Point-to-point Responses to Reviewers' Comments

Review #1

As this is a second review, I have focused my efforts on the elements changed in response to my original review.

First, I appreciate the effort put in by the reviewers to address many of my previous concerns. The comparison to land use regression models and to the average CMAQ output is welcome, as is the additional clarity on how R2 and RMSE are calculated. I was particularly happy to see that the authors have conducted paired sensitivity tests with CMAQ and FastCTM.

However, I was disappointed to see that the comparison between FastCTM and the LUR methods appears to be entirely qualitative, with no numerical assessment of the relative performance. Furthermore, the extremely low R2 values achieved by land use regression are surprising. Studies such as Zhang et al (2021) have achieved R2 values of ~0.8 with simple generalized linear models. For this comparison to be useful, I recommend that the authors first establish that their point of comparison (the land use regression model) is a fair representation of recent efforts in this area, achieving reasonable levels of accuracy. It would then be more helpful to compare the accuracy of their LUR models to CMAQ quantitatively, perhaps using the suite of metrics already recommended for air quality comparison by Huang et al. (2021). Said metrics could also be used as a point of comparison for FastCTM against the literature.

Response: We have added statistical metrics of R2, NME and NMB for comparing FastCTM and LUR models as the reviewer suggested. The LUR model used in this study (random forest and XGBoost) represent widely-used approaches in recent literature, as demonstrated in the comprehensive review of popular LUR models by Ma et al. (2024).. The related revision in lines 313-316 are as follows,

"These LUR models were developed using the same input meteorological data, emission, and geophysical variables as FastCTM to ensure fair comparison. When compared with the FastCTM model, the performance of the LUR models was found to be significantly inferior as demonstrated in the Table. 1 and Figure S10 – S12 in the SI. For example, R^2 values for FastCTM range from 0.68-0.90, whereas the LUR models only achieve 0.06-0.33."

***Table 1**. Performance metrics of LUR models and FastCTM compared against CMAQ*

Variable	*Model*	*RMSE*	*R^2*	*NMB*
$PM_{2.5}$	*FastCTM*	*8.78*	*0.81*	*-0.15*
	Liner Model	*35.05*	*0.09*	*-0.24*
	Random Forest	*33.08*	*0.19*	*-0.25*
	XGBoost	*33.02*	*0.14*	*-0.12*
PM_{10}	*FastCTM*	*11.58*	*0.80*	*-0.17*
	Liner Model	*44.66*	*0.10*	*-0.23*
	Random Forest	*45.07*	*0.19*	*-0.33*
	XGBoost	*44.53*	*0.15*	*-0.21*
SO_2	*FastCTM*	*4.51*	*0.80*	*0.09*
	Liner Model	*39.42*	*0.14*	*-1.18*
	Random Forest	*25.74*	*0.33*	*-0.65*
	XGBoost	*25.57*	*0.26*	*-0.60*
NO_2	*FastCTM*	*4.24*	*0.83*	*0.04*
	Liner Model	*21.42*	*0.27*	*-0.30*
	Random Forest	*25.13*	*0.16*	*-0.58*
	XGBoost	*23.88*	*0.15*	*-0.43*
CO	*FastCTM*	*51.84*	*0.90*	*0.01*
	Liner Model	*427.67*	*0.03*	*6.38*
	Random Forest	*83.25*	*0.08*	*1.32*
	XGBoost	*70.06*	*0.06*	*1.10*
O_3	*FastCTM*	*11.46*	*0.68*	*0.02*
	Liner Model	*357.97*	*0.09*	*-0.46*
	Random Forest	*285.16*	*0.19*	*-0.21*
	XGBoost	*291.58*	*0.15*	*-0.22*

However, through in-depth literature search and review, we found that LUR models have been seldom used in the way that the CTM models have been used for air quality simulations and forecasts. CTM models are able to simulate air pollutant concentrations given initial conditions, meteorological conditions and emissions, etc. However, LUR models are typically used to estimate air pollutant concentrations across complete spatial fields by interpolating from discrete station observations using land use and other covariates as predictors. For example, Zhang et al. (2021) used a GLM-based LUR model to predict 10km resolution gridded daily

concentrations of six criteria pollutants in Beijing, supervised by observations from 35 stations. Similarly, Wong et al. (2021) predicted daily $PM_{2.5}$ concentration at 50m resolution in Taiwan using observations from 73 monitoring sites.

LUR models face significant challenges in accurately predicting air pollutant concentrations when relying solely on static relationships between pollutant concentrations and supporting variables, particularly without corresponding in-situ observations. This challenge becomes especially pronounced for hourly air pollutant predictions, as concentrations can change rapidly between consecutive hours even when meteorological conditions show minimal variation. Consequently, there are virtually no successful studies demonstrating LUR models' capability for this type of application, as confirmed by the comprehensive review by Ma et al. (2024)."

Ma, X., Zou, B., Deng, J., Gao, J., Longley, I., Xiao, S., Guo, B., Wu, Y., Xu, T., Xu, X., Yang, X., Wang, X., Tan, Z., Wang, Y., Morawska, L., and Salmond, J.: A comprehensive review of the development of land use regression approaches for modeling spatiotemporal variations of ambient air pollution: A perspective from 2011 to 2023, Environment International, 183, 108430, https://doi.org/10.1016/j.envint.2024.108430, 2024.

Similarly, the comparison of FastCTM versus the simple average of CMAW 2018-2022 does not contain any quantitative assessment. The authors state on the basis of line graphs in Figure 5 that "it becomes evident that the predictions made by FastCTM in 2023 align more closely with the actual CMAQ forecasts", but that is an assertion and not evidence. While I am inclined to agree in the case of PM2.5, the performance for ozone does not appear to be particularly improved – however without any statistical assessment it is hard to say. I am also concerned by the significant gaps in the data shown in Figure 5. Why are some days simply not shown for CMAQ 2023 or FastCTM? These gaps must be explained to ensure a fair and transparent comparison.

Response: We added quantitative assessment for the difference and correlations between FastCTM and CMAQ simulations. The related description was revised in this section (Lines 335-342) as follows.

"*As illustrated in Figure 5, the predictions made by FastCTM in 2023 align more closely with the actual CMAQ forecasts for that year with* R^2 *= 0.94 and 0.72 respectively for* $PM_{2.5}$ *and* O_3*, rather than with the forecasts generated from the training data of 2018-2022 with* R^2*=0.54 and 0.59. The NMB was also lower between FastCTM and CMAQ for the same year 2023. These*

results not only validate the adaptive learning capabilities of the FastCTM model but also indicate that the model is not using a simplistic approach of averaging concentrations from the previous five years based on time of day. Hourly time series plots of air pollutant concentrations (Figure S6 in the SI) further demonstrate that FastCTM appears to incorporate real-time meteorological feedback, adjust for shifts in emission patterns, and leverage its learned relationships to provide more accurate and contemporaneous predictions."

Regarding the missing gaps, these occur because the CMAQ simulations are incomplete due to data unavailability. We have added an explanation for this in the caption of Figure 5 and the Model Evaluation section as follows.

"FastCTM was assessed against CMAQ simulations using the same input emission data and meteorological fields. Starting from 0:00 local time on each day, the CMAQ model simulated 120-hour forecasts in one cycle. There are 139 cycles in the evaluation year of 2023 due to data unavailability in the remaining days."

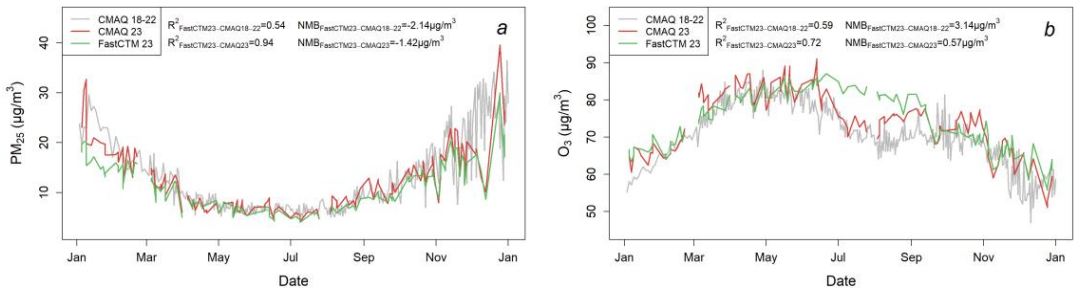

Figure 1: The daily FastCTM forecasts compared with CMAQ forecasts, respectively in the training period of 2018-2022 and evaluation period of 2023 for (a) $PM_{2.5}$ and (b) O_3. The gaps for FastCTM and CMAQ in 2023 are due to data unavailability in these days.

The lack of quantitative analysis further extends to the valuable paired sensitivity tests comparing the response of CMAQ and FastCTM to emissions perturbations (second half of Section 3.2). Showing 24 bar charts (each with 11 box-and-whisker elements) side by side is certainly transparent, but it is very difficult to assess the degree to which the model is better or worse at predicting certain changes. The analysis only states that changes are "comparable", have "similar" IQRs, or are in "good agreement". What is needed is a robust comparison which seeks to identify the conditions under which FastCTM performs well (as defined by some reasonable quantitative standard), and when it performs poorly. For example, it seems that NO2 benefits are routinely overestimated whereas changes in SO2 seem to be generally underestimated, but with some dependence on the specific scenario. In many ways the comparisons for ozone and PM2.5 are encouraging, suggesting that FastCTM may have more

skill in predicting the response to policy than it does in predicting baseline concentrations. These would be very useful features to understand quantitatively, but as it stands the manuscript falls short by providing almost no numerical analysis of the model's performance for this test.

Response: As the reviewer suggested, we added quantitative analysis of the FastCTM's responses to emission changes under different scenarios. Specifically, correlation coefficients (R) were calculated for each air pollutant and each scenario at 139 stations to evaluate consistency between FastCTM and CMAQ in simulating emission interventions. Meanwhile, the boxplots of responses in Figures 9 and 10 have been moved to the supplementary material. The comparative analysis between FastCTM and CMAQ has been revised accordingly to better reflect FastCTM's capabilities in replicating CMAQ model performance, as follows,

"The results indicated that, overall, the FastCTM simulations due to emissions changes were in good agreement with those of CMAQ, as reflected in two aspects. The correlation coefficient R values are around 0.9 for SO_2, NO_2 and O_3 in both summer and winter months. For $PM_{2.5}$ and PM_{10}, FastCTM exhibited higher consistency with CMAQ in July than in January, with R values around 0.6 for most cases. For CO, FastCTM has much better performance in January than in July, with R values of approximately 0.8 and 0.2. Considering that CO concentration changes are mostly due to physical dispersion and transport, the decreased performance is probably due to increased vertical mixing in summer, which is not fully represented in the 2D scheme of FastCTM. Specifically, in January 2019, except NO_2, FastCTM responded to emission changes with an interquartile range (IQR, 25% - 75% percentile) similar to that of CMAQ (Figure S16). In July 2019, as depicted in Figure S17, all the criteria pollutants except CO demonstrated a comparable degree of response to emission reductions."

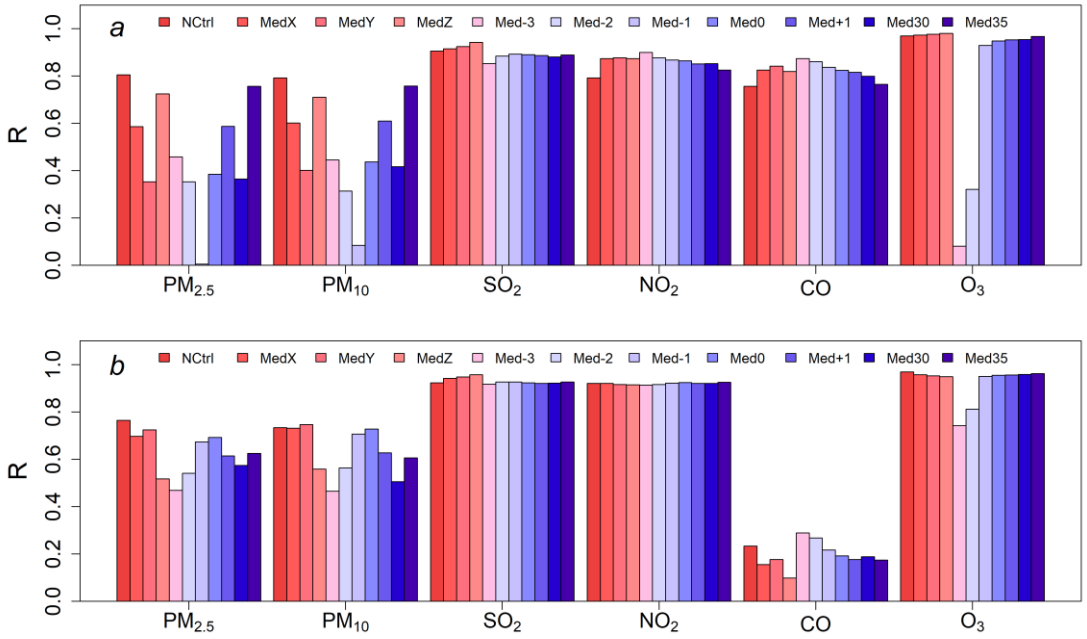

Figure 2: Coefficient of correlation (R) for responses of FastCTM and CMAQ to different emission scenarios and different air pollutants in January, 2023 (panel a) and July, 2023 (panel b).

Regarding the training of the five operators, I was glad to see that there is more description but the new analysis of these operators raises some concerns. Figure 12 seems to show that CMAQ and FastCTM differ markedly in their estimation of the relative contribution of different processes to (what I assume is) the local rate of change of PM2.5, but again there is no quantitative assessment; just a qualitative assertion that the models are comparable but consistent.

Response: We added a description comparing process contributions of the two models with statistical analysis.

"Simulated contributions of five major processes to hourly $PM_{2.5}$ concentration changes are compared between FastCTM and CMAQ at 139 stations (Figure S15) in the Sichuan-Chongqing region from October 12, 2024 to October 16, 2024, as shown in boxplots of Figure 11. Overall, the simulation results of the process contributions by FastCTM and its parent model CMAQ were relatively consistent. Higher degrees of consistency were found in simulations of emissions, advection processes, and diffusion processes between the two models. Contributions from chemical reactions of FastCTM exhibited overestimation compared to CMAQ, while contributions from deposition were underestimated. The differences in the simulated deposition and reaction contributions between the two models could be due to incomplete representation of influencing factors, given the complexity of the two processes. In general, the consistency between the two models provides confidence in the reliability of FastCTM for simulating and understanding the complex interplay of atmospheric processes that govern $PM_{2.5}$ levels."

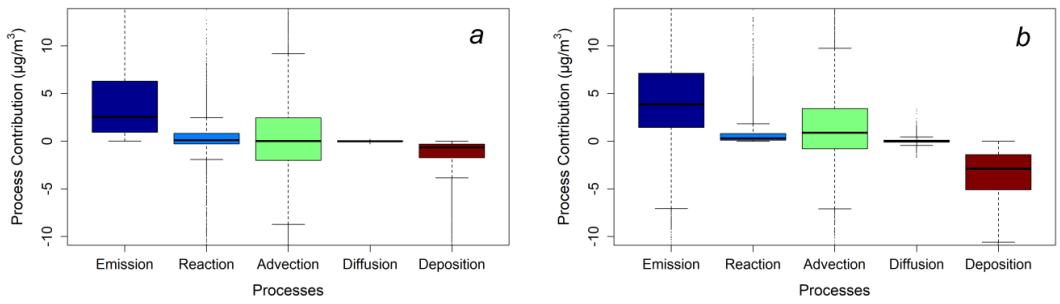

Figure 3: Boxplots of contributions from five major atmospheric processes at 139 evaluation stations from October 13, 2024, to October 16, 2024, simulated by (a) CMAQ

and (b) FastCTM.

The lack of quantitative analysis throughout the manuscript, but in particular in the new sections, means that I cannot recommend this article for publication and recommend a substantial restructuring to focus on reproducible, quantitative assessments of accuracy. I would strongly recommend that the authors seek to reduce the length of the manuscript by condensing the many qualitative comparisons and multi-panel plots into quantitative analyses with comparisons of key metrics of performance. Without these it is very difficult to evaluate the performance of the model against either the parent CMAQ model or existing reduced-order models, and therefore whether there is sufficient novelty.

Response: As the reviewer kindly suggested, we added more quantitative analysis regarding accuracy evaluation, emission sensitivity analysis and process contribution assessment. Qualitative analysis, such as Figures 9, 10, and 12 in the previous version, have been moved to the supplementary material and replaced by more-straightforward comparisons with statistical metrics.

Finally, the manuscript still needs some superficial improvement. The edits appear to have introduced numerous new grammatical errors which I would recommend the authors seek to correct. I would also ask that the authors incorporate higher-resolution images, as some (including Figure 12) are almost unreadable (at least in the PDFs I have access to).

Response: We conducted a thorough check on the manuscript to address grammatical accuracy, clarity, and consistency with scientific writing conventions. Images were originally generated in high resolution. We further improved their quality by revising the axis name, labels and legends.

Huang, L., Zhu, Y., Zhai, H., Xue, S., Zhu, T., Shao, Y., ... & Li, L. (2021). Recommendations on benchmarks for numerical air quality model applications in China–Part 1: PM 2.5 and chemical species. Atmospheric Chemistry and Physics, 21(4), 2725-2743.

Zhang, L., Tian, X., Zhao, Y., Liu, L., Li, Z., Tao, L., ... & Luo, Y. (2021). Application of nonlinear land use regression models for ambient air pollutants and air quality index. Atmospheric Pollution Research, 12(10), 101186.

Review #2

I thank the authors for addressing most of my comments. Here are additional comments:

From comments to reviewer #1:

For the scheme, masses are conserved. However, FastCTM as a whole is not mass conserved, because it also consists other neural network modules such as reaction and deposition.

--> This should be said explicitly; in other words, mass is not conserved by the model

Response: We added the description explicitly, as follows in Section 2.2,

"With the scheme, this transport module itself is mass conserved, even though FastCTM is not mass conserved as a whole."

From comments to reviewer #2:

"Even though these successful applications using deep learning methods to simulate individual atmospheric chemical and physical processes, there is an missing gap in coupling these NN operator replacements together as an complete deep learning based CTM. "

--> I don't think this point is properly motivated still. CTMs are slow but certain components (chemistry, transport) are much slower than others (deposition, emissions). If anything, the previous citations suggest that we do not want a complete deep learning based CTM due to accumulating errors and oversimplifying of the physical schemes. Just because you did the work in creating FastCTM does not mean the motivation and utility is intrinsic.

Response: We appreciate the reviewer's critical perspective on the motivation for developing a complete deep learning-based CTM by coupling individual neural network (NN) operators. This concern prompts us to clarify the unique value of such an integrated framework beyond merely combining existing components, and to address the potential challenges of error accumulation and physical oversimplification.

First, while prior work has successfully applied deep learning to simulate individual processes (e.g., chemical reactions, deposition; Kelp et al., 2022; Silva et al., 2019; Xia et al., 2024), these efforts remain fragmented. A "complete" deep learning CTM is necessary to capture these interdependencies—individual NN operators, in isolation, cannot replicate the holistic dynamics of air quality evolution that are critical for applications such as pollution episode

attribution, emission sensitivity analysis, or real-time forecasting. For example, understanding how a regional emission reduction affects $PM_{2.5}$ requires not only simulating emissions but also how those reduced species are transported, chemically transformed, and deposited—interactions that only a unified framework can model coherently.

Second, we acknowledge the risk of accumulating errors in full deep learning-based models, which the reviewer rightfully highlights. FastCTM's design explicitly addresses this by adopting a principle-informed structure (Section 2.2), where each module (transport, reaction, deposition, etc.) is constrained by the governing physical/chemical equations (e.g., Eq. 1). This distinguishes it from unconstrained "black-box" deep learning models, which are more prone to oversimplification and error propagation. As demonstrated by Sturm and Wexler (2020), physics-constrained machine learning frameworks can mitigate such risks by ensuring process interactions align with fundamental principles. Our results indicated that FastCTM maintains high agreement with CMAQ across long-term simulations (Section 3.1) and exhibits good consistent sensitivity to meteorology and emissions (Section 3.2).

Third, the utility of FastCTM stems from its ability to bridge the gap between the computational efficiency of deep learning and the functional completeness of traditional CTMs. Traditional CTMs, while accurate, are computationally prohibitive for many practical use cases—for example, high-resolution ensemble forecasting or rapid evaluation of hundreds of emission control scenarios (Efstathiou et al., 2024). Individual NN operators, while fast, cannot replace a full CTM in these contexts because they lack the integrated process chain needed to simulate end-to-end air quality. FastCTM addresses this by providing a unified tool that retains the multi-functionality of traditional CTMs (process attribution, sensitivity analysis) while achieving GPU-accelerated speeds (Section 3), making it feasible for applications where both speed and completeness are critical.

In summary, the motivation for FastCTM lies not in "intrinsic" value from coupling components, but in addressing a critical gap: the need for an efficient, integrated model that captures interconnected atmospheric processes—constrained by physical principles to avoid oversimplification—for real-world air quality management and research. This aligns with recent calls in the literature for physics-informed machine learning models that retain the interpretability and reliability of traditional models while leveraging the speed of deep learning (Irrgang et al., 2021; Reichstein et al., 2019).

References

Efstathiou, C. I., et al. (2024). Enabling high-performance cloud computing for the Community Multiscale Air Quality Model (CMAQ) version 5.3.3: performance evaluation and benefits for the user community. *Geoscientific Model Development*, 17, 7001–7027.

Irrgang, C., et al. (2021). Towards neural Earth system modelling by integrating artificial intelligence in Earth system science. *Nature Machine Intelligence*, 3, 667–674.

Kelp, M. M., et al. (2022). An online-learned neural network chemical solver for stable long-term global simulations of atmospheric chemistry. *Journal of Advances in Modeling Earth Systems*, 14, e2021MS002926.

Reichstein, M., et al. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, 566, 195–204.

Silva, S. J., et al. (2019). A deep learning parameterization for ozone dry deposition velocities. *Geophysical Research Letters*, 46, 983–989.

Sturm, P. O., & Wexler, A. S. (2020). A mass- and energy-conserving framework for using machine learning to speed computations: a photochemistry example. *Geoscientific Model Development*, 13, 4435–4442.

Xia, Z., et al. (2024). Advancing Photochemistry Simulation in WRF-Chem V4.0: Artificial Intelligence PhotoChemistry (AIPC) Scheme with Multi-Head Self-Attention Algorithm.

"Interpretations of deep learning network are also widely vowed to improve their applications in earth system science and climate studies"

--> Again, perhaps a translation issue here, but I do not know what interpretations mean here. I also don't think this work will 'widely vowed to improve' climate studies, which are way ahead of their adoption of ML in their modeling approaches compared to air quality

Response: We appreciate the reviewer's insightful comment, which highlights the need for clearer terminology and contextualization. We apologize for the imprecision and confusion in our original statement. The term "interpretations" refers to model interpretability—i.e., the ability to trace predictions to underlying processes or mechanisms, rather than treating the model as an opaque "black box". This aligns with the core design of FastCTM: its principle-informed structure represents five modular processes, each with physically constrained formulations (Eqs. 3 – 14). As demonstrated in Section 3.3, this enables quantification of individual process contributions to pollutant concentration changes (e.g., Figure 10 shows transport dominating $PM_{2.5}$ changes in a pollution episode, while deposition offsets chemical production). Such interpretability distinguishes FastCTM from black-box deep learning models, enabling error attribution and physical insight.

We also acknowledge the reviewer's point about the relative maturity of machine learning adoption in climate studies compared to air quality modeling. Our intention was not to suggest that this work would revolutionize climate studies, but rather to position FastCTM within the broader context of interpretable machine learning in Earth system science. Interpretability is a shared priority across earth system science, including climate research. As noted by Irrgang et al. (2021), neural network applications in earth system modeling (whether for air quality, climate, or other domains) increasingly demand interpretability to ensure physical consistency and facilitate knowledge discovery. FastCTM contributes to this broader effort by demonstrating how principle-constrained architectures can retain machine learning efficiency while enabling process-level analysis.

We refined this sentence to: "*Enhancing the interpretability of deep-learning models is critical for advancing their application in Earth-system science, including climate and air-quality studies.*"

"FastCTM's design supports incremental integration of additional species (e.g., via user-defined modules) without overhauling the core framework. Future versions will explore adding VOCs and secondary organics to address broader research needs."

--> Can this be expanded on? How is this the case. If you wanted to add VOC species won't those species have to be added in multiple modules (e.g., emissions, chemistry)? Or can this be fine-tuned like in transformer-like models (also referred to as transfer learning in different settings). If so, how can you do this? It is unclear to me how this can be incrementally added without retraining the entire pipeline if they are trained together, even though each individual process is discretized differently.

Response: We appreciate the reviewer's critical question regarding the incremental integration of additional species (e.g., VOCs) into FastCTM. As noted, adding new species such as VOCs—which participate in multiple processes (emissions, chemical reactions, transport, etc.)—requires updates across multiple modules, and retraining of the modified model with the new variables is indeed necessary. However, if added species are chemical inactive, which did not participate in chemistry, we could freeze parameters in the chemical module to fine-tune FastCTM to adapt to new input. We revised the original description that clarify the need for retraining when adding new species, while still highlighting the model's modular design for targeted updates to avoid confusions, as follows,

"*FastCTM's modular, principle-informed architecture facilitates targeted updates to integrate additional species (e.g., VOCs or secondary organics) by focusing modifications on relevant*

processes rather than overhauling the entire framework. However, adding new species, especially those participating in multiple atmospheric processes, requires updating associated modules and retraining the model with the expanded set of variables to ensure the model learns the new species' interactions with existing pollutants and processes. Future work will explore such expansions, leveraging the framework's modularity to streamline updates while retraining to incorporate the new species and their dynamics."

"Response: As reviewer kindly pointed, FastCTM possibly "

--> You used this response for reviewer #1 which works well. But this is not the same thing I am asking. You can't see any diurnal errors here. Choose week long time series and show how we get to diurnally varying RMSEs. Furthermore, Figure 5 is a bit messy. There are many trajectories and it's hard to see the difference between CMAQ 2023 and FastCTM 2023.

Response: We apologize for not responding adequately to this specific concern in the previous revision. As suggested, we compared CMAQ forecasts and FastCTM forecasts by plotting time series of hourly pollutant concentrations averaged for all China at two winter dates, in Figure S6 of the supplementary material. It is clear that FastCTM did not predict the same values given the time of day, as the reviewer kindly asked in the previous revision. We can also conclude this from Figure 5. In this figure, daily average FastCTM simulations in 2023 correlate much better with CMAQ simulations in 2023 rather than in the training period of 2018-2022, indicating FastCTM is not just borrowing time-dependent rules from the training dataset. We also revised Figure 5 by adding statistical metrics for the time series.

Descriptions in the manuscript are revised accordingly, as follows in Lines 273-274 and 336-343:

"Hourly RMSE values show clear diurnal variation with higher RMSE values in the nighttime than daytime, which could be due to higher hourly concentrations of air pollutants at night, except for O_3 (Figure S5 of SI)."

"As illustrated in Figure 5, the predictions made by FastCTM in 2023 align more closely with the actual CMAQ forecasts for that year with R^2 = 0.94 and 0.72 respectively for $PM_{2.5}$ and O_3, rather than with the forecasts generated from the training data of 2018-2022 with R^2=0.54 and 0.59. The NMB was also lower between FastCTM and CMAQ for the same year 2023. These results not only validate the adaptive learning capabilities of the FastCTM model but also indicate that the model is not using a simplistic approach of averaging concentrations from the previous five years based on time of day. Hourly time series plots of air pollutant concentrations (Figure S6 in the SI) further demonstrate that FastCTM appears to incorporate real-time

meteorological feedback, adjust for shifts in emission patterns, and leverage its learned relationships to provide more accurate and contemporaneous predictions."

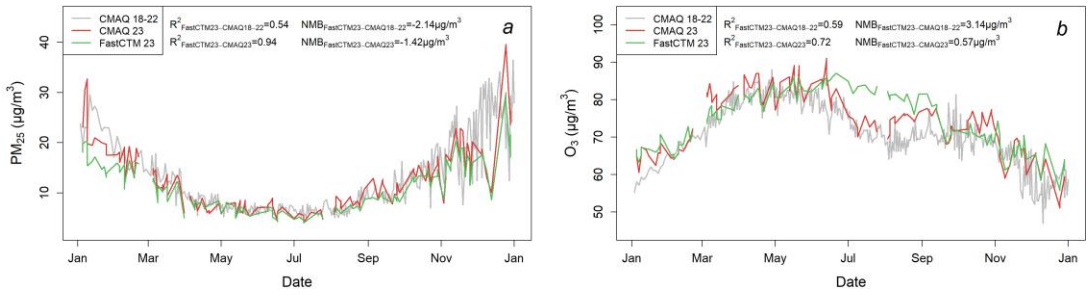

Figure 4: The daily FastCTM forecasts comparing to CMAQ forecasts respectively in training period of 2018-2022 and evaluation period of 2023 for (a) $PM_{2.5}$ and (b) O_3. The gaps for FastCTM and CMAQ in 2023 are due to data unavailability in these days.

"it is a notable finding that the MAE values tend to be higher in polluted areas. This can be attributed to the complex and dynamic nature of pollutant interactions in such regions. In polluted environments, there are often multiple sources of emissions, complex chemical reactions, and variable meteorological conditions that can lead to greater discrepancies between the model - predicted and actual pollutant concentrations. Conversely, the NMAE values exhibit an opposite trend, being lower in polluted areas. In these regions, the NMAE values typically hover around 0.2, in contrast to the relatively higher values of approximately 1 in cleaner areas."

--> I think this is a normal finding. If your FastCTM model is able to show the tendency effect of individual processes why can't you discuss errors in terms of this as well? This would actually be interesting and useful if you can definitively say that transport or diffusion is the problem in urban areas. Not many ML error analyses are able to do such comparisons

Response: We agree with the reviewer that higher MAE in regions with elevated pollutant concentrations is a normal and expected finding, rather than "notable" as initially described. This phenomenon arises because high-pollution areas are characterized by more intense atmospheric processes—including stronger emissions, more vigorous chemical reactions, and complex transport dynamics. Minor inaccuracies in modeling these intensified processes (e.g., small deviations in reaction rates) can be amplified, leading to larger absolute discrepancies in simulated concentrations, which directly contribute to higher MAE values. Conversely, the lower NMAE in these regions reflects the model's ability to capture the relative magnitude of pollution levels, as the large baseline concentrations normalize the impact of absolute errors.

We greatly appreciate the reviewer's insightful suggestion to decompose errors into contributions from specific processes (transport, diffusion, chemistry, etc.). This would indeed be a valuable and innovative contribution to ML-based air quality modeling, as few studies have been able to provide such process-specific error attribution. FastCTM's modular architecture theoretically enables this type of analysis by allowing us to isolate individual process contributions to prediction errors—for instance, determining whether transport errors dominate in urban areas or if chemical reaction uncertainties are the primary source of discrepancies in specific regions.

However, implementing such a comprehensive process-oriented error analysis would require extensive retrospective simulations with systematic perturbation of individual modules, along with detailed validation against process-specific observational data (e.g., tracer studies, chamber experiments). This would demand substantial computational resources and time that extend beyond the scope of the current study. We recognize this as a compelling direction for future research, where we can systematically conduct process-oriented error attribution by leveraging long-term historical datasets, enhanced computational capabilities, and potentially collaborating with observational campaigns designed to isolate individual process contributions.

For the current manuscript, we revised the relevant text to clarify that the observed MAE patterns in polluted areas are expected phenomena arising from the amplification of minor process-related uncertainties in regions with intense atmospheric activity, while acknowledging the potential for future process-specific error analysis enabled by FastCTM's modular design.

"The spatial distributions of the mean absolute error (MAE) and the normalized mean absolute error (NMAE) are presented in Figure 3. For all six pollutants under consideration, MAE values tend to be higher in polluted areas. In polluted environments, there are often multiple sources of emissions, complex chemical reactions, and variable meteorological conditions that can lead to greater discrepancies between the predicted concentrations between the two models."

"It should also be noted that atmospheric physical and chemical processes are defined in principles-guided neural network modules in FastCTM. Their specific formulation was learned and optimized to minimize the sum of loss errors of all species concentrations, rather than being supervised by data of actual internal processes in CMAQ. The actual contributions of each process to pollutant concentration changes can be calculated using the integrated process rate (IPR) analysis and integrated reaction rate (IRR) analysis tools within CMAQ. Future studies could use these IPR and IRR results to supervise the simulated processes in FastCTM to further improve its simulation accuracy and robustness."

I find most of the figures quite poorly made. The legends, texts, and labels are much too small:

Fig 2: legends too small. Are these all forecasts across all time in 2023?

3: color bar, axes labels too small

4: Not colorblind friendly, should be "R-squared"

5: see above comment in text, hard to see difference between 2023 CMAQ, FastCTM

9, 10: way too many subplots and its actually pretty hard to tell the difference between CMAQ and FastCTM in this format

12: too small, figure and labels are fuzzy

Response: Thanks. Figures are re-plot with larger axis, labels and legend texts, to improve quality and readability. Figures with more quantitative analysis are also added.

Fig.2: legend, axis tick name and label name are enlarged, evalution time are added. Colors of data points are also revised, as follows,

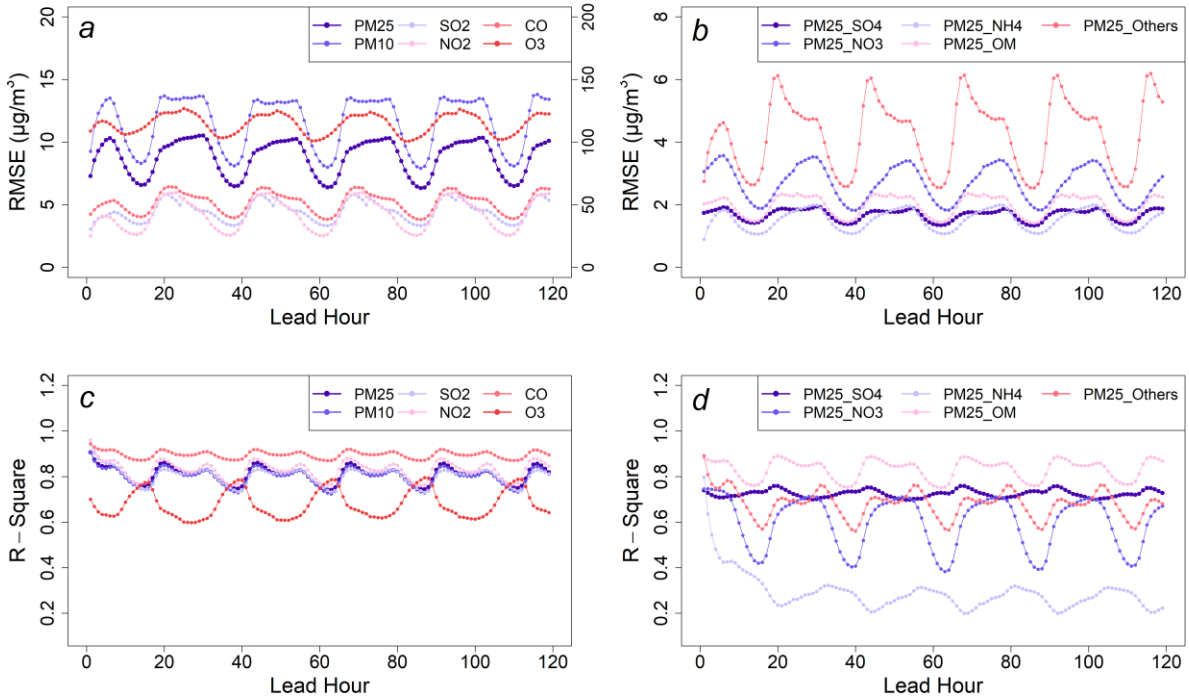

Figure 5: The evaluation performances of FastCTM forecasts against CMAQ forecasts in 2023. Panel (a) and (b) respectively show RMSE values of criteria pollutants and the $PM_{2.5}$ components. Panel (c) and (d) respectively show R^2 values. It should be noted that RMSE value of CO corresponds to the right axis in panel (a).

Fig.3 color bar and axes labels are enalarged as follows,

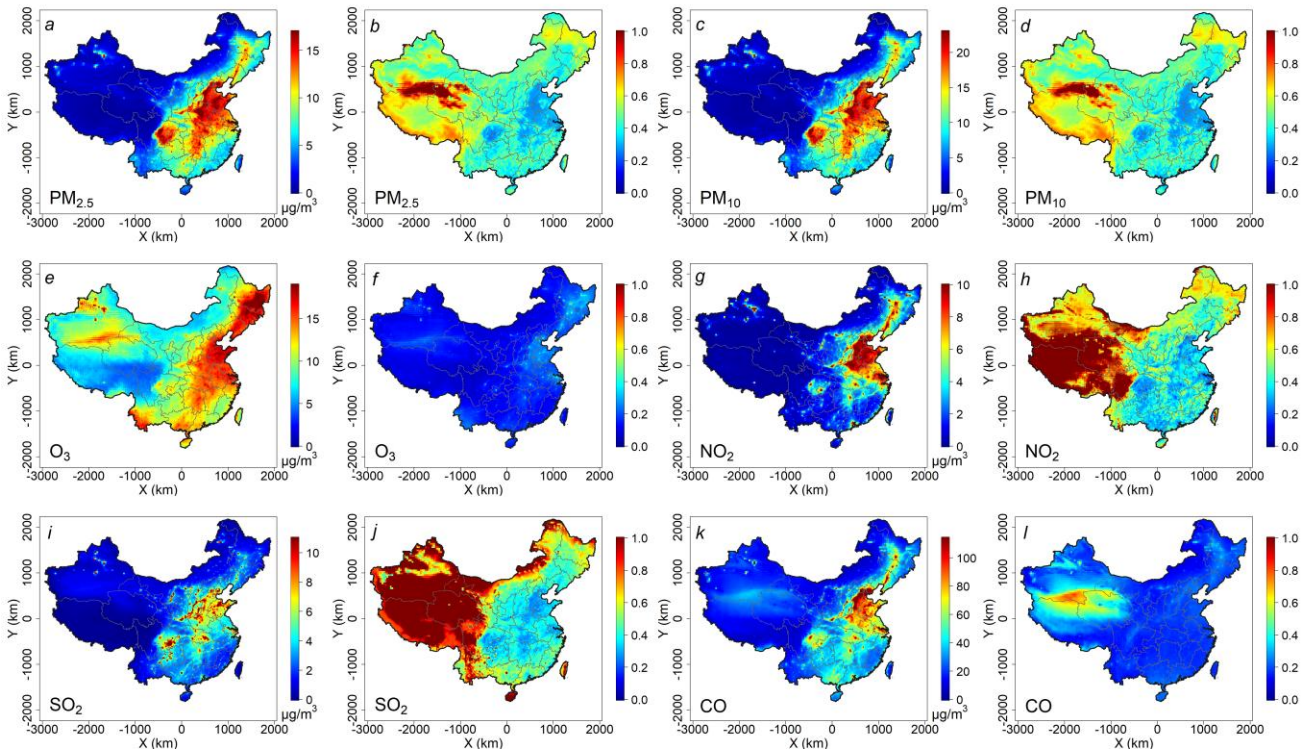

Figure 6: Spatial distribution of mean absolute error (panels a, c, e, g, i, and k) and normalized mean absolute error for the six criteria pollutants (panels b, d, f, h, j, and l) of FastCTM compared with CMAQ in 2023.

Fig.4: Colors are changed and label name are revised, as follows,

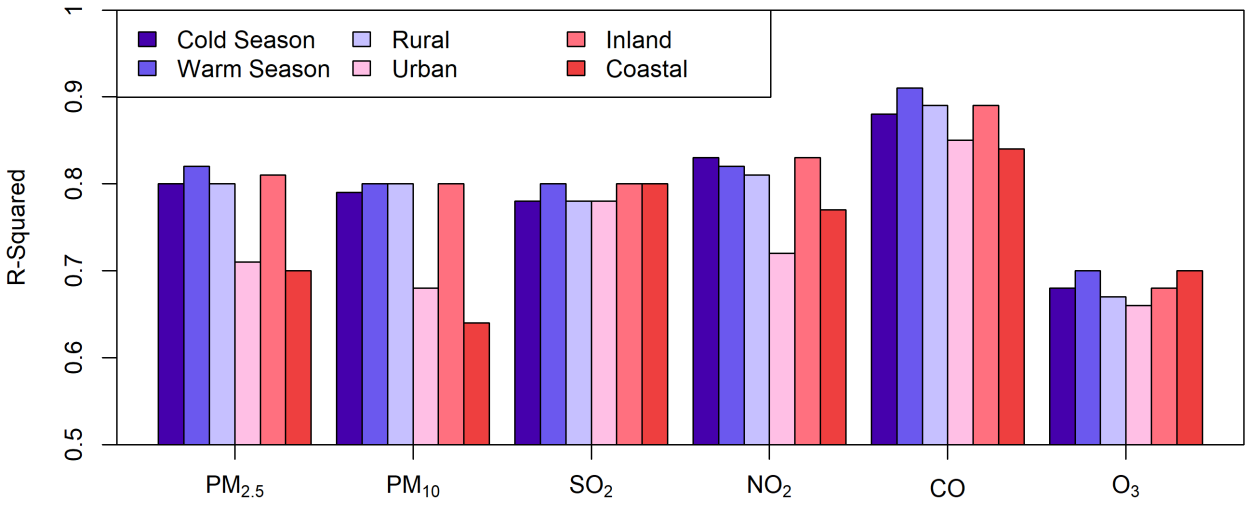

Figure 7: The mean evaluation R^2 values for all 119 leading hours of FastCTM forecasts in warm/cold seasons, rural/urban areas, and coastal/inland areas.

Fig.5: reivsed as shown above.

Correlation anlaysis are exhibited to better demonstrate the relation and difference between FastCTM and CMAQ, as follows. Fig.9 and 10 are moved to the supplementary material.

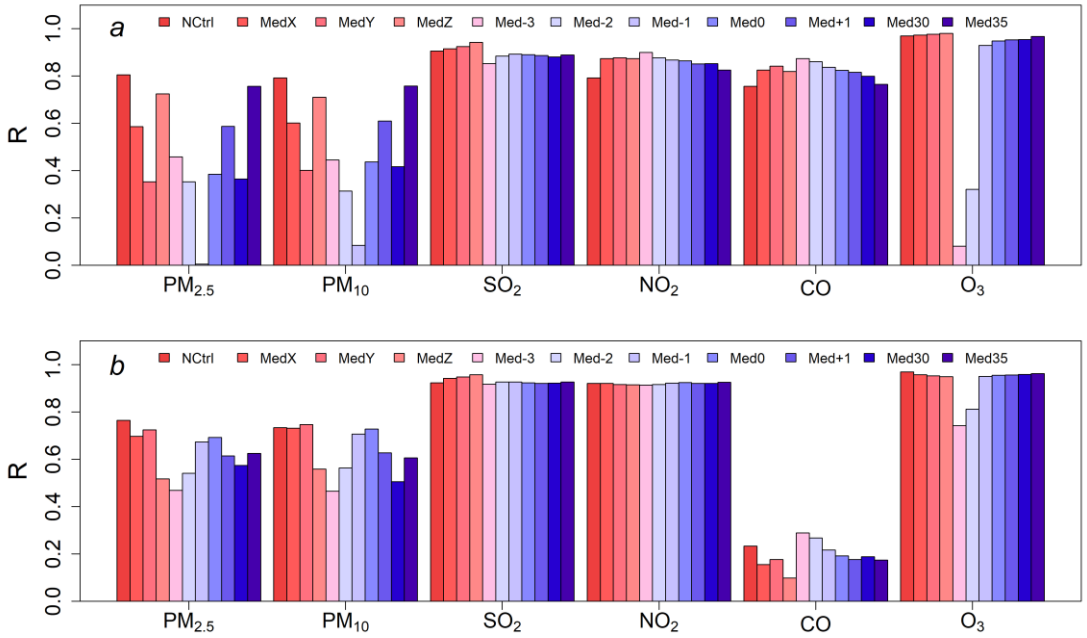

Figure 8: Coefficient of correlation (R) for responses of FastCTM and CMAQ to different emission scenarios and different air pollutants in January, 2023 (panel a) and July, 2023 (panel b).

Fig. 12 is replaced with quantative boxplots to better compare process contributions from different processes, as follows,

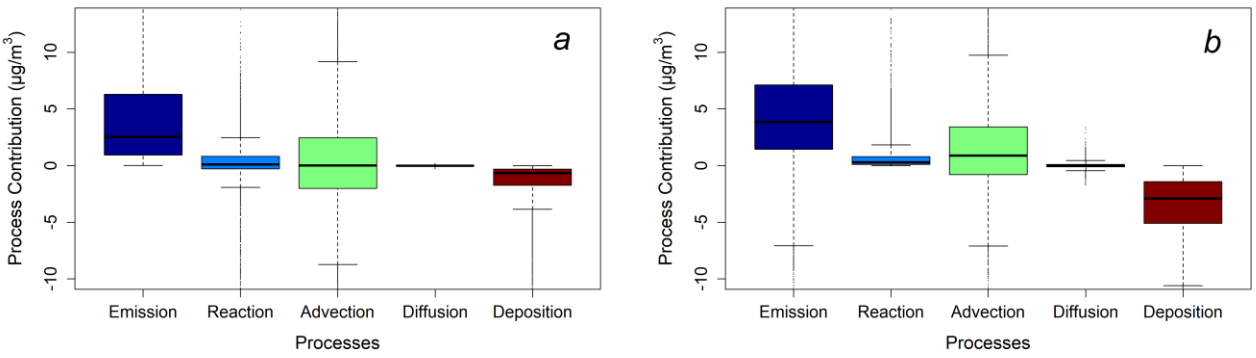

Figure 9: Boxplots of hourly PM2.5 contribution changes from five major atmospheric processes at 139 evaluation stations from October 13, 2024, to October 16, 2024, simulated by (a) CMAQ and (b) FastCTM.