atmosphere

MDPI

Article

Quantifying the Impact of Multiple Factors on Air Quality Model Simulation Biases Using Machine Learning

Chunying Fan [1], Ruilin Wang [2,*], Ge Song [1], Mengfan Teng [1], Maolin Zhang [1], Huangchuan Liu [1], Zhujun Li [1], Siwei Li [1,3,*] and Jia Xing [4]

[1] Hubei Key Laboratory of Quantitative Remote Sensing of Land and Atmosphere, School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China; fancy99@whu.edu.cn (C.F.); gesong@whu.edu.cn (G.S.); tengmengfan@whu.edu.cn (M.T.); maolinzhang@whu.edu.cn (M.Z.)
[2] Institute of Software, Chinese Academy of Sciences, Beijing 100864, China
[3] Perception and Effectiveness Assessment for Carbon-neutrality Efforts, Engineering Research Center of Ministry of Education, Institute for Carbon Neutrality, Wuhan University, Wuhan 430072, China
[4] Department of Civil and Environmental Engineering, The University of Tennessee, Knoxville, TN 37996, USA; jxing3@utk.edu
* Correspondence: wangruilin@iscas.ac.cn (R.W.); siwei.li@whu.edu.cn (S.L.)

Abstract: Accurate air pollutant prediction is essential for addressing environmental and public health concerns. Air quality models like WRF-CMAQ provide simulations, but often show significant errors compared to observed concentrations. To identify the sources of these model biases, we applied the XGBoost machine learning algorithm to assess the performance of WRF-CMAQ in predicting air pollutants across two regions in China. XGBoost models trained with observations achieved high accuracy ($R > 0.95$), indicating that the selected features effectively capture pollutant variations. When trained on WRF-CMAQ inputs, XGBoost still improved performance but revealed biases linked to both model inputs (10–60%) and mechanisms (1–30%). Analysis identified previous-hour pollutant levels as the largest bias contributor, followed by meteorological variables. The study highlights the need for improving both model inputs and mechanisms to enhance future air quality predictions and support pollution control strategies.

Keywords: air quality; simulation; bias; machine learning; prediction

Citation: Fan, C.; Wang, R.; Song, G.; Teng, M.; Zhang, M.; Liu, H.; Li, Z.; Li, S.; Xing, J. Quantifying the Impact of Multiple Factors on Air Quality Model Simulation Biases Using Machine Learning. *Atmosphere* **2024**, *15*, 1337. https://doi.org/10.3390/atmos15111337

Academic Editor: Alexandra Monteiro

Received: 24 September 2024
Revised: 28 October 2024
Accepted: 5 November 2024
Published: 7 November 2024

1. Introduction

Heavy air pollution has been a critical issue with far-reaching implications for environmental sustainability and public health in various regions worldwide [1–3]. Among these, the Beijing–Tianjin–Hebei (BTH) and Yangtze River Delta (YRD) regions in China are two examples of heavily polluted areas that have drawn public attention to this issue. Accurate prediction of pollutant concentrations plays a crucial role in addressing this challenge and designing effective mitigation strategies. Traditional methods for predicting pollutant concentrations include statistical and dispersion models [4]. The former establishes relationships between pollutant concentrations and influencing factors using regression analysis [5–7]. However, these models suffer from limitations such as limited spatial coverage due to sparse monitoring station distribution, data gaps, and difficulties in capturing non-linear and complex relationships. Air Quality models simulate pollutant dispersion and transport based on emission sources, meteorological conditions, and terrain characteristics [8–10]. However, these models rely on accurate meteorological data, have simplified representations of atmospheric processes, and struggle to capture localized variations and short-term fluctuations in pollutant concentrations. Such uncertainties from model inputs and physical mechanism result in considerable biases in simulating air pollutant concentrations with air quality models.

Recently, machine learning techniques have displayed advantages in handling complex and nonlinear relationships [11–13] and have been widely used in air quality prediction

with air quality modeling. For example, Xue [14] introduced a data-fusion algorithm to estimate ozone (O_3) combining observations and simulations. Xiao [15] applied four gap-filling strategies to a 1-km resolution random forest prediction model of $PM_{2.5}$ (fine particulate matter with a diameter of 2.5 μm or smaller) daily concentration in the BTH and YRD regions in 2013, combining ground observation, air quality model simulation and satellite aerosol optical depth. Machine learning's ability to identify complex patterns and relationships can significantly enhance our understanding of the sources of air quality modeling biases [16]. By linking the intricate interactions between pollutant sources, meteorological variables, and their impacts on pollutant concentrations, the reliability and precision of pollutant predictions in air quality models will be enhanced [17]. This approach not only captures the general trends but also accounts for the complex interactions that traditional methods might overlook, leading to more accurate and reliable air quality predictions.

As mentioned previously, the simulated pollutant concentrations with air quality models often result in substantial errors when directly applied due to variations in accuracy compared to observations [14,18]. Some studies have attempted to diagnose and improve the underlying mechanism in air quality model with machine learning methods. For instance, Keller [19] and Yin [20] analyzed the important factors on O_3 modeling biases. They found that certain meteorological conditions and precursor emissions were key influencing O_3 levels, which are critical for refining air quality models to better predict ozone concentrations. Liu and Xing [21] analyzed the relative contribution of the variables in $PM_{2.5}$ biases with a neural network, which revealed that incorporating machine learning techniques can effectively identify and reduce biases in $PM_{2.5}$ predictions. Ye [22] diagnosed the model bias in estimating O_3 and attempted to improve the physical and chemical representations in air quality model. Xu [18] calibrated simulation to match site observation by a random forest algorithm, demonstrating the potential of machine learning techniques in adjusting outputs to better align with observed pollution levels and thereby reducing prediction errors. Recognizing the crucial role of comprehending data discrepancies in pollutant prediction, it is equally important to quantify the contributions from individual factors, encompassing both inputs and the model mechanism.

To improve comprehension of these data discrepancies, here we present a comprehensive investigation into the modeling biases in predicting the concentration of four key pollutants in two regions with machine learning. We conduct contribution analysis to identify variables that have the most significant impact on the predictions, and sensitivity analysis to evaluate the effects of different perturbations added to input data on the results over time. This paper presents a novel approach to analyzing the effects of biases in air quality model simulation using machine learning. With insights gained through the analysis, this study aims to provide valuable guidance for improving the accuracy of concentration prediction and the reliability of air quality model simulations, ultimately contributing to the effective mitigation of heavy air pollution.

2. Materials and Methods

2.1. Study Region

This study focuses on two heavily polluted regions, namely the BTH and YRD regions in China, using hourly data from the year 2015 for analysis. As shown in Figure 1, the BTH region encompasses the capital city of Beijing, along with the surrounding provinces of Tianjin and Hebei, while the YRD region includes Shanghai and the surrounding provinces along the Yangtze River.

Both regions are densely populated and heavily industrialized, with a concentration of manufacturing, energy production, and transportation activities. These factors contribute to significant emissions of pollutants, including nitrogen dioxide (NO_2), sulfur dioxide (SO_2), O_3, and $PM_{2.5}$. Taking the pollution level of $PM_{2.5}$ in these regions as an example, their annual average concentrations reached 75 and 52 $\mu g/m^3$ respectively in 2015, which were far beyond the recommended annual average of 10 $\mu g/m^3$ in the World Health

Organization (WHO) (2017) Air Quality Guidelines. Therefore, it is necessary to make accurate predictions and effective mitigation strategies.

Figure 1. Study region.

2.2. Data

2.2.1. Simulation Data

The air quality model we used in this study is the Community Multiscale Air Quality (CMAQ). The simulated meteorological fields needed by the CMAQ model (version 5.3.2) are provided by the version 4.3 of the Weather Research and Forecasting (WRF) model. The input meteorological data in WRF are derived from FNL(Final) data of the National Centers for Environmental Prediction at a spatial resolution of 0.25° by 0.25° and a temporal resolution of 6 h from the Global Data Assimilation System (GDAS). The suite of parameterizations is shown in Table S1. To eliminate the effect of initial conditions, a 5-day spin-up simulation was performed.

In this study, the variables used in the study are 10 m U-component wind speed(U), 10 m V-component wind speed (V), temperature at 2 m (T), precipitation (PREP), relative humidity (RH), planetary boundary layer height (PBLH) and surface pressure (SP) as shown in Table 1. The CMAQv5.2 model was used to estimate the pollutant concentrations. The simulation region covers East Asia with a 27 km horizontal spatial resolution and 23 vertical layers. The 5-day simulation spin-up strategy was adopted as in other studies. The unit of pollutant concentrations is converted from part per billion volume (ppbV) to micrograms per cubic meter ($\mu g \cdot m^{-3}$) for the coordination between the simulation and observation.

Table 1. Summary of the input variables.

Type	Variable	Abbr.	Simulation (Unit)		Observation (Unit)	
Pollutant	Surface NO_2 concentration	NO_2	CMAQ	ppbV	In-site	$\mu g \cdot m^{-3}$
	Surface SO_2 concentration	SO_2		ppbV		$\mu g \cdot m^{-3}$
	Surface O_3 concentration	O_3		ppbV		$\mu g \cdot m^{-3}$
	Surface $PM_{2.5}$ concentration	$PM_{2.5}$		ppbV		$\mu g \cdot m^{-3}$
Meteorology	10 m U-component wind speed	U	WRF	m/s	ERA5	m/s
	10 m V-component wind speed	V		m/s		m/s
	2 m temperature	T		K		K
	total precipitation	TP		mm		mm
	Relative Humidity/2 m Dewpoint temperature	RH/DT		%		K
	Planetary boundary layer height	PBLH/PB		m		m
	Surface pressure	SP		Pa		Pa

2.2.2. Observation Data

The meteorological data used in this study are obtained from the European Center for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5) dataset, which provides a reanalysis rather than direct observations, serving as a "best guess" based on model output and observations. The ERA5 data are widely recognized for their high special and temporal resolution, which is often not available with traditional observational network. The variables selected in ERA5 are U, V, T, total precipitation (PREP), 2 m dewpoint temperature (DT), PBLH, and SP as shown in Table 1, which are the same as meteorological simulation of WRF except DT. For the coordination, the dewpoint temperature together with the temperature at 2 m in ERA5 is converted into relative humidity, which exists in the WRF simulation. The hourly in-situ measurements of four pollutants are collected from the China National Environmental Monitoring Centre (http://beijingair.sinaapp.com/, accessed on 29 August 2024).

2.3. Methods

As shown in Figure 2, we designed a schematic diagram of the main tasks of the research. The feasibility of the methodology assumes that the surrogate model can learn the prediction mechanism of pollutant predictions well based on observations. We utilized the eXtreme Gradient Boosting (XGBoost) as the surrogate model. It should be noted that, different from previous studies, we trained the XGBoost with observations rather than the simulations data, considering that the uncertainties from the model itself relate to the physical mechanism. This design enables us to quantify the uncertainties from both inputs and model mechanism.

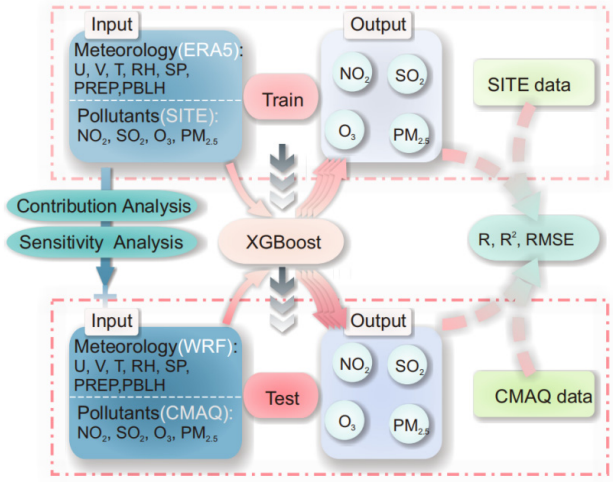


Figure 2. Method diagram of the study. The dotted box at the top of the figure represents the stage of training XGBoost models with ERA5 data and in situ observation data as input. The dotted box at the bottom of the figure represents the stage of testing XGBoost models using WRF-CMAQ simulation as input. *R*, R^2, and *RMSE* between in situ data and CMAQ simulation data are the indexes to estimate the performance of the models. By combining observation and simulation as input, contribution analysis and sensitivity analysis are conducted.

The process mainly consists of two parts: the training stage and the test analysis stage. In the training stage, the observed concentration of each pollutant (NO_2, SO_2, O_3, or $PM_{2.5}$) of the current hour is set as the label, while the referential feature variables as input consist of relevant observed meteorological data (U, V, T, TP, RH, PBLH, and SP) of the current hour and four pollutant data (NO_2, SO_2, O_3, and $PM_{2.5}$) of the previous hour. All of these are the variables that appear in the WRF-CMAQ model to study the relationship between the pollutant concentration and the features. In the test analysis phase, we carry out contribution analysis and sensitivity analysis respectively by interactively inputting simulations and observations, with details as follows.

2.3.1. Model Structure

In this study, the XGBoost is adopted as the surrogate model of prediction, which is developed from traditional gradient boosting algorithms [23]. The main idea of XGBoost is to incorporate several optimizations and enhancements, making it more powerful and efficient. Due to its strong predictive capacities and faster training speed with less overfitting on processing massive data, XGBoost has gained popularity in various domains, such as environmental research [24–26]. The loss function is set to *RMSE* and the learning rate is 0.1, with max depth 5 and boost round number 2000. To overcome underestimation, the objective is set to "reg: squared error" and the weight is squared with the target value.

We comprehensively evaluate the performances of three different outcomes (see Figure 2). The performance of the model at the training stage against the observation is marked as OML-OS. Next, we replace inputs with the corresponding simulations to evaluate the performance of the outcomes against pollutant simulations (marked as SML-SC). Moreover, the accuracy of the simulation against the observation is called SC-OS. In theory, the deficiency of SML-SC relative to OML-OS should be attributed mainly to systematic bias related to model itself, while the difference between OML-OS and SC-OS is considered to be caused by the input data bias, for they differ only in the input data. In this way, we can distinguish the effects of system bias and data bias. We train 10 models for each pollutant individually to verify the stability of the surrogate model.

To evaluate the performance of the models, some common statistical indexes are used, i.e., correlation coefficient (R), coefficient of determination (R^2), and Root mean square error ($RMSE$). The formulas are shown in (1)–(3).

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

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

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

where i is the number of samples, n is the total number, y_i is the target value of sample i, $\hat{y}_i$ is the source value of sample i, and $\overline{y}$ is the average value of the target set. The closer R is to -1, the more negatively correlated they are, and vice versa. The closer R^2 is to 1, the better the fitting effect.

2.3.2. Data Preprocessing

ERA5 data are resampled by linear interpolation to the same resolution as the simulated data. We match the site data to the grid cells consistent with the simulated data based on latitude and longitude and then remove all items containing missing values from variables, ensuring a high degree of data integrity and consistency.

In the test stage, all the simulations are taken as input when testing model performance with simulation, while simulations matching the observation grid are input when contribution analysis and sensitivity contribution are performed. Moreover, for long-term prediction analysis, it is necessary to find data not missing within 24 h.

In this study, 70% of the observed data is randomly selected as the dataset when training each model. Moreover, 70% of the dataset is selected as the training set, and the remaining 30% is selected as the test set.

2.3.3. Contribution Analysis Method

A contribution analysis is performed to represent the influence of individual predictors on the pollutant concentration. The Shapley Additive exPlanations (SHAP) method is used to interpret the trained model in this study, which has been widely used in previous studies [27–29]. SHAP provides a detailed interpretation of how each predictor affects the

model's output, both in terms of magnitude (how much it influences the prediction) and direction (whether it increases or decreases the predicted value). This approach, based on the Shapley value from game theory, fairly distributes the contribution of each feature to the final output. SHAP values are calculated for each case, allowing for a local interpretation of the model's behavior, where the magnitude represents the predictor's importance and the direction reflects the fluctuation in output. By comparing the model's predictions across all possible permutations of covariates, SHAP isolates the individual contribution of each variable. Meanwhile, the simulations of all variables are replaced with the observations one by one to quantify the effect of data bias on results.

2.3.4. Sensitivity Analysis Method

To delineate the effects of perturbation of different variables on prediction results. the simulations of each variable will be multiplied by different factors as disturbed input in each test group. Since different variables have different dimensions and data ranges, standardized factors are used to reduce these impacts. New input value is computed as below:

$$x^i_{new} = x^i_{ori} \times (1 + \frac{f \cdot X^i_{std}}{X^i_{mean}}) \quad (4)$$

where i is the variable i; x^i_{ori} is the original value of the variable, that is, the simulated value; f is the variation of the standard deviation, which is set as -1 and 1 in different groups; X^i_{std} and X^i_{mean} are the standard deviation value and the mean value of the variable. The relative delta values between the results of perturbations with different magnifications are then calculated.

3. Results

3.1. Modeling Biases of WRF-CMAQ

The performance of the WRF-CMAQ model in simulating air quality was evaluated in this study. From Figure 3, we can see that the WRF-CMAQ model demonstrated acceptable performance in simulating air quality, with R ranging from 0.35 to 0.59 across different pollutants. These results indicate that the model is generally capable of capturing the trends in air quality, with some variations in performance for different pollutant species. However, the scatter plots showed that the simulation substantially underestimates the pollutant concentrations, especially SO_2 concentration in the YRD region. The underestimation of pollutant concentrations may be due to the limitations in the model's representation of the complex atmospheric processes and the uncertainties in the input data. Therefore, further improvements in the model parameterization and input data are essential to enhance the accuracy of WRF-CMAQ in predicting air quality.

Figure 4 illustrates the diurnal and monthly variations in the bias of different pollutants as simulated by the WRF-CMAQ model in the two regions. The line chart in the figure showed the observed and simulated concentrations of pollutants, while the heat maps represent the hourly biases of simulated values from observed values. Specifically, we found that simulations tended to overestimate the value of NO_2 at night and underestimate its value during midday, especially in YRD region. The value of SO_2 was often underestimated. Nevertheless, there was substantial overestimation over 50 $\mu g/m^3$ in BTH region in December. For O_3, we observed that the bias was more centralized in overestimation during midday when photochemical reactions are more active throughout the year, except for the underestimation at night from April to October. In terms of $PM_{2.5}$, predominantly negative biases were observed most of the time, particularly in BTH region, while the positive biases were observed during summer nights.

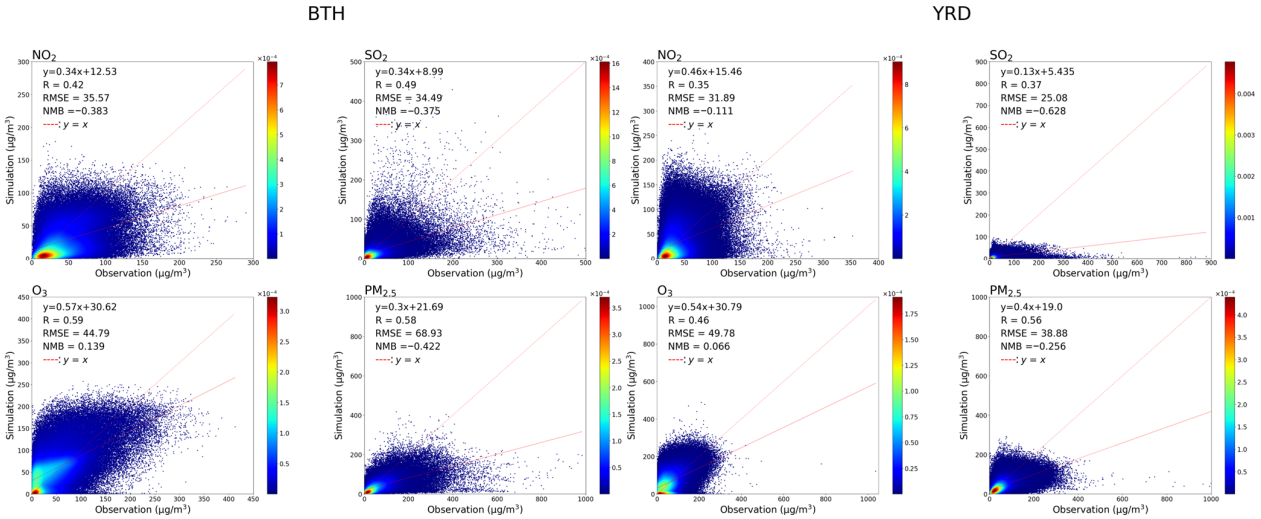


Figure 3. Scatter plots of WRF-CMAQ simulations versus observations.

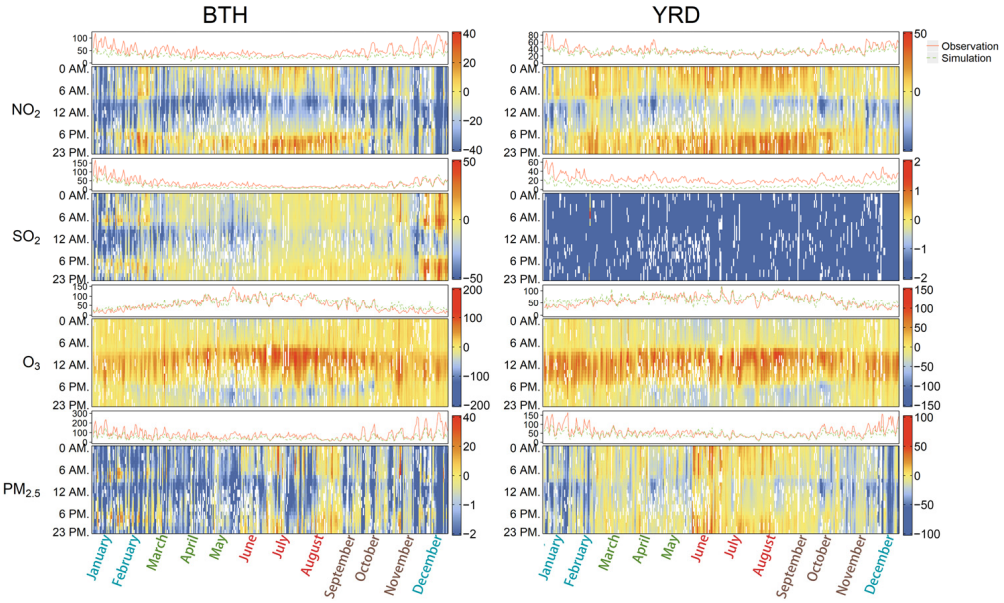


Figure 4. Simulation bias of four pollutants in two regions. The *x*-axis represents the month (January–December), and the *y*-axis indicates the concentration levels in μg/m^3 (line chart) and the time of day (0 AM, 6 AM, 12 AM, 6 PM, 23 PM in the heat map). In the line charts, red lines denote simulated values, while green lines represent observed values. In the heat maps, colors denote biases in μg/m^3.

3.2. Model Performance

The XGBoost models are trained to predict the hourly concentrations of air pollutants with observations. This study evaluated the performance of the XGBoost models, illustrated in Figure 5, under different conditions and analyzed the results for different pollutants and regions. In OML-OS, the values of R were about 0.95 and R^2 were over 0.85 except SO_2, which meant that the XGBoost models performed well in predicting the pollutant concentration using observations. On the other hand, the models had satisfying performances in SML-SC, with R between 0.92 and 0.97 and R^2 between 0.83 and 0.94.

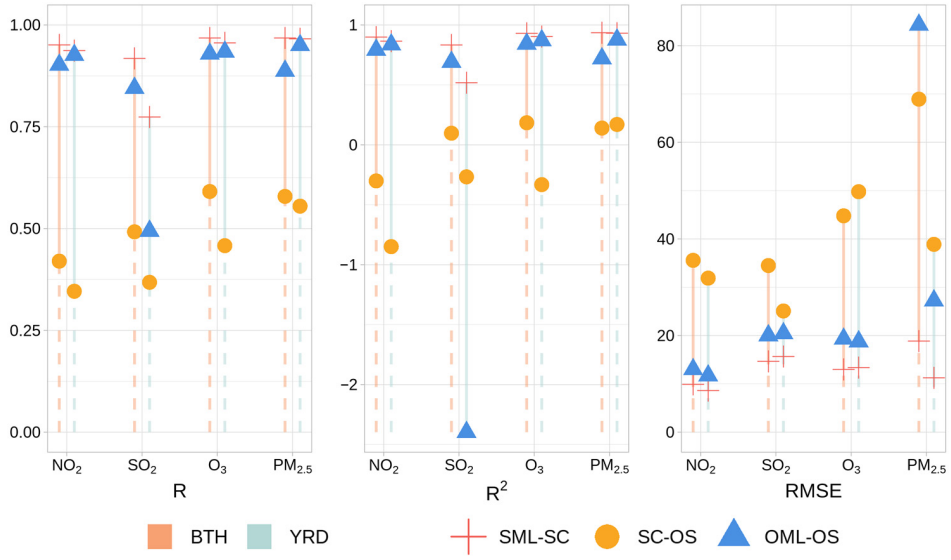
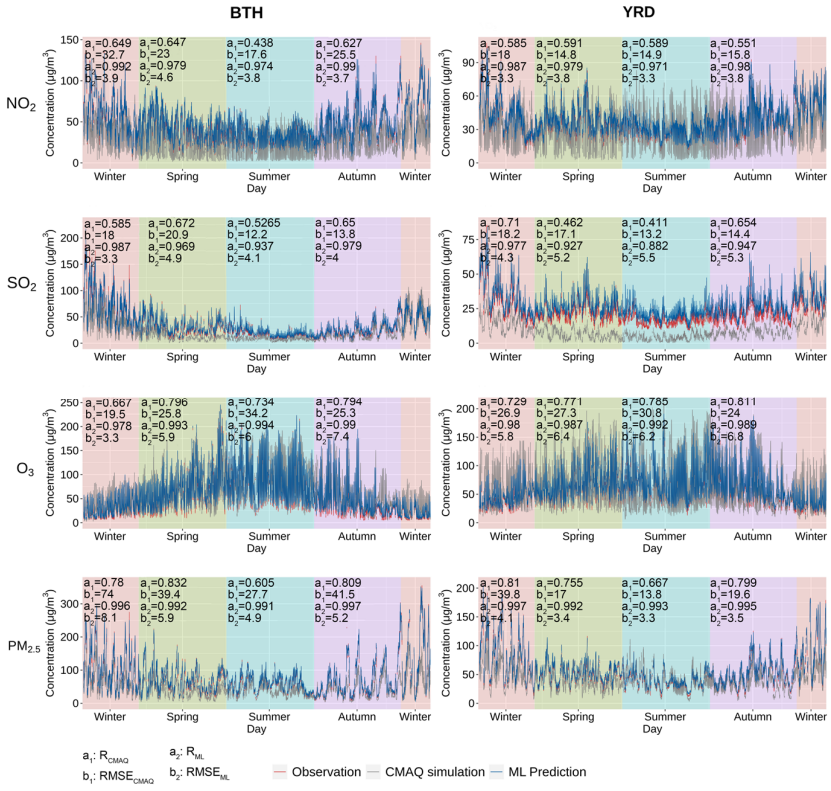


Figure 5. *R*, R^2, and *RMSE* of SC-OS, SML-SC, and OML-OS on pollutants. The lines and markers on the left of each pollutant represent the values of BTH, while the ones on the right represent the values of YRD.

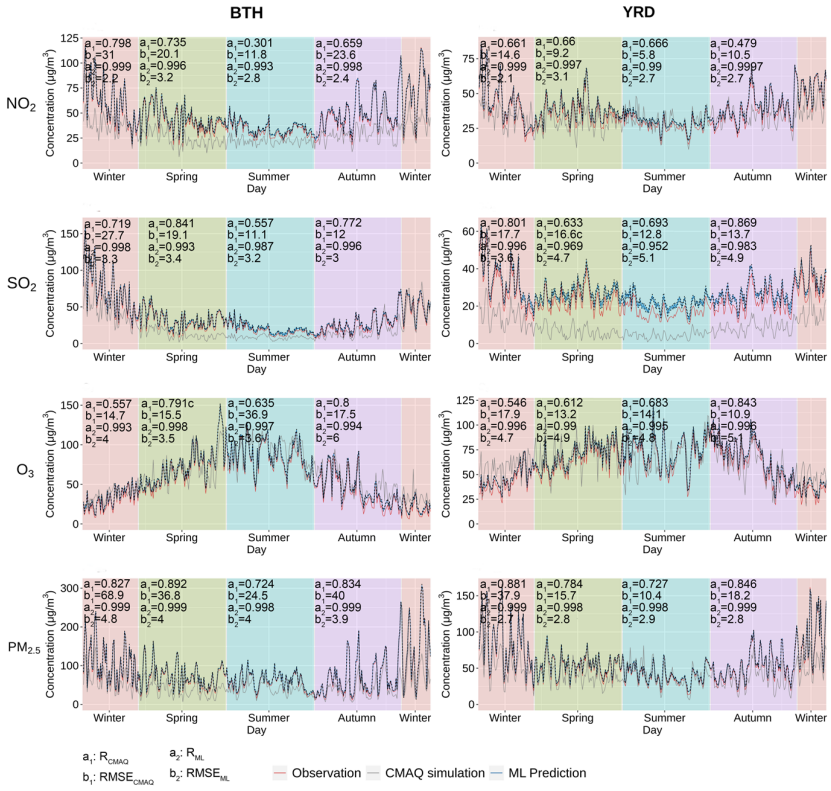
The SC-OS analysis showed a large deviation between the simulations of WRF-CMAQ and observations, as can be seen in Figure 3. After replacing the input from simulation to observation, the accuracy changed from SC-OS to OML-OS, reflecting the contribution of input bias. In data bias, the difference values of *R* were overall between 0.4–0.5, with NO_2 showing the most improvement, especially in YRD with a value of 0.59. Moreover, the difference values of R^2 were all over 0.7. The amounts of improvement of *R* and R^2 on NO_2, O_3, and $PM_{2.5}$ in YRD were greater than those in BTH. Nevertheless, the *RMSE* experienced a greater reduction in BTH, indicating more obvious improvement.

The performance of SML-SC was inferior to that of OML-OS, indicating the presence of systematic bias in the WRF-CMAQ model. Due to the mechanism of model-based prediction, the model cannot obtain the ideal accuracy (namely that of OML-OS) consistent with the simulated target pollutant by inputting the simulated variables. Among the systematic biases, SO_2 exhibited the largest bias, suggesting that the WRF-CMAQ model has a harder time accurately predicting SO_2 levels compared to other pollutants. The complex chemistry and high reactivity of SO_2 in the atmosphere might provide possible explanations for the higher bias and make it more challenging for the model to accurately simulate its concentration, which was consistent with previous research [30]. In contrast, O_3, a type of secondary pollutant, showed less systematic bias due to its well-developed chemical mechanism. The increases of *RMSE* from OML-OS to SML-SC in BTH except SO_2 were greater than those in YRD.

We also analyzed the seasonal prediction accuracy against observation at hourly and daily scales to explore its temporal variability. As can be seen from Figure 6, the XGBoost prediction result with observation input had low uncertainty and high consistency. Notably, the accuracy of the two kinds of prediction data showed significant differences across different seasons. CMAQ simulation had the lowest accuracy in summer, with *R* ranging from 0.2 to 0.8. In contrast, the *R* values of XGBoost prediction were almost all above 0.95, suggesting that the machine learning approach outperformed the CMAQ simulation in terms of accuracy and reliability in predicting air pollutant levels. Furthermore, both CMAQ simulation and machine learning prediction showed higher *RMSE* values in the BTH area compared to the YRD area across all reasons.



(**a**)



(**b**)

Figure 6. Seasonal variability of predictions of WRF-CMAQ and XGBoost at hourly scale (**a**) and daily scale (**b**).

Overall, our results demonstrated the promising performance of the XGBoost models in predicting pollutant levels with observations and highlighted the importance of evaluating the performance of WRF-CMAQ simulation in predicting air pollutant concentrations. Table 2 shows the contribution of systematic bias and input data bias to total bias in terms of R. The distribution of model mechanism bias is about 1–30% (specifically, 1–28% in the two regions). In addition, the uncertainty contribution of inputs ranges from about 10% to 60% (specifically, 12–58% in the two regions).

Table 2. The percentages of systematic bias and input data bias in terms of R.

Region	Pollutant	Systematic Bias (%)	Input Data Bias (%)
BTH	NO_2	5	48
	SO_2	8	35
	O_3	4	34
	$PM_{2.5}$	8	31
YRD	NO_2	1	58
	SO_2	28	12
	O_3	3	47
	$PM_{2.5}$	2	39

However, as illustrated in Figure 7, the SML-SC scatter plots showed obvious outliers, especially for the SO_2 model and $PM_{2.5}$ model in YRD. The XGBoost model, when applied to simulated data, faced challenges in accurately forecasting these extreme deviations. A potential reason for this is that the model is primarily trained on observed pollutant data that both consist of low and high concentration values, while simulated pollutant data of CMAQ are underestimated compared with the observations. Therefore, the predictions of XGBoost models may be higher than the simulated data, which are the outliers.

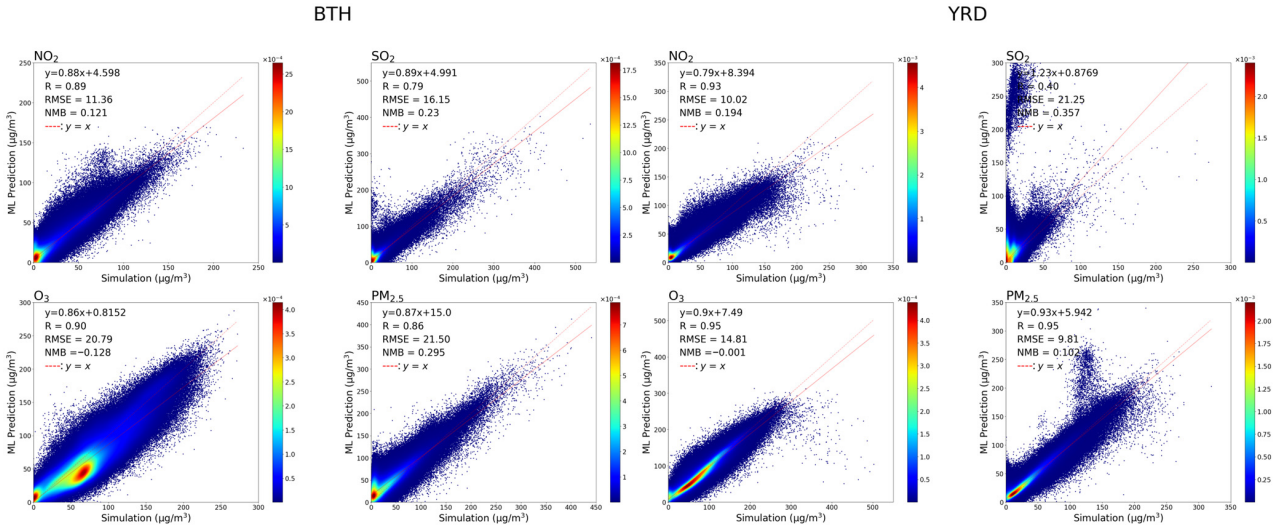


Figure 7. Scatter plots of WRF-CMAQ simulations versus XGBoost prediction with simulation input.

3.3. Contribution Analysis

To gain further insight into the input biases contributed by each variable, we presented SHAP values in terms of average contribution as feature importance, as shown in Figure 8. Feature importance rankings among models were consistent for each pollutant prediction experiment. Our analysis revealed that the primary features were the pollutants themselves from the previous hour under all conditions. However, there were slight differences in the importance ranking of other features of each pollutant prediction between the two regions. For instance, $PM_{2.5}$, NO_2, and T ranked as the top three features in both regions, while SP ranked fourth in the BTH region and U ranked fourth in the YRD region for

$PM_{2.5}$ prediction models. Interestingly, we observed that the model uncertainties in the importance value of the primary feature were the largest. We also presented the local impacts of values of covariates in Figure S1. The results further support the information in Figure 8. As shown in Figure S1, O_3 was found to be negatively correlated with the prediction of NO_2, while NO_2 was positively correlated with the prediction of O_3.

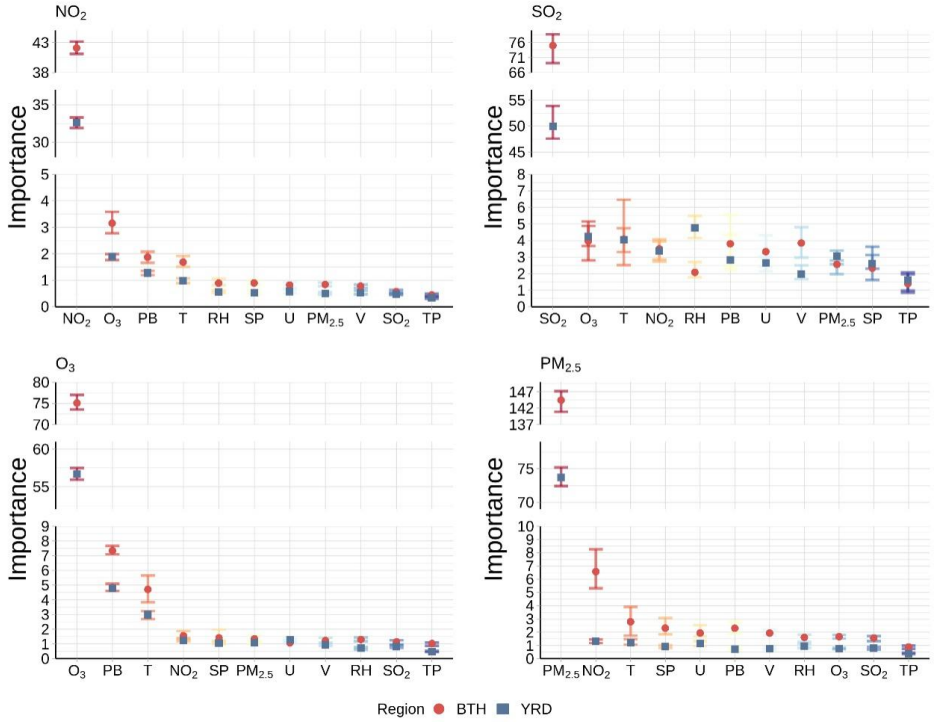


Figure 8. Importance of features on pollutant prediction in SHAP. The lines og different colors represent uncertainty of different variables.

On the other hand, we found that TP received the lowest rating in importance across all prediction models. This low rating may be attributed to several data quality issues, such as measurement errors. This finding underlines the importance of ensuring data quality when developing accurate models for predicting air pollutant concentrations.

Figure 9 presents the cumulative contribution of U, V, T, SP, RH, TP, PB, NO_2, SO_2, O_3, and $PM_{2.5}$. Our findings revealed that for all pollutants, the previous-hour concentration for each pollutant (noted as the primary feature) brought the most substantial improvement (by 90.2%, 70.6%, 91.2%, and 82.8% of improvement ratios on *R* in BTH respectively, and 93.5%, 67.6%, 92.4%, and 95.7% of those in YRD). Though the primary feature plays the most important role in predicting the following hour concentration, the cumulative contribution of other variables cannot be ignored, accounting for up to 30% of total impacts. The uncertainty of SO_2 was the greatest, which might be attributed to the model limitations and data quality, such as that of SO_2. In terms of regional differences, we observed that the models for SO_2 performed better in BTH from the perspective of R. Moreover, the cumulative contribution of other variables based on the primary feature, i.e., $PM_{2.5}$, were more obvious in BTH for concentration prediction of $PM_{2.5}$.

We further compared the cumulative contribution difference between the primary feature and all features together, by replacing the simulations of the primary feature with the observations (marked as PO) and replacing the simulations of all variables (marked as AO) with the observations. This helped us understand how much better the predictions were when we used observed data instead of simulated data across various pollution levels. The results of this experiment are shown in Figure 10. To better highlight the details of the proportion distribution with concentration, the data of the first 75% of concentration were selected for analysis. As can be seen from the percentage chart, the cumulative improvement

of AO was higher than that of PO and the uncertainty was reduced, indicating that using all variables in the model (AO) led to more accurate predictions and less uncertainty compared to using only the primary feature (PO). The proportion of improvement fluctuated in the low concentration range, and then increased as the concentration increased, possibly because the WRF-CMAQ model tended to underestimate at high concentration as shown in Figure 3. Moreover, the variation trends of primary pollutants with concentration were almost the same in PO and AO. In general, the uncertainty of the improvement percentage in YRD was greater.

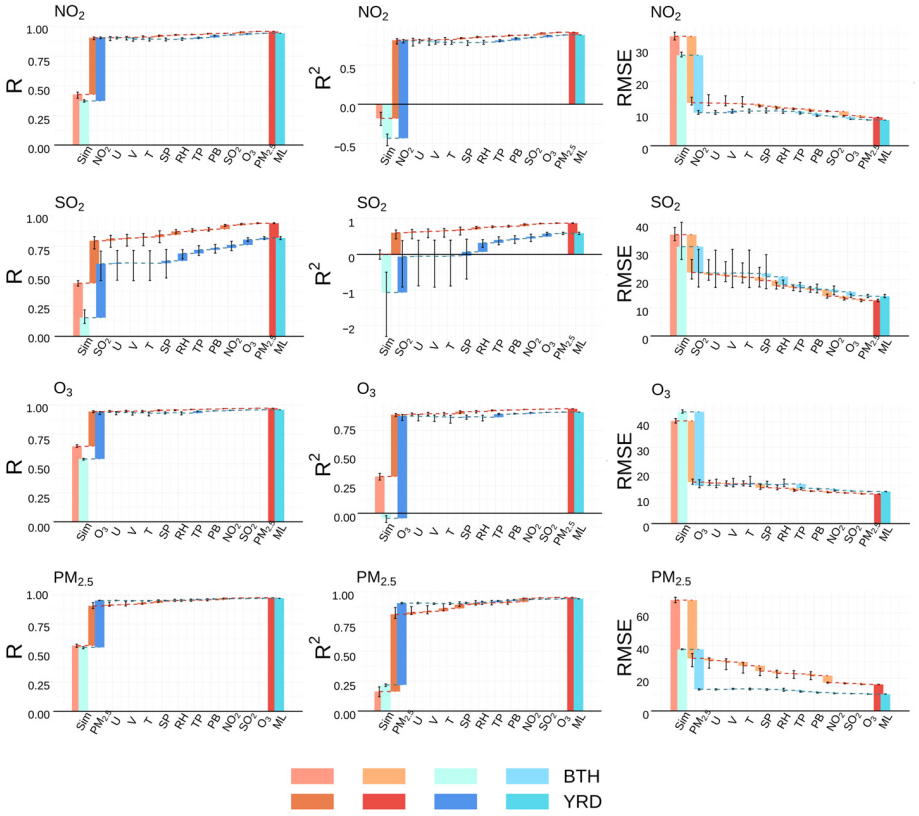


Figure 9. Cumulative contribution by changing the simulation to the observation of each variable one by one. ML means the performance of the model with all observed input. The lines represent uncertainty of different variables.

Specifically, the percentages of improvement for NO_2 exceeded 60% both in PO and AO as a whole, and the percentage at low concentration in AO reached up to 90%, revealing an obvious improvement. For SO_2, the percentage range was 40% to 80% in BTH, while it was 50% to 70% in YRD. Moreover, the percentage of PO exhibited a large degree of uncertainty within YRD. For O_3, the percentages decreased at low concentrations followed by an increase at higher concentrations, implying the important of other features (including precursor and meteorological variables) in predicting high O_3 concentration. However, there was a slight decrease in the percentage of PO within the range of 60–80 $\mu g/m^3$. $PM_{2.5}$ showed distinct patterns in the two regions. In BTH, both PO and AO experienced a decline in the percentages at low concentrations followed by a subsequent increase, whereas in YRD, it first increased and then decreased at low concentrations. Such results suggest that accurate initial concentrations of pollutants (itself also precursors) and meteorological factors are both important to achieve the best prediction in the following hours.

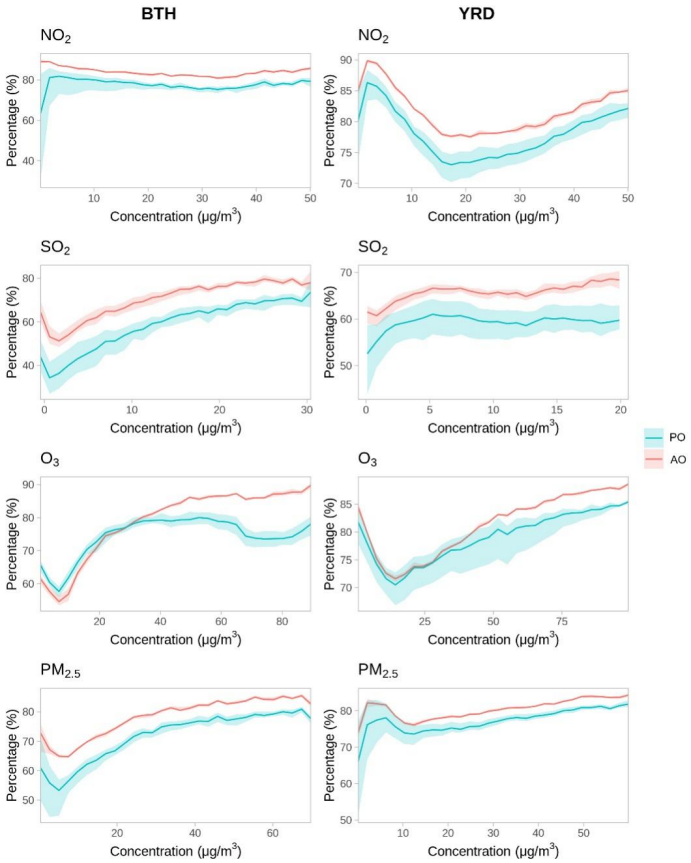


Figure 10. The percentage of hours with better results against simulations. The shaded regions represent the uncertainty ranges of 10 models for each pollutant.

3.4. Sensitivity Analysis

To obtain the sensitivity to data at different times, we firstly acquired the predictions of pollutant levels with small perturbations to a variable (by increasing it by $f = 1$ and decreasing it by $f = -1$) according to formula (4) and assessed their impact on different periods of the day based on the delta values between their prediction errors. Figure 11 revealed that altering a single variable within the simulation caused the prediction error to fluctuate in a pattern that was dependent on the time of day. Moreover, each of the four pollutants exhibited a distinct pattern of variation in response to these perturbations.

As shown in Figure 11, for NO_2 sensitivity, the error deltas induced by U variable disturbance were the highest. Moreover, the error deltas from each variable were generally smaller during midday and kept in the same direction all day. Regarding SO_2, the variation characteristics of the errors were consistent with each other in the two regions with morning and evening peaks, while larger deltas were observed in BTH. The magnitudes are related to the baseline concentration of each pollutant. In terms of O_3, the disturbance contributions were comparable between the two regions. Unlike other pollutants, the errors influenced by variable disturbance grew larger during midday, which aligned with the temperature variation characteristics. The variation features of $PM_{2.5}$ were similar to those of SO_2 with significantly larger error deltas observed in BTH. Notably, the prominent fluctuations driven by NO_2 and SO_2 during midday suggested the occurrence of secondary aerosol production from photochemical reactions around noon. On the whole, among non-primary variables, perturbations in wind, temperature, and surface pressure resulted in substantial fluctuations in delta of prediction errors, indicating that the prediction is highly sensitive to such changes. The results highlighted the intricate relationship between pollutant sensitivities and various variable perturbations.

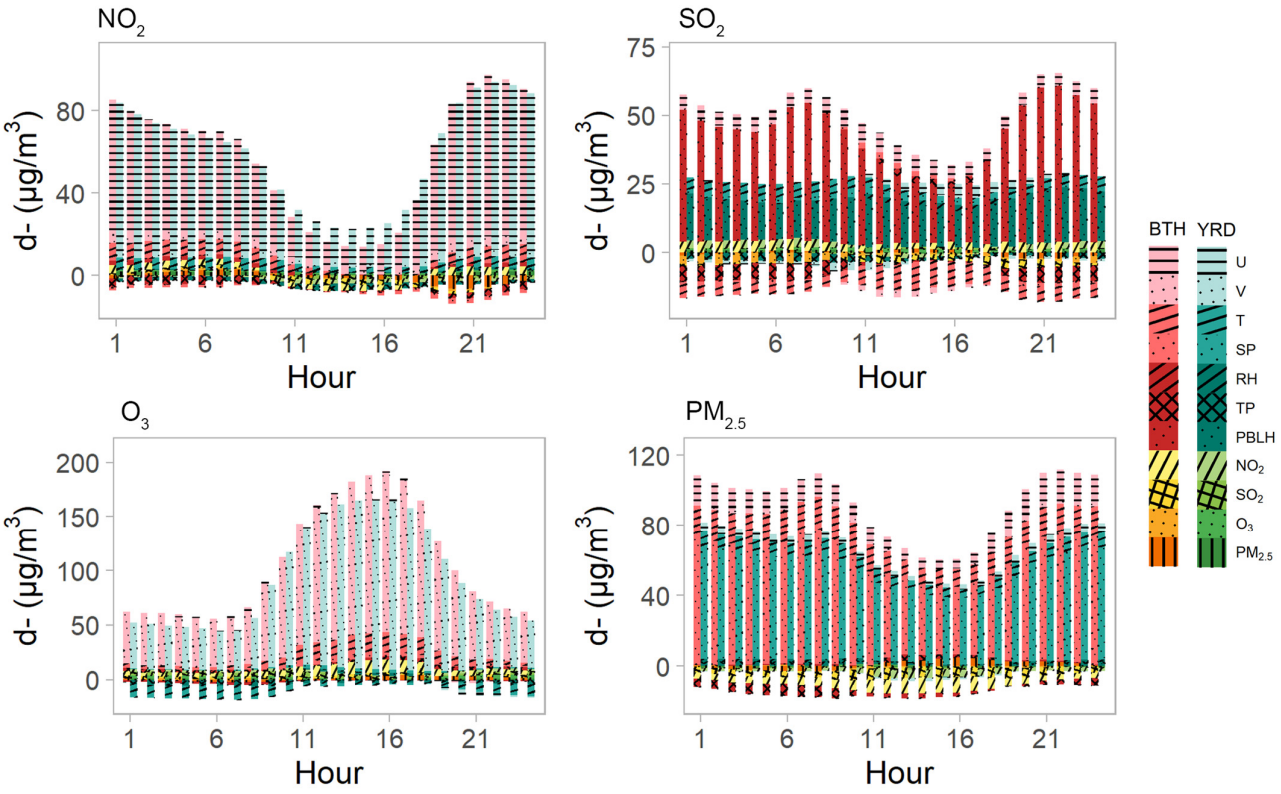


Figure 11. The delta value between the prediction errors with small perturbation to a variable at $f = 1$ and $f = -1$ at different local times of the day.

4. Discussion

This study aimed to comprehensively investigate the effects of data discrepancies between WRF-CMAQ simulations and site observations on predicting the concentration of four key pollutants in the BTH and YRD regions of China. The XGBoost machine learning algorithm, known for its efficiency and effectiveness on such issues, was utilized to conduct the analyses. Our analyses included examining the model's performance across seasons and locations, conducting feature contribution analysis to identify significant variables, and performing disturbance sensitivity to evaluate the effects of perturbations on the results over time.

The WRF-CMAQ model captured general trends of several air pollutant levels but underestimated pollutant concentrations, as the model's grid resolution and the parameterizations used for physical processes might not fully capture the small-scale processes that influence pollutant dispersion and chemistry [31]. Uncertainties in emission inventories, meteorological inputs, and chemical reactions within the model could also contribute to the underestimation [32,33]. Moreover, different pollutants exhibited varying biases throughout the data and year, which could be influenced by their specific emission sources, chemical transformations, and atmospheric conditions. For example, the underestimation of NO_2 at night could be due to the model's inability to capture nighttime emissions accurately, such as from traffic or industrial sources [34]. The biases in SO_2 outside the noon hours could be related to the model's representation of chemical reactions and the formation of secondary pollutants [35]. The model misses key reaction pathways that are important for SO_2 and its spatial resolution is insufficient to resolve the microenvironments where SO_2 is converted to secondary pollutants. The severe biases in O_3 at high concentrations in the YRD region could be attributed to the complex interplay between precursor emissions, meteorological conditions, and local atmospheric chemistry. The biases in $PM_{2.5}$ during summer nights could be influenced by factors like regional transport, meteorological conditions, and the representation of aerosol processes in the model [36,37].

As for machine learning, the XGBoost models performed well in predicting pollutant concentrations using observations, with R about 0.95 and R^2 over 0.85. However, when using simulation data as input, the models showed systematic biases between 1–28%, especially for SO_2 whose systematic bias is 28%. Nevertheless, the machine learning approach still outperformed the CMAQ simulation in terms of accuracy and reliability, suggesting that they were able to capture additional information or patterns that improved the accuracy and reliability of the predictions.

Feature importance analysis revealed that the pollutants themselves from the previous hour were primary features in all models. This is not surprising, given that pollutant levels often exhibit a high degree of auto-correlation over time. Essentially, the concentration of pollutants from the last hour is a strong indicator of the current hour's levels. This can be attributed to the enduring presence of pollution sources and the way meteorological conditions, such as wind and temperature, affect the dispersion and accumulation of pollutants in atmosphere. Additionally, the analysis also revealed that data quality issues, such as outliers, or inaccuracies in data, affected the importance of certain variables, highlighting the need for ensuring data quality for accurate predictions. Cumulative contribution analysis demonstrated that the primary feature had the most significant impact, but other variables also contributed to prediction accuracy. This suggested that a combination of quality-controlled data multiple variables, including meteorological and pollutant-related factors, was necessary for capturing the complex dynamics of pollutant concentrations accurately. Sensitivity analysis showed that changes in a single variable within the simulation resulted in prediction errors that varied according to the time of day. In addition, the different patterns observed for each pollutant and variable indicated that the sensitivity of predictions to disturbances varies depending on the specific pollutant and meteorological factor. Moreover, meteorological data, such as wind, temperature, and surface pressure disturbances would cause significant errors for prediction. It could be due to the sensitivity of pollutant dispersion and transport to meteorological conditions, which can greatly influence the spread and concentration of pollutants in atmosphere. Inaccurate or uncertain meteorological data can lead to errors in predicting pollutant behavior.

Our study also considers the influence of variable disturbances on prediction errors. Firstly, input variables in pollutant prediction models are subject to various disturbances or uncertainties. By examining how these disturbances affect the accuracy of predictions, we can gain a better understanding of the robustness and reliability of the model in different scenarios. Secondly, analyzing these influences helps identify which input variables are more sensitive to disturbances and require greater attention in terms of data quality assurance. All in all, this work provides insights into the relationships between variable data discrepancies, including those of pollutants and meteorological factors, and prediction accuracy. It demonstrates that the influence of these variables is not static but variables across different pollutants, These findings contribute to the development of more accurate and effective models adapting to the data discrepancy, which also underscores the complexity of pollutant prediction and encourages future exploration in this area.

However, there are some limitations in this study. The study focused on specific regions, namely the BTH and the YRD regions in China. The findings may not be directly applicable to other regions with different geographical, meteorological, and emission characteristics. Therefore, caution should be exercised when generalizing the results to other areas. Moreover, the research process likely involved certain assumptions and simplifications to make the analysis feasible. These assumptions may introduce uncertainties and limitations in the results. It is important to acknowledge these assumptions and consider their potential impact on the findings. However, all in all, by identifying the data discrepancy effects between simulation and observation in current approaches, we hope that this research serves as a foundation for future investigation and provides valuable insights for future studies aiming to enhance the accuracy and effectiveness of air quality predictions.

5. Conclusions

In conclusion, the study's results provide insights into the contribution of each variable to input biases and emphasize the importance of addressing issues of input bias both for meteorological data and pollutant data and accurately representing meteorological factors in air quality models. The slight differences in feature importance rankings for each pollutant prediction between the two regions suggest the need for region-specific patterns of air pollutants to achieve more accurate predictions. This can help policymakers formulate targeted strategies to mitigate air pollution levels in different regions.

The research serves as a foundation for future investigation and studies aiming to enhance the accuracy and effectiveness of air quality predictions. By identifying the effects of data discrepancies between simulation and observation, this study contributes to the understanding of the complexity of pollutant prediction and encourages further exploration in this area.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/atmos15111337/s1, Figure S1: Local impact by covariate values; Table S1: The parameter setting of WRF.

Author Contributions: Conceptualization, R.W. and J.X.; methodology, R.W.; software, C.F.; validation, G.S., M.T. and M.Z.; formal analysis, C.F. and G.S.; investigation, J.X.; resources, J.X.; data curation, H.L. and Z.L.; writing—original draft preparation, C.F.; writing—review and editing, R.W. and C.F.; visualization, C.F.; supervision, S.L.; project administration, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Open Research Program of the International Research Center of Big Data for Sustainable Development Goals, Grant No. CBAS2022ORP01.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The simulation data used in this study are not publicly available at this time, as they are part of ongoing research that has yet to be fully published. Once these related studies are completed and published, the simulation data will be made accessible upon reasonable request. However, the observational data utilized in this research are publicly available and can be accessed from the ERA5 reanalysis dataset (https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5, accessed on 29 August 2024) and from the respective official websites of the monitoring stations.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Anenberg, S.C.; Henze, D.K.; Tinney, V.; Kinney, P.L.; Raich, W.; Fann, N.; Malley, C.S.; Roman, H.; Lamsal, L.; Duncan, B.; et al. Estimates of the Global Burden of Ambient PM2.5, Ozone, and NO_2 on Asthma Incidence and Emergency Room Visits. *Environ. Health Perspect.* **2018**, *126*, 107004. [CrossRef] [PubMed]
2. Lelieveld, J.; Evans, J.S.; Fnais, M.; Giannadaki, D.; Pozzer, A. The Contribution of Outdoor Air Pollution Sources to Premature Mortality on a Global Scale. *Nature* **2015**, *525*, 367–371. [CrossRef] [PubMed]
3. Maji, K.J.; Ye, W.-F.; Arora, M.; Nagendra, S.M.S. Ozone Pollution in Chinese Cities: Assessment of Seasonal Variation, Health Effects and Economic Burden. *Environ. Pollut.* **2019**, *247*, 792–801. [CrossRef] [PubMed]
4. Pak, U.; Ma, J.; Ryu, U.; Ryom, K.; Juhyok, U.; Pak, K.; Pak, C. Deep Learning-Based PM2.5 Prediction Considering the Spatiotemporal Correlations: A Case Study of Beijing, China. *Sci. Total Environ.* **2020**, *699*, 133561. [CrossRef] [PubMed]
5. Di Carlo, P.; Pitari, G.; Mancini, E.; Gentile, S.; Pichelli, E.; Visconti, G. Evolution of Surface Ozone in Central Italy Based on Observations and Statistical Model. *J. Geophys. Res.* **2007**, *112*, 2006JD007900. [CrossRef]
6. Hu, X.; Waller, L.A.; Al-Hamdan, M.Z.; Crosson, W.L.; Estes, M.G.; Estes, S.M.; Quattrochi, D.A.; Sarnat, J.A.; Liu, Y. Estimating Ground-Level PM2.5 Concentrations in the Southeastern U.S. Using Geographically Weighted Regression. *Environ. Res.* **2013**, *121*, 1–10. [CrossRef]
7. Jeong, J.I.; Park, R.J.; Yeh, S.-W.; Roh, J.-W. Statistical Predictability of Wintertime PM2.5 Concentrations over East Asia Using Simple Linear Regression. *Sci. Total Environ.* **2021**, *776*, 146059. [CrossRef]

8. David, L.M.; Ravishankara, A.R.; Brewer, J.F.; Sauvage, B.; Thouret, V.; Venkataramani, S.; Sinha, V. Tropospheric Ozone over the Indian Subcontinent from 2000 to 2015: Data Set and Simulation Using GEOS-Chem Chemical Transport Model. *Atmos. Environ.* **2019**, *219*, 117039. [CrossRef]
9. Cheng, F.-Y.; Feng, C.-Y.; Yang, Z.-M.; Hsu, C.-H.; Chan, K.-W.; Lee, C.-Y.; Chang, S.-C. Evaluation of Real-Time PM2.5 Forecasts with the WRF-CMAQ Modeling System and Weather-Pattern-Dependent Bias-Adjusted PM2.5 Forecasts in Taiwan. *Atmos. Environ.* **2021**, *244*, 117909. [CrossRef]
10. Christian, K.E.; Brune, W.H.; Mao, J. Global Sensitivity Analysis of the GEOS-Chem Chemical Transport Model: Ozone and Hydrogen Oxides during ARCTAS (2008). *Atmos. Chem. Phys.* **2017**, *17*, 3769–3784. [CrossRef]
11. Gui, K.; Che, H.; Zeng, Z.; Wang, Y.; Zhai, S.; Wang, Z.; Luo, M.; Zhang, L.; Liao, T.; Zhao, H.; et al. Construction of a Virtual PM2.5 Observation Network in China Based on High-Density Surface Meteorological Observations Using the Extreme Gradient Boosting Model. *Environ. Int.* **2020**, *141*, 105801. [CrossRef] [PubMed]
12. Wang, A.; Xu, J.; Tu, R.; Saleh, M.; Hatzopoulou, M. Potential of Machine Learning for Prediction of Traffic Related Air Pollution. *Transp. Res. Part D Transp. Environ.* **2020**, *88*, 102599. [CrossRef]
13. Liu, X.; Lu, D.; Zhang, A.; Liu, Q.; Jiang, G. Data-Driven Machine Learning in Environmental Pollution: Gains and Problems. *Environ. Sci. Technol.* **2022**, *56*, 2124–2133. [CrossRef] [PubMed]
14. Xue, T.; Zheng, Y.; Geng, G.; Xiao, Q.; Meng, X.; Wang, M.; Li, X.; Wu, N.; Zhang, Q.; Zhu, T. Estimating Spatiotemporal Variation in Ambient Ozone Exposure during 2013–2017 Using a Data-Fusion Model. *Environ. Sci. Technol.* **2020**, *54*, 14877–14888. [CrossRef] [PubMed]
15. Xiao, Q.; Geng, G.; Cheng, J.; Liang, F.; Li, R.; Meng, X.; Xue, T.; Huang, X.; Kan, H.; Zhang, Q.; et al. Evaluation of Gap-Filling Approaches in Satellite-Based Daily PM2.5 Prediction Models. *Atmos. Environ.* **2021**, *244*, 117921. [CrossRef]
16. Zaytar, M.A.; El Amrani, C. Machine Learning Methods for Air Quality Monitoring. In Proceedings of the 3rd International Conference on Networking, Information Systems & Security, Marrakech, Morocco, 31 March–2 April 2020; ACM: New York, NY, USA; pp. 1–5.
17. Sarker, I.H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Comput. Sci.* **2021**, *2*, 420. [CrossRef]
18. Xu, R.; Ye, T.; Yue, X.; Yang, Z.; Yu, W.; Zhang, Y.; Bell, M.L.; Morawska, L.; Yu, P.; Zhang, Y.; et al. Global Population Exposure to Landscape Fire Air Pollution from 2000 to 2019. *Nature* **2023**, *621*, 521–529. [CrossRef]
19. Keller, C.A.; Evans, M.J.; Knowland, K.E.; Hasenkopf, C.A.; Modekurty, S.; Lucchesi, R.A.; Oda, T.; Franca, B.B.; Mandarino, F.C.; Díaz Suárez, M.V.; et al. Global Impact of COVID-19 Restrictions on the Surface Concentrations of Nitrogen Dioxide and Ozone. *Atmos. Chem. Phys.* **2021**, *21*, 3555–3592. [CrossRef]
20. Yin, H.; Lu, X.; Sun, Y.; Li, K.; Gao, M.; Zheng, B.; Liu, C. Unprecedented Decline in Summertime Surface Ozone over Eastern China in 2020 Comparably Attributable to Anthropogenic Emission Reductions and Meteorology. *Environ. Res. Lett.* **2021**, *16*, 124069. [CrossRef]
21. Liu, J.; Xing, J. Identifying Contributors to $PM_{2.5}$ Simulation Biases of Chemical Transport Model Using Fully Connected Neural Networks. *J. Adv. Model. Earth Syst.* **2023**, *15*, e2021MS002898. [CrossRef]
22. Ye, X.; Wang, X.; Zhang, L. Diagnosing the Model Bias in Simulating Daily Surface Ozone Variability Using a Machine Learning Method: The Effects of Dry Deposition and Cloud Optical Depth. *Environ. Sci. Technol.* **2022**, *56*, 16665–16675. [CrossRef] [PubMed]
23. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA; pp. 785–794.
24. Hu, L.; Wang, C.; Ye, Z.; Wang, S. Estimating Gaseous Pollutants from Bus Emissions: A Hybrid Model Based on GRU and XGBoost. *Sci. Total Environ.* **2021**, *783*, 146870. [CrossRef] [PubMed]
25. Ma, J.; Cheng, J.C.P.; Xu, Z.; Chen, K.; Lin, C.; Jiang, F. Identification of the Most Influential Areas for Air Pollution Control Using XGBoost and Grid Importance Rank. *J. Clean. Prod.* **2020**, *274*, 122835. [CrossRef]
26. Pan, B. Application of XGBoost Algorithm in Hourly PM2.5 Concentration Prediction. *IOP Conf. Ser. Earth Environ. Sci.* **2018**, *113*, 012127. [CrossRef]
27. Kim, M.; Brunner, D.; Kuhlmann, G. Importance of Satellite Observations for High-Resolution Mapping of near-Surface NO_2 by Machine Learning. *Remote Sens. Environ.* **2021**, *264*, 112573. [CrossRef]
28. Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *arXiv* **2017**, arXiv:1705.07874. [CrossRef]
29. Lundberg, S.M.; Nair, B.; Vavilala, M.S.; Horibe, M.; Eisses, M.J.; Adams, T.; Liston, D.E.; Low, D.K.-W.; Newman, S.-F.; Kim, J.; et al. Explainable Machine-Learning Predictions for the Prevention of Hypoxaemia during Surgery. *Nat. Biomed. Eng.* **2018**, *2*, 749–760. [CrossRef]
30. Zhang, S.; Xing, J.; Sarwar, G.; Ge, Y.; He, H.; Duan, F.; Zhao, Y.; He, K.; Zhu, L.; Chu, B. Parameterization of Heterogeneous Reaction of SO_2 to Sulfate on Dust with Coexistence of NH_3 and NO_2 under Different Humidity Conditions. *Atmos. Environ.* **2019**, *208*, 133–140. [CrossRef]
31. Tao, H.; Xing, J.; Zhou, H.; Pleim, J.; Ran, L.; Chang, X.; Wang, S.; Chen, F.; Zheng, H.; Li, J. Impacts of Improved Modeling Resolution on the Simulation of Meteorology, Air Quality, and Human Exposure to PM2.5, O_3 in Beijing, China. *J. Clean. Prod.* **2020**, *243*, 118574. [CrossRef]

32. Sistla, G.; Zhou, N.; Hao, W.; Ku, J.-Y.; Rao, S.T.; Bornstein, R.; Freedman, F.; Thunis, P. Effects of Uncertainties in Meteorological Inputs on Urban Airshed Model Predictions and Ozone Control Strategies. *Atmos. Environ.* **1996**, *30*, 2011–2025. [CrossRef]
33. Shan, Y.; Liu, J.; Liu, Z.; Shao, S.; Guan, D. An Emissions-Socioeconomic Inventory of Chinese Cities. *Sci. Data* **2019**, *6*, 190027. [CrossRef] [PubMed]
34. Shen, Y.; Jiang, F.; Feng, S.; Zheng, Y.; Cai, Z.; Lyu, X. Impact of Weather and Emission Changes on NO_2 Concentrations in China during 2014–2019. *Environ. Pollut.* **2021**, *269*, 116163. [CrossRef] [PubMed]
35. Zhang, Z.; Li, H.; Ho, W.; Cui, L.; Men, Q.; Cao, L.; Zhang, Y.; Wang, J.; Huang, C.; Lee, S.; et al. Critical Roles of Surface-Enhanced Heterogeneous Oxidation of SO_2 in Haze Chemistry: Review of Extended Pathways for Complex Air Pollution. *Curr. Pollut. Rep.* **2024**, *10*, 70–86. [CrossRef]
36. Chen, Z.; Chen, D.; Zhao, C.; Kwan, M.; Cai, J.; Zhuang, Y.; Zhao, B.; Wang, X.; Chen, B.; Yang, J.; et al. Influence of Meteorological Conditions on PM2.5 Concentrations across China: A Review of Methodology and Mechanism. *Environ. Int.* **2020**, *139*, 105558. [CrossRef]
37. Chen, H.; Xu, Y.; Zhong, S.; Mo, Y.; Zhu, S. Mapping Nighttime PM2.5 Concentrations in Nanjing, China Based on NPP/VIIRS Nighttime Light Data. *Atmos. Environ.* **2023**, *303*, 119767. [CrossRef]