Supplement of

Fireworks promote the secondary formation of atmospheric imidazoles rather than their direct emission: Insights from firework source samples and machine learning

Chen Chen[1], Weibo Li[1], Jierui Hou[1], Qizong Wang[1], Xiaodi Liu[1,2,3], Can Wu[3], Xin Zhang[1,2], Zhanfang Hou[1,2], Jingjing Meng[1,2*]

[1] School of Geography and Environment, Liaocheng University, Liaocheng 252000, China

[2] Institute of Huanghe Studies, Liaocheng University, Liaocheng 252000, China

[3] School of Geographic Sciences, Key Lab of Geographic Information Science of the Ministry of Education, East China Normal University, Shanghai 200241, China

*Corresponding author: Jingjing Meng

E-mail address: mengjingjing@lcu.edu.cn

Text S1. Model construction, parameter setting, and applicability assessment

This study used field observations collected in the fireworks (FW) episode, comprising 48 valid hourly samples. With the first 20 observations used as the initial training window, the walk-forward evaluation produced approximately 28 one-step-ahead out-of-sample predictions. Counts of imidazole-containing particles were used as the response variable, and 10 physicochemical parameters were selected as predictors, including relative humidity (RH), corrected aerosol acidity (R_{ra}'), O_3, nitrate, sulfate, ammonium, aerosol liquid water content (ALWC), NO_x, carbonyl compounds, and amines. Three proxy variables derived from time, namely Hour, Cos_Hour, and Sin_Hour, were further added to the predictor matrix to capture temporal decay at the episode scale and diurnal cyclic variations in aerosol chemistry, even though they were not part of the original physicochemical variable set. Incorporation of these temporal terms brings two major benefits: first, Hour encodes the overall temporal decay trend of total imidazoles across the fireworks-influenced episode; second, Cos_Hour and Sin_Hour decompose diurnal periodic signals to represent cyclical meteorological-photochemical forcing (e.g., boundary-layer evolution, diurnal photochemical activity). Adding these temporal proxies allows the model to separate physical-driven temporal variations from chemical precursor effects and improves overall predictive capacity without requiring additional direct chemical measurements. A multi-model comparison framework was constructed, and predefined attribution rules were established to identify key drivers under small-sample conditions. With these three temporal features included, the total number of predictors reached 13; the sample-to-variable ratio was approximately 4, below the recommended threshold for conventional regression models; hence this was treated as a typical small-sample regression problem. Therefore, a modeling strategy combining regularization and latent-variable dimensionality reduction was adopted to mitigate overfitting risk, and generalization performance was interpreted with appropriate caution. Corresponding visualizations are shown in Fig. S6, Fig. S7, and Fig. 6. All input features (both physicochemical predictors and time-derived Hour, Cos_Hour, Sin_Hour) were Z-score standardized. In the walk-forward procedure, the scaler was fitted only on the training

window at each step and then applied to the next prediction point, so no future information entered the standardization. To alleviate the highly skewed distribution of the target variable and heteroscedasticity originating from high-concentration samples, total imidazole concentrations were natural log-transformed, while model performance was also evaluated on the original concentration scale for direct comparison. Model performance was evaluated using a rolling-origin (walk-forward) one-step-ahead scheme. Starting from the 20th observation, each model was refitted using all preceding observations and then used to predict the next hour, so that all prediction points were strictly out-of-sample and no future information leaked into model training. Model evaluation simultaneously reported R^2, RMSE, and MAE for both original-scale and log-transformed outputs to enable multidimensional performance comparison. Quantitative test-set performance statistics for all models are summarized in Fig. S6.

Six types of base models and two ensemble schemes were constructed. The linear models included Ridge regression, Huber robust regression, and partial least squares regression (PLSR); the nonlinear models included Gaussian process regression (GPR), support vector regression (SVR), and gradient boosting regression (GBR). On this basis, a simple-average ensemble (Ensemble_Old) was constructed as the control, and an inverse-MSE sliding-window weighted dynamic ensemble (Ensemble_New) was developed. Model hyperparameters were set to conservative, literature-informed values to limit overfitting under the small-sample condition; no data-driven hyperparameter tuning was performed, and the walk-forward prediction points were used only for out-of-sample evaluation. Ridge regression introduced L_2 regularization to mitigate parameter instability caused by feature collinearity. Huber robust regression used epsilon = 1.35 and a combined loss to reduce the influence of high-concentration outliers on parameter estimation. PLSR compressed collinear features through latent-variable decomposition; its standardized regression coefficients were used for feature interpretation. GPR used an RBF kernel and provided both predictive means and 95% prediction intervals for uncertainty characterization. SVR and GBR employed an RBF kernel and a tree ensemble structure, respectively, to capture potential nonlinear response relationships among variables. The dynamic weighted ensemble

Ensemble_New calculated normalized weights from the reciprocal MSE of each base model within a sliding window, adaptively adjusting the contributions of base models under different conditions. Ensemble_Old directly performed an equal-weight simple average of the four base models (Ridge, PLSR, GPR, and SVR).

On the original scale, Ridge regression was the best-performing single model under walk-forward out-of-sample evaluation (Fig. S6a–c; $R^2 = 0.86$, RMSE = 57.39, MAE = 44.04); after logarithmic transformation, its log-scale R^2 increased to 0.867. GPR ($R^2 = 0.84$) and Ensemble_New ($R^2 = 0.83$) ranked next, followed by Huber ($R^2 = 0.83$) and PLSR ($R^2 = 0.73$), whereas nonlinear models such as GBR ($R^2 = 0.68$) and SVR ($R^2 = 0.619$) did not exhibit the expected advantages. Log transformation notably improved the walk-forward R^2 for most linear models and GBR, with increments ranging from approximately 0.03 to 0.12 (e.g., PLSR +0.070 and GBR +0.116), confirming its effectiveness in mitigating the skewness and high-concentration heteroscedasticity of total imidazoles for these algorithms. However, the improvement for Ridge was marginal (only +0.011), and the performance of GPR (−0.020) and SVR (−0.053) slightly decreased on the log scale. The best model achieved $R^2 > 0.85$, with a large effect size, exceeding levels commonly reported in comparable machine learning studies in this field. Under the present observation conditions, the model showed stable predictive capability. This conclusion was further supported by residual diagnostics (Fig. S6d, e) and segmented diagnostics (Fig. S6f), with Ridge maintaining R^2 values of 0.635 and 0.756 in the early and late portions of the walk-forward prediction period, respectively. Time-series simulations showed that Ridge and Ensemble_New reproduced the overall decline of total imidazoles from ~370 to ~50 during the walk-forward prediction period (Fig. S7a). The inverse-MSE dynamic weighting strategy increased the ensemble fitting slope from 0.828 to 0.915 and reduced the intercept from 71.3 to 50.7 (Fig. S7b, c), partially alleviating systematic overestimation at low values and underestimation at high values. GPR provided adaptive 95% confidence intervals that characterize time-varying prediction uncertainty and can serve as an auxiliary criterion under extreme conditions; dynamic ensemble weights evolved with environmental conditions (Fig. S7d), but owing to sample

limitations, the overall accuracy of the dynamic ensemble did not surpass that of the Ridge single model. After log transformation, most models showed improved fitting ability, indicating that the preprocessing effectively alleviated deficiencies in the data distribution. In terms of residual distribution, the residuals were generally randomly distributed around zero without a global unidirectional bias, although some heteroscedasticity was present, i.e., residual dispersion increased at higher predicted values. This phenomenon is common in environmental time-series observations and does not necessarily indicate model failure; rather, it reflects more complex reaction processes and increased predictive uncertainty under high-concentration conditions. The prediction intervals provided by GPR explicitly characterize this uncertainty. When the test set was divided chronologically into earlier and later segments for segmented performance testing, the linear models showed improved fitting in the later low-concentration segment. The dynamic weighted ensemble maintained reasonable predictive accuracy across different stages, and the ensemble weights varied systematically with conditions rather than fluctuating randomly, which is consistent with the model adapting to changing environmental conditions; this pattern alone, however, cannot exclude overfitting to local samples. The feature-interpretation results from the standardized PLSR coefficients, permutation importance across all six base models, and the Huber model corroborated one another: carbonyl compounds and organic amines were consistently identified as the highest-rank predictors, supporting the robustness of the core predictor ranking. The relative importance of some variables, such as nitrate, sulfate, ammonium, RH and NO_x, fluctuated with preprocessing methods and model loss functions, mainly owing to multicollinearity within the feature set. Based on the above multidimensional tests, under the seasons and pollution conditions covered by this observation, the overall model predictions were reliable and can be used for quantitative estimation and predictor screening for total imidazoles.

The inherent constraints of data-driven models should be objectively recognized. These boundaries limit the applicable scope of the models but do not negate their value in this study. The limited number of walk-forward prediction points restricts the precision of performance estimates to some extent; however, the relatively high

coefficient of determination represents a large effect size, indicating that the statistical associations between the features and the target are real. Future work can use Bootstrap resampling to obtain confidence intervals for evaluation metrics and time-series block cross-validation to further verify model robustness. Although this study identified and partially mitigated covariate shift between the earlier training window and the later walk-forward prediction points through log transformation, dynamic weighting, and segmented performance diagnostics, distributional differences remain: the earlier observations contain high-concentration peak samples, whereas the later prediction points are dominated by low-to-medium concentrations. Therefore, the model is not recommended for unconstrained extrapolation to pollution scenarios substantially different from the training set. Multicollinearity exists within the feature set. This issue mainly interferes with mechanistic interpretation and has limited impact on predictive capability, causing the importance rankings of some secondary features to fluctuate. Accordingly, all conclusions of this study describe only statistical associations and should not be interpreted as strict causal mechanisms; the model is intended for predictor screening and approximate concentration estimation within the observed range, not for mechanistic attribution. The model still achieved good simulation performance with the selected input features. These constraints are common practical issues in machine learning studies based on field observations. This study has identified and fully disclosed them, and the applicable operational boundaries should be strictly observed when using the model results.

Table S1 Marker ions selected for imidazole-containing particle types.

Species	m/z	Marker ion	Area	Relative area	Function
Elemental carbon	12 ± 0.5	$[C]^+$	50	0.005	or
	24 ± 0.5	$[C_2]^+$	50	0.005	or
	36 ± 0.5	$[C_3]^+$	50	0.005	or
	48 ± 0.5	$[C_4]^+$	50	0.005	or
	60 ± 0.5	$[C_5]^+$	50	0.005	or
	−12	$[C]^-$	50	0.005	or
	−24	$[C_2]^-$	50	0.005	or
	−36	$[C_3]^-$	50	0.005	or
	−48	$[C_4]^-$	50	0.005	or
	−60	$[C_5]^-$	50	0.005	or
Organic carbon	27	$[C_2H_3]^+$	50	0.005	or
	43	$[C_2H_3O]^+$	50	0.005	or
Elemental carbon and organic carbon	12	$[C]^+$	50	0.005	or
	24	$[C_2]^+$	50	0.005	or
	36	$[C_3]^+$	50	0.005	or
	37	$[C_3H]^+$	50	0.005	or
	43	$[C_2H_3O]^+$	50	0.005	or
Levoglucosan	39	$[K]^+$	1500		and
	113, 115	$[K_2Cl]^+$	50	0.005	or
	213	$[K_3SO_4]^+$	50	0.005	or
	−26	$[CN]^-$	50	0.005	or
	−59	$[C_2H_3O_2]^-$	50	0.005	or
	−73	$[C_3H_3O_2]^-$	50	0.005	or
Heavy metal	206–208	$[Pb]^+$	50	0.005	or
	64, 66	$[Zn]^+$	50	0.005	or
	63, 65	$[Cu]^+$	50	0.005	or
	56	$[Fe]^+$	50	0.005	or
	55	$[Mn]^+$	50	0.005	or
	71	$[MnO]^+$	50	0.005	or
	51	$[V]^+$	50	0.005	or
	67	$[VO]^+$	50	0.005	or
Dust	24	$[Mg]^+$	1000		
	27	$[Al]^+$	1500		
	40	$[Ca]^+$	2500		
	56	$[CaO]^+$/ $[Fe]^+$	100	0.05	
	-76	$[SiO_3]^-$	500	0.05	
K	39	$[K]^+$	No other intense peaks were observed.		

Table S2 Classification and proportions of total imidazole-containing particle subtypes from firework-related (FW-related) and non-firework (non-FW) sources.

Source	Type	Subtype	Proportion
FW-related	EC	EC-nit-sul	1.6%
		EC-NaK-CN	3.5%
		KEC-CN	3.1%
	ECOC	ECOC-nit-sul	3.4%
	OC	KOC	13.0%
	Metal	Metal-AlCu	25.7%
		Metal-Ba	2.6%
	K	K-nit	12.6%
		Subtotal	65.4%
non-FW	ECOC	ECOC-sul	11.1%
	OC	OC-PAH	1.1%
		KOC-CN	13.8%
	Metal	K-Ca-nit	10.7%
		Metal-Fe	0.8%
	K	K-LEV	2.9%
		K-nit-sul	4.3%
		Subtotal	34.6%

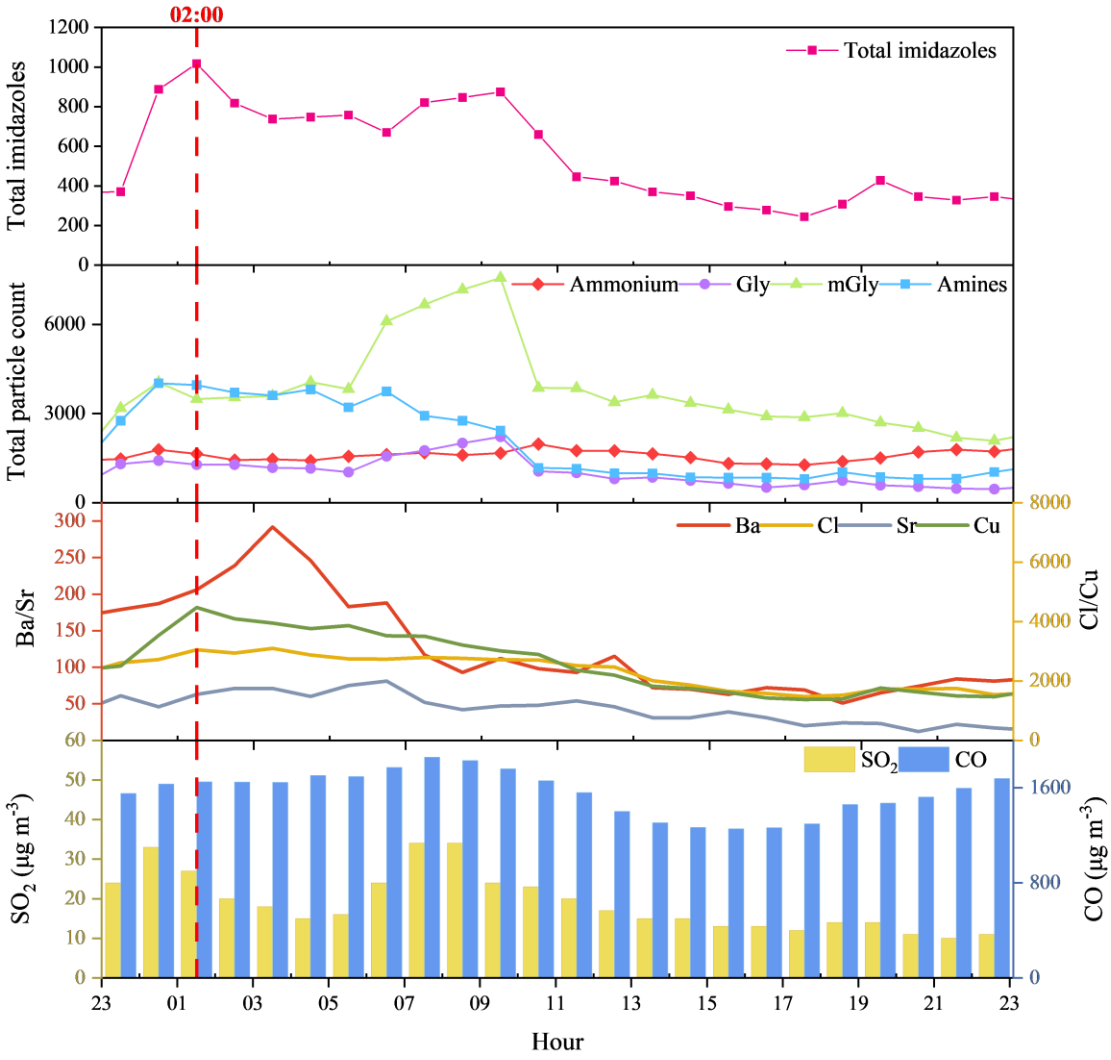


Figure S1. Hourly variations in total imidazoles, imidazole precursors, metals, and gaseous pollutants during the intense firework emission episode (from 00:00 to 23:00 on 22 January).

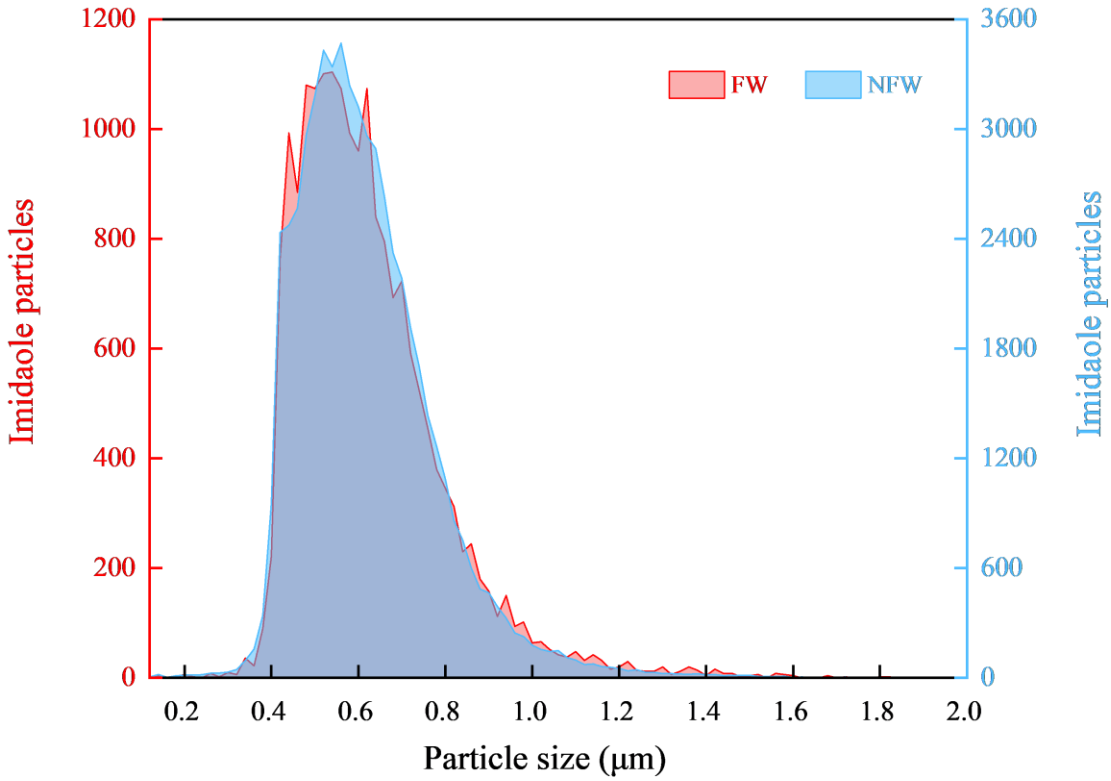


Figure S2. Unscaled size distribution of imidazole-containing particles during the FW and NFW episodes.

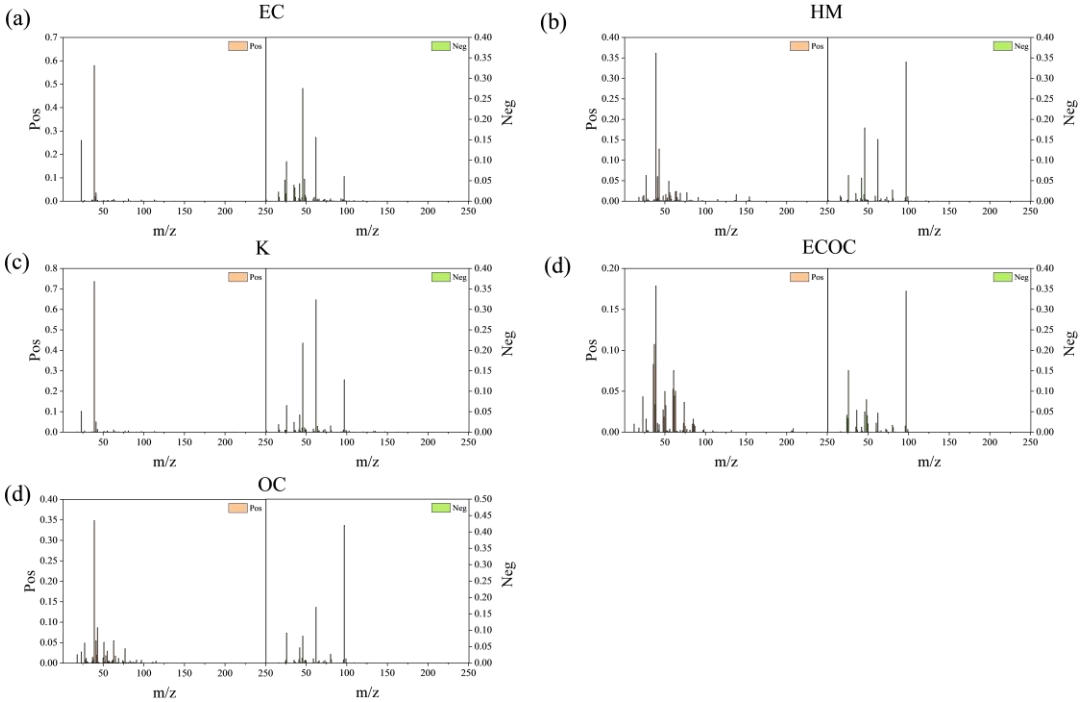
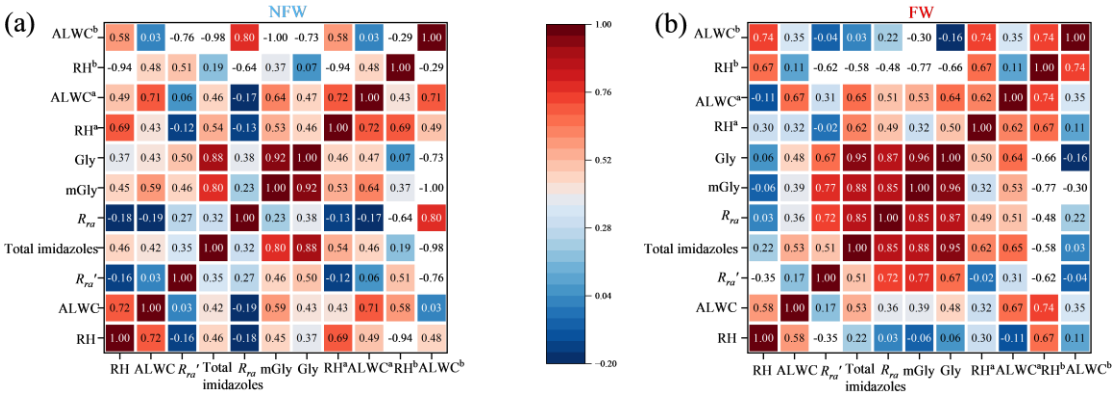


Figure S3. Mass spectra of the identified types of imidazole-containing particles in the FW episode.



186

187 **Figure S4.** Correlation heatmap of environmental factors. ALWC^a and RH^a denote
188 ALWC and RH, respectively, measured under RH < 80%. ALWC^b and RH^b denote
189 ALWC and RH, respectively, measured under RH > 80%.

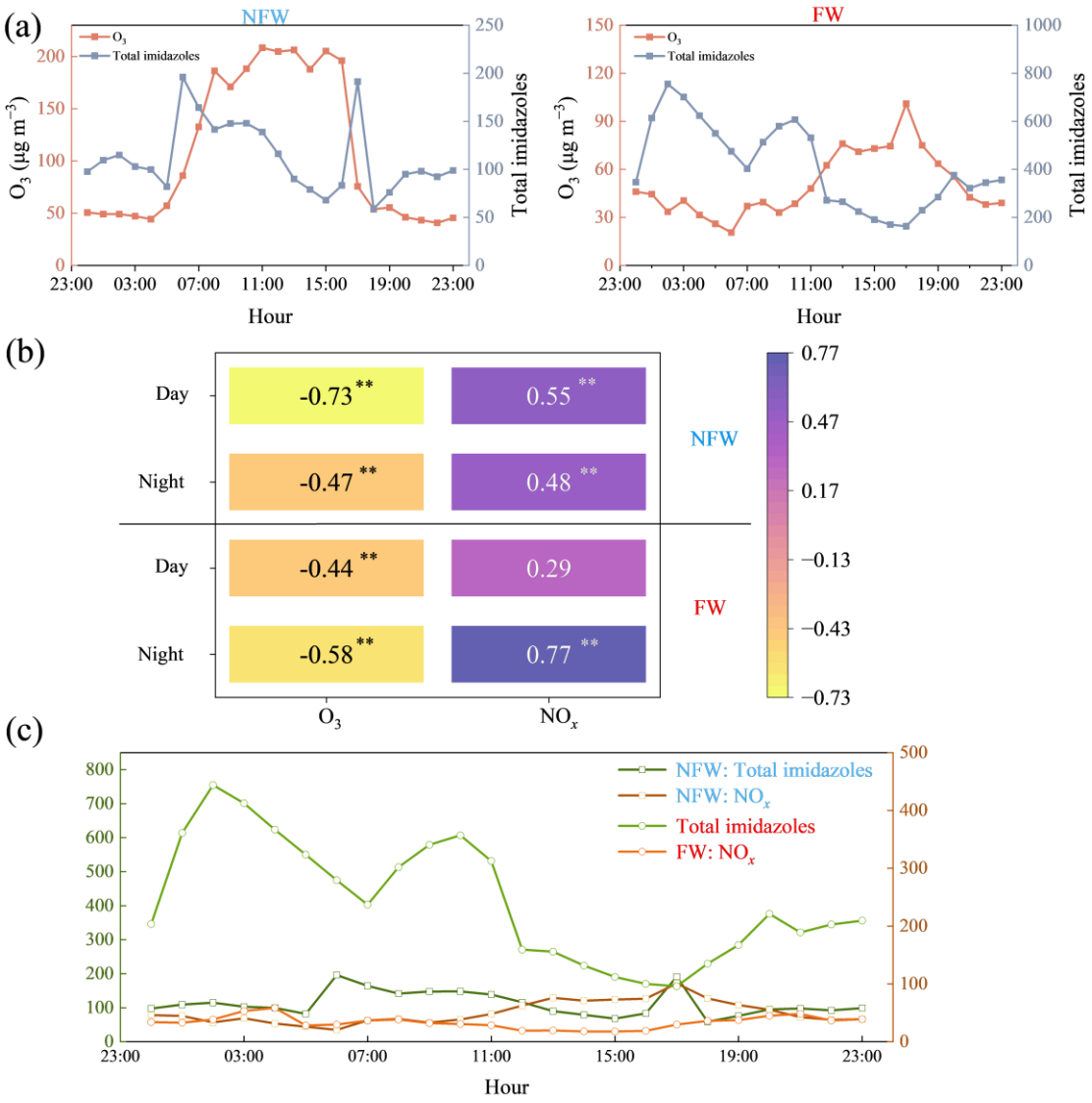


Figure S5 (a) Hourly variations of O_3 and total imidazoles during the FW and NFW episodes; (b) heatmaps of diurnal correlations of O_3 and NO_x with total imidazole-containing particles during the FW and NFW episodes; (c) hourly variations in NO_x and imidazole-containing particles during the FW and NFW episodes.

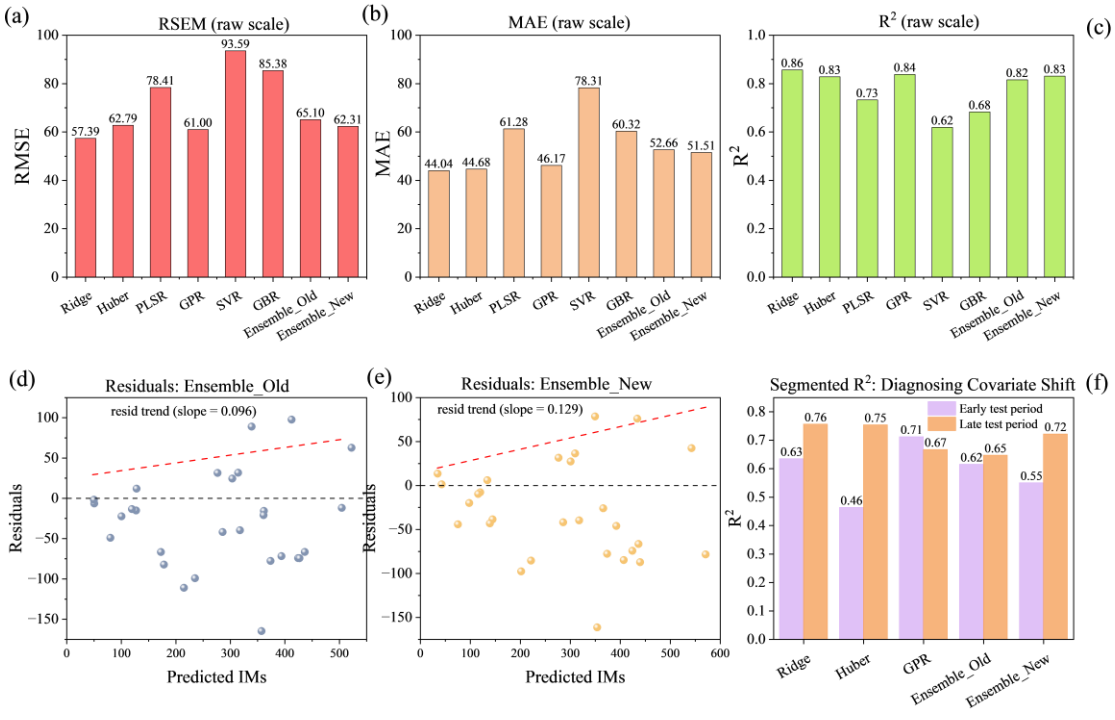
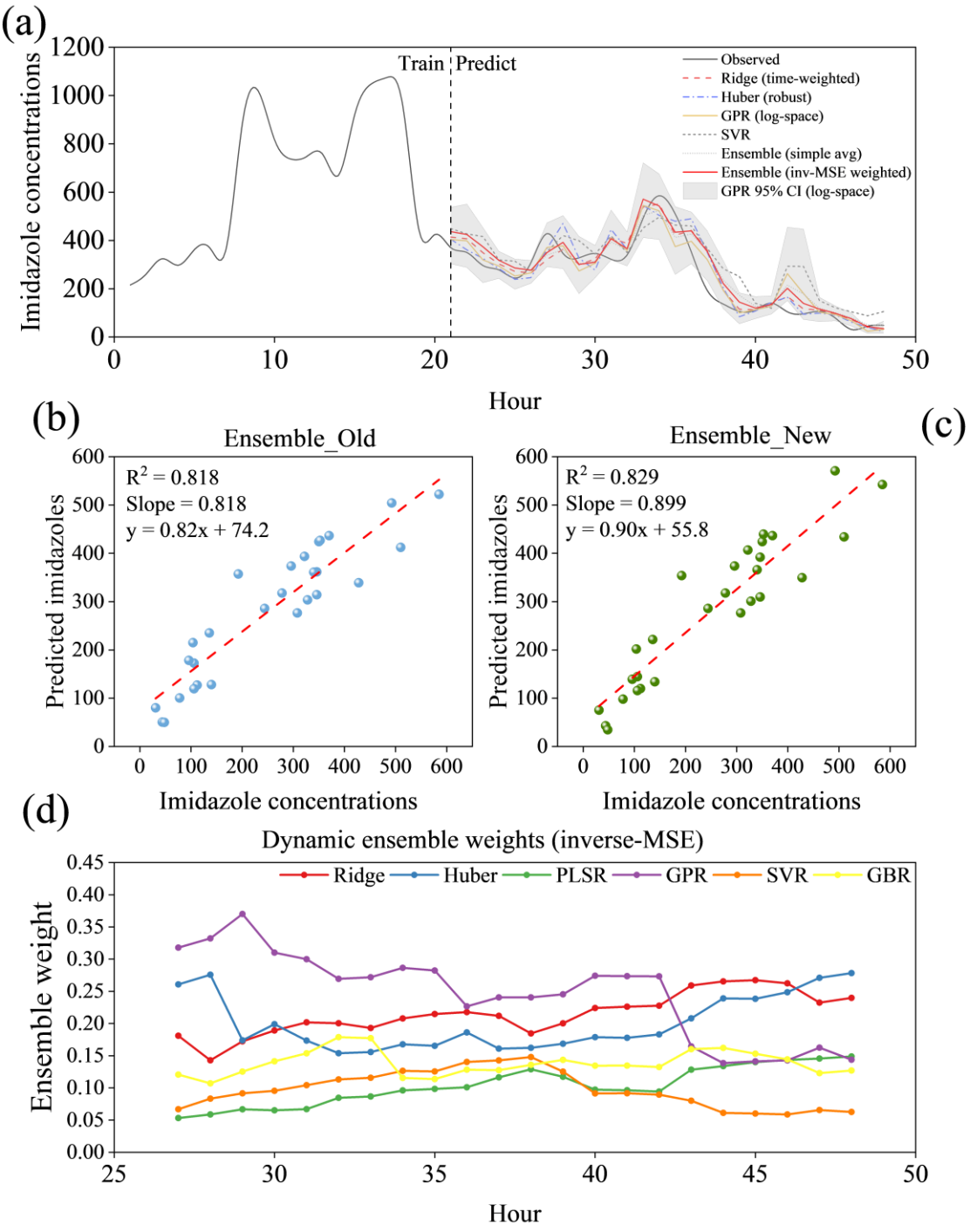


Figure S6 Performance evaluation and temporal diagnostic analysis of regression models for particulate total imidazole prediction under firework emission scenarios. (a–c) Comparison of raw-scale predictive metrics, including root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2), across six base models (Ridge, Huber, PLSR, GPR, SVR, GBR) and two ensemble models; (d–e) residual distributions versus predicted total imidazole concentrations for the old and new ensemble models, with red dashed lines representing the linear trend of residuals to detect systematic prediction bias; (f) segmented R^2 values for the early and late test periods, applied to diagnose temporal covariate shift in model generalization performance.



206

207 **Figure S7** Temporal fitting performance, predictive accuracy validation, and dynamic
208 weight evolution of multi-model ensemble frameworks for particulate total imidazole
209 prediction in the FW episode. (a) Time-series comparison between observed total
210 imidazole concentrations and multi-model predictions across the training and test
211 periods, covering time-weighted Ridge regression, robust Huber regression, log-space
212 GPR, SVR, simple-averaging ensemble and inverse-MSE weighted ensemble; the grey
213 shaded area indicates the 95% confidence interval of the log-space GPR model. (b–c)

Ensemble_Old and Ensemble_New, with red dashed lines representing the linear regression fits, annotated with R^2, regression slope and linear fitting equation; (d) temporal evolution of dynamic weights for six base models (Ridge, Huber, PLSR, GPR, SVR, GBR) in the inverse-MSE weighted ensemble scheme during the test period.