Journal of Environmental Management 303 (2022) 114210

Contents lists available at ScienceDirect

Journal of Environmental Management

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

Improvements of response surface modeling with self-adaptive machine learning method for $PM_{2.5}$ and O_3 predictions

Jinying Li [a,b], Youzhi Dai [a], Yun Zhu [b,*], Xiangbo Tang [c], Shuxiao Wang [d], Jia Xing [d], Bin Zhao [d], Shaojia Fan [e], Shicheng Long [b], Tingting Fang [b]

[a] College of Environment and Resources, Xiangtan University, Xiangtan, 411105, China
[b] College of Environment and Energy, South China University of Technology, Guangzhou Higher Education Mega Center, Guangzhou, 510006, China
[c] School of Frontier Crossover Studies, Hunan University of Technology and Business, Changsha, 410205, China
[d] State Key Joint Laboratory of Environmental Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing, 100084, China
[e] Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Sun Yat-Sen University, Zhuhai, 519000, China

ARTICLE INFO

Keywords:
Air quality model
Response surface model
Emissions control
Machine learning
Fine particles
Ozone

ABSTRACT

Quickly quantifying the $PM_{2.5}$ or O_3 response to their precursor emission changes is a key point for developing effective control policies. The polynomial function-based response surface model (pf-RSM) can rapidly predict the nonlinear response of $PM_{2.5}$ and O_3 to precursors, but has drawbacks of overload computation and marginal effects (relatively larger prediction errors under strict control scenarios). To improve the performance of pf-RSM, a novel self-adaptive RSM (SA-RSM) was proposed by integrating the machine learning-based stepwise regression for establishing robust models to increase the computational efficiency and the collinearity diagnosis for reducing marginal effects caused by overfitting. The pilot study case demonstrated that compared with pf-RSM, SA-RSM can effectively reduce the training number by 70% and 40% and the fitting time by 40% and 52%, and decrease the prediction error by 49% and 74% for $PM_{2.5}$ and O_3 predictions respectively; moreover, the isopleths of $PM_{2.5}$ or O_3 as a function of their precursors generated by SA-RSM were more similar to those derived by chemical transport model (CTM), after successfully addressing the marginal effect issue. With the improved computation efficiency and prediction performance, SA-RSM is expected as a better scientific tool for decision-makers to make sound $PM_{2.5}$ and O_3 control policies.

1. Introduction

Ambient fine particulate matters ($PM_{2.5}$) and ozone (O_3) have been regarded as major air quality evaluation indicators around the world because of their significant effects on human health and eco-environment (Cohen et al., 2017; Forouzanfar et al., 2016; Fuhrer et al., 2016; Murray et al., 2020). For both developing and developed countries like China and USA, the attainment of stringent ambient $PM_{2.5}$ and O_3 standards still requires different levels of reductions in precursors emissions (Wang et al., 2012; Zhang et al., 2020). The complexity of the physical and photochemical processes involving various precursors in the atmosphere leads to the strong nonlinear chemistry during the $PM_{2.5}$ and O_3 formation (Chen et al., 2019; Cohan et al., 2005; El-Harbawi, 2013; Lu et al., 2021). Therefore, it is important to accurately predict the nonlinear effects of precursor emission changes on $PM_{2.5}$ and O_3 concentrations to support the policymakers in developing effective control strategies for $PM_{2.5}$ and O_3.

Chemical transport models (CTMs) are commonly used to simulate the atmospheric processes and have been an essential tool for supporting air quality management by quantifying the impact of various emission sources on $PM_{2.5}$ and O_3 concentrations (Chatani et al., 2020; Duan et al., 2021; Lin et al., 2005). However, the methods of directly using the air quality model to quantify the impact are computationally expensive and do not allow rapid prediction of air quality responses to emission changes in the wide range of possible emission scenarios that are of interest to policymakers (Liu et al., 2021; Xing et al., 2011).

To improve the efficiency of air quality prediction, a response surface model (RSM) was firstly developed by the USEPA to characterize the relationship between the pollutant concentrations and precursors emissions using a limited number of CTM simulation experimental samples combined with advanced statistical methods (USEPA, 2006; Xing et al., 2011). RSM has the advantage of quickly predicting the

[*] Corresponding author.
E-mail address: zhuyun@scut.edu.cn (Y. Zhu).

https://doi.org/10.1016/j.jenvman.2021.114210
Received 26 August 2021; Received in revised form 24 November 2021; Accepted 28 November 2021

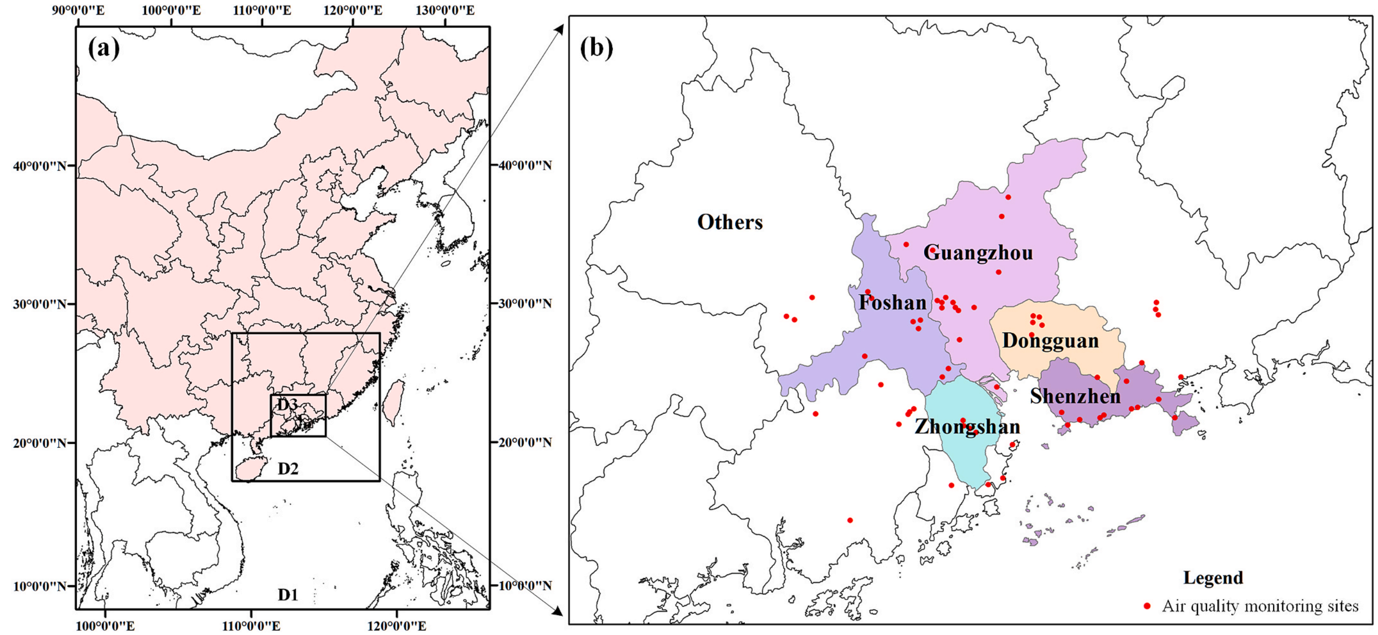

Fig. 1. (a) Three nested domains with 27 km, 9 km, and 3 km resolutions and (b) the innermost D3 domain. The red points are the locations of the state controlled air quality monitoring sites. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

response of pollutant concentrations to a wide range of emission changes (Wang et al., 2011). The RSM method has been continuously enhanced over the past several years to be more efficient and accurate, such as an extended version of the response surface model (ERSM) for multi-regional pollutant predictions (Xing et al., 2017; Zhao et al., 2015, 2017), a polynomial function-based RSM (pf-RSM) to substantially reduce the case number required to build the RSM and improve the ability for nonlinearity quantification (Xing et al., 2018), a deep learning approach (DeepRSM) based on convolutional neural networks (CNNs) to estimate the coefficients of the pf-RSM polynomial function using the ambient concentrations of chemical indicators simulated by CTM (Xing et al., 2020). However, the establishment of a single-regional ERSM and pf-RSM still requires at least 20 control scenarios to be simulated by the CTM (Xing et al., 2017, 2018). Although the DeepRSM substantially reduces the required case number of CTM simulations to only two (base and control scenarios respectively), there are plenty of scenarios of CTM simulations and pf-RSM runs that need to be trained preliminarily to build the DeepRSM, and the issue of high computing cost was still not well addressed (Xing et al., 2020). In addition, the accuracy of current RSM (e.g., pf-RSM and DeepRSM) is susceptible to the overfitting issues caused by the fixed fitting parameters by human experience. For instance, it has been previously reported that the RSM's predictions were relatively large biased from the CMAQ's simulations under strict emission control scenarios, and RSM also exhibited a relatively poor performance for predictions in the marginal areas where the precursor emission ratio is close to zero (Xing et al., 2011). The additional margin processing in the sampling method can improve the performance of RSM's predictions but only when the number of training samples is large enough (Xing et al., 2011). Hence, it is usually difficult to further reduce the number of CTM simulations while still achieving the high accuracy of RSM's predictions.

To address the aforementioned limitations, this study integrated the machine learning method named "stepwise regression" and the statistic method named "collinearity diagnosis" to improve pf-RSM forming a novel self-adaptive RSM (SA-RSM) (Chen et al., 2013; Cheng et al., 2012; Dormann et al., 2013; Hwang et al., 2015; Kerckhoffs et al., 2019; Salmerón Gómez et al., 2016; Stewart, 1987). The stepwise regression method can intelligently build a robust polynomial function-based model to reduce the error range caused by the fixed fitting parameters and improve the computation efficiency (Chen et al., 2013; Kerckhoffs et al., 2019). The collinearity diagnosis can be used for diagnosing the results of stepwise regression to effectively address the overfitting issues and reduce the marginal error (Li et al., 2011). The performance of SA-RSM is demonstrated in this presented study by comparatively evaluating the predictions of SA-RSM and pf-RSM against the CTM simulations in a case study over the Pearl River Delta region (PRD), China. The innovative SA-RSM are expected to provide better user experience for decision-makers and scientists to make sound emission control policies for lowing ambient $PM_{2.5}$ and O_3 concentration.

2. Methods

2.1. Model data and configuration setup

The ambient air quality was simulated by the Community Multiscale Air Quality Modeling System (CMAQ) in this study, and the meteorological conditions predicted by the Weather Research and Forecasting Model (WRF) were provided as the input for CMAQ simulations. Three nested model domains (D1, D2, and D3) with horizontal resolutions of 27 km, 9 km, 3 km, respectively, were utilized in the WRF-CMAQ modeling system (Fig. 1a), and 14 vertical layers were set for all domains. The innermost D3 region was divided into six sub-regions denoted as Guangzhou (GZ), Shenzhen (SZ), Foshan (FS), Dongguan (DG), Zhongshan (ZS), and Others (OTH) (Fig. 1b). Tsinghua University provided emission inventories for D1 and D2 domains (Wang et al., 2014; Zhao et al., 2018), and the 2017 emission inventory for the D3 domain was developed by the collaborative team from Tsinghua University and South China University of Technology. Initial conditions for D2 and D3 were generated from the simulations of D1 and D2, respectively, and a 5-day dormancy period was set to reduce the effect of initial conditions on simulation results. Version 6 of the Carbon Bond Mechanism (CB6) for aerosol extensions was utilized in the CMAQ meteorological chemistry module, and the AREO5 aerosol mechanism was employed in the aerosol module. Simulation periods of January and July 2017 were selected as the representative polluted months for $PM_{2.5}$ and O_3, respectively. The performance of CMAQ for $PM_{2.5}$ and O_3 simulations was evaluated by comparing the simulated baseline concentrations with the ground-based observations, and results showed that the

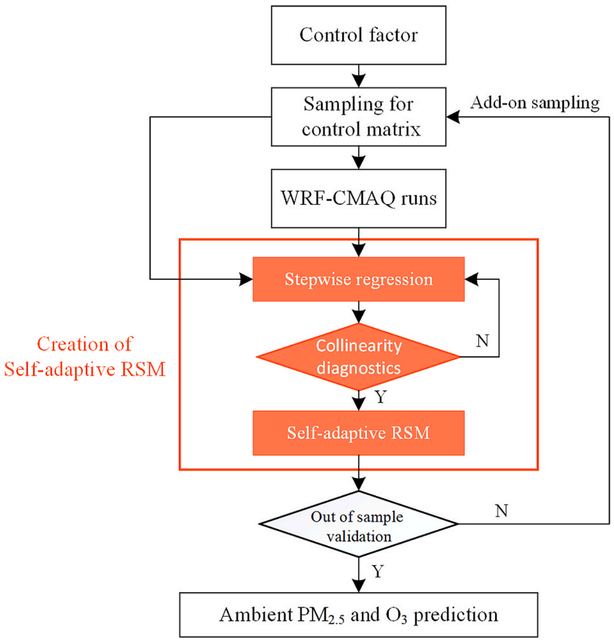

Fig. 2. The flow chart of the self-adaptive RSM (SA-RSM) for $PM_{2.5}$ and O_3 predictions.

correlation coefficient (R) between the model simulations and the observations for $PM_{2.5}$ and O_3 was greater than or equal to 0.4 and 0.6 in selected representative sites (Fig. S1 and Fig. S2), which generally satisfied the criteria of R for $PM_{2.5}$ and O_3 recommended by Emery et al. (2017). The NMB was within 25% for selected representative site (Fig. S1 and Fig. S2), which was also comparable to those in other publications (Fang et al., 2020; Pan et al., 2020). Consequently, the model performance was acceptable in this study.

2.2. *Development of self-adaptive response surface model*

The process of building the SA-RSM for $PM_{2.5}$ and O_3 predictions based on the machine learning approach is shown in Fig. 2. The emission control factors for $PM_{2.5}$ and O_3 were selected first and then sampled to form the control matrix consisting of various control scenarios. Next, the ambient $PM_{2.5}$ and O_3 concentrations in the PRD region under various scenarios were simulated using the WRF-CMAQ system. Then, the self-adaptive RSM (SA-RSM) was constructed based on the WRF-CMAQ simulations using the stepwise regression method combined with the collinearity diagnosis. Finally, the performance of SA-RSM was evaluated through the out of sample validation; if the validation results were desirable, the established SA-RSM can be applied for ambient $PM_{2.5}$ and O_3 prediction, otherwise, the SA-RSM will be re-fitted by increasing the number of samples.

2.2.1. *Stepwise regression*

Stepwise regression is a statistical method that has been previously successfully applied to enhance the precision and applicability of ambient pollutant concentrations forecasting models (Chen et al., 2013; Cheng et al., 2012). To illustrate the calculation process of the stepwise regression in this study, the establishment of the nonlinear relationship between $PM_{2.5}$ and its precursors was conducted as an example case. First, 15 terms constructing the polynomial responsive function of $PM_{2.5}$ were selected (Table S1), in which all product terms (Table S2) were regarded as the selectable terms; then the $PM_{2.5}$ concentrations were fitting with each selectable term separately using the linear regression function, and an F-test was performed on each selectable term to determine the magnitude of its ability for $PM_{2.5}$ prediction, in which the term with the largest F-test (Table S3) value will be chosen as the selected terms (Fig. 3. Step 1). Second, a partial F-test was performed on each selectable term to determine the magnitude of its ability for further enhancement of $PM_{2.5}$ prediction, in which the term with a partial F-test value greater than the threshold (with the initial value of 2) was considered as the desirable term, and the term with the largest partial F-test value among all the desirable terms was furtherly moved to the selected terms (Fig. 3. Step 2); then step 2 was repeated until there were more than two selected terms forming the initial polynomial responsive function. Next, the partial F-test was sequentially implemented for each selected term, and the term failed the partial F-test will be removed from the selected terms determined in the former steps (Fig. 3. Step 3). Finally, both step 2 and step 3 were repeated until all the selected terms passed the partial F-test and all the selectable terms failed the partial F-test, when the polynomial responsive function of $PM_{2.5}$ to its precursor emissions can be established by all the selected terms.

2.2.2. *Collinearity diagnostics*

Collinearity (also multicollinearity) is a phenomenon in which one independent variable can be linearly predicted from other independent variables in a multiple regression model (Stewart, 1987). Affected by the collinearity, the estimations of the multiple regression model will significantly and erratically change due to the small changes in the training data, thus reducing the stability and prediction performance of the regression model (Magel et al., 1987). Since different terms in the polynomial response function generated by the stepwise regression contain the same independent variable, as shown in step 4 of Fig. 3 where there are five terms in the polynomial response function that contain NO_x (i.e., $E_{NO_x}{}^4$, $E_{NO_x}{}^2$, E_{NOx}, $E_{NO_x}E_{NH_3}$, $E_{NO_x}E_{VOC}$), and the terms contain the same independent variables may have a high degree of collinearity, which will greatly decrease the prediction performance of the RSM established by the stepwise regression model (Dormann et al., 2013; Stewart, 1987). Therefore, utilization of the stepwise regression method alone to determine the combination of various desirable terms is insufficient to establish a robust RSM, and the evaluations of collinearity between different terms selected by the stepwise regression are necessary. Two commonly used statistical indexes, condition number and variance inflation factor (VIF) were utilized to diagnose the collinearity in this study (Jou et al., 2014; Salmerón et al., 2018), and the polynomial functions obtained by the stepwise regression must satisfy the requirement of both the condition number and VIF to pass the collinearity diagnosis. The final SA-RSM can be constructed based on these polynomial functions that pass the collinearity diagnosis. The collinearity diagnosis process is performed as below:

First, considering each term in the polynomial functions attained by the stepwise regression as an independent variable, a linear function of $y = Ax$ is established, and the condition number (κ) in this linear function can be calculated by Equation (1) subsequently.

$$\kappa(A) = \|A^{-1}\| \cdot \|A\| \tag{1}$$

where $\kappa(A)$ is condition number of the linear function of $y = Ax$; A is the coefficient matrix of the linear function of $y = Ax$.

Second, the VIF of this linear function can be estimated by Equation (2).

$$VIF = \max_i \{\frac{1}{1 - R_i^2}\} \ i \in 1, \cdots, n \tag{2}$$

where R_i^2 is the determination coefficient of the linear function composed of the i-th independent variable and the other independent variables; n is the number of independent variables.

If the condition number $\kappa(A)$ is greater than 100 or the VIF value is greater than 10 (Lazaridis, 2007; Salmerón et al., 2018), the selected terms (selected independent variable) will be considered as failing the collinearity diagnosis. Then, the threshold of the partial F-test value will be scaled up to 1.5 times of the original value (Hwang et al., 2015), and the stepwise regression will be restarted until the evaluation result of

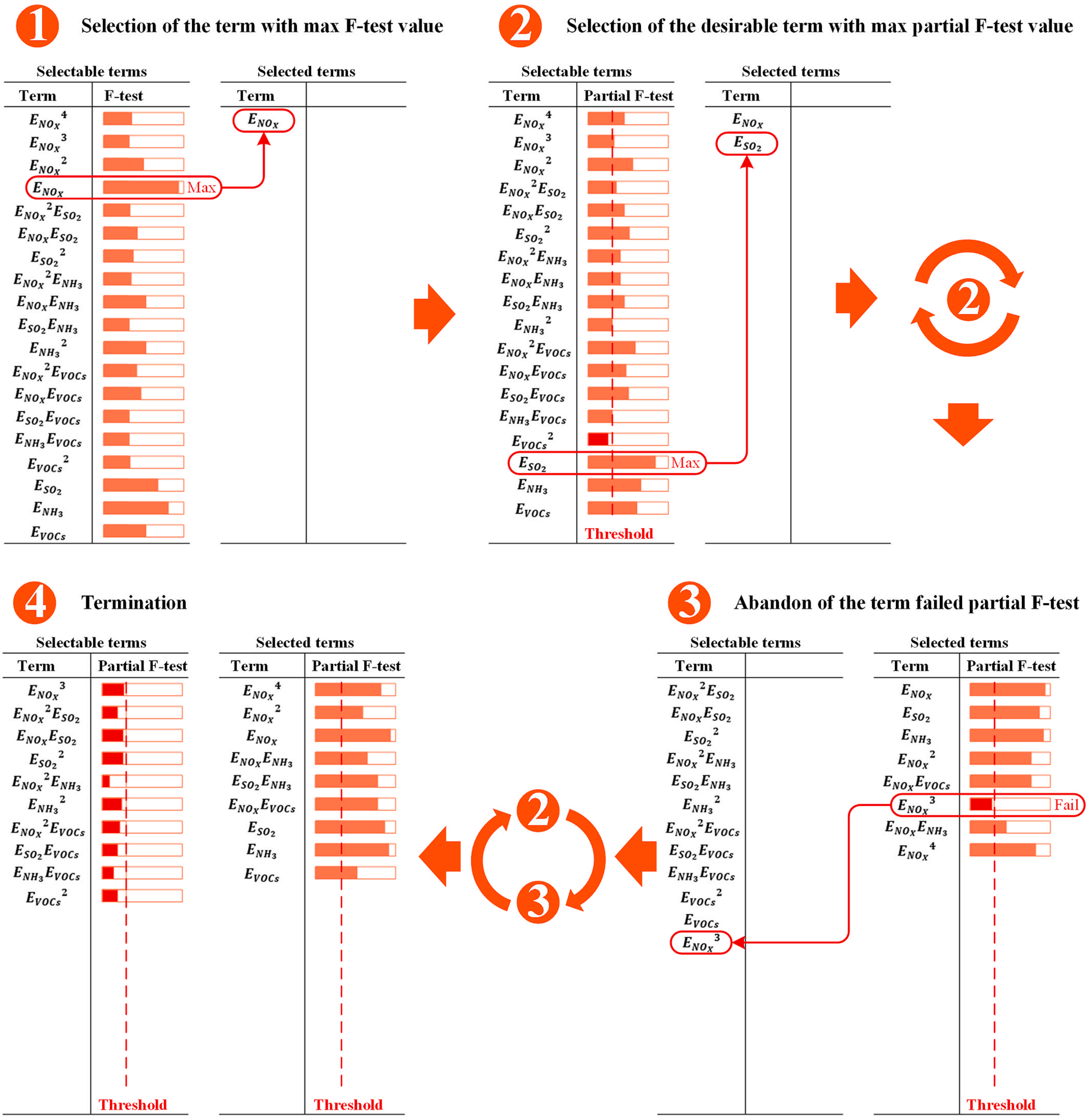

Fig. 3. The stepwise regression process for $PM_{2.5}$ analysis.

this selected terms passes the collinearity diagnosis. Finally, the SA-RSM will be established based on all the selected terms in the polynomial functions that passed the collinearity diagnosis.

3. Results and discussion

3.1. Improved training efficiency

To examine the improvement on the prediction accuracy of SA-RSM, the mean relative errors (MRE) of SA-RSM and pf-RSM for $PM_{2.5}$ and O_3 prediction under different numbers of training samples were compared in Fig. 4a. It can be seen that the SA-RSM can achieve comparable performance to the pf-RSM with a smaller number of training samples for both the $PM_{2.5}$ and O_3 predictions. For example, the MRE of pf-RSM for $PM_{2.5}$ and O_3 predictions were 1.5% and 2.5% respectively under 20 training samples, while to obtain the comparable MRE performance, the required training samples for SA-RSM to fit the $PM_{2.5}$ and O_3 are only 6 and 12 respectively. The minimum training samples of pf-RSM is 15 because the polynomial function used to construct the pf-RSM has 15 terms (Table S1), while SA-RSM, due to the improved training efficiency, can get the effective terms with fewer (e.g., 6) samples than the pf-RSM (e.g., 15). In addition, we notice that the MRE of either the SA-RSM or pf-RSM for O_3 is larger than that for $PM_{2.5}$ under the same number of training samples. It was because that the O_3 was an entirely

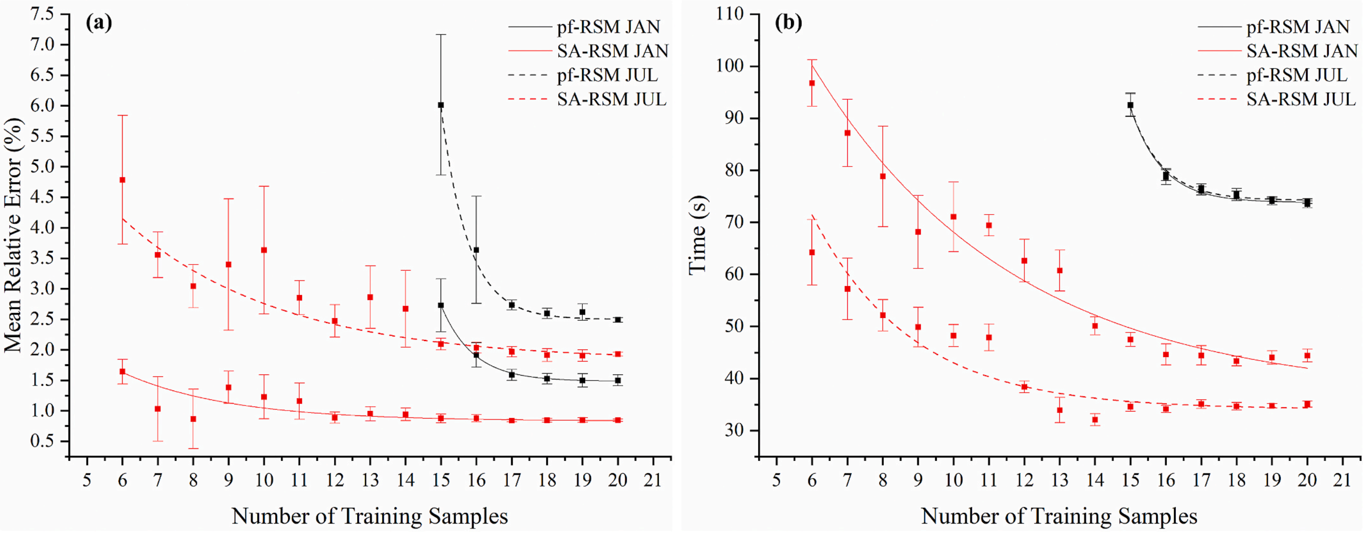

Fig. 4. Mean relative error (a) and time consumed (b) of SA-RSM and pf-RSM for $PM_{2.5}$ and O_3 predictions under different numbers of training samples. Error bars indicate standard deviation of five repetitions.

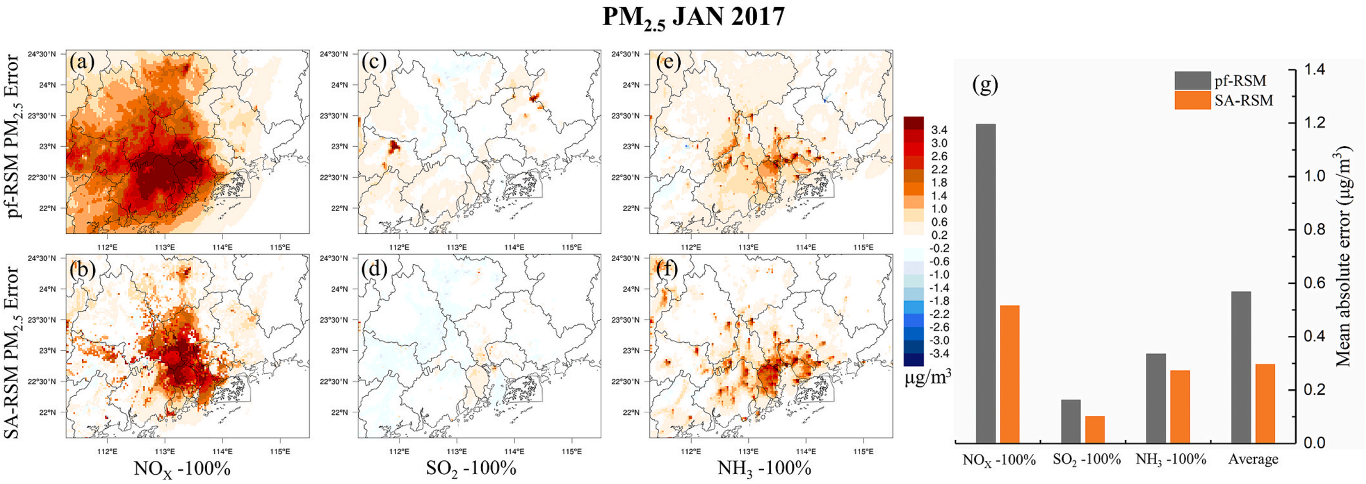

Fig. 5. (a)–(f) Spatial distribution of the error and (g) the MAE of pf-RSM and SA-RSM for $PM_{2.5}$ predictions across the entire domain in January 2017, PRD. The MAE is defined as the mean absolute error of the predicted monthly averaged concentration compared with that simulated by the CMAQ. Both pf-RSM and SA-RSM were established by 20 training samples.

secondary pollutant generated by the complex photochemical reactions among different precursors, while the $PM_{2.5}$ was partially emitted from some primary sources (e.g., the primary $PM_{2.5}$ emissions contributed 67.9–78.6% to $PM_{2.5}$ concentrations in the PRD as reported by our previous research) (Li et al., 2021). Hence, O_3 exhibited a stronger nonlinear relationship with its precursors compared to $PM_{2.5}$, making the ambient O_3 more difficult to predict than the $PM_{2.5}$ for both SA-RSM and pf-RSM.

The variation of SA-RSM's and pf-RSM's fitting time for $PM_{2.5}$ and O_3 were also compared under different numbers of training samples, as shown in Fig. 4b. It is found that the fitting time of SA-RSM is significantly lower than that of pf-RSM under the same number of training samples; for example, under the 20 training samples, the fitting time for $PM_{2.5}$ and O_3 is reduced from 73.5 s to 74.1 s by the pf-RSM to 44.4 s and 35.5 s by the SA-RSM, with a decrease rate of 40% and 52% respectively. It is also noticed that the pf-RSM take similar time to fit the $PM_{2.5}$ and O_3, while the SA-RSM consumes obviously less time to fit the O_3 compared with the $PM_{2.5}$. This was because the polynomial function used by the pf-RSM for both the $PM_{2.5}$ and O_3 fitting consisted of 15 fixed terms (Table S1), which caused the similar computation time; while the SA-RSM optimized the fitting process through the stepwise regression to effectively choose useful terms, and there were more kinds of $PM_{2.5}$ precursors (i.e., NO_x, SO_2, NH_3, VOC) than O_3 precursors (i.e., NO_x, VOC), making the optimization process for $PM_{2.5}$ fitting more complicated and time-consuming.

3.2. Enhanced prediction accuracy

To evaluate the enhancement on the performance of SA-RSM for $PM_{2.5}$ predictions, the spatial distribution of the mean absolute error (MAE) of pf-RSM and SA-RSM were compared under the 100% emission reductions of NO_x, SO_2 and NH_3 (Table S4), individually. As shown in Fig. 5, the prediction error of SA-RSM mainly lay in the central areas of the PRD, and the MAE value of SA-RSM is lower than that of pf-RSM in over 70% areas of the PRD. The advantage of SA-RSM is especially obvious under the 100% NO_x emission reduction scenario, in which the MAE over the entire D3 region is largely reduced from 1.19 μg/m^3 by the pf-RSM to 0.51 μg/m^3 by the SA-RSM (with a decrease rate of 57%). Moreover, under the 100% emission reductions of SO_2 and NH_3, the MAE for $PM_{2.5}$ predictions is also reduced from 0.16 μg/m^3 and 0.33 μg/

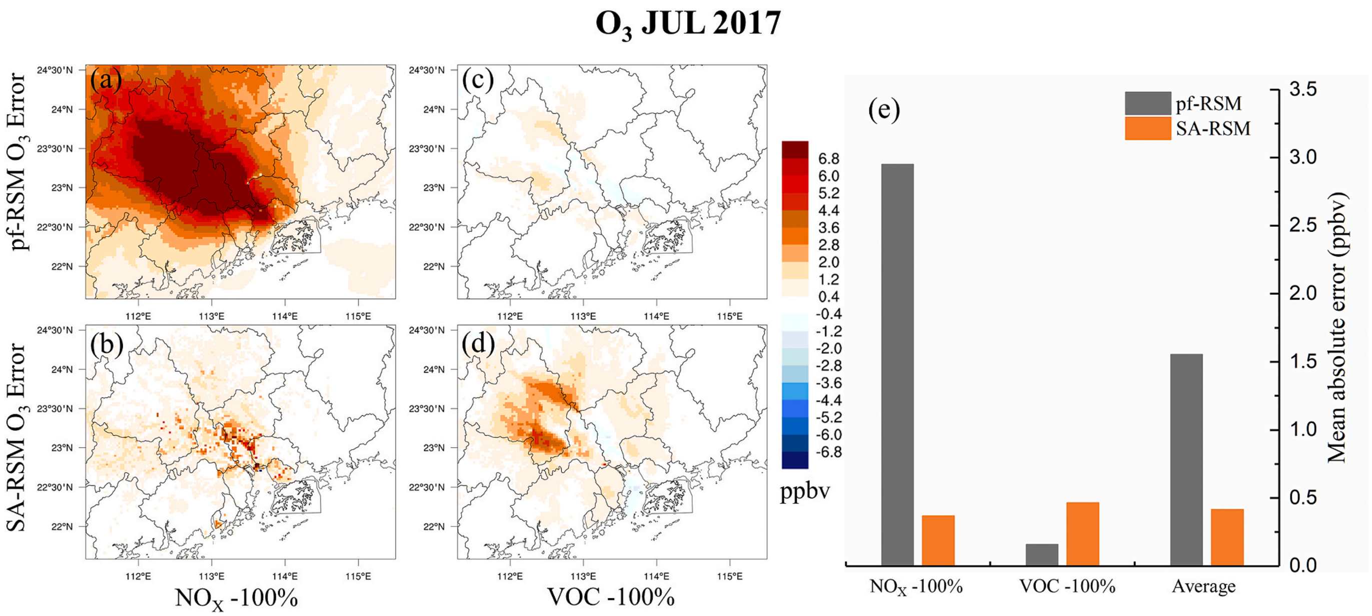

Fig. 6. (a)–(d) Spatial distribution of the error and (e) the MAE of pf-RSM and SA-RSM for O_3 predictions across the entire domain in July 2017, PRD. The MAE is defined as the mean absolute error of the predicted monthly averaged daily 8 h-max concentration compared with that simulated by the CMAQ. Both pf-RSM and SA-RSM were established by 20 training samples.

m^3 by the pf-RSM to 0.10 μg/m^3 and 0.27 μg/m^3 by the SA-RSM (with a decrease rate of 38% and 18%). As a result, the average MAE under NO_x, SO_2 and NH_3 individual 100% emission reduction scenarios is reduced from 0.57 μg/m^3 to 0.29 μg/m^3 with an average decrease rate of 49%. In addition, to evaluate the improvement of SA-RSM for predicting individual secondary $PM_{2.5}$ components (i.e., nitrate, sulfate, and

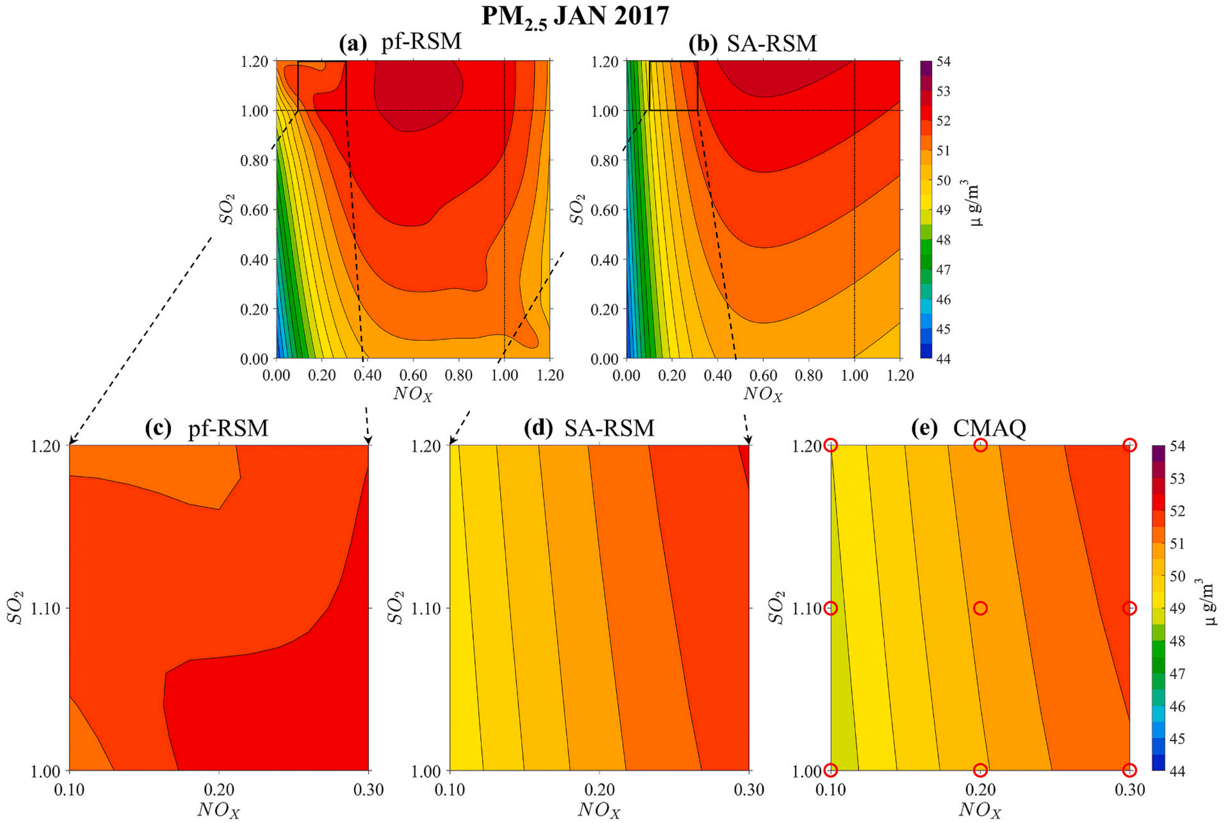

Fig. 7. Isopleths of the $PM_{2.5}$ response to NO_x and SO_2 emission changes in January 2017 in the entire D3 domain predicted by pf-RSM (a) and SA-RSM (b), respectively, and the $PM_{2.5}$ isopleths in the marginal areas predicted by pf-RSM (c) and SA-RSM (d), respectively, and the $PM_{2.5}$ isopleths in the marginal areas directly interpolated from the 9 CMAQ simulation scenarios as marked by the red circle (e). The predicted $PM_{2.5}$ response is monthly averaged across all monitoring sites of the PRD; the x and y-axes represent the emission ratio of NO_x and SO_2 respectively for the entire D3 domain, and the baseline emission ratio is equal to 1. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

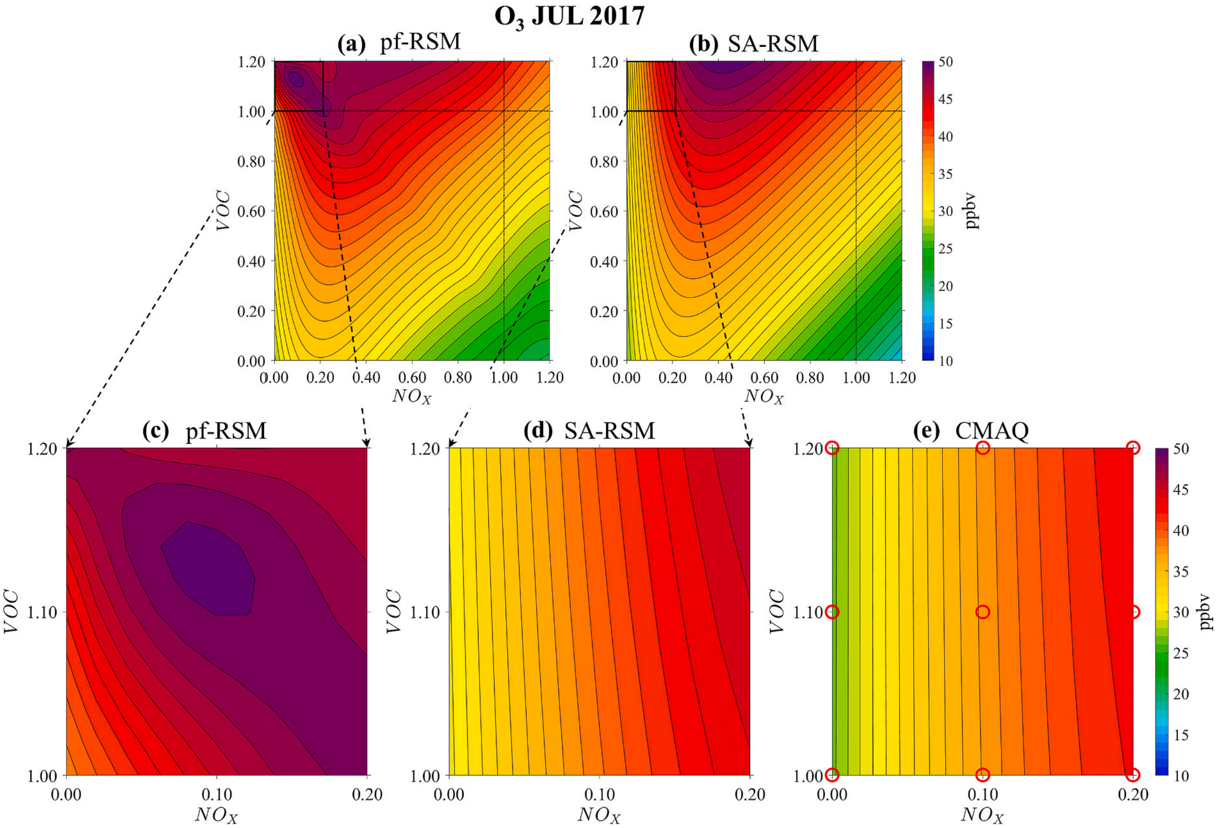

Fig. 8. The same as Fig. 7 but for the O_3 response to NO_x and VOC emission changes in July 2017.

ammonium), the overall MAE of pf-RSM and SA-RSM under 5 randomly selected control scenarios (Table S5) for $PM_{2.5}$ component predictions were compared in Fig. S3. It can be seen that the MAE of SA-RSM for each $PM_{2.5}$ component was also obviously lower than that of the pf-RSM, indicating that the SA-RSM had an improvement over the pf-RSM in both the total $PM_{2.5}$ and $PM_{2.5}$ components predictions.

The enhancement in the performance of SA-RSM for O_3 predictions is also evaluated by comparing the prediction error of pf-RSM and SA-RSM under the 100% emission reductions of NO_x and VOC respectively. As shown in Fig. 6, in July, under the 100% NO_x emission reduction scenario, SA-RSM can overall reduce the O_3 prediction error over the entire D3 domain, especially at the downwind areas of the PRD affected by the southeast monsoon in July. The MAE is significantly reduced from 2.95 ppbv by the pf-RSM to 0.37 ppbv by the SA-RSM (with a decrease rate of 87%). Since the polynomial function formed the pf-RSM contained some NO_x terms with high degrees for capturing the nonlinear response of $PM_{2.5}$ or O_3 to NO_x emissions (e.g., the $PM_{2.5}$ fitting equation contained a NO_x term with 4th power, and the O_3 fitting equation contained a NO_x term with 5th power), as shown in Table S1. These higher degrees will lead to a significant overfitting phenomenon and increase the error of pf-RSM under NO_x reduction scenarios (Xing et al., 2018), while the SA-RSM can effectively address this overfitting issue through the stepwise regression method combined with the collinearity diagnosis.

Under the 100% VOC reduction scenario, both pf-RSM and SA-RSM exhibit a desirable performance for O_3 predictions, with the MAE of 0.16 ppbv and 0.46 ppbv respectively, but pf-RSM perform a little better than SA-RSM in this condition (Fig. 6e). It was because that on the one hand, the VOC contribution to O_3 in July was small (less than 15%) as reported by Fang et al. (2020), therefore, the slight response of O_3 to VOC emission changes in July was a little hard to be identified by the stepwise regression and causing a slight error; on the other hand, SA-RSM was based on the iterative calculations to select the appropriate polynomial responsive function for establishing the responsive relationship between VOC and O_3, causing the error accumulation during the iterative calculation; while the pf-RSM was fit with the fixed function through once calculation and the error arising from a slight VOC contribution was small and will not accumulate. Although the MAE of SA-RSM is slightly larger than that of pf-RSM for VOC reduction scenario, the average MAE of SA-RSM under NO_x and VOC individual emission reduction scenarios is still greatly reduced by 74% compared to pf-RSM (from 1.55 ppbv by the pf-RSM to 0.41 ppbv by the SA-RSM).

We further performed the October PRD case, when both VOC and NO_x had a significant contribution on O_3 (the contribution of VOC was more than 52%) (Fang et al., 2020). As shown in Fig. S4, SA-RSM significantly reduces the prediction error of pf-RSM over the entire region (especially at the downwind areas affected by the northeast monsoon in October 2017) under NO_x and VOC individual reduction scenarios (reduced by 90% and 77% respectively).

3.3. Resolved marginal effects issue

The current RSM (e.g., pf-RSM and DeepRSM) was reported to have relatively poor performance at the marginal areas (Xing et al., 2011, 2018), for example the areas of NO_X emissions nearing 0 and SO_2 or VOCs emissions nearing 1.2 as shown in Fig. 7 or Fig. 8. In order to investigate the improvement of SA-RSM in addressing the marginal effects issue, the $PM_{2.5}$ isopleths predicted by pf-RSM and SA-RSM at the marginal areas were compared with that directly derived from the CMAQ simulations. The results show that the distribution of the response surface in the $PM_{2.5}$ isopleths obtained by pf-RSM and SA-RSM is generally similar (Fig. 7a and b), but the varying trend of the pf-RSM's prediction is distorted especially at the marginal areas, which shows significant overfitting phenomenon (Fig. 7c). In contrast, the SA-RSM predicted response surface is similar to CMAQ (Fig. 7e) which is much smoother than that predicted by the pf-RSM (Fig. 7c and d). The isopleths' shape of the $PM_{2.5}$ as a function of NO_x and SO_2 emissions created by SA-RSM is also similar to the reported ones in the publications (Blanchard et al., 2007; Chen et al., 2017).

Fig. 8 shows the isopleths of the O_3 as a function of NO_x and VOC emission changes in PRD predicted by pf-RSM and SA-RSM respectively, and the predictions of two methods at the marginal areas were also compared with the CMAQ simulations. The overall trend of the response surfaces established by SA-RSM and pf-RSM is similar (Fig. 8a and b), but the pf-RSM still exhibits significant overfitting at marginal areas (Fig. 8c), while the response surface of O_3 established by SA-RSM is smoother and more consistent with the results of CMAQ simulation at the marginal areas (Fig. 8c, d and e) and is more consistent with the previously reported O_3 isopleths (Guo et al., 2019; Luo et al., 2021).

4. Conclusions

In this study, a novel SA-RSM was developed by integrating the machine learning-based methods of stepwise regression and the statistic method of collinearity diagnosis. SA-RSM was successfully applied for $PM_{2.5}$ and O_3 predictions in the Pearl River Delta region, China.

Our results suggested that there are three main improvements of SA-RSM over the pf-RSM. First, SA-RSM can effectively reduce the computational burden and improve the training efficiency; for $PM_{2.5}$ and O_3 predictions, the training samples were reduced by 70% (from 20 to 6) and 40% (from 20 to 12), and the fitting time was decreased by 40% (from 73.5 s to 44.4 s) and 52% (from 74.1 s to 35.5 s) respectively. Second, SA-RSM was also able to significantly improve the overall prediction accuracy for $PM_{2.5}$ and O_3 compared with the pf-RSM, with the average MAE reduced by 49% (from 0.57 $\mu g/m^3$ to 0.29 $\mu g/m^3$) and 74% (from 1.55 ppbv to 0.41 ppbv) for $PM_{2.5}$ and O_3 predictions, respectively. Lastly, SA-RSM was demonstrated to well resolve the marginal effects issue of pf-RSM by avoiding the overfitting phenomenon, and generated the isopleths of $PM_{2.5}$ or O_3 as a function of their precursors that were more similar to the ones derived by CTM. Accordingly, the newly developed SA-RSM can achieve an effective improvement on the computation efficiency and prediction performance over the pf-RSM, and the application of SA-RSM is expected to provide a better user experience for decision-makers and scientists to form sound emission control policies for lowing ambient $PM_{2.5}$ and O_3.

Credit author statement

Jinying Li: Conceptualization, Methodology, Modeling, Software, Investigation, Writing – original draft.**Youzhi Dai**: Writing – review & editing. **Yun Zhu**: Resources, Writing – review & editing, Supervision, Project administration, Data curation. **Xiangbo Tang**: Writing – review & editing. **Shuxiao Wang**: Resources, Writing – review & editing, Data curation. **Jia Xing**: Resources, Writing – review & editing, Data curation. **Bin Zhao**: Writing – review & editing. **Shaojia Fa**n: Writing – review & editing. **Shicheng Long**: Validation, Formal analysis, Visualization, Software. **Tingting Fang**: Validation, Formal analysis, Visualization, Software.

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 work was supported by the Science and Technology Program of Guangzhou, China (No. 202002030188), the National Natural Science Foundation of China (No. 72174060), the Innovation Group Project of Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai) (No. 311021001). The authors sincerely thank Carey Jang and James Kelly (U.S. Environmental Protection Agency) for their assistance in preparation and management of this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jenvman.2021.114210.

References

Blanchard, C.L., Tanenbaum, S., Hidy, G.M., 2007. Effects of sulfur dioxide and oxides of nitrogen emission reductions on fine particulate matter mass concentrations: regional comparisons. J. Air Waste Manag. Assoc. 57 (11), 1337–1350. https://doi.org/10.3155/1047-3289.57.11.1337.

Chatani, S., Yamaji, K., Itahashi, S., Saito, M., Takigawa, M., Morikawa, T., Kanda, I., Miya, Y., Komatsu, H., Sakurai, T., et al., 2020. Identifying key factors influencing model performance on ground-level ozone over urban areas in Japan through model inter-comparisons. Atmos. Environ. 223, 117255. https://doi.org/10.1016/j.atmosenv.2019.117255.

Chen, T.F., Chang, K.H., Tsai, C.Y., 2017. Modeling approach for emissions reduction of primary PM2.5 and secondary PM2.5 precursors to achieve the air quality target. Atmos. Res. 192, 11–18. https://doi.org/10.1016/j.atmosres.2017.03.018.

Chen, Y., Fung, J.C.H., Chen, D.H., Shen, J., Lu, X.C., 2019. Source and exposure apportionments of ambient PM2.5 under different synoptic patterns in the Pearl River Delta region. Chemosphere 236, 124266. https://doi.org/10.1016/j.chemosphere.2019.06.236.

Chen, Y.Y., Shi, R.H., Shu, S.J., Gao, W., 2013. Ensemble and enhanced PM10 concentration forecast model based on stepwise regression and wavelet analysis. Atmos. Environ. 74, 346–359. https://doi.org/10.1016/j.atmosenv.2013.04.002.

Cheng, S.Y., Zhou, Y., Li, J.B., Lang, J.L., Wang, H.Y., 2012. A new statistical modeling and optimization framework for establishing high-resolution PM10 emission inventory – I. Stepwise regression model development and application. Atmos. Environ. 60, 613–622. https://doi.org/10.1016/j.atmosenv.2012.07.056.

Cohan, D.S., Hakami, A., Hu, Y.T., Russell, A.G., 2005. Nonlinear response of ozone to emissions:source apportionment and sensitivity analysis. Environ. Sci. Technol. 39 (17), 6739–6748. https://doi.org/10.1021/es048664m.

Cohen, A.J., Brauer, M., Burnett, R., Anderson, H.R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., et al., 2017. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. Lancet 389 (10082), 1907–1918. https://doi.org/10.1016/S0140-6736(17)30505-6.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36 (1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x.

Duan, W.J., Wang, X.Q., Cheng, S.Y., Wang, R.P., Zhu, J.X., 2021. Influencing factors of PM2.5 and O3 from 2016 to 2020 based on DLNM and WRF-CMAQ. Environ. Pollut. 285, 117512. https://doi.org/10.1016/j.envpol.2021.117512.

El-Harbawi, M., 2013. Air quality modelling, simulation, and computational methods: a review. Environ. Rev. 21 (3), 149–179. https://doi.org/10.1139/er-2012-0056.

Emery, C., Liu, Z., Russell, A.G., Odman, M.T., Yarwood, G., Kumar, N., 2017. Recommendations on statistics and benchmarks to assess photochemical model performance. J. Air Waste Manag. Assoc. 67 (5), 582–598. https://doi.org/10.1080/10962247.2016.1265027.

Fang, T.T., Zhu, Y., Jang, J.C., Wang, S.X., Xing, J., Chiang, P.C., Fan, S.J., You, Z.Q., Li, J.Y., 2020. Real-time source contribution analysis of ambient ozone using an enhanced meta-modeling approach over the Pearl River Delta Region of China. J. Environ. Manag. 268, 110650. https://doi.org/10.1016/j.jenvman.2020.110650.

Forouzanfar, M.H., Afshin, A., Alexander, L.T., Anderson, H.R., Bhutta, Z.A., Biryukov, S., Brauer, M., Burnett, R., Cercy, K., Charlson, F.J., et al., 2016. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet 388 (10053), 1659–1724. https://doi.org/10.1016/S0140-6736(16)31679-8.

Fuhrer, J., Val Martin, M., Mills, G., Heald, C.L., Harmens, H., Hayes, F., Sharps, K., Bender, J., Ashmore, M.R., 2016. Current and future ozone risks to global terrestrial biodiversity and ecosystem processes. Ecol. Evol. 6 (24), 8785–8799. https://doi.org/10.1002/ece3.2568.

Guo, H., Chen, K.Y., Wang, P.F., Hu, J.L., Ying, Q., Gao, A.F., Zhang, H.L., 2019. Simulation of summer ozone and its sensitivity to emission changes in China. Atmospheric Pollut. Res. 10 (5), 1543–1552. https://doi.org/10.1016/j.apr.2019.05.003.

Hwang, J.-S., Hu, T.-H., 2015. A stepwise regression algorithm for high-dimensional variable selection. JSCS 85 (9), 1793–1806. https://doi.org/10.1080/00949655.2014.902460.

Jou, Y.J., Huang, C.C.L., Cho, H.J., 2014. A VIF-based optimization model to alleviate collinearity problems in multiple linear regression. Comput. Stat. 29 (6), 1515–1541. https://doi.org/10.1007/s00180-014-0504-3.

Kerckhoffs, J., Hoek, G., Portengen, L., Brunekreef, B., Vermeulen, R.C.H., 2019. Performance of prediction algorithms for modeling outdoor air pollution spatial surfaces. Environ. Sci. Technol. 53 (3), 1413–1421. https://doi.org/10.1021/acs.est.8b06038.

Lazaridis, A., 2007. A note regarding the condition number: the case of spurious and latent multicollinearity. Qual. Quantity 41 (1), 123–135. https://doi.org/10.1007/s11135-005-6225-5.

Li, Y.T., Tsung, F., 2011. Detecting and diagnosing covariance matrix changes in multistage processes. IIE Trans. 43 (4), 259–274. https://doi.org/10.1080/0740817X.2010.521805.

Li, Z.F., Zhu, Y., Wang, S.X., Xing, J., Zhao, B., Long, S.C., Li, M.H., Yang, W.W., Huang, R.L., Chen, Y., 2021. Source contribution analysis of PM2.5 using response surface model and particulate source apportionment Technology over the PRD region. China. Sci. Total Environ. 151757. https://doi.org/10.1016/j.scitotenv.2021.151757.

Lin, C.J., Ho, T.C., Chu, H.W., Yang, H., Chandru, S., Krishnarajanagar, N., Chiou, P., Hopper, J.R., 2005. Sensitivity analysis of ground-level ozone concentration to emission changes in two urban regions of southeast Texas. J. Environ. Manag. 75 (4), 315–323. https://doi.org/10.1016/j.jenvman.2004.09.012.

Liu, C., Zhang, H.R., Cheng, Z., Shen, J.Y., Zhao, J.H., Wang, Y.C., Wang, S., Cheng, Y., 2021. Emulation of an atmospheric gas-phase chemistry solver through deep learning: case study of Chinese Mainland. Atmospheric Pollut. Res. 12 (6), 101079. https://doi.org/10.1016/j.apr.2021.101079.

Lu, M.M., Tang, X., Feng, Y.C., Wang, Z.F., Chen, X.H., Kong, L., Ji, D.S., Liu, Z.R., Liu, K. X., Wu, H.J., et al., 2021. Nonlinear response of SIA to emission changes and chemical processes over eastern and central China during a heavy haze month. Sci. Total Environ. 788, 147747. https://doi.org/10.1016/j.scitotenv.2021.147747.

Luo, H.H., Zhao, K.H., Yuan, Z.B., Yang, L.F., Zheng, J.Y., Huang, Z.J., Huang, X.B., 2021. Emission source-based ozone isopleth and isosurface diagrams and their significance in ozone pollution control strategies. J. Environ. Sci. 105, 138–149. https://doi.org/10.1016/j.jes.2020.12.033.

Magel, R.C., Hertsgaard, D., 1987. A collinearity diagnostic for nonlinear regression. Commun. Stat. Simulat. Comput. 16 (1), 85–97. https://doi.org/10.1080/03610918708812579.

Murray, C.J.L., Aravkin, A.Y., Zheng, P., 2020. Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 396 (10258), 1223–1249. https://doi.org/10.1016/S0140-6736(20)30752-2.

Pan, Y.Z., Zhu, Y., Jang, J.C., Wang, S.X., Xing, J., Chiang, P.C., Zhao, X.T., You, Z.Q., Yuan, Y.Z., 2020. Source and sectoral contribution analysis of PM2.5 based on efficient response surface modeling technique over Pearl River Delta Region of China. Sci. Total Environ. 737, 139655. https://doi.org/10.1016/j.scitotenv.2020.139655.

Salmerón Gómez, R., García Pérez, J., López Martín, M.D.M., García, C.G., 2016. Collinearity diagnostic applied in ridge estimation through the variance inflation factor. J. Appl. Stat. 43 (10), 1831–1849. https://doi.org/10.1080/02664763.2015.1120712.

Salmerón, R., García, C.B., García, J., 2018. Variance inflation factor and condition number in multiple linear regression. JSCS 88 (12), 2365–2384. https://doi.org/10.1080/00949655.2018.1463376.

Stewart, G.W., 1987. Collinearity and least squares regression. Stat. Sci. 2 (1), 68–84. https://doi.org/10.1214/ss/1177013439.

USEPA, 2006. Technical Support Document for the Proposed PM NAAQS Rule: Response Surface Modeling: Office of Air Quality Planning and Standards. https://www.epa.gov/sites/production/files/2020-10/documents/pmnaaqs_tsd_rsm_all_021606.pdf.

Wang, S., Xing, J., Jang, C., Zhu, Y., Fu, J.S., Hao, J., 2011. Impact assessment of ammonia emissions on inorganic aerosols in east China using response surface modeling technique. Environ. Sci. Technol. 45 (21), 9293–9300. https://doi.org/10.1021/es2022347.

Wang, S.X., Hao, J.M., 2012. Air quality management in China: issues, challenges, and options. J. Environ. Sci. 24 (1), 2–13. https://doi.org/10.1016/S1001-0742(11)60724-9.

Wang, S.X., Zhao, B., Cai, S.Y., Klimont, Z., Nielsen, C.P., Morikawa, T., Woo, J.H., Kim, Y., Fu, X., Xu, J.Y., et al., 2014. Emission trends and mitigation options for air pollutants in East Asia. Atmos. Chem. Phys. 14 (13), 6571–6603. https://doi.org/10.5194/acp-14-6571-2014.

Xing, J., Ding, D., Wang, S., Zhao, B., Jang, C., Wu, W., Zhang, F., Zhu, Y., Hao, J., 2018. Quantification of the enhanced effectiveness of NOx control from simultaneous reductions of VOC and NH3 for reducing air pollution in the Beijing–Tianjin–Hebei region, China. Atmos. Chem. Phys. 18 (11), 7799–7814. https://doi.org/10.5194/acp-18-7799-2018.

Xing, J., Wang, S.X., Jang, C., Zhu, Y., Hao, J.M., 2011. Nonlinear response of ozone to precursor emission changes in China: a modeling study using response surface methodology. Atmos. Chem. Phys. 11 (10), 5027–5044. https://doi.org/10.5194/acp-11-5027-2011.

Xing, J., Wang, S.X., Zhao, B., Wu, W.J., Ding, D., Jang, C., Zhu, Y., Chang, X., Wang, J. D., Zhang, F.F., et al., 2017. Quantifying nonlinear multiregional contributions to ozone and fine particles using an updated response surface modeling technique. Environ. Sci. Technol. 51 (20), 11788–11798. https://doi.org/10.1021/acs.est.7b01975.

Xing, J., Zheng, S.X., Ding, D., Kelly, J.T., Wang, S.X., Li, S.W., Qin, T., Ma, M.Y., Dong, Z.X., Jang, C., et al., 2020. Deep learning for prediction of the air quality response to emission changes. Environ. Sci. Technol. https://doi.org/10.1021/acs.est.0c02923.

Zhang, F.F., Xing, J., Zhou, Y., Wang, S.X., Zhao, B., Zheng, H.T., Zhao, X., Chang, H.Z., Jang, C., Zhu, Y., et al., 2020. Estimation of abatement potentials and costs of air pollution emissions in China. J. Environ. Manag. 260, 110069. https://doi.org/10.1016/j.jenvman.2020.110069.

Zhao, B., Wang, S.X., Xing, J., Fu, K., Fu, J.S., Jang, C., Zhu, Y., Dong, X.Y., Gao, Y., Wu, W.J., et al., 2015. Assessing the nonlinear response of fine particles to precursor emissions: development and application of an extended response surface modeling technique v1.0. Geosci. Model Dev. (GMD) 8 (1), 115–128. https://doi.org/10.5194/gmd-8-115-2015.

Zhao, B., Wu, W., Wang, S., Xing, J., Chang, X., Liou, K.N., Jiang, J.H., Gu, Y., Jang, C., Fu, J.S., et al., 2017. A modeling study of the nonlinear response of fine particles to air pollutant emissions in the Beijing–Tianjin–Hebei region. Atmos. Chem. Phys. 17 (19), 12031–12050. https://doi.org/10.5194/acp-17-12031-2017.

Zhao, B., Zheng, H.T., Wang, S.X., Smith, K.R., Lu, X., Aunan, K., Gu, Y., Wang, Y., Ding, D., Xing, J., 2018. Change in household fuels dominates the decrease in PM2. 5 exposure and premature mortality in China in 2005–2015. Proc. Natl. Acad. Sci. U.S.A. 115 (49), 12401–12406. https://doi.org/10.1073/pnas.1812955115.