Authors' Response to Reviews of

A Transformer-based agent model of GEOS-Chem v14.2.2 for informative prediction of $PM_{2.5}$ and O_3 levels to future emission scenarios: TGEOS v1.0

Dehao Li, Jianbing Jin*, Guoqiang Wang, Mijie Pang, Hong Liao*

Geoscientific Model Development Discussions, 10.5194/egusphere-2025-2186

RC: ***Reviewers' Comment***, AR: Authors' Response, ☐ Manuscript Text

1. Overview

Response to Referee 2: We would like to thank the referee for the careful review throughout the paper and the in-depth comments that help to improve our paper.

2. Major concerns

RC: ***The extensive critique of Response Surface Models (RSM) in Sections 1 appears disconnected from the proposed Transformer-based TGEOS framework. While RSMs rely on empirical statistical approximations to reduce dimensionality, TGEOS operates as a pure deep learning emulator that directly maps high-dimensional inputs to outputs. Thus, positioning TGEOS as addressing core RSM challenges misrepresents its paradigm. The review should focus on deep learning emulator challenges and explicitly contextualize innovations against relevant works like NN-CTM (Huang et al., 2021). Crucially, benchmarking against only architecturally inferior models (RF/MLP) – rather than comparable deep learning approaches like CNN-based Deep-RSM (Xing et al., 2020) or NN-CTM – undermines claims of Transformer superiority.***

AR: We thank the reviewer for this comment. In this research, RSM techniques were selected for discussion due to its ongoing development and established reliability within existing CTM simulators. RSMs were constructed based on the nonlinear relationship between emissions and concentrations using statistical methods, enabling rapid estimation of pollutant concentrations under varying emission scenarios. This characteristic makes RSMs closely aligned with the TGEOS model used in this study at the application level, leading us to focus primarily on the limitations of RSMs. In addition, we did not compare TGEOS with the previous DL-based emulators, as TGEOS differs from these models in terms of time resolution, learning objectives, and applicable scenarios, making direct comparison infeasible. For example, DeepRSM uses CMAQ as its target and is designed specifically for response prediction under a uniform regional emission coefficient (Xing et al., 2020), limiting its applicability to more detailed emission scenarios like DPEC-SSP/DPEC-CA scenarios used in this study. Following the reviewer's comments, we have revised the original manuscript to ensure that the discussion encompasses existing CTM simulator technologies more broadly, rather than focusing solely on RSMs. Details are shown in blew.

Introduction (L59-L106)

~~To address the computational challenge and efficiently retrieve the nonlinear relationship between~~

~~emissions and concentrations, data-driven statistical emulators have been proposed to accelerate numerical simulations (Castruccio et al., 2014). A reliable emulator can accurately depict intricate relationships between inputs and outputs, such as from emissions to concentrations. It can also faithfully approximate the fundamental mechanisms of atmospheric models, thereby generating numerical simulations that exhibit a high degree of consistency to the model (Salman et al., 2024). Among all the emulators, Response Surface Model (RSM) is the most widely used method. It is a statistical method developed by the US EPA (EPA, 2006) that uses the maximum likelihood estimation - empirical best linear unbiased predictors (MLE-EBLUPs) technique (Santner et al., 2003) to establish the complex relationships between emission rates of several pollutants and the responses they produce on the pollutant concentrations by fitting response surfaces of the nonlinear system (Box and Draper, 2007), and provide best estimate of the pollutant. When given some unknown emission scenarios, RSM can rapidly retrieve the changes of aimed concentrations without additional CTM simulation involved (Wang et al., 2011). RSM technique has been successfully employed in the response modeling of $PM_{2.5}$ (Wang et al., 2011) and ozone (Xing et al., 2011) to precursor emissions in China for typical regions. Since conventional RSM commonly requires a large number of CTM simulations to fit reliable response surfaces (Xing et al., 2011; Zhao et al., 2015), notable advances focusing on enhancements in both efficiency and accuracy in RSM technology have been achieved (Li et al., 2022). For example, Extended Response Surface Models (ERSMs) (Zhao et al., 2015; Xing et al., 2017) allow for the incorporation of a greater number of variables and geographical regions, improving alignment with independent CTM simulations compared with traditional RSM (Zhao et al., 2015; Xing et al., 2017). Moreover, the polynomial function based RSM (pf-RSM) is capable of quantifying the nonlinear relationships between air pollutant concentrations and precursor emissions by fitting CTM simulations to a series of polynomial functions and mitigating the computational burden through decreasing the number of required CTMs up to 60% (Xing et al., 2018). Recently, many studies have used novel machine learning techniques to accelerate the modeling process of RSM by further reducing the number of required CTMs. For instance, Deep-RSM, developed by Xing et al. (2020) using convolution neural networks (CNN), requires only two CTM cases (i.e., base and control scenarios) to startup the model; Self-adaptive RSM (SA-RSM, Li et al. (2022)) further reduces the number of required CTMs for pf-RSM modeling by employing a stepwise regression method to estimate the coefficients of polynomial functions.~~

~~Although existing RSM techniques exhibit more efficiency than traditional CTM in predicting the response of pollutant concentrations to a wide range of emission changes, there are still several issues to be addressed. Firstly, due to the structural limitations that restrict the model from executing multi-target predictions, existing techniques focus mainly on the response of average of the target pollutants over a period of time, such as the monthly average (Huang et al., 2021). However, predicting the singular monthly average of pollutant concentrations may overlook critical variations throughout the month, such as extreme values (Guo et al., 2020; Zhao et al., 2022). Therefore, these approaches fall short in providing a comprehensive evaluation of future pollution states, including the ability to identify potential extreme pollution events under various emission scenarios. Secondly, RSM techniques rely on the polynomial assumption, leading to its disadvantage to cope with high-dimension problems. As the number of input variables increases, the complexity of RSM model grows, necessitating a larger number of samples for accurate fitting (Zhao et al., 2015) and potentially leading to multi-collinearity issues (Xing et al., 2018). This limitation restricts the applicability of RSM to more intricate emission scenarios. Therefore, existing RSM studies have primarily concentrated on emissions of a few major pollutants and the add-up emissions, failing to address air quality response under more detailed~~

~~scenarios that incorporate sectoral emissions and a broader range of emission species. While ERSM considers emission sectors (Zhao et al., 2015), the inherent limitations of RSM in handling high-dimensional data result in a substantial requirement for CTM samples, thus confining its application to modeling studies in smaller areas. Thirdly, current RSMs e.g. pf-RSM (Xing et al., 2018) and SA-RSM (Li et al., 2022) account for each spatial grid independently while neglect the impact of surrounding emissions, which have been shown to affect local pollutant concentrations (Cheng et al., 2019). While ERSM (Zhao et al., 2015) has considered regional transport of emissions, it requires a substantial number of scenario simulations to ensure the accuracy of the model (Zhao et al., 2015; Xing et al., 2017). For example, modeling for a middle-scale region typically necessitates hundreds of scenarios as support (Zhao et al., 2015). The computational burden significantly limits the application of this technology on a national scale. In summary, given that existing techniques inadequately address the challenges associated with high temporal-resolution prediction, inapplicability of multivariate scenarios, and negligence of emission transport, developing a comprehensive national-level "emission-concentration" predictive model poses a significant challenge.~~

To overcome the computational challenge and efficiently retrieve the nonlinear relationship between emissions and concentrations, data-driven statistical emulators have been proposed to accelerate numerical simulations (Castruccio et al., 2014). As a simplified-form of CTM, a reliable emulator can effectively capture the intricate relationships between important CTM inputs and concentration outputs, and rapidly estimate "CTM-aligned" concentrations of pollutants. Response Surface Model (RSM), served as statistical surrogates developed by the US EPA (EPA, 2006) to establish the relationships between emission rates and the concentration responses of CTM, has been continuously developed since the past decade. RSM techniques have been successfully employed in the response modeling of PM$_{2.5}$ (Wang et al., 2011) and O$_3$ (Xing et al., 2011) to precursor emissions in China for typical regions. To address the inherent computational burden stemmed from considerable advanced CTM supports for model building (Xing et al., 2011), optimized versions of conventional RSM were developed, such as ERSM (Zhao et al., 2015; Xing et al., 2017) and pf-RSM (Xing et al., 2018). Recently, novel machine learning (ML) techniques, for its well performance in simulating complex non-linear relationships in atmospheric systems (Liu et al., 2021) and dealing with tasks involving multiple variables and objectives (Masmoudi et al., 2020; Huang et al., 2021), have been employed in RSM techniques to further optimize modeling efficiency and estimation accuracy of RSMs (Xing et al., 2020; Li et al., 2022). Based on this advantage, many studies have attempted to build effective emulators using pure ML method (Huang et al., 2021; Zhang et al., 2023a). For example, Zhang et al. (2023a) used ResCNN framework to predict annual PM$_{2.5}$ concentration from fossil energy use and reveal the co-benefits of the energy transition, demonstrating the potential of ML method in addressing the emulator modeling task.

Although existing CTM emulators exhibit more efficiency than traditional CTM in estimating the pollutant concentrations to a wide range of emission changes, there are still several issues to be addressed. Firstly, due to the computing limitations (Liu et al., 2022), the temporal resolution for some emulators was constrained with annual scale, which greatly prevent these emulators from providing detailed estimations of air pollutants such as extreme values throughout the year (Guo et al., 2020; Zhao et al., 2022). Secondly, while some emulators have the ability to offer concentration estimations with finer temporal resolution, they still have limitations. On one hand, RSM-based emulators rely on the polynomial assumption, leading to its disadvantage to cope with high-dimension problems. As the number of input variables increases, the complexity of RSM model grows, necessitating a larger number of samples for accurate fitting (Zhao et al., 2015) and potentially leading to multi-collinearity

> issues (Xing et al., 2018). In the revised manuscript, we will provide examples (BTH, YRD) to avoid ambiguity. This limitation restricts the applicability of RSM-based emulators to more intricate emission scenarios. Therefore, existing RSM studies have primarily concentrated on emissions of a few major pollutants and the add-up emissions (Xing et al., 2020), failing to address air quality response under more detailed scenarios that incorporate sectoral emissions and a broader range of emission species. On the other hand, some emulators were constructed based on in-situ observations using ML method (Zhang et al., 2023a), which is easy to employ and more convenient than those RSM-based emulators. However, these models are constrained by the limited number of observational data stations and are therefore unable to effectively assess air quality in regions where observational infrastructure is lacking (Xu et al., 2022). Furthermore, due to insufficient observational data, these models often do not have enough representative samples to achieve accurate model fitting, which leads to suboptimal predictive performance (Tang et al., 2024). In addition, traditional ML models, such as Multi-Layer Perceptron (MLP) and Random Forest (RF), may not fully capture the nonlinear relationships in complex atmospheric variables (Masmoudi et al., 2020; Natarajan et al., 2024; Abuouelezz et al., 2025), which further undermine their predictions. Thirdly, some current emulators account for each spatial grid or observation site independently while neglect the impact of surrounding emissions (Xing et al., 2018; Li et al., 2022; Zhang et al., 2023a), which have been shown to affect local pollutant concentrations (Cheng et al., 2019). Although certain studies have employed convolutional neural network (CNN) architectures capable of capturing local features to develop models (Xing et al., 2020; Huang et al., 2021; Liu et al., 2022), the computational resource constraints have hindered these "face-to-face" models from processing large volumes of feature inputs. As a result, the application of such models is limited in terms of emission details and research domain. In summary, given that existing techniques inadequately address the challenges associated with high temporal-resolution prediction, inapplicability of multivariate scenarios, and negligence of emission transport, it still be a significant challenge to develop a comprehensive emulator using more advanced method.

Additionally, due to the distinct form of the input data derived from TGEOS, RF and MLP-rather than CNN architecture that has been employed for emulator building (Xing et al., 2020; Huang et al., 2021)-were selected for model comparison. Previous models represent both input and output features in matrix forms (Xing et al., 2020; Huang et al., 2021), facilitating a "face-to-face" modeling approach that is well-suited for CNNs, which is commonly used in image processing (Li et al., 2021). In contrast, the input of the TGEOS consists of sequential samples from individual grids, for each sample containing 1045 features (mentioned in Table 2 of the manuscript), making it incompatible with the CNN framework and thus not considered.

To highlight the advantages of the Transformer architecture, an attempt was made to construct a CNN-based model for comparative analysis. The basic architecture of this model is illustrated in Figure 1. In this CNN-based model, we transformed the feature input of each sample from its original dimension of (1, 1045) into a matrix format of (9, 116). For the temporal features (i.e., month information corresponding to each scenario in this study), we individually convert them into embedding vectors—following an approach commonly used in NLP (Stankevičius and Lukoševičius, 2024)—and subsequently concatenate these vectors with the flattened output of the final convolutional layer of the CNN before feeding them into the fully connected layer. The same training and test set of TGEOS were used for model training and validation.

As illustrated in Figure 2, the performance of the four models on the test set is compared. On one hand, compared to the previously selected MLP and RF models, the CNN-based model demonstrates superior performance, characterized by higher R values as well as lower MAE. This advantage can be attributed to the CNN's local convolution kernel, which is capable of capturing patterns among adjacent data points. On the other hand, when compared to TGEOS employed in this study, the CNN-based model underperforms across

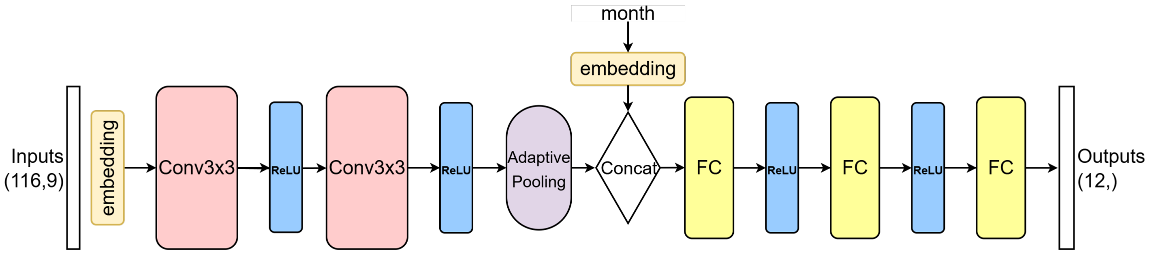

Figure 1: Basic architecture of CNN-based model.

all evaluation metrics. This is primarily due to TGEOS's self-attention mechanism, which enables more effective dynamic and global modeling. In contrast to CNNs, which are constrained by fixed convolution kernels and limited network depth, Transformer-based TGEOS exhibits a stronger capacity for capturing complex relationships in high-dimensional data.

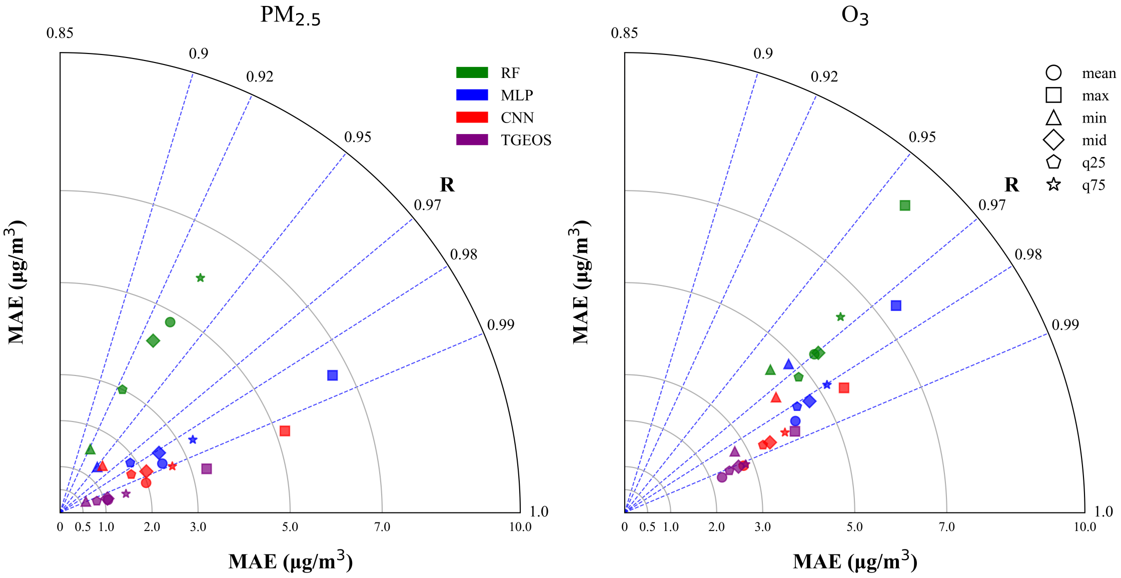

Figure 2: Predictive performance of four models, with green represents RF predictions, blue denotes MLP predictions, red denotes CNN predictions, and purple indicates TGEOS predictions. All indicators are averaged in national scale and computed based on the six test scenarios.

We have updated the "Comparison of Different Machine Learning Models" section of the manuscript to include a discussion on the CNN model. The details are provided below:

> To validate the performance of the TGEOS model in "emission-concentration" modeling against other machine learning models, ~~two widely used machine learning models, including multilayer perceptrons (MLP) and random forests (RF)~~ three widely used machine learning frameworks, namely Multilayer Perceptrons (MLP), Random Forests (RF), and Convolutional Neural Network (CNN) employed in previous studies (Xing et al., 2020; Xing et al., 2021), were simultaneously employed based on the

multi-scenario dataset mentioned in Section 2.1. For each ML model, we identified the model ~~that demonstrated optimal fitting performance for testing after conducting a series of parameter tuning experiments.~~ with the best combination of hyperparameters after fine-tuning process based on Optuna tool. The MLP model uses 4 hidden layers with 2048, 1024, 512, and 256 neurons, applying ReLU activation and Dropout to prevent overfitting. The optimizer is Adam with a learning rate of $1e^{-3}$, and the loss function is Mean Squared Error (MSE). Training uses a batch size of 1024 and 100 epochs, with a learning rate scheduler to adjust the learning rate dynamically. The RF model uses 300 trees with a maximum depth of 25, a minimum sample split of 4, and a minimum sample per leaf of 2. It uses parallel computation with all CPU cores and performs feature selection by choosing the top 500 important features. The CNN model consists of two convolutional layers followed by fully connected layers, with an additional embedding layer to incorporate month information. The first convolutional layer applies 32 filters of size 3×3 (padding = 1) to the single-channel input, followed by a second convolutional layer with 128 filters of the same size. Both convolutional layers use ReLU activations. The output feature maps are then processed by an adaptive average pooling layer to reduce the spatial resolution to 29×3. To integrate temporal information, a month embedding layer maps month indices (1–12) to a 4-dimensional vector. The pooled convolutional features are flattened and concatenated with the month embedding, forming the input to a three-layer fully connected network: the first linear layer maps the concatenated vector to 256 units, the second reduces it to 64 units, and the final output layer produces 12 regression targets. ReLU activation functions are applied after the first and second fully connected layers. For each model, hyperparameters were obtained after fine-tuning based on Optuna tool.

Table S2 and S3 summarize the performance of the three models on the test set. We found that TGEOS outperforms the other two models in both R^2 and MAE metrics. To clearly illustrate the predictive performance of different models, we presented a modified Taylor diagram (Taylor, 2005; Fang et al., 2023) in Fig. 10. This diagram simultaneously displays the Mean Absolute Error (MAE) and correlation coefficient (R) for predictions of $PM_{2.5}$ and O_3 indicators from three models in various regions. Our findings indicate that the Random Forest (RF) model performs the poorest. This is primarily due to its reliance on feature importance assessments during feature selection, which overlooks potential underlying features in the data, adversely affecting the model's fitting capability. Additionally, the RF model is sensitive to the distribution of training data, leading to limited extrapolation abilities and poor predictive performance for extreme values. In contrast, the Multi-Layer Perceptron (MLP) shows a significant improvement in predictive performance relative to the RF model. Leveraging its multi-layer neural network structure, the MLP can more effectively learn complex relationships between multiple features. But this layered structure can struggle when dealing with high-dimensional feature spaces, especially for highly stochastic indicators such as maximum values, where the MLP still exhibits considerable prediction errors. Compared to the previously selected MLP and RF models, the CNN-based model demonstrates superior performance, characterized by higher R values as well as lower MAE. This advantage can be attributed to the CNN's local convolution kernel, which is capable of capturing patterns among adjacent data points.

Conversely, the Transformer-based TGEOS model demonstrates superior performance compared to the other models, exhibiting higher R values (exceeding 0.98 and 0.97) and lower MAE values (less than 4.0 g/m^3 for the majority indicators). These results suggest a higher degree of reliability and accuracy in its predictions. For several indicators where MLP performs poorly, TGEOS demonstrates substantial improvements. The superiority of the Transformer model can be attributed to its greater number of parameters and more complex architecture, which leverage powerful feature extraction capabilities and self-attention mechanisms, allowing it to adapt to intricate patterns and relationships.

In contrast to CNN, which are constrained by fixed convolution kernels and limited network depth, TGEOS exhibits a stronger capacity for capturing complex relationships in high-dimensional data. ~~Consequently, in high-dimensional tasks like air quality modeling, Transformer models have proven to be more advantageous compared to their counterparts.~~

RC: ***The exclusive use of 2017 MERRA-2 meteorology across all 36 emission scenarios creates critical limitations. (1) Artificial performance inflation: Model validation (Section 3) only tests emission sensitivity under identical meteorological conditions, ignoring O’s established sensitivity to temperature/radiation. This likely overstates accuracy for real-world applications where meteorology co-varies. (2) Unverified generalizability: No experiments challenge the model with meteorological variability (e.g., heatwaves), leaving robustness under climate fluctuations untested. (3) Neglect of emission-climate feedbacks: The abstract positions TGEOS for "future emission scenarios", yet fixed meteorology cannot capture feedbacks like emission-driven aerosol-radiation interactions affecting O_3. Given the study’s policy-assessment ambitions, this design flaw is critical. Cross-meteorological sensitivity tests should quantify key indicator fluctuations to establish operational reliability.***

AR: We appreciate the reviewer for the insightful comment. At the beginning, we fully acknowledge that the use of a fixed 2017 meteorological field limits TGEOS’s ability to capture meteorology–emission interactions or emission–climate feedbacks. Indeed, this design precludes assessing the influence of future climate variability (e.g., temperature/radiation changes, extreme events) on pollutant concentrations.

However, the primary objective of TGEOS in this study is to support air quality predictions under future emission scenarios, with a specific focus on isolating the concentration responses attributable to emission changes. So we intentionally did not include climate change effects in this work. The “fixed meteorology (based on a certain meteorological year or meteorological field) with different emission scenarios” framework has been widely adopted in future air quality assessment studies. For example, Shi et al. (2021) simulated future air quality in China under carbon neutrality using the WRF-Comprehensive Air Quality Model with Extensions (WRF-CAMx) with the meteorology fixed at 2019; Liu et al. (2022) adopted machine learning approach to explore the interaction patterns between air-quality improvement and climate change mitigation in China, using 2017 meteorology ; Wang et al. (2023) used GEOS-Chem model with identical meteorology of year 2015 to assess the changes in concentrations of $PM_{2.5}$ and O_3 and associated health impacts. In addition, many studies, namely He et al. (2018), Xiao et al. (2021) and Bhattarai et al. (2024), have also employed similar designs to quantify the contribution of emissions to future air pollutant concentrations. This approach enables a clean separation of emission-driven changes in pollutant levels, thereby allowing a clearer estimation of mitigation benefits without the confounding influence of meteorological variability.

Additionally, simultaneously incorporating both emissions and meteorology changes in future projections is a technically challenging task. From the modeling perspective, it requires training the emulator to represent complex, nonlinear interactions between meteorology and emissions across a much larger parameter space. From the data perspective, the resolution of the currently available future meteorological data is excessively coarse (native 100-km resolution), and the number of available meteorological variables is limited (mainly focus on temperature and precipitate (Zhang et al., 2025)) , which makes it challenging to comprehensively represent the future meteorological field required for GEOS-Chem input.

To provide a preliminary evaluation of TGEOS’s generalization capability across different meteorology scenarios, we designed a cross-meteorology sensitivity experiment, as summarized in Table 1.

* **Group 1:** The baseline TGEOS, trained exclusively on scenarios with 2017 meteorology (identical to the model described in the manuscript).

- **Group 2:** An extended TGEOS, trained on scenarios with 2017 meteorology and supplemented by nine additional scenarios with 2014 meteorology.
- **Test set:** Six scenarios with 2022 meteorology.

The results of the two groups are shown in Fig. 3 and Fig. 4. Although the baseline TGEOS (Group 1) exhibited reduced performance when applied to 2022 meteorology compared to its performance under 2017 meteorology (R^2> 0.9 in the manuscript), it still achieved acceptable predictive skill, with R^2 values exceeding 0.8. In contrast, the extended TGEOS (Group 2) consistently outperformed Group 1, demonstrating that incorporating more meteorological conditions into the training set could stably enhance the model's robustness under unseen climatic conditions. It should be noted that reconciling the combined effects of emissions and meteorology is inherently challenging, and a sufficiently large number of simulated samples across diverse meteorology scenarios is essential to achieve reliable predictions. Here only a limited set of samples was used to illustrate the feasibility of the proposed approach. Developing the capability for rapid air quality predictions across varying meteorological scenarios will be the central focus of our next work.

Table 1: Design of cross-meteorology sensitivity experiments.

Experiment	Scenario number	Description
Group1	36	36 scenarios as depicted in Table 1 of the manuscript
Group2	45	36 scenarios of Group1, 9 scenarios with emissions of SSP1_2030, SSP1_2040, SSP1_2050, SSP4_2030, SSP4_2040, SSP4_2050, SSP5_2030, SSP5_2040, SSP5_2050 with meteorology of 2014
Test	6	6 scenarios with emissions of SSP2_2030, SSP2_2040, SSP2_2050, SSP3_2030, SSP3_2040, SSP3_2050, with meteorology of 2022

Finally, we have revised in Conclusion section of the manuscript to emphasize the limitation of the meteorology-fixed methodology.

> The TGEOS model still have some limitations to be improved. Firstly, it should be noted that the predictions generated by TGEOS remain incapable of accurately representing actual air pollutant concentrations, even though TGEOS is highly consistent with GEOS-Chem, since systematic biases have been demonstrated to exist within GEOS-Chem itself (Travis and Jacob, 2019; Miao et al., 2020). Therefore, correcting errors in TEGOS based on near-real observations or reanalysis data is of paramount importance and constitutes a priority for our subsequent research. ~~Additionally, due to the considerable effect of meteorological conditions on the generation (Shi et al., 2020), spatiotemporal patterns (Zhang et al., 2013; Chen et al., 2020), and concentration levels (Wang et al., 2019) of $PM_{2.5}$ and O_3 concentrations, and meteorological conditions other than 2017 are not considered in this study. Consequently, there is also a need to incorporate various climate scenarios that represent meteorological variations to enhance the TGEOS's predictive capability regarding future air quality under more complex scenarios with variations in emissions and meteorology.~~ Additionally, in order to isolate the impact of emission changes to future air quality as previous studies did (Shi et al., 2021; Shi et al., 2023), the meteorology used in this study was fixed at 2017. The identified meteorology limits TGEOS's ability to generate reliable estimations in cross-meteorology scenarios, and to capture meteorology–emission interactions

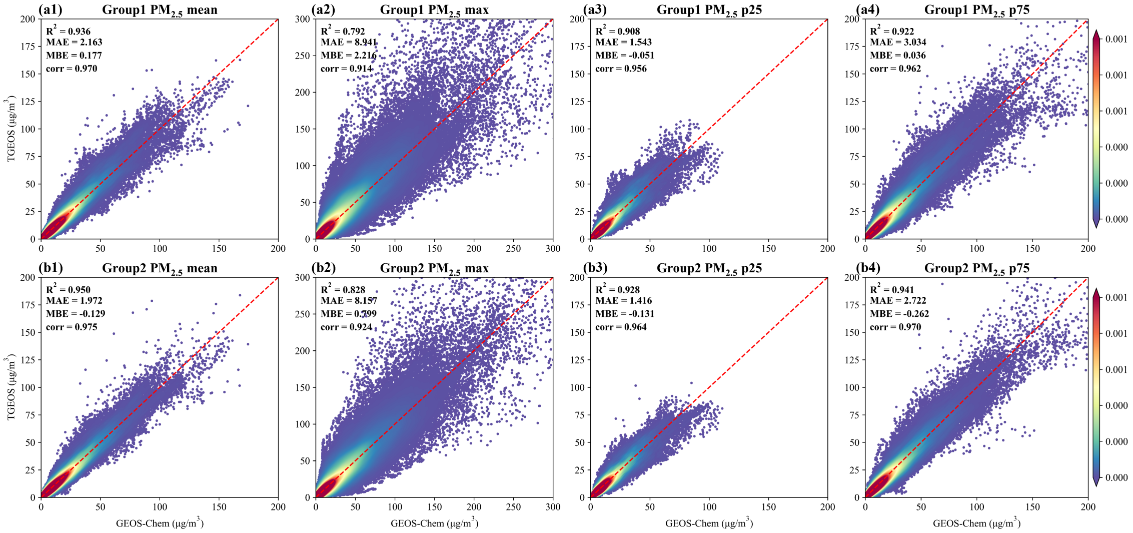

Figure 3: Results of two experiments for $PM_{2.5}$ in SSP3_2030 scenario. a1 to a4 shows predictions of mean, max, 25 percentile and 75 percentile based on TGEOS with Group1 training strategy. b1 to b4 shows predictions using TGEOS with Group2 training strategy.

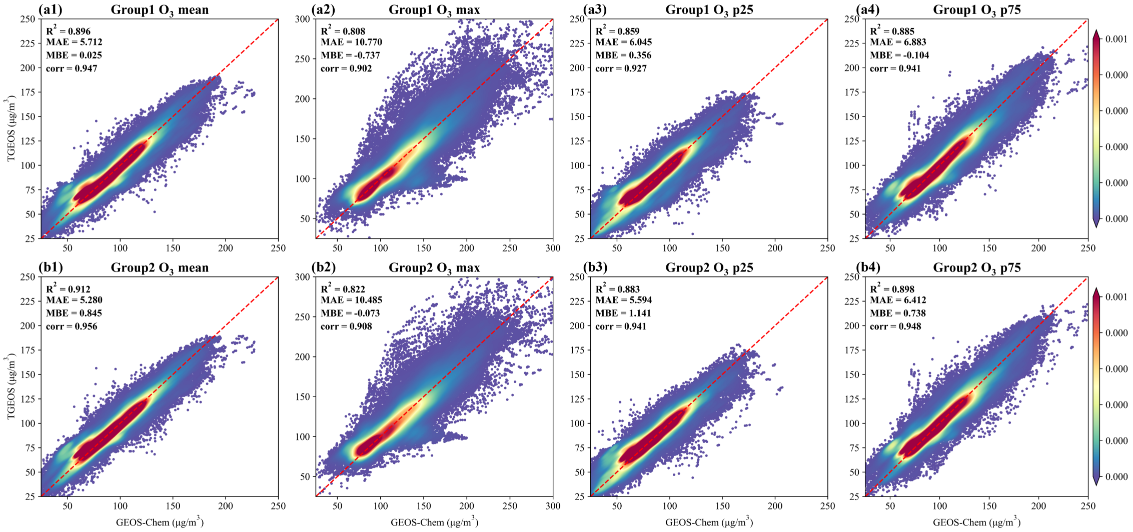

Figure 4: Same as Fig. 3, but for O_3 predictions in SSP3_2030 scenario.

and “emission–climate” feedbacks. Consequently, there is also a need to incorporate various climate scenarios that represent meteorological variations to enhance the TGEOS’s predictive capability regarding future air quality under more complex scenarios with variations in emissions and meteorology.

RC: ***The dataset design introduces potential leakage between training (DPEC-SSP SSP1/4/5 + DPEC-CA + tuning) and testing (DPEC-SSP SSP2/3) sets. (1) Structural homology: All scenarios derive from the DPEC framework, sharing inherent inventory structures, sectoral mappings, and spatial patterns. (2) Meteorological invariance: Identical 2017 meteorology further constrains emission-concentration mapping diversity. (3) Unverified independence: The sole qualitative comparison (otp2030 vs. SSP2_2050) lacks quantitative validation. No analysis demonstrates statistical separability between DPEC-SSP, DPEC-CA, and tuning scenarios. Sampling only three years (2030/2040/2050) within these correlated DPEC trajectories risks distributional overlap. This undermines claims of rigorous holdout testing, especially given the high R² values (0.96+) that may reflect dataset artifacts rather than true generalizability.***

AR: Thank the reviewer for this comment. We acknowledge that in the test set (SSP2/SSP3), certain grid cells exhibit emission or concentration levels very similar to those in the training set. This is primarily because, within the DPEC framework, emission changes are implemented in a spatially coherent and relatively smooth manner, which reflects the realistic evolution characteristics of emissions in the future related to policies. As a result, even though these scenarios differ in their long-term emission trajectories, localized changes remain gradual, leading to the potential convergence between training and testing sets. Furthermore, in regions with intrinsically low emission levels, such as western China, differences in both emissions and concentrations across scenarios are rather minimal. Consequently, the model can achieve high predictive accuracy in these areas even under unknown scenarios, which partly contributes to the strong performance observed in the test set.

As discussed in Section 2.1.1 of the manuscript, we had added 10 random scenarios based on perturbation method to enhance generalizability of TGEOS. These scenarios are independent from DEPC framework and we believe the model has the ability to make preliminary predictions under emission scenarios out of DPEC framework. Thus, we further evaluated TGEOS under two sets of randomly perturbed emission scenarios using same tuning method but with different scales (0.9 and 1.1) to assess TGEOS’s robustness beyond the structured future scenarios. As presented in Fig. 5 and 6. For O_3, the model maintained high predictive accuracy across both perturbation sets, demonstrating strong generalization to altered emission patterns. For $PM_{2.5}$, although overall performance remained high (R² > 0.89), the predictive skill declined compared to that achieved on the future-scenario test set, with noticeable biases in high-concentration predictions. This poorer performance is likely due to the more complex formation pathways of $PM_{2.5}$ as a secondary pollutant (Shi et al., 2024), combined with the fact that the training set primarily comprised 25 future scenarios with relatively smooth spatial emission variations. We anticipate that incorporating additional perturbed-emission scenarios into the training set would further enhance TGEOS’s robustness and predictive capability.

In response to the reviewer’s concern regarding potential data leakage, we specifically examined eight key emission variables that predominantly influence $PM_{2.5}$ and O_3 concentrations (Pinder et al., 2007; Wang et al., 2013; Lu et al., 2019; Skyllakou et al., 2021; Lai et al., 2021), and analyzed the distributions across three emission scenario sets: DPEC-SSP, DPEC-CA, and tuning scenarios. As illustrated in Fig.7 and Fig.8, although the magnitudes are generally comparable, noticeable differences among the three curves are evident for each emission variable, thereby confirming the separability of these datasets. For the otp2030 and SSP2_2050 scenarios, we further conducted the Kolmogorov–Smirnov (K–S) test, with the results summarized in Table 2. Given the large sample size, the p-values for all emission variables are approximately zero (Demir, 2022), making the KS statistic (D-value) a more meaningful indicator. Our analysis shows

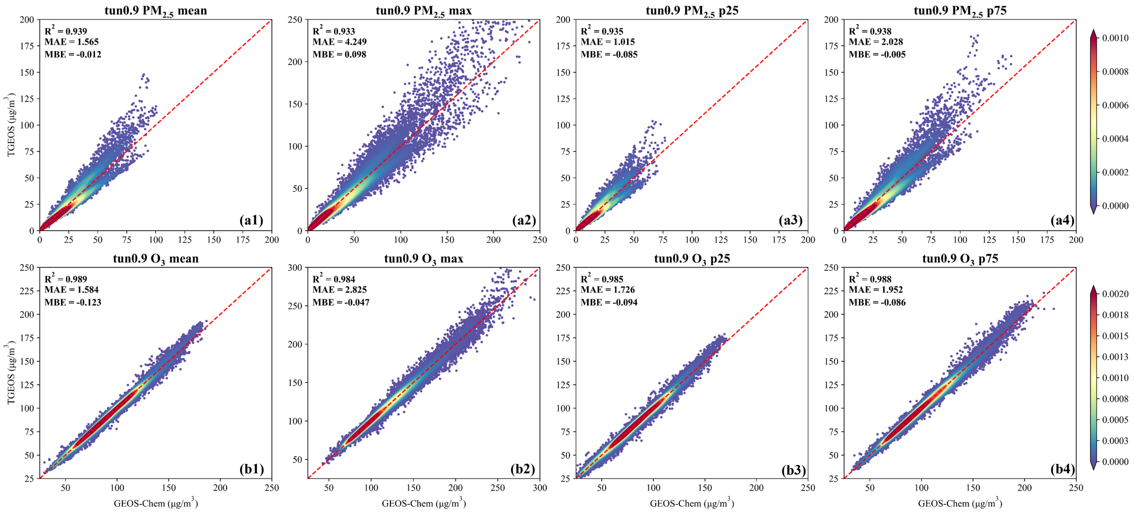

Figure 5: Density scatter plots between GEOS-Chem simulations and TGEOS predictions for four indicators of PM$_{2.5}$ and O$_3$ concentrations in tun0.9 scenario, where a1 to a4 denotes the mean, maximum, 25th percentile, and 75th percentile of January PM$_{2.5}$ concentration; b1 to b4 denotes the corresponding statistics for July O$_3$ concentration.

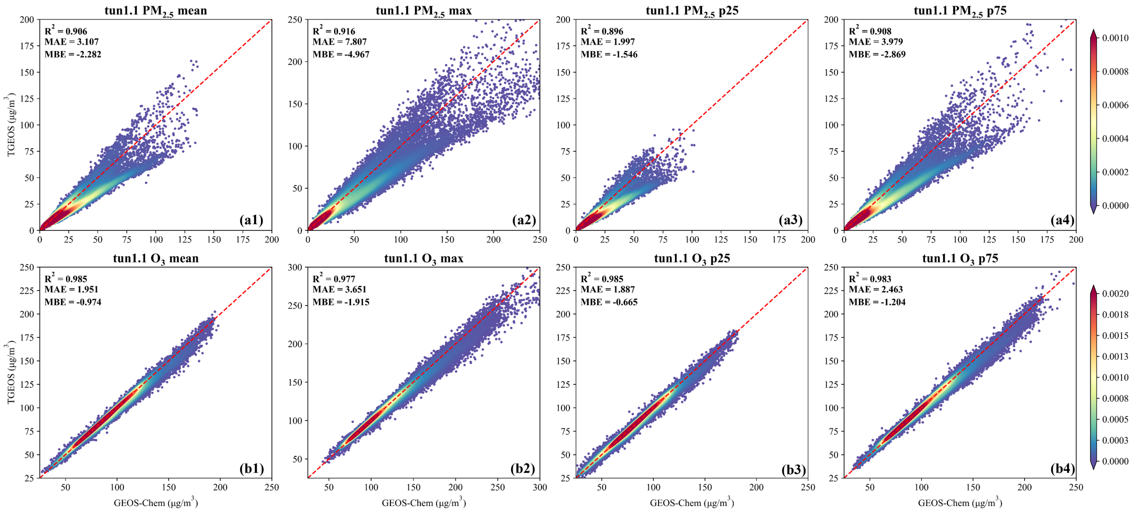

Figure 6: Same as Fig. 5 but for tun1.1 scenario.

that the emissions of the two scenarios differ to varying extents, with all D values being greater than zero. As discussed earlier, the convergence of future emission trends and the presence of grid cells less sensitive to emission perturbations contribute to the similarity among future emission scenarios. This similarity is reflected in relatively small D values (typically < 0.3). Nevertheless, for most emission variables, D values exceed 0.1, suggesting that non-negligible differences still exist between the two scenarios.

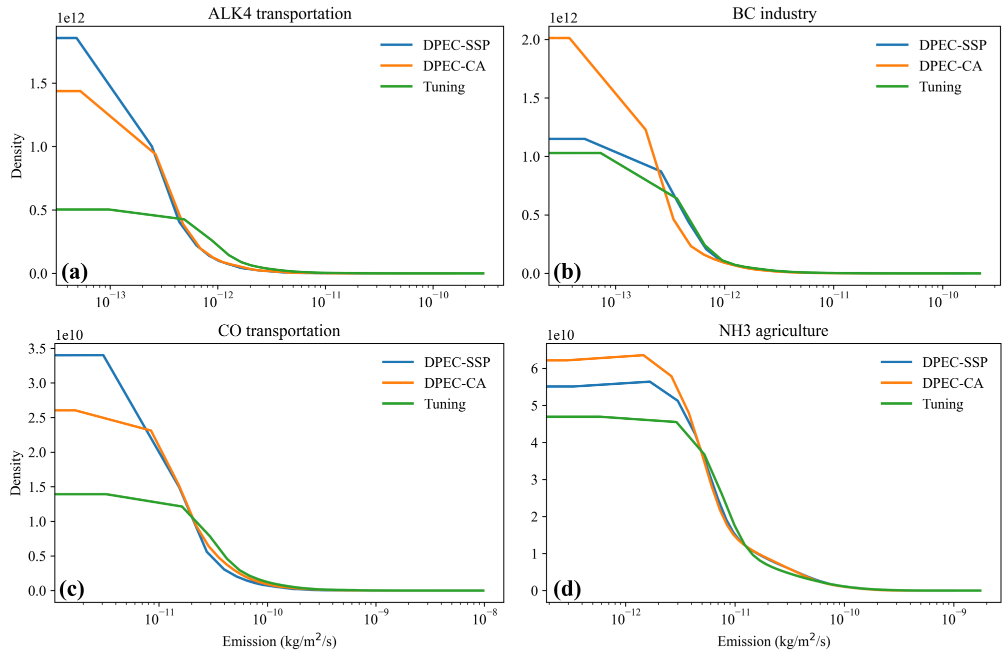

Figure 7: Kernel Density Estimation (KDE) curves for ALK4 transportation (a), BC industry (b), CO transportation (c), and NH_3 agriculture (d) emissions in three scenario sets on semi-logarithmic scales.

We have revised the responding part of the manuscript to present the the difference between training and test set, and compare two scenarios based on spatial distribution and K-S test. Details are illustrated in blew.

3.1 Differences between training and test set (L265-L274)

~~In case of potential data leakage due to similar concentration and emission levels in some emission scenarios, we analyzed the spatial distribution of the mean values of $PM_{2.5}$ and O_3 under the stochastically selected SSP2_2050 scenario of the test set, along with the otp2030 scenario of the training set that exhibiting the highest similarity in concentration levels of SSP2_2050 scenario. Focusing on the North China Plain (NCP) region where both $PM_{2.5}$ and O_3 pollution are severe, the absolute difference in concentration between two scenarios was demonstrated. Concurrently, distributions of six emission variables of two scenarios with significant impacts on $PM_{2.5}$ and O_3 concentrations (Hu et al., 2023), as well as the differences, were also analyzed. The spatial distribution of the concentration, emission, and absolute difference levels of $PM_{2.5}$ are shown in Fig. 2, and those of O_3 are illustrated in Fig. S1. These pictures indicated that the concentrations of pollutants, as well as emission variables, of the training and test set are exclusive despite some distributional similarities, particularly for samples from highly polluted regions.~~

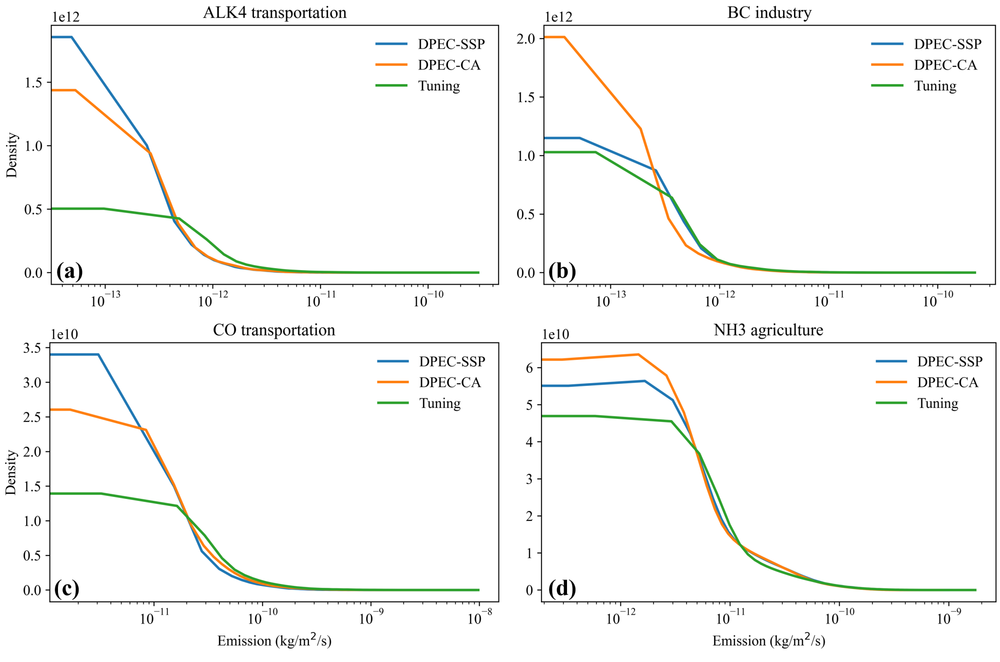

Figure 8: Same as Fig. 7, but for NO transportation (a), SO$_2$ power (b), PM$_{2.5}$ residential (c), and XYLE industry (d).

Since emission trajectories with different reduction rates may converge at certain time horizons, there exists a potential risk of data leakage arising from similarities in emission and concentration levels across scenarios. To address this concern, we analyzed the Kernel Density Estimation (KDE) curves for six key emission variables, which strongly influence PM$_{2.5}$ and O$_3$ concentrations (Hu et al., 2023), of the training and test set, as illustrated in Fig. S5. The results indicate that, although the general distribution trends are similar, the densities at different emission levels vary significantly between the two. Furthermore, focusing on the North China Plain (NCP) where both PM$_{2.5}$ and O$_3$ pollution are particularly severe, we examined the spatial distribution of mean PM$_{2.5}$ and O$_3$ concentrations, six critical emissions as well as corresponding absolute differences under the stochastically selected SSP2_2050 test scenario, in comparison with a training scenario (otp2030). The otp2030 scenario was selected by calculating the Euclidean distance between the mean PM$_{2.5}$ and O$_3$ values of SSP2_2050 and those of each training scenario, and identifying the scenario with the minimum distance. The results are illustrated in Fig. 2 for PM$_{2.5}$ and Fig. S4 for O$_3$. These pictures indicated that the concentrations of pollutants, as well as emission variables, of the training and test set are exclusive despite some distributional similarities, particularly for samples from highly polluted regions.

In addition, we conducted Kolmogorov-Smirnov (K-S) tests on a total of 12 emission variables, comprising the aforementioned six emissions as well as an additional set of six emissions, with the results summarized in Table 2. Given the large sample size, the p-values for all emission variables are approximately zero (Demir, 2022), making the KS statistic (D-value) a more meaningful indicator. Our analysis shows that the emissions of the two scenarios differ to varying extents, with all D values

Table 2: Results of K-S test for eight emission variables of otp2030 and SSP2_2050 scenarios.

Variable	K-S stat	p-value
NO transportation	0.220	0.000
NO industry	0.109	0.000
NO power	0.100	0.000
NH_3 agriculture	0.071	0.000
SO_2 power	0.461	0.000
BC industry	0.224	0.000
OC residential	0.347	0.000
CO transportation	0.255	0.000
$PM_{2.5}$ residential	0.348	0.000
XYLE industry	0.019	0.085
ALK3 transportation	0.393	0.000
ALK4 transportation	0.433	0.000

being greater than zero. It is noteworthy that emission changes are primarily concentrated in major emission regions of eastern China, whereas in many western and southern regions the variations across scenarios are negligible. This spatial heterogeneity implies the presence of redundant samples in the dataset, which could in turn contribute to statistical similarities between scenarios when comparisons are made (D-value < 0.3). Nevertheless, for most emission variables, D values exceed 0.1, suggesting that certain differences still exist between the two scenarios.

RC: ***The claim that TGEOS predicts "probability distributions" is inconsistent with its methodology. (1) Temporal distribution gap: The model outputs six statistical indicators (e.g., monthly percentiles) per grid cell. However, no validation confirms these reconstruct temporal distributions at individual locations. (2) Spatial vs. temporal conflation: Section 3.3 analyzes spatial probability distributions aggregated across regions, which fundamentally differ from the temporal distributions implied by Section 3.2's grid-level statistics. This conceptual ambiguity obscures what "probability distribution" signifies in results. While the six indicators efficiently summarize central tendency and spread, presenting them as full probability distributions overstates methodological capabilities without empirical proof of distributional accuracy at the intended spatiotemporal scale.***

AR: We appreciate the reviewer's comment and agree that our terminology could be made more precise. Our primary goal was not to reconstruct full temporal probability distributions from limited statistics, but rather to predict a set of monthly-scale concentration indicators that closely match those derived from GEOS-Chem simulations under future emission scenarios. These six indicators are computed from daily mean concentrations within each month and are intended to capture essential statistical properties of pollutant levels, such as the 75th percentile and maximum concentration, that are particularly relevant for policy evaluation and extreme-event assessment (Reich et al., 2012; Zhang et al., 2022).

While this approach necessarily ignores the exact day-to-day temporal sequence, it enables a fast but informative representation of pollution conditions at the monthly scale. For example, in a given GEOS-Chem simulation, the peak $PM_{2.5}$ concentration for a future scenario may occur on 23 January. Although TGEOS cannot specify this exact date, since obtaining daily-resolved future concentrations is not technically feasible, it can still capture the magnitude of this potential extreme event through the prediction of monthly maximum. Moreover, we fit approximate probability distribution curves using these indicators to provide intuitive visualization of distribution patterns and potential extremes for future $PM_{2.5}$ and O_3. These PDF fittings are not intended as complete reconstruction of the true distribution, but rather as means of summarizing the predicted concentration characteristics that close to GC outputs.

Regarding Section 3.2, our intention was to demonstrate the spatial agreement between TGEOS predictions and GEOS-Chem simulations, including both elevated pollution months and other seasons. We were not implying temporal distribution reconstruction in this section. Given that the model did not incorporate initial concentration fields, each month's prediction was treated independently.

To avoid ambiguity, we have revised the manuscript to replace "predict probability distributions" with "predict key statistical summaries of monthly concentration distributions," and explicitly clarify the application scope for the TGEOS.

1 Introduction (L129-L131)

~~First, TGEOS is able to predict the probability distribution of future air quality under different emission scenarios. Compared to solely average estimated by previous RSM methods, probability distribution can provide informative frequency distributions of pollutants (yang et al., 2022).~~ First, TGEOS is able to predict the key statistical indicators of monthly concentrations, and provide approximations of probability distributions under different emission scenarios. Compared to solely average estimated by previous RFMs, probability distribution can provide informative frequency distributions of pollutants (yang et al., 2022). .

3. Minor concerns

RC: ***Lines 8: The term "online predictions" is ambiguous and potentially misleading. The claimed probability distributions are reconstructed from six statistical indicators rather than dynamically generated in real-time.***

AR: We appreciate the reviewer's comment and agree that the term "online predictions" may be ambiguous in this context. Our intention was not to imply dynamical or streaming predictions, but rather that the model timely generates monthly statistical indicators for each grid cell given future emission scenarios. These indicators reflect statistical features of $PM_{2.5}$ and O_3 concentrations to comprehensively present future air pollutants. To avoid confusion, we have revised the wording in the manuscript to use a more precise term. Details are shown below:

1. Abstract (L8)

In this study, an informative future air quality prediction model "TGEOS v1.0" based on the Transformer framework is developed as an efffcient agent model of GEOS-Chem v14.2.2. TGEOS is able to ~~swiftly and accurately conduct online predictions of probability distributions for $PM_{2.5}$ and O_3 concentrations~~ efficiently estimate key statistical indicators of $PM_{2.5}$ and O_3 concentrations under future emission scenarios and capture potential extreme pollution events.

2. Prediction of probability distribution of PM$_{2.5}$ and O$_3$ (Section 2.3, L401)

Furthermore, the TGEOS model shows a high level of similarity to the GC model in predicting pollutant distribution and extreme events, making it a valuable tool for ~~online~~ rapid assessments of related emission reduction policies to enhance decision-making efficiency.

3. Conclusion (L435)

TGEOS has successfully established the complex relationship between precursor emissions and concentrations of PM$_{2.5}$ and O$_3$ pollutants, which can be used for rapid ~~online~~ assessment of the effects of different emission control schemes.

RC: ***Lines 12: The interpretation of correlation coefficients (0.97 for PM$_{2.5}$, 0.96 for O$_3$) is unclear. Specify whether these values represent spatial correlation between models, probability distribution accuracy, or overall model performance metrics.***

AR: We thank the reviewer for pointing out the ambiguity. The reported correlation coefficients (0.97 for PM$_{2.5}$ and 0.96 for O$_3$) in Section 3.3 refer to the model performance of averaged TGEOS predictions across all grid cells within each key pollution region. Specifically, for each region, we computed the Pearson correlation coefficient between predicted and simulated values at each grid cell, and then averaged these values. The 0.97 (PM$_{2.5}$) and 0.96 (O$_3$) values correspond to the lower limit of calculated regional averages across the key regions. To avoid conclusion, we have revised the definition of these correlation coefficients to more understandable metrics representing the overall model performance for national winter PM$_{2.5}$ and summer O$_3$ in the test set. We have revised the corresponding part of the manuscript, with details shown in blew.

Abstract (L11-L13)

The spatial and probability distributions predicted by TGEOS are in good agreement with GEOS-Chem, with the general correlation coefficients for PM$_{2.5}$ and O$_3$ exceed ~~0.97 and 0.96, respectively.~~ 0.98 in high-pollution months.

RC: ***Line 56: The 350-hour computational benchmark lacks critical context. Specify whether this duration includes: (a) Standalone nested-domain simulation (0.5°×0.625° over China); (b) Coupled global (2°×2.5°) + nested simulations (0.5°×0.625° over China); (c) Hardware specifications (CPU/GPU model, and software)***

AR: We thank the reviewer for pointing out the lack of computational context. The reported 350-hour computational benchmark refers to a standalone nested-domain GEOS-Chem simulation over China at 0.5° × 0.625° resolution. This duration was measured from the 2017 baseline scenario simulation and is representative for all nested-domain simulations in our study. All simulations used identical configurations: GEOS-Chem version 14.2.2, compiled with OpenMP parallelization (OMP_NUM_THREADS = 32), executed on a Linux cluster node equipped with two Intel(R) Xeon(R) E5-2620 v4 CPUs (8 cores per socket, 32 logical processors total) and 62 GB RAM. The manuscript Line 56 was revised as:

Typically, for GEOS-Chem version 14.2.2, ~~on a computational cluster utilizing 32 cores~~ on a computational cluster mentioned in Section 2.1.2 , a ~~1-year~~ one-year standalone full-chem nested simulation of China

> at a resolution of 0.5 × 0.625 requires approximately 350 hours, and this duration is expected to increase when conducting simulations at finer resolutions or over extended time periods.

And the specifications of computation device in this study has been listed in Section 2.1.2 of the manuscript:

> **2.1.2. GEOS-Chem configuration (L190-L191)**
>
> For anthropogenic emissions out of China, we used data from the Community Emissions Data System (CEDS) inventory (Hoesly et al., 2018). All the GC simulations in this research relied on identical configurations: GEOS-Chem version 14.2.2, compiled with OpenMP parallelization (OMP_NUM_THREADS = 32), executed on a Linux cluster node equipped with two Intel(R) Xeon(R) E5-2620 v4 CPUs (8 cores per socket, 32 logical processors total) and 62 GB RAM.

RC: ***Line 102: The term "middle-scale region" requires quantitative definition.***

AR: We thank the reviewer for pointing out the ambiguity. By "middle-scale region," we refer to mega-urban agglomerations like the North China Plain (NCP) and Yangtze River Delta (YRD) regions in China, typically covering multiple cities and spanning several hundred kilometers across, and suffering from severe air pollution. We have avoided such vague expressions in the revised manuscript

RC: ***Table 1: The relationship between DPEC-SSP (socioeconomic pathways) and DPEC-CA (policy scenarios) remains unexplained. Justify scenario combinations' scientific relevance to climate modeling objectives. Expand all acronyms (e.g., SSP1-5, SSP1-26-BHE control, early_peak-net_zero-clean_air control) in table/footnotes.***

AR: We thank the reviewer for this comment. On one hand, the DPEC-SSP scenarios (SSP1–SSP5) are derived from the Shared Socioeconomic Pathways (SSP) framework and represent long-term global socioeconomic trajectories (e.g., low-emission sustainable development under SSP1, fossil-fuel intensive pathway under SSP5). Within DPEC, five CMIP6 climate scenarios (i.e., SSP1-26, SSP2-45, SSP3-70, SSP4-60, SSP5-85) and three groups of pollution control scenarios (i.e., Business-As-Usual, BAU; Enhanced-control-policy, ECP; Best-Health-Effect, BHE) are considered to represent five future emission scenarios in China under different socioeconomic and technological assumptions (Tong et al., 2020). On the other hand, the DPEC-CA scenarios are based on SSP1 assumption without considering additional climate or air pollution control policies, and constructed to reflect policy-driven "clean air" emission control pathways in China like carbon peak policies and carbon neutrality target (Cheng et al., 2023). For example, the on-time peak-clean air (mentioned as "otp" in the manuscript) scenarios will implement the Best-Health-Effect (BHE) local pollution control by 2060 but further deploy carbon reduction measures in 2020-2030, to achieve the carbon peak around 2030. Compared with DPEC-SSP, DPEC-CA scenarios represent more strict emission control policies and can provide more "low-value" samples for the whole training set. By combining DPEC-SSP and DPEC-CA, the model will consider both global socioeconomic influences and China-specific policy interventions to the emission levels, which strengthen the predictive ability of TGEOS for future air quality assessments.

In addition, we would like to clarify that the objective of this study is air quality modeling rather than climate modeling. While climate scenarios can influence long-term air quality, the present work does not explicitly incorporate climate variability or emission–climate feedbacks. Instead, we focus on future emission-driven changes in pollutant concentrations. The DPEC-SSP and DPEC-CA scenarios were selected because they provide a scientifically consistent and policy-relevant representation of China's future emissions at high spatial resolution, including both baseline socioeconomic pathways and alternative policy controls. These scenarios are widely recognized as a reliable basis for projecting China's future emission trajectories (Cheng et al., 2021). By using them, our study can efficiently and systematically explore the potential air quality outcomes under different emission futures. Therefore, the integration of DPEC-SSP and DPEC-CA in this work is motivated not by climate modeling objectives, but by the need to ensure credible emission inputs for rapid air quality assessments.

Acronym expansion (to be added in table footnotes): SSP1-5: five CMIP6 climate scenarios under Shared Socioeconomic Pathways (SSPs).

SSP1-26-BHE control: scenario combined with SSP1-26 climate scenario and Best-Health-Effect (BHE) pollution control scenario.

SSP2-45-ECP control: scenario combined with SSP2-45 climate scenario and Enhanced-control-policy (ECP) pollution control scenario.

SSP3-70-BAU control: scenario combined with SSP3-70 climate scenario and Business-As-Usual (BAU) pollution control scenario.

SSP4-60-BAU control: scenario combined with SSP4-60 climate scenario and Business-As-Usual (BAU) pollution control scenario.

SSP5-85-BHE control: scenario combined with SSP5-85 climate scenario and Best-Health-Effect (BHE) pollution control scenario.

clean_air: scenarios shares the same socio-economic development and energy transitions as the SSP1, but the optimal end-of-pipe pollution control will be implemented during 2020–2060 to explore the contribution and potential of stricter clean air actions to future air quality improvement.

on-time_peak_clear-air: scenarios driven by SSP1 socio-economic development, and will deploy various carbon reduction measures in 2020–2030 to achieve carbon peak around 2030, but there are no additional climate targets after 2030 to reinforce the low-carbon transition; configurations on end-of-pipe pollution control is consistent with the clean air scenario.

early_peak-net_zero-clean_air: scenarios driven by SSP1 with intensified carbon reduction measures in 2020–2030 to boost an earlier carbon peak around 2025, and after 2030, more ambitious low-carbon transition will be intensified to achieve carbon neutrality by 2060 (Cheng et al., 2023).

The Detailed explanation of clean_air, on-time_peak_clear-air, and early_peak-net_zero-clean_air scenarios are discussed in Cheng et al. (2023).

We have revised the corresponding part of the manuscript, with details shown in blew.

2.1.1 Multi-scenario inventory (L154-L166)

As a prerequisite to simulate future air quality, we produced a multi-scenario emission inventory of 36 emission scenarios, including 24 future emission scenarios, 11 fine-tuned scenarios and 1 background scenario. Detailed information on the inventory is shown in Table 1. We first used 24 future emission scenarios based on the DPEC (Dynamic Projection model for Emissions in China) platform (`http://meicmodel.org.cn`) to initially construct the data set. As a dynamic model developed by Tsinghua University (Tong et al., 2020), DPEC can reflect the dynamic changes of China's future emissions under various socioeconomic and policy control scenarios, and provide detailed gridded emission data, including emissions with different control scenarios, emission sectors and spatial coordinate information. ~~Specifically, we designed two scenario sets of~~

> ~~DPEC-SSP and DPEC-CA to represent emission scenarios under different social-economic scenarios and different emission reduction policies, respectively. DPEC-SSP was selected from DPECv1.0 (Tong et al., 2020) and consists of five sub-scenario sets (SSP1-5). Data of 2030, 2040 and 2050 were selected in each sub-scenario, and each of them was treated as an independent emission scenario for our multi-scenario inventory. DPEC-CA was selected from DPECv1.2 (Cheng et al., 2023) and was composed of three sub-scenario sets including "clean air", "on-time peak-clean air", and "early peak-net zero-clean air". Similarly, we selected three years of data for each sub-scenario set as independent emission scenarios.~~ The DPEC-provided scenarios are widely recognized as a reliable basis for projecting China's future emission trajectories (Cheng et al., 2021). Firstly, we constructed a scenario set named "DPEC-SSP" to represent emission scenarios under different socio-economic scenarios and different emission control policies. DPEC-SSP was selected from DPECv1.0 (Tong et al., 2020) and consists of five sub-scenario sets (SSP1 to SSP5) as combinations of different climate scenarios (i.e., SSP1-26, SSP2-45, SSP3-70, SSP4-60, SSP5-85) and pollution control scenarios (i.e., Business-As-Usual, Enhanced-control-policy, Best-Health-Effect). Sectoral emissions of 2030, 2040 and 2050 were selected in each sub-scenario, and each of them was treated as an independent emission scenario for the multi-scenario inventory. Furthermore, to capture policy-driven "clean air" pathways such as carbon peaking and carbon neutrality targets in China, another scenario set, DPEC-CA, was developed based on DPECv1.2 Cheng et al., 2023) dataset. The DPEC-CA was composed of three sub-scenario sets including "clean air", "on-time peak-clean air", and "early peak-net zero-clean air". Each of these scenario was constructed on SSP1 assumption, without introducing additional climate or air pollution control policies, to reflect different short-term carbon emission reduction policies in China. Compared with DPEC-SSP, DPEC-CA scenarios represent more stringent emission control policies and can provide more "low-value" samples to enrich training set. Similar to DPEC-SSP, three years (2030, 2040, and 2050) were selected from each DPEC-CA sub-scenario, with each year treated as an independent emission scenario.

RC: ***Lines 167-172: Explain how MEIC-2017-based perturbations enhance generalizability for 2030-2050 predictions. Address potential biases from applying contemporary (2017) emission factors to distant future scenarios (13-33 year gap).***

AR: We thank the reviewer for this question. All future emission scenarios in our study are constructed based on the 2017 MEIC inventory, with each grid cell treated independently. Introducing perturbations relative to MEIC-2017 enhances model generalizability by providing multiple emission levels per grid cell in the training data, thereby expanding coverage of the input space and reducing the risk of extrapolation to unseen values, especially for those predictions under high emission scenarios. We acknowledge that this approach may introduce biases for distant future years, as spatial emission patterns may change over time. However, the primary aim of this study is to quantify the concentration response to changes in emission magnitudes rather than to provide precise forecasts of absolute future air quality.

RC: ***Text S1 (Line 16): Correct "Fig ??" with the appropriate figure identifier. Verify all figure citations for accuracy.***

AR: We thank the reviewer for pointing this out. The placeholder "Fig ??" in Text S1 (Line 16) will be replaced with the correct identifier. We will also carefully review all figure citations throughout the manuscript and supplementary material to ensure their accuracy.

> For instance, in the tun0.6 scenario ~~depicted in Fig ??~~ mentioned in Table 1 , coefficient matrix for each emission variable was created from multiple sampling of the MND with 0.6 as the mean.

RC: ***Lines 173-179: Clarify why DPEC-SSP/CA emissions were scaled using 2017 MEIC ratios rather than used directly. Define the "five sectors" referenced. Justify setting coefficient maximums to 2.0. Quantify the percentage of coefficients exceeding this threshold and discuss sensitivity to alternative values (e.g., 1.5 or 2.5). Specify what constitutes the "original inventory."***

AR: We appreciate the reviewer's request for clarification and provide the following details:

(1) Rationale for not directly using DPEC-SSP/CA emissions: The DPEC-SSP/CA scenario inventories are provided in units of tons per grid cell, whereas the "original inventory" referred to in our manuscript is expressed in kg/m²/s. To reconcile these formats, we obtained DPEC and MEIC inventories (both in tons per grid cell) from their respective official sources and computed a grid-wise ratio matrix. This matrix captures future emission trends for each sector and species, and was then applied to the MEIC emissions in rate form (kg/m²/s) used for GEOS-Chem input.

(2) Definition of "five sectors": These correspond to power, industry, residential, transportation, and agriculture.

(3) Justification for a maximum coefficient of 2.0: Inspection of the raw ratio matrices showed that over 80% of grid cells had coefficients below 2.0, with most values in key emission regions concentrated between 0.3 and 1.5. Coefficients exceeding 2.0 occurred almost exclusively in very low-emission grid cells (typically < 0.1 ton per grid), where differences between DPEC and MEIC inventories in absolute terms are small but amplified in ratio form. Thus, slightly adjusting the coefficient threshold has negligible influence on the resulting GEOS-Chem simulations. Although some previous studies have adopted lower thresholds (e.g., 1.2, (Xing et al., 2011)), other research suggests that future emissions in certain heavily polluted regions could plausibly exceed 1.5 Brean et al. (2023), making 2.0 a more flexible yet still reasonable limit.

We revised Section 2.1 to include these clarifications, with details below:

2.1.1 Multi-scenario inventory (L173-L179)

~~For each emission scenario, we divided it by the 2017 MEIC inventory to obtain a series of monthly coefficient matrices for the emissions of various species in five sectors. It is worth noting that since the units of DEPC and MEIC data are tons per grid, significant variation are exhibited between adjacent grids. Thus, we set the maximum value of each coefficient matrix to 2.0 when making the DPEC scenarios to avoid abnormal emission coefficients due to magnitude differences. Subsequently, we multiplied the generated coefficient matrices to the corresponding part of the original inventory used for GEOS-Chem input to obtain the multi-scenario inventory that reflect the control of each scenario. The new inventory was employed in GEOS-Chem simulating to obtain PM$_{2.5}$ and O3 concentrations under future emission scenarios.~~ Since the unit of DPEC-SSP/CA emissions is t/grid, which is incompatible for GEOS-Chem running, we used MEIC inventory with t/grid as unit at 2017 as a benchmark (denoted as b-MEIC), and make elementwise divisions between DPEC and b-MEIC to obtain a series of monthly emission coefficient matrices for various species in five sectors, namely power, industry, residential, transportation, and agriculture. Since the majority of grids with emission factor smaller than 2.0 (>80%) and to prevent abnormal values due to magnitude difference of two inventories, the threshold of emission factors was artificially set to 2.0. Subsequently, we took the Schur product of the coefficient matrices and corresponding part of MEIC inventory used for GEOS-Chem input, with unit of kg/m²/s, to generate emission inventories projected with DPEC-SSP/CA.

RC: ***Line 184-185: Report the spin-up time for the global GEOS-Chem simulations providing boundary conditions.***

AR: We appreciate the reviewer's comment. For the global GEOS-Chem simulations (2° × 2.5°) that provided boundary conditions to the nested-domain runs, we used a spin-up period of 6 months to minimize the influence of initial conditions.

RC: ***Lines 185-187: Elaborate how MERRA-2 meteorology was integrated into GEOS-Chem. Confirm whether meteorology was prescribed identically across all 36 emission scenarios. Discuss the limitation of using static 2017 meteorology for future scenarios, as it ignores potential meteorology-emission feedbacks and climate variability, particularly for ozone sensitivity.***

AR: We thank the reviewer for the comment. The MERRA-2 meteorological fields were integrated into GEOS-Chem through the standard HEMCO interface, providing assimilated meteorology at 0.5° × 0.625° (nested domain) and 2° × 2.5° (global) resolution, with a 3-hour temporal resolution. All 36 emission scenarios in this study were simulated with identical 2017 MERRA-2 meteorology. As mentioned in major comment 2, the methodology of varying emission under fixed meteorological condition is commonly applied in exploring future air quality. This design was intentional to isolate the impact of emissions on concentrations by removing meteorological variability, allowing for a more controlled assessment of emission–concentration relationships. We acknowledge that using fixed-year meteorology does not capture climate variability or emission–meteorology feedbacks, such as the influence of temperature and radiation changes on ozone formation, or aerosol–radiation interactions under future scenarios. We have revised the manuscript to explicitly state this limitation in the part of Conclusion Section and note that incorporating variable meteorology in future work would enable the evaluation of combined emission and climate drivers on air quality.

4 Conclusions (L457-L462)

~~Additionally, due to the considerable effect of meteorological conditions on the generation (Shi et al., 2020), spatiotemporal patterns (Zhang et al., 2013; Chen et al., 2020), and concentration levels (Wang et al., 2019) of PM2.5 and O3 concentrations, and meteorological conditions other than 2017 are not considered in this study. Consequently, there is also a need to incorporate various climate scenarios that represent meteorological variations to enhance the TGEOS's predictive capability regarding future air quality under more complex scenarios with variations in emissions and meteorology.~~ Additionally, in order to isolate the impact of emission changes to future air quality as previous studies did (Shi et al., 2020; Wang et al., 2023), the meteorology for GC simulations of all scenarios was fixed at 2017. The identified meteorology limits TGEOS's ability to generate reliable estimations in cross-meteorology scenarios, and to capture meteorology–emission interactions and "emission–climate" feedbacks. Consequently, there is also a need to incorporate various climate scenarios that represent meteorological variations to enhance the TGEOS's predictive capability regarding future air quality under more complex scenarios with variations in emissions and meteorology.

RC: ***Lines 190-191: Justify the use of different emission inventories for China (MEIC) and other regions (CEDS). Address potential inconsistencies in source sectors, speciation, or spatial/temporal resolution between inventories.***

AR: We appreciate the reviewer's comment. As this study focuses on the China region, we used the MEIC inventory because it provides more accurate and precise emissions for China, with sectoral and species definitions consistent with those in the DPEC inventory (Li et al., 2017; Tong et al., 2020). This consistency facilitates the construction of multiple emission scenarios for our model. The CEDS inventory was used only for the global 2° × 2.5° simulations to generate boundary conditions for the nested-domain runs. These global simulations were performed with fixed emissions and served as a common set of boundary fields for all 36 emission scenarios.

RC: ***Lines 192-203: Describe GEOS-Chem inputs/outputs (emissions, meteorology, concentrations) in Section 2.1, and TGEOS training inputs/outputs in Section 2.2. Specify the spatiotemporal resolution of all TGEOS input features (emissions, meteorology) and output targets. Clarify if training is grid-cell-based. If so, justify why only the 8 nearest neighbors are sufficient to represent regional transport. List the "8 key meteorological parameters" explicitly. State that "dust components were excluded" means $PM_{2.5}$ predictions exclude dust aerosols—highlight that this deviates from standard $PM_{2.5}$ definitions and significantly impacts regions like Northwest China.***

AR: We thank the reviewer for the detailed suggestions. We will revise Sections 2.1 and 2.2 to clarify the following: (1) GEOS-Chem inputs/outputs: As we mentioned in Section 2.1.2, the nested-domain GEOS-Chem simulations (0.5° × 0.625° over China) used anthropogenic emissions from multi-scenario inventory (0.25° × 0.25° over China) as substitutes for MEIC in China, along with identical biogenic emissions from MEGAN and other natural sources. MERRA-2 meteorological fields in 2017 provided the meteorology inputs, along with identical boundary condition and restart files as the concentration inputs. The model outputs included hourly concentrations of O_3 and daily concentrations of $PM_{2.5}$ components.

(2) TGEOS training inputs/outputs: As we described in Table 2, the TGEOS model was trained on a grid-cell basis, with each sample containing local and surrounding (8-neighbor) emissions for 105 sectoral anthropogenic sources, 9 key meteorological parameters, spatial coordinates (latitude, longitude), as well as the simulating months. The emissions and meteorology were interpolated into 0.5° × 0.625° for model input. Output targets were 12 monthly statistical indicators (mean, min, max, median, 25 and 75 percentiles) for $PM_{2.5}$ and O_3 at the same grid cell with spatial resolution of 0.5° × 0.625°. The temporal resolution of all input features was monthly, with concentration data derived from hourly/daily GEOS-Chem outputs and meteorology data stemmed from MERRA-2 reanalysis after interpolation.

(3) Spatiotemporal resolution: We adopted the 8 nearest neighbors to efficiently represent regional transport influence while keeping feature dimensionality tractable. On one hand, we focused on short-range regional transport effects in this study. Given the relatively coarse spatial resolution of 0.5° × 0.625°, the 8 surrounding grid cells already cover a substantial geographic area of 200–250 km^2, within which emissions typically exert the most significant influence on local pollutant concentrations (Liu et al., 2019). On the other hand, taking more grids into account could bring about redundancy features and thus affect model performance.

(4) There are actually 9 meteorological parameters mentioned in this article, namely 2-meter air temperature (T2M), 10-meter northward wind (V10M), 10-meter eastward wind (U10M), planetary boundary layer height (PBLH), 2-meter specific humidity (QV2M), total precipitation (PRECTOT), relative humidity (RH), evaporation from turbulence (EVAP), and surface pressure (PS). We have corrected the text in Line 195 to match Table 2, now referring to “9 meteorological parameters” for consistency and accuracy.

(4) By “dust components were excluded,” we mean that $PM_{2.5}$ predictions in this study exclude dust aerosol species (DST1–DST4 in GEOS-Chem). We acknowledge that this differs from the standard $PM_{2.5}$ definition and may lead to underestimation in dust-influenced regions such as Northwest China. However, our research focused on variations of anthropogenic emissions to air pollutants concentrations, large predictive bias may arise in the northern China such as NCP and FWP when considering the impact of dust.

We have revised the corresponding part of the manuscript, with details shown in blew.

2.1.2 GEOS-Chem configuration

The GEOS-Chem ~~chemistry~~ chemical transport model (http://www.geos-chem.org, version 14.2.2) was used to simulate the spatiotemporal distribution of surface $PM_{2.5}$ and O_3 concentrations under different emission scenarios based on year 2017. The nested model was configured with a horizontal resolution 0.5° latitude by 0.625° longitude covering China (from 17.5 to 54°N and 72 to 136°E) and 47 vertical layers. Boundary condition files for model startup were offered by 1-year global GC simulation with a horizontal resolution of 2° latitude by 2.5° longitude. ~~Assimilated meteorological data from the NASA Global Modeling and Assimilation Office's Modern-Era Retrospective analysis for Research and Applications Version 2 (MERRA-2) (Gelaro et al., 2017) were selected as meteorology fields entering the model.~~ Assimilated meteorological data from the NASA Global Modeling and Assimilation Office's Modern-Era Retrospective analysis for Research and Applications Version 2 (MERRA-2) (Gelaro et al., 2017) were selected as meteorology fields entering the model through HEMCO interface, with the 3-hour temporal resolution and 0.5° × 0.625° spatial resolution. Consistent with previous studies on future air quality assessment (Shi et al., 2021; Shi et al., 2022; Shi et al., 2023), the meteorological inputs were fixed at 2017 for all simulations in this study to isolate concentration changes attributable solely to emission variations. The Multi-resolution Emission Inventory for China (MEIC, http://meicmodel.org/) (Li et al., 2017) and the multi-scenario emission inventories with a horizontal resolution of 0.25° latitude by 0.25° (as detailed in 2.1.1) were used as the monthly anthropogenic emissions to simulate $PM_{2.5}$ and O_3 concentrations under various emission scenarios. For anthropogenic emissions out of China, we used data from the Community Emissions Data System (CEDS) inventory (Hoesly et al., 2018). All the GC simulations in this research relied on identical configurations: GEOS-Chem version 14.2.2, compiled with OpenMP parallelization (OMP_NUM_THREADS = 32), executed on a Linux cluster node equipped with two Intel(R) Xeon(R) E5-2620 v4 CPUs (8 cores per socket, 32 logical processors total) and 62 GB RAM.

2.1.3 Multi-scenario dataset

~~As demonstrated in Table 2, Sectoral emission data for 26 precursors across all grid cells in each scenario were utilized as training features. To account for pollutant transport from adjacent areas, emission data from the 8 grid cells surrounding the target grid were also incorporated as training features. Concurrently, 8 key meteorological parameters, previously demonstrated to exhibit significant correlations with $PM_{2.5}$ and O_3 concentrations (Shi et al., 2020; Zhang et al., 2022a), for both the target and neighboring grids derived from MERRA-2 data from 2017 were included as training features. Furthermore, local and peripheral spatial location information was incorporated as a training feature to enable the model to capture spatial patterns in pollutant emissions and concentrations. Finally, twelve statistical indicators, including the 25 quantile, 75 quantile, median, average, maximum, and minimum values that derived from the daily averaged concentrations of $PM_{2.5}$ and O_3 of a month, were utilized as training targets to represent the probability distribution of pollutant concentrations. It should be noted that this study mainly concentrates on the prediction of air quality in different scenarios of anthropogenic emissions, so the dust components were excluded during subsequent data processing.~~ Before model training and evaluation, we constructed a multi-scenario dataset by combining emissions with the corresponding GC simulations, as summarized in Table 1. Each sample in this dataset is defined on a grid-cell basis. All emission data were interpolated to a spatial resolution of 0.5° × 0.625° to match the GC simulations. Detailed descriptions of the input features and output targets are provided in Table 2. For each grid cell, 105 sectoral emissions were selected as predictors to represent local emission conditions. At this resolution, a 3 × 3 grid domain already covers a geographic extent of approximately 200–250 km^2, within which emissions typically exert the most pronounced influence on local pollutant concentrations (Liu et al., 2019). To account for regional transport of precursors, we also incorporated the emissions of the eight neighboring cells. In addition, nine key meteorological variables, previously identified as strongly correlated with $PM_{2.5}$ and O_3

concentrations (Shi et al., 2020; Zhang et al., 2022a), were included for each local and neighboring grids, based on the 2017 MERRA-2 reanalysis processed into monthly averages. Spatial information of both the local and adjacent grids was further incorporated to enable the model to capture spatial heterogeneity in emissions and pollutant concentrations. The training targets consisted of twelve statistical indicators on a monthly scale, including the 25th and 75th percentiles, median, mean, maximum, and minimum, derived from the daily averaged concentrations of $PM_{2.5}$ and O_3 in the GC outputs for each scenario. It is important to emphasize that our analysis focuses on air quality responses to anthropogenic emission changes. Consequently, dust-related components were excluded during preprocessing, since dust intrusions can introduce large predictive biases in northern and western China, where they make substantial contributions to $PM_{2.5}$ concentrations (Pang et al., 2023).

RC: ***Lines 231-232: The claim that "pollutants generally conform to either a standard normal or skewed distribution" lacks validation.***

AR: We thank the reviewer for pointing this out. In this study, our assessment of distribution type was based on the comparison between the mean and median of the pollutant concentrations. For O_3, the mean and median values were nearly identical, indicating symmetry consistent with a normal distribution. For $PM_{2.5}$, the mean was consistently greater than the median, indicating positive skewness consistent with a right-skewed Gamma distribution. These findings are in line with previous studies such as Zhang et al. (2018); Zeng et al. (2021), which also reported normal-like distributions for O_3 and skewed Gamma-like distributions for $PM_{2.5}$ in similar contexts. We will revise the manuscript to explicitly describe this assessment method and reference the supporting literature.

~~We selected six indicators for each pollutant as our prediction targets since the distributions of these two pollutants generally conform to either a standard normal or skewed distribution. The probability distribution curve would be quantified with these 6 indicators in the following test.~~ Previous research has indicated that $PM_{2.5}$ and O_3 concentrations tend to follow characteristic statistical patterns Zhang et al. (2018); Zhang et al. (with $PM_{2.5}$ generally displaying a right-skewed Gamma-like distribution and O_3 approximating a normal distribution. This distinction is also evident from the comparison of their mean and median values. Based on this insight, we used the TGEOS-predicted statistical indicators to approximate regional probability distribution curves. Specifically, the mean, 25th, and 75th percentiles were applied to capture the overall shape of the distributions, while the minimum and maximum values were incorporated to constrain their ranges.

RC: ***Lines 239-242: Justify: (1) Why SSP1/SSP5 were excluded despite representing critical low/high-emission pathways; (2) Whether "low/high" refers to base year or future projections.***

AR: We thank the reviewer for this comment. In this study, SSP1, SSP4, and SSP5 were included in the training set to enrich the range of low and high emission samples available for model learning. These scenarios were therefore not used for testing, unlike SSP2 and SSP3, which served as independent test scenarios to evaluate model generalization. The terms “low” and “high” emission levels are defined relative to the 2017 baseline year. SSP2 generally represents a lower-emission trajectory, while SSP3 represents a higher-emission trajectory compared to 2017 (Tong et al., 2020).

RC: ***Line 267-268: Define the quantitative metric used to identify otp2030 as the scenario "most similar" to SSP2_2050.***

AR: We thank the reviewer for this observation. In the manuscript, the scenario "most similar" to SSP2_2050, namely otp2030, is identified based on a quantitative similarity metric. Specifically, we calculate the Euclidean distance between mean values of $PM_{2.5}$ and O_3 in SSP2_2050 scenario and those in each training scenario. The otp2030 scenario exhibits the minimum distance to SSP2_2050, indicating the highest similarity in the selected features. We will revise the manuscript to explicitly state the quantitative metric.

RC: ***Line 305-315: Explain why a single fixed initial condition (from the "background scenario") was used for all GEOS-Chem simulations despite varying emissions. This likely introduces errors, especially when simulated concentrations diverge significantly from initial states. Clarify why SSP3 concentrations align better with this initial state than SSP2.***

AR: We thank the reviewer for this comment. We acknowledge that employing fixed initial conditions may introduce biases, particularly during the early stages of each simulation. Nevertheless, such effects are primarily relevant for short-term or global-scale applications. These discrepancies are typically dampened through multiple physical and chemical adjustment processes, enabling the model to converge toward a state consistent with the prescribed emissions and meteorology. Previous studies have shown that the influence of initial condition discrepancies is largely dissipated within approximately 10 days of simulation spin-up in regional-scale simulations (Appel et al., 2012; Demir, 2022). Therefore, the impact of initialization errors on the monthly statistical indicators considered in this study is expected to be acceptable.

Furthermore, the objective of this study is not to reproduce short-term dynamical variability but to provide rapid predictions of concentration distributions under future emission scenarios. The use of a single fixed initial condition ensures consistency across all 36 emission scenarios, such that differences in simulated pollutant concentrations can be attributed solely to emission perturbations rather than heterogeneous initial states. Within this framework, adopting fixed initial conditions represents a scientifically defensible simplification that balances computational feasibility with robustness of the comparative analysis.

Regarding the comment that pollutant concentrations under the SSP3 scenario aligns more closely with the initial model conditions compared to the SSP2 scenario. This is due to the fact that the SSP3 scenario (SSP3-70-BAU) employs a Business-As-Usual (BAU) emission control strategy, which implies a consistently increasing and high-emission trajectory. The initial conditions of the model are derived from the background scenario with the simulation year of 2017. Consequently, the projected emission levels in the SSP3 scenario exhibit a relatively small deviation from the pollutant concentrations in 2017 (Tong et al., 2020). In contrast, the SSP2 scenario (SSP2-45-ECP) implements the Enhanced Control Policy (ECP), which enforces stringent emission reductions (Tong et al., 2020), thereby leading to notable discrepancies between the simulated pollutant concentrations and the initial conditions during the early simulation period.

The error graphs of PM2.5 indicators for SSP2$_40andSSP3_40areshowninFig.3a3tod3andFig.S2a3tod3.Wefoundthemode$

RC: ***Lines 319-320: Supplement Figures with equivalent spatial maps of the original GEOS-Chem simulated seasonal indicators for direct comparison with TGEOS predictions.***

AR: We appreciate the reviewer's suggestion. Due to space limitation, we did not demonstrate these GEOS-Chem simulated indicators for direct comparison with TGEOS predictions. Subsequently, we have supplemented the current figures with corresponding spatial maps of these indicators simulated by GEOS-Chem, enabling a direct side-by-side comparison with TGEOS predictions.

RC: ***Line 342-343: Detail the probability distribution fitting procedure: Specify the distribution type fitted to the regional data. Clarify the data used: Is the PDF based on daily concentrations across all grid cells within a region over a month? List exactly which of the 12 indicators were used as distribution parameters.***

AR: We thank the reviewer for requesting clarification. In our analysis: For $PM_{2.5}$, we fitted a right-skewed

Gamma distribution; for O_3, we fitted a normal distribution. For each region, the probability density function (PDF) was fitted using the monthly-scale indicators averaged over all grid cells in the region. These monthly indicators were computed from daily mean concentrations at each grid cell and served as prediction targets for the model. The fitting procedure primarily used the 25th percentile, 75th percentile, and mean as parameters to characterize the distribution shape, with the maximum and minimum values used to constrain the distribution boundaries.

3.3 Prediction of probability distribution of $PM_{2.5}$ and O_3 (L342-L343)

~~For each region, we calculated the average of twelve statistical indicators across all grid points and subsequently utilized these averaged indicators to fit the probability distribution curves for $PM_{2.5}$ and O_3.~~ For each region, the probability density function (PDF) curves were fitted using the TGEOS-predicted monthly indicators averaged over all grid cells in the region. For $PM_{2.5}$, we fitted a right-skewed gamma distribution; for O_3, we fitted a normal distribution. The fitting procedure primarily used the 25th percentile, 75th percentile, and mean as parameters to characterize the distribution shape, with the maximum and minimum values used to constrain the distribution boundaries. These probability distribution curves derived from monthly statistical indicators can be used to preliminarily assess the overall distribution of pollutant concentrations for a given month or quarter under various future emission scenarios.

RC: ***Lines 344-345: Define the geographical boundaries for NCP, YRD, FWP, and SCB regions.***

AR: We thank the reviewer for the suggestion. In our study, the four key regions are defined by rectangular lat–lon boundaries, as follows:

- **NCP (North China Plain):** 34–42° N, 113–120° E;
- **YRD (Yangtze River Delta):** 26–34° N, 115–123° E;
- **FWP (FenWei Plain):** 33–38° N, 103–114° E;
- **SCB (Sichuan Basin):** 26–34° N, 103–107° E.

We have revised the corresponding part of the manuscript, with details shown in blew.

3.3 Prediction of probability distribution of $PM_{2.5}$ and O_3 (L340-L342)

In this study, we focus on the probability distributions predicted by TGEOS for four key polluted areas: ~~the North China Plain (NCP), Yangtze River Delta (YRD), Fenwei Plain (FWP), and Sichuan Basin (SCB).~~ the North China Plain (NCP, 34–42° N, 113–120° E), Yangtze River Delta (YRD, 26–34° N, 115–123° E), Fenwei Plain (FWP, 33–38° N, 103–114° E), and Sichuan Basin (SCB, 26–34° N, 103–107° E).

RC: ***Lines 364-367: The statement "O_3 concentrations are relatively less influenced by emissions" due to meteorology dominance is misleading. Precursor emissions (NO_x, VOCs) critically influence O_3 formation. The core limitation is the use of identical 2017 meteorology for all scenarios, preventing assessment of emission impacts under varying meteorology.***

AR: We thank the reviewer for this observation. We agree that O_3 concentrations are strongly influenced by precursor emissions like NO_x and VOCs, and that our original wording may have overstated the dominance of meteorology. Our intent was to highlight that, in this study, the use of identical 2017 MERRA-2 meteorology for all scenarios constrains the variability in meteorological drivers, making the relative differences between scenarios primarily emission-driven. We revised the sentence to more accurately reflect this relationship. Detail are blew:

> ~~Compared to the notable variations observed in $PM_{2.5}$ levels, O_3 concentrations are relatively less influenced by emissions. This is largely due to the fact that ozone levels are predominantly determined by meteorological conditions, particularly air temperature, while all scenarios in this study are modeled using meteorological data from 2017, signiffcant fluctuations are not expected.~~ Since the meteorological conditions for all scenarios are fixed in 2017, these concentration variations can be attributed to changes in emissions.

RC: ***Line 369: The notation "a2 to d2" lacks corresponding figure identification.***

AR: We thank the reviewer for this observation. The notation "a2 to d2" refers to subpanels in Figure S17 and S20. We have revised the text to explicitly link these notations to their corresponding figure, and check the entire manuscript to ensure all subpanel references are clearly identified.

RC: ***Line 370: Define "high-emission samples". Quantify how many scenarios/samples represent high emissions within the training dataset.***

AR: We define high-emission samples as those whose total anthropogenic emissions of precursors exceed the 95th percentile of the corresponding emission distribution across the entire training dataset. In the training dataset, this criterion corresponds to 61888 high-emission samples approximately 4.6% of all training samples.

RC: ***Line 372-373: The attribution of O_3 underestimation in SSP2_2050 to high precursor emissions (ALK4, ALK5, TOLU) appears inconsistent. If elevated emissions cause this systemic bias, why is a similar or stronger underestimation not observed under the even higher emissions of SSP3_2050 (Fig. S11)?***

AR: We thank the review for this comment and acknowledge that our previous explanation was misleading. To address this, we carefully examined the emission distributions in the SSP2_2050 scenario compared with the training set. As shown in Fig. 9, we found that several precursor emissions, such as CO residential and NO transportation, exhibited substantially higher density in the low-emission range (left tail of the distribution) under SSP2_2050 (orange line) than in the training set (blue line). This indicates that the model has limited training experience in this regime, which is likely a primary cause of the systematic underestimation. Consistently, our residual analysis showed that the mean residuals of six residential emissions of CO, BC, $PM_{2.5}$, PM_{10} and SO_2 were significantly below zero in the low-emission regime (Fig. 10), in line with their density distributions (Fig. 9). Therefore, we infer that the anomalously low values of these emissions contributed to the underestimation of O_3 concentrations in the SSP2_2050 scenario. Although no studies have demonstrated a direct causal link between these emissions and O_3 concentration, but from a modeling perspective, this phenomenon can be attributed to distributional shifts between training and test data rather than to the physical or chemical effects of these species. Specifically, the SSP2_2050 scenario contains a much larger proportion of samples in the low-emission regime, where training samples are sparse. In this regime, the model is forced to extrapolate beyond its well-constrained domain and tends to regress toward the mean patterns learned from the entire training set. As a result, even for features only weakly related to O_3, the distribution mismatch acts as an indicator of domain shift and leads to systematic deviations in model predictions. We have revised the corresponding part of the manuscript, with details shown in blew.

Additionally, we observed that the model performs slightly poorly in predicting the probability distribution of pollutants under certain high emission scenarios (a2 to d2). As discussed in section 3.3, this discrepancy arises from the limited number of high-emission samples in the dataset, which undermines the model's generalization capabilities. It is also important to emphasize that when predicting O_3 levels under the SSP2_2050 scenario, ~~the model demonstrates a systemic underestimation, as shown in Fig. S11 (a1 to d1). A detailed analysis of the emission data for this scenario indicates that several key precursors for O_3 generation, including ALK4, ALK5, and TOLU, exhibit relatively high emission levels in the SSP2_2050 scenario. The scarcity of high-emission samples for these species prevents the model from adequately recognizing the importance of high-emission features to the target variable, leading to a tendency for the model to underestimate predictions.~~ It is also important to emphasize that when predicting O_3 levels under the SSP2_2050 scenario, TGEOS shows a clear underestimation in the YRD region (Fig. S20 d1). To investigate this, we compared the emission distributions of SSP2_2050 and the training set. As shown in Fig. S9, several precursor emissions (e.g., CO residential, NO transportation) exhibit much higher densities in the low-emission range under SSP2_2050 (orange line) than in the training set (blue line), where the model has limited training experience. Residual analysis further confirms that the mean residuals of multiple residential emissions (CO, BC, $PM_{2.5}$, PM_{10}, SO_2) are significantly below zero in this regime (Fig. S10), consistent with their density distributions. We therefore attribute the underestimation of O_3 to distributional shifts between training and test data rather than to the direct physical or chemical effects of these species. In particular, the SSP2_2050 scenario contains a substantially larger fraction of samples in the low-emission regime, forcing the model to extrapolate beyond its well-constrained domain and regress toward mean patterns learned from the training set, thereby inducing these prediction biases.

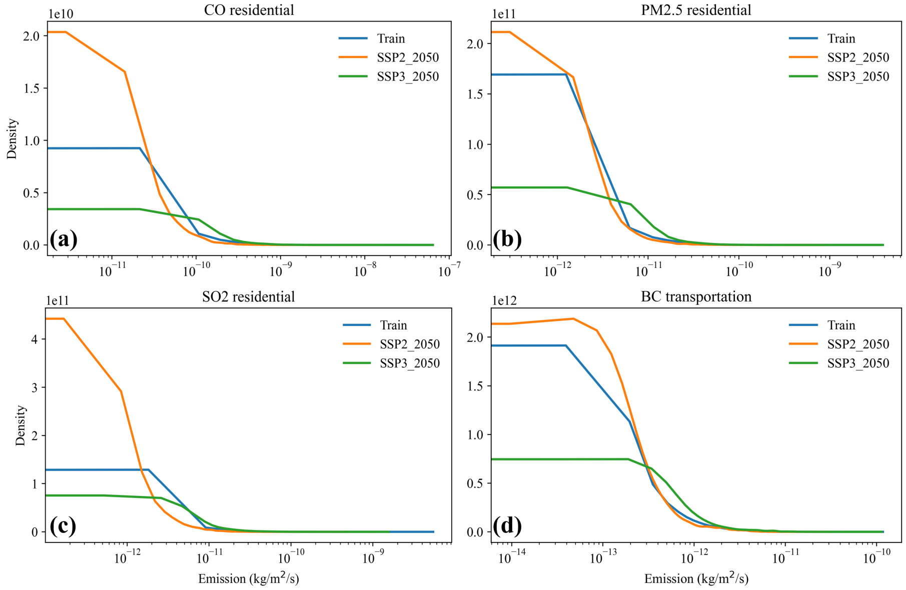

Figure 9: Kernel Density Estimation (KDE) curves for residential emissions of CO (a), $PM_{2.5}$ (b), SO_2 (c), and BC (d) emissions in SSP2_2050 scenario (orange line) and training set (blue line) on semi-logarithmic scales. For convenience, only four emission variables were displayed.

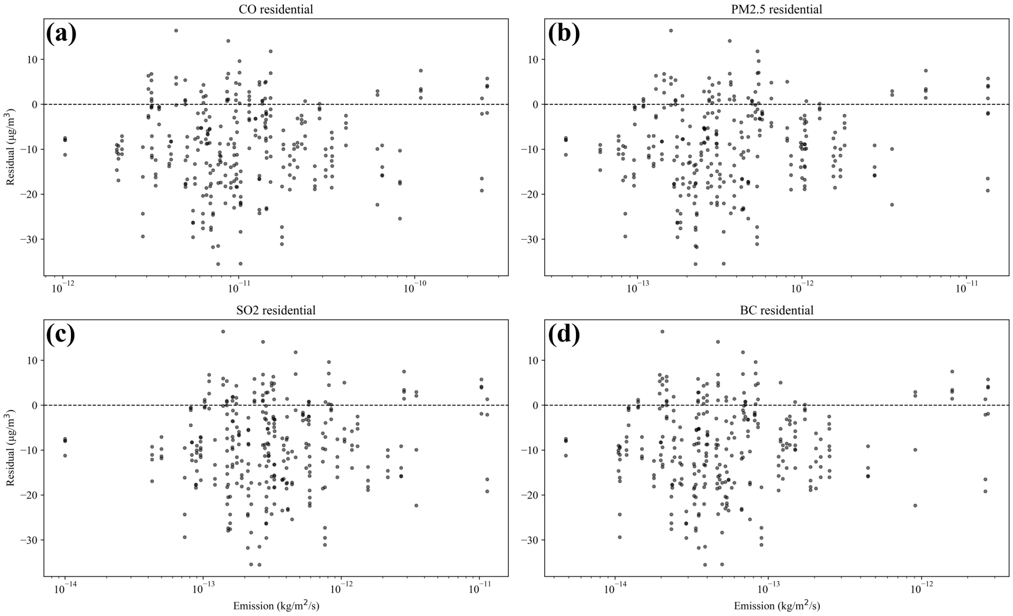

Figure 10: Scatterplot of residual distribution for CO (a), $PM_{2.5}$ (b), SO_2 (c), and BC (d) emissions in SSP2_2050 scenario.

RC: ***Lines 384-385: Justify comparing extreme event probability changes to the "base scenario" (presumably 2017). Since meteorology is identical (2017), changes are solely emission-driven—state this explicitly to clarify the comparison's purpose.***

AR: We thank the reviewer for the comment. The "base scenario" refers to the background scenario with 2017 simulated with emissions and MERRA-2 meteorology in 2017. Since all scenarios in this study were simulated under identical meteorological conditions, the differences in extreme event probabilities are solely attributable to changes in emissions. The purpose of this comparison is therefore to isolate the impact of emission changes on the probability of extreme pollution events, without the confounding influence of meteorological variability.

> Since the future scenarios were made based on 2017 background scenario and identical meteorology used in the dataset, the concentration changes are solely emission-driven. Our findings indicate that under low-emission scenarios, the incidence of extreme $PM_{2.5}$ events decreased most significantly in the SCB and YRD regions, as illustrated in Fig. 9 b1 and c1.

RC: ***Line 385: The notation "(b1) and (c1)" lacks corresponding figure identification.***

AR: We thank the reviewer for this observation. The notation "(b1)" and "(c1)" refers to subpanels in Figure 9 in the manuscript. We have revised the text to explicitly link these notations to their corresponding figure, and check the entire manuscript to ensure all subpanel references are clearly identified. We have revised the corresponding part of the manuscript.

Our findings indicate that under low-emission scenarios, the incidence of extreme $PM_{2.5}$ events decreased most significantly in the SCB and YRD regions, as illustrated in ~~(b1) and (c1)~~ Fig. 9 b1 and c1 .

RC: ***Line 437: Provide hardware specifications for the 2.51-second/year prediction benchmark (e.g., CPU/GPU model, and software).***

AR: We thank the reviewer for the comment. The 2.51-second/year prediction benchmark refers to the inference time of the trained TGEOS model to one-year test scenario, obtained on a GPU-equipped server. Specifically, the benchmark was measured on an NVIDIA GeForce RTX 4080 with 31 GB memory, using PyTorch 2.3.1 with CUDA 12.4 and Python 3.11.5, running under Ubuntu 20.04.6. We have added these hardware and software specifications to the revised manuscript.

Model training and evaluation (section 2.2.2)

In this study, four machine learning models were employed independently to evaluate the performance for each kind of model structure. Except for the TGEOS model discussed in this paper, three ML models, namely RF, MLP and CNN, which had demonstrated good performance in air quality modeling (Huang et al., 2021; Fang et al., 2023), were simultaneously employed based on the same training strategy. Training and evaluation of four models were conducted on a GPU-equipped server. Specifically, the benchmark was measured on an NVIDIA GeForce RTX 4080 with 31 GB memory, using PyTorch 2.3.1 with CUDA 12.4 and Python 3.11.5, running under Ubuntu 20.04.6.

RC: ***Ensure consistent formatting throughout: Use subscripts for chemical species (e.g., O_3, $PM_{2.5}$) and superscripts for statistical terms (e.g., R^2). Thoroughly check all text, figures, and tables.***

AR: We thank the reviewer for this observation. We will thoroughly check the entire manuscript, including the main text, figures, and tables, to ensure consistent formatting. Specifically, chemical species will be presented with subscripts (e.g., $PM_{2.5}$ and O_3) and statistical terms with superscripts (e.g., R^2). All inconsistencies will be corrected in the revised version.

References

Appel, K. W., Chemel, C., Roselle, S. J., Francis, X. V., Hu, R.-M., Sokhi, R. S., Rao, S., and Galmarini, S.: Examination of the Community Multiscale Air Quality (CMAQ) model performance over the North American and European domains, Atmospheric environment, 53, 142–155, 2012.

Bhattarai, H., Tai, A. P., Martin, M. V., and Yung, D. H.: Responses of fine particulate matter (PM2. 5) air quality to future climate, land use, and emission changes: Insights from modeling across shared socioeconomic pathways, Science of the total Environment, 948, 174 611, 2024.

Brean, J., Rowell, A., Beddows, D. C., Shi, Z., and Harrison, R. M.: Estimates of future new particle formation under different emission scenarios in Beijing, Environmental Science & Technology, 57, 4741–4750, 2023.

Cheng, J., Tong, D., Liu, Y., Yu, S., Yan, L., Zheng, B., Geng, G., He, K., and Zhang, Q.: Comparison of current and future PM2. 5 air quality in China under CMIP6 and DPEC emission scenarios, Geophysical Research Letters, 48, e2021GL093 197, 2021.

Cheng, J., Tong, D., Liu, Y., Geng, G., Davis, S. J., He, K., and Zhang, Q.: A synergistic approach to air pollution control and carbon neutrality in China can avoid millions of premature deaths annually by 2060, One Earth, 6, 978–989, 2023.

Demir, S.: Comparison of normality tests in terms of sample sizes under different skewness and Kurtosis coefficients, International Journal of Assessment Tools in Education, 9, 397–409, 2022.

Fang, L., Jin, J., Segers, A., Liao, H., Li, K., Xu, B., Han, W., Pang, M., and Lin, H. X.: A gridded air quality forecast through fusing site-available machine learning predictions from RFSML v1. 0 and chemical transport model results from GEOS-Chem v13. 1.0 using the ensemble Kalman filter, Geoscientific Model Development, 16, 4867–4882, 2023.

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., et al.: The modern-era retrospective analysis for research and applications, version 2 (MERRA-2), Journal of climate, 30, 5419–5454, 2017.

He, H., Liang, X.-Z., and Wuebbles, D. J.: Effects of emissions change, climate change and long-range transport on regional modeling of future US particulate matter pollution and speciation, Atmospheric Environment, 179, 166–176, 2018.

Hu, W., Zhao, Y., Lu, N., Wang, X., Zheng, B., Henze, D. K., Zhang, L., Fu, T.-M., and Zhai, S.: Changing responses of PM2. 5 and ozone to source emissions in the Yangtze River Delta using the adjoint model, Environmental Science & Technology, 58, 628–638, 2023.

Huang, L., Liu, S., Yang, Z., Xing, J., Zhang, J., Bian, J., Li, S., Sahu, S. K., Wang, S., and Liu, T.-Y.: Exploring deep learning for air pollutant emission estimation, Geoscientific Model Development Discussions, 2021, 1–22, 2021.

Lai, A., Lee, M., Carter, E., Chan, Q., Elliott, P., Ezzati, M., Kelly, F., Yan, L., Wu, Y., Yang, X., et al.: Chemical investigation of household solid fuel use and outdoor air pollution contributions to personal PM2. 5 exposures, Environmental Science & Technology, 55, 15 969–15 979, 2021.

Li, M., Liu, H., Geng, G., Hong, C., Liu, F., Song, Y., Tong, D., Zheng, B., Cui, H., Man, H., et al.: Anthropogenic emission inventories in China: a review, National Science Review, 4, 834–866, 2017.

Li, Z., Liu, F., Yang, W., Peng, S., and Zhou, J.: A survey of convolutional neural networks: analysis, applications, and prospects, IEEE transactions on neural networks and learning systems, 33, 6999–7019, 2021.

Liu, Y., Zheng, M., Yu, M., Cai, X., Du, H., Li, J., Zhou, T., Yan, C., Wang, X., Shi, Z., et al.: High-time-resolution source apportionment of PM< sub> 2.5</sub> in Beijing with multiple models, Atmospheric Chemistry and Physics, 19, 6595–6609, 2019.

Liu, Z., Dong, M., Xue, W., Ni, X., Qi, Z., Shao, J., Guo, Y., Ma, M., Zhang, Q., and Wang, J.: Interaction patterns between climate action and air cleaning in China: a two-way evaluation based on an ensemble learning approach, Environmental Science & Technology, 56, 9291–9301, 2022.

Lu, X., Zhang, L., Chen, Y., Zhou, M., Zheng, B., Li, K., Liu, Y., Lin, J., Fu, T.-M., and Zhang, Q.: Exploring 2016–2017 surface ozone pollution over China: source contributions and meteorological influences, Atmospheric Chemistry and Physics, 19, 8339–8361, 2019.

Pang, M., Jin, J., Segers, A., Jiang, H., Fang, L., Lin, H. X., and Liao, H.: Dust storm forecasting through coupling LOTOS-EUROS with localized ensemble Kalman filter, Atmospheric Environment, 306, 119 831, 2023.

Pinder, R. W., Adams, P. J., and Pandis, S. N.: Ammonia emission controls as a cost-effective strategy for reducing atmospheric particulate matter in the eastern United States, 2007.

Reich, B., Cooley, D., Foley, K., Napelenok, S., and Shaby, B.: Extreme value analysis for evaluating ozone control strategies, The annals of applied statistics, 7, 739, 2012.

Shi, J., Deng, L., Du, W., Han, X., Zhong, Y., Rao, W., Xie, H., Xiang, F., Ning, P., and Tian, S.: Chemical characteristics, sources, and formation mechanisms of PM2. 5 before, during, and after the Spring Festival in a plateau city of Southwest China, Atmospheric Environment, 338, 120 788, 2024.

Shi, X., Zheng, Y., Lei, Y., Xue, W., Yan, G., Liu, X., Cai, B., Tong, D., and Wang, J.: Air quality benefits of achieving carbon neutrality in China, Science of the Total Environment, 795, 148 784, 2021.

Skyllakou, K., Rivera, P. G., Dinkelacker, B., Karnezi, E., Kioutsioukis, I., Hernandez, C., Adams, P. J., and Pandis, S. N.: Changes in PM 2.5 concentrations and their sources in the US from 1990 to 2010, Atmospheric Chemistry and Physics, 21, 17 115–17 132, 2021.

Stankevičius, L. and Lukoševičius, M.: Extracting sentence embeddings from pretrained transformer models, Applied Sciences, 14, 8887, 2024.

Taylor, K. E.: Taylor diagram primer, Work. Pap, pp. 1–4, 2005.

Tong, D., Cheng, J., Liu, Y., Yu, S., Yan, L., Hong, C., Qin, Y., Zhao, H., Zheng, Y., Geng, G., et al.: Dynamic projection of anthropogenic emissions in China: methodology and 2015–2050 emission pathways under a range of socio-economic, climate policy, and pollution control scenarios, Atmospheric Chemistry and Physics, 20, 5729–5757, 2020.

Wang, S., Wu, D., Wang, X.-M., Fung, J. C.-H., and Yu, J. Z.: Relative contributions of secondary organic aerosol formation from toluene, xylenes, isoprene, and monoterpenes in Hong Kong and Guangzhou in the Pearl River Delta, China: an emission-based box modeling study, Journal of Geophysical Research: Atmospheres, 118, 507–519, 2013.

Wang, Y., Liao, H., Chen, H., and Chen, L.: Future projection of mortality from exposure to PM2. 5 and O3 under the carbon neutral pathway: roles of changing emissions and population aging, Geophysical Research Letters, 50, e2023GL104 838, 2023.

Xiao, Q., Zheng, Y., Geng, G., Chen, C., Huang, X., Che, H., Zhang, X., He, K., and Zhang, Q.: Separating emission and meteorological contribution to PM 2.5 trends over East China during 2000–2018, Atmospheric Chemistry and Physics Discussions, 2021, 1–32, 2021.

Xing, J., Wang, S., Jang, C., Zhu, Y., and Hao, J.: Nonlinear response of ozone to precursor emission changes in China: a modeling study using response surface methodology, Atmospheric Chemistry and Physics, 11, 5027–5044, 2011.

Xing, J., Zheng, S., Ding, D., Kelly, J. T., Wang, S., Li, S., Qin, T., Ma, M., Dong, Z., Jang, C., et al.: Deep learning for prediction of the air quality response to emission changes, Environmental science & technology, 54, 8589–8600, 2020.

Zeng, L., Yang, Y., Wang, H., Wang, J., Li, J., Ren, L., Li, H., Zhou, Y., Wang, P., and Liao, H.: Intensified modulation of winter aerosol pollution in China by El Niño with short duration, Atmospheric Chemistry and Physics, 21, 10 745–10 761, 2021.

Zhang, F., Li, X., Wang, X., Tan, M., and Xin, L.: CMIP6-driven 10 km super-resolution daily climate projections with PET estimates in China, Scientific Data, 12, 720, 2025.

Zhang, J., Gao, Y., Luo, K., Leung, L. R., Zhang, Y., Wang, K., and Fan, J.: Impacts of compound extreme weather events on ozone in the present and future, Atmospheric Chemistry and Physics, 18, 9861–9877, 2018.

Zhang, X., Xiao, X., Wang, F., Brasseur, G., Chen, S., Wang, J., and Gao, M.: Observed sensitivities of PM2. 5 and O3 extremes to meteorological conditions in China and implications for the future, Environment International, 168, 107 428, 2022.