Point-to-Point Responses to Reviewer's Comments

We would like to thank for reviewer's thoughtful comments on our manuscript.

The authors describe a new, neural network-based reduced-order model of atmospheric chemistry and transport (FastCTM) which has been trained using an extensive dataset of output from CMAQ. FastCTM uses a novel and interesting approach, building physics-informed networks for five separate operators. The authors show that FastCTM is able to reproduce the general patterns of concentrations calculated by CMAQ for an out-of-training-data year (2023), and that the sensitivities of FastCTM to key meteorological variables or nation-wide changes in emissions mostly follow expected patterns. If FastCTM can be shown to be reliable in policy-relevant contexts then it could be a very useful tool.

This approach to modelling is interesting, and this methodological advance has the potential to significantly accelerate air quality scenario analysis. A CTM which can respect key physical constraints (e.g. mass conservation) while also accurately reproducing the effect of different perturbations to emissions and meteorological fields would have great value. However, the manuscript as written does not quite live up to this promise. Along with some minor concerns, the key challenge is that the authors do not show evidence that this new model can fulfil the roles of a CTM and produce accurate results for one of the most common use cases (i.e. understanding the effects of different perturbations). I explain this concern in more detail below, and until this concern is addressed I do not believe the manuscript should be accepted for publication by GMD.

Major comments

The most significant concern relates to the validation/evaluation of the model. The authors appear to have trained the five physical operators based on several years of output from the CMAQ chemistry transport model. While I have some questions regarding the training process, I will take it as read for the moment that the training was done in such a way as to avoid overfitting. However, the verification of the model rests on its ability to predict, from the 2018-2022 data, the performance in 2023. This approach is inadequate for two reasons. First, the authors do not compare the performance of the model to simpler approaches with the same data such as generalized additive models, gradient boosting, or linear regression with land use (see e.g. Wong et al., 2021 and Cheng et al., 2021). Without such a comparison to evaluate how such models would have performed in predicting 2023, it is difficult to say what the magnitude of FastCTM's advance is. This is exacerbated by the relatively shallow quantitative assessment in section 3.1.1. RMSE and R2 values are provided, but it is not clear how these were calculated; given that these are calculated as a function of time, are these calculated based on the difference in each of the 158,742 grid cells between CMAQ and FastCTM? A deeper analysis which investigates how model performance varies between (e.g.) rural and urban areas, coastal and inland areas, winter and summer, and so on would provide a much more robust test of the model's ability to predict the effect of changing meteorology. This could be informed by (e.g.) taking the difference of FastCTM for 2023 against CMAQ for 2023, and comparing that to the difference between CMAQ for 2023 and the average of CMAQ from 2018 to 2022. This would at least demonstrate whether FastCTM provides more explanatory power for the mean atmospheric state than taking the average concentration from the previous five years.

Response: Thanks. We agree with the review's point that the current analysis does not adequately demonstrate that FastCTM has actual capabilities to simulate air quality changes by learning and representing physical and chemical processes. As the reviewer suggested, more tests, comparisons, and analyses are performed.

(1) For comparing with simpler machine learning models, we tested three models of Linear Regression, Random Forest, and XGBoost with the same train and test dataset. The results are added in the manuscript as follows,

To validate FastCTM model, three land use regression (LUR) models were constructed, namely the linear regression model, the random forest model (with the number of trees set at 500), and the XGBoost model (with the booster specified as gbtree). These LUR models were developed using the same input meteorological data, emission, and geophysical variables. When compared with the FastCTM model, the performance of the LUR models was found to be significantly inferior (as demonstrated in Figure S10 – S12 in the SI). This outcome is, in fact, anticipated when we consider the complex nature of air quality dynamics. Air quality is not a static entity, but it varies both spatially and temporally. For instance, the transport of air pollution is a highly dynamic process that hinges on wind fields and air pollution concentrations in a reciprocal manner. The wind direction and speed dictate the trajectory along which pollutants travel, while the existing pollutant concentrations in different regions influence the overall dispersion and mixing patterns. LUR models, which predominantly rely on local input data (Wong et al., 2021; Cheng et al., 2021), struggle to capture these intricate, non-local interactions. They lack the capacity to account for the far-reaching effects such as wind-driven pollutant transport and the consequential changes in air quality over larger geographical areas.

The supplementary Figures demonstrate the performances of three machine learning models are displayed in the following part.

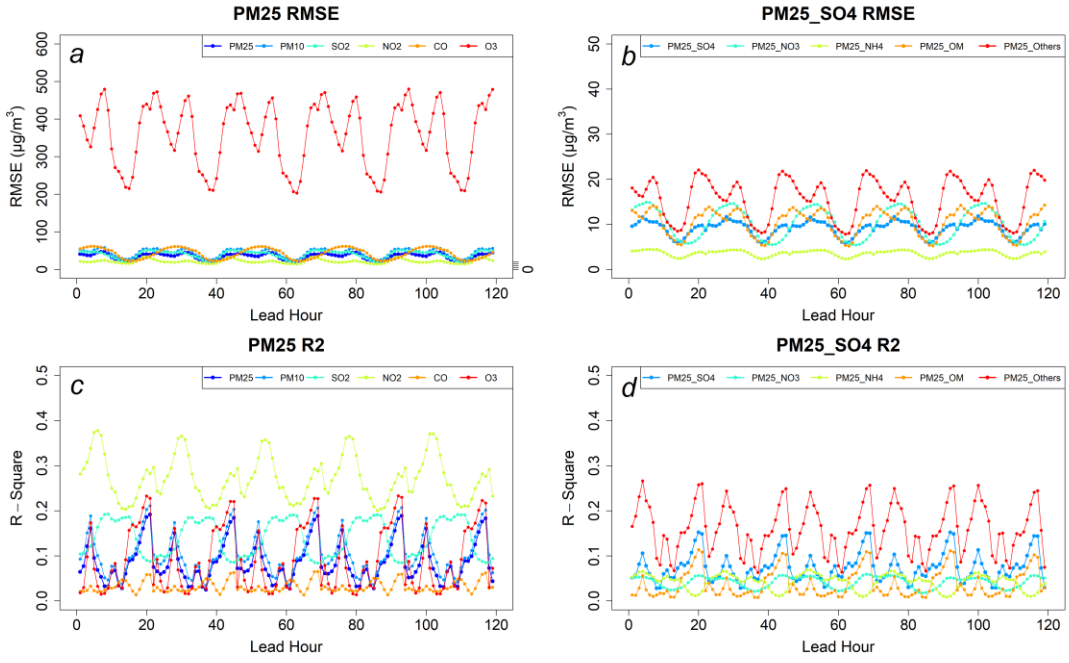

Figure S10: The evaluation performances of linear regression forecasts against CMAQ forecasts in 2023. Panel (a) and (b) respectively show RMSE values of criteria pollutants and the $PM_{2.5}$ components. Panel (c) and (d) respectively show R^2 values. It should be noted that the RMSE value of CO corresponds to the right axis in panel (a).

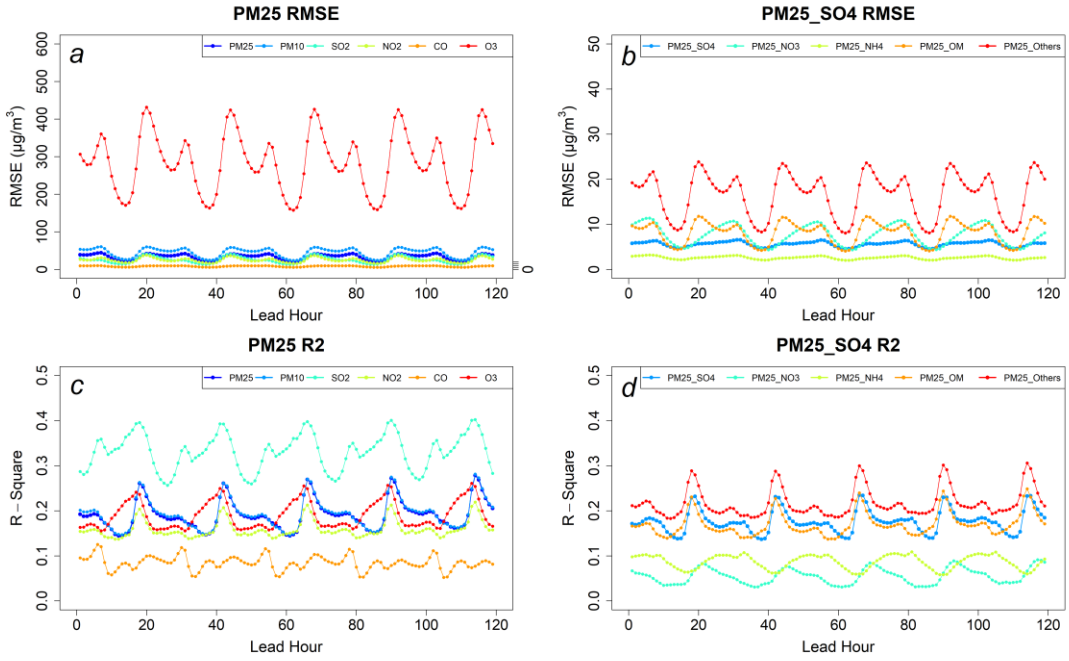

Figure S11: The evaluation performances of random forest forecasts against CMAQ forecasts in 2023. Panel (a) and (b) respectively show RMSE values of criteria pollutants and the $PM_{2.5}$ components. Panel (c) and (d) respectively show R^2 values. It should be noted that the RMSE value of CO corresponds to the right axis in panel (a).

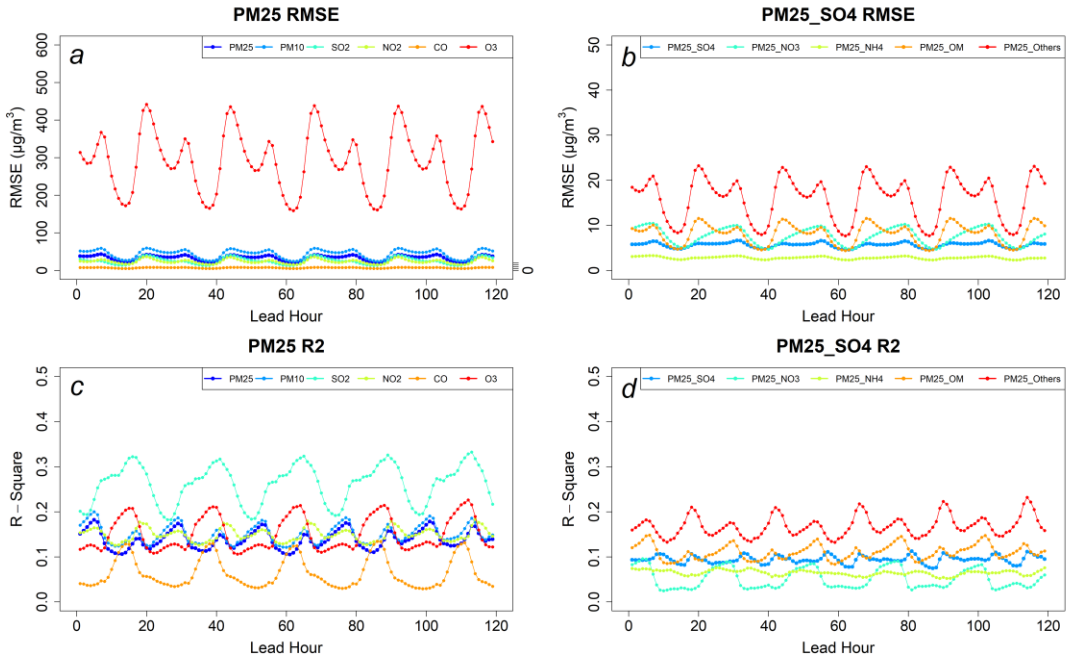

Figure S12: The evaluation performances of XGboost forecasts against CMAQ forecasts in 2023. Panel (a) and (b) respectively show RMSE values of criteria pollutants and the $PM_{2.5}$ components. Panel (c) and (d) respectively show R^2 values. It should be noted that the RMSE value of CO corresponds to the right axis in panel (a).

(2) As for the calculation process for the metrics of RMSE and R2, they are elaborated in section 2.5 Model Evaluation, as follows.

The metrics of root mean square error (RMSE) and coefficient of determination (R^2) were calculated daily in each of 119 leading hours on the difference in each of the 158,742 grid cells between CMAQ and FastCTM. Therefore, 119 static values for each metric of R^2 and RMSE were obtained on each day of the independent test year of 2023. The statistical values on each day are then averaged for the same leading hour for comparison.

(3) As the reviewer suggested, evaluations of FastCTM compared to CMAQ in rural/urban, inland/coastal areas, and cold/warm seasons are further performed. Generally, they have similar performances in comparative areas or seasons. However, FastCTM exhibited lower correlations in urban areas and coastal areas. In urban areas, emission sources and chemical processes are more complex than that in rural areas, making it harder for FastCTM to simulate due to its 2D setting and fewer chemicals considered than CMAQ. It is also true for FastCTM's performance in coastal areas, where meteorological conditions are more varied in time. Related discussions and results are added in section 3.1.1. and in the supplementary material as follows.

Defining the warm season as the months from April to September and the winter and cold season as the remaining months, the FastCTM model exhibited comparable performances. As shown in Figure 4 (with detailed information in Figure S7 in the SI), the coefficient of determination R^2 values for the six criteria pollutants were 0.82, 0.8, 0.8, 0.82, 0.91, and 0.7 in the warm season, and 0.8, 0.79, 0.78, 0.83, 0.88, and 0.68 in the cold season, respectively. To assess the performance variations of FastCTM across different spatial locations, comparative evaluations were carried out in urban and rural areas as well as in inland and coastal regions. Generally, FastCTM demonstrated slightly higher accuracies in rural areas compared to urban areas (as presented in Figure S8 in the SI). This outcome is reasonable given the more intricate emission and chemical processes prevalent in urban settings (Guo et al., 2014). Similarly, FastCTM exhibited comparable performances in inland areas to those in coastal areas, with the exception of $PM_{2.5}$ and PM_{10} (Figure S9 in the SI).

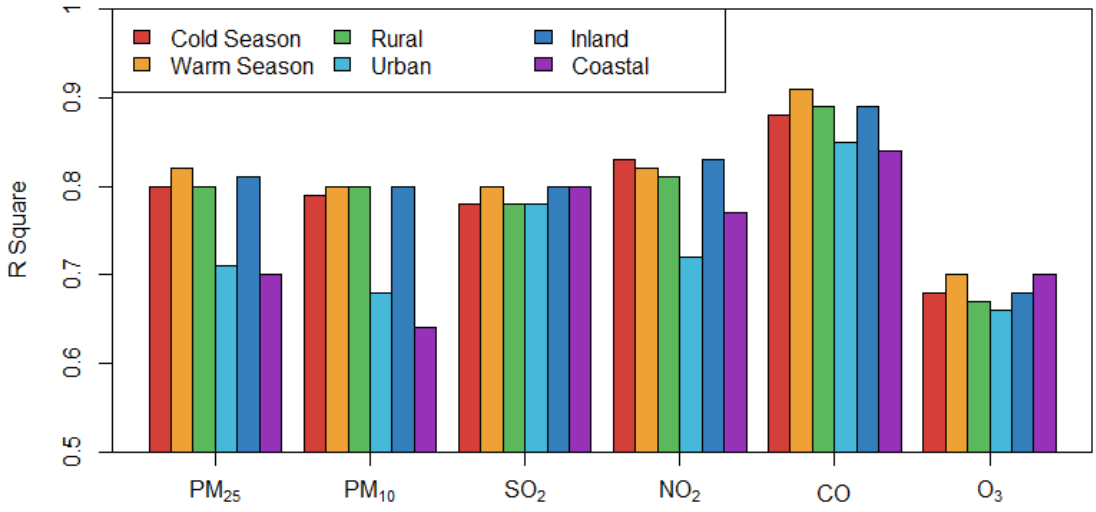

Figure 1: The mean evaluation R^2 values for all 119 leading hours of FastCTM forecasts in warm/cold seasons, rural/urban areas and coastal/inland areas.

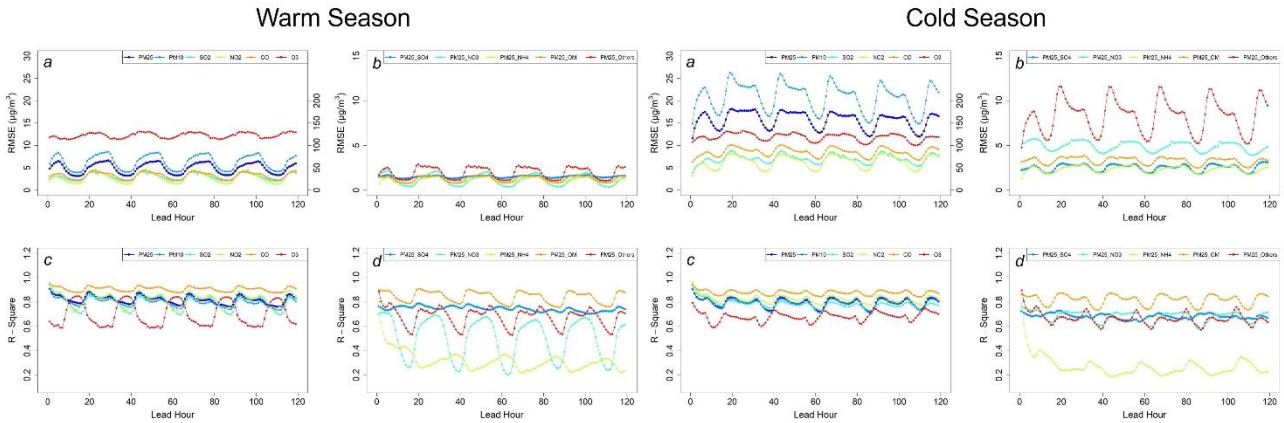

Figure S2: The evaluation performances of random forest forecasts against CMAQ forecasts in warm season of 2023. Panel (a) and (b) respectively show RMSE values of criteria pollutants and the $PM_{2.5}$ components of. Panel (c) and (d) respectively show R^2 values. It should be noted that RMSE value of CO corresponds to the right axis in panel (a).

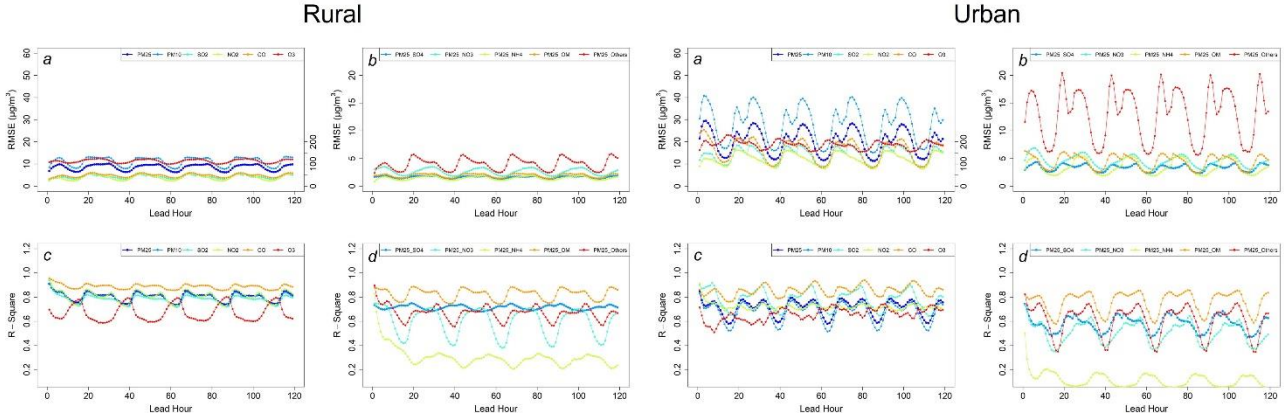

Figure S3: The evaluation performances of FastCTM forecasts against CMAQ forecasts in rural and urban areas in 2023.. Panel (a) and (b) respectively show RMSE values of criteria pollutants and the $PM_{2.5}$ components of. Panel (c) and (d) respectively show R^2 values. It should be noted that RMSE value of CO corresponds to the right axis in panel (a).

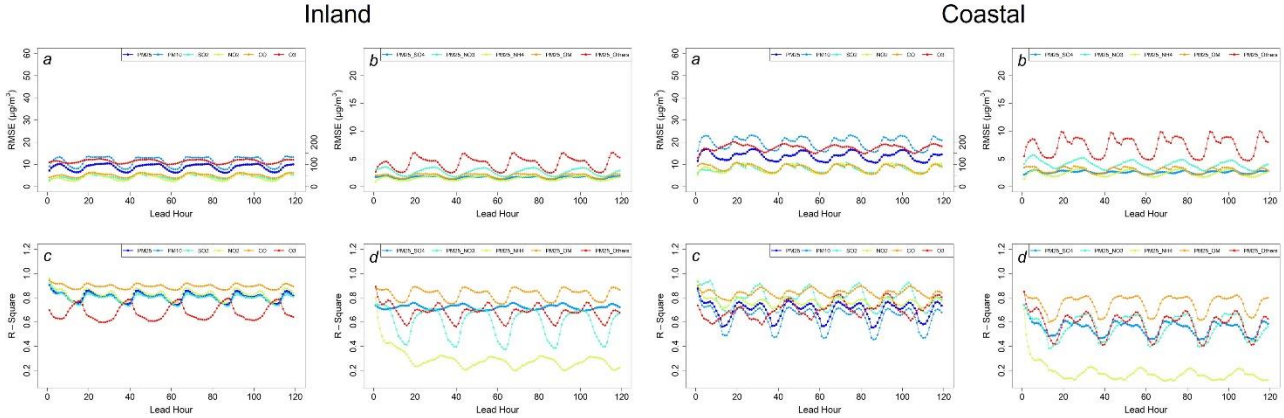

Figure S4: The evaluation performances of FastCTM forecasts against CMAQ forecasts in inland and coastal areas in 2023. Panel (a) and (b) respectively show RMSE values of criteria pollutants and the $PM_{2.5}$ components of. Panel (c) and (d) respectively show R^2 values. It should be noted that RMSE value of CO corresponds to the right axis in panel (a).

(4) As the reviewer has kindly pointed out, FastCTM may have taken average pollutant concentration from five-year training data in 2018-2022. In order to verify if FastCTM was able to predict air quality based on given meteorological conditions and emissions, daily average FastCTM simulation in the fifth leading day (leading hours 96-119) in the test year of 2023 is compared with daily average CMAQ simulations in 2023 and in the training years of 2018-2022. Results revealed that FastCTM forecasts are generally in good correlation with CMAQ forecasts in 2023, rather than that in 2018-2022. It means FastCTM has learned the evolution rules of air pollutant concentrations, instead of just giving average air pollutant concentration according to time of the year. Related results have been added in the manuscript in section 3.1.1, as follows.

Annually, the daily air quality typically exhibits similar fluctuations to those in other years, which can be primarily attributed to the cyclical nature of meteorological conditions and pollutant emission patterns. The FastCTM model was trained using a comprehensive dataset spanning five years, from 2018 to 2022. It was crucial to rule out the possibility that the model was merely reproducing historical averages during the test year of 2023. The daily national average concentrations of $PM_{2.5}$ and O_3 in 2023, as predicted by FastCTM, were similarly compared with those simulated by CMAQ in the same test year, as well as with the CMAQ forecasts from the training years of 2018-2022. As illustrated in Figure 4, it becomes evident that the predictions made by FastCTM in 2023 align more closely with the actual CMAQ forecasts for that year, rather than with the forecasts generated from the training data of 2018-2022. This finding not only validates the adaptive learning capabilities of the FastCTM model but also indicates that the model is not resorting to a simplistic approach of taking the average concentration from the previous five years based on the time of day. Instead, it is likely to incorporate real-time meteorological feedback, adjusting for shifts in emission patterns, and leveraging its learned relationships to provide more accurate and contemporaneous predictions.

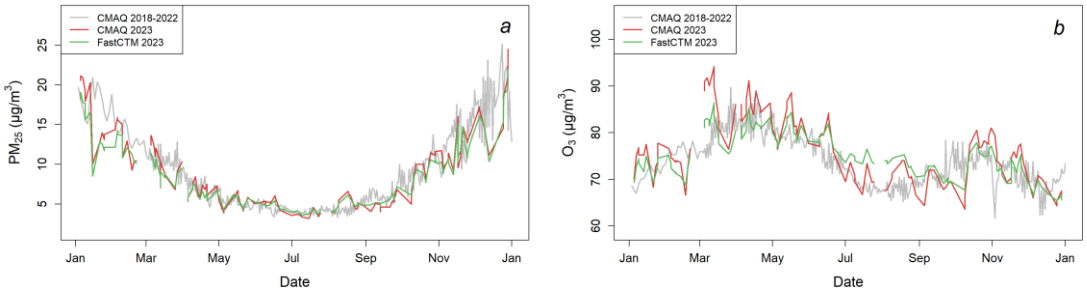

Figure 5: The timeseries of FastCTM forecasts against CMAQ forecasts..

Second, and perhaps more importantly, the function of FastCTM is to reproduce the results of high-fidelity CTMs at a fraction of the computational cost – specifically to support air pollution simulations, sensitivity analysis, and internal process analysis (abstract lines 30-32). The comparison to 2023 only tells us that FastCTM can reproduce the general pattern of air pollution in 2023, but does not tell us whether FastCTM will accurately predict the effect of interventions. The sensitivity tests in section 3.2 have no basis for comparison, and are in any case so broad (representing nationwide changes in temperature, PBL height, or emissions) that they are a limited test of the CTM's capabilities. At the very minimum, an evaluation is needed which shows that FastCTM's trends actually match the underlying trends in WRF-CMAQ; this should be straightforward for the emissions cases. Since the

goal of FastCTM is to reduce computational costs, it is critical that FastCTM be shown to be faithful to its parent model for realistic applications such as projecting the impact of a change in emissions. Going further and comparing sensitivities for local or single-sector emissions changes would provide even more powerful proof, and I strongly recommend that the authors consider such a comparison.

Without these kinds of quantitative comparisons I can only judge the model's success based on data such as Figure 3, where I am concerned because the patterns do not – speaking qualitatively – appear to match that well between CMAQ and FastCTM. I am particularly concerned that the model may be mostly reproducing emissions maps and historical scalings, rather than accurately representing chemistry and transport (especially given that transport is 2-D only). A more critical, quantitative analysis of the models strengths and weaknesses would be necessary before I would recommend its use in a scientific or regulatory context.

Response: We agree with the reviewer's comment that more analysis is needed to verify FastCTM's capabilities to project the impact of a change in emissions. Since the emissions are the same for each year from 2018-2023, it is not possible to test FastCTM's trends to that of WRF-CMAQ. Instead, we added a comparison between FastCTM and CMAQ under 11 emission scenarios in the winter month of January 2019 and in the summer month of July 2019. The results signified that the FastCTM simulations manifested a high level of concordance with those of CMAQ, which was manifested in two principal aspects. Firstly, similar to CMAQ, the FastCTM model forecasted positive responses to increased emissions in the no-control (NCtrl) scenario and negative responses in the other emission-controlled scenarios. This implies that when emissions were unrestricted and increased, as in the NCtrl scenario, both models detected a corresponding upward trend in pollutant levels. Conversely, in scenarios where emissions were curbed, they both predicted a decline. Secondly, in scenarios characterized by more substantial emission reductions, the FastCTM model simulated a more pronounced decrease in air pollutant concentrations. This is of particular significance as it shows the model's sensitivity to the magnitude of emission interventions. It suggests that the FastCTM model is not only capable of discerning changes in emission scenarios but can also reflect the degree of impact on air quality, thereby reinforcing its reliability and utility in simulating air quality dynamics in tandem with CMAQ. Related results in the manuscript are shown as follows.

The sensitivities of FastCTM simulations to emission interventions were contrasted with those of CMAQ. Specifically, CMAQ was employed to simulate 11 emission scenarios over the two-month periods of January and July 2019 in Southwest China (Huang et al., 2022). The alterations in emissions relative to the base case are presented in Table 1. Among these scenarios, 10 involved reduced emissions of major species, with only the no-control scenario exhibiting increased emissions. Utilizing the identical emissions and meteorological data, FastCTM also conducted simulations, which were then compared to those of CMAQ. For the 11 scenarios in question, the changes in air pollutant concentrations relative to the base case at the locations of 139 national air quality monitoring stations (Figure S14 in the SI) were extracted and compared in the winter month of January 2019 (Figure 9) and in summer month of July 2019 (Figure 10). The results indicated that, overall, the FastCTM simulations were in good agreement with those of CMAQ reflected in two aspects. First, FastCTM predicted positive responses to increased emissions in the nocontrol (NCtrl) scenario and negative responses to other emission-controlled scenarios just as CMAQ. Second, FastCTM simulated larger air pollutant concentration decrease in those scenarios with higher emission reductions. Specifically,

in January 2019, with the exception of NO_2, FastCTM responded to emission changes with an interquartile range (IQR, 25% - 75% percentile) similar to that of CMAQ (Figure 9). For NO_2, in the same emission reduction scenarios, FastCTM simulated lower NO_2 values. In the summer month of July 2019, as depicted in Figure 10, all the criteria pollutants except CO demonstrated a comparable degree of response to emission reductions. The comparison suggests that the FastCTM model is not only capable of discerning changes in emission scenarios but can also reflect the degree of impact on air quality, thereby reinforcing its reliability and utility in simulating air quality dynamics in tandem with CMAQ. It should be noted that in both months, FastCTM exhibited slightly larger median values, suggesting its greater sensitivity to emission interventions.

***Table 1**. The emission change details of emission scenarios*

Scenario	abbreviation	Sector	NO_x	VOCs	SO_2	CO	$PM_{2.5}$	PMC
nocontrol	NCtrl	Industrial	30%	30%	30%	30%	30%	30%
		Traffic	20%	20%	20%	20%	20%	20%
medianX	MedX	Industrial	-36%	-35%	-48%	-23%	-9%	-9%
		Traffic	-40%	-10%	0	-26%	-10%	-10%
medianY	MedY	Industrial	-26%	-20%	-38%	-13%	-4%	-4%
		Traffic	-30%	0%	0	-16%	-5%	-5%
medianZ	MedZ	Industrial	-36%	-10%	-48%	-23%	-9%	-9%
		Traffic	-40%	0%	0	-26%	-10%	-10%
median-3	Med-3	Industrial	-10%	-10%	-18%	0	0	0
		Traffic	-10%	0%	0	0	0	0
median-2	Med-2	Industrial	-16%	-20%	-28%	-3%	0	0
		Traffic	-20%	0%	0	-6%	0	0
median-1	Med-1	Industrial	-26%	-35%	-38%	-13%	-4%	-4%
		Traffic	-30%	-10%	0	-16%	-5%	-5%
median0	Med0	Industrial	-36%	-50%	-48%	-23%	-9%	-9%
		Traffic	-40%	-20%	0	-26%	-10%	-10%
median+1	Med+1	Industrial	-46%	-65%	-58%	-33%	-19%	-19%
		Traffic	-50%	-30%	0	-36%	-20%	-20%
median2030	Med30	Industrial	-55%	-70%	-80%	-40%	-40%	-40%
		Traffic	-60%	-40%	0	-40%	-40%	-40%

median2035	***Med35***	*Industrial*	*-80%*	*-80%*	*-90%*	*-60%*	*-50%*	*-50%*
		Traffic	*-80%*	*-60%*	*0*	*-60%*	*-50%*	*-50%*

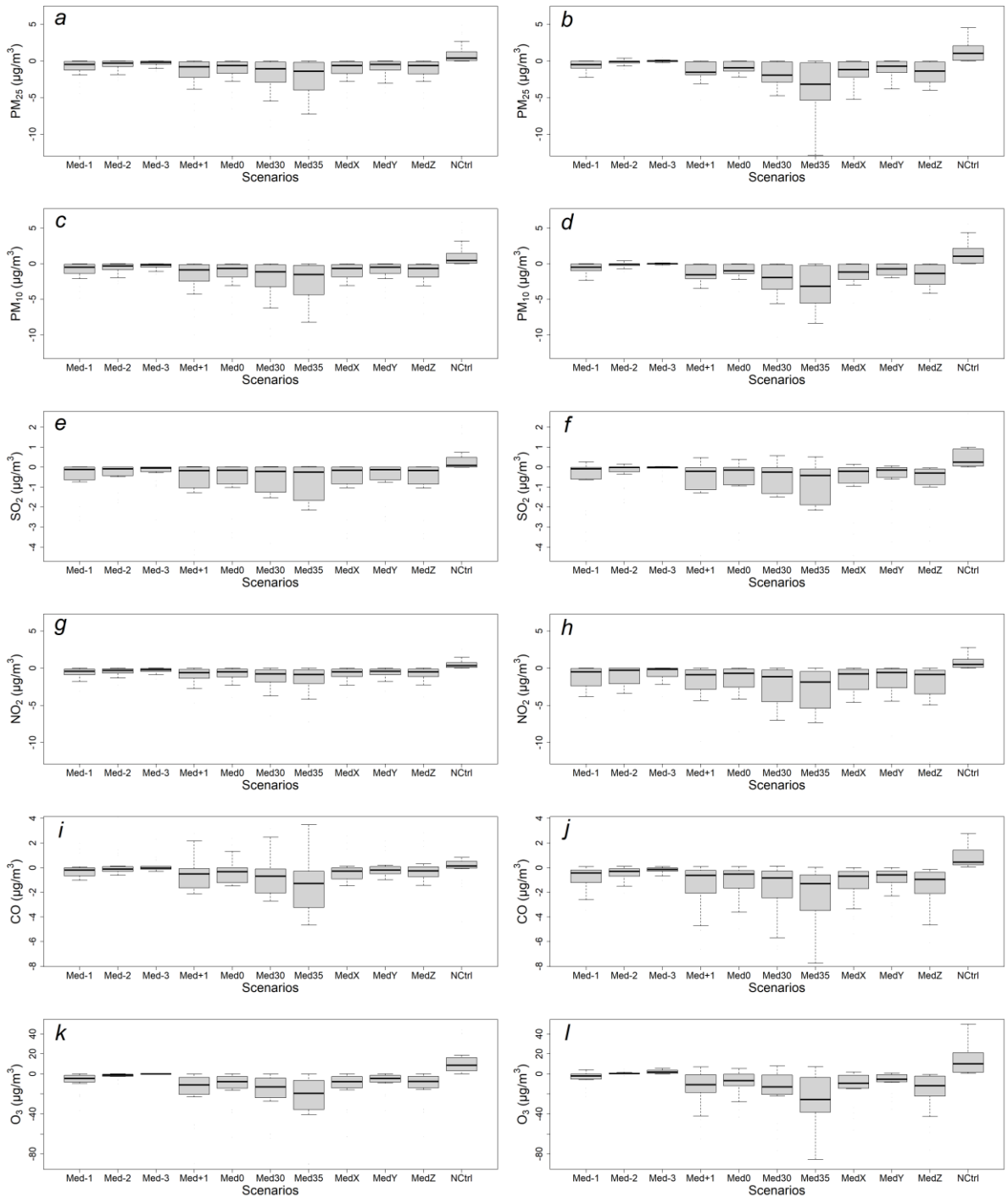

Figure 9: Air pollutant concentration changes in terms of base case simulated by CMAQ (subplots of a, c, e, g, i, and k in the first column) and by FastCTM (subplots of b, d, f, h, j, and l in the second column) in January 2019.

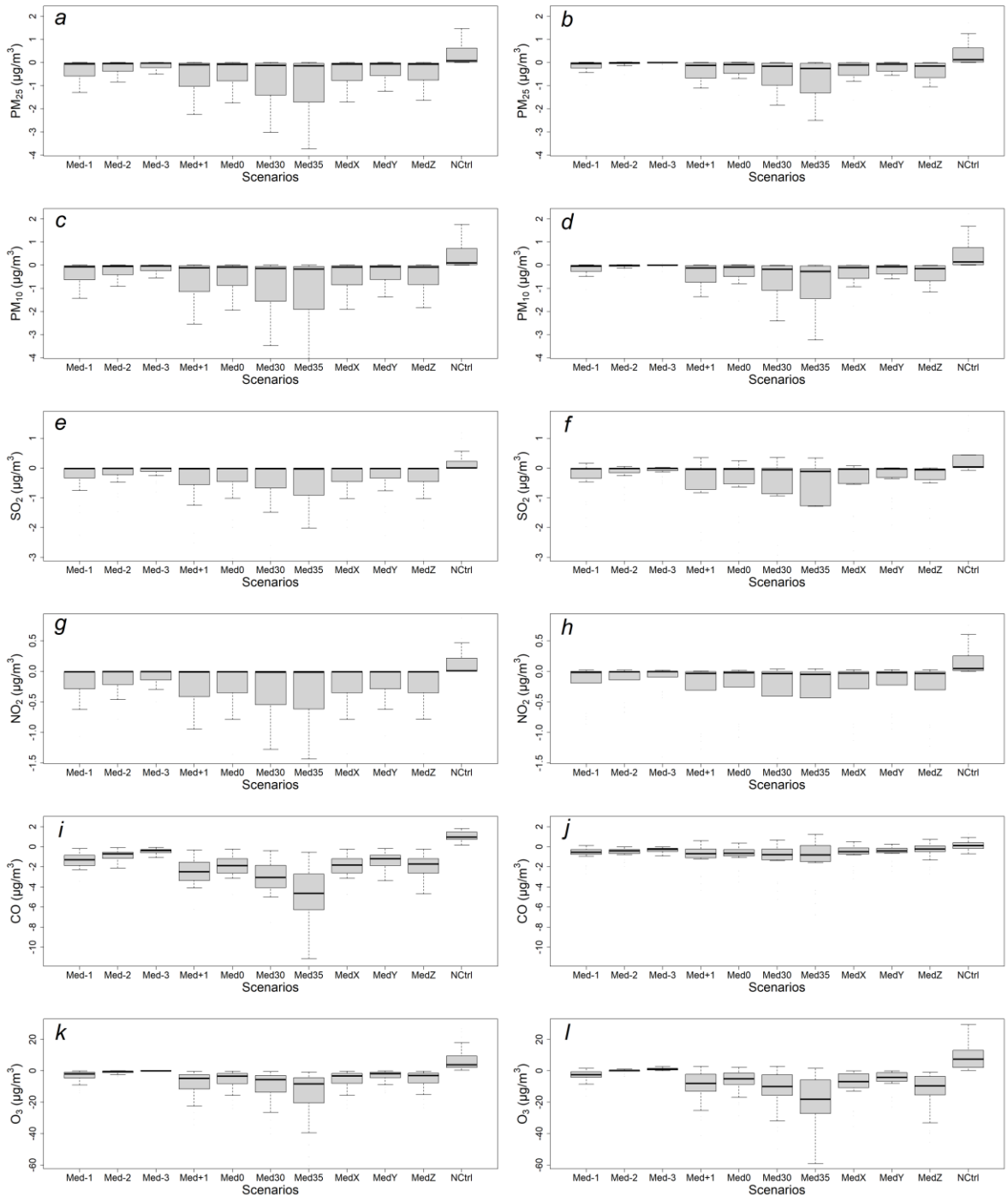

Figure 10: Air pollutant concentration changes in terms of base case simulated by CMAQ (subplots of a, c, e, g, i, and k in the first column) and by FastCTM (subplots of b, d, f, h, j, and i in the second column) in July 2019.

Minor comments

The description of the five operators does not quite seem to verify that physical constraints are being satisfied but this may simply be a misinterpretation on my part. For example, can you confirm that the method you used to generate the convolution kernels (Eq 5-7) for transport ensures mass conservation? This seems to depend on Ci being in units of molec/cm3 rather than ppbv, but the units of Ci are not

clearly specified.

Response: We added descriptions for the model framework in Section 2.3. The unit for all pollutants is μg/m^3. We used an upwind-scheme to simulate diffusion and advection processes. For the scheme, masses are conserved. However, FastCTM is not mass conserved, because it also includes other neural network modules such as reaction and deposition. These deep learning modules are learned to minimize the loss function of mean squared error. The revised model description and figures are shown as follows.

Instead, we use a 1-hour initial pollutant concentration (J=1) to simulate 24-hour air quality pollutants (K=24), to ensure FastCTM is dedicated to learning air quality changes between neighboring two hours as shown in Figure 1a. In other words, at time $t = 0$, FastCTM predicted K-hour air pollutant concentrations of $C_{t=0}, C_{t=1}, \ldots, C_{t=K-1}$, given the input air pollutant concentration in the previous hour $C_{t=-1}$ and corresponding meteorological data and emissions at time $t = 0, 1, \ldots,$ K-1. The unit of concentrations is μg/m^3 for all pollutants.

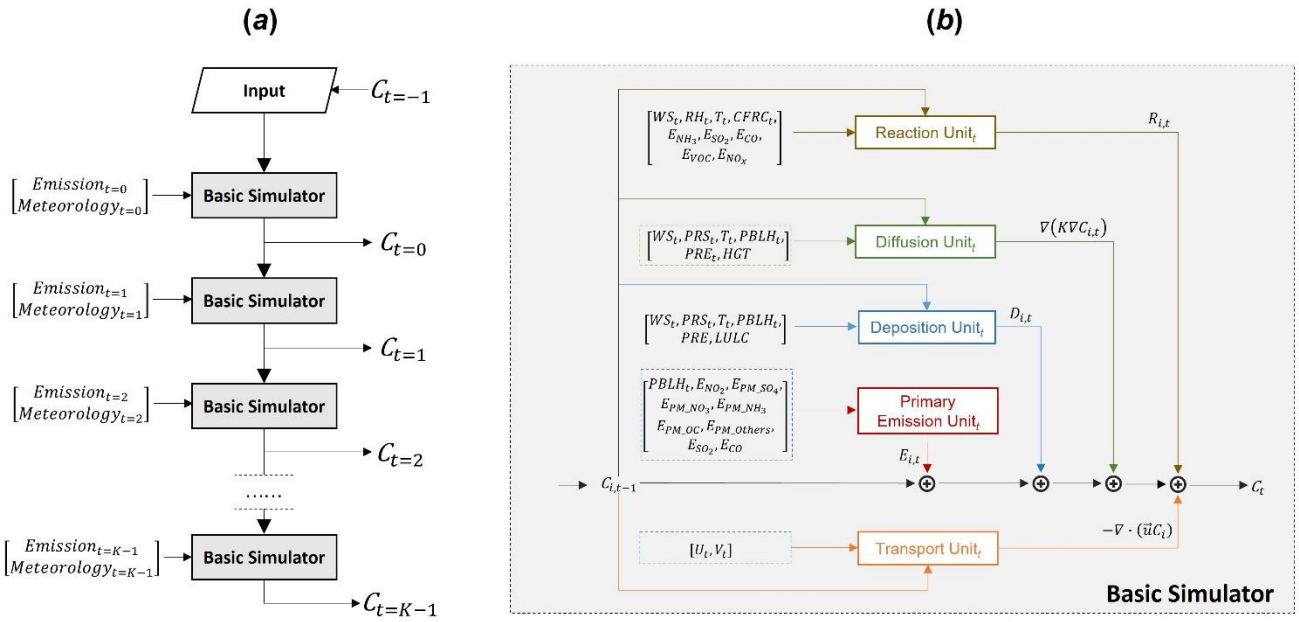

Figure 5: (a) General model workflow, and (b) the basic simulator module structure at the time step t of deep learning simulation model FastCTM designed according to Eq.1. Arrows and boxes with different colours represent calculation modules of different atmospheric physical and chemical processes.

A related concern is that surface layer winds are used and treated conservatively, which neglects the fact of rapid vertical mixing. Can the authors provide evidence that the surface winds (which would be expected to be slower than the mean wind speed in the boundary layer) are accurately predicting pollutant motion? It seems that any model which is designed to predict transport using only the horizontal near-surface winds will underestimate overall transport. Should the model not be using the PBL-averaged horizontal winds in Eq. 4 instead of the surface winds at 10-meter height (lines 83-84)?

Response: We appreciate the reviewer's critical observation regarding the use of surface-layer winds (10 m) in FastCTM's transport module and the potential underestimation of pollutant transport. We

agree that PBL-averaged horizontal winds (also called transport winds) could better predict vertical transportation. We are going to apply our FastCTM model to 3D dimensions in the future version. This will enable a more realistic simulation of both horizontal and vertical transport processes. The relevant section of the manuscript has been revised accordingly as follows.

FastCTM will also extend to 3D dimension to improve its representation for processes such as vertical mixing, vertical wind gradient and in-cloud chemistries.

It would be helpful to get more detail on how components such as the diffusion encoder were trained. Currently the manuscript states that 5 years of data (2018 – 2022 inclusive) were used in training, but not how the five different models were trained using that data. A naïve assessment would assume that all five sub-models were trained based simply on hour-to-hour pollutant concentrations, but that would suggest that the models were each trying to represent all atmospheric processes simultaneously.

Response: The five modules in FastCTM are defined in the form of operator, where operator parameters are estimated, rather than in the form of pure predictor mapping concentrations from one hour to the next. For example, in the diffusion module, FastCTM learns to encode diffusion coefficient K from meteorological conditions before performing an upwind finite difference procedure to solve the diffusion process $\nabla(K\nabla C_i)$. It's also the same for processes such as reaction, advection, and deposition. Therefore, it is impossible for one process to represent all atmospheric processes simultaneously. The independent contribution of each process is depicted in Figure 12 of section 3.3. Each process exhibited its patterns of contribution to hourly air pollutant concentration changes, constrained by the form of the operator in the processes. The related description was added in Section 2.3 Model Training.

Even though five modules are defined in FastCTM, individual processes are not trained separately. The model was trained as a whole with hour-to-hour air pollutant concentrations, while each process could learn its parameters under the constrains of its dedicated formulation. Specifically, FastCTM was tuned to minimize the loss function $\mathcal{L}$, which was determined to be L2 loss (Bühlmann and Yu, 2003) of the regularized mean squared error (MSE) as shown in Eq. 15.

Line 172 says that the reaction encoder in Equation 12 "has the same structure as that of reaction and deposition encoder models (Eq. 10)". This is recursive, but also Eq. 10 refers to the diffusion module?

Response: This error was revised as follows in the corresponding section,

Therefore, the reaction rate constant k is simulated using a spatial encoder function $\mathrm{Encoder}$ as shown in Eq. 12, which has the same structure as that of diffusion encoder modules (Eq. 10).

On line 194, "We did not use the fixed area as that in the previous studies (Xing et al., 2022)" – can you elaborate? It was not clear to me what this meant.

Response: Revised as follows,

The random sampling tactics would help the model learn inherent physical and chemical principles model rather than just statistical spatiotemporal autocorrelations using data in constant spatial area (Xing et al., 2022). Besides, the spatio-temporal random samples contain varied emissions which would improve FastCTM adaptation to changing emission levels.

The y-axis labels on Figure 5 say "Percentage", but from context it appears these must really be the factor difference from the baseline (as all cross at 1.0).

Response: Revised.

Finally, there are numerous minor grammatical errors (e.g. L14: simulations and managements; L67: interpretations of the FastCTM are also widely vowed; L70: including and major; and so on). This is not important for judgment of the paper's appropriateness for publication, but I recommend the authors take another look at the paper to correct such minor issues.

Response: We have read through the manuscript and made a thorough revision paying particular attention to the grammar.

Citations

Wong et al., "Using a land use regression model with machine learning to estimate ground level PM2.5." Environ. Pollut., 2021.

Cheng et al. "Influence of weather and air pollution on concentration change of PM2. 5 using a generalized additive model and gradient boosting machine." Atmospheric environment, 2021.