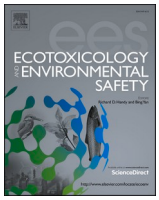
Ecotoxicology and Environmental Safety 309 (2026) 119659

Contents lists available at ScienceDirect

Ecotoxicology and Environmental Safety

journal homepage: www.elsevier.com/locate/ecoenv



High-resolution mapping of allergenic pollen risk across China using ensemble machine learning

Jie Yin [a,b,c,d,1], Yuan Zhang [c,d,e,f,1], Yifei Du [a,b,c,d], Yuhui Ouyang [c,d,e,f], Chengshuo Wang [c,d,e,f], Zhiqi Ma [g], Hongtian Wang [h], Shengzhi Sun [a,b,c,d,*], Luo Zhang [c,d,e,f,**], Rui Chen [a,b,c,d,*]

[a] School of Public Health, Capital Medical University, Beijing 100069, China
[b] Beijing Key Laboratory of Environment and Aging, Capital Medical University, Beijing 100069, China
[c] Beijing Laboratory of Allergic Diseases, Beijing Municipal Education Commission, Beijing 100069, China
[d] Laboratory for Clinical Medicine, Capital Medical University, Beijing 100069, China
[e] Department of Allergy, Beijing Tongren Hospital, Capital Medical University, Beijing 100730, China
[f] Department of Otolaryngology Head and Neck Surgery, Beijing Tongren Hospital, Capital Medical University, Beijing 100730, China
[g] Department of Otorhinolaryngology, Affiliated Hangzhou First People's Hospital, School of Medicine, Westlake University, Hangzhou City, Zhejiang Province 310000, China
[h] Department of Allergy, Beijing Shijitan Hospital, Capital Medical University, Beijing, China

ARTICLE INFO

Edited by Dr. RENJIE CHEN

Keywords:
Allergenic airborne pollen
Machine learning
Spatiotemporal distribution
China

ABSTRACT

Airborne pollen is a key environmental allergen affecting millions across China. As pollen levels and allergy prevalence continue to rise under rapid urbanization and climate change, developing spatially explicit, long-term pollen datasets becomes increasingly important for public health and ecological risk assessment. In this study, we developed a novel ensemble machine learning framework integrating random forest and gradient boosting models to estimate daily tree and herbaceous pollen concentrations across mainland China from 2011 to 2023. Models were trained using daily pollen data from 27 monitoring sites during 2019–2024 and a rich set of predictors, including meteorological, vegetation, land use, and spatiotemporal variables. By applying the trained models to historical environmental datasets, we reconstructed nationwide daily pollen concentrations for 2011–2023 to extend the temporal coverage beyond the observational record. The models achieved high accuracy, with R^2 values of 0.90 (tree) and 0.89 (herbaceous), and root mean square errors of 0.58 and 0.49, respectively. Tree pollen peaked in early spring in eastern, northeastern, central, and southwestern regions, while herbaceous pollen peaked in late summer in northern and northwestern areas. Seasonal timing, temperature, and vegetation indices were key drivers, with short-term lagged temperature (0–7 days) strongly influencing predictions. This study provides the first nationwide, long-term, daily pollen dataset for China derived from observation-based modeling and historical reconstruction, serving as an important resource for ecological research and public health applications. The established modeling framework offers a robust foundation for pollen exposure assessment, allergy forecasting, and climate-responsive risk management of aeroallergens under changing environmental conditions.

1. Introduction

Airborne pollen is a major environmental trigger for allergic diseases, impacting 20–30 % of the global population (Pawankar et al., 2013), with approximately 200 million affected individuals in China alone (Zhou et al., 2022). The rising prevalence of allergic diseases in recent decades has been linked to climate-related changes in pollen dynamics, including increased pollen production, extended pollen

* Corresponding authors at: School of Public Health, Capital Medical University, Beijing 100069, China
** Corresponding author at: Department of Allergy, Beijing Tongren Hospital, Capital Medical University, Beijing 100730, China
E-mail addresses: shengzhisun@ccmu.edu.cn (S. Sun), dr.luozhang@139.com (L. Zhang), ruichen@ccmu.edu.cn (R. Chen).
[1] These authors contributed equally to this work

https://doi.org/10.1016/j.ecoenv.2025.119659
Received 5 December 2025; Received in revised form 29 December 2025; Accepted 29 December 2025

seasons, and enhanced allergenicity (Damialis et al., 2019). These changes not only pose direct public health threats but also represent important ecological indicators of vegetation response to climate and land use change (Ziska et al., 2019). This highlights the critical need for advanced spatiotemporal modeling and real-time pollen forecasting to inform allergy prevention, public health strategies, and urban ecological planning.

Airborne pollen concentrations are significantly influenced by meteorological factors, including temperature, humidity, and wind speed and direction (Maya-Manzano et al., 2017). Temperature generally exhibits a positive association with pollen levels, whereas humidity and precipitation are typically negatively correlated (Valipour Shokouhi et al., 2024a; Rahman et al., 2020; Lo et al., 2021; Khwarahm et al., 2014; Ritenberga et al., 2018; Tseng et al., 2018). Additionally, remote sensing data provide essential information for pollen modelling (Schnake-Mahl and Sommers, 2017; Li et al., 2019). Vegetation indices, such as normalized difference vegetation index (NDVI), captures vegetation activity and phenological stages linked to pollen release, while land cover and topographic data characterize source habitats and dispersal pathways (Valipour Shokouhi et al., 2024a; Lugonja et al., 2019). These spatially continuous indicators complement ground observations and improve the characterization of airborne pollen dynamics.

Many studies have developed predictive models that incorporate phenological progress and environmental variables to estimate the spatiotemporal distribution of airborne pollen (Valipour Shokouhi et al., 2024a, 2024b; Lo et al., 2021). While traditional statistical models such as multiple linear regression provide interpretability, they often fail to capture the complexities of nonlinear relationships between pollen and environmental factors (Cotos-Yanez et al., 2004; Hjort et al., 2016). Recently, advanced machine learning algorithms such as random forest, gradient boosting, and artificial neural networks, have been demonstrated superior predictive capabilities (Valipour Shokouhi et al., 2024b; Liu et al., 2022; Puc, 2012; Ouyang et al., 2025a; Ruan et al., 2024), particularly when utilizing ensemble methods that integrate multiple approaches.

In Europe and North America, several well-established operational pollen forecasting systems provide daily or near-real-time predictions for major allergenic pollen types. For example, the System for Integrated modeLing of Atmospheric coMposition (SILAM) developed by the Finnish Meteorological Institute (FMI) and operationalized by the Copernicus Atmosphere Monitoring Service (CAMS) deliver regional pollen forecasts across Europe using advanced atmospheric composition models (System, 2025; Forecasting, 2025). Similarly, in the United States, daily pollen forecasts and concentration maps are publicly available through platforms such as Pollen.com (Pollen.com, 2025). These systems typically integrate real-time meteorological data, vegetation indices, and process-based or statistical modeling approaches to support public health alerts and allergy prevention. However, such comprehensive forecasting infrastructure remains largely underdeveloped across much of Asia, particularly in China. Existing efforts are often limited to local-scale studies or short-term monitoring, lacking long-term, high-resolution datasets or national coverage. Moreover, few systems fully leverage multi-source environmental data and advanced modeling techniques.

Beyond forecasting, pollen datasets also serve as ecological indicators that reflect vegetation composition, phenology, and climate interactions at landscape and national scales. This study aims to develop a novel, machine learning framework to estimate daily tree and herbaceous pollen concentrations across mainland China from 2011 to 2023 at a 10-km spatial resolution. By integrating diverse environmental data, including meteorological, vegetation, geographic, and temporal variables, this study investigates the relationships between key environmental factors and pollen concentrations. Ultimately, it seeks to create the first long-term, high-resolution daily pollen dataset for China, providing a valuable tool for ecological monitoring, landscape management, and public health applications.

2. Methods

2.1. Pollen data

Daily airborne pollen concentrations were measured at 27 monitoring sites established by Beijing Tongren Hospital, covering a diverse range of ecological and climatic zones across mainland China (Fig. 1). Monitoring was conducted during the active pollen seasons from March to October each year between 2019 and 2024. Detailed monitoring periods for each site are provided in Supplementary Table S1.

Pollen sampling employed Durham-type samplers based on the gravimetric method, with collection slides replaced every 24 h. Although Durham traps are known to have lower sampling efficiency compared to volumetric methods such as Hirst-type samplers, they remain widely used in China due to their simplicity, cost-effectiveness, and the historical continuity of their use. To ensure data consistency and reliability, all monitoring sites followed standardized protocols for slide preparation, staining, and pollen identification. Slides were stained with alkaline fuchsin to enhance pollen grain visibility and examined under a light microscope (Olympus BX-51, 200 ×) for manual counting and taxonomic classification.

Because gravimetric methods can underestimate atmospheric pollen concentrations due to meteorological influences, we additionally converted Durham-derived counts to volumetric-equivalent concentrations. This conversion was derived from an accuracy evaluation conducted by our research team, which compared simultaneous measurements from Durham-type samplers and a newly developed volumetric suction sampler (Ouyang et al., 2025b). The study demonstrated a strong correlation ($R^2 = 0.7605$) and established the following linear conversion equation:

$$Y = 2.065X - 3.962$$

where X represents pollen concentrations collected by the Durham sampler (grains/1000 mm^2), and Y represents estimated volumetric concentrations (grains/m^3). This conversion was uniformly applied across all sites to improve comparability with volumetric-based measurements and reduce uncertainty associated with gravimetric sampling.

Mainland China typically experiences two distinct pollen seasons annually. The spring season is primarily characterized by tree pollens, while the autumn season is marked by elevated concentrations of allergenic herbaceous pollen. In this study, we estimated total pollen concentrations for tree and herbaceous pollen, respectively. Tree pollen includes taxa such as *Cupressaceae* (cypress), *Salicaceae* (willow and poplar), *Ulmaceae* (elm), *Betulaceae* (birch), *Pinaceae* (pine), and *Oleaceae* (white ash), which predominantly contribute to pollen levels during the spring and early summer. In contrast, herbaceous pollen mainly consists of taxa from *Asteraceae* (including both *Artemisia* and non-*Artemisia* species), *Moraceae* (genus *Humulus*), *Poaceae* (grasses), and *Chenopodiaceae* (goosefoot), which are more abundant in late summer and autumn.

2.2. Explanatory variables

The explanatory variables used in this study were derived from atmospheric reanalysis datasets and satellite remote sensing products. We incorporated a range of variables potentially influencing pollen concentrations, including meteorological variables, vegetation-related variables, land use types, spatial and temporal features (Table 1).

Daily meteorological variables, including ambient temperature, precipitation, relative humidity, surface pressure, wind speed, and wind direction, were extracted for buffers of 1 km, 5 km, and 10 km surrounding each pollen monitoring site. Wind speed and direction at 10 m were derived from the eastward (U) and northward (V) wind

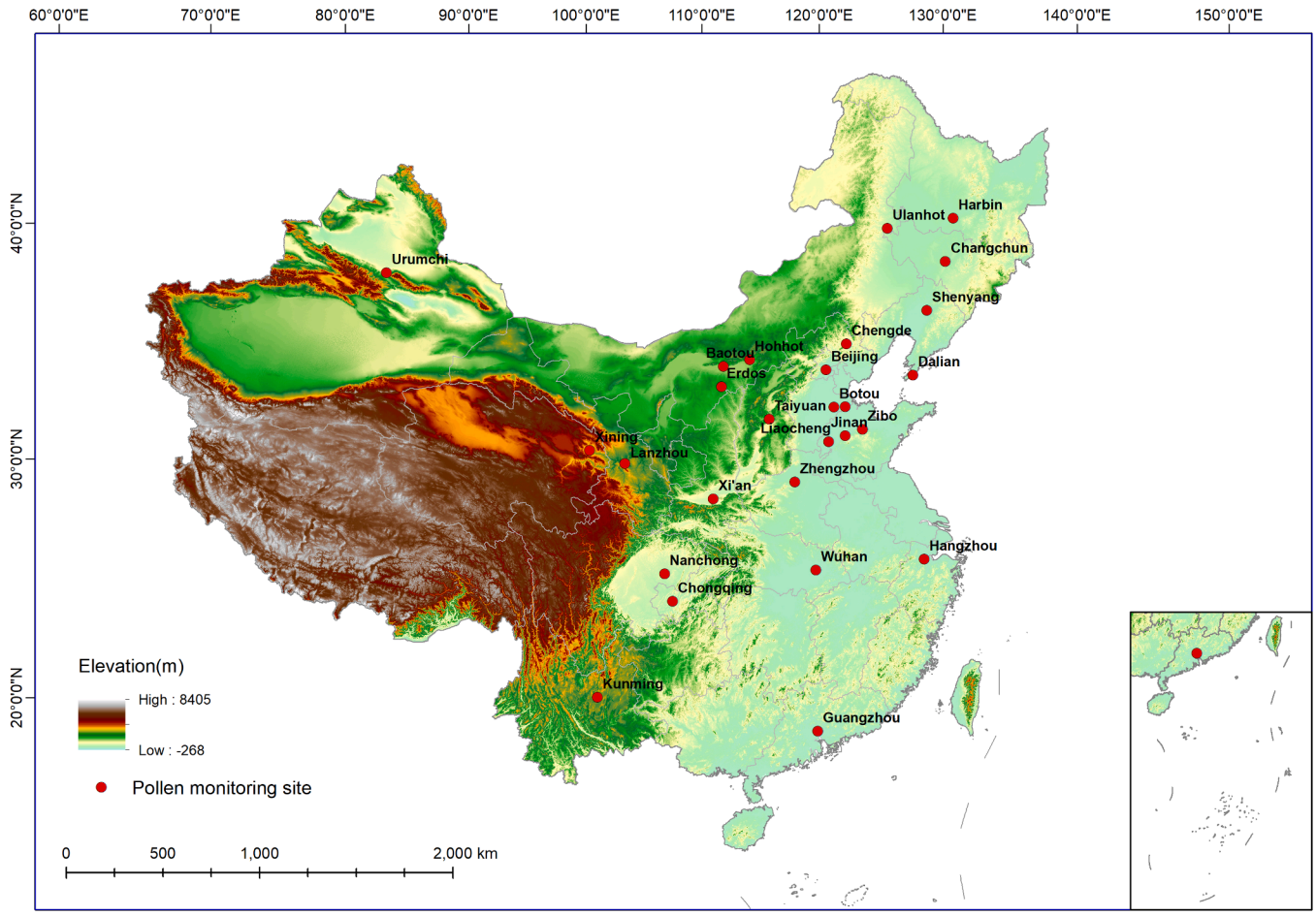


Fig. 1. Spatial distribution of pollen monitoring stations.

components. To account for the lagged effects of weather on pollen emission and dispersion, we collected daily values of each meteorological variable for the day of monitoring (lag 0) as well as for the preceding 1–7 days (lags 1–7).

In addition, we incorporated vegetation-related factors, including vegetation indices and land use characteristics, to reflect plant growth conditions and surrounding land cover. Specifically, we extracted the normalized difference vegetation index (NDVI), enhanced vegetation index (EVI), and leaf area index (LAI) for both high and low vegetation types across the 1 km, 5 km, and 10 km buffer zones. We also calculated the proportion of major land use types (e.g., cropland, forest, shrubland, grassland, and water) within each buffer zone to estimate their potential contributions to local pollen emissions. Spatial features included longitude, latitude, and elevation for each monitoring site. Temporal features such as day of the year, week of the year, month, and season. In addition, we created a binary indicator (1 = peak season, 0 = non-peak season) to distinguish periods of high and low pollen activity. The peak season for tree pollen was defined as March to June, while for herbaceous pollen, it was defined as July to October.

To ensure spatial and temporal alignment across datasets, bilinear interpolation was applied to resample meteorological variables and LAI data to a uniform spatial resolution of 1 km. Additionally, NDVI and EVI data, initially available at a 16-day temporal resolution, were resampled to daily values using spatiotemporal interpolation techniques. These preprocessing steps were implemented to maintain consistency with daily pollen concentration data and to enhance the accuracy of subsequent model training.

2.3. Statistical methods

In this study, we used ensemble machine learning techniques to model the relationship between a comprehensive set of 23 spatiotemporal variables and daily airborne pollen concentrations. Separate estimate models were developed for tree and herbaceous pollen to account for their distinct ecological and phenological characteristics.

Prior to model development, we applied natural logarithmic transformation with an offset of 1 to daily pollen concentrations to reduce right-skewness. Potential outliers were identified using the conventional 1.5 ×interquartile range (IQR) method, and observations with missing values were excluded to ensure consistent model training.

2.4. Model development

For the models, we used two machine learning algorithms: the Random Forest Regressor (RFR) and the Gradient Boosting Regressor (GBR). The Random Forest Regressor was chosen to capture the complex nonlinear relationships between environmental factors and pollen concentration while effectively reducing the risk of overfitting. The formula for the Random Forest Regressor is expressed as:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x) \tag{1}$$

where: $\hat{y}$ represents the estimated pollen concentration, T is the number of trees, and $f_t(x)$ denotes the estimation from the t-th tree.

The Gradient Boosting Regressor was used to construct additive models by sequentially minimizing the residual errors from preceding

Table 1
Summary of input variables and data sources for pollen concentration estimation.

Variable	Unit	Spatial resolution	Temporal resolution	Temporal scope	Data source
Meteorology					
Temperature	°C	1 km	Daily	2011–2024	ERA5-Land Daily Aggregated
Relative humidity	%				
Precipitation	mm				
Surface pressure	kPa				
Wind speed	m/s				
Wind direction	°				
Vegetation					
LAI_low	–	1 km	Daily	2011–2024	ERA5-Land Daily Aggregated
LAI_high					
NDVI	–	1 km	16 days	2011–2024	MODIS/Terra Vegetation Indices
EVI	–				
Land use type					
Cropland	%	30 m	Annual	2011–2024	Annual China Land Cover Dataset (Yang and Huang, 2024)
Forest					
Shrub					
Grassland					
Water					
Spatial					
Longitude	–	–	–	–	–
Latitude					
Elevation	m	1 km	–	–	Shuttle Radar Topography Mission
Temporal					
Day of the year	–	–	–	–	–
Week of the year					
Month					
Season					
Peak season					

Abbreviations: LAI_low = leaf area index for low vegetation; LAI_high = leaf area index for high vegetation; NDVI = normalized difference vegetation index; EVI = enhanced vegetation index.

trees, thereby improving accuracy. Its formulation is as follows:

$$\widehat{y} = \sum_{m=1}^{M} \eta h_m(x) \quad (2)$$

where $\widehat{y}$ is the final output, M is the number of boosting iterations, η is the learning rate, $h_m(x)$ is the fitted tree at iteration m.

To enhance model robustness and generalization, we implemented a Voting Regressor that combines the outputs of the Random Forest Regressor and the Gradient Boosting Regressor by averaging their estimations.

2.5. Model optimization

To optimize model performance, we used feature selection and hyperparameter tuning. Feature selection combined recursive feature elimination with variable importance rankings from the Random Forest algorithm to retain only the most informative features. Hyperparameter tuning was performed using grid search with 10-fold cross-validation. For the Random Forest Regressor, we tested tree counts of 200, 500, and 1000; maximum depths of 10, 15, and 20; and minimum samples for splitting and leaf nodes set at (5, 10) and (2, 5), respectively. For the Gradient Boosting Regressor, we evaluated the number of boosting iterations (100, 200, 300), learning rates (0.01, 0.1, 0.2), and maximum tree depths (3, 5, 7). To prevent overfitting, early stopping was implemented by monitoring validation loss during training. Model performance was evaluated using the coefficient of determination (R^2) and root mean square error (RMSE).

2.6. Nationwide daily pollen maps

To generate nationwide daily pollen concentration maps, we applied the optimized tree and herbaceous pollen models, selected based on maximized R^2 and minimized RMSE from 10-fold cross-validation, to gridded environmental variable datasets with a spatial resolution of 10 km. Environmental variables from March to October for each year between 2011 and 2023 were used as inputs to estimate daily pollen concentrations across China. Separate raster maps were produced for tree and herbaceous pollen, both maintaining a consistent spatial resolution of 10 km.

2.7. Sensitivity analyses

To evaluate the robustness of the model and examine its dependence on temporal features, we conducted two sensitivity analyses. First, to assess the extent to which model performance relied on calendar-based temporal features, we re-trained the models after completely excluding all temporal variables, including day of the year, week of the year, month, and season. Second, to examine the model's ability to generalize across time and support historical reconstruction, we trained the models using data from 2020 to 2024 and evaluated their performance on an independent dataset from 2019.

3. Results

3.1. Seasonal patterns of pollen

Tree and herbaceous pollen displayed complementary seasonal patterns, with tree pollen predominating in the spring and herbaceous pollen becoming more prevalent in late summer to early autumn (Fig. 2). Tree pollen exhibited an early-season peak, with concentrations starting to rise in early March, reaching a sharp maximum in early to mid-April, and then gradually declining by June. In contrast, herbaceous pollen was sparsely present in early spring but began to rise significantly in July, peaking sharply from late August to early September, before tapering off by October. These seasonal trends were consistent across all monitoring sites except Guangzhou, where herbaceous pollen concentrations were slightly higher than tree pollen in spring.

3.2. Model evaluation and validation

The models demonstrated high accuracy in estimating daily pollen concentrations. For tree pollen, the overall R^2 reached 0.90 with an RMSE of 0.58, while the herbaceous pollen model achieved an R^2 of 0.89 and an RMSE of 0.49. Model performance remained robust during peak pollen seasons (R^2 = 0.82–0.88) and non-peak periods (R^2 = 0.81–0.89) (Table 2). Ten-fold cross-validation further confirmed the stability of the models, with consistent results across folds (Supplementary Table S2).

Scatter plots of observed versus estimated values showed that most estimates closely followed the 1:1 reference line (dashed red), suggesting minimal bias across most concentration ranges (Fig. 3). However, greater dispersion was observed at higher concentration ranges, with both models tending to slightly underestimate extreme values. Daily predicted values showed consistent spatiotemporal trends with observations in most cities, offering a more detailed appraisal of model performance at the city level (Supplementary Figures S1 and S2).

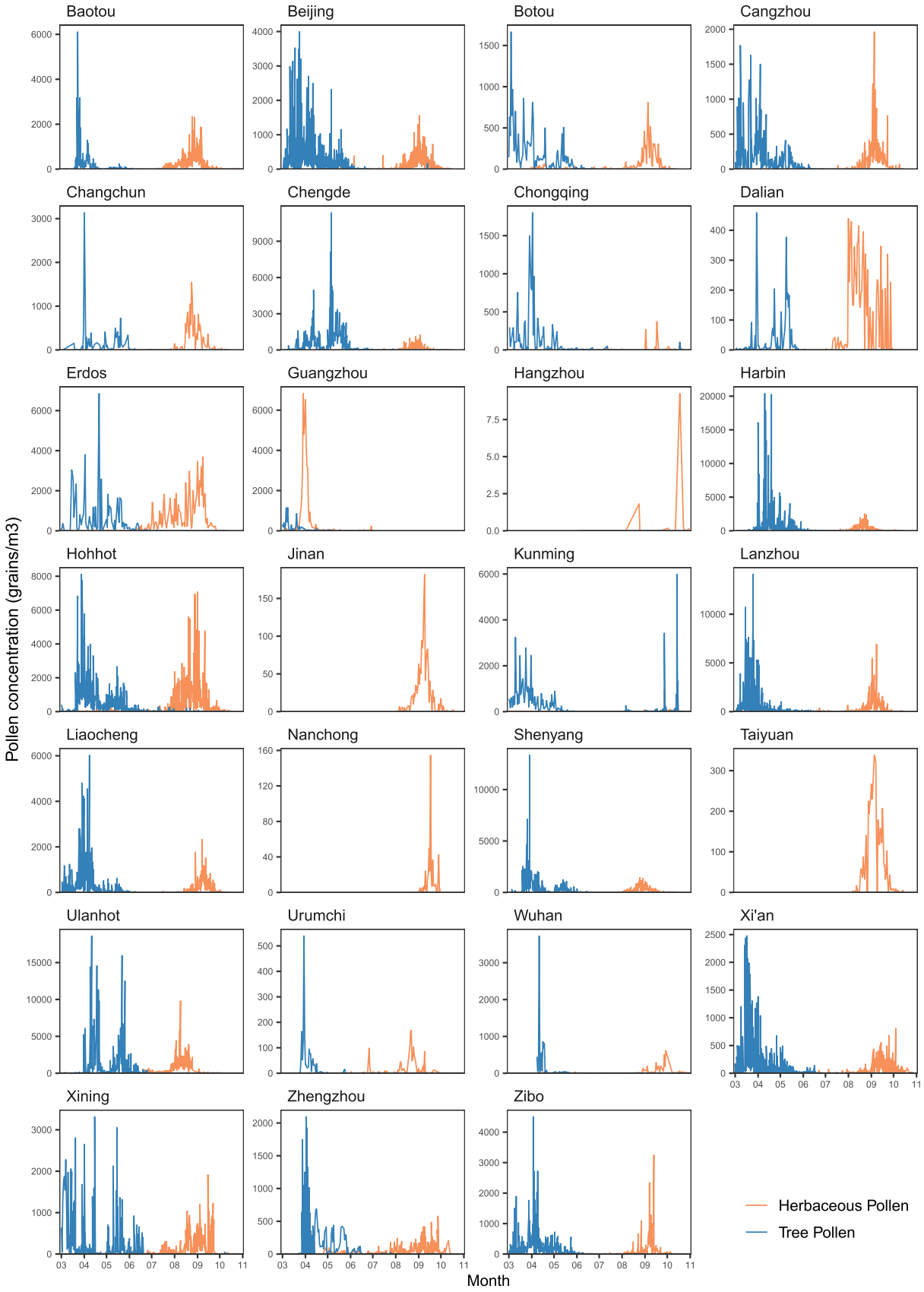


Fig. 2. Daily average pollen concentrations of tree and herbaceous plants at each monitoring site from 2019 to 2024. Note: Data for Guangzhou covers only the first pollen season in 2024, and data for Taiyuan, Jinan, Nanchong, and Hangzhou covers only the latter half of the pollen season in 2023.

Table 2
Cross-validated performance of the tree and herbaceous pollen models.

Pollen types	Time period	R^2	RMSE
Tree pollen	Overall	0.90	0.58
	Peak season	0.88	0.63
	Non-peak season	0.81	0.35
Herbaceous pollen	Overall	0.89	0.49
	Peak season	0.82	0.38
	Non-peak season	0.89	0.53

Abbreviation: R^2 = coefficient of determination; RMSE = root mean square error.

3.3. Feature importance of explanatory variables

Feature importance analysis revealed that both tree and herbaceous pollen models were primarily driven by temporal variables, with day of year and week of year ranking highest (Fig. 4). Day of year alone contributed 48.16 % and 25.96 % of the total importance in the tree and herbaceous pollen models, respectively. Vegetation-related variables, including LAI, NDVI, and EVI within 5–10 km buffers, as well as meteorological variables such as temperature, relative humidity, and surface pressure (including short-term lags), also showed high importance.

Despite these similarities, distinct differences in feature importance were observed between the two models. For tree pollen, forest coverage within 5 km buffers and temperature with a 7-day lag were the most influential vegetation and meteorological variables, indicating that stronger source vegetation and cumulative thermal dependencies. In contrast, herbaceous pollen concentration was primarily driven by day of the year and latitude, suggesting stronger geographic dependence. Low vegetation LAI within 10 km and EVI within 5 km were the most important vegetation variables for herbaceous pollen, suggesting that local vegetation density and greenness play key roles in shaping its variation.

3.4. Relationships between key variables and pollen concentration

Tree pollen concentrations peaking around the 90th day of the year and declining steadily thereafter (Fig. 5**a**). Temperature with a 7-day lag was negatively associated with tree pollen concentration, particularly above 15 °C. Vegetation-related effects were modest: tree pollen showed a slight increase with higher EVI values. Forest cover within a 5 km buffer exhibited only limited and inconsistent influence, with minor increases at low levels and little change beyond. Notably, increasing LAI of low vegetation within a 10 km buffer was associated with reduced tree pollen concentrations.

In contrast, herbaceous pollen remained low until around the 190th day of the year, then rose sharply to peak between days 225 and 250, aligning with elevated values in the late summer to early autumn weeks (Fig. 5**b**). herbaceous pollen was negatively correlated with latitude below 30°, but positively correlated above 30°, suggesting differing regional dynamics. Higher low-vegetation LAI within a 10 km radius was associated with reduced herbaceous pollen levels, whereas EVI generally showed a positive association with herbaceous pollen. Surface pressure exhibited a threshold response: below 89 kPa, herbaceous pollen concentration increased with decreasing pressure, while above this threshold, concentrations dropped abruptly and then remained nearly constant.

3.5. Nationwide daily pollen concentration maps

Nationwide daily pollen concentrations from March to October (2011–2023) were estimated at a 10 km resolution, with spatial patterns illustrated using the average monthly pollen concentrations calculated from daily predictions during the 2023 pollen season (Fig. 6). Tree pollen concentrations were high from March to May, with extensive coverage in eastern, northeastern, central, and southwestern China. From June to September, concentrations significantly decreased across most regions, with a slight resurgence observed in October, primarily in the southwestern areas. In contrast, herbaceous pollen concentrations remained low from March to June. Beginning in July, concentrations increased, reaching widespread high values in August and September, particularly in the northern and northwestern regions. By October, herbaceous pollen levels had largely diminished.

3.6. Sensitivity analyses

First, after excluding all temporal features, model performance remained high, with the R^2 values of 0.88 for tree pollen and 0.85 for herbaceous pollen, and RMSE values of 0.63 and 0.57, respectively. Model performance remained robust during both peak pollen seasons (R^2 = 0.83–0.85) and non-peak periods (R^2 = 0.79–0.84) (Supplementary Table S3). Feature importance analysis revealed that temperature-related variables were the dominant predictors for tree pollen, whereas vegetation-related variables contributed most strongly to herbaceous pollen estimation (Supplementary Figure S3). Second, models trained on data from 2020 to 2024 and evaluated on the independent 2019 dataset achieved R^2 values of 0.72 for tree pollen and 0.71 for herbaceous pollen. Estimated pollen concentrations were generally consistent with observed values in 2019 (Supplementary Figure S4), supporting the model's capacity to extrapolate across years.

4. Discussion

In this study, we developed and validated machine learning models to estimate daily concentrations of tree and herbaceous pollen across mainland China at a 10 km resolution. The models achieved high accuracy and revealed distinct seasonal and spatial patterns: tree pollen peaked in early spring, primarily in eastern, northeastern, central, and southwestern regions, while herbaceous pollen was most abundant in late summer, especially in northern and northwestern areas of China. Seasonal timing emerged as a dominant feature for both pollen types, while the associated environmental predictors differed, with tree pollen closely associated with source vegetation and temperature, and herbaceous pollen exhibited stronger correlations with latitude and vegetation index.

Numerous approaches have been developed to estimate airborne pollen concentrations. Early studies primarily relied on statistical models such as multiple linear regression (Ritenberga et al., 2016), generalized additive models (Cotos-Yanez et al., 2004), or time series methods (Rojo et al., 2017), which provided interpretable relationships but often struggled to capture complex nonlinear interactions between environmental factors and pollen levels. Process-based models, including phenology-driven and temperature accumulation models (Kmenta et al., 2017; Garcia-Mozo et al., 2000), have been used to simulate flowering periods and emission timing for specific taxa. However, these models typically require detailed plant physiological data and are less scalable for large-scale operational forecasting.

In parallel, mechanistic modeling approaches based on pollen emission, dispersion, and deposition processes have been developed and operationalized, particularly in Europe and North America (Siljamo et al., 2013). These models simulate the entire pollen life cycle by incorporating explicit parameterizations of emission based on meteorological conditions and phenology, atmospheric transport driven by wind fields, and removal through dry and wet deposition (Siljamo et al., 2013; Zink et al., 2012). Notable examples include SILAM, and COSMO-ART (Consortium for Small-scale Modelling-Aerosols and Reactive Trace gases) (Vogel et al., 2009), which are directly coupled with numerical weather prediction frameworks. Comprehensive air quality models like CMAQ (Community Multiscale Air Quality Modeling System) can be applied to pollen simulation (Ren et al., 2022). While these physics-based models provide valuable insights into large-scale pollen transport and interannual variability, they rely on extensive

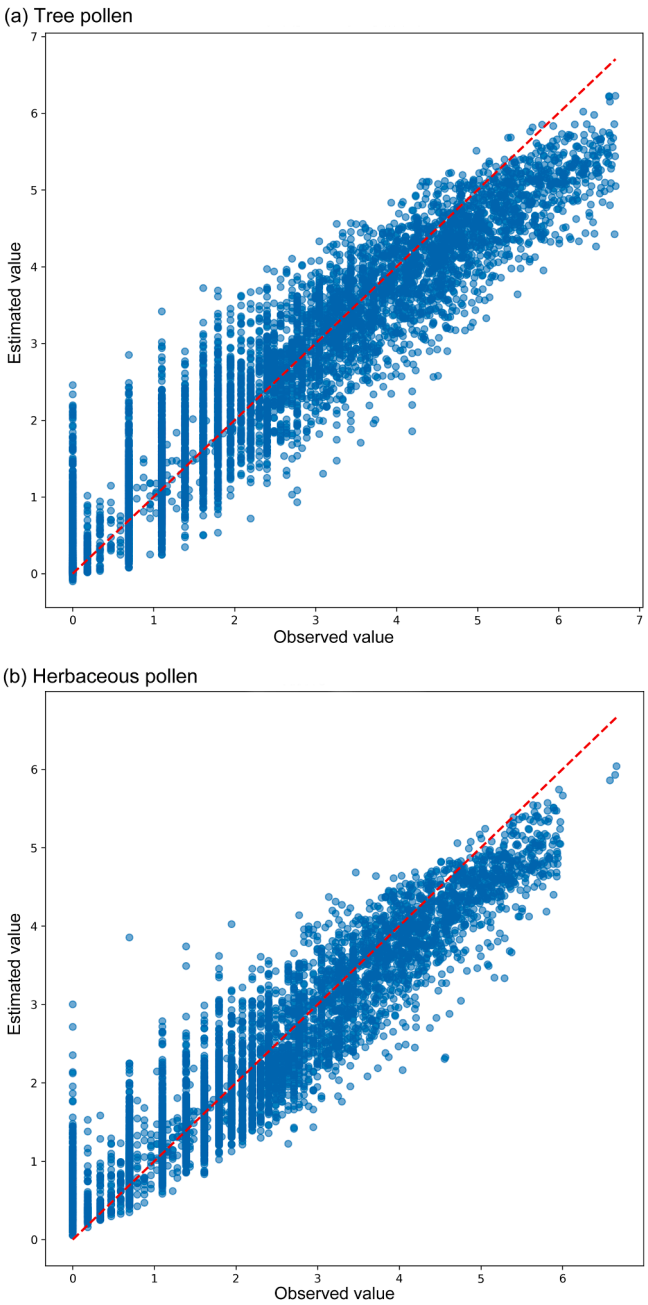


Fig. 3. Observed and estimated pollen concentrations for the tree and herbaceous pollen models. The dashed red line represents the 1:1 reference line. Note: Both observed and estimated values are log-transformed daily pollen concentrations.

high-resolution input data (e.g., detailed pollen source maps, species-specific emission potentials, and complex particle-process parameterizations), which can limit their applicability for operational forecasting in data-scarce regions.

Recently, machine learning approaches have emerged as a powerful alternative for modeling pollen concentrations. Algorithms such as random forest, support vector machines (Zhao et al., 2018), and artificial neural networks (Puc, 2012) have shown clear advantages in capturing nonlinear relationships and improving model accuracy. Unlike mechanistic models that reproduce underlying processes, machine learning methods infer empirical associations directly from observational data, making them particularly suitable for regions with limited knowledge of pollen source distributions or species-specific emission parameters. Studies conducted in Switzerland have demonstrated that machine learning models outperform linear models, with the random forest algorithm showing superior spatiotemporal performance and achieving R^2 values ranging from 0.84 to 0.91 for daily pollen concentrations[6]. More recent advancements indicate that ensemble models integrating multiple machine learning algorithms can further enhance model performance. One such study reported that a model combining six machine learning algorithms achieved R^2 values of 0.86 and 0.91 for birch and grass pollen, respectively (Valipour Shokouhi et al., 2024b). Our model similarly used this data-driven method. Rather than explicitly simulating emission and transport, it learns the complex associations between multi-source environmental predictors (meteorological conditions, vegetation indices, land use) and observed pollen concentrations. This approach enables effective prediction even in data-scarce regions and facilitates large-scale applications. Compared with existing studies in China, our approach substantially extends spatial coverage and temporal continuity while maintaining high predictive skill (Ouyang

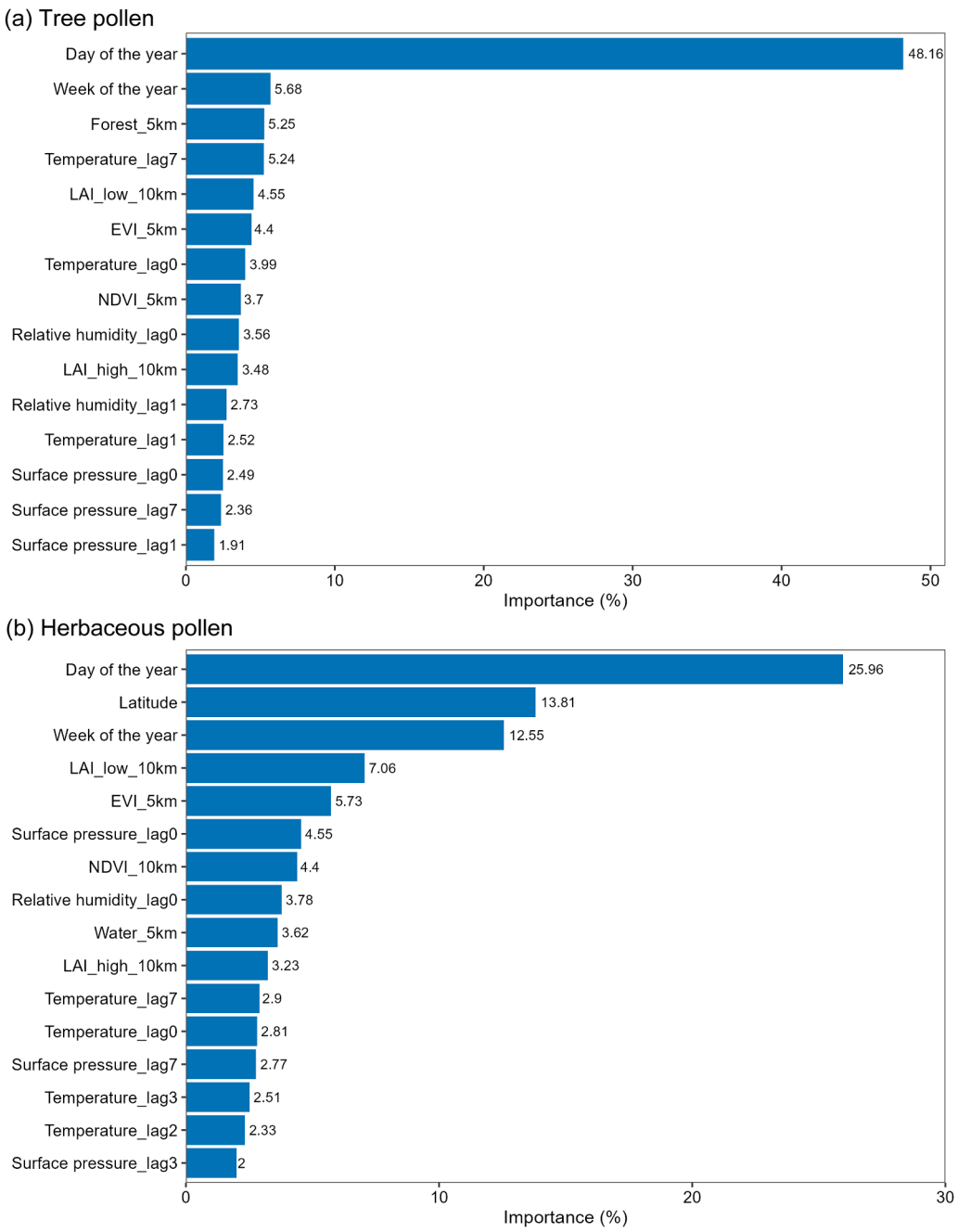


Fig. 4. Relative importance of variables in the tree and herbaceous pollen estimations. Abbreviations: NDVI = normalized difference vegetation index (5-km radius); EVI = enhanced vegetation index (5-km radius); LAI = leaf area index (low/high vegetation, 10-km radius); lag0/lag1/lag2/lag3/lag7 = current day/1-day/2-day/3-day/7-day lagged variables.

et al., 2025a; Li et al., 2025).

Notably, this is the first study to incorporate lagged meteorological effects into pollen estimation models at a national scale in China. Previous studies have primarily relied on same-day meteorological conditions (Valipour Shokouhi et al., 2024a, 2024b; Ravindra et al., 2022), potentially overlooking the delayed responses of pollen production and release to environmental drivers. Our results highlight the importance of antecedent weather conditions, with lagged temperature (0–7 days) identified as a significant feature for both pollen types (Aguilera et al., 2014).

Seasonal timing variables, including day of year, consistently ranked among the most important predictors, reflecting the inherently phenology-driven nature of pollen dynamics (Li et al., 2022). However, sensitivity analyses demonstrated that even after completely excluding all calendar-based temporal variables, model performance remained robust, with meteorological and vegetation-related predictors effectively reconstructing pollen variability. This finding indicates that temporal variables primarily function as parsimonious proxies for phenological alignment rather than dominating model structure. The distinct predictor importance profiles observed for tree versus herbaceous pollen further reflect differences in their ecological controls, supporting the value of plant-type-specific modeling strategies.

This study has several notable strengths. First, to the best of our knowledge, it is the first to develop large-scale models estimating daily airborne pollen concentrations for both tree and herbaceous plants across mainland China, incorporating multidimensional spatiotemporal variables. Previous studies have primarily focused on individual cities, with a lack of nationwide assessments. Second, by integrating random

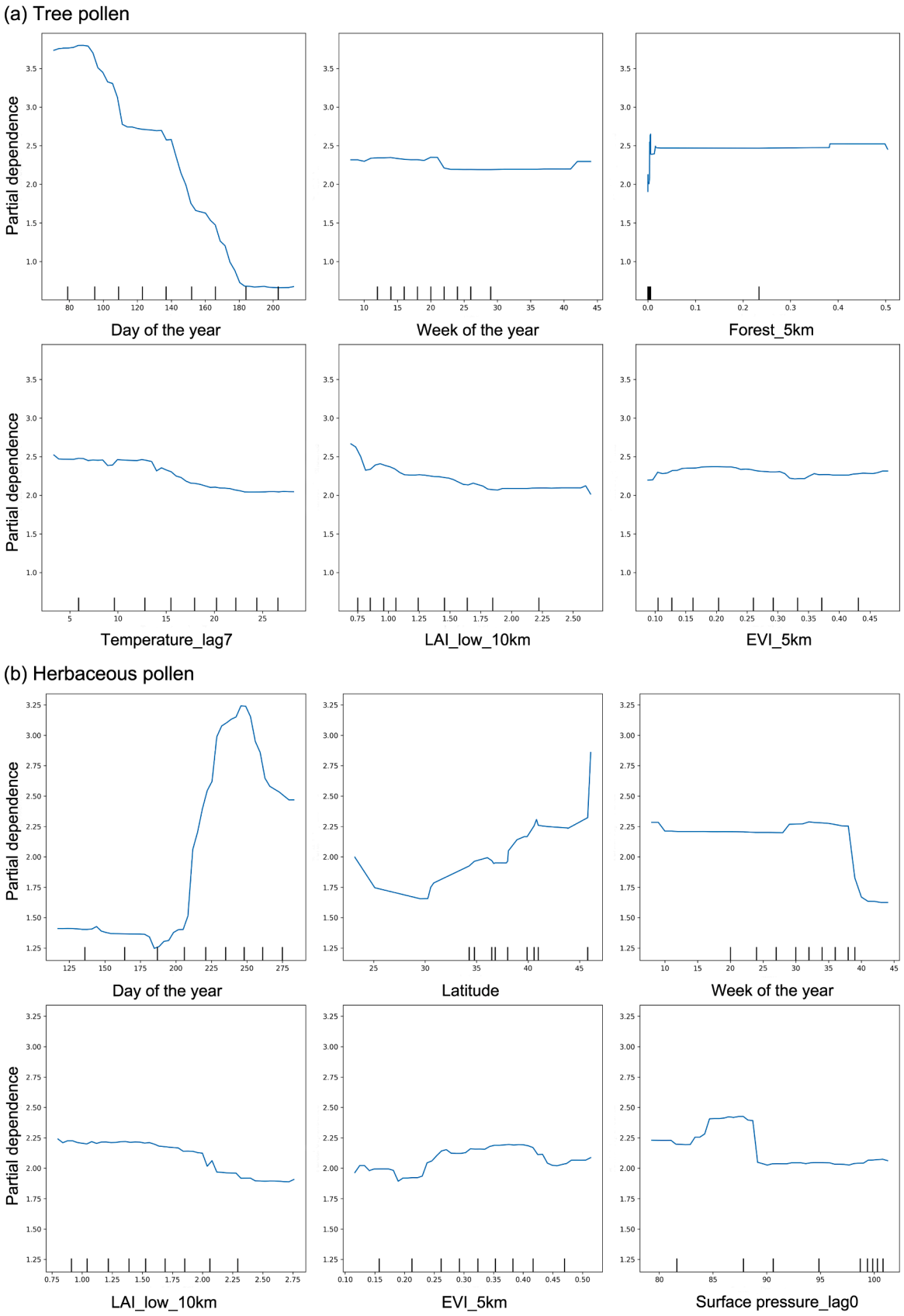


Fig. 5. Partial dependence plots of key variables for tree and herbaceous pollen. Abbreviations: EVI = enhanced vegetation index (5-km radius); LAI_low = leaf area index for low vegetation (10-km radius); lag0 /lag7 = current day/7-day lagged variables.

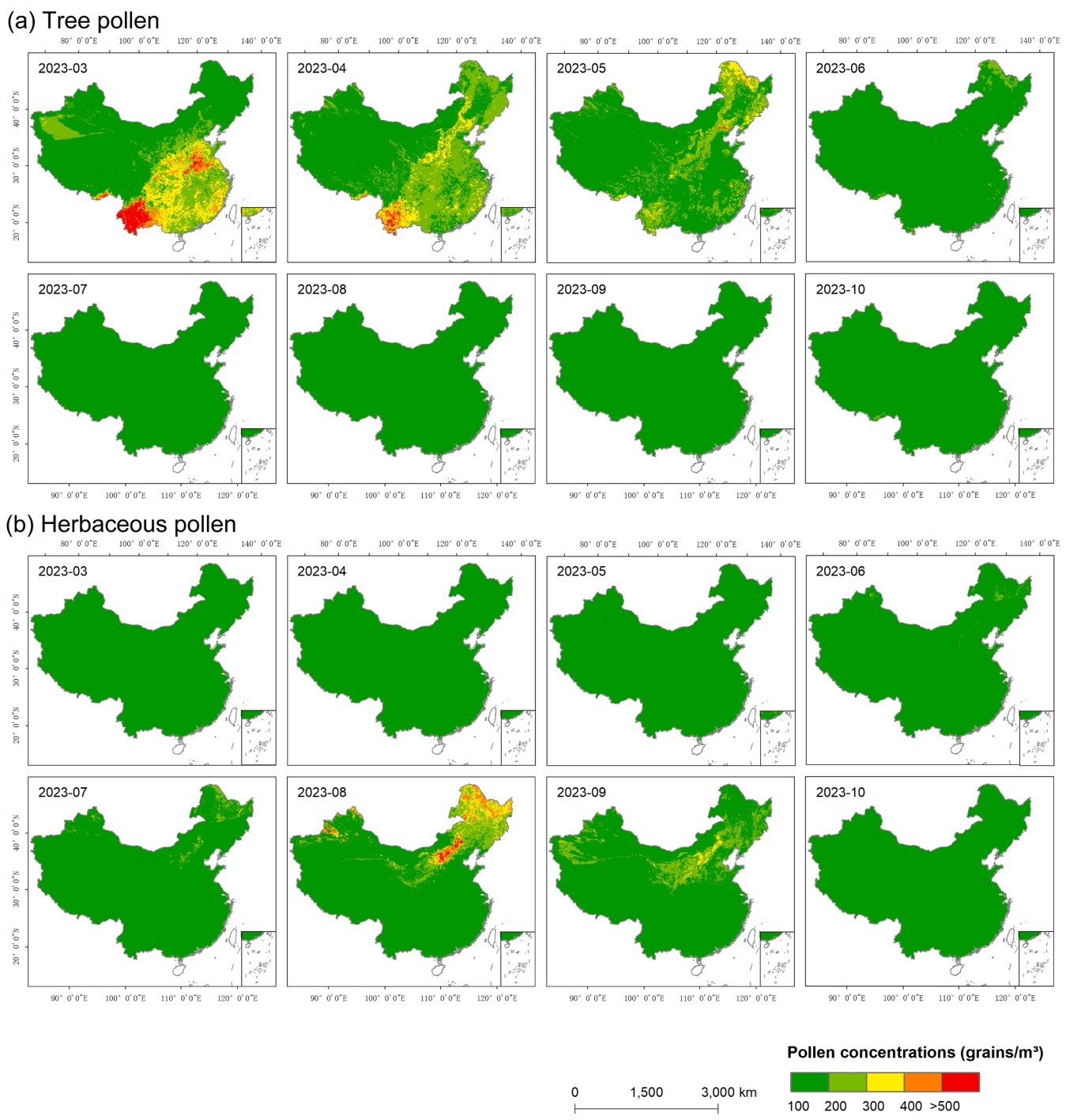


Fig. 6. Spatial distribution of average monthly tree and herbaceous pollen concentrations, calculated from daily estimates during the 2023 pollen season.

forest and gradient boosting regression, our ensemble model effectively captures complex nonlinear relationships and temporal lags in meteorological and phenological variables, substantially enhancing the model's accuracy and robustness. Third, the models account for the lagged effects of meteorological variables, which are often overlooked in prior research but are crucial for understanding pollen dynamics.

Despite the strong overall performance of our model, several sources of uncertainty should be considered when interpreting the reconstructed pollen fields. First, the historical estimates implicitly rely on the assumption that the empirical relationships learned between environmental predictors and pollen concentrations remain approximately stable over time. Although year-specific meteorological and vegetation inputs were used for each reconstructed year, ongoing climate change, land-use modification, and urban expansion may induce gradual non-stationarity in pollen-environment relationships, particularly in regions experiencing rapid environmental transitions (Wei et al., 2021; Picornell et al., 2023). Similar concerns regarding temporal transferability have been raised in large-scale environmental exposure reconstructions based on machine learning models (Valipour Shokouhi et al., 2024a; Wei et al., 2021). In this study, reduced predictive performance in independent hold-out years provides a quantitative indication of uncertainty associated with temporal transferability. As a result, reconstructed pollen fields are expected to be more reliable for capturing relative spatial contrasts and seasonal variability than for

representing exact year-specific absolute concentrations, with potentially higher uncertainty in earlier periods.

Second, spatial uncertainty arises from the uneven distribution of monitoring stations across China and the country's pronounced climatic and ecological heterogeneity. Monitoring sites are concentrated in eastern and central regions, whereas western China, characterized by complex terrain, arid and high-altitude climates, and distinct vegetation regimes, is relatively under-sampled. Predictions in these regions therefore rely on spatial extrapolation across climatic zones and increasing distances from monitoring locations, which may reduce robustness. Previous large-scale environmental modeling studies have shown that prediction uncertainty increases with distance to monitors and when extrapolating across poorly represented climate regimes (Valipour Shokouhi et al., 2024a; Adams-Groom et al., 2017). Consequently, pollen estimates for western and high-altitude regions should be interpreted with greater caution, particularly regarding absolute concentration levels.

Third, additional uncertainty is introduced by the measurement framework itself. Pollen observations were collected using Durham-type gravimetric samplers, and gravimetric counts were converted to volumetric-equivalent concentrations using an empirical calibration equation. While this conversion facilitates comparability with internationally used volumetric units, applying a uniform conversion across diverse climatic and ecological contexts may introduce spatially heterogeneous biases. Variations in pollen morphology, deposition velocity, wind conditions, and humidity can all influence gravimetric-to-volumetric relationships (Weber, 2003; Suanno et al., 2021). Such conversion-related uncertainty is likely to affect absolute concentration estimates more strongly than relative spatial or seasonal patterns, which are the primary focus of the present analysis.

Finally, the aggregation of pollen into broad tree and herbaceous categories introduces a form of structural uncertainty. Dominant allergenic taxa within each group differ substantially in phenology, climatic sensitivity, and allergenic potency, and aggregation may smooth sharp peaks associated with highly allergenic species such as *Betulaceae* or *Artemisia*, potentially underestimating short-term exposure risk (Lo et al., 2021; Sofiev et al., 2024). Species-level modeling could improve predictive specificity, enhance biological interpretability, and increase relevance for clinical and epidemiological applications. However, such modeling was not feasible in this study due to limitations in species-resolved data availability and spatial continuity. As more detailed and long-term taxon-level monitoring data become available, future work will prioritize species-specific extensions to further refine exposure assessment and health-oriented applications.

The high-resolution, daily pollen concentration fields developed in this study provide a valuable resource for public health research and practice, particularly in regions lacking routine pollen monitoring. By offering spatially continuous exposure estimates, the dataset can support epidemiological studies of pollen-related health outcomes, including allergic rhinitis, asthma exacerbations, and other environmentally sensitive conditions, using time-series or cohort-based designs. In addition, the reconstructed historical pollen fields enable assessment of long-term and seasonal exposure patterns, facilitating investigations of temporal variability and potential links with climate and land-use change. The modeling framework also has potential for operational application, as it could be coupled with meteorological forecasts to support pollen risk mapping and early warning, thereby complementing existing air quality management and public health decision-making systems.

5. Conclusion

This study developed and validated machine learning models to estimate daily concentrations of tree and herbaceous pollen across China, leveraging a comprehensive set of meteorological, vegetation, land use type, spatial and temporal variables. The models demonstrated strong performance in estimating pollen concentrations, with high accuracy and robustness across different pollen types. Nationwide daily allergenic pollen concentration maps from 2011 to 2023 were generated at a 10 km spatial resolution, providing a detailed depiction of seasonal and regional pollen dynamics. These findings enhance our understanding of allergenic airborne pollen behavior in both space and time, offering valuable resources for public health planning, allergy prevention, and ecological forecasting in the context of changing environmental conditions.

Abbreviations

SILAM	System for Integrated modeLing of Atmospheric coMposition
FMI	Finnish Meteorological Institute
CAMS	Copernicus Atmosphere Monitoring Service
NDVI	Normalized difference vegetation index
EVI	Enhanced vegetation index
LAI	Leaf area index
IQR	Interquartile range
RFR	Random Forest Regressor
GBR	Gradient Boosting Regressor
R^2	Coefficient of determination
RMSE	Root mean square error
COSMO-ART	Consortium for Small-scale Modelling-Aerosols and Reactive Trace gases
CMAQ	Community Multiscale Air Quality Model

CRediT authorship contribution statement

Jie Yin: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Yuan Zhang:** Writing – review & editing. **Yifei Du:** Writing – review & editing. **Yuhui Ouyang:** Writing – review & editing. **Chengshuo Wang:** Writing – review & editing. **Zhiqi Ma:** Writing – review & editing. **Hongtian Wang:** Writing – review & editing. **Shengzhi Sun:** Writing – review & editing. **Luo Zhang:** Writing – review & editing, Supervision. **Rui Chen:** Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study was financially supported by National Key R&D Program of China (2022YFC2504100), Beijing Nova Program (20250484793), National Science Fund for Distinguished Young Scholars (82025031), Key Program of the National Natural Science Foundation of China (82230109), Beijing Outstanding Young Scientist Program (JWZQ20240101024), Young Beijing Scholars Project and the Chinese Institutes for Medical Research, Beijing (CX23YZ01), China Postdoctoral Science Foundation (2024M752180), Postdoctoral Fellowship Program of CPSF (GZC20241095).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ecoenv.2025.119659.

Data Availability

Data will be made available on request. The pollen monitoring data are not publicly available due to data sharing agreements with local hospitals. However, the national daily gridded pollen concentration dataset is available from the corresponding author upon reasonable

request with a signed data access agreement.

References

Adams-Groom, B., Skjøth, C.A., Baker, M., Welch, T.E., 2017. Modelled and observed surface soil pollen deposition distance curves for isolated trees of Carpinus betulus, Cedrus atlantica, Juglans nigra and Platanus acerifolia. Aerobiologia 33 (3), 407–416. https://doi.org/10.1007/s10453-017-9479-1.

Aguilera, F., Ruiz, L., Fornaciari, M., et al., 2014. Heat accumulation period in the Mediterranean region: phenological response of the olive in different climate areas (Spain, Italy and Tunisia). Int J. Biometeorol. 58 (5), 867–876. https://doi.org/10.1007/s00484-013-0666-7.

Cotos-Yanez, T.R., Rodriguez-Rajo, F.J., Jato, M.V., 2004. Short-term prediction of Betula airborne pollen concentration in Vigo (NW Spain) using logistic additive models and partially linear models. Int J. Biometeorol. 48 (4), 179–185. https://doi.org/10.1007/s00484-004-0203-9.

Damialis, A., Traidl-Hoffmann, C., Treudler, R., 2019. Climate Change and Pollen Allergies (Published online). Biodivers. Health Face Clim. Change 47–66. https://doi.org/10.1007/978-3-030-02318-8_3.

Forecasting pollen to alleviate allergy suffering. Accessed May 28, 2025. ⟨https://stories.ecmwf.int/forecasting-pollen-to-alleviate-allergy-suffering/⟩.

Garcia-Mozo, H., Galan, C., Gomez-Casero, M.T., Dominguez, E., 2000. A comparative study of different temperature accumulation methods for predicting the start of the *Quercus* pollen season in Cordoba (South West Spain). Grana 39 (4), 194–199. https://doi.org/10.1080/00173130051084322.

Hjort, J., Hugg, T.T., Antikainen, H., et al., 2016. Fine-Scale Exposure to Allergenic Pollen in the Urban Environment: Evaluation of Land Use Regression Approach. Environ. Health Perspect. 124 (5), 619–626. https://doi.org/10.1289/ehp.1509761.

Khwarahm, N., Dash, J., Atkinson, P.M., et al., 2014. Exploring the spatio-temporal relationship between two key aeroallergens and meteorological variables in the United Kingdom. Int J. Biometeorol. 58 (4), 529–545. https://doi.org/10.1007/s00484-013-0739-7.

Kmenta, M., Bastl, K., Berger, U., et al., 2017. The grass pollen season 2015: a proof of concept multi-approach study in three different European cities. World Allergy Organ J. 10, 31. https://doi.org/10.1186/s40413-017-0163-2.

Li, J., An, X., Sun, Z., et al., 2025. Construction and application of a pollen emissions model based on phenology and random forests. Atmos. Chem. Phys. 25 (6), 3583–3602. https://doi.org/10.5194/acp-25-3583-2025.

Li, L., Hao, D., Li, X., et al., 2022. Satellite-based phenology products and in-situ pollen dynamics: A comparative assessment. Environ. Res. 204, 111937. https://doi.org/10.1016/j.envres.2021.111937.

Li, X., Zhou, Y., Meng, L., Asrar, G., Sapkota, A., Coates, F., 2019. Characterizing the relationship between satellite phenology and pollen season: a case study of birch. Remote Sens Environ. 222, 267–274. https://doi.org/10.1016/j.rse.2018.12.036.

Liu, T.L., Flückiger, B., De Hoogh, K., 2022. A comparison of statistical and machine-learning approaches for spatiotemporal modeling of nitrogen dioxide across Switzerland. Atmos. Pollut. Res 13 (12), 101611. https://doi.org/10.1016/j.apr.2022.101611.

Lo, F., Bitz, C.M., Hess, J.J., 2021. Development of a Random Forest model for forecasting allergenic pollen in North America. Sci. Total Environ. 773, 145590. https://doi.org/10.1016/j.scitotenv.2021.145590.

Lugonja, P., Brdar, S., Simović, I., et al., 2019. Integration of in situ and satellite data for top-down mapping of Ambrosia infection level. Remote Sens Environ. 235, 111455. https://doi.org/10.1016/j.rse.2019.111455.

Maya-Manzano, J.M., Sadyś, M., Tormo-Molina, R., et al., 2017. Relationships between airborne pollen grains, wind direction and land cover using GIS and circular statistics. Sci. Total Environ. 584-585, 603–613. https://doi.org/10.1016/j.scitotenv.2017.01.085.

Ouyang, Y., Yang, J., Zhang, J., Yan, Y., Zhang, L., 2025b. Accuracy evaluation of airborne pollen concentrations monitored by a new volumetric suction pollen monitor. Zhonghua Er Bi Yan Hou Tou Jing Wai Ke Za Zhi 60 (5), 527–531. https://doi.org/10.3760/cma.j.cn115330-20240501-00257.

Ouyang, Y., Yin, Z., Yan, Y., 2025a. Spring and summer-autumn pollen grading and forecasting model based on daily visits of allergic rhinitis patients. Zhonghua Er Bi Yan Hou Tou Jing Wai Ke Za Zhi 60 (03), 313–320.

Ruby Pawankar, Giorgio Walter Canonica, Stephen T. Holgate, Richard F. Lockey, Michael S. Blaiss WAO White Book on Allergy: Update 2013. Published online 2013. ⟨https://allergypaais.org/wp-content/themes/twentytwentyone/pdf/WhiteBook2-2013-v8.pdf⟩.

Picornell, A., Ruiz-Mata, R., Rojo, J., et al., 2023. Applying wind patterns and land use to estimate the concentrations of airborne pollen of herbaceous taxa in a statistical framework. Urban Clim. 49, 101496. https://doi.org/10.1016/j.uclim.2023.101496.

Pollen.com. Accessed May 28, 2025. ⟨https://www.pollen.com/⟩.

Puc, M., 2012. Artificial neural network model of the relationship between Betula pollen and meteorological factors in Szczecin (Poland). Int J. Biometeorol. 56 (2), 395–401. https://doi.org/10.1007/s00484-011-0446-1.

Rahman, A., Luo, C., Chen, B., et al., 2020. Regional and seasonal variation of airborne pollen and spores among the cities of South China. Acta Ecol. Sin. 40 (4), 283–295. https://doi.org/10.1016/j.chnaes.2019.05.012.

Ravindra, K., Goyal, A., Mor, S., 2022. Influence of meteorological parameters and air pollutants on the airborne pollen of city Chandigarh, India. Sci. Total Environ. 818, 151829. https://doi.org/10.1016/j.scitotenv.2021.151829.

Ren, X., Cai, T., Mi, Z., Bielory, L., Nolte, C.G., Georgopoulos, P.G., 2022. Modeling past and future spatiotemporal distributions of airborne allergenic pollen across the contiguous United States. Front Allergy 3, 959594. https://doi.org/10.3389/falgy.2022.959594.

Ritenberga, O., Sofiev, M., Kirillova, V., Kalnina, L., Genikhovich, E., 2016. Statistical modelling of non-stationary processes of atmospheric pollution from natural sources: example of birch pollen. Agric. Meteor. 226-227, 96–107. https://doi.org/10.1016/j.agrformet.2016.05.016.

Ritenberga, O., Sofiev, M., Siljamo, P., et al., 2018. A statistical model for predicting the inter-annual variability of birch pollen abundance in Northern and North-Eastern Europe. Sci. Total Environ. 615, 228–239. https://doi.org/10.1016/j.scitotenv.2017.09.061.

Rojo, J., Rivero, R., Romero-Morte, J., Fernández-González, F., Pérez-Badia, R., 2017. Modeling pollen time series using seasonal-trend decomposition procedure based on LOESS smoothing. Int J. Biometeorol. 61 (2), 335–348. https://doi.org/10.1007/s00484-016-1215-y.

Ruan, W., Li, Z., Sun, Z., et al., 2024. Enhancing Pollen Prediction in Beijing, a Chinese Megacity: Leveraging Ensemble Learning Models for Greater Accuracy. Aerosol Air Qual. Res 24 (11), 240123. https://doi.org/10.4209/aaqr.240123.

Schnake-Mahl, A.S., Sommers, B.D., 2017. Health care in the suburbs: an analysis of suburban poverty and health care access. Health Aff. (Millwood) 36 (10), 1777–1785. https://doi.org/10.1377/hlthaff.2017.0545.

Siljamo, P., Sofiev, M., Filatova, E., et al., 2013. A numerical model of birch pollen emission and dispersion in the atmosphere. Model evaluation and sensitivity analysis. Int J. Biometeorol. 57 (1), 125–136. https://doi.org/10.1007/s00484-012-0539-5.

Sofiev, M., Palamarchuk, J., Kouznetsov, R., et al., 2024. European pollen reanalysis, 1980–2022, for alder, birch, and olive. Sci. Data 11 (1), 1082. https://doi.org/10.1038/s41597-024-03686-2.

Suanno, C., Aloisi, I., Fernández-González, D., Del Duca, S., 2021. Monitoring techniques for pollen allergy risk assessment. Environ. Res 197, 111109. https://doi.org/10.1016/j.envres.2021.111109.

System for Integrated modeLling of Atmospheric coMposition. Accessed May 28, 2025. ⟨https://silam.fmi.fi/⟩.

Tseng, Y.T., Kawashima, S., Kobayashi, S., Takeuchi, S., Nakamura, K., 2018. Algorithm for forecasting the total amount of airborne birch pollen from meteorological conditions of previous years. Agric. Meteor. 249, 35–43. https://doi.org/10.1016/j.agrformet.2017.11.021.

Valipour Shokouhi, B., De Hoogh, K., Gehrig, R., Eeftens, M., 2024b. Spatiotemporal modelling of airborne birch and grass pollen concentration across Switzerland: a comparison of statistical, machine learning and ensemble methods. Environ. Res 263, 119999. https://doi.org/10.1016/j.envres.2024.119999.

Valipour Shokouhi, B., De Hoogh, K., Gehrig, R., Eeftens, M., 2024a. Estimation of historical daily airborne pollen concentrations across Switzerland using a spatio temporal random forest model. Sci. Total Environ. 906, 167286. https://doi.org/10.1016/j.scitotenv.2023.167286.

Vogel B., Vogel H., Baumer D., et al. The comprehensive model system COSMO-ART – Radiative impact of aerosol on the state of the atmosphere on the regional scale. Atmos Chem Phys. Published online 2009.

Weber, R.W., 2003. Meteorologic variables in aerobiology. Immunol. Allergy Clin. North Am. 23 (3), 411–422. https://doi.org/10.1016/S0889-8561(03)00062-6.

Wei, J., Li, Z., Lyapustin, A., et al., 2021. Reconstructing 1-km-resolution high-quality $PM_{2.5}$ data records from 2000 to 2018 in China: spatiotemporal variations and policy implications. Remote Sens Environ. 252, 112136. https://doi.org/10.1016/j.rse.2020.112136.

Yang, J., Huang, X., 2024. The 30 m annual land cover datasets and its dynamics in China from 1985 to 2023. Earth Syst. Sci. Data 13 (1), 3907–3925. https://doi.org/10.5281/zenodo.12779975.

Zhao, W., Wang, J., Yu, D., Zhang, G., 2018. Prediction of Daily Pollen Concentration using Support Vector Machine and Particle Swarm Optimization Algorithm.Int. J. Perform. Eng. 14 (11), 2808–2819. https://doi.org/10.23940/ijpe.18.11.p27.28082819.

Zhou, Y., Dai, J., Liu, H., Liu, X., 2022. Tourist risk assessment of pollen allergy in tourism attractions: a case study in the Summer Palace, Beijing, China. Front Public Health 10, 1030066. https://doi.org/10.3389/fpubh.2022.1030066.

Zink, K., Vogel, H., Vogel, B., Magyar, D., Kottmeier, C., 2012. Modeling the dispersion of Ambrosia artemisiifolia L. pollen with the model system COSMO-ART. Int J. Biometeorol. 56 (4), 669–680. https://doi.org/10.1007/s00484-011-0468-8.

Ziska, L.H., Makra, L., Harry, S.K., et al., 2019. Temperature-related changes in airborne allergenic pollen abundance and seasonality across the northern hemisphere: a retrospective data analysis. Lancet Planet Health 3 (3), e124–e131. https://doi.org/10.1016/S2542-5196(19)30015-4.