Supplementary Information

Meteorology Shapes Urban Particle Nucleation, While Pollution-Related Factors Modulate Growth

Shi Yin,* Shiyu Zhan, Mingrui Jiang, Yadong Wang, Yunxi Lu, Yongjun Hu, Christian George, Hartmut Herrmann, Lin Du,* Maofa Ge*

*Corresponding author. Email: yinshi@m.scnu.edu.cn; lindu@sdu.edu.cn; gemaofa@iccas.ac.cn

This Supplementary Information file includes:

Supplementary Methods

Measurements

The Beijing observation site was located on the roof of an academic building on the Peking University campus (16). Particle number size distributions (PNSDs) were measured continuously with a Fast Mobility Particle Sizer (FMPS; TSI 3091) over 5.6–560 nm. Measurements were conducted from 16 to 23 December 2011 and from 18 to 27 April 2012 at 1 s time resolution. The FMPS data were corrected following Zimmerman et al. (41); the upper midpoint limit of 523 nm was adjusted to 857 nm, and the scaling-down coefficient for total particle number concentration was 1.28 (16). At the same rooftop site, SO_2, NO_2, NO, O_3, CO_2 and CO were measured every minute. A nearby meteorological station measured air temperature (T), relative humidity (RH), wind speed (WS) and wind direction (WD). $PM_{2.5}$ and PM_{10} data were obtained from the Climate Prediction Center of the National Oceanic and Atmospheric Administration.

The Qingdao observation sites were located at the Qingdao campus of Shandong University (Site A, 120.69°E, 36.37°N) and Ocean University of China (Site B, 120.67°E, 36.35°N) (43). The two sites are approximately 3 km apart and are suburban locations with few industrial emissions and no obvious local point sources. Observations at Site A were conducted from 1 July to 19 August 2019. Observations at Site B were conducted from 11 November 2019 to 25 February 2020, from 22 March to 8 April 2020 and from 6 to 30 June 2020. Because NPF events can occur over spatial scales of tens to hundreds of kilometres and the two sites are close to one another, we treated the Qingdao observations as being influenced by broadly similar weather processes and air masses. PNSDs were measured with a GRIMM scanning mobility particle sizer (SMPS) over 10.8–500 nm. NO, SO_2, CO and O_3 were measured with Thermo Fisher Scientific 42i, 43i, 48i and 49i analysers, respectively, and $PM_{2.5}$ was measured with a Thermo 5014i monitor. Surface net solar radiation (SNSR) data for both Beijing and Qingdao were obtained from the European Centre for Medium-Range Weather Forecasts.

Data preprocessing

Data consolidation and interpolation

Meteorological, pollution and PNSD measurements were first consolidated into a unified time series. Measurements recorded within the same minute were combined. Because the Beijing PNSD data were recorded at 1 s resolution, PNSD values were averaged to 1 min intervals to reduce high-frequency numerical fluctuations and align them with the other variables.

Handling missing data

Data integrity is paramount in our analysis. To maintain data integrity and enhance the robustness of our models, we systematically removed rows containing missing values. This step ensures that our modeling efforts are based on complete and reliable data.

Data standardization

Different machine-learning algorithms have different requirements for input scaling. Random Forest and XGBoost are tree-based algorithms and do not require feature standardization, whereas Ridge regression and fully connected neural networks benefit from standardized inputs. For Ridge regression and neural-network modelling, input features were standardized with the StandardScaler function in scikit-learn (38). StandardScaler transforms each feature to have a mean of 0 and a standard deviation of 1. For a sample x, the standardized value z is calculated as: $z = (x - \mu) / \sigma$

where μ is the mean of the training samples, and σ is the standard deviation of the training samples.

Machine learning

Evaluation metrics

We used the coefficient of determination (R^2) and root-mean-square error (RMSE) to evaluate predictive performance. R^2 measures the fraction of variance explained by the model and is defined as one minus the ratio of the residual sum of squares to the total sum of squares. Values closer to 1 indicate better model fit. RMSE measures the typical prediction error in the original units of the target variable. The two metrics are defined as follows:

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

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

Model selection

Models for particle number concentrations at individual particle sizes were developed in Python. We compared four commonly used algorithms: Ridge regression and Random Forest from scikit-learn, XGBoost from the XGBoost library and fully connected neural networks implemented with Keras/TensorFlow (39). Ridge regression is a regularized linear model. Random Forest is an ensemble of decision trees fitted to bootstrap samples of the training data (40). XGBoost is a gradient-boosting algorithm that iteratively combines weak learners and uses regularization to reduce overfitting (37).
The fully connected neural network was a feedforward model with multiple dense layers and was used to capture nonlinear relationships in tabular data.

Model training and hyperparameter optimization

For each particle-size model, the data were randomly split into a training set (80%) and a test set (20%). Hyperparameters were tuned within the training set using grid search with 3-fold cross-validation. In this procedure, the training data were divided into three folds; the model was trained on two folds and validated on the remaining fold, and this process was repeated until each fold had served once as the validation set.

After selecting the best hyperparameters, we trained a final model using the full training set and evaluated its performance on the independent 20% test set using R^2 and RMSE. All models used in this study followed this procedure. The scikit-learn functions train_test_split and GridSearchCV were used for model splitting and hyperparameter optimization (38). This train-test split evaluates out-of-sample reconstruction within the observed campaigns and should not be interpreted as prospective forecasting performance under independent future conditions.

Interpretation notes

SHAP values quantify contributions to model predictions within the measured covariate space and should not be interpreted as direct causal effects. For correlated predictors, attribution can be distributed among covarying variables; therefore, individual feature rankings are interpreted together with grouped predictor classes and physical context. The random train-test split evaluates out-of-sample reconstruction within the observed campaigns rather than prospective forecasting under independent future conditions.

Supplementary Figures and Tables

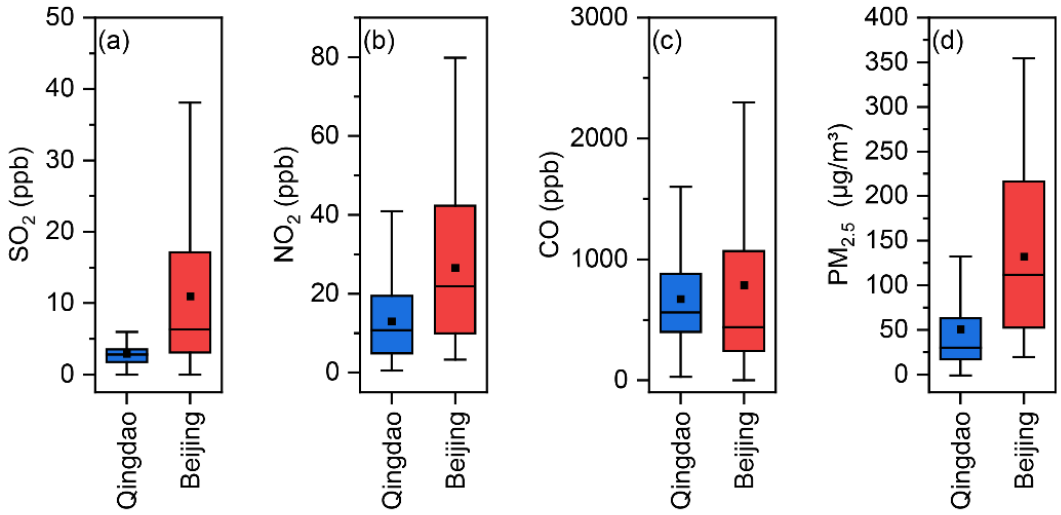

Fig. S1. Comparison of SO_2, NO_2, CO and $PM_{2.5}$ concentrations between Qingdao and Beijing during NPF and non-NPF days. For each box plot, the lower and upper box boundaries indicate the 25th and 75th percentiles, the inner line indicates the median and the circle indicates the mean. Whiskers extend to 1.5 × IQR below the first quartile and above the third quartile.

Supplementary discussion

During the Qingdao observation period, the mean concentrations of SO_2, NO_2, CO and $PM_{2.5}$ were 2.9 ppb, 13.0 ppb, 672.5 ppb and 50.4 μg m^{-3}, respectively. The corresponding values in Beijing were higher: 11.0 ppb for SO_2, 26.6 ppb for NO_2, 787 ppb for CO and 132.1 μg m^{-3} for $PM_{2.5}$. These contrasts support the use of Qingdao and Beijing as relatively cleaner coastal and more polluted inland urban observational cases, respectively, while the main text interprets inter-site differences cautiously because the datasets also differ in observation period and instrumentation.

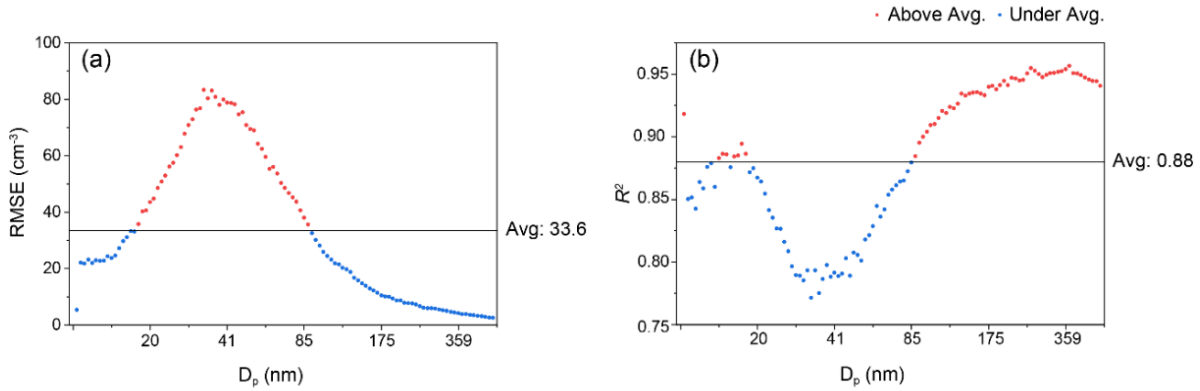

Fig. S2. Performance of XGBoost size-bin models for Qingdao. (a) Distribution of RMSE values across 10–500 nm. RMSE was calculated on the 20% test set. (b) Distribution of R^2 values for the same models on the 20% test set.

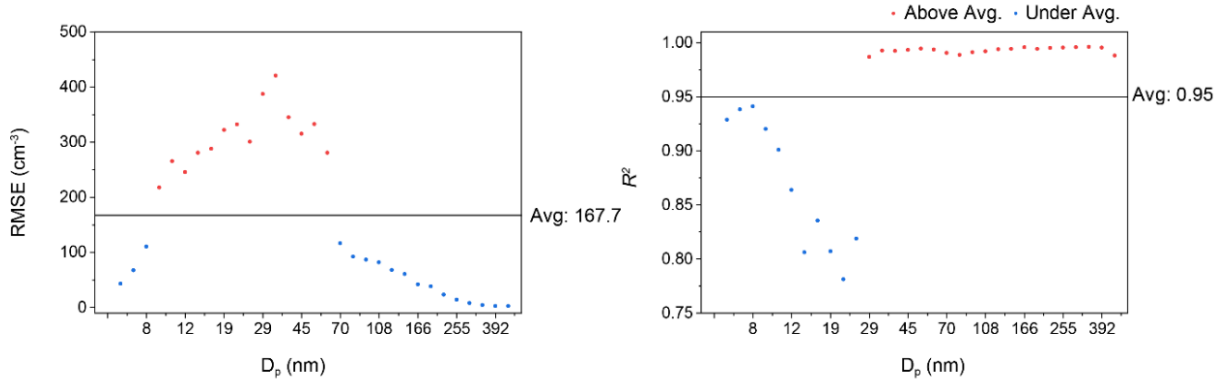

Fig. S3. Performance of XGBoost size-bin models for Beijing. (a) Distribution of RMSE values across 10–500 nm. RMSE was calculated on the 20% test set. (b) Distribution of R^2 values for the same models on the 20% test set.

Supplementary discussion

Figure S2 summarizes the performance of the 109 XGBoost size-bin models used to predict PNSDs in Qingdao. Higher R^2 values and lower RMSE values indicate better test performance. Models for PNCs in the 20–85 nm range had RMSE values that exceeded the average of the 109 models by 33.6 cm^{-3}, and their R^2 values were below the average of 0.95, indicating weaker performance in this highly variable size range. Many observed PNSD values below 20 nm and above 85 nm were close to zero, which reduced prediction difficulty in those ranges and contributed to higher apparent accuracy. Similar patterns were observed for the 31 Beijing PNSD models (Fig. S3). In Beijing, RMSE values for PNCs in the 8–75 nm range exceeded the average by 167.7 cm^{-3}, likely because of stronger short-term variability in that size range. Overall, the Qingdao and Beijing models showed reliable test performance ($R^2 > 0.75$), supporting the subsequent model-interpretation analyses.

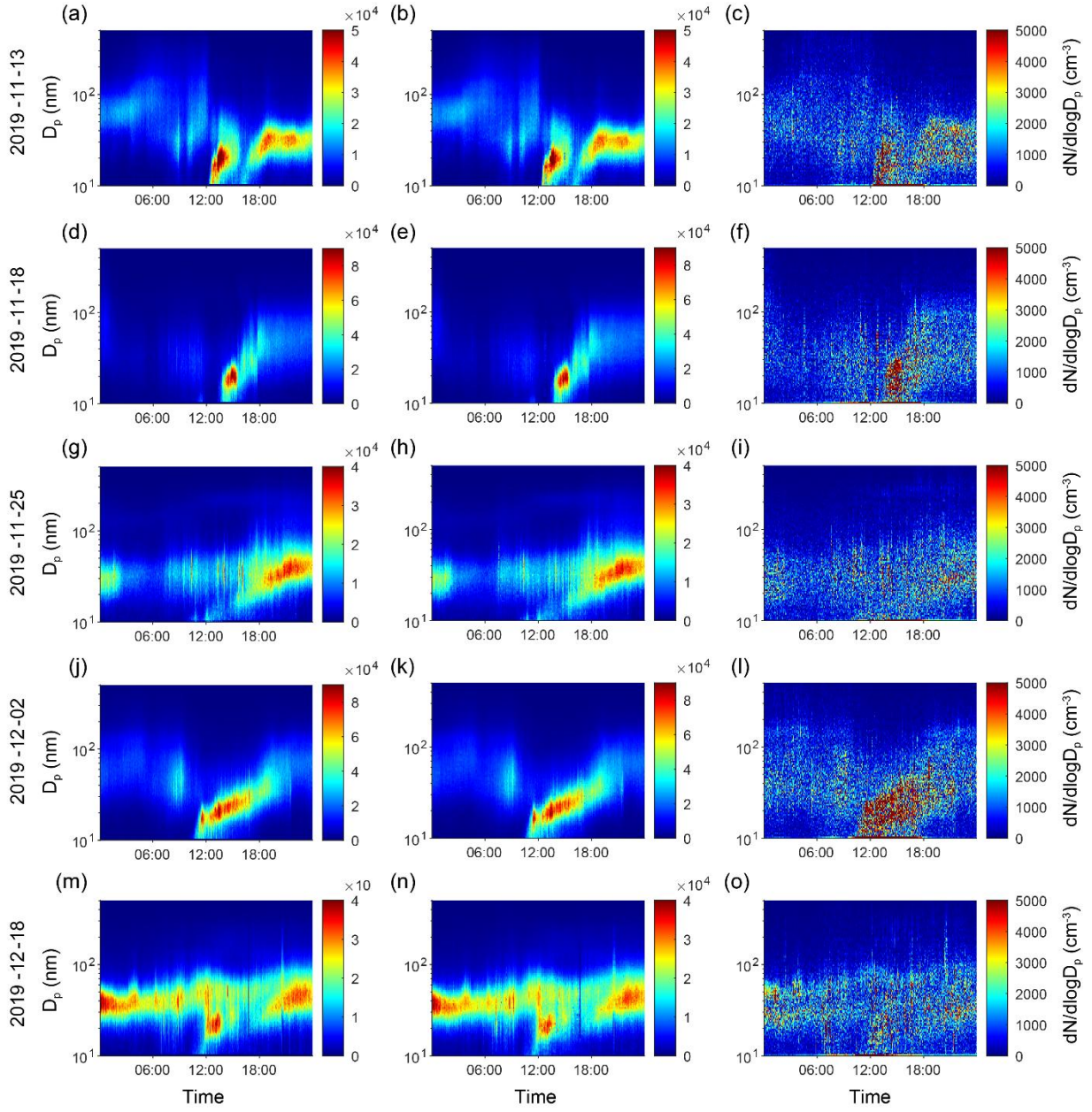

Fig. S4. Comparison between observed particle number size distributions and XGBoost predictions for Qingdao. Panels show observed and predicted PNSDs, together with the corresponding absolute errors, for 13 November (a–c), 18 November (d–f), 25 November (g–i), 2 December (j–l) and 18 December (m–o) 2019.

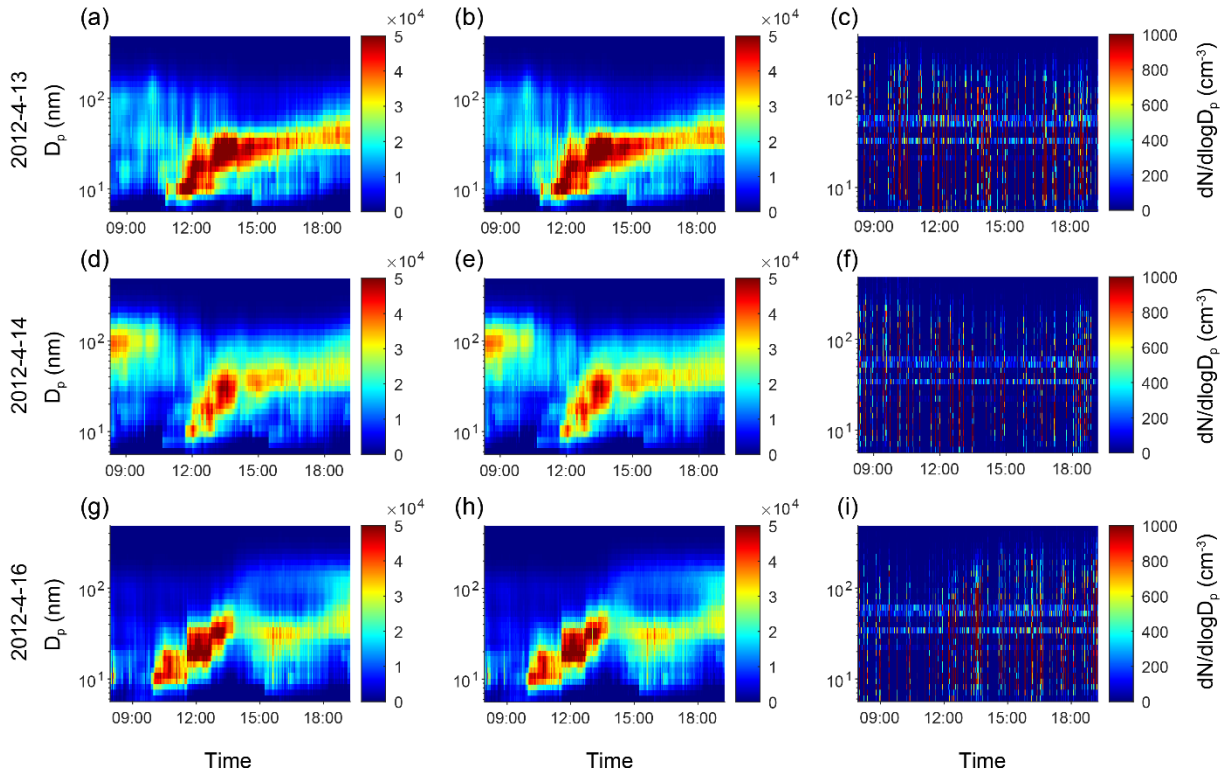

Fig. S5. Comparison between observed particle number size distributions and XGBoost predictions for Beijing. Panels show observed and predicted PNSDs, together with the corresponding absolute errors, for 13 April (a–c), 14 April (d–f) and 16 April (g–i) 2012.

Supplementary discussion

To evaluate size-bin prediction during representative event periods, we compared observed and predicted PNSDs for typical NPF days in Qingdao and Beijing (Figs. S4 and S5). The close agreement between observations and predictions indicates that the multiple size-bin models captured the main temporal and size-dependent variability. The third column in each figure shows absolute errors, calculated as the absolute difference between observed and predicted values for the same time and size bin. Absolute errors were generally below 5,000 cm^{-3}, supporting the use of the trained models for subsequent size-resolved SHAP attribution.

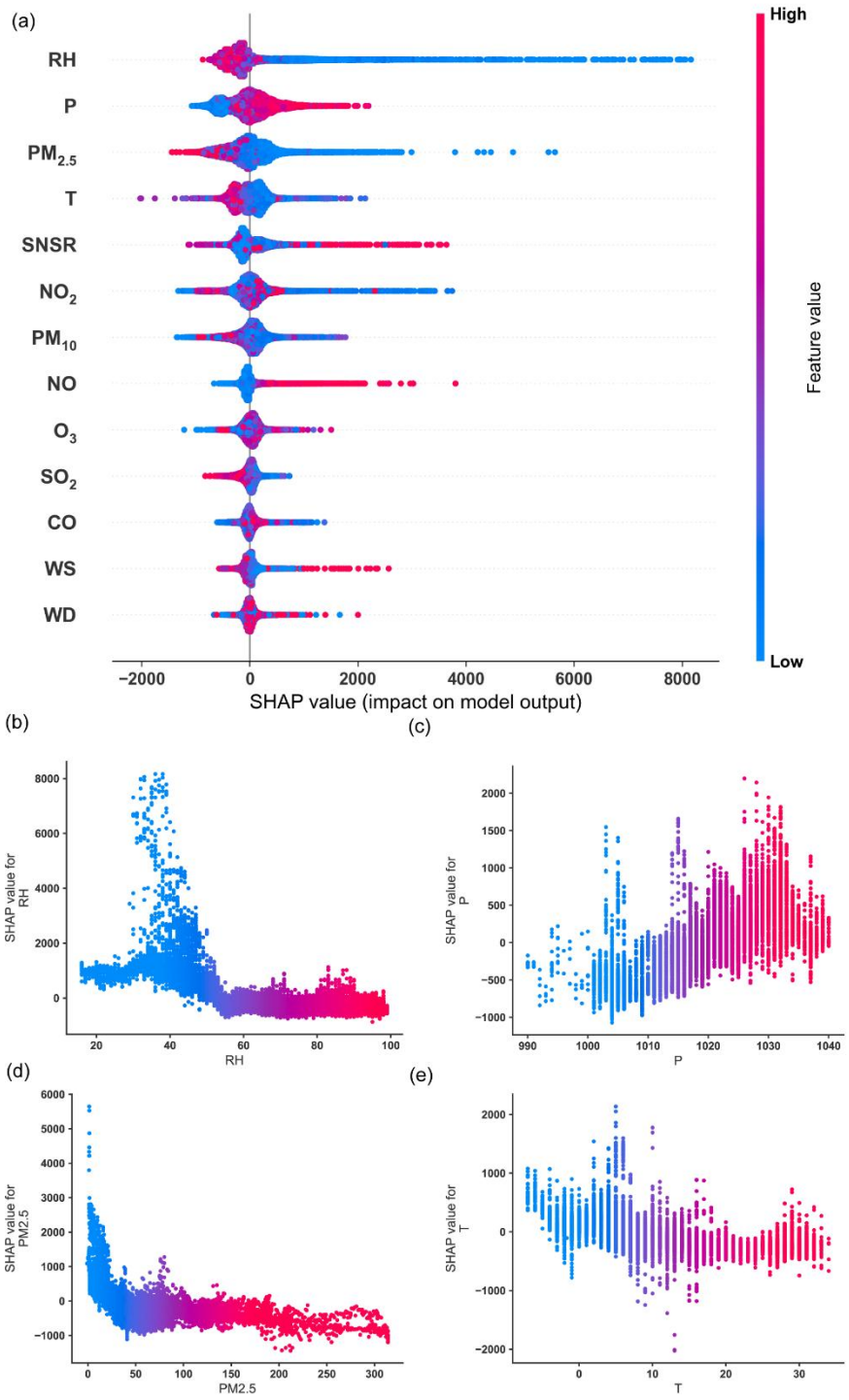

Fig. S6. SHAP dependence plots for the Qingdao nucleation-mode XGBoost model. (a) SHAP summary plot for all input variables, ordered by mean absolute |SHAP| value. Dependence plots show the relationships between SHAP values and (b) RH, (c) P, (d) $PM_{2.5}$ and (e) T.

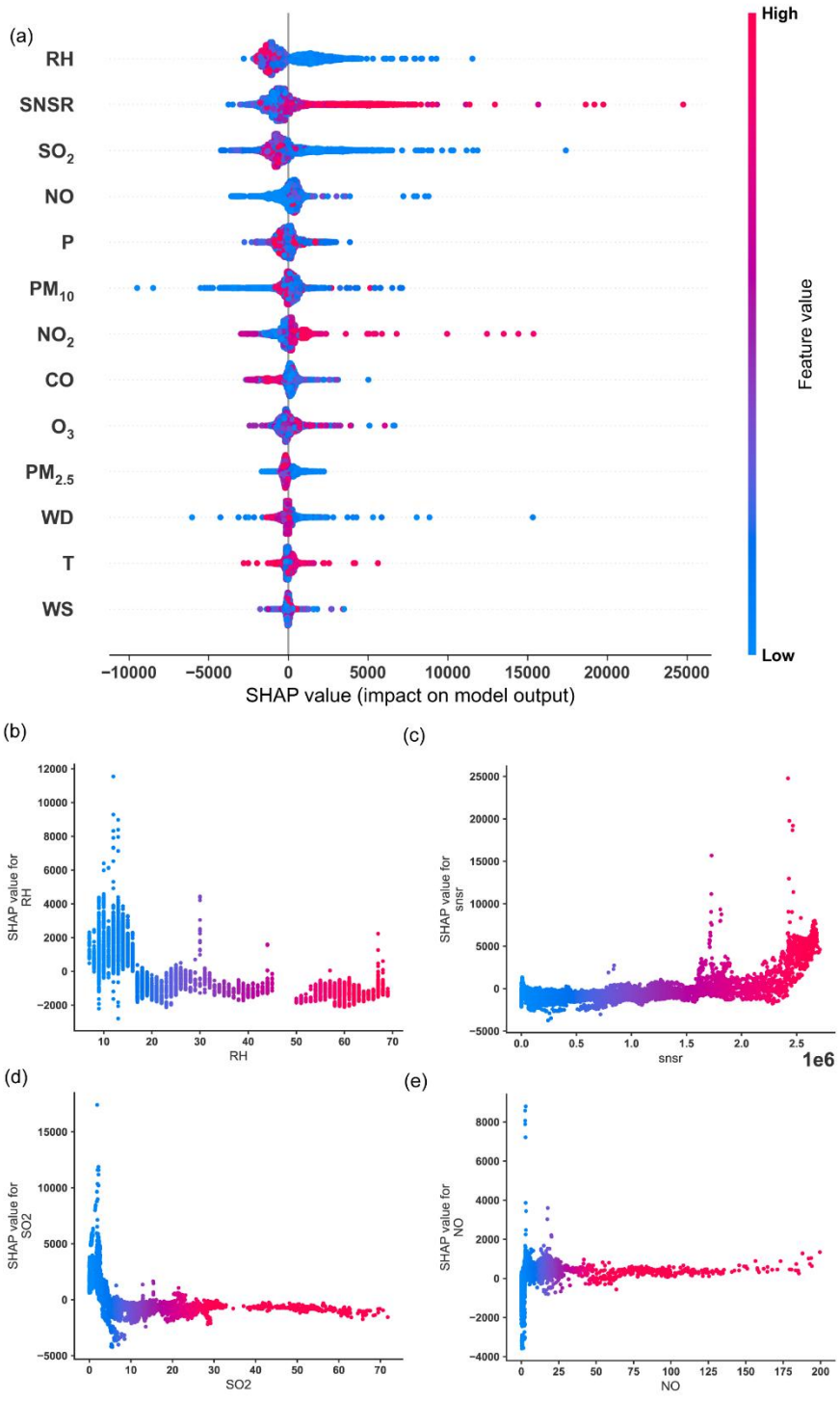

Fig. S7. SHAP dependence plots for the Beijing nucleation-mode XGBoost model. (a) SHAP summary plot for all input variables, ordered by mean absolute |SHAP| value. Dependence plots show the relationships between SHAP values and (b) RH, (c) SNSR, (d) SO_2 and (e) NO.

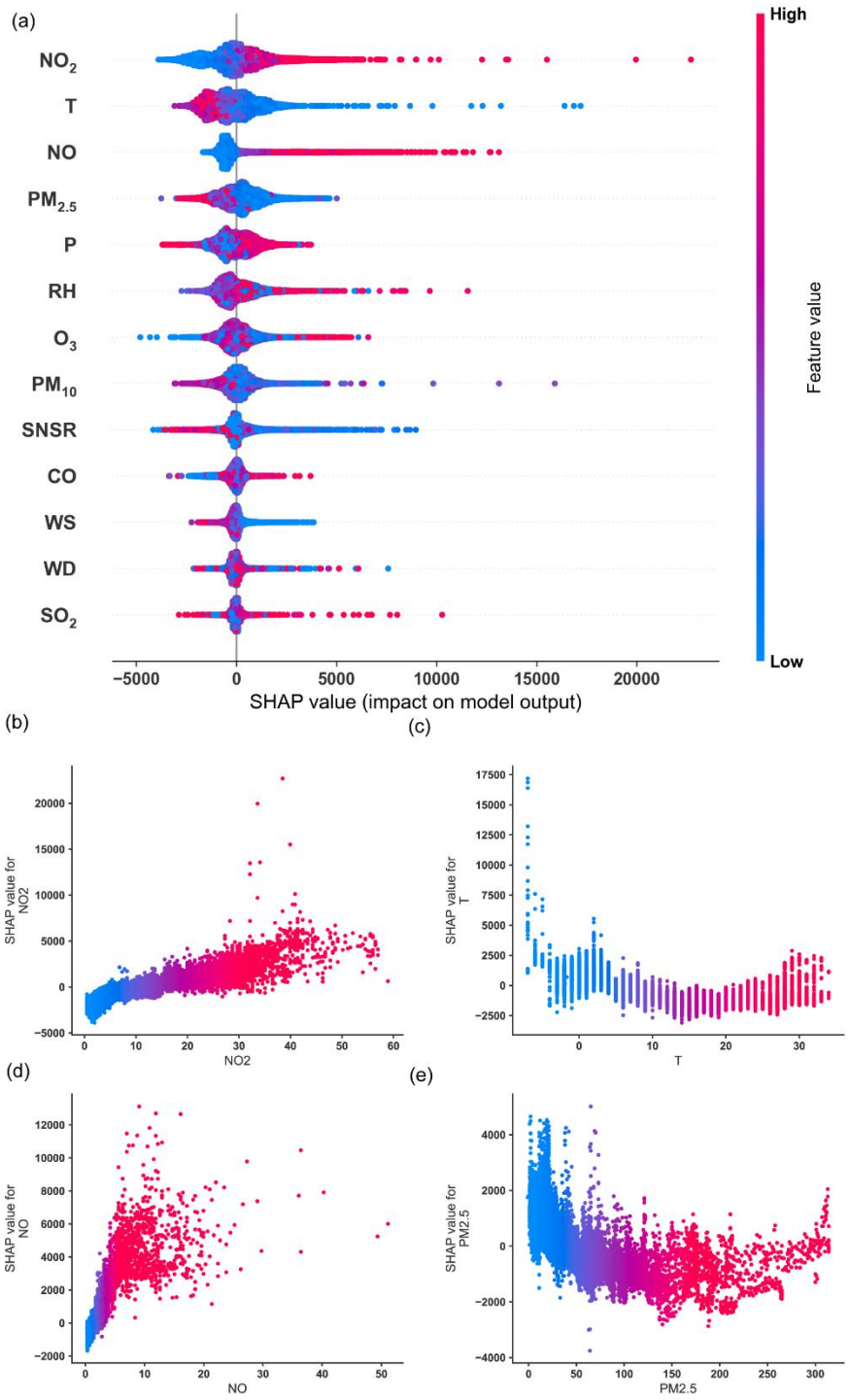

Fig. S8. SHAP dependence plots for the Qingdao Aitken-mode XGBoost model. (a) SHAP summary plot for all input variables, ordered by mean absolute |SHAP| value. Dependence plots show the relationships between SHAP values and (b) NO_2, (c) T, (d) NO and (e) $PM_{2.5}$.

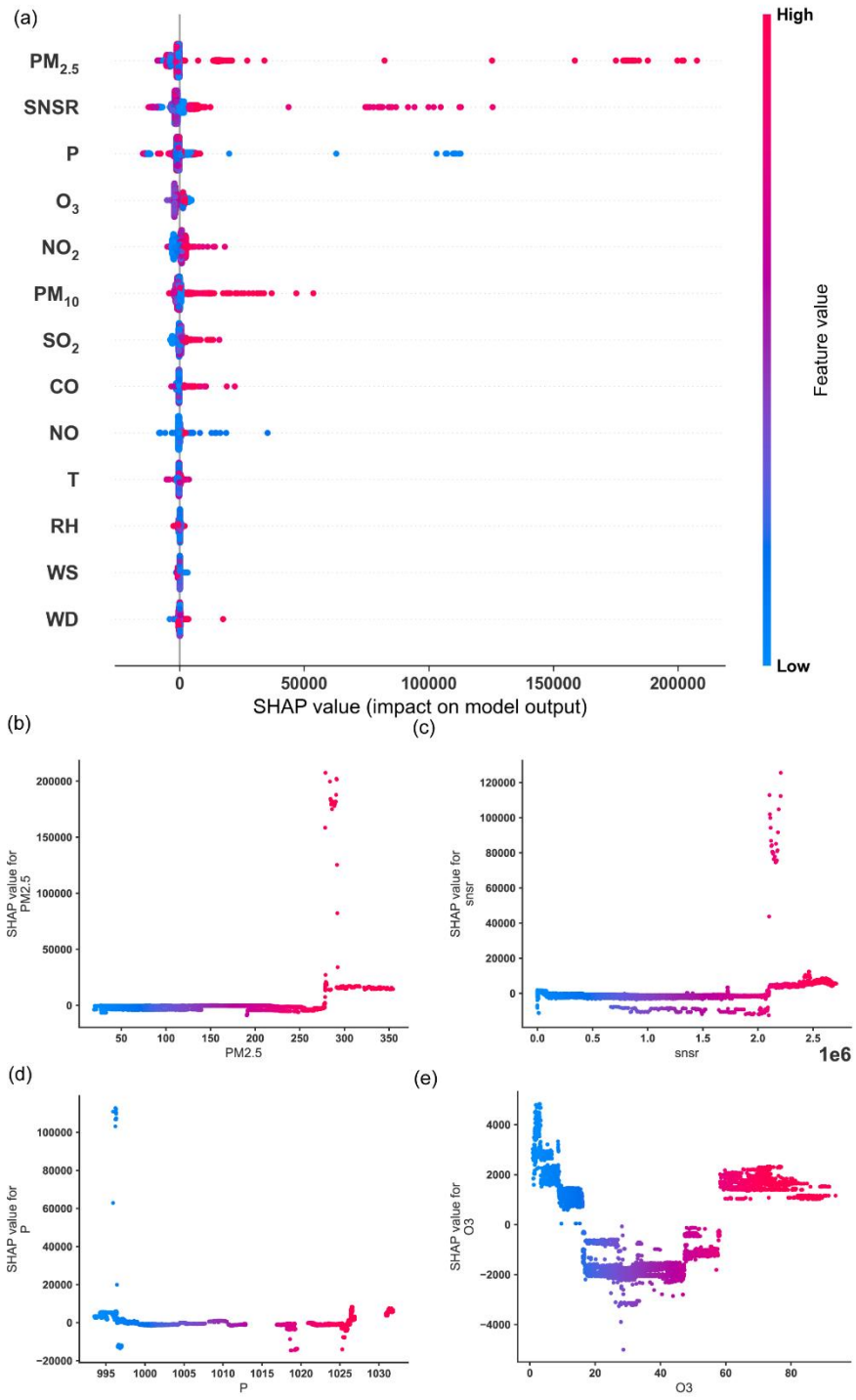

Fig. S9. SHAP dependence plots for the Beijing Aitken-mode XGBoost model. (a) SHAP summary plot for all input variables, ordered by mean absolute |SHAP| value. Dependence plots show the relationships between SHAP values and (b) $PM_{2.5}$, (c) SNSR, (d) P and (e) O_3.

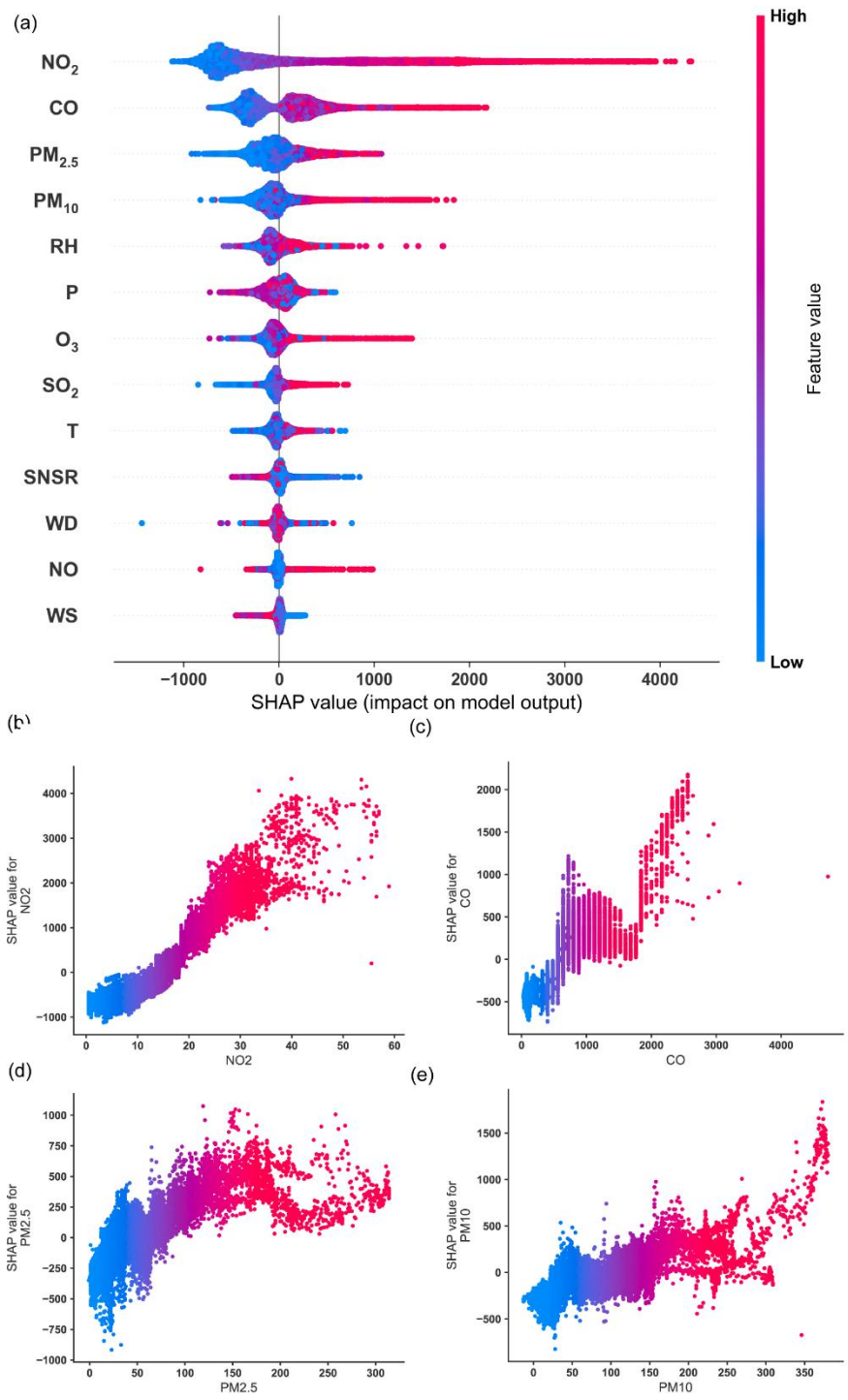

Fig. S10. SHAP dependence plots for the Qingdao accumulation-mode XGBoost model. (a) SHAP summary plot for all input variables, ordered by mean absolute |SHAP| value. Dependence plots show the relationships between SHAP values and (b) NO_2, (c) CO, (d) $PM_{2.5}$ and (e) PM_{10}.

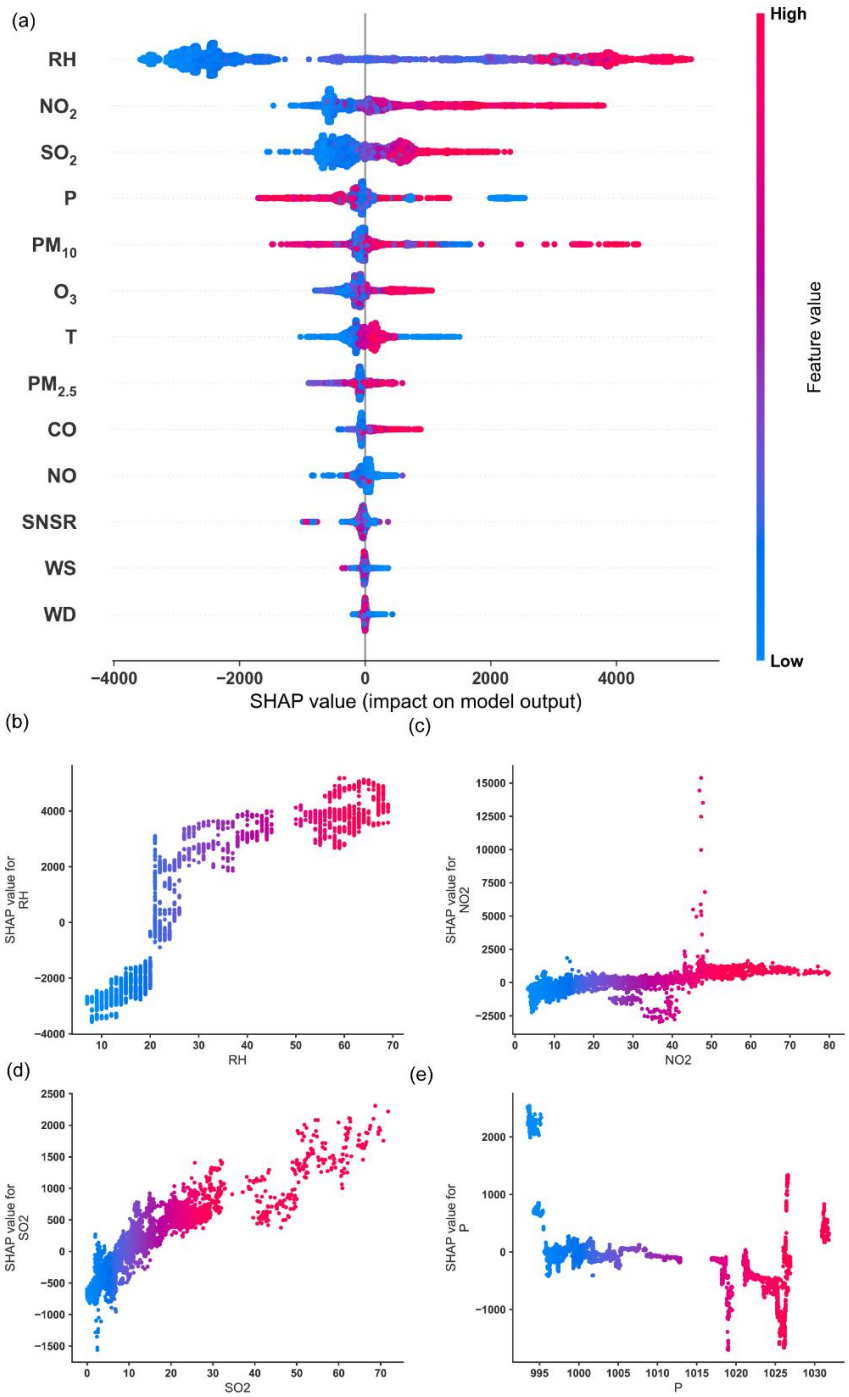

Fig. S11. SHAP dependence plots for the Beijing accumulation-mode XGBoost model. (a) SHAP summary plot for all input variables, ordered by mean absolute |SHAP| value. Dependence plots show the relationships between SHAP values and (b) RH, (c) NO_2, (d) SO_2 and (e) P.

Supplementary discussion

The SHAP dependence plots visualize how selected variables contribute to model predictions within the trained XGBoost models. Each plot shows the variable value on the x-axis and the corresponding SHAP value on the y-axis. The plots therefore provide model-specific attribution patterns similar in purpose to

partial-dependence analysis, but they should not be interpreted as isolated causal effects because predictors can covary.

In Qingdao (Fig. S6), RH showed negative SHAP contributions to nucleation-mode PNC under low-humidity conditions (RH < 40%). Pressure was positively associated with nucleation-mode PNC, indicating the contribution of pressure-related atmospheric states. $PM_{2.5}$ showed negative contributions to nucleation-mode PNC, consistent with pre-existing particles acting as sinks. For Aitken-mode particles, $PM_{2.5}$ also showed negative contributions, whereas NO and NO_2 showed positive contributions (Fig. S8). Accumulation-mode PNC was positively associated with CO, NO_2, SO_2 and O_3 concentrations (Fig. S10).

In Beijing (Fig. S7), RH showed negative SHAP contributions to nucleation-mode PNC under low-humidity conditions (RH < 20%). SNSR showed positive contributions to nucleation-mode PNC, consistent with the role of solar radiation in nucleation-mode particle variability. CO and NO_2 showed positive contributions to Aitken-mode PNC, whereas RH, CO and O_3 showed positive contributions to accumulation-mode PNC (Figs. S9 and S11). Together, these results support the main-text interpretation that meteorological and pollution-related variables have size- and mode-dependent attribution patterns.

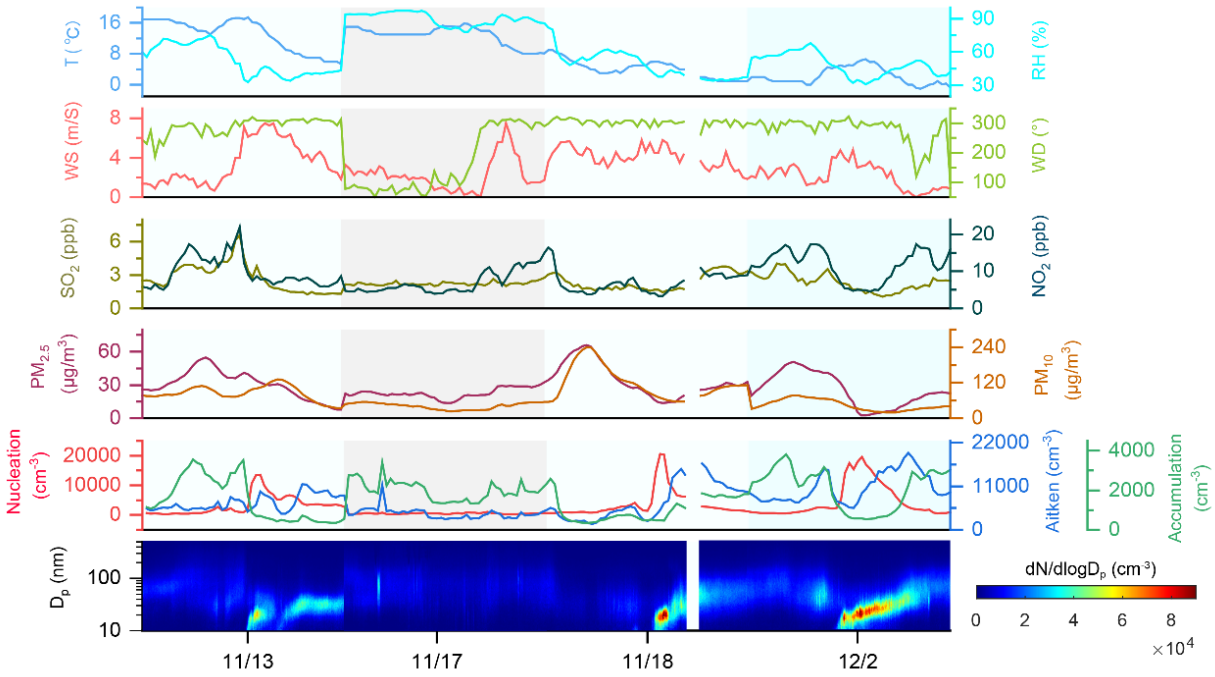

Fig. S12. Particle number size distributions and associated variables for three NPF event days and one non-event day in Qingdao. Time series are shown for meteorological parameters, pollutants, and nucleation-, Aitken- and accumulation-mode particle number concentrations on 13 November, 17 November, 18 November and 2 December 2019. Event days are indicated with light blue shading, and the non-event day is indicated with light grey shading.

Supplementary discussion

To examine atmospheric conditions during NPF events, we plotted daily variations in meteorological conditions and pollutant concentrations for representative NPF and non-NPF days in Qingdao (Fig. S13). November 13, November 18 and December 2 were NPF event days, and November 17 was a non-event

day. The events occurred after decreasing RH, and their onset coincided with the lowest RH. The lower RH on event days than on the non-event day indicates that NPF events in Qingdao were associated with low humidity, consistent with observations in Xiamen, another coastal city in China (42). The dominant wind direction on event days was northwest, whereas the non-event day was dominated by east and east-northwest winds. SO_2, NO_2, $PM_{2.5}$ and PM_{10} concentrations decreased sharply before NPF events, indicating that these events occurred under relatively low-pollution conditions. Before event onset, accumulation-mode PNC decreased sharply. When NPF events occurred, nucleation-mode PNC increased rapidly and then declined as particles grew through condensation and coagulation (12). The subsequent increase in Aitken-mode PNC reflects growth of nucleation-mode particles, followed by increased accumulation-mode PNC. These observations describe event-associated environmental states; the magnitude of each contribution is quantified using explainable machine learning in the main text.

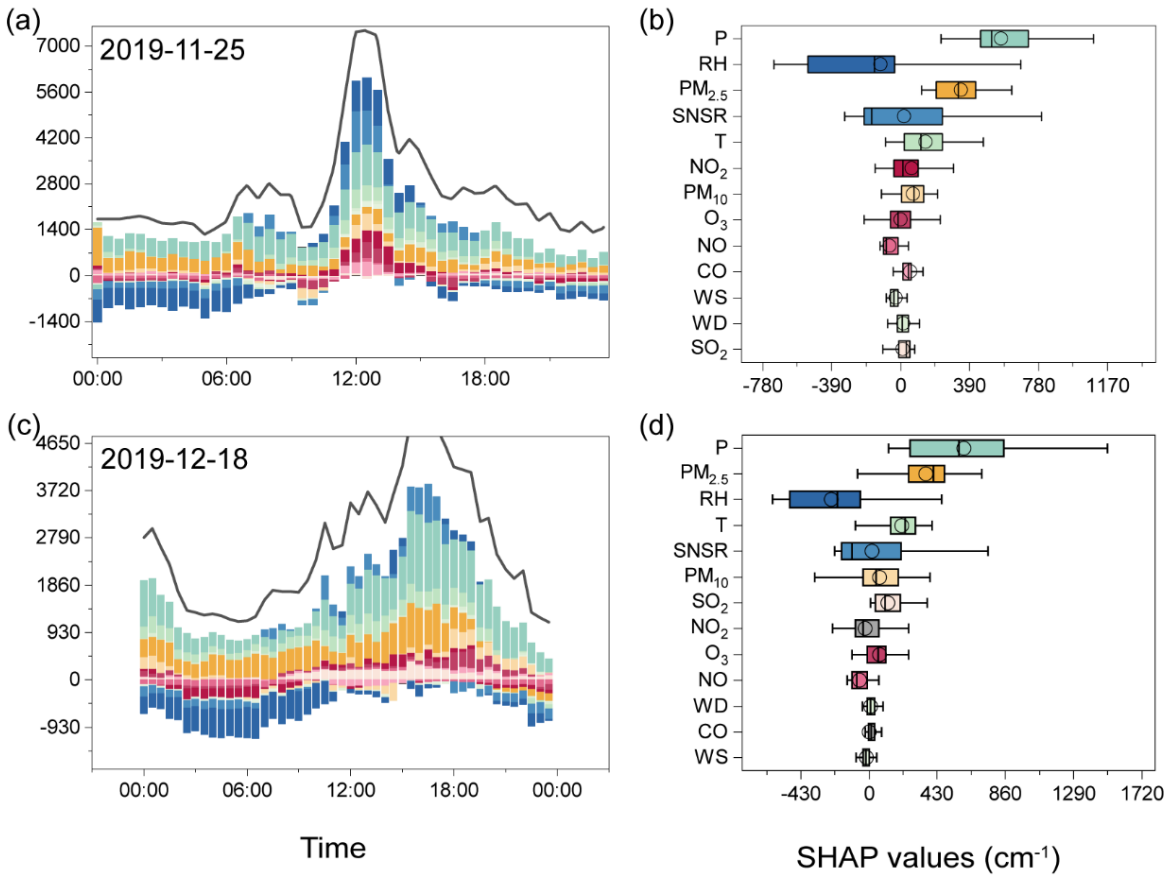

Fig. S13. Time-resolved SHAP values from XGBoost models for nucleation-mode particle number concentration. Factor-specific SHAP values, averaged every 30 min, are shown for 25 November and 18 December in Qingdao. Box plots summarize the same data ordered by mean absolute |SHAP| value. For each box plot, the lower and upper box boundaries indicate the 25th and 75th percentiles, the inner line indicates the median, the circle indicates the mean and whiskers extend to 1.5 × IQR below the first quartile and above the third quartile.

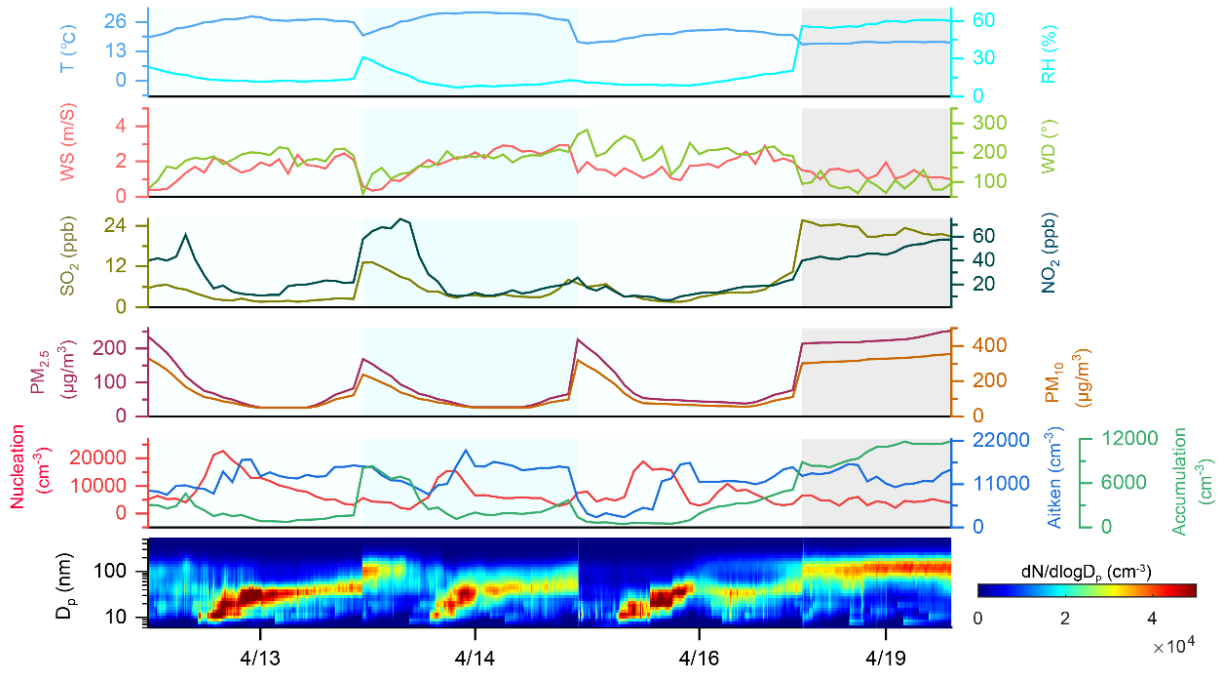

Fig. S14. Particle number size distributions and associated variables for three NPF event days and one non-event day in Beijing. Time series are shown for meteorological parameters, pollutants, and nucleation-, Aitken- and accumulation-mode particle number concentrations on 13 April, 14 April, 16 April and 19 April 2012. Event days are indicated with light blue shading, and the non-event day is indicated with light grey shading.

Supplementary discussion

Daily variations in meteorological conditions and pollutant concentrations during representative NPF and non-NPF days in Beijing are shown in Fig. S15. April 13, April 14 and April 16 were classified as NPF event days, whereas April 19 was classified as a non-event day. RH was notably lower on event days, remaining below 30%, consistent with previous observations in Beijing (43–45). Concentrations of SO_2, NO_2, $PM_{2.5}$ and PM_{10} also decreased sharply during NPF events, supporting the main-text interpretation that observable nucleation-mode bursts were associated with meteorological states favorable for NPF and reduced pre-existing particle loading.

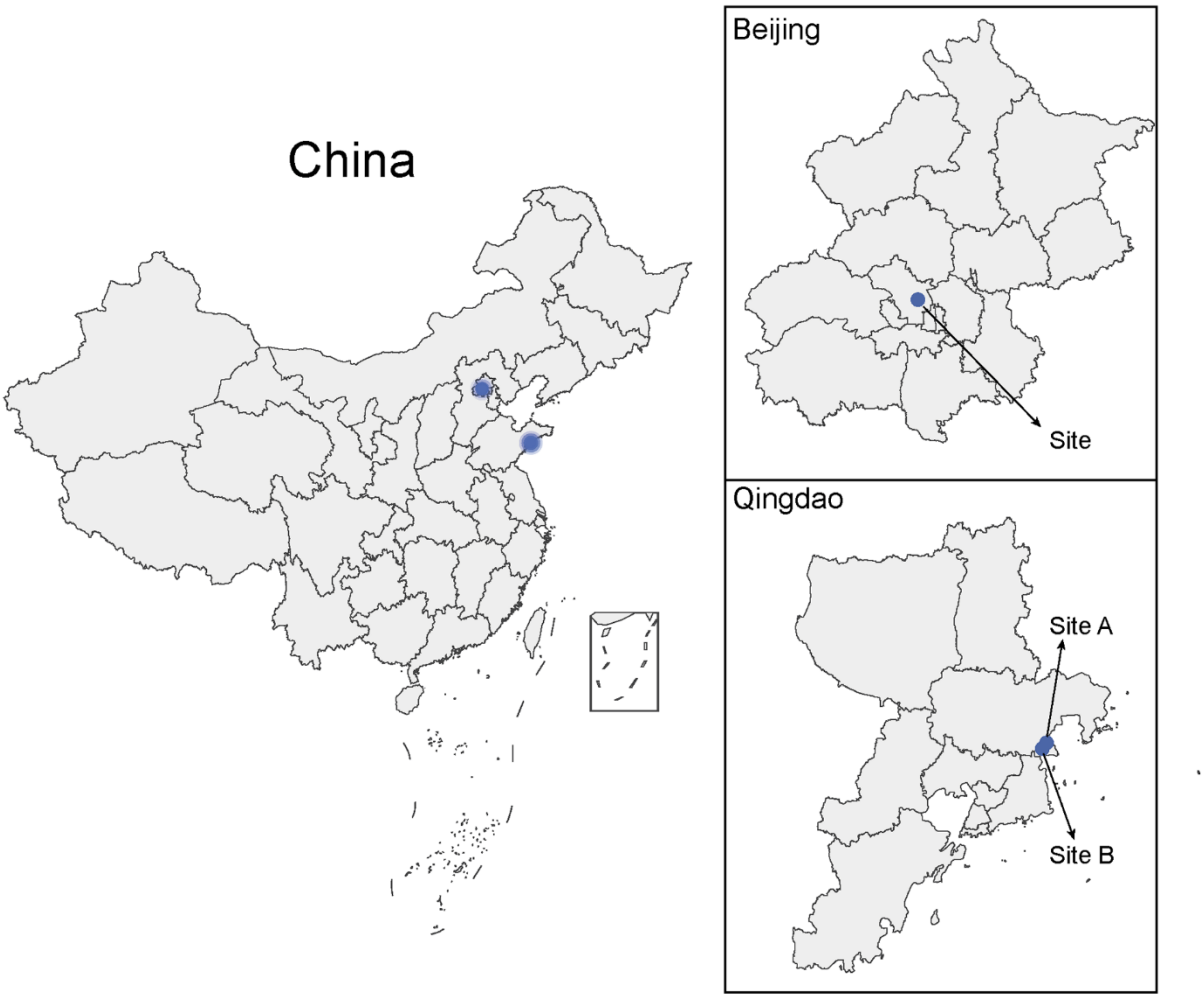

Fig. S15. Sampling sites in Qingdao and Beijing.

Table S1. Predictor variables used in the machine-learning models, including units and brief descriptions.

Variable	Unit	Description
RH	%	Relative humidity
T	°C	Air temperature
P	hPa	Air pressure
WS	m/s	Wind speed
WD	°	Wind direction
SNSR	J m^{-2}	Surface net solar radiation
NO_2	ppb	The concentration of nitrogen dioxide
NO	ppb	The concentration of nitrogen monoxide
CO	ppb	The concentration of carbon monoxide
O_3	ppb	The concentration of ozone
SO_2	ppb	The concentration of sulfur dioxide
$PM_{2.5}$	μg m^{-3}	Particulate matter with diameter < 2.5 μm
PM_{10}	μg m^{-3}	Particulate matter with diameter < 10 μm

Supplementary discussion

We considered six meteorological variables (RH, T, P, WS, WD and SNSR) because these variables can influence NPF occurrence, frequency and intensity. RH affects condensation and evaporation involving gases and aerosol particles. Classical theories suggest that water vapour can participate in nucleation (46), whereas field and chamber studies often show that high RH inhibits NPF (43, 47), indicating that the role of RH is context dependent. Pressure can affect the concentration and activity of gas-phase species involved in NPF reactions, although its direct influence is often secondary to its covariation with temperature and humidity. WS and WD affect transport and dispersion of aerosol particles and precursor gases and can influence mixing between polluted and cleaner air masses. Solar radiation drives photochemical reactions that produce reactive intermediates and low-volatility vapours involved in NPF (48). We considered five trace gases (SO_2, CO, NO, NO_2 and O_3) because they are commonly used to evaluate air quality and identify pollution sources in China (36). SO_2 is an important precursor of H_2SO_4 formation (49, 50). CO is commonly used as a primary-pollution indicator. NO and NO_2 reflect urban nitrogen chemistry and traffic-related combustion influence, and NO_2 can contribute to HNO_3 formation through atmospheric reactions (51–53). HNO_3 can contribute to particle growth, for example by combining with ammonia through acid-base stabilization to form ammonium nitrate particles (54). O_3 is a secondary pollutant and an indicator of atmospheric oxidative capacity; it can also contribute indirectly to NPF through reactions with volatile organic compounds. High $PM_{2.5}$ and PM_{10} concentrations can enhance coagulation and scavenging losses of newly formed particles, thereby suppressing NPF (20). These variables are therefore interpreted in the main text as measured meteorological or pollution-related proxies, not as isolated mechanistic agents.

Table S2. R^2 and RMSE values for the four machine-learning models on the Qingdao test set. To compare model performance across particle size, one size bin was selected every ten bins from the 109 Qingdao size bins.

Diameter (nm)	**RR**		**FCNN**		**RF**		**XGBOOST**	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
10.94	0.21	50.06	0.77	24.95	0.89	18.61	0.85	21.68
15.68	0.21	74.76	0.86	28.88	0.91	25.84	0.88	29.68
22.47	0.25	110.74	0.69	65.02	0.87	46.45	0.84	50.77
32.2	0.28	142.86	0.75	80.57	0.81	73.51	0.79	76.76
46.14	0.29	141.25	0.70	94.75	0.83	69.57	0.80	74.60
66.12	0.42	102.70	0.77	61.09	0.87	49.36	0.84	53.61
94.75	0.58	61.33	0.86	35.59	0.91	27.79	0.90	30.11
135.77	0.70	35.40	0.91	19.17	0.95	15.11	0.93	16.69
194.56	0.77	18.50	0.93	10.12	0.95	8.36	0.94	9.34
278.81	0.82	11.09	0.95	5.85	0.96	5.28	0.95	5.90
399.54	0.84	6.52	0.95	3.78	0.96	3.31	0.95	3.59
Average	0.49	68.66	0.83	39.07	0.90	31.20	0.88	21.68

Table S3. R^2 and RMSE values for the four machine-learning models on the Beijing test set. To compare model performance across particle size, one size bin was selected every three bins from the 31 Beijing size bins.

Diameter (nm)	**RR**		**FCNN**		**RF**		**XGBOOST**	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
10.8	0.38	663.97	0.81	372.88	0.93	229.44	0.90	265.46
16.5	0.32	585.69	0.52	642.44	0.84	281.94	0.84	287.99
25.5	0.31	585.24	0.76	303.72	0.84	282.96	0.82	300.58
39.2	0.05	3874.27	0.86	1403.54	0.99	470.23	0.99	344.71
60.4	0.07	3426.56	0.85	1476.63	0.98	473.55	0.99	280.87
93.1	0.64	553.43	0.96	173.04	1.00	60.61	0.99	87.16
143.3	0.80	355.08	0.98	110.27	1.00	39.21	0.99	60.88
220.7	0.86	123.66	0.98	49.86	1.00	15.67	1.00	23.58
339.8	0.85	26.68	0.99	8.41	1.00	3.85	1.00	4.36
Average	0.48	1132.73	0.86	504.53	0.95	206.39	0.95	183.96

Table S4. Performance metrics of XGBoost models trained to predict particle number concentrations in each mode in Beijing and Qingdao. Models were trained using the training set, and R^2 and RMSE were evaluated on the test set.

Mode	Qingdao		Beijing	
	R^2	RMSE	R^2	RMSE
Nucleation	0.89	642.58	0.83	2077.30
Aitken	0.81	2175.55	0.98	2857.05
Accumulation	0.96	323.18	0.99	285.75

Supplementary discussion

Six XGBoost models were trained to predict particle number concentrations in three particle modes for Qingdao and Beijing. Each dataset was split into an 80% training set and a 20% test set. Hyperparameters were tuned using 3-fold cross-validation on the training set. The parameter ranges used in GridSearchCV were n_estimators: [100, 200, 300, 400, 500], max_depth: [3, 5, 7, 9, 11], min_child_weight: [3, 4, 5, 6], gamma: [0, 0.1, 0.2, 0.3] and subsample: [0.6, 0.7, 0.8, 0.9]. The parameters are defined below.

- n_estimators: number of trees in the XGBoost model. Increasing this value can improve performance but also increases computational cost.
- max_depth and min_child_weight: maximum tree depth and minimum number of instances required in each child node, respectively. These parameters control tree complexity and help reduce overfitting.
- gamma: minimum loss reduction required to split a leaf node. This regularization parameter penalizes splits that do not substantially improve model fit.
- subsample: fraction of samples used to train each tree. Lower values increase stochasticity and may improve generalization in noisy datasets.

After selecting the hyperparameters that optimized cross-validation performance on the training set, the XGBoost model was trained using the full training set. Final performance was evaluated on the 20% test set using R^2 and RMSE. These mode-specific models had test R^2 values higher than 0.81 (Table S4), indicating strong predictive performance within the observed campaigns.

Supplementary References

41. Zimmerman, N. et al. A source-independent empirical correction procedure for the fast mobility and engine exhaust particle sizers. *Atmos. Environ.* **100**, 178–184 (2015).

42. Wang, J. et al. Particle number size distribution and new particle formation in Xiamen, the coastal city of Southeast China in wintertime. *Sci. Total Environ.* **826**, 154208 (2022).

43. Cai, R. et al. Aerosol surface area concentration: a governing factor in new particle formation in Beijing. *Atmos. Chem. Phys.* **17**, 12327–12340 (2017).

44. Shang, D. et al. New particle formation occurrence in the urban atmosphere of Beijing during 2013–2020. *J. Geophys. Res. Atmos.* **128**, e2022JD038334 (2023).

45. Peng, Y. et al. Different characteristics of new particle formation events at two suburban sites in northern China. *Atmosphere* **8**, 258 (2017).

46. Duplissy, J. et al. Effect of ions on sulfuric acid-water binary particle formation: 2. Experimental data and comparison with QC-normalized classical nucleation theory. *J. Geophys. Res. Atmos.* **121**, 1752–1775 (2016).

47. Li, X., Chee, S., Hao, J., Abbatt, J. P. D., Jiang, J. & Smith, J. N. Relative humidity effect on the formation of highly oxidized molecules and new particles during monoterpene oxidation. *Atmos. Chem. Phys.* **19**, 1555–1570 (2019).
48. Kulmala, M. et al. Direct observations of atmospheric aerosol nucleation. *Science* **339**, 943–946 (2013).
49. Boy, M. et al. Sulphuric acid closure and contribution to nucleation mode particle growth. *Atmos. Chem. Phys.* **5**, 863–878 (2005).
50. Xiao, S. et al. Strong atmospheric new particle formation in winter in urban Shanghai, China. *Atmos. Chem. Phys.* **15**, 1769–1781 (2015).
51. Ramazan, K. et al. New experimental and theoretical approach to the heterogeneous hydrolysis of NO_2: key role of molecular nitric acid and its complexes. *J. Phys. Chem. A* **110**, 6886–6897 (2006).
52. Liu, L. et al. The role of nitric acid in atmospheric new particle formation. *Phys. Chem. Chem. Phys.* **20**, 17406–17414 (2018).
53. Li, P. et al. Formation of nitrogen-containing gas phase products from the heterogeneous (photo)reaction of NO_2 with gallic acid. *Commun. Chem.* **6**, 198 (2023).
54. Wang, M. et al. Rapid growth of new atmospheric particles by nitric acid and ammonia condensation. *Nature* **581**, 184–189 (2020).