Supplement of Geosci. Model Dev., 15, 7791–7807, 2022
https://doi.org/10.5194/gmd-15-7791-2022-supplement

Geoscientific Model Development
Open Access
EGU

Supplement of

Development of a regional feature selection-based machine learning system (RFSML v1.0) for air pollution forecasting over China

Li Fang et al.

Correspondence to: Jianbing Jin (jianbing.jin@nuist.edu.cn) and Hong Liao (hongliao@nuist.edu.cn)

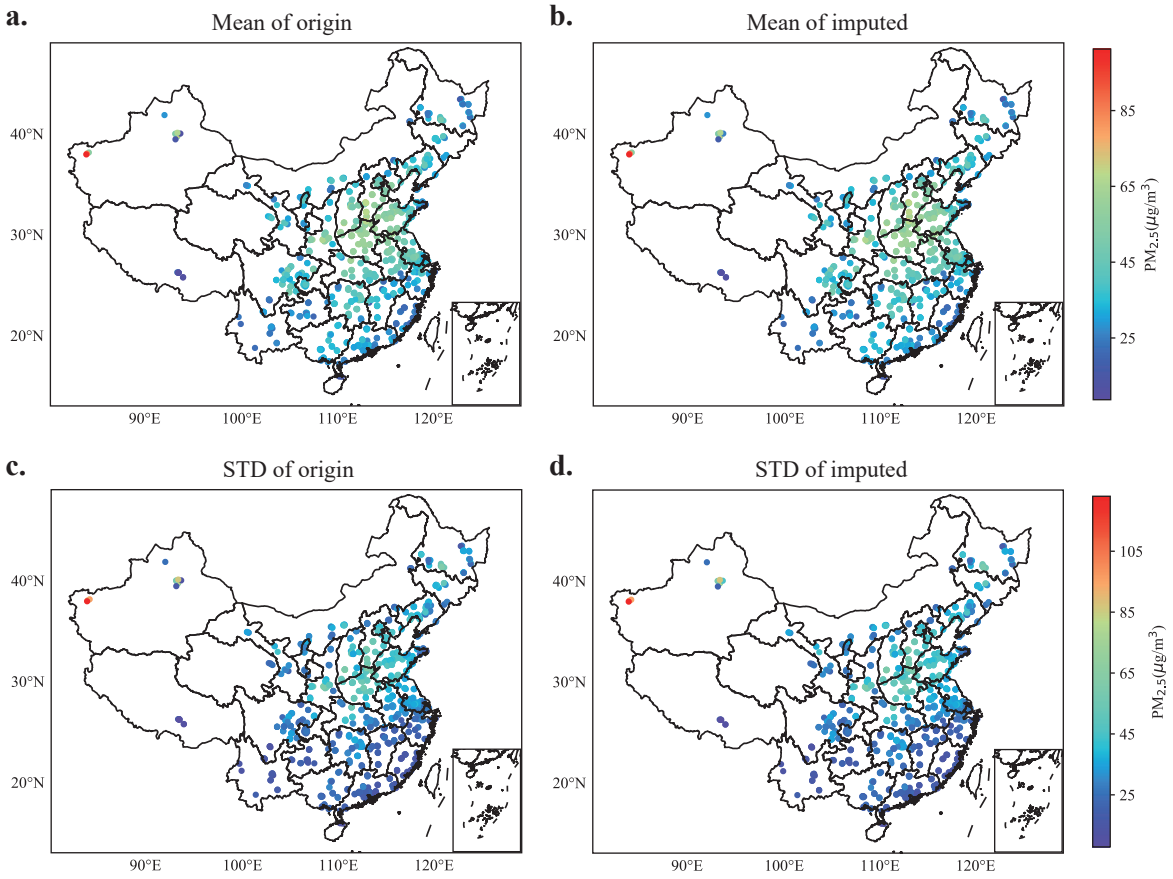

Figure S1. Spatial distribution of original and interpolated time series as for mean and standard deviation. Panel a and c are the mean and STD of time series of whole original $PM_{2.5}$ used in this work. Panel b and d are the same with a and c but with imputation. There is no significant change in colors between the distribution of original and imputed time series which implies the interpolation method is reliable.

Algorithm 1 KNN interpolation based on IDW

Initialization: read input site E
Calculate its' distance D_s with surrounding sites and construct distance matrix
if $D_s < 0.8$ radius **then**
    Get all proper sites and count the amount F
    if $F_s > 4$ **then**
        Randomly select 4 sites
    else if $F_s < 2$ **then**
        Drop E
    else
        Select all proper sites
    end if
end if
while exist missing values **do**
    IDW
end while

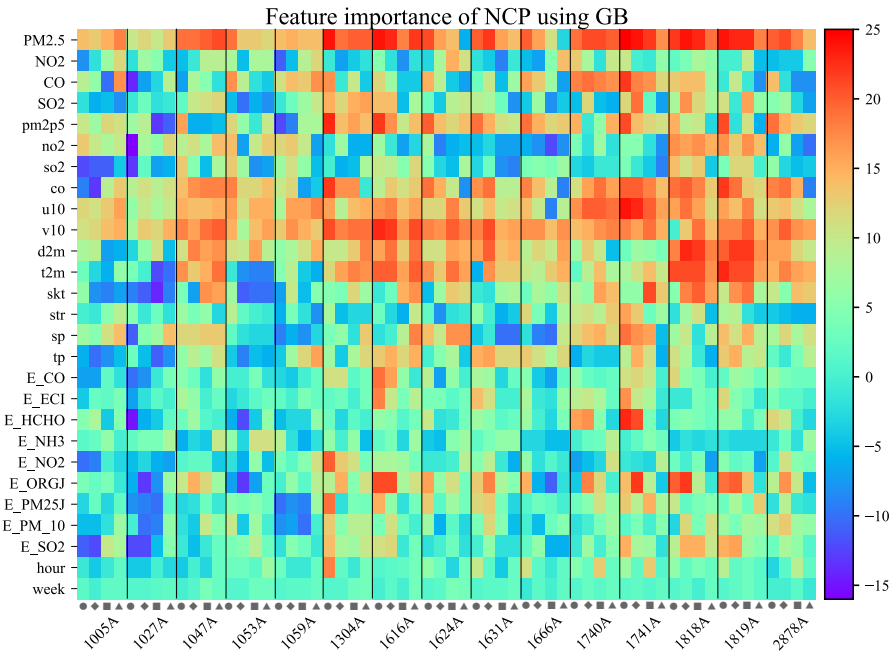

Figure S2. Heatmap of all empirical features with random 15 monitoring stations in NCP and four predicting horizons. The circle, diamond, square and triangle represent four predicting horizons 6, 12, 18 and 24 h respectively. The heatmap is based on ranking the SAGE analysis of features training by GB. The warmer the color tone on the whole rows, the more important the corresponding feature.

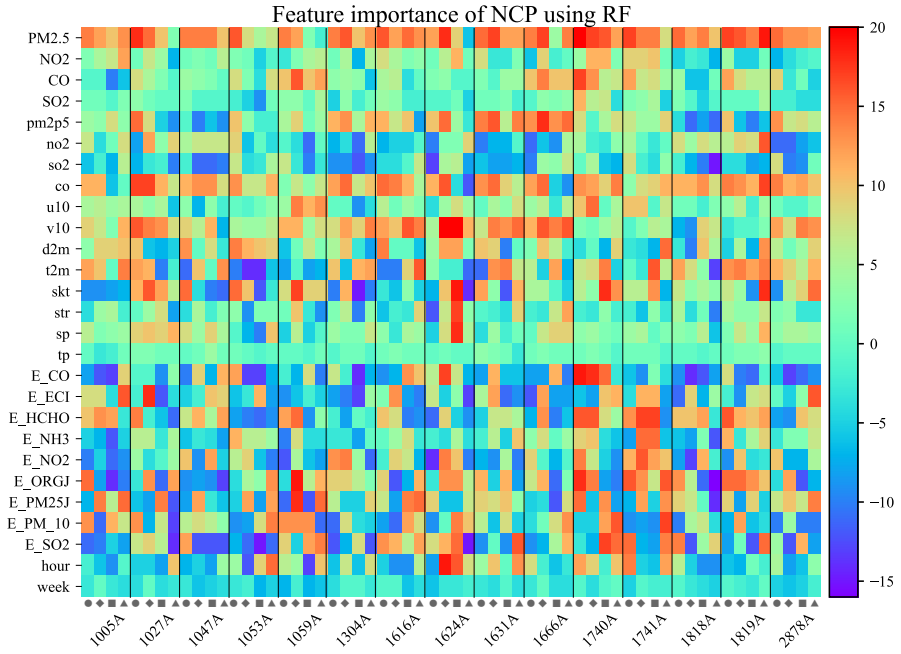

Figure S3. Heatmap of all empirical features with random 15 monitoring stations in NCP and four predicting horizons. The circle, diamond, square and triangle represent four predicting horizons 6, 12, 18 and 24 h respectively. The heatmap is based on ranking the SAGE analysis of features training by RF. The warmer the color tone on the whole rows, the more important the corresponding feature.

Table S1. Summary of prediction performance in the time period of April, 2020.

Region	**Metric**	**Predicting horizon**			
		6		18	
		standardML	**RFSML**	**standardML**	**RFSML**
NCP	RMSE	17.71	12.20	22.11	16.71
	MAE	14.06	9.30	17.86	13.19
	R	0.71	0.83	0.50	0.69
PRD	RMSE	10.70	7.78	13.17	11.10
	MAE	8.51	5.74	10.38	8.39
	R	0.83	0.90	0.70	0.77
SCB	RMSE	13.29	10.37	17.02	13.51
	MAE	10.13	7.63	13.11	10.20
	R	0.72	0.81	0.53	0.66
YRD	RMSE	14.08	10.43	18.67	14.41
	MAE	11.27	8.09	14.76	11.48
	R	0.75	0.87	0.51	0.74
FWP	RMSE	16.26	13.24	19.80	16.44
	MAE	12.69	10.14	15.65	12.97
	R	0.66	0.73	0.47	0.60
REST	RMSE	21.59	17.89	26.01	22.25
	MAE	14.29	10.50	17.48	13.62
	R	0.68	0.79	0.48	0.66

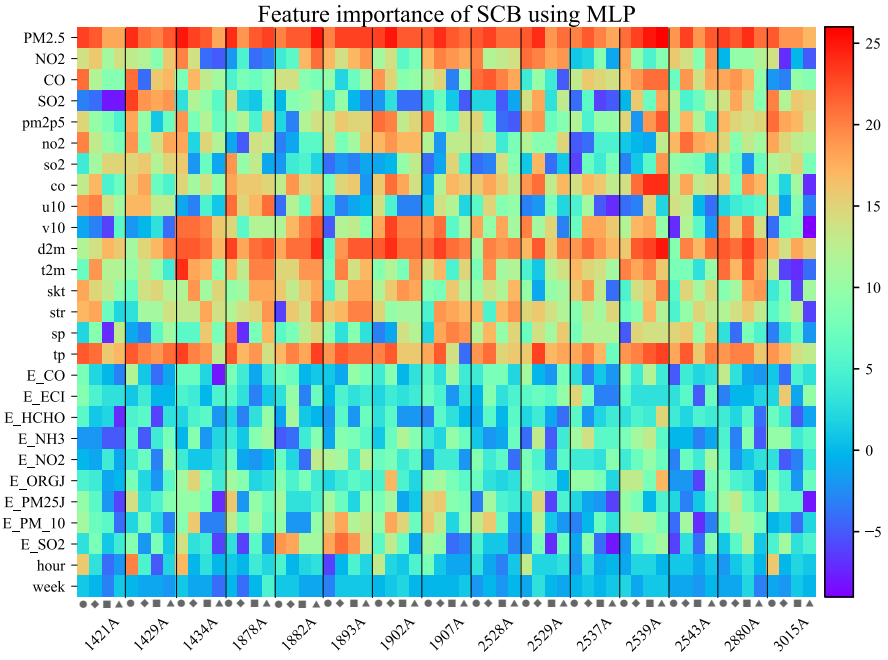

Figure S4. Heatmap of all empirical features with random 15 monitoring stations in SCB and four predicting horizons. The circle, diamond, square and triangle represent four predicting horizons 6, 12, 18 and 24 h respectively. The heatmap is based on ranking the SAGE analysis of features training by MLP. The warmer the color tone on the whole rows, the more important the corresponding feature.

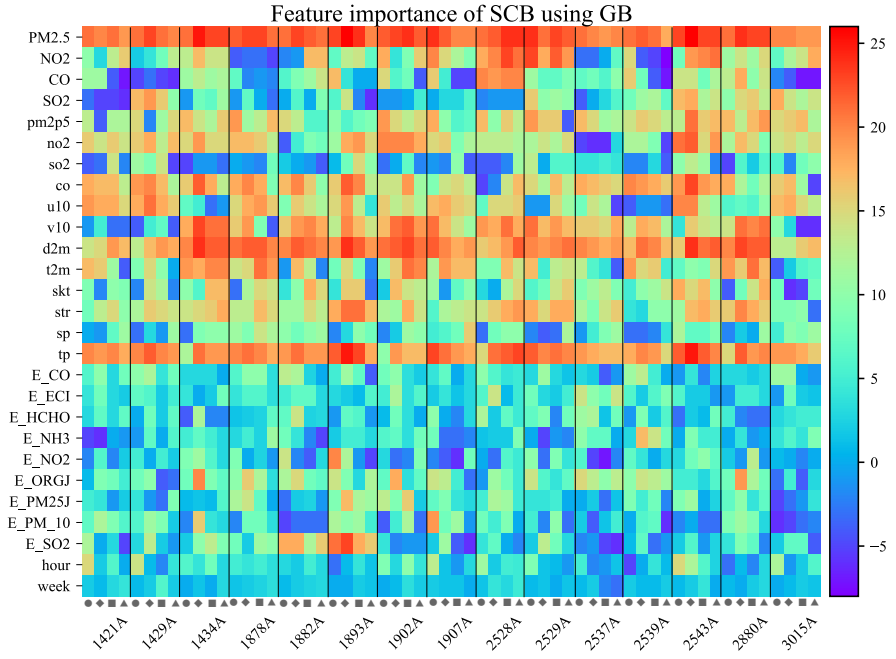

Figure S5. Heatmap of all empirical features with random 15 monitoring stations in SCB and four predicting horizons. The circle, diamond, square and triangle represent four predicting horizons 6, 12, 18 and 24 h respectively. The heatmap is based on ranking the SAGE analysis of features training by GB. The warmer the color tone on the whole rows, the more important the corresponding feature.

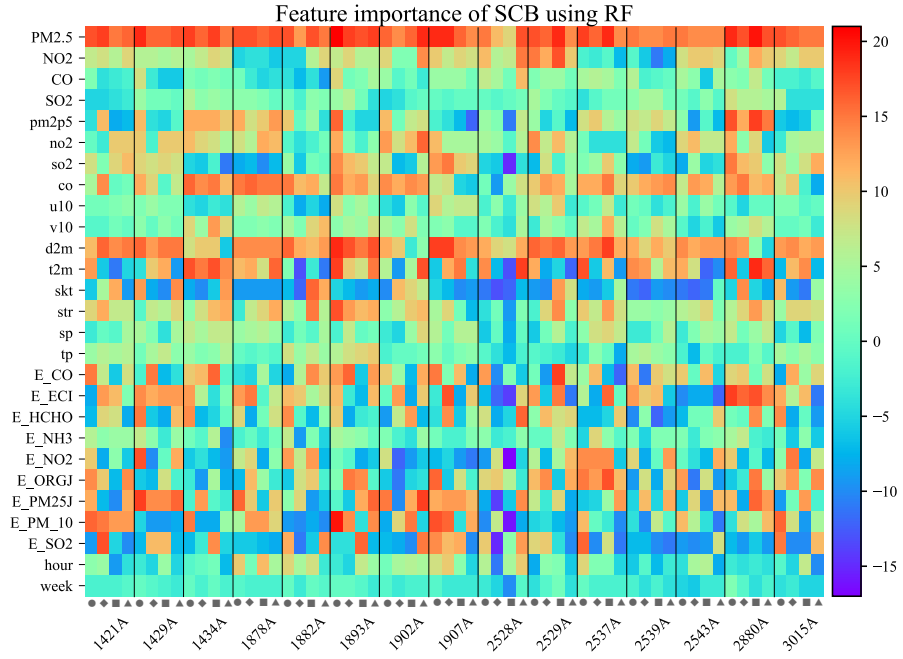

Figure S6. Heatmap of all empirical features with random 15 monitoring stations in SCB and four predicting horizons. The circle, diamond, square and triangle represent four predicting horizons 6, 12, 18 and 24 h respectively. The heatmap is based on ranking the SAGE analysis of features training by RF. The warmer the color tone on the whole rows, the more important the corresponding feature.

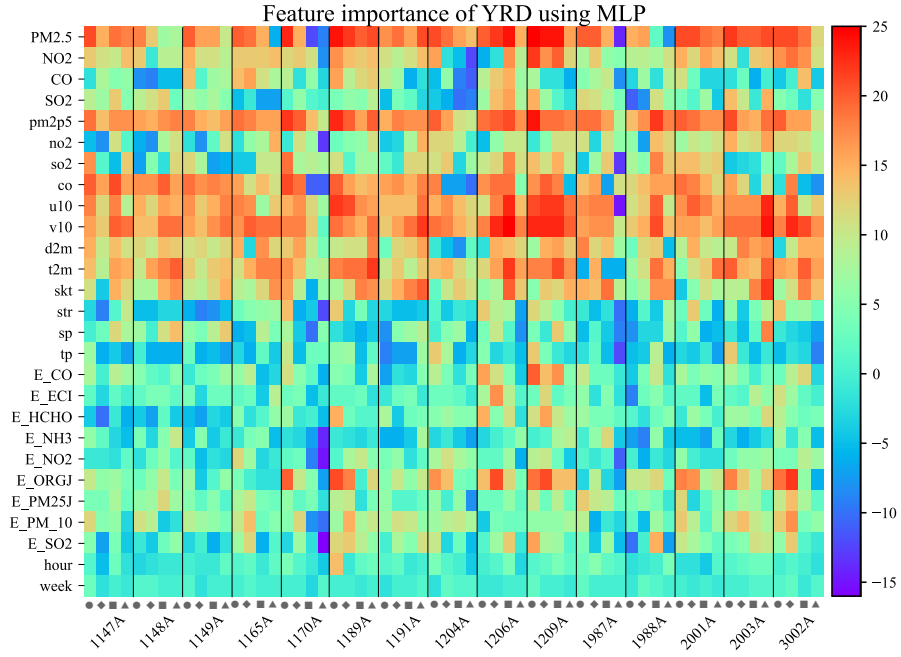

Figure S7. Heatmap of all empirical features with random 15 monitoring stations in YRD and four predicting horizons. The circle, diamond, square and triangle represent four predicting horizons 6, 12, 18 and 24 h respectively. The heatmap is based on ranking the SAGE analysis of features training by MLP. The warmer the color tone on the whole rows, the more important the corresponding feature.

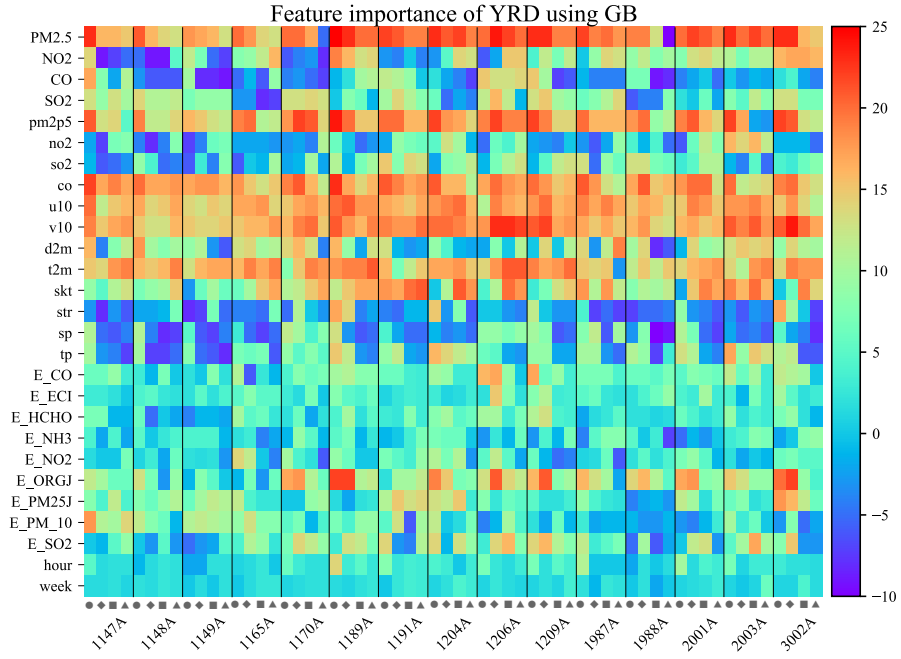

Figure S8. Heatmap of all empirical features with random 15 monitoring stations in YRD and four predicting horizons. The circle, diamond, square and triangle represent four predicting horizons 6, 12, 18 and 24 h respectively. The heatmap is based on ranking the SAGE analysis of features training by GB. The warmer the color tone on the whole rows, the more important the corresponding feature.

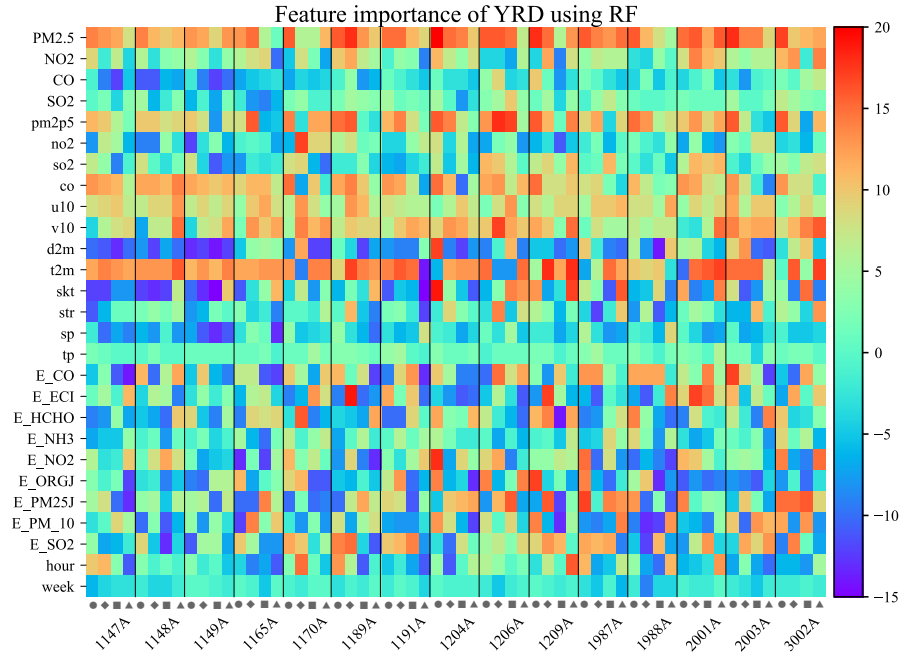

Figure S9. Heatmap of all empirical features with random 15 monitoring stations in YRD and four predicting horizons. The circle, diamond, square and triangle represent four predicting horizons 6, 12, 18 and 24 h respectively. The heatmap is based on ranking the SAGE analysis of features training by RF. The warmer the color tone on the whole rows, the more important the corresponding feature.

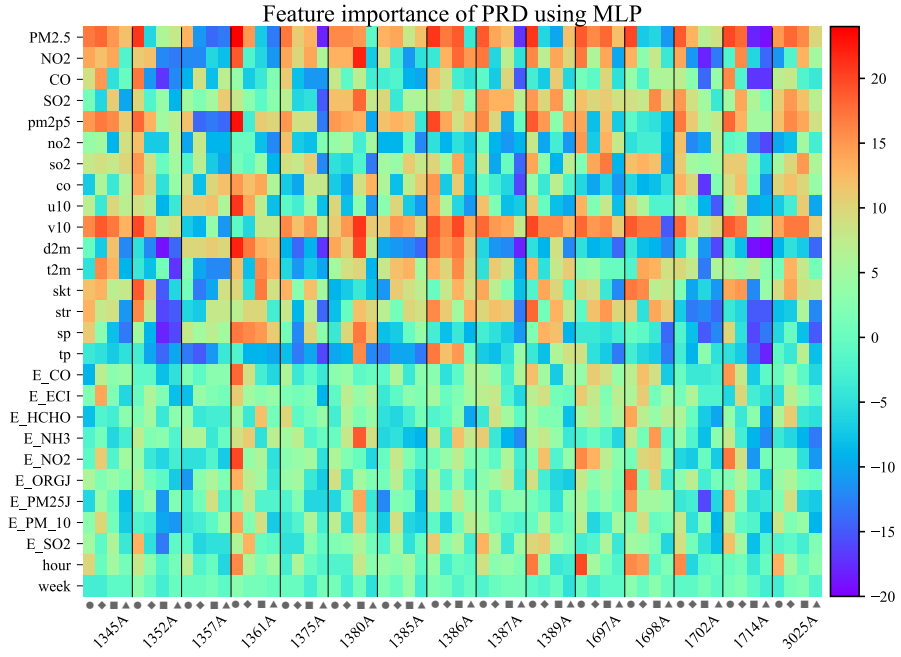

Figure S10. Heatmap of all empirical features with random 15 monitoring stations in PRD and four predicting horizons. The circle, diamond, square and triangle represent four predicting horizons 6, 12, 18 and 24 h respectively. The heatmap is based on ranking the SAGE analysis of features training by MLP. The warmer the color tone on the whole rows, the more important the corresponding feature.

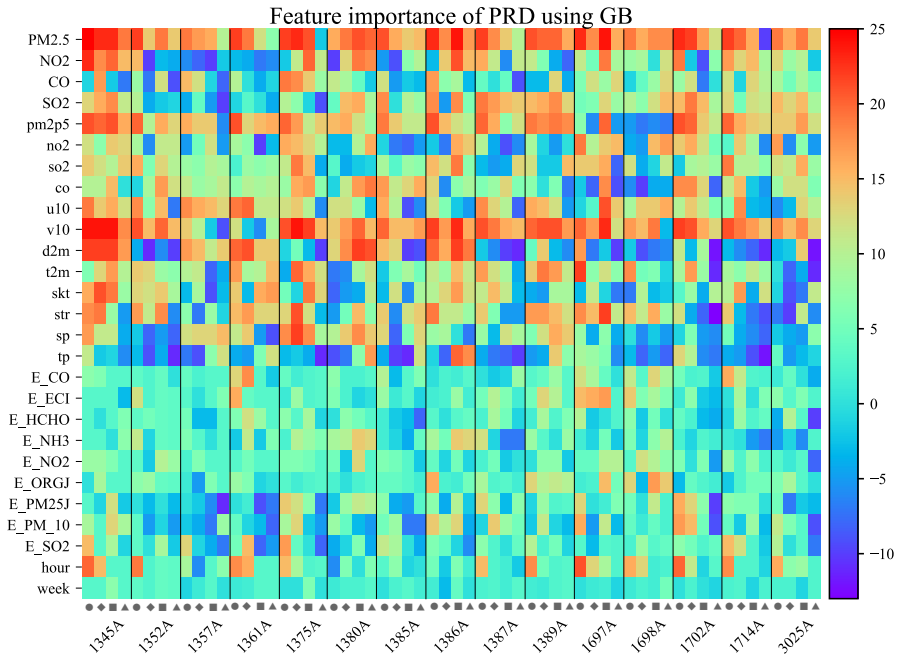

Figure S11. Heatmap of all empirical features with random 15 monitoring stations in PRD and four predicting horizons. The circle, diamond, square and triangle represent four predicting horizons 6, 12, 18 and 24 h respectively. The heatmap is based on ranking the SAGE analysis of features training by GB. The warmer the color tone on the whole rows, the more important the corresponding feature.

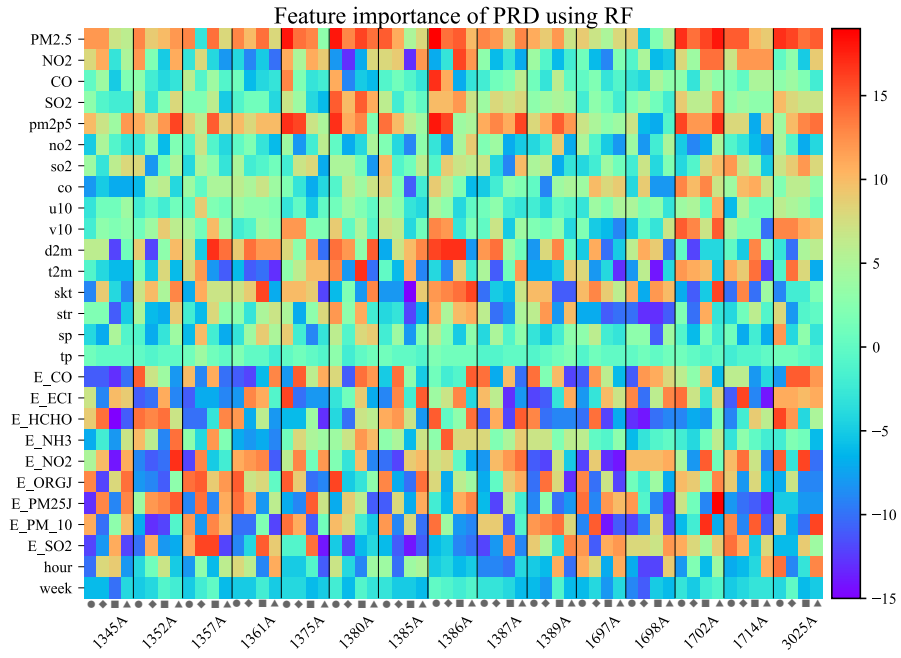

Figure S12. Heatmap of all empirical features with random 15 monitoring stations in PRD and four predicting horizons. The circle, diamond, square and triangle represent four predicting horizons 6, 12, 18 and 24 h respectively. The heatmap is based on ranking the SAGE analysis of features training by RF. The warmer the color tone on the whole rows, the more important the corresponding feature.

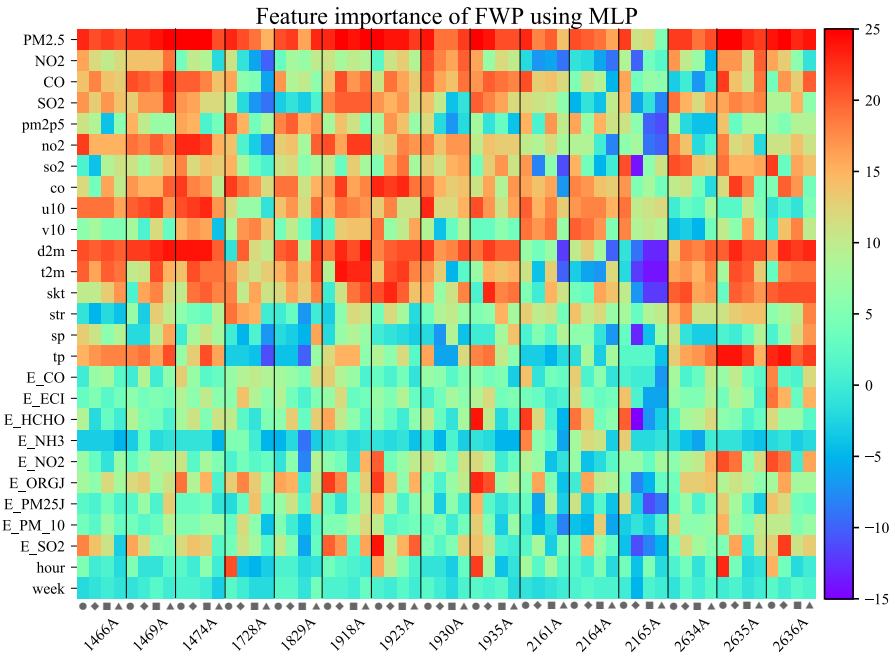

Figure S13. Heatmap of all empirical features with random 15 monitoring stations in FWP and four predicting horizons. The circle, diamond, square and triangle represent four predicting horizons 6, 12, 18 and 24 h respectively. The heatmap is based on ranking the SAGE analysis of features training by MLP. The warmer the color tone on the whole rows, the more important the corresponding feature.

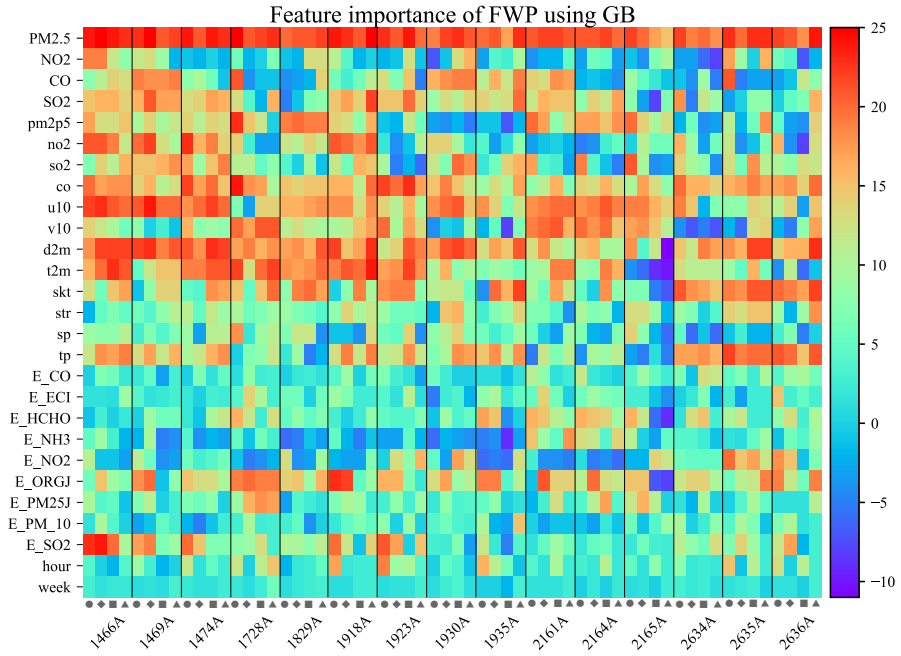

Figure S14. Heatmap of all empirical features with random 15 monitoring stations in PRD and four predicting horizons. The circle, diamond, square and triangle represent four predicting horizons 6, 12, 18 and 24 h respectively. The heatmap is based on ranking the SAGE analysis of features training by GB. The warmer the color tone on the whole rows, the more important the corresponding feature.

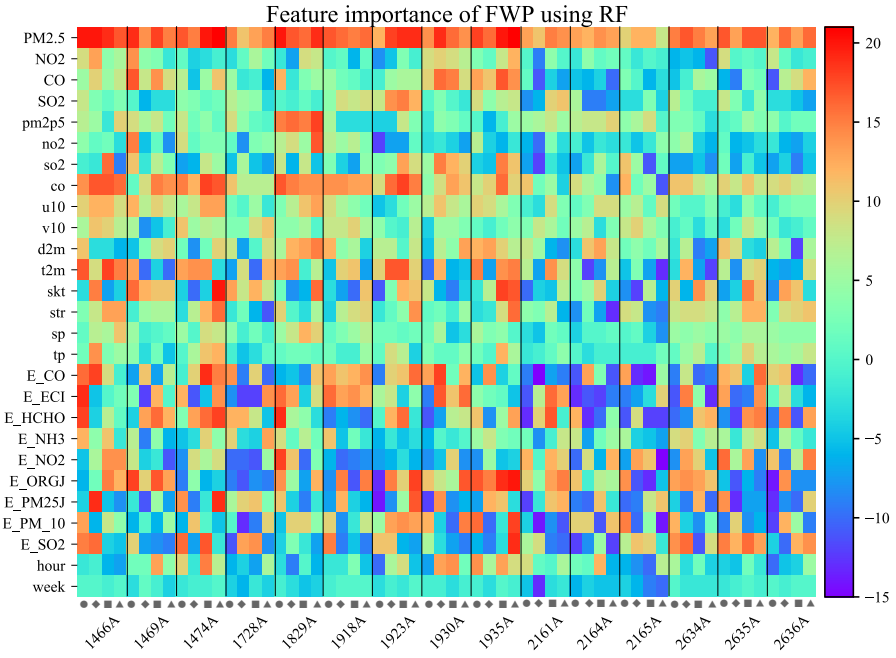

Figure S15. Heatmap of all empirical features with random 15 monitoring stations in PRD and four predicting horizons. The circle, diamond, square and triangle represent four predicting horizons 6, 12, 18 and 24 h respectively. The heatmap is based on ranking the SAGE analysis of features training by RF. The warmer the color tone on the whole rows, the more important the corresponding feature.

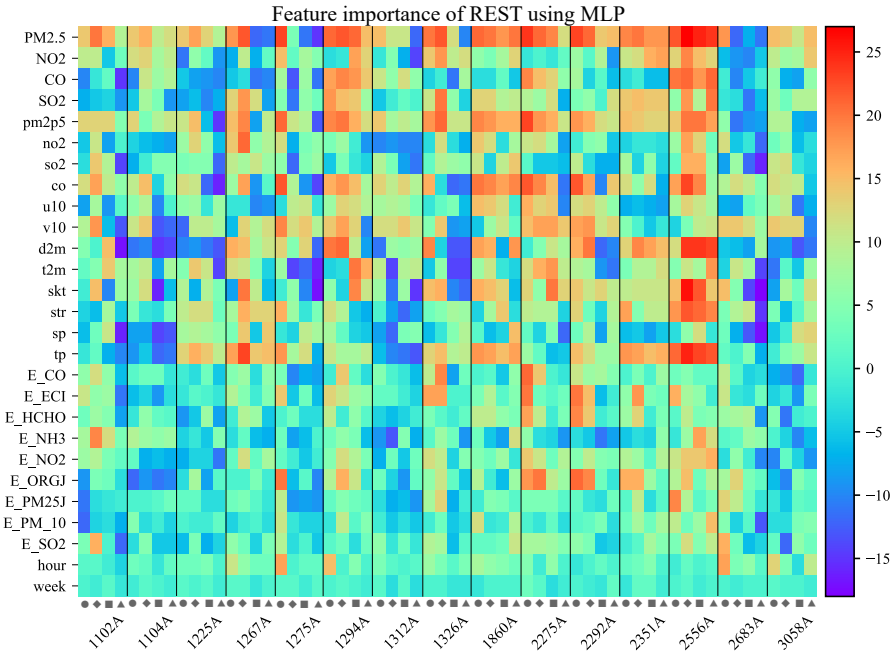

Figure S16. Heatmap of all empirical features with random 15 monitoring stations in the rest area of China and four predicting horizons. The circle, diamond, square and triangle represent four predicting horizons 6, 12, 18 and 24 h respectively. The heatmap is based on ranking the SAGE analysis of features training by MLP. The warmer the color tone on the whole rows, the more important the corresponding feature.

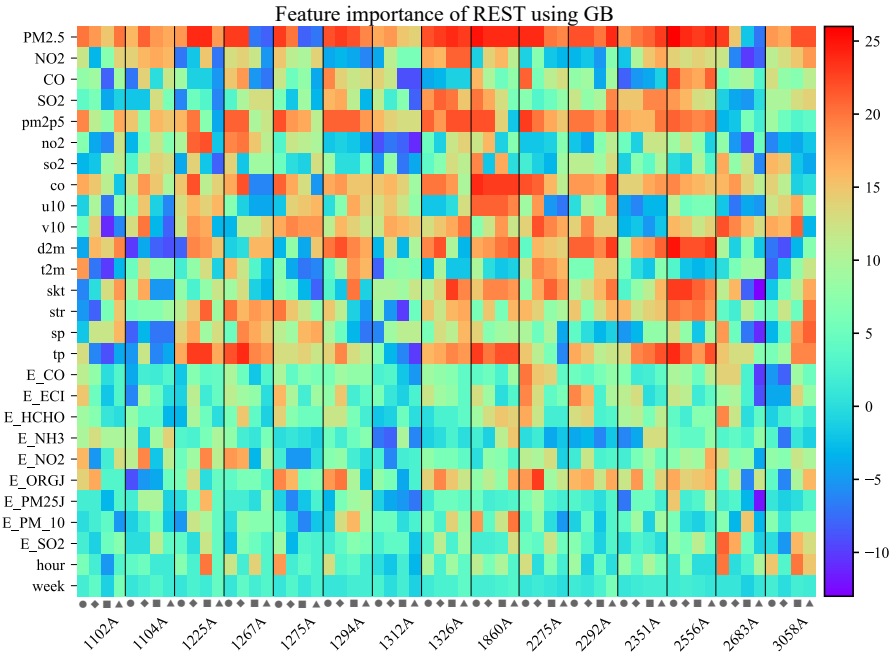

Figure S17. Heatmap of all empirical features with random 15 monitoring stations in the rest area of China and four predicting horizons. The circle, diamond, square and triangle represent four predicting horizons 6, 12, 18 and 24 h respectively. The heatmap is based on ranking the SAGE analysis of features training by GB. The warmer the color tone on the whole rows, the more important the corresponding feature.

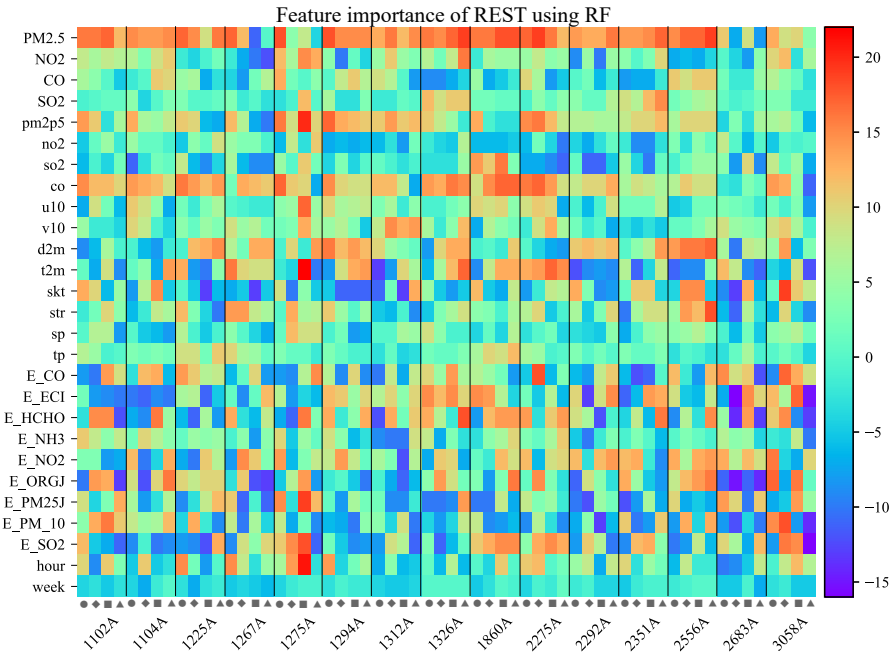

Figure S18. Heatmap of all empirical features with random 15 monitoring stations in the rest area of China and four predicting horizons. The circle, diamond, square and triangle represent four predicting horizons 6, 12, 18 and 24 h respectively. The heatmap is based on ranking the SAGE analysis of features training by RF. The warmer the color tone on the whole rows, the more important the corresponding feature.

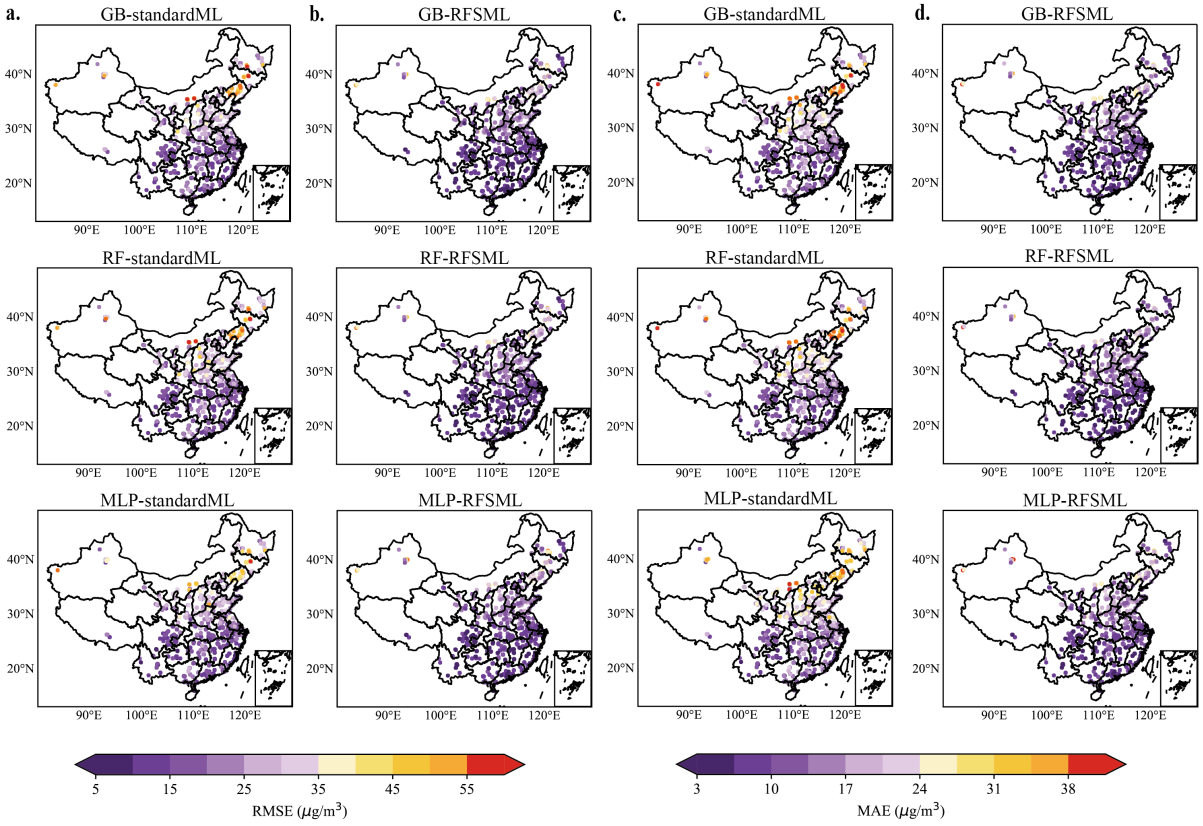

Figure S19. Spatial distribution of RMSE and MAE in a predicting horizon of 6 hours. Panel a and c are results of standard machine learning system while panel b and d are results of RFSML. The cooler the color tone, the lower the RMSE and MAE, thus the better predicting performance.

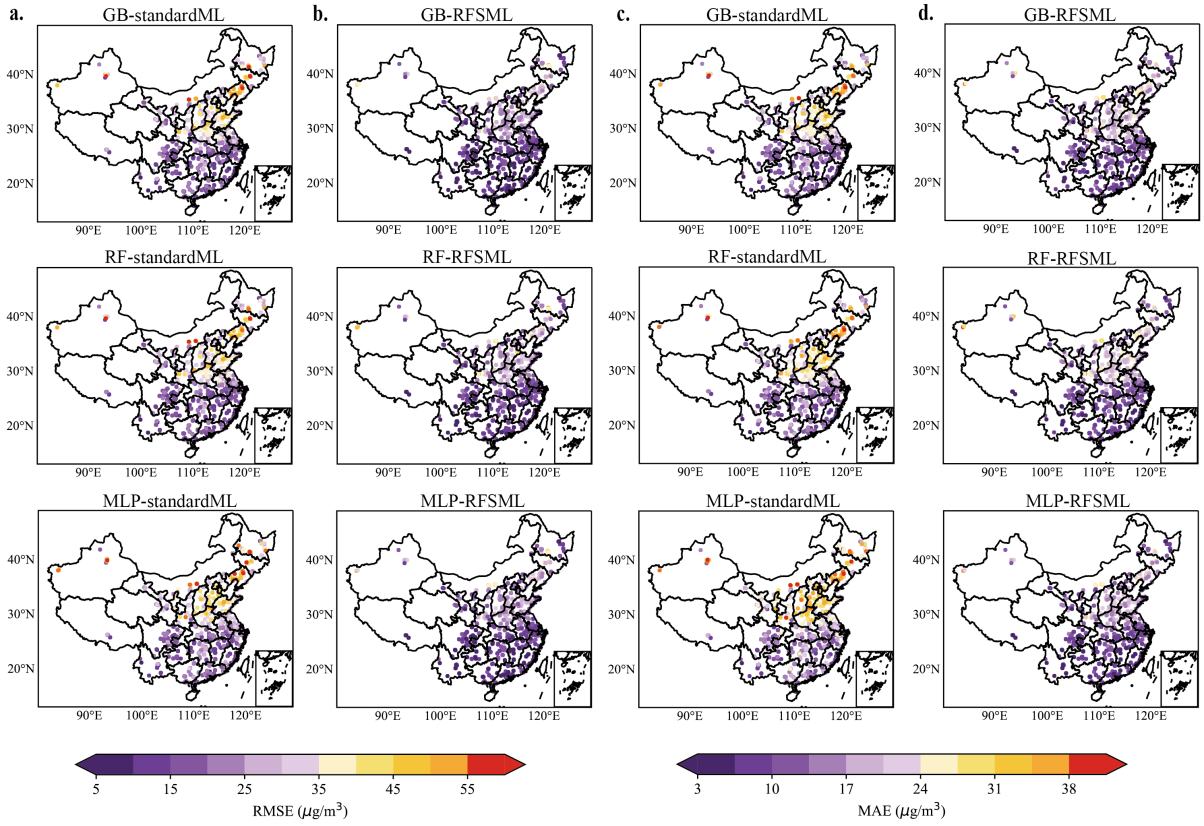

Figure S20. Spatial distribution of RMSE and MAE in a predicting horizon of 18 hours. Panel a and c are results of standard machine learning system while panel b and d are results of RFSML. The cooler the color tone, the lower the RMSE and MAE, thus the better predicting performance.

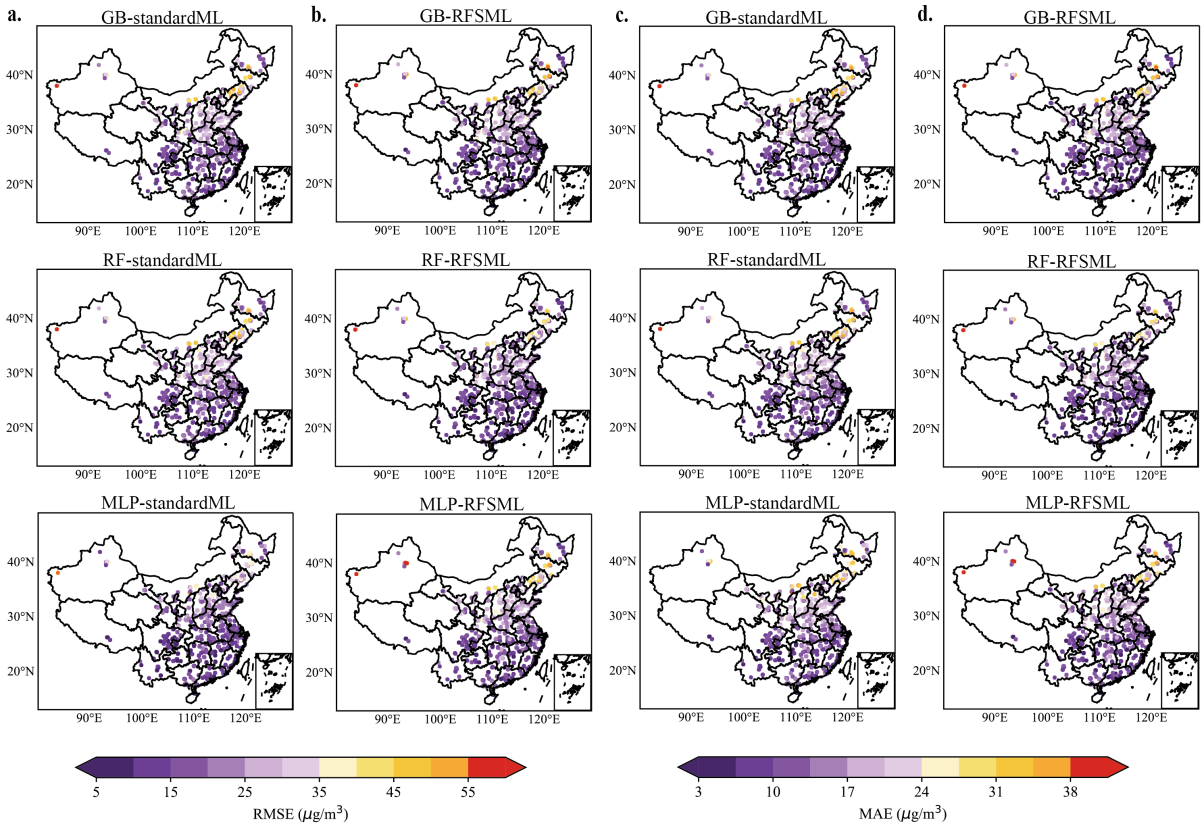

Figure S21. Spatial distribution of RMSE and MAE in a predicting horizon of 24 hours. Panel a and c are results of standard machine learning system while panel b and d are results of RFSML. The cooler the color tone, the lower the RMSE and MAE, thus the better predicting performance.

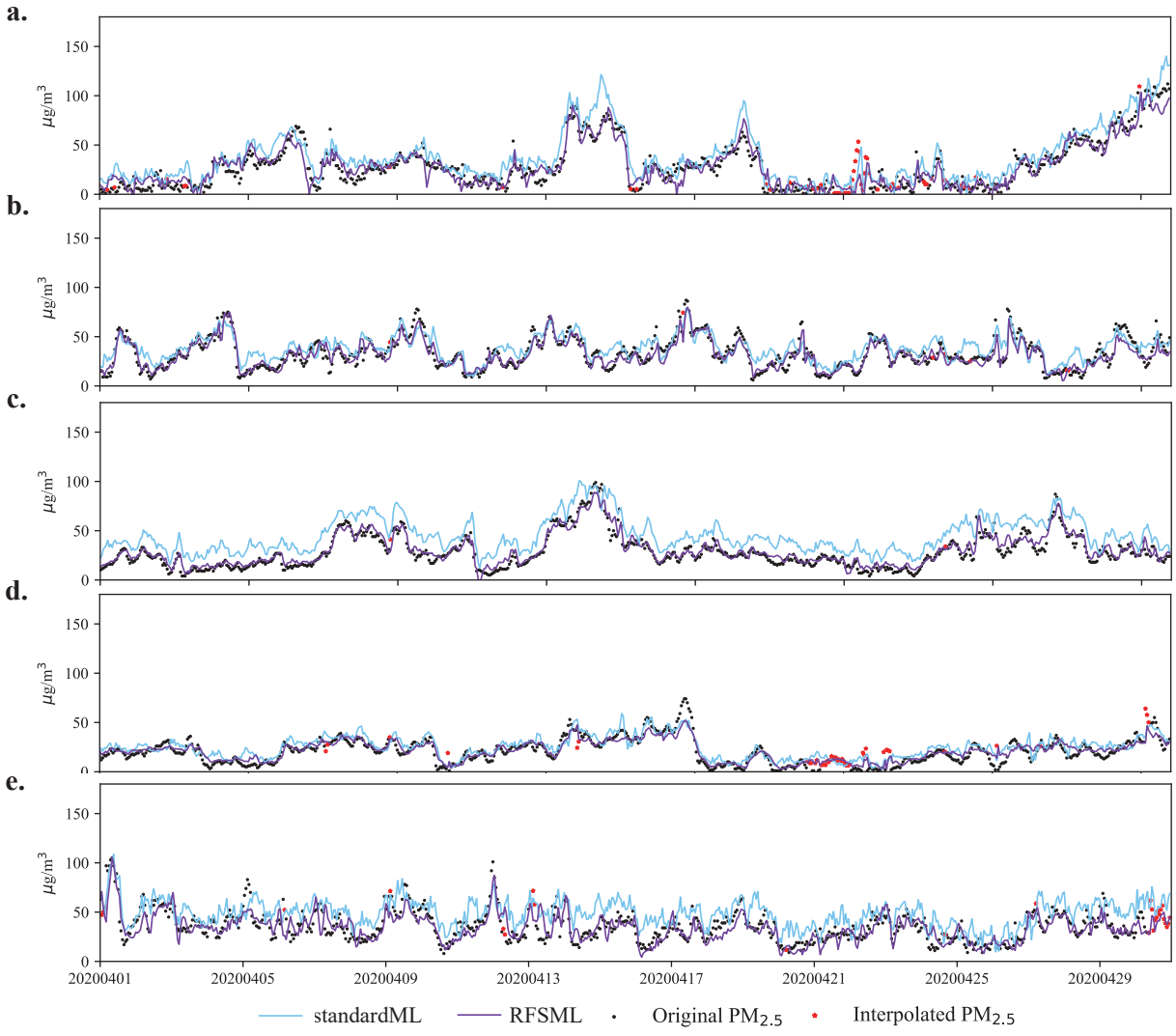

Figure S22. Time series of a prediction horizon of 6 hours in five mega-city cluster regions. The black dots and red pentacles represent original and interpolated $PM_{2.5}$ respectively. The solid lines with light sky blue and dark violet represent prediction of standard machine learning system and RFSML respectively. Panel a, b, c, d and e represent a random site in NCP, YRD, PRD, SCB and FWP respectively.

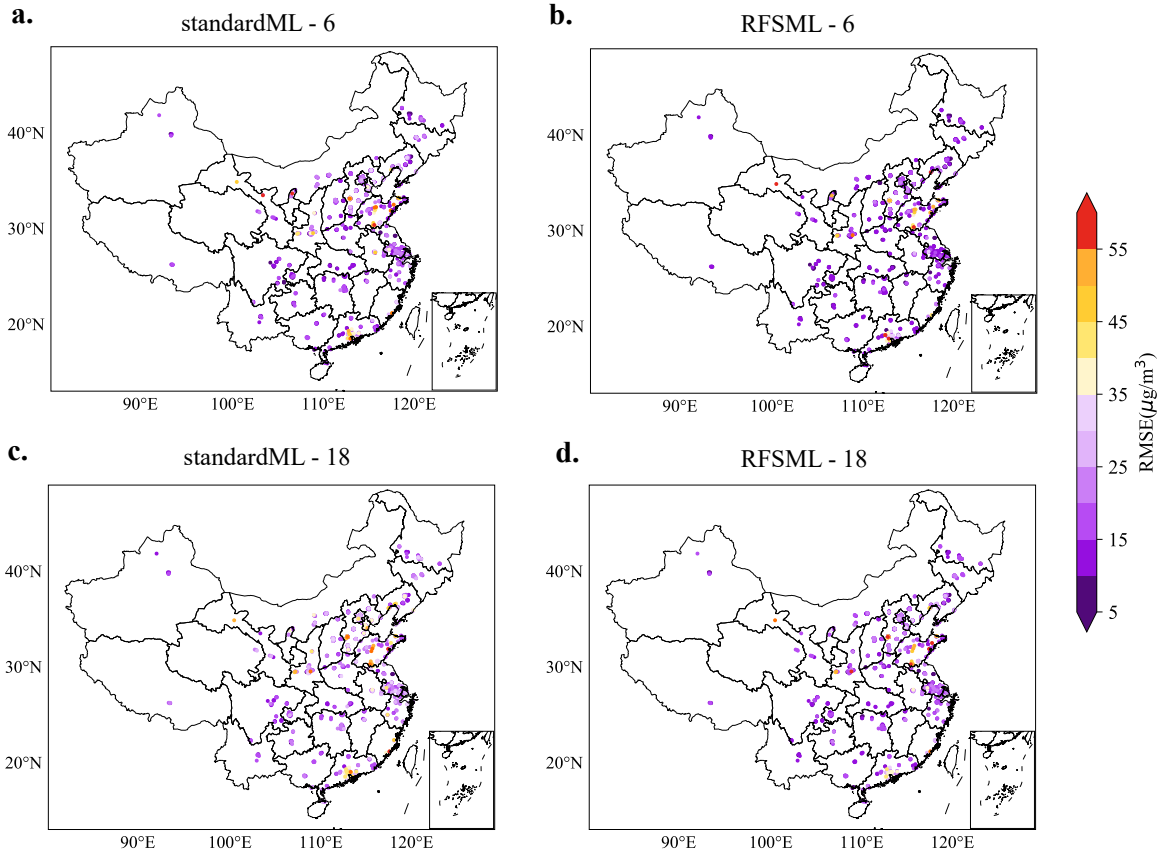

Figure S23. Spatial distribution of RMSE in a predicting horizon of 6 and 18 hours. Panel a and c are results of standard machine learning system while panel b and d are results of RFSML. The cooler the color tone, the lower the RMSE, thus the better predicting performance.

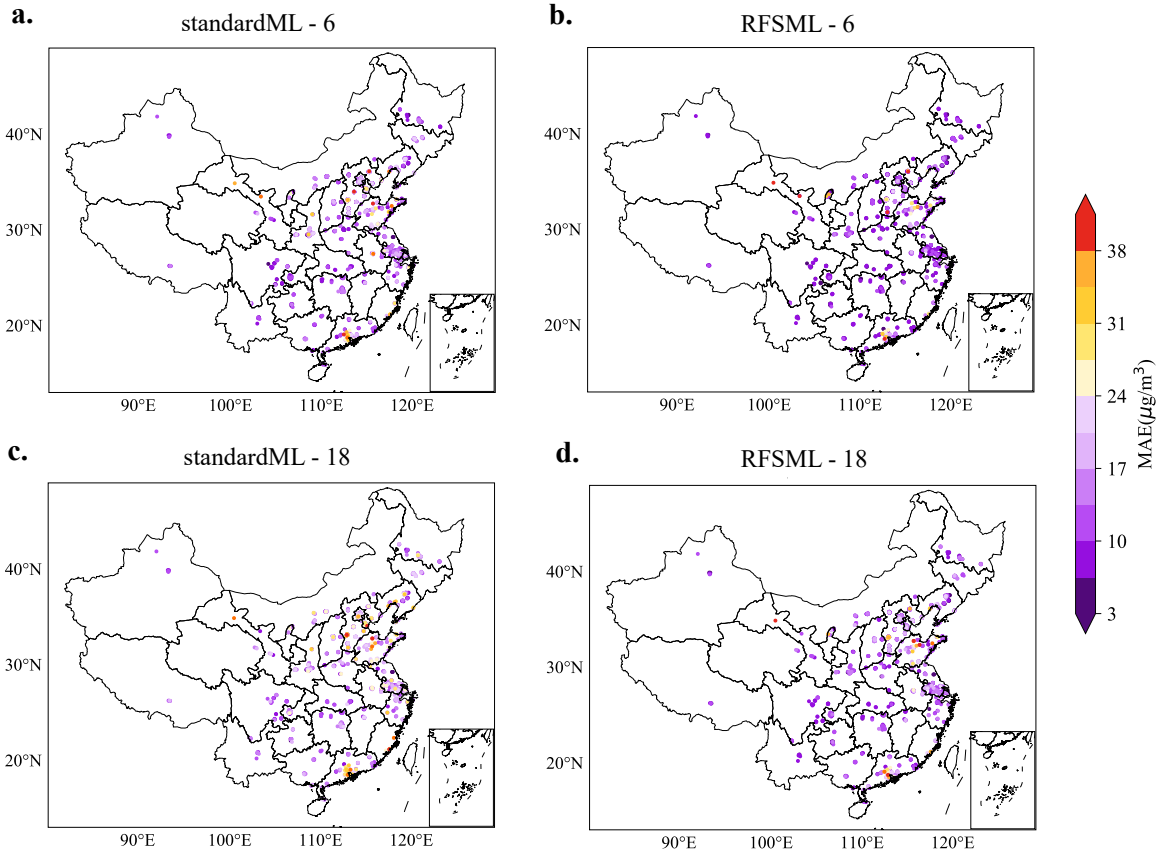

Figure S24. Spatial distribution of MAE in a predicting horizon of 6 and 18 hours. Panel a and c are results of standard machine learning system while panel b and d are results of RFSML. The cooler the color tone, the lower the MAE, thus the better predicting performance.