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 1: We sincerely thank the reviewer for the careful evaluation of our manuscript and for the constructive comments provided throughout the review process. We appreciate the reviewer's recognition of the methodological contribution of this work and their support for publication.

2. Major concerns

RC: ***In principle, I am supportive of the publication of this manuscript, as it presents a methodologically interesting contribution. However, I do not fully agree with the claimed novelties and conclusions regarding computational efficiency, which require revision before the paper can be accepted. While the authors analyze that the computational complexity of a single CNN layer is generally lower than that of a Transformer layer, they subsequently introduce architecture-specific assumptions, namely that CNNs necessarily require much deeper architectures with large hidden dimensions, whereas sequence-based Transformers do not, and use this to conclude that the proposed Transformer approach offers a more scalable solution. I find this line of reasoning insufficiently general and potentially misleading. In particular, the need for deeper or wider CNN architectures depends strongly on task formulation, architectural design, and representational choices, and does not constitute a universal limitation of CNN-based approaches. As such, claims of superior scalability or computational efficiency should be carefully qualified and framed as empirical observations under specific model configurations, rather than as a general advantage of the proposed method.***

AR: Thank you for the insightful comments and for acknowledging the methodological interest of our work. We agree that conclusions about the superior computational efficiency or scalability of Transformer-based models over CNNs are not reliable beyond the specific methodological framework considered in this study. In response, we have removed the discussion and claims related to general computational efficiency and scalability from the manuscript, and refocused the presentation on the technical contribution of the proposed modeling framework. The revised manuscript emphasizes that our intention is not to claim a universal advantage of Transformer architectures, but rather to demonstrate that, for the specific task formulation and data structure considered here, the Transformer-based TGEOS model exhibits favorable predictive performance. Finally, we appreciate the reviewer again for these insightful comments, which have helped us substantially improve the clarity and framing of the manuscript.

We have revised the corresponding part of the manuscript, details are shown in below:

1 Introduction

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). This limitation restricts the applicability of these emulators to more intricate emission scenarios. Therefore, existing RSM-based emulators 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 studies directly used in-situ observations as targets based on ML method (Du et al., 2023; 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.~~ 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), their applicability to detailed emission response and broader research domains has so far been limited. 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.

2.2.1 Model architecture

~~In previous emulator modeling, field-to-field modeling using the convolution neural networks (CNN) architecture has been widely used because of the efficient usage of capturing the spatial relationship between features and concentrations (Xing et al., 2020; Huang et al., 2021; Liu et al., 2022). However, both model inputs and outputs are represented as high-resolution 2D matrices in these approaches, which require significant GPU memory and computational resources, especially when the number of input variables increases. As a result, these model were limited to a few kinds of emission species (Xing et al., 2020), or solely average values (Liu et al., 2022). In contrast, our dataset includes over 100 variables, including sectoral emissions and multiple meteorological parameters. Representing these as full spatial fields and training a surface-to-surface model was not feasible under our available computational resources. To clarify this, we provided a direct comparison between a hypothetical global CNN that operates on the full spatial field (surface-to-surface mapping) and Transformer used in this study in Text S3. For importantly, the number of scenarios in our dataset is limited, which could only yield fewer than 500 samples totally for training and significantly fall short of the requirements for robust model development. Therefore, instead of modeling spatial fields directly, we adopted a high-dimensional sequential modeling strategy. Our dataset is not field-based but rather consists of structured multivariate sequences, in which spatial and feature-level information (e.g., emissions, meteorology, and concentrations at 3×3 neighborhood with 9 grid cells) is flattened and treated as a sequence of tokens fed into TGEOS model. This approach offers a more scalable solution while preserving the ability to capture complex relationships among variables across grid points.~~ Previous emulator studies have often adopted field-based modeling strategies, in which both inputs and outputs are represented as spatially explicit two-dimensional fields (Xing et al., 2020; Huang et al., 2021; Liu et al., 2022). While effective in data-rich settings, such formulations typically require a large number of training samples to robustly learn high-dimensional spatial mappings. In the present study, the number of available scenarios is limited, yielding fewer than 500 samples in total for training. This sample size is insufficient to support stable training of high-capacity field-to-field models, particularly when the input space includes more than 100 variables spanning sectoral emissions and multiple meteorological parameters. Under this data regime, directly modeling full spatial fields would substantially increase the risk of overfitting and unstable generalization. Therefore, rather than adopting a purely field-based representation, we reformulate the problem as a high-dimensional sequential learning task. The dataset is organized as structured multivariate sequences, in which spatial and feature-level information, such as emissions, meteorology, and concentrations over a 3×3 neighborhood (9 grid cells), is flattened and treated as a sequence of tokens input to the TGEOS model. This formulation is better aligned with the available sample size and enables more efficient utilization of limited training data, while still preserving key spatial and cross-variable dependencies among neighboring grid points.

4 Conclusions

~~In this study, we developed a Transformer-based information prediction model "TGEOS v1.0" that serve as a GEOS-Chem proxy model to represent future air quality under different emission scenarios. Based on GEOS-Chem version 14.2.2, 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. With exceptional computational efficiency, TGEOS can perform one-year predictions in about 2.51 seconds.~~ In this study, we develop a Transformer-based informative prediction model, TGEOS v1.0, which serves as a GEOS-Chem agent model to represent future air quality under different emission scenarios. Built upon simulations from GEOS-Chem version 14.2.2, TGEOS learns the complex relationships between

precursor emissions and the resulting concentrations of $PM_{2.5}$ and O_3. Once trained, the model enables rapid online assessment of the impacts of alternative emission control strategies, producing one-year predictions in approximately 2.51 seconds. Compared to previous studies that focus solely on average predication, TGEOS can predict the probability distribution of $PM_{2.5}$ and O_3 concentrations in different regions. Leveraging the strengths of high-dimensional data modeling inherent in the Transformer model, TGEOS is capable to provide more accurate predictions based on more detailed emission scenarios that take into account multiple precursor species, emission sectors, and adjacent emissions.