The Journal of High School Research

Prediction of Pollen Concentration in Six Urban Districts of Beijing Using an LSTM-Based Model

Kuang Yuchen*

Submitted: 05 September 2026 **Accepted:** 09 October 2026 **Publication date:** 10 November 2026

DOI: 10.70671/2bn84f97

Abstract: Airborne pollen can cause various allergic symptoms and impact the quality of life of a considerable part of the global population. Therefore, it is crucial to forecast airborne pollen concentrations and disseminate this information to protect public health and improve the quality of life of the overall population. In this study, a data-oriented approach was adopted to develop a daily pollen concentration prediction model using a Long Short-Term Memory (LSTM) neural network. The prediction model is based on the pollen concentration time-series dataset of six urban districts of Beijing (Chaoyang, Haidian, Fengtai, Shijingshan, Changping, and Shunyi) recorded from 2021 to 2025. The pollen concentration dataset shows that the data characteristics of the six districts are different, even though they are all located in Beijing and close to each other. Therefore, different look_back values were set for the six districts in the LSTM-based prediction model to achieve better prediction performance. In the pollen concentration prediction tests for the six districts, the model successfully predicted pollen concentrations in 2025 and achieved R^2 values between 0.4 and 0.7. Moreover, this work also studied the seasonal variation of detailed pollen species concentrations (including nine woody and four herbaceous plants) in Changping District. The daily airborne pollen concentration prediction model can provide a reference for local pollen allergy prevention.

Author keywords: Pollen concentration prediction; LSTM model; Beijing urban districts; seasonal pollen variation; pollen allergy prevention

Introduction

The growth and reproduction of plants are critical for ecosystem stability, and the seasonal dispersion of airborne pollen is an important part of plant reproduction. However, allergies caused by airborne pollen, such as allergic rhinitis and asthma, have become a common health issue and affect human lives globally.[1,2] It is reported that 10–30% of the global population suffers from pollen-induced allergies.[3] Pollen allergies cannot be completely cured, but there are several treatments that can relieve pollen allergy symptoms.[4] Therefore, it is meaningful to provide pollen concentration forecasts to pollen allergy sufferers during the pollen season to help them avoid exposure to high concentrations of pollen. The airborne pollen concentration is mainly influenced by plant species, flowering time, and meteorological factors. Many pollen season environmental variables such as air temperature, sunlight duration, humidity, precipitation, wind vector, and other meteorological factors have a significant influence on pollen concentrations, and many studies have demonstrated that they are important predictors of pollen prediction models.[5–7] Currently, observation-oriented and source-based approaches are the two main methods for pollen concentration prediction model construction.[8] Observation-oriented models usually conduct pollen concentration prediction based on site-specific environmental data, such as the average pollen concentration on the same day in previous years.[9] The commonly used observation-oriented pollen concentration prediction models include regression, time-series analyses, process-based phenological models, and models based on machine learning techniques.[8] Different from the observation-oriented approach, source-based pollen concentration prediction models are usually used to quantify and predict the temporal and spatial distributions of airborne pollen near the surface.[10] These models are based on Chemistry Transport Models and need the source conditions and diffusion calculations for pollen dispersal calculation.[8,11]

In recent years, machine learning methods have been successfully applied in pollen concentration prediction models as a new technique and have achieved excellent model prediction performance.[12,13] Many machine learning techniques have been developed, but not all of them are suitable for pollen concentration prediction models. Among these methods, recurrent neural networks (RNNs) are commonly used for time-series prediction models and have advantages over other neural networks in processing time-series data.[14] Compared with other neural networks, the single-layer networks of RNN have loops, which enable information to be transferred between networks.[15] The structure of an RNN

*Corresponding Author: Kuang Yuchen.
Email: anitakuang2028@126.com
The Experimental High School Attached to Beijing Normal University, No.14, Erlong Road, Xicheng District, Beijing 100032, China

allows it to learn information from the previous data and combine the related information with the current input to make decisions. These features make RNNs very suitable for tasks that require learning from previous data, such as time-series analysis. However, the relevant information may appear further in the past when the time gap becomes longer, and RNNs may not effectively connect this information under such conditions.[16] Long short-term memory (LSTM) networks were developed to overcome this issue. An LSTM network is a variation of RNN and was first proposed in 1997.[17] A typical LSTM architecture consists of a memory cell and four functional layers, namely, the input, forget gate, update, and output layers, which jointly regulate information flow between networks.[15] Compared with conventional RNNs, the LSTM architecture is capable of learning long-term dependencies and effectively addressing complex time-series problems. Several studies have reported the application of LSTMs to pollen concentration prediction and have achieved satisfactory performance.[18,19]

In this study, an LSTM network-based pollen concentration prediction model was developed to predict the pollen concentration in the urban area of Beijing. The pollen concentration time-series data from 2021 to 2025 of six urban districts of Beijing, namely, Chaoyang, Haidian, Fengtai, Shijingshan, Changping, and Shunyi, were used in this research to develop the prediction model. These data show that the pollen concentration distribution in the urban area of Beijing has obvious seasonality, with obvious peaks from March to May and from August to September each year. The results also showed that the intensity and duration of the pollen concentration peaks varied across the six districts. In addition, the pollen concentration data for different woody and herbaceous plants in the urban area of Beijing (Changping District) were collected from 2021 to 2025. Similar to the total pollen concentration, the distribution of pollen concentration of different plants also has seasonality. Specifically, woody plants contributed the majority of total pollen concentrations in spring (March–June), while herbaceous plants contributed most of the pollen in autumn. Different from published works.[5,19,20] this study only used the time series of pollen concentrations of six districts and did not use meteorological factors as predictors in the development of the prediction model, since it is very difficult to obtain the detailed meteorological factor data for each district. Some studies have shown that an LSTM network can still achieve good prediction results using only pollen concentration time-series data.[16,18] Although adding meteorological factors as predictors in the prediction model can improve forecast performance, access to these data may be limited in some cases. To evaluate the predictive performance of the model, the data from 2021 to 2024 were used as the training dataset, and the latest data, from 2025, were used as the test dataset. The results show that the model demonstrated moderate predictive capability across the six districts. Currently, most of the studies focus on forecasting the overall pollen concentration of Beijing. However, pollen concentrations can exhibit considerable differences across regions on the same day due to environmental and meteorological variations. The district-based pollen concentration forecasting model developed in this research can help people suffering from pollen allergies in different areas of Beijing reduce their exposure to pollen environments and take preventive measures in advance.

Methodology

Study area

The study was conducted in Beijing (39°56′ N, 116°20′ E), which is the capital city of China and covers an administrative area of about 16,411 km^2. Beijing is surrounded by mountains on the west, north, and northeast, with overall terrain high in the northwest and low in the southeast. About 62% of Beijing's area is mountainous, and the remaining 38% consists of plains.[21] The climate of Beijing belongs to the warm temperate, semi-humid, and semi-arid monsoon climate, which gives Beijing four distinct seasons.[22] The average annual temperature in Beijing is about 13°C. The coldest month is January, with an average temperature of around −3°C, and the warmest month is July, with an average temperature of almost 27°C. The average annual rainfall in Beijing is about 570–600 mm and is concentrated in June, July, and August. In the past few decades, Beijing has carried out a series of urban greening projects. As of 2024, the overall forest coverage rate of Beijing was 44.95%, with an urban greening coverage rate of 49.84%. These projects also enriched the plant species in Beijing. Many of these plants have the ability to spread allergenic pollen, which has a significant impact on the lives of residents during the pollen season.[23] Currently, Beijing has 16 administrative districts (Fig. 1) and a population of over 20 million.[22] In this work, the districts with relatively large populations in Beijing (Chaoyang, Haidian, Fengtai, and Changping) and two of their adjacent districts (Shijingshan and Shunyi) were selected as the research areas.



Figure 1. The administrative map of Beijing and the selected research districts[24]

Pollen data collection

In this work, the pollen concentration data were collected from the official WeChat account, since it is very difficult to find detailed, district-level pollen concentration data on the official government websites. Beijing is one of the few cities in China that publishes pollen concentration forecasts. During the pollen season (from March to October each year), the Beijing Meteorological Service publishes the pollen concentration measurement data for the previous day and forecasts for the following day for each district every day through its affiliated official WeChat account (Pollen Monitoring and Forecast). Their daily pollen samples were collected using the Durham pollen sampler based on gravity sedimentation and counted using the unit of grains/1000 mm^2.[25] There are also some private social media accounts that provide pollen forecast services and publish pollen concentrations, but their data may not be as reliable as the government data. Therefore, the daily pollen concentration data of the six districts over the past five years (2021–2025) used in this study were finally collected from the official WeChat account. Among them, the 2021–2024 data were used as the training dataset for the pollen concentration prediction model, and the remaining 2025 data were used as the test dataset to test the model prediction performance. In addition, the daily pollen concentration data of nine major woody plants (Cypress, Poplar, Willow, Elm, Ash, Pine, *Juglans*, *Betula*, and Fagaceae) and four major herbaceous plants (Chenopodiaceae, *Artemisia*, *Humulus*, and Poaceae) that spread pollen in Beijing from 2021 to 2025 were also collected (the data were obtained from the Changping District Observatory).

Model development process

Processing of pollen data before model training

Fig. 2 shows the development process of the LSTM-based pollen concentration prediction model. The first step for model building is to create the database of daily pollen concentrations of six districts from 2021 to 2025. Next, the obtained pollen concentration time-series data will be imported into the LSTM model for model training and evaluation. However, there are some issues in the collected data caused by the format used by the official WeChat account to report pollen concentrations. Therefore, the obtained pollen concentration time-series data were first processed before they were used for model training. The purpose of data processing is to replace some uncertain data values and avoid data discontinuity in datasets to ensure they are suitable for model training.

In 2016, the China Meteorological Administration released the five-level classification (1–5) of the meteorological index of pollen allergy, ranging from highly unlikely to highly likely to trigger pollen allergy symptoms (Table 1).[26] As shown in Table 2, each grade of the meteorological index of pollen allergy corresponds to a pollen concentration grade (1 to 5), which is a pollen concentration range. The grade classification of pollen concentration for woody plants and herbaceous plants is in different numerical ranges due to their differences in pollen production. The official Pollen Monitoring and Forecast WeChat account also divided daily pollen concentrations into five grades based on the same numerical range of this classification. They just replaced the unit of grains/cm with 1000 grains/1000 mm^2 based on their pollen count method. For pollen concentration grades 1 to 4, the official WeChat account will usually publish the corresponding specific pollen concentration value or a value bar that can be used to estimate specific values. However, when the grade reaches 5 (very high), the official WeChat account usually publishes >800 grains/1000 mm^2 instead of a specific value. This substitution is particularly common during the peak pollen concentration periods (such as April to May and September to October each year), resulting in many values being recorded as >800 in the dataset instead of specific values. This release format continued until August 25, 2025, and changed to >300 after that. Therefore, the pollen concentration data collection period for model development was also truncated at this date to ensure data consistency, as part of the data after this date may range from 300 to 800 grains/1000 mm^2 in the new release format. Obviously, these uncertain values that are greater than 800 cannot be used directly for model training and testing. In some time periods, specific values corresponding to >800 will be released. Based on the collected data, these values are distributed in the range of 800 to above 1000 grains/1000 mm^2, but most of these values are in the range of 800–2000 grains/1000 mm^2. Therefore, values reported as >800 were replaced with a specific value during data preprocessing. For convenience, the replacement operation function has been integrated into the model algorithm in the initial part. After many tests, the results show that 1000 is a suitable replacement value. When using 1000 as the replacement value, the model achieved a good fitting performance for both values >800 grains/1000 mm^2 and the low pollen concentrations. When the replacement value was below 1000, the model showed a relatively poor fitting performance for values >800 grains/1000 mm^2. As the replacement increases, the fit to high pollen concentrations continues to improve,

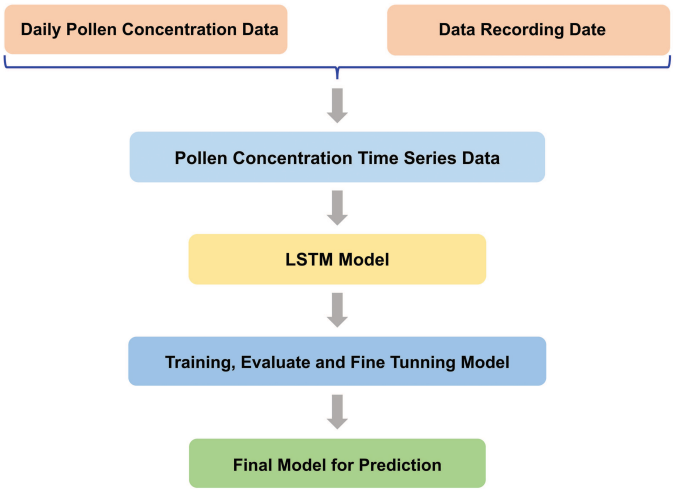


Figure 2. Schematic diagram of the LSTM-based pollen concentration prediction model development process

Table 1. Current meteorological index of pollen allergy classification and grade descriptions released by the China meteorological administration[26]

Meteorological index of pollen allergy grade	Grade description
1	Highly Unlikely to Trigger Pollen Allergy Symptoms
2	Unlikely to Trigger Pollen Allergy Symptoms
3	Can Trigger Pollen Allergy Symptoms
4	Likely to Trigger Pollen Allergy Symptoms
5	Highly Likely to Trigger Pollen Allergy Symptoms

Table 2. Numerical classifications of woody and herbaceous plant pollen concentrations corresponding to the meteorological index of pollen allergy in Table 1 (unit: grains/cm^3)[26]

Pollen concentration grade	Grade description	Pollen concentration value (Woody plants)	Pollen concentration value (Herbaceous plants)
1	Low	≤100	≤50
2	Low to Moderate	101∼250	51∼100
3	Moderate	251∼400	101∼150
4	High	401∼800	151∼300
5	Very High	>800	>300

but the fitting performance for low pollen concentrations begins to deteriorate. In addition, June to July each year is the pollen concentration trough period in Beijing. During these two months, the official WeChat account only releases data for one of the six study districts (six districts are selected in rotation). Therefore, the specific value of pollen concentration reported for the selected district was collected, and the pollen concentration grades for the other five districts were assumed to be the same as that of the selected forecast district. This assumption is based on records from multiple years showing that pollen concentrations in all six districts generally remained at the lowest grade (same grade) during these two months. Afterwards, the median value (as an integer) of the pollen concentration range corresponding to the grade of the selected district was used as the pollen concentration value of the other five districts to fill missing pollen concentration values in June and July. To further validate this assumption, pollen concentration data from other social media platforms in the same periods were also checked to make sure that the assigned values for the other five districts fell within the same pollen concentration grade range as those reported by these platforms. This step was performed simultaneously during the data collection process.

Model training and evaluation

After preparing the pollen concentration time-series data, the datasets were imported into the LSTM model for model training. The gated architecture of the LSTM model enables it to capture long-term dependencies, allowing it to adjust its predictions by learning from historical patterns.[27] Besides, LSTM can also recognize the complex and nonlinear patterns in data that involve long-term dependencies.[28] There are no fixed rules for LSTM model parameter settings, which usually depend on the specific application and data. In this study, the single-variable pollen time series is modeled via a sliding-window supervised-learning setup, using different sequence lengths for each district. For each district, a MinMaxScaler (0–1) is fitted on the training dataset and then applied to transform the test dataset. The model predictor is a two-layer LSTM (hidden units per layer = 50), followed by a dense readout (units = 1) that maps the last time-step output to a one-step-ahead prediction. The model parameters are initialized with PyTorch defaults: LSTM weights from $U\left(-\frac{1}{\sqrt{50}}, \frac{1}{\sqrt{50}}\right)$ with zero biases, and the linear layer uses Kaiming-Uniform weights ($a = \sqrt{5}$) with bias from $U\left(-\frac{1}{\sqrt{50}}, \frac{1}{\sqrt{50}}\right)$. After multiple tests, the number of training epochs and the batch size were set to 20 and 64, respectively. After training, the model will generate one-step predictions on training and test windows (next-day prediction). For the LSTM model, an important parameter is the look_back value, which is the number of days that the model needs to look back when making predictions. The value of the time step look_back needs to be set individually for each district as it will notably affect the model fitting and prediction performance. In the model, the look_back values for each district will first be scanned from 5 to 90 (with a step size of 5) to select the optimal values. The root mean square error (RMSE), mean absolute error (MAE), and R-square (R^2) metrics were used to evaluate and select the best look_back values for each district and other model configurations. Their calculation formulas are as follows[5,27,29]:

$$\mathrm{RMSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n}} \sum_{i=1}^{n} \left(\mathbf{y}_i - \hat{\mathbf{y}}_i\right)^2} \tag{1}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

Here, n is the number of pollen concentration observations, y_i is the actual observed pollen concentration, $\hat{y}_i$ is the pollen concentration value predicted by the model, and $\bar{y}$ is the average of the actual observed pollen concentrations. Generally, RMSE is more sensitive to large differences between predicted and actual values because the error is squared in the formula. MAE can provide a more robust and intuitive measure of the difference between the actual value and the predicted value. R can quantify the proportion of the total variation in the data that is explained by the model.[5] An R^2 value of 0 means that using the average of the observed pollen concentrations in the evaluation dataset as a fixed prediction gives the same mean squared error as the model. A value of 1 indicates perfect agreement between the predicted values and the actual data. Typically, smaller RMSE and MAE values, along with an R^2 value closer to 1, correspond to superior model predictive performance.

Discussions and Recommendations for Future Work

Variation of total pollen concentration in six districts

The pollen season in Beijing usually lasts from early March to mid-October. Fig. 3 presents the daily pollen concentration recorded in Chaoyang District during 2024 (March 1–October 14). It is easy to see that the pollen season mainly includes two peak periods (March–May and August–October) and one trough period (June–July). Specifically, the peak pollen concentrations are recorded from mid-March to mid-May and from mid-August to mid-September. During the peak pollen concentration periods, the pollen concentration could be above 800 grains/1000 mm^2 (replaced by 1000 grains/1000 mm^2) for several consecutive days, resulting in multiple small plateaus in the daily pollen concentration curve. People who are allergic to pollen should take effective protective measures during these two time periods. In addition, the pollen concentration peaks in spring are more intense and longer-lasting than the peaks in autumn, which could be attributed to the differences in flowering plant types during spring and autumn. Pollen concentrations in Chaoyang District are lowest in June and July. Additionally, Beijing usually has abundant rainfall during these two months, which further reduces pollen concentration and suppresses pollen transmission. After mid-October, Beijing will gradually enter winter, and the low temperature will last until March of the next year.

Fig. 4 shows the daily pollen concentrations in Chaoyang (CY), Haidian (HD), Fengtai (FT), Shijingshan (SJS), Changping (CP), and Shunyi (SY) from 2021 to 2025. In Fig. 4, all the data were combined in the order of days (a total of 1075 days), since the dates of pollen concentration data for each year are not continuous with the previous year and the next year. Similar to Chaoyang, the daily pollen concentration distribution in the other five districts also has obvious peak and trough periods. However, the distribution of pollen concentrations during peak periods varies among the six districts due to the differences in their natural environments. Among them, Shijingshan has the most days with high pollen concentration (Grade 5) during the spring and autumn peak periods. Chaoyang and Haidian districts have stable peak pollen concentrations in spring and autumn in 2021–2022. But their autumn pollen peak intensity decreased in 2023 and 2024, especially in Haidian. It is worth noting that the autumn pollen concentration in Fengtai has been low in the past few years, although

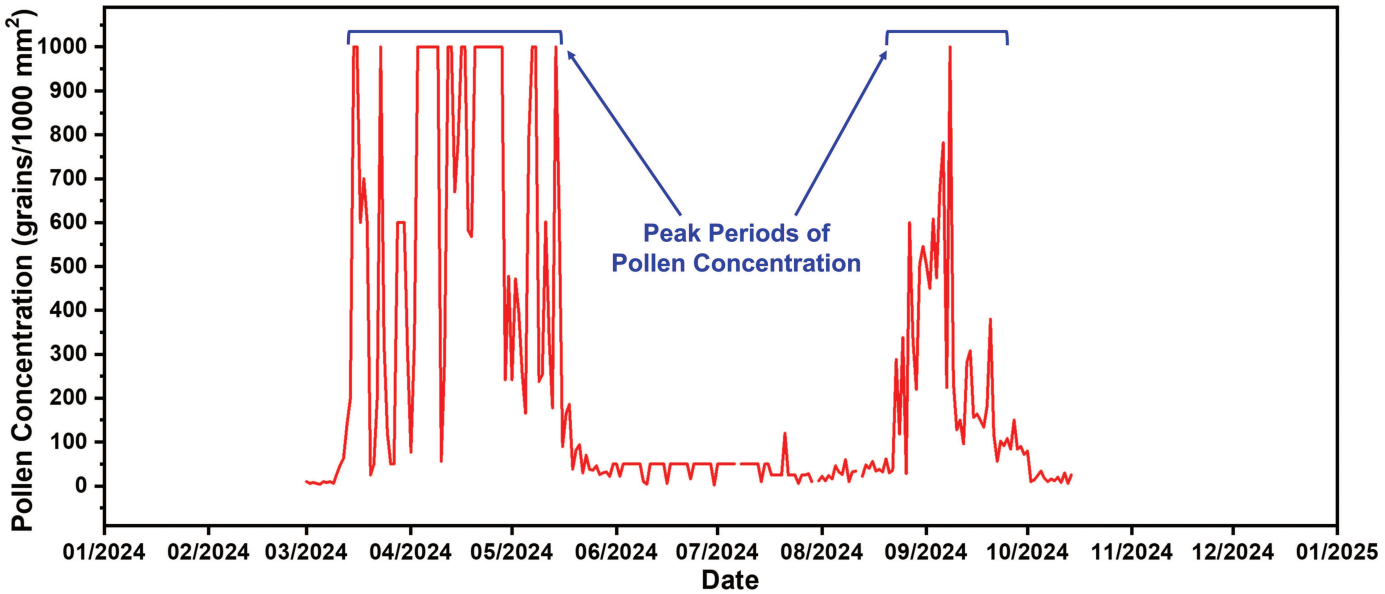


Figure 3. Daily pollen concentration in Chaoyang District in 2024

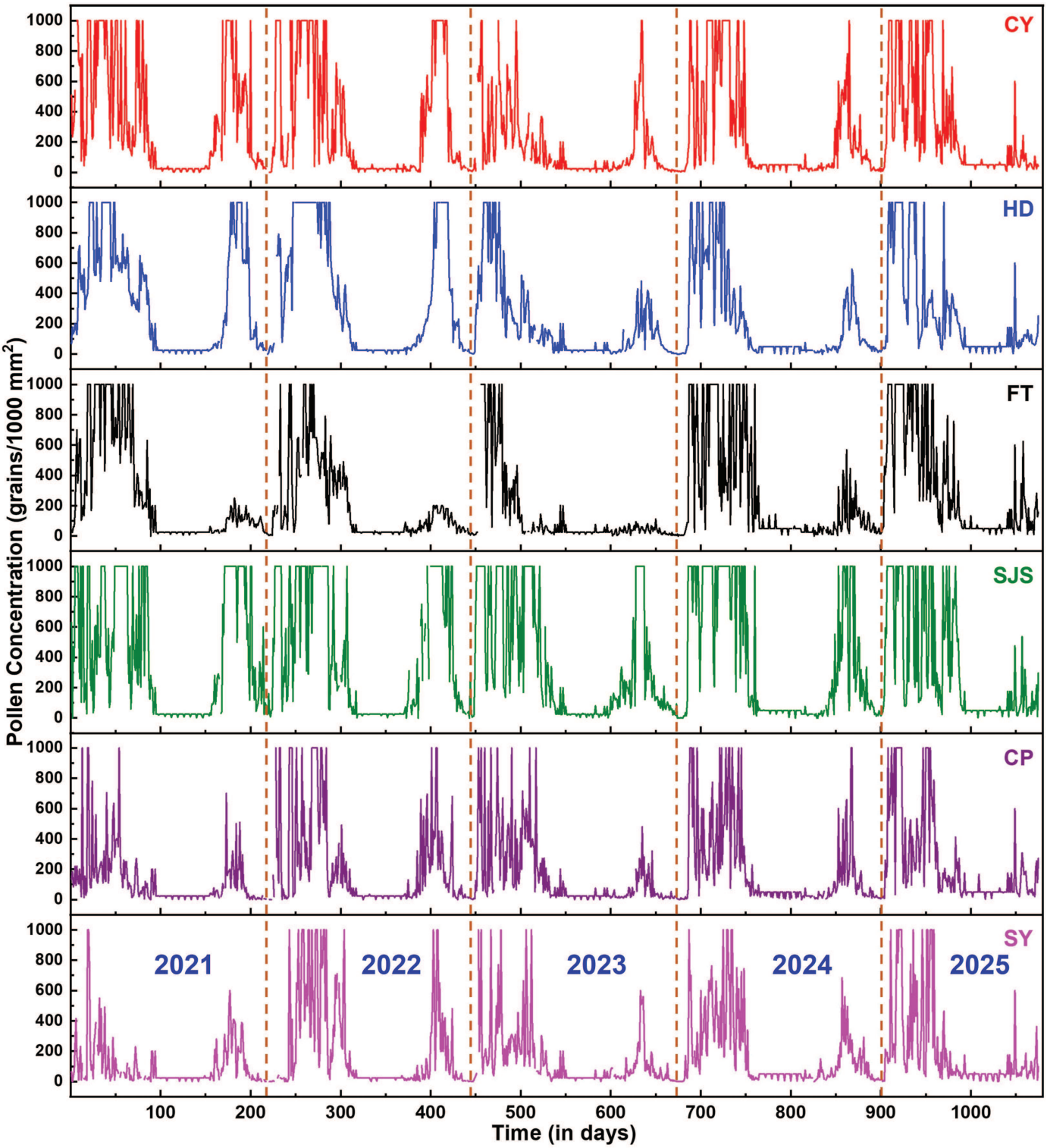


Figure 4. Daily pollen concentrations in six districts from 2021 to 2025

it has a relatively strong spring pollen concentration peak period every year. Changping and Shunyi have relatively lower-intensity spring and autumn pollen peaks among the six districts. A split in the spring pollen peak can also be observed in these two districts, which is very obvious in 2023–2025. In these three years, it is easy to see that the spring peak period has a trend to be split into two peaks spanning from approximately early March to mid-April and late April to mid-May, respectively. These results indicate that pollen concentrations can vary substantially across districts within the same city. Compared with studies using city-average pollen concentration data for prediction, the regional variation in pollen concentration highlights the need for district-level prediction research.

Variation of woody and herbaceous plant pollen concentrations

In Beijing, pollen allergies can be caused by airborne pollen from many plant species. Among these plants, some species of woody plants and herbaceous plants have been identified as the sources of pollen allergies.[23,30] This work summarized the pollen concentrations of nine woody (Cypress, Poplar, Willow, Elm, Ash, Pine, *Juglans*, *Betula*, and Fagaceae) and four herbaceous (Chenopodiaceae, *Artemisia*, *Humulus*, and Poaceae) plant species with allergenic potential in Beijing. Fig. 5a shows the daily pollen concentration distribution of these plants in Changping District during 2024. As shown in Fig. 5a, the flowering period of the nine woody plants is mainly distributed from March to June, and they can produce a lot of airborne pollen during these four months. Besides, the peak pollen concentrations of these plants also vary in time. From early to mid-March, Cypress, Poplar, and Elm are the main pollen contributors. The peak pollen concentration periods of Willow and Ash range from late March to early April, with lower peak concentrations than those of Cypress and Elm. Pine has the highest observed pollen concentrations among the nine woody plants, with the peak season spanning from early April to mid-May, while *Juglans* and Fagaceae have the lowest pollen concentrations among the nine plants. The peak pollen period of *Betula* extends from late April to early May. Pine and *Betula* are also the main contributors to the total pollen concentration peak in late spring. For the four herbaceous plant species, their peak pollen season is concentrated in the period from mid-August to early September. Among them, the pollen concentration of *Artemisia* is much higher than that of the other three plants. Chenopodiaceae and *Humulus* have similar peak pollen concentrations, while Poaceae has the lowest

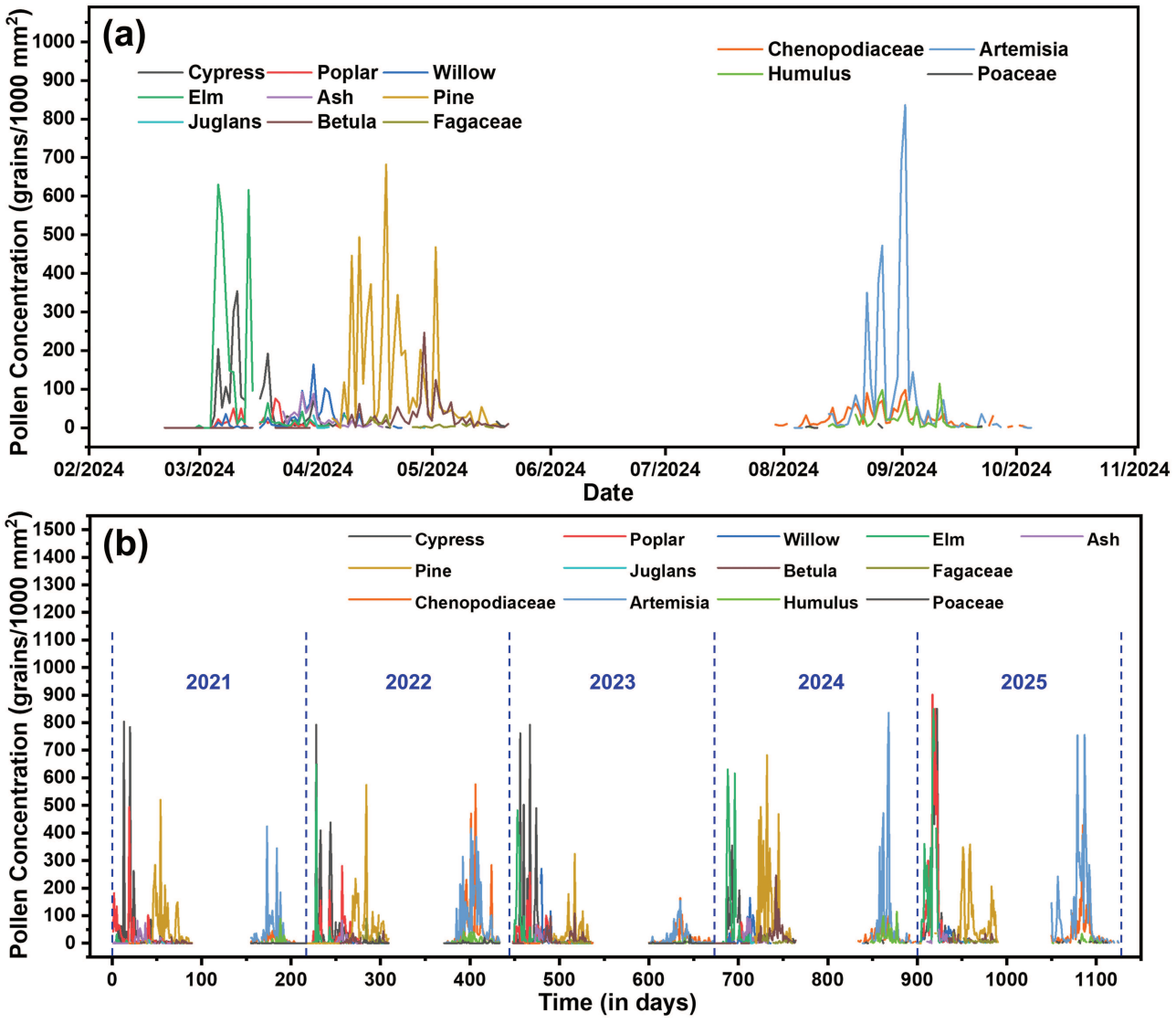


Figure 5. (a) Variation in airborne pollen concentrations for different plants in 2024 (Changping District). These plants include nine woody plant species (Cypress, Poplar, Willow, Elm, Ash, Pine, *Juglans*, *Betula*, and Fagaceae) and four herbaceous plant species (Chenopodiaceae, *Artemisia*, *Humulus*, and Poaceae). (b) Pollen concentrations of nine woody and four herbaceous airborne pollen plants from 2021 to 2025 (Changping District)

pollen concentration among the four herbaceous plants. Compared with woody plants that bloom in spring, the peak pollen concentration period of herbaceous plants is relatively short, which is also consistent with the spring and autumn total pollen concentration distribution in Changping District during 2024 (Fig. 4).

Fig. 5b presents the daily pollen concentrations of nine woody and four herbaceous plant species from 2021 to 2025 (March 5, 2021, to October 14, 2025) in the order of days (a total of 1125 days). It can be seen that the peak period of pollen concentration of woody plants is longer than that of herbaceous plants every year. In spring, the woody plant pollen peak period is divided into two main peaks: one represented by Cypress and Elm from early to mid-March, and the other by Pine from early April to mid-May, which can be used to explain the spring total pollen concentration split observed in Changping District from 2021 to 2025 (Fig. 4). In autumn every year, there is only one pollen concentration peak that is dominated by *Artemisia* and Chenopodiaceae. The plant species distribution across the six study districts in Beijing is similar.[31] Therefore, it can be inferred that the other five districts have similar distributions of woody and herbaceous pollen concentrations. This interpretation is consistent with the similar timing of spring and autumn pollen peaks across the six districts from 2021 to 2025.

Pollen concentration prediction performance of the LSTM-based model

As discussed above, there are some differences in the daily pollen concentrations in 2021–2025 for the six districts, although they are all located in Beijing. Therefore, the optimal look_back values for each district were selected during model development to achieve the best prediction performance. In this study, the final look_back values were selected based on the 2021–2024 training data, and the data for 2025 were kept separate. Table 3 lists the look_back values for each district set in the model and the corresponding RMSE, MAE, and R of the fitted and predicted values. The optimal look_back value usually should be the value that corresponds to the smallest RMSE and MAE values and the largest R value for model fitting. However, the results showed that the best fitting performance often appeared near the optimal value during tests. For the developed model, it is expected that it can fit and predict the peak and trough periods of pollen concentration for each district, especially for the peak period when pollen concentration is higher than 800 grains/1000 mm^2 (replaced by 1000 grains/1000 mm^2 in the datasets). During model training, the ideal performance for the model is to give a fitted value (predicted value for 2021–2024 data) higher than 800 grains/1000 mm^2 during the peak period. Therefore, a fitted value >800 grains/1000 mm^2 given by the model can still be defined as correct from the perspective of pollen concentration grade (Grade 5). In contrast, if the model cannot fit pollen concentration values above 800 grains/1000 mm^2 during model training, it means that the model may perform poorly in predicting high pollen concentration values. In addition, the model fitted values during the trough periods should also be close to the actual pollen concentrations or belong to the same pollen concentration grade range (still can be considered a correct prediction). Based on these considerations and the fitting performance during model training, the look_back values for the six districts were finally set to 45 (CY), 15 (HD), 20 (FT), 65 (SJS), 40 (CP), and 50 (SY), respectively.

Fig. 6 shows the model fitted (2021–2024) and predicted (2025) pollen concentrations together with the actual pollen concentration values from 2021 to 2025 for the six districts. In this study, the pollen concentration values were selected as the output for model fitting and prediction, given that they can vary considerably even within the same pollen allergy grade. The predicted pollen concentrations can also be converted into the corresponding pollen allergy grades using the ranges listed in Table 2. During the training phase, the peaks and troughs of the model fitted values match the actual values well in each year. However, the fitting performance of the model varied across the six districts. In detail, the model fitted pollen concentrations better in Chaoyang, Haidian, Fengtai, and Shijingshan than in Changping and Shunyi. The RMSE and MAE of Changping and Shunyi are not the largest among the six districts. But the R values of the fitted values for these four districts are all above 0.6, which are much higher than the R for Changping (0.42) and Shunyi (0.47) (Table 3). Moreover, the model cannot fit the high pollen concentration values in these two districts well and

Table 3. Model look_back value settings for the six districts and the corresponding RMSE, MAE, and R^2 of the LSTM model fitted and predicted pollen concentration values

District	look_back value (Days)	LSTM model fitted value			LSTM model predicted value		
		RMSE	MAE	R^2	RMSE	MAE	R^2
CY	45	201.44	130.30	0.62	173.80	111.98	0.55
HD	15	140.02	88.52	0.83	186.36	108.07	0.60
FT	20	170.87	95.63	0.69	194.59	132.39	0.64
SJS	65	224.90	148.10	0.69	175.77	109.21	0.49
CP	40	194.60	114.74	0.42	153.74	94.81	0.56
SY	50	170.22	101.39	0.47	154.89	91.61	0.45

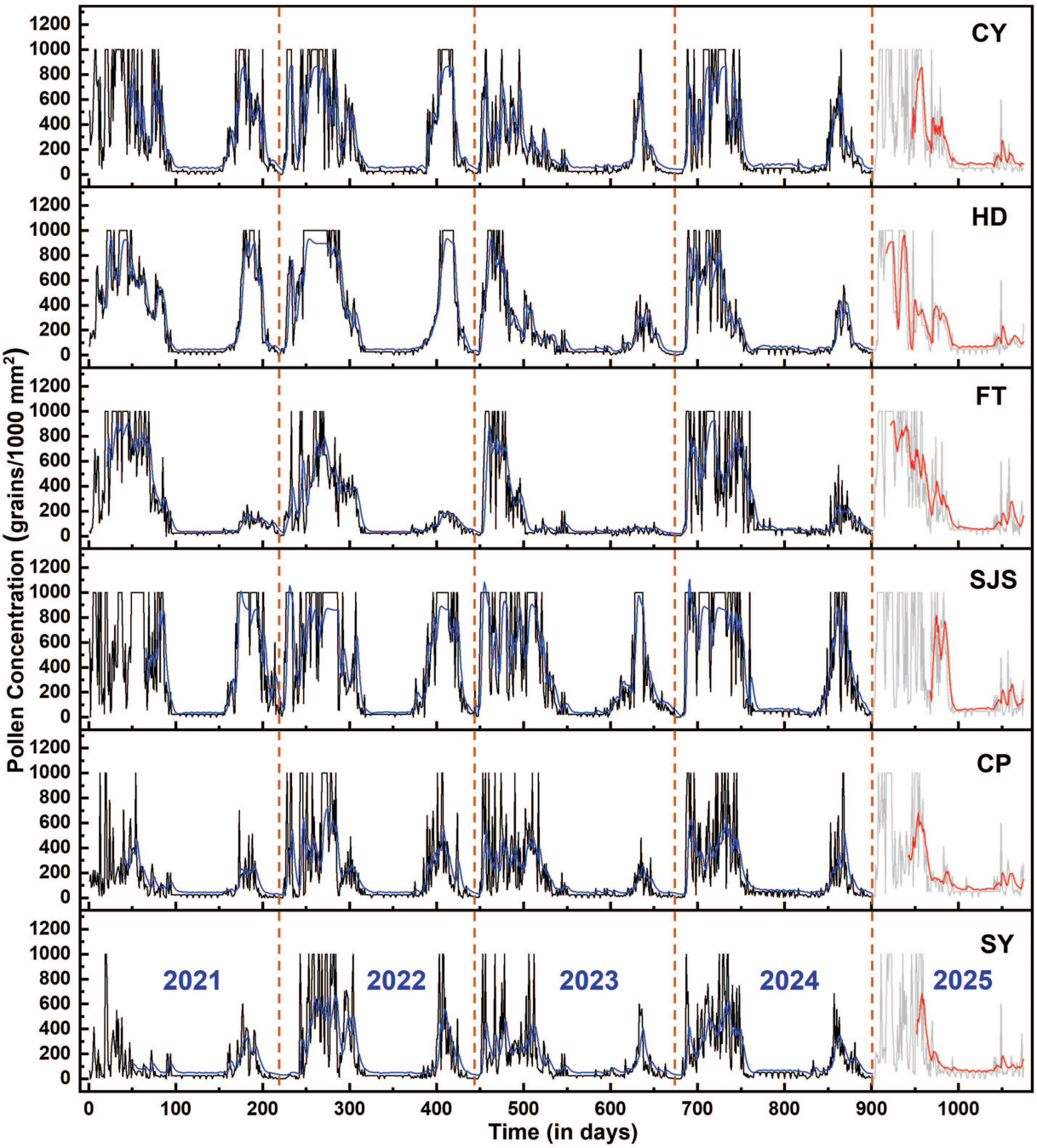


Figure 6. Daily pollen concentrations in six districts from 2021 to 2025 (black and grey lines) and LSTM-based model-fitted (blue line) and predicted (red line) pollen concentration values

cannot give any fitted values higher than 800 grains/1000 mm^2, which shows the impact of pollen concentration data differences in different districts on the model performance. Compared with the other four districts, Changping and Shunyi have relatively few days with pollen concentrations above 800 grains/1000 mm^2, which is not conducive to the model capturing the distribution characteristics of high pollen concentration values.

In the 2025 pollen concentration prediction test, the model successfully predicted the peak seasons in spring and autumn (until August 25) and the trough periods from June to July in six districts. During the test, it can be seen that the model does not generate predicted values starting from the first data point in 2025 for the six districts. This is because the first look_back observations in the 2025 test dataset were used as the input window. Therefore, predictions began after this initial window. Similar to the fitted values generated during training, the model successfully predicted pollen concentrations above 800 grains/1000 mm^2 in Chaoyang, Haidian, Fengtai, and Shijingshan. The model achieved test R values of 0.60 and 0.64 in the prediction for Haidian and Fengtai, respectively. There are no predicted values higher than 800 grains/1000 mm^2 in the predictions for Changping and Shunyi, which can be attributed to the relatively poor

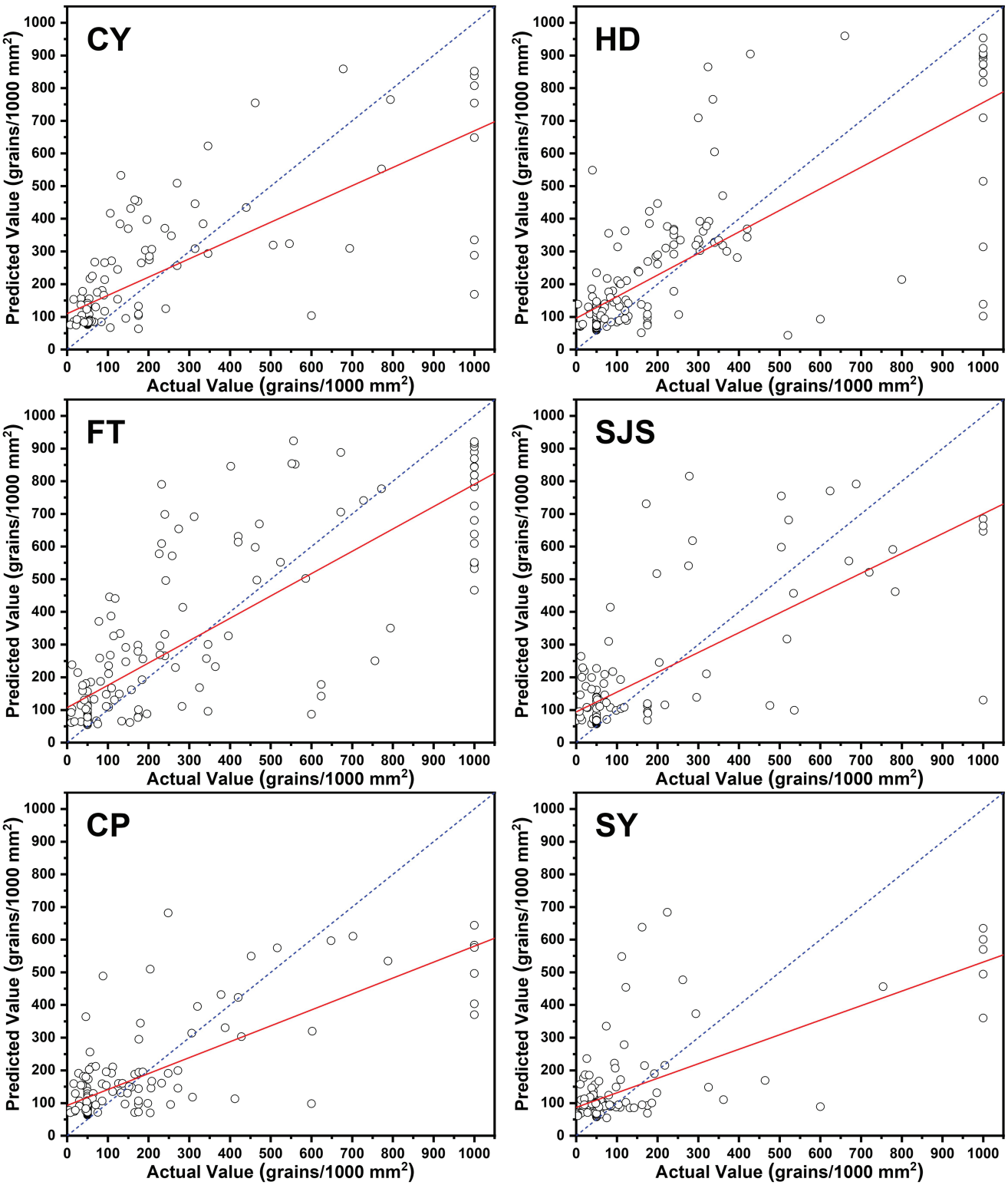


Figure 7. Scatter plots of actual and predicted pollen concentrations in 2025 for the six districts. The red solid line is the linear fitting line of the predicted pollen concentration values based on the least squares method, and the blue dashed line represents the perfect fit line ($\mathbf{y = x}$)

model performance for these two districts in the training phase. However, the model achieved an R (0.56) in the prediction for Changping, which indicates that the predicted value still has a good match with the actual value. In addition, Changping and Shunyi also had the two lowest RMSE and MAE values among the six districts. It is easy to see that the pollen concentration data have strong seasonality from Fig. 6, so some simple methods may also be able to capture these seasonal variations and achieve the prediction of pollen concentrations. Table 4 summarizes the RMSE, MAE, and

Table 4. RMSE, MAE, and R^2 of model predicted pollen concentration values using the moving average method and the historical same-date mean method

District	Moving average method Predicted value			Historical same-date mean method Predicted Value		
	RMSE	MAE	R^2	RMSE	MAE	R^2
CY	189.04	104.31	0.47	196.46	109.94	0.43
HD	207.57	115.65	0.50	242.22	156.73	0.32
FT	202.96	133.23	0.61	217.63	141.88	0.55
SJS	173.01	96.80	0.50	203.09	119.14	0.32
CP	159.35	95.23	0.53	162.60	102.02	0.51
SY	168.98	91.47	0.34	169.88	91.13	0.33

R^2 values of predicted pollen concentration values when using the moving average method and the historical same-date mean method. These two relatively simple methods were still able to predict pollen concentrations, which can be attributed to the strong seasonality of the pollen concentration data. However, the LSTM method achieved the lowest RMSE and MAE values and the highest R^2 values overall (Table 3) compared with them, even though these two methods can slightly outperform the LSTM model in some districts. Besides, the LSTM model will update its internal states through gating mechanisms and nonlinear operations when processing a sequence, enabling it to capture more complex temporal relationships in historical data. When the seasonality of pollen concentrations becomes less obvious or changes substantially, the LSTM model may perform better than these two methods. Therefore, this study still used the LSTM model for pollen concentration prediction. Fig. 7 presents the scatter plots between the observed actual values and the model predicted values for the six districts. The data points for Chaoyang, Haidian, Fengtai, and Shijingshan are distributed throughout the graph (especially the area above 800 grains/1000 mm^2), and their predicted value fitting lines are closer to the perfect fit line than the fitting lines of Changping and Shunyi districts. Changping and Shunyi have dense point distribution in the low pollen concentration areas, and the point distribution in high pollen concentration areas is very sparse. These results indicate that the model performs well for low pollen concentrations in these two districts, although it has poor performance in high pollen concentration prediction.

The prediction results for the six districts demonstrate the potential of the model in predicting pollen concentrations. It is worth mentioning that the LSTM-based prediction model in this study was developed using only the daily pollen concentration time series data. Compared with studies that added meteorological variables as predictors,[19,20,29] the model developed in this study can still achieve comparable prediction performance. Even with limited data, it is still able to provide useful pollen concentration predictions. However, the current model performs relatively poorly in predicting high pollen concentrations, especially for the data from Changping and Shunyi. This is mainly associated with the lack of accurate high pollen concentration values for the six districts in the datasets. In addition, the change in the data release format of the data source (WeChat official account) in 2025 also limited the data collection, and the datasets could only be updated to August 25, 2025. These issues could be mitigated by collecting more pollen concentration data and further optimizing the model configuration in future research. In summary, the current model successfully achieved the aim of predicting pollen concentrations across the six districts and may provide a useful reference for people suffering from pollen allergies.

Recommendations for future work

Although the present study demonstrated the potential of using an LSTM-based model for district-level pollen concentration prediction, there remains room for further improvement. Future research should first expand the dataset by obtaining more pollen concentration data (such as the data from before 2021) and continuing data collection over more pollen seasons. Besides, district-level meteorological data, including temperature, humidity, precipitation, wind speed and direction, and sunshine duration, could be used as additional predictors to further improve model prediction performance. It is anticipated that many difficulties could be encountered in collecting these data. Therefore, exploring different variants of LSTM based on the accessible data to improve prediction performance could also be a potential direction for future research. Additionally, a multivariate model that jointly learns from data across all six districts could be explored. The pollen concentration data point distributions for Changping and Shunyi are relatively sparse at high pollen concentrations. Sharing information across nearby districts may help improve predictions for these two districts. Finally, although the predicted pollen concentrations have been evaluated using RMSE, MAE, and R^2, the accuracy of the corresponding pollen allergy grade predictions has not been assessed. Future work could evaluate overall classification accuracy, performance on days with high or very high pollen concentrations, and recall for Grade 5 events. Confusion matrices could also help identify which pollen allergy grades are most frequently misclassified.

Conclusions

In this study, an LSTM-based pollen concentration prediction model was developed to predict the pollen concentrations in six districts of Beijing. The model demonstrated moderate predictive capability and captured broad seasonal patterns across the six districts, although prediction of extreme pollen concentrations (>800 grains/1000 mm^2) remains limited. In addition, this work also studied the pollen concentration variation of nine woody and four herbaceous plant species in Changping District from 2021 to 2025. The research on these plants indicates that the total pollen concentrations in spring and autumn were mainly contributed by woody and herbaceous plants, respectively. The pollen of these plants can easily cause allergies, and the study of the pollen periods of these plants can provide a more detailed reference for people with pollen allergies. Overall, the district-level pollen concentration predictions, together with information on the pollen periods of different plant species, may provide a useful reference for the public, especially individuals suffering from pollen allergies.

Acknowledgements

I would like to thank Ms. Nancy Hua for her valuable suggestions during this study. I would also like to express my deep gratitude to my family for their continuous support and encouragement throughout this research.

References

[1] Gao Z, Fu WY, Sun Y, et al. Artemisia pollen allergy in China: component-resolved diagnosis reveals allergic asthma patients have significant multiple allergen sensitization. *Allergy*. 2019;74(2):284–293. doi:10.1111/all.13597.

[2] Nur Husna SM, Tan H-TT, Md Shukri N, Mohd Ashari NS, Wong KK. Allergic rhinitis: a clinical and pathophysiological overview. *Front Med*. 2022;9:874114. doi:10.3389/fmed.2022.874114.

[3] Berezhanskiy PV, Mahmoudizeh A, Fakhri Y. Pollen exposure and allergy risk: a systematic review and meta-analysis. *Int J Environ Health Res*. 2025.1–15. doi:10.1080/09603123.2025.2533354.

[4] Bergmann K-C, Berger M, Klimek L, et al. Nonpharmacological measures to prevent allergic symptoms in pollen allergy: a critical review. *Allergol Select*. 2021;5(1):349. doi:10.5414/ALX02294E.

[5] Muzalyova A, Brunner JO, Traidl-Hoffmann C, Damialis A. Forecasting Betula and Poaceae airborne pollen concentrations on a 3-hourly resolution in Augsburg, Germany: toward automatically generated, real-time predictions. *Aerobiologia*. 2021;37(3):425–446. doi:10.1007/s10453-021-09699-3.

[6] Myszkowska D, Majewska R. Pollen grains as allergenic environmental factors–new approach to the forecasting of the pollen concentration during the season. *Ann Agric Environ Med*. 2014;21(4). doi:10.5604/12321966.1129914.

[7] Piotrowska K. Forecasting the Poaceae pollen season in eastern Poland. *Grana*. 2012;51(4):263–269. doi:10.1080/00173134.2012.659204.

[8] Maya-Manzano JM, Smith M, Markey E, Hourihane Clancy J, Sodeau J, O´Connor DJ. Recent developments in monitoring and modelling airborne pollen, a review. *Grana*. 2021;60(1):1–19. doi:10.1080/00173134.2020.1769176.

[9] Šikoparija B, Marko O, Panić M, Jakovetić D, Radišić P. How to prepare a pollen calendar for forecasting daily pollen concentrations of *Ambrosia, Betula* and Poaceae? *Aerobiologia*. 2018;34(2):203–217. doi:10.1007/s10453-018-9507-9.

[10] Verstraeten WW, Dujardin S, Hoebeke L, et al. Spatiotemporal monitoring and modelling of birch pollen levels in Belgium. *Aerobiologia*. 2019;35(4):703–717. doi:10.1007/s10453-019-09607-w.

[11] Sofiev M, Belmonte J, Gehrig R, et al. Airborne pollen transport. In: Sofiev M, Bergmann K-C (Eds.), *Allergenic Pollen: A Review of the Production, Release, Distribution and Health Impacts* (pp. 127–159). Dordrecht: Springer; 2013. doi:10.1007/978-94-007-4881-1_5

[12] Makra L, Coviello L, Gobbi A, et al. Forecasting daily total pollen concentrations on a global scale. *Allergy*. 2024;79(8):2173–2185. doi:10.1111/all.16227.

[13] Mills SA, Maya-Manzano JM, Tummon F, MacKenzie AR, Pope FD. Machine learning methods for low-cost pollen monitoring–model optimisation and interpretability. *Sci Total Environ*. 2023;903(4):165853. doi:10.1016/j.scitotenv.2023.165853.

[14] Tagliabue LC, Cecconi FR, Rinaldi S, Ciribini ALC. Data driven indoor air quality prediction in educational facilities based on IoT network. *Energy Build*. 2021;236(1):110782. doi:10.1016/j.enbuild.2021.110782.

[15] Zhu Y, Al-Ahmed SA, Shakir MZ, Olszewska JI. LSTM-based IoT-enabled CO_2 steady-state forecasting for indoor air quality monitoring. *Electronics*. 2022;12(1):107. doi:10.3390/electronics12010107.

[16] Navares R, Aznarte JL. Predicting air quality with deep learning LSTM: towards comprehensive models. *Ecol Inf*. 2020;55(1):101019. doi:10.1016/j.ecoinf.2019.101019.

[17] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735–1780. doi:10.1162/neco.1997.9.8.1735.

[18] Hurtado S, Antequera-Gómez ML, Barba-González C, Picornell A, Navas-Delgado I. e-Science workflow: a semantic approach for airborne pollen prediction. *Knowl-Based Syst*. 2024;284(12):111230. doi:10.1016/j.knosys.2023.111230.

[19] Picornell A, Hurtado S, Antequera-Gómez ML, et al. A deep learning LSTM-based approach for forecasting annual pollen curves: Olea and Urticaceae pollen types as a case study. *Comput Biol Med*. 2024;168:107706. doi:10.1016/j.compbiomed.2023.107706.

[20] Astray G, Amigo Fernández R, Fernández-González M, Dias-Lorenzo DA, Guada G, Rodríguez-Rajo FJ. Machine learning to forecast airborne parietaria Pollen in the North-West of the Iberian Peninsula. *Sustainability*. 2025;17(4):1528. doi:10.3390/su17041528.

[21] Jia P, Zhuang D, Wang Y. Impacts of temperature and precipitation on the spatiotemporal distribution of water resources in Chinese mega cities: the case of Beijing. *J Water Clim Change*. 2017;8(4):593–612. doi:10.2166/wcc.2017.038.

[22] General Information about Beijing. The people's government of Beijing municipality; 2025. https://www.beijing.gov.cn/.

[23] Hou Y-B, Sun J-L. Common pollen and related allergen components in patients with allergic diseases in the Beijing area. *Front Allergy*. 2024;5:1478392. doi:10.3389/falgy.2024.1478392.

[24] Beijing Administrative Map. Beijing geospatial information service platform; 2025. https://beijing.tianditu.gov.cn/bzdt/.
[25] Miki K, Kawashima S, Kobayashi S, Takeuchi S, Tseng Y-T, Nakamura K. Evaluating inaccurate pollen concentrations caused by turbulence using passive sampler. *Aerobiologia*. 2022;38(1):1–12. doi:10.1007/s10453-021-09728-1.
[26] Meteorological Index of Pollen Allergy. National public service platform for standards information; 2016. https://hbba.sacinfo.org.cn/stdDetail/20d486639a4318d6b33665d6669b387f.
[27] Lv S, Cheng B, Li J, Shen Q. Temperature time series prediction using convolutional bidirectional LSTM optimized by FPA. *2024 IEEE 4th International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)*; December 06–08, 2024; Chongqing, China.
[28] Al Mehedi MA, Amur A, Metcalf J, McGauley M, Smith V, Wadzuk B. Predicting the performance of green stormwater infrastructure using multivariate long short-term memory (LSTM) neural network. *J Hydrol*. 2023;625:130076. doi:10.1016/j.jhydrol.2023.130076.
[29] Cordero JM, Rojo J, Gutiérrez-Bustillo AM, Narros A, Borge R. Predicting the *Olea* pollen concentration with a machine learning algorithm ensemble. *Int J Biometeorol*. 2021;65(4):541–554. doi:10.1007/s00484-020-02047-z.
[30] Zhaobin S, Yuxin Z, Xingqin A, et al. Effects of airborne pollen on allergic rhinitis and asthma across different age groups in Beijing, China. *Sci Total Environ*. 2024;912:169215. doi:10.1016/j.scitotenv.2023.169215.
[31] Li X, Li X, Zhang M, Luo Q, Li Y, Dong L. Urban park attributes as predictors for the diversity and composition of spontaneous plants—a case in Beijing, China. *Urban Forest Urban Green*. 2024;91:128185. doi:10.1016/j.ufug.2023.128185.

About the Author



Kuang Yuchen is an 11th-grade student at Experimental High School Attached to Beijing Normal University. Her research interests lie at the intersection of physics, data science, and the quantitative analysis of real-world phenomena. Her research includes computational environmental modeling, with the goal of exploring how computational methods can help address problems with direct relevance to everyday life, as well as hands-on experimental studies in fluid dynamics and the optical measurement of thin liquid films. In addition to her research, she also leads her school's Physics Tournament community. She organizes academic workshops and mentors junior students to help them develop their research skills. She hopes to further explore how mathematical modeling, experimentation, and rigorous observation can be integrated to unpack complex physical and environmental systems.