the model [47].

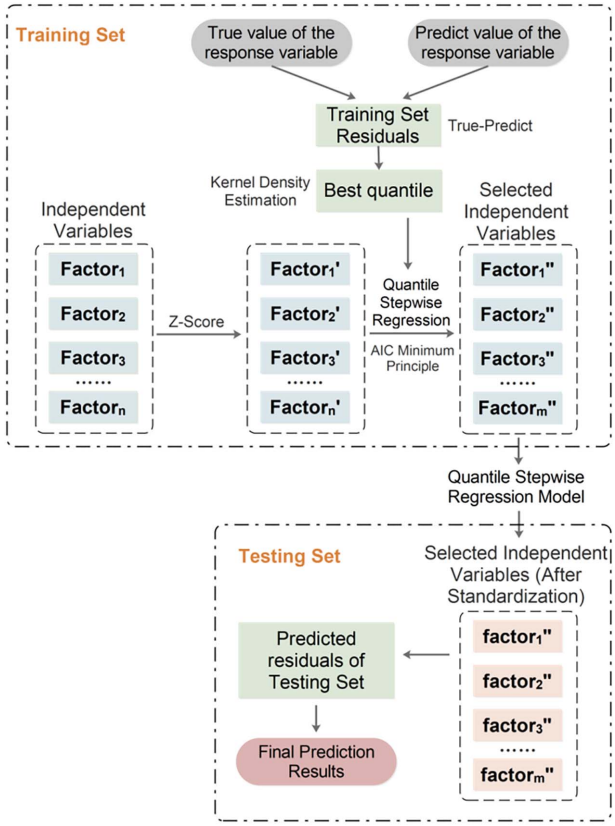


Fig. 2. Residual correction process of KDE-SQR model.

The calculation formula of AIC is as follows:

$$\mathrm{AIC} = 2k - 2\log(L) \tag{6}$$

where L is the likelihood function value of the model on a given data set, and k is the number of parameters in the model (including the intercept term). The goal of AIC is to minimize its value, so the model with a smaller AIC value is considered better.

In stepwise regression, AIC is used to evaluate the goodness of fit of the model after adding or removing variables at each step, so as to select the best model. Specifically, stepwise regression selects the final model by continuously adding or removing variables and comparing the AIC values of different models.

The KDE method is used to estimate the optimal quantile of the QSR model. KDE is a nonparametric statistical method used to estimate the probability density function of a continuous random variable [48]. Given a set of sample data $(x_1, x_2, \ldots, x_n)$. The density estimate at point x is given by the following formula:

$$\mathrm{PDF} = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{7}$$

where n is the sample size and h is the bandwidth (smoothing parameter), which determines the degree of smoothing. The larger (smaller) the bandwidth, the higher (lower) the degree of smoothing. K is the kernel function, which is used to calculate the contribution of each data point. Here, a Gaussian kernel is used. The KDE of different quantiles is calculated, and the quantile with the largest KDE is the "optimal quantile."

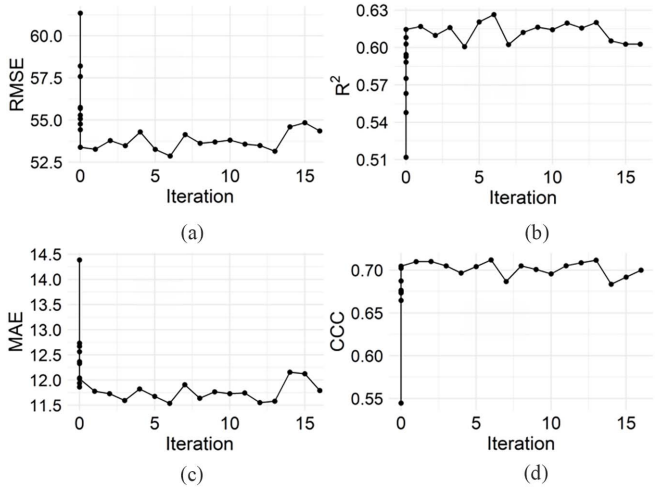


Fig. 3. Changes in the evaluation index values of the RF model hyperparameter optimization with the number of iterations. (a) RMSE. (b) R^2. (c) MAE. (d) CCC.

In general, the residual correction process of the KDE-QSR model is as follows:

1) Calculate the residual of the training set;
2) Use the KDE method to obtain the "optimal quantile";
3) Perform Z-score standardization on all predictive variables in the training set;
4) Use the QSR method to screen suitable predictive variables and establish a model;
5) Perform Z-score standardization on all predictive variables in the test set;
6) Use the model obtained from (4) to predict the residual of the test set;
7) Add the residual obtained from (6) to the prediction result of the test set to obtain the corrected prediction result.

D. Model Evaluation

We choose four indicators to evaluate the accuracy of model prediction. They are root mean square error (RMSE), mean absolute error (MAE), R-squared (R^2), and concordance correlation coefficient (CCC).

IV. Results and Analysis

A. Prediction Effect of RF Model

Using the indicators in Table I as predictive variables, the RF model was constructed to predict the daily incidence scale. The Bayesian algorithm was used to optimize the hyperparameters of the model. The sample selection adopted repeated sampling with replacement, so the model hyperparameter "bootstrap" value was "True." The values of hyperparameters such as the maximum number of features (max_features), n_estimators number of decision trees (n_estimators), and minimum number of samples to be split (min_samples_split) were iteratively optimized. The changes in the values of RMSE, R^2, MAE, and CCC with the increase in the number of iterations are shown in Fig. 3. Under different hyperparameter combinations, the RMSE ranged from 52.86 to 61.34, the R^2 ranged from 0.51 to 0.63, the MAE ranged from 11.54 to 14.39, and the CCC ranged from 0.54 to 0.71. In this study, 10 random samplings were performed before the Bayesian optimization started to initially explore the parameter space, and 16 iterations were finally performed to optimize the hyperparameters. The optimal parameters were obtained at the 6th iteration. The optimal parameter selection under each metric is shown in Table II. We choose the optimal hyperparameter combination: [max_features: 8, n_estimators: 424, min_samples_splits: 4].

TABLE II
Parametric Optimization of RF With Bayesian Optimization Method

Parameters	Researcher Range	Optimization Value Based on RMSE	Optimization Value Based on R^2	Optimization Value Based on MAE	Optimization Value Based on CCC
max_ features	[2,22]	8	8	8	8
n_ estimators	[100, 600]	424	424	424	424
min_ samples_ split	[2,30]	4	4	4	4
Bootstrap	True				

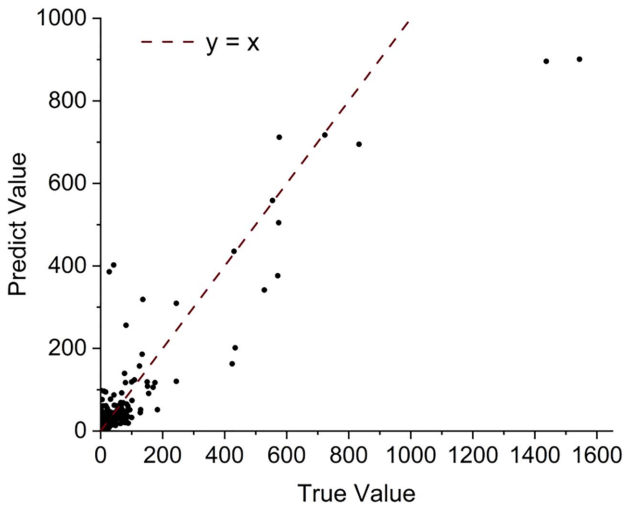


Fig. 4. Prediction results of RF model.

TABLE III
Evaluation Index Values of RF Module Prediction Effect

RMSE	MAE	R^2	CCC
34.45	9.27	0.82	0.89

Fig. 4 and Table III show the prediction effect of the RF model. The purpose of this model is to quickly estimate the approximate scale of the disease at the beginning of an infectious disease outbreak, rather than to make an accurate prediction of the number of cases. From this perspective, the RF model can effectively achieve the goal of qualitatively predicting the scale of the disease. The R^2 of the model prediction results reached 0.82, and the CCC reached 0.89, indicating that

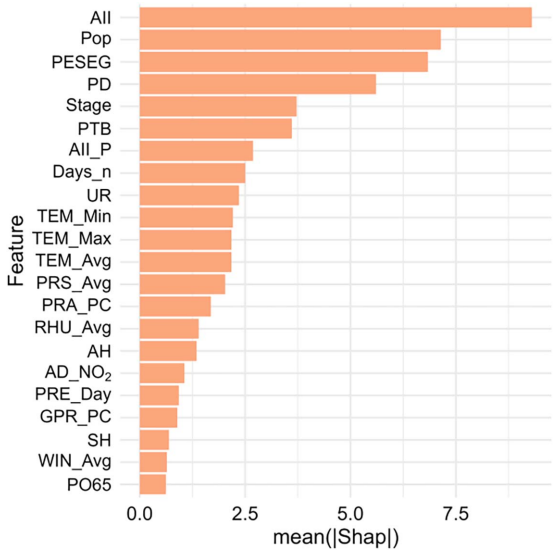


Fig. 5. Feature importance sorting with SHAP method.

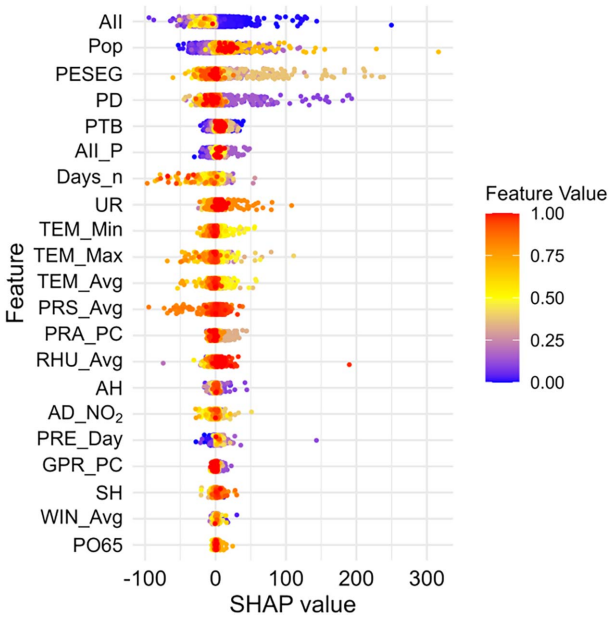


Fig. 6. SHAP summary plot for the RF model ("SHAP value" represents the contribution of each feature to the prediction result of a specific sample. "Feature value" represents the value of the sample on this feature which has been normalized).

the true value and the predicted value of the model have a good correlation and a high goodness of fit. The model has a good prediction effect for smaller-scale outbreaks (less than 1000). However, for outbreaks greater than 1000, the estimated value is low, which increases the RMSE.

B. Feature Importance

The importance ranking of predictor variables is obtained using the SHapley Additive exPlanations (SHAP) method. AII was the most significant feature, followed by Pop, followed by PESEG and PD. The importance of these four indicators is much higher than other indicators (Fig. 5). This shows that the

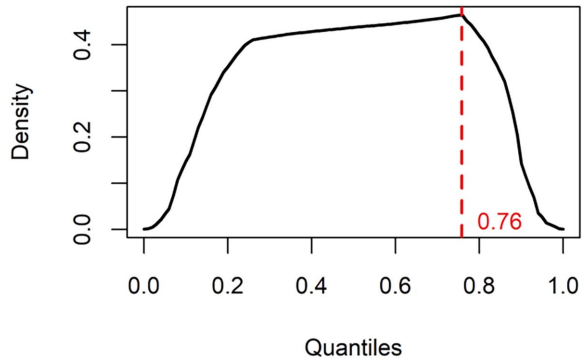


Fig. 7. Probability density distribution of each quantile of the residual of the training set.

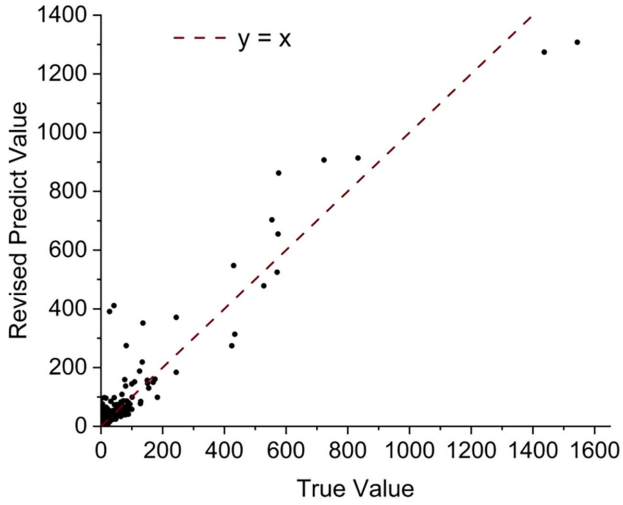


Fig. 8. Prediction effect after residual correction by KDE-QSR.

TABLE IV
EVALUATION INDEX VALUES OF RF-KDE-QSR MODULE PREDICTION EFFECT

RMSE	MAE	R^2	CCC
25.37	7.54	0.90	0.95

population mobility between cities, city size, and investment in welfare expenditures such as science and education have greatly affected the scale of the outbreak. Among meteorological elements, temperature is the most important variable.

Fig. 6 shows the impact of the values of continuous predictor variables on the response variable. For AII, when it is a low value, it corresponds to a positive change in the scale of the outbreak. This is because China has taken the most stringent control measures since the outbreak of COVID-19 [6], [28]. Since the beginning of the event, the influx of people into the city is strictly controlled. So higher AII tends to appear at the beginning and end of the event, and the mid-term urban migration is at a low value. In an infectious disease outbreak, according to the law of infectious disease transmission, the high value of daily incidence occurs in the middle of the event. A higher PESEG corresponds to a smaller scale of outbreak. Higher

TABLE V
COMPARISON OF EVALUATION INDEX VALUES OF RF AND RF-KDE-QSR MODELS PREDICTION EFFECT IN FIVE EXPERIMENTS

Index	RMSE		MAE		R^2		CCC	
Model	RF	RF-KDE-QRS	RF	RF-KDE-QRS	RF	RF-KDE-QRS	RF	RF-KDE-QRS
No.1	33.61	26.36	9.22	7.61	0.80	0.88	0.88	0.94
No.2	36.07	30.39	9.86	8.46	0.88	0.92	0.81	0.87
No.3	40.73	35.71	10.23	8.72	0.72	0.79	0.83	0.89
No.4	30.34	28.12	9.38	7.49	0.74	0.78	0.86	0.91
No.5	43.99	40.84	10.93	9.83	0.75	0.79	0.84	0.87

levels of social development are conducive to reducing the incidence of COVID-19 [49]. A high population size (high Pop value) has a positive impact on the scale of the outbreak, but the impact of urban population density is the opposite. In large cities, people have more opportunities to meet each other, which increases the probability of infection [50]. However, the density of people in cities may not be a risk factor for the spread of infectious diseases if cities take good prevention and control measures [51]. A moderate TEM_Avg corresponds to a positive change in the scale of the outbreak [8].

C. Results After Residual Correction of KDE-QSR Model

The KDE-QSR method is used to correct the residuals of the RF model. First, the probability density of each quantile of the training set is calculated, and the peak of the probability density distribution is found. The peak is taken at the 0.76 quantile (Fig. 7). Therefore, the 0.76 quantile is taken for quantile stepwise regression.

After residual correction, the prediction accuracy of the testing set was significantly improved (Fig. 8, Table IV). The residual correction was most effective for large-scale samples with an incidence of more than 1000 cases (Fig. 8). Overall, R^2 and CCC increased by 0.08 and 0.06, respectively. RMSE and MAE decreased by 26.36% and 18.66%, respectively, compared with before residual correction (Table IV).

The training set and testing set were randomly divided into 75:25 ratios for testing, and the test was repeated five times. The prediction results of the RF model and the RF-KDE-QSR model were recorded respectively (Fig. R3). The prediction accuracy index values showed that the RF-KDE-QSR model had an average R^2 increase of 0.05 (0.04–0.08) and an average CCC increase of 0.05 (0.03–0.06) compared with the RF model in five predictions. The RMSE decreased by 12.82% (7.16–21.57%) and the CCC decreased by 15.33% (10.06–20.15%) on average (Table V). The prediction effect was significantly improved.

V. CONCLUSION

We constructed an RF-KDE-QSR model to make a preliminary estimate of the scale of the COVID-19 outbreak and correct the estimated results through meteorological indicators, socioeconomic indicators, and population mobility indicators. This model can be used to predict the normalization of COVID-19 prevention and control and other infectious disease outbreaks. Its advantage is that it does not rely on disease information. In the early stage of the disease, when there is a lack of disease-related data, it can roughly estimate the scale of the disease using only easily available social and natural data, which can buy time for timely response to infectious diseases and provide an effective reference. The model is effective in estimating the scale of the disease (in the above six tests, R^2 is greater than 0.70 before correction and greater than 0.75 after correction). In addition, the correction of the KDE-QSR model effectively improves the prediction accuracy of the RF model.

VI. DISCUSSION

We selected city-specific indicators for prediction, considering factors like city size and development level, making our approach more refined than previous studies. The research method of this work can also be extended to the spread prediction of other infectious diseases affected by seasonal climate change and population mobility, such as influenza [52], especially for countries and regions with insufficient medical data. In the process of model promotion, it is necessary to carefully verify and adjust according to these disease characteristics.

There is still scope for further improvement in our model. Factors such as changes in drug prevention and treatment methods [53] should also be considered for their impact on the incidence. In addition, there are strong correlations between prediction variables like TEM_Avg, TEM_Max, TEM_Min, and AH due to the feedback between water vapor and temperature [54], [55]. To avoid excluding important factors, we retained all indicators, as the RF model handles collinearity well but increases computational cost. Future work should explore these relationships between variables to reduce dimensionality and simplify the model. Additionally, our dataset has limitations, with a small sample size for large-scale outbreaks during the study period, leading to less accurate RF model predictions for large outbreaks (though it can estimate their general scale). Further expansion and enrichment of the dataset are still needed to improve prediction accuracy. Our model primarily provides a preliminary estimate of daily outbreak size, and more precise predictions require detailed case information.

REFERENCES

[1] WHO, C. O. *World Health Organization. Responding to Community Spread of COVID-19*. Reference WHO/COVID-19/community_transmission/2020.1, 2020.

[2] A. Chhabra et al., "Sustainable and intelligent time-series models for epidemic disease forecasting and analysis," *Sustain. Technol. Entrepreneurship*, vol. 3, no. 2, 2024, Art. no. 100064.

[3] Q. Liu, L. Cao, and X.-Q. Zhu, "Major emerging and re-emerging zoonoses in China: A matter of global health and socioeconomic development for 1.3 billion," *Int. J. Infectious Diseases*, vol. 25, pp. 65–72, 2014.

[4] A. Sharifi and A. R. Khavarian-Garmsir, "The COVID-19 pandemic: Impacts on cities and major lessons for urban planning, design, and management," *Sci. Total Environ.*, vol. 749, 2020, Art. no. 142391.

[5] A. Josephson, T. Kilic, and J. D. Michler, "Socioeconomic impacts of COVID-19 in low-income countries," *Nature Human Behaviour*, vol. 5, no. 5, pp. 557–565, 2021.

[6] L. Tan et al., "Assessing the impacts of COVID-19 on the industrial sectors and economy of China," *Risk Anal.*, vol. 42, no. 1, pp. 21–39, 2022.

[7] Z. J. Jia, S. Y. Wen, and B. Q. Lin, "The effects and reacts of COVID-19 pandemic and international oil price on energy, economy, and environment in China," *Appl. Energy*, vol. 302, pp. 117612, 2021.

[8] Y. R. Wu and C. R. Huang, "Climate change and vector-borne diseases in China: A review of evidence and implications for risk management," *Biology*, vol. 11, no. 3, pp. 370, 2022.

[9] J. T. Liu et al., "Impact of meteorological factors on the COVID-19 transmission: A multi-city study in China," *Sci. Total Environ.*, vol. 726, 2020, Art. no. 138513.

[10] S. Xiao et al., "Meteorological conditions are heterogeneous factors for COVID-19 risk in China," *Environ. Res.*, vol. 198, 2021, Art. no. 111182.

[11] S. Sannigrahi et al., "Examining the association between socio-demographic composition and COVID-19 fatalities in the European region using spatial regression approach," *Sustain. Cities Soc.*, vol. 62, 2020, Art. no. 102418.

[12] J. B. Dowd et al., "Demographic science aids in understanding the spread and fatality rates of COVID-19," *Proc. Nat. Acad. Sci.*, vol. 117, no. 18, pp. 9696–9698, 2020.

[13] Z. X. Xie et al., "Spatial and temporal differentiation of COVID-19 epidemic spread in mainland China and its influencing factors," *Sci. Total Environ.*, vol. 744, pp. 140929, 2020.

[14] C. W. Liu et al., "The impact of crowd gatherings on the spread of COVID-19," *Environ. Res.*, pp. 113604, 2022.

[15] B. Lim and S. Zohren, "Time-series forecasting with deep learning: A survey," *Philos. Trans. R. Soc. A*, vol. 379, no. 2194, 2021, Art. no. 20200209.

[16] Q. Zhou et al., "Interrupted time series analysis using the ARIMA model of the impact of COVID-19 on the incidence rate of notifiable communicable diseases in China," *BMC Infectious Diseases*, vol. 23, no. 1, pp. 375, 2023.

[17] R. Yaesoubi and T. Cohen, "Generalized Markov models of infectious disease spread: A novel framework for developing dynamic health policies," *Eur. J. Oper. Res.*, vol. 215, no. 3, pp. 679–687, 2011.

[18] Y. Wang et al., "Prediction and analysis of COVID-19 daily new cases and cumulative cases: time series forecasting and machine learning models," *BMC Infectious Diseases*, vol. 22, no. 1, pp. 495, 2022.

[19] M. Erraguntla, J. Zapletal, and M. Lawley, "Framework for infectious disease analysis: A comprehensive and integrative multi-modeling approach to disease prediction and management," *Health Inform. J.*, vol. 25, no. 4, pp. 1170–1187, 2019.

[20] J. Huang et al., "An overview for monitoring and prediction of pathogenic microorganisms in the atmosphere," *Fundam. Res.*, vol. 4, no. 3, pp. 430–441, 2024, doi: 10.1016/j.fmre.2023.05.022.

[21] J. Huang et al., "Global prediction system for COVID-19 pandemic," *Sci. Bull.*, vol. 65, no. 22, pp. 1884, 2020.

[22] S. He, Y. Peng, and K. Sun, "SEIR modeling of the COVID-19 and its dynamics," *Nonlinear Dyn.*, vol. 101, pp. 1667–1680, 2020.

[23] J. Huang et al., "Development of the second version of global prediction system for epidemiological pandemic," *Fundam. Res.*, vol. 4, no. 3, pp. 516–526, 2024, doi: 10.1016/j.fmre.2023.02.030.

[24] H. Li et al., "Impact of human mobility on the epidemic spread during holidays," *Infectious Disease Modelling*, vol. 8, no. 4, pp. 1108–1116, 2023, doi: 10.1016/j.idm.2023.10.001.

[25] J. Huang et al., "Multi-source dynamic ensemble prediction of infectious disease and application in COVID-19 case," *J. Thoracic Disease*, vol. 15, no. 7, pp. 4040–4052, 2023, doi: 10.21037/jtd-23-234.

[26] K. Paarporn et al., "Networked SIS epidemics with awareness," *IEEE Trans. Computat. Social Syst.*, vol. 4, no. 3, pp. 93–103, 2017.

[27] Z. Zeng, L. Wang, and H. Zhang, "Containment of SARS-CoV-2 delta strain in Guangzhou, China by quarantine and social distancing: A modelling study," *Sci. Rep.*, vol. 12, no. 1, 2022, Art. no. 21096.

[28] Z. Yang et al., "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions," *J. Thoracic Disease*, vol. 12, no. 3, pp. 165–174, 2020.

[29] H. B. Syeda et al., "Role of machine learning techniques to tackle the COVID-19 crisis: Systematic review," *JMIR Medical Inform.*, vol. 9, no. 1, 2021, Art. no. e23811.

[30] H. R. Niazkar and M. Niazkar, "Application of artificial neural networks to predict the COVID-19 outbreak," *Global Health Res. Policy*, vol. 5, pp. 1–11, 2020.

[31] J. Luo et al., "Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms," *Results Phys.*, vol. 27, 2021, Art. no. 104462.

[32] A. Y. Yeung, F. Roewer-Despres, L. Rosella, and F. Rudzicz, "Machine learning–based prediction of growth in confirmed COVID-19 infection cases in 114 countries using metrics of nonpharmaceutical interventions and cultural dimensions: Model development and validation," *J. Medical Internet Res.*, vol. 23, no. 4, 2021, Art. no. e26628.

[33] S. H. Gökler, "Prediction of Covid-19 confirmed cases and deaths using hybrid support vector machine-Taguchi method," *Comput. Ind. Eng.*, vol. 191, 2024, Art. no. 110103.

[34] J. Karmokar et al., "An assessment of meteorological parameters effects on COVID-19 pandemic in Bangladesh using machine learning models," *Environmental Sci. Pollut. Res.*, vol. 29, no. 44, pp. 67103–67114, 2022.

[35] X. Lian et al., "Environmental indicator for COVID-19 non-pharmaceutical interventions," *Geophys. Res. Lett.*, vol. 48, no. 2, 2021, Art. no. e2020GL090344.

[36] L. Luo et al., "Crucial control measures to contain China's first delta variant outbreak," *Natl. Sci. Rev.*, vol. 9, no. 4, 2022, Art. no. nwac004.

[37] Z. Tan et al., "The first two imported cases of SARS-CoV-2 omicron variant—Tianjin municipality, China, December 13, 2021," *China CDC Weekly*, vol. 4, no. 4, pp. 76, 2022.

[38] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[39] P. Kontschieder, S. R. Bulò, H. Bischof, and M. Pelillo, "Structured class-labels in random forests for semantic image labelling," in *IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Barcelona, Spain, Nov. 2011, pp. 2190–2196.

[40] R. Yu, Y. Yang, L. Yang, and G. Han, "RAQ-A random forest approach for predicting air quality in urban sensing systems," *Sensors*, vol. 16, pp. 86, 2016.

[41] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, *Applied Linear Statistical Models*, 4th ed. McGraw-Hill Irwin, 2004.

[42] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, vol. 821, John Wiley & Sons, 2012,

[43] R. Koenker and G. Bassett, Jr., "Regression quantiles," *Econometrica: J. Econometric Soc.*, vol. 46, pp. 33–50, 1978.

[44] R. Koenker and K. F. Hallock, "Quantile regression," *J. Econ. Perspectives*, vol. 15, no. 4, pp. 143–156, 2001.

[45] J. Zietz, E. N. Zietz, and G. S. Sirmans, "Determinants of house prices: A quantile regression approach," *J. Real Estate Finance Econ.*, vol. 37, pp. 317–333, 2008.

[46] L. Zhang et al., "Probability density forecasting of wind speed based on quantile regression and kernel density estimation," *Energies*, vol. 13, no. 22, pp. 6125, 2020.

[47] M. Wang et al., "A comparison of approaches to stepwise regression on variables sensitivities in building simulation and analysis," *Energy Build.*, vol. 127, pp. 313–326, 2016.

[48] W. J. Lee et al., "Monitoring of a machining process using kernel principal component analysis and kernel density estimation," *J. Intell. Manuf.*, vol. 31, pp. 1175–1189, 2020.

[49] D. Chang et al., "The determinants of COVID-19 morbidity and mortality across countries," *Sci. Rep.*, vol. 12, no. 1, pp. 5888, 2022.

[50] S. Angel and A. Blei, "Covid-19 thrives in larger cities, not denser ones," *J. Extreme Events*, vol. 7, no. 4, 2020, Art. no. 2150004.

[51] A. R. Khavarian-Garmsir, A. Sharifi, and N. Moradpour, "Are high-density districts more vulnerable to the COVID-19 pandemic?," *Sustain. Cities Soc.*, vol. 70, 2021, Art. no. 102911.

[52] Q. Liu et al., "Changing rapid weather variability increases influenza epidemic risk in a warming climate," *Environ. Res. Lett.*, vol. 15, no. 4, 2020, Art. no. 044004.
[53] Z. Zeng et al., "The impact of medical resources and oral antiviral drugs on SARS-CoV-2 mortality—Hong Kong SAR, China, 2022," *China CDC Weekly*, vol. 6, no. 21, pp. 469, 2024.
[54] J. Sun et al., "A quasi-linear relationship between planetary outgoing longwave radiation and surface temperature in a radiative-convective-transportive climate model of a gray atmosphere," *Adv. Atmos. Sci.*, vol. 41, no. 1, pp. 8–18, Jan. 2024.
[55] M. Cai et al., "The quasi-linear relation between planetary outgoing longwave radiation and surface temperature: A climate footprint of radiative and nonradiative processes," *J. Atmos. Sci.*, vol. 80, no. 9, pp. 2131–2146, 2023.

Chuwei Liu received the bachelor's degree in atmospheric science from Nanjing University of Information Science and Technology, Nanjing, China, in 2019. She is currently working toward the Ph.D. degree in climatology with the College of Atmospheric Science, Lanzhou University, Lanzhou, China.

Jianping Huang received the Ph.D. degree in weather dynamics from Lanzhou University, Lanzhou, China, in 1988.

Currently, he is a Professor with the College of Atmospheric Science, Lanzhou University and a Director of Collaborative Innovation Center for Western Ecological Safety, Lanzhou University. He has long been dedicating to the study of long-term climate prediction, dust–cloud interaction, and semi-arid climate change by combining field observations and theoretical study. Since the COVID-19 pandemic, he led his team to establish a global prediction system for COVID-19 pandemic by the combination of epidemic model and statistical–dynamic climate prediction methods.

Siyu Chen received the Ph.D. degree in atmospheric physics and atmospheric environment from the College of Atmospheric Science, Lanzhou University, Lanzhou, China, in 2014.

Currently, she is a Professor with the College of Atmospheric Science, Lanzhou University and a Young Changjiang Scholar of the Ministry of Education. She has long been dedicating to the study of air pollution, the interaction between atmospheric environment and climate change, aerosol physical processes and climate change, and health risks. Since the COVID-19 pandemic, she and her team have done many research on environmental and health risks and achieved important results.

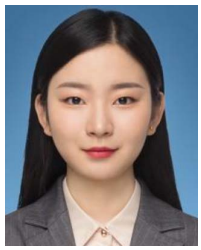
Jiaqi He received the master's degree in engineering in computer technology in 2022 from the School of Information Science and Engineering, Lanzhou University, Lanzhou, China, where he is currently working toward the Ph.D. degree in resource and environmental engineering with the College of Earth and Environmental Science, Lanzhou University.

Shikang Du received the master's degree in engineering in computer technology in 2022 from the School of Information Science and Engineering, Lanzhou University, Lanzhou, China, where he is currently working toward the Ph.D. degree in resource and environmental engineering with the College of Earth and Environmental Science, Lanzhou University.

Nan Yin received the master's degree in software engineering in 2023 from the College of Computer Science & Engineering, Northwest Normal University, Lanzhou, China. She is currently working toward the Ph.D. degree in resource and environmental engineering with the College of Earth and Environmental Science, Lanzhou University, Lanzhou.

Chao Zhang received the Ph.D. degree in engineering mechanics from the College of Civil Engineering and Mechanics, Lanzhou University, Lanzhou, China, in 2022.

Currently, he is a Postdoctoral Researcher with the College of Atmospheric Science, Lanzhou University. His research interests include the mechanism of near-surface dust transport process and virus aerosol transmission.



Danfeng Wang received the master's degree in english translation from the School of Foreign Languages and Literatures, Lanzhou University, Lanzhou, China, in 2020.

Currently, she is a Research Assistant with the Collaborative Innovation Center for Western Ecological Safety, Lanzhou University.