Artificial Intelligence, Machine Learning, & Robotics in Business

Efficient Weather and Air Quality Forecasts Through AI

Jia Xing[1], Joshua S. Fu*[1,2]

[1]University of Tennessee-Knoxville, Knoxville, Tennessee, USA; [2]Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA

***Correspondence: jsfu@utk.edu, Department of Civil and Environmental Engineering, the University of Tennessee, Knoxville, TN 37996, USA**

Abstract: Accurate and timely weather forecasts and air quality predictions are essential for designing effective strategies to manage weather-related events and air pollution. These forecasts also play a key role in understanding aerosol-meteorology interactions within weather systems. However, traditional numerical methods, such as chemical transport models (CTMs), are computationally intensive, and their high resource demands limit their practical use in real-time air quality management and weather forecasting. In response to these challenges, we develop a novel approach called DeepCTM4D, which leverages deep learning to replicate CTM simulations, enhancing the computational efficiency of meteorology and air quality modeling in the four-dimensional chemistry space. The DeepCTM4D model is trained to accurately predict atmospheric chemical concentrations based on inputs such as precursor emissions, meteorological factors, and initial conditions. The key advantage of DeepCTM4D lies in its ability to efficiently identify the main drivers of pollution formation and assess how changes in emissions and meteorological conditions influence air quality. The relationships between emissions, meteorology, and concentration that DeepCTM4D captures align with established atmospheric chemistry mechanisms, further supporting the model's scientific validity. Overall, DeepCTM4D offers a promising solution for simulating complex atmospheric processes, providing policymakers with critical information needed to design effective pollution control strategies and weather-caused events. This AI-driven model can also be integrated into global weather and air quality forecasting systems, serving as a powerful tool for more efficient, real-time predictions.

Keywords: Air Quality, Machine Learning, DeepCTM4D, Weather Forecast, Four-Dimension

1 INTRODUCTION

In the context of global climate change, weather forecasts and atmospheric management face significant challenges due to inefficient and inaccurate predictions of extreme events such as air pollution, and weather anomalies. The complexity of chemical interactions makes it difficult for traditional numerical models to provide timely and accurate estimates in weather forecasts and air quality prediction, which require better processes for dynamics, physics, and chemistry. Addressing the challenges demands a more efficient approach to enhance the model speed and forecast accuracy.

Advanced machine learning methods effectively capture the complex, nonlinear relationships within the atmospheric system (Cabaneros et al., 2019; Kelp et al., 2020; Xing et al., 2020) and offer great potential for predicting atmospheric chemical concentrations using emissions and meteorological data much efficiently by avoiding the computationally intensive processes of traditional chemical transport models (CTMs). However, limitations arise from the chaotic nature of the atmosphere, where small errors grow over time, leading to significant biases during long-term forecasts with frequent spatial interactions. Additionally, these models require handling large datasets, straining memory, and computational resources. Uncertainties in initial conditions and emission estimates from traditional models can also affect prediction accuracy. Our previous work has advanced the development of the deep-learning-based CTM (DeepCTM) across various domains in East Asia and over the contiguous United States (CONUS). Progressing from the ResNet (Xing et al., 2020) to UNet-LSTM (Huang et al., 2021; Xing et al., 2022), and most recently to the ConvLSTM (Xing et al., 2024), the model has evolved significantly. It is now able to address long-term error propagation, manage memory demands for multiple VOC/PM species, update emissions efficiently, and integrate near-real-time data from ground-based and satellite observations (Xing et al., 2022). However, our previous work has not fully incorporated vertical information into the system. While vertical profiles have been studied recently for NO_2 (Li and Xing, 2024), they have not yet been explored for other species such as O_3 or $PM_{2.5}$. Additionally, the multiple species in VOC and PM emission profiles have not been adequately considered in previous model designs. Here we propose a new model structure of four-dimensional DeepCTM (DeepCTM4D) to address the above limitations.

2 DATA AND MODEL

2.1 Training data

We leveraged hourly, spatially-resolved input and output data from simulations of the US EPA Community Multiscale Air Quality model (CMAQ) (Appel et al., 2013) over the U.S. 12 km Continental US (CONUS) domain for 2019 and 2020 to train the model.

The feature inputs to DeepCTM4D will mostly remain the same as in previous applications. The emissions data include layer-specific profiles for five species: NOx, VOC, SO_2, NH_3, and $PM_{2.5}$, with plume-rise effects accounted for point sources. The meteorological data including 3D variables of U- and V-winds (UW, VW), water vapor (QV), cloud water mixing ratio (QC), air temperature (TA), mid-layer height above ground (ZH), 3D resolved cloud fraction (CRFAC), and air pressure (PRES), as well as 2D variables crucial for simulating near-surface chemical dispersion and chemistry, including planetary boundary layer height (PBL), 10-meter wind speed (WS), short-wave radiation (SWR), convective velocity scale (WSTAR), 2-meter temperature (T2), humidity (Q2), Leaf Area Index (LAI), and vegetation coverage (VEG). These variables are input at corresponding hourly intervals to ensure temporal alignment with model requirements. Additionally, we incorporate static geographical data, such as terrain height (HT), land-water

mask, and land use category, which will provide spatial context that remains consistent over time. This will support more accurate predictions of air pollutant dispersion in relation to topographical features. The prediction targets six key chemical species: NO_2, formaldehyde (FORM), O_3, SO_2, NH_3, and $PM_{2.5}$. For training, we used data from the 1st to the 25th day of each month, leaving the remainder for testing.

We employed data augmentation by randomly cropping feature maps to a size of 60 rows by 60 columns. This approach prioritizes local spatial dependencies, reflecting the fact that atmospheric processes typically occur within a limited spatial range over a relatively short period of time (i.e., one hour). Random cropping increases sample variability, improves model focus on localized patterns, and significantly reduces memory requirements by avoiding the use of full high-resolution maps. The training process optimizes Mean Squared Error (MSE) as the loss function over 5000 epochs, a duration empirically determined to ensure robust performance on both training and testing datasets. The learning rate initiates at 0.001 and decays linearly to zero by the end of training, supporting stable convergence. We employ the Adam optimizer (Kingma and Ba, 2014) to enhance model convergence.

2.2 The DeepCTM4D Model

To address the spatiotemporal variation of atmospheric chemical concentrations, meteorological, and topographical variables, we used a hybrid model combining 3D-ResNet (Tran et al., 2018) and ConvLSTM (Shi et al., 2015) architectures (Figure 1). Such design aligns closely with the governing physical principles of atmospheric processes, effectively capturing both spatial and temporal dependencies in the data. Particularly, the 3-D CNN leverages surrounding information from both horizontal and vertical neighborhood grid cells, capturing interactions between downwind/upwind and upper/lower areas.

In the ConvLSTM structure, the concentrations predicted in the previous time step serve as initial conditions for forecasting concentrations in the subsequent hour. This approach enables DeepCTM4D to make predictions based on past predictions, introducing a temporal dependency that reflects the chemical variations driven by atmospheric physics (Xing et al., 2024). This design aligns with the continuity of atmospheric processes, which modulate concentrations over short intervals, such as one hour, by considering the two roles of previous-hour concentration (IC) in the model structure:

(1) Perturbation Baseline: modulated based on changes (perturbations) resulting from atmospheric dynamics estimated within the ConvLSTM cell and added to the initial concentration.

(2) Atmospheric Processes Integration: passed into the ConvLSTM cell, where it interacts with meteorology variables, to simulate atmospheric physical and chemical processes.

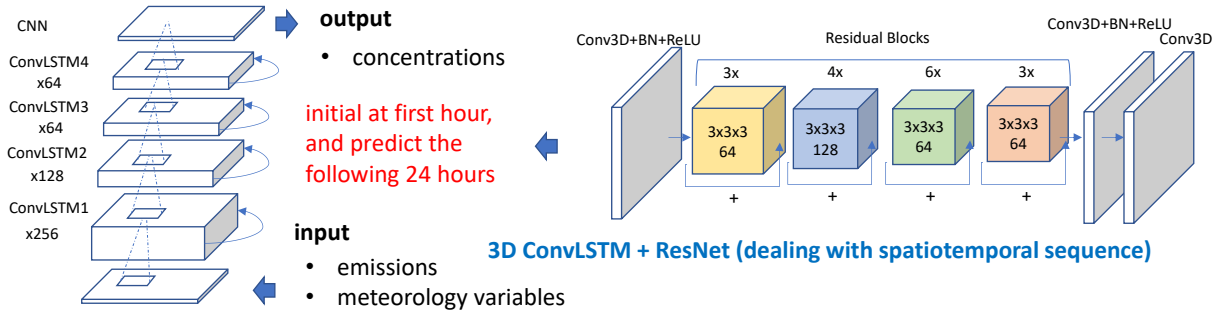
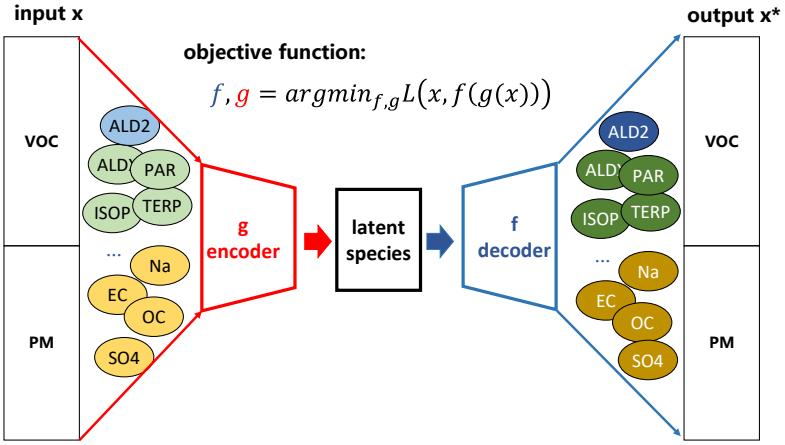


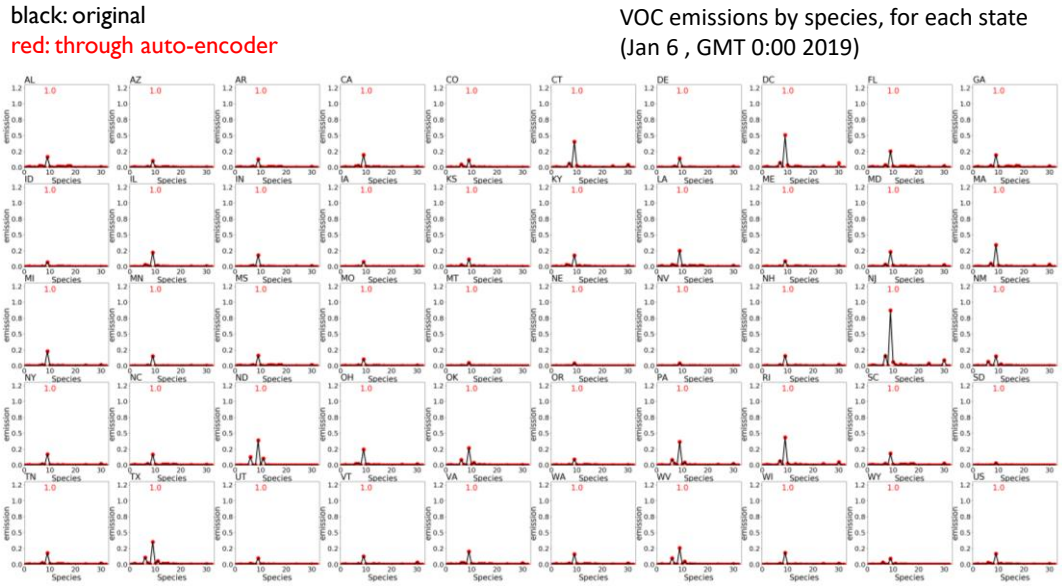
Figure 1. The DeepCTM4D for predicting the following hour's concentration

As a result, the prediction for the upcoming hour is closely linked to the prediction for the previous hour, simplifying the training process and enhancing the model's stability. The hidden layers in ConvLSTM retain historical information, helping to mitigate error propagation during predictions over relatively long periods.

To address the numerous VOC and PM species in their emission profile, we leveraged the encoder-decoder architecture in DeepCTM to effectively reduce the number of VOC species and PM components by compressing them into a limited set of latent variables. The encoder processes the original species, condensing them into a compact latent space for calculating concentrations in the next time step. The decoder then reconstructs the original species from these latent variables in the output (see **Figure 2**). This approach significantly improves training efficiency and reduces the computational burden, enabling the model or CTM to handle chemical species more effectively without processing each individually.



(a) illustrate the auto-encoder structure



(b) performance in decoding the VOC emission profile by states (using 2 latent variables to represent the original 33 VOCs, the red number represents the R^2)

Figure 2. Using Auto-encoder to compress the VOC/PM species number

Figure 2b illustrates an example where two latent species, compressed from the original 33 VOC species, are used to decode back to the original 33 VOC species using an autoencoder model. The well-trained autoencoder model effectively identifies the dominant species in each state and reconstructs the original 33 VOC species with high accuracy, achieving an R^2 value close to 1.

3 RESULTS

3.1 Predicting the surface concentration across time

The performance of DeepCTM4D in predicting surface concentrations of six species is given in Figure 3. Overall, DeepCTM4D effectively captures the temporal variations, achieving strong performance, with RMSE of 0.08 ppb, 0.04 ppb, 0.02 ppb, 0.1 ppb, 0.05 ppb, and 0.04 μg m^{-3} for NO_2, HCHO, O_3, SO_2, NH_3 and $PM_{2.5}$ respectively.

The RMSE is quite similar on both the training and test datasets, for all species demonstrating the model's generalization ability in predicting a different day that is not included in the training dataset.

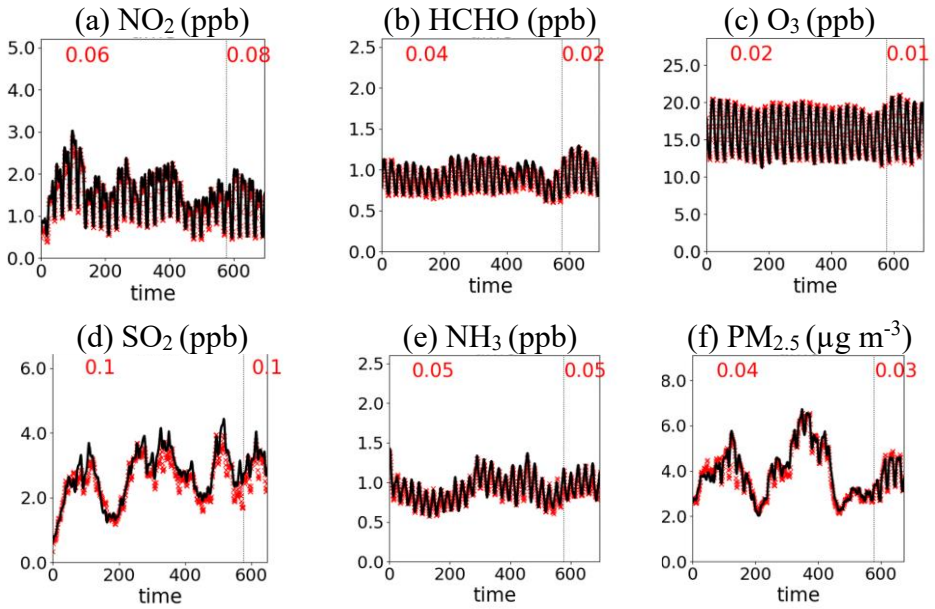


Figure 3. Performance in predicting time series for hourly surface concentration across one month (the first 25 days for training and last 5 days for testing, the red number represents RMSE; reinitialize every day; domain average in January 2019)

3.2 Predicting the surface concentration across horizontal space

Figure 4 compares the DeepCTM4D-predicted spatial distribution of six key species with CMAQ simulations, taking one day as one example (January 6, 2019, for NO_2, SO_2, and $PM_{2.5}$; and July 6, 2019, for HCHO, O_3, and NH_3).

The model is initialized at hour 0 and predicts the following 24 hours. In the first hour, DeepCTM4D provides predictions very close to CMAQ, as the changes within one hour are minimal and primarily influenced by the initial conditions used as input. Over time, DeepCTM4D continues to produce spatial patterns closely resembling those of CMAQ for all species,

demonstrating its capability to capture concentration modulations driven by emissions and meteorological factors.

The R is larger than 0.6, 0.4, 0.7, 0.5, 0.5, and 0.7 for NO_2, HCHO, O_3, SO_2, NH_3, and $PM_{2.5}$ respectively, across the 24-hour prediction.

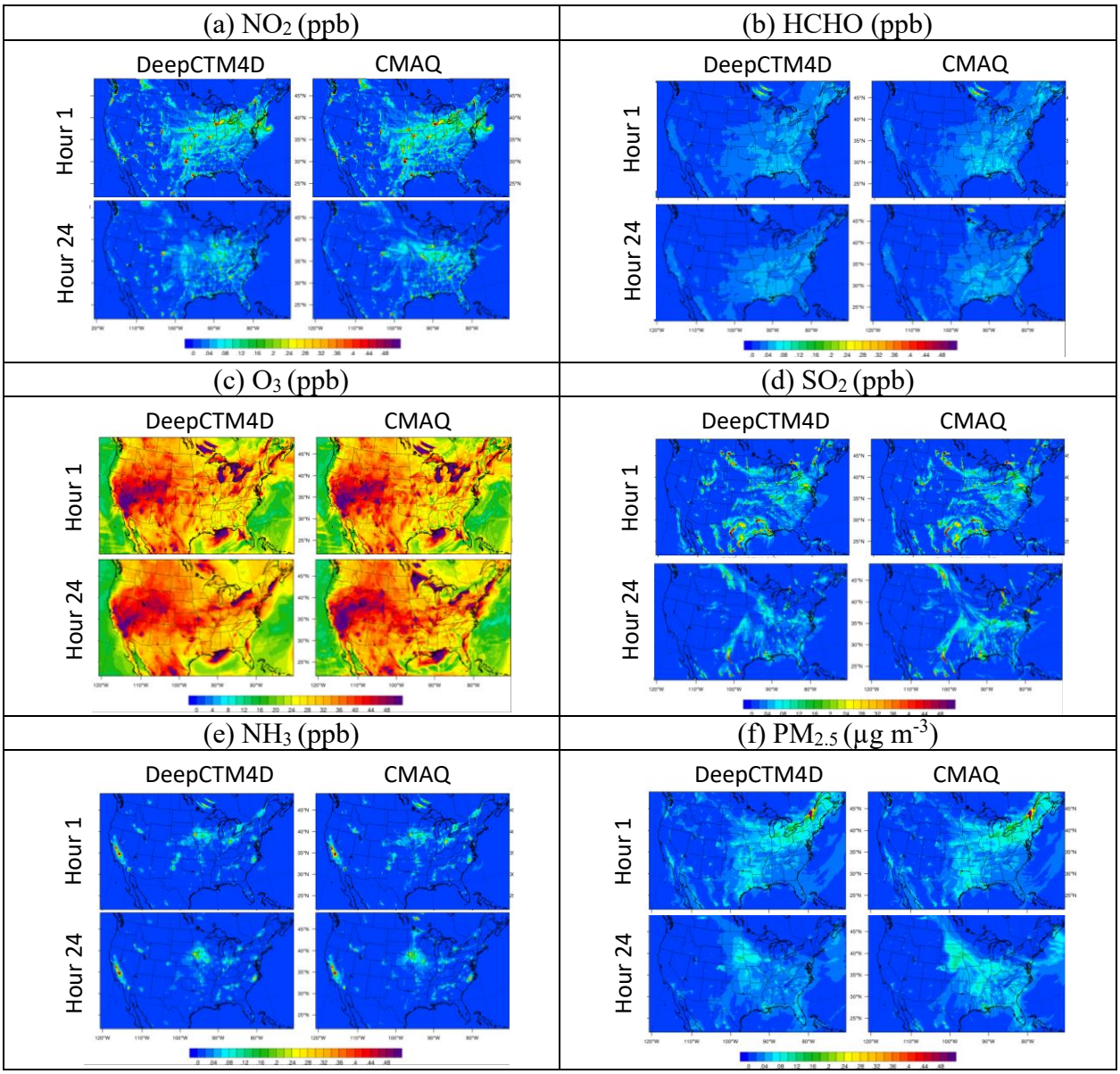


Figure 4. Performance in predicting spatial distribution of surface concentration across a day

3.3 Predicting the vertical profiles

The DeepCTM4D-predicted vertical profiles of NO_2, O_3, and $PM_{2.5}$ were further investigated by comparison with CMAQ. Overall, the model successfully captures the vertical distribution of

these species, particularly the high concentrations near the ground for NO_2 and $PM_{2.5}$, which decrease with height, and the opposite trend is observed for O_3, with a high R^2 value close to 1.

However, the model faces challenges in reproducing the top layers, where larger discrepancies are observed for NO_2 and $PM_{2.5}$. This may require future improvements, such as adding additional constraints on the top layer for these species.

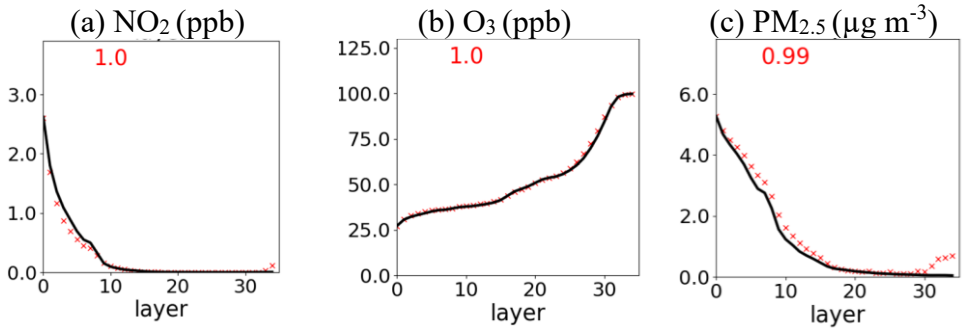


Figure 4. Performance in predicting vertical profiles (January 6 GMT 1:00; the WRF-CMAQ model defines 35 vertical layers, corresponding to approximate heights of 20, 70, 120, 180, 250, 320, 380, 400, 490, 660, 800, 920, 1,140, 1,400, 1,700, 2,100, 2,800, 3,500, 3,900, 4,300, 4,800, 5,200, 5,600, 6,100, 6,700, 7,200, 7,800, 8,400, 9,000, 9,600, 10,000, 11,000, 12,000, 13,500, and 15,000 meters above ground level)

3.4 Identify the contributors to the prediction

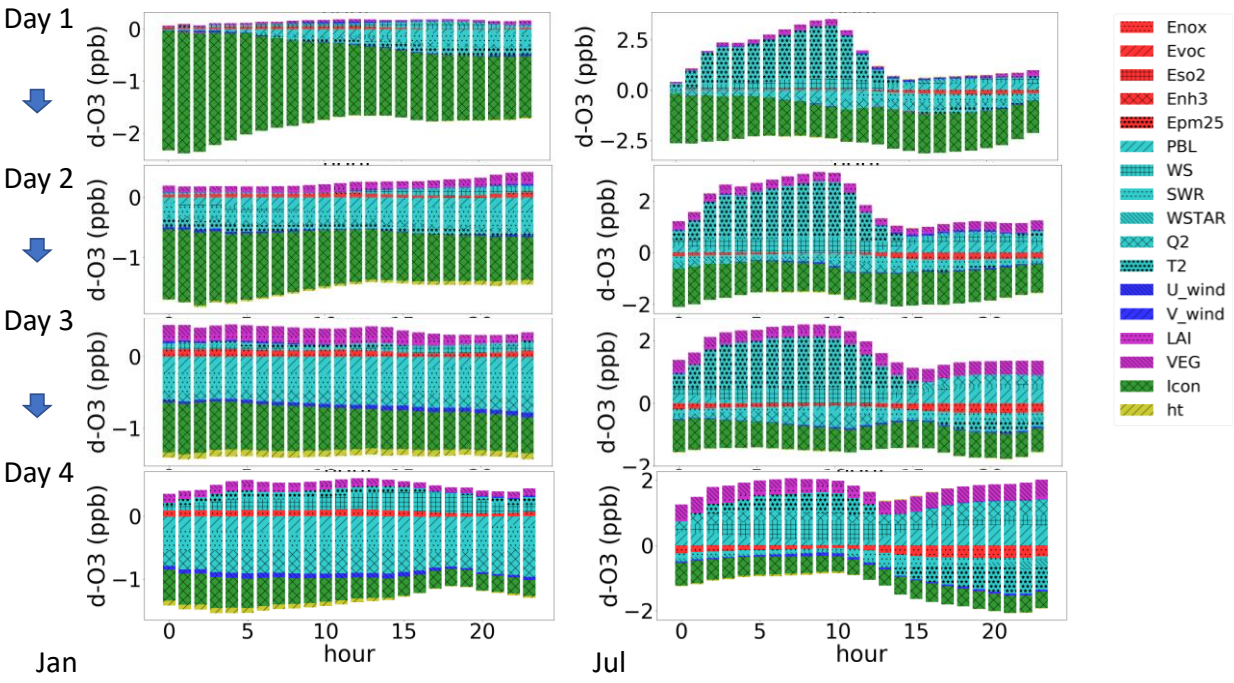


Figure 5. Contributors to the surface O_3 concentration along with prediction time

We further investigated the contribution of each feature to changes in concentrations and its variation over the prediction period. This was done by modulating each feature—reducing it by 20% (reducing T2 by 2°C)—and calculating the difference from the baseline as the feature's contribution. As an example, we present O_3 predictions for one day in January and one in July.

Overall, the initial condition has the largest impact but diminishes over time, as the lifetime of O_3 is long and the concentration in the next hour is similar to the previous one. NOx emissions start to exert a notable influence, with negative impacts on O_3 in winter and positive impacts in summer, reflecting the nonlinear chemistry of O_3 transitioning from VOC-limited to NOx-limited regimes. LAI/VEG has a negative impact due to dry deposition, as increased vegetation enhances the O_3 sink through this process. Meteorological variables show either negative or positive impacts.

These results reveal the complex interactions among emissions, meteorological variables, and concentrations. The DeepCTM4D model effectively captures these dynamics and reproduces variations in O_3 concentrations driven by these factors.

4 CONCLUSION

This study presents a novel method, DeepCTM4D, which utilizes deep learning by integrating a 3D-ResNet with ConvLSTM to emulate CMAQ simulations for four-dimensional air pollution concentrations across space and time. This approach significantly enhances the computational efficiency of atmospheric chemistry modeling. An application over the CONUS domain demonstrates that DeepCTM4D can accurately reproduce CMAQ simulations for six key species, capturing variations across time, space, and vertical profiles. These advancements position DeepCTM4D as a robust tool for operational forecasting systems, improving atmospheric chemistry predictions, weather feedback, and overall modeling efficiency and accuracy.

Future applications of DeepCTM4D could involve coupling with meteorological weather forecasting to enhance air quality and weather predictions across short-term, sub-seasonal, and long-term scales. Additionally, by integrating observational data from satellites and ground-based measurements, the model could deliver accurate and efficient predictions to support targeted strategies for addressing air pollution and climate challenges at local scales. This capability would contribute to global efforts to mitigate climate change, improve environmental health, and foster sustainable development. In addition to the conclusions above, GPU and memory resources are crucial for enhancing the efficiency and accuracy of weather and air quality forecasts, especially when handling large datasets during training.

Acknowledgment.
This work was partly supported by Microsoft Climate Research Initiative Program, and the Bellagio Center Residency Program by the Rockefeller Foundation.

REFERENCE

Appel, K. W. et al., (2013) Evaluation of dust and trace metal estimates from the Community Multiscale Air Quality (CMAQ) model version 5.0. Geoscientific Model Development, 6, (4), 883-899.

Cabaneros, S.M.S. et al., (2019) A review of artificial neural network models for ambient air pollution prediction. Environmental Modelling & Software, 119, 285-304

Huang, L. et al., (2021) Exploring deep learning for air pollutant emission estimation, Geosci. Model Dev., 14, 4641–4654

Kelp, M. M. et al., (2020) Toward Stable, General Machine-Learned Models of the Atmospheric Chemical System. Journal of Geophysical Research: Atmospheres, 125(23), e2020JD032759.

Kingma, D. P. and Ba, J. (2014) Adam: A method for stochastic optimization. arXiv preprint, arXiv:1412.6980

Li, S. and Xing, J. (2024) DeepSAT4D: Deep learning empowers four-dimensional atmospheric chemical concentration and emission retrieval from satellite. The Innovation Geoscience, 2024, 2(1), 100061-1

Shi, X.; Chen, Z.; Wang, H.; Yeung, D. Y.; Wong, W. K.; Woo, W. C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems, 2015, 28.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y. and Paluri, M., 2018. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 6450-6459).

Xing, J. et al., (2020) Deep learning for prediction of the air quality response to emission changes. Environmental science & technology, 54(14), 8589-8600.

Xing, J. et al., (2022) Mimicking atmospheric photochemical modeling with a deep neural network, Atmospheric Research, DOI: 10.1016/j.atmosres.2021.105919

Xing, J. et al., (2024) A physically constrained deep-learning fusion method for estimating surface NO_2 concentration from satellite and ground monitors, Environmental Science & Technology, https://doi.org/10.1021/acs.est.4c07341.

Dr. Jia Xing has over 18 years of research experience in atmospheric chemistry and physics modeling, with a focus on understanding the sources and distributions of atmospheric chemicals and their impacts on human health, weather, climate change, and ecosystems. He developed the machine-learning-based reduced-form atmospheric chemistry model including DeepRSM, DeepCTM, DeepSAT4D, DeepMMF for global applications in air quality forecasting and management. He also made contributions to advancing atmospheric chemistry modeling, particularly in refining key chemical mechanisms such as stratosphere-troposphere exchange, HONO chemistry, and sulfate heterogeneous reactions. He has authored more than 200 peer-reviewed papers in leading journals such as *Proceedings of the National Academy of Sciences, Nature Sustainability, and Nature Geoscience,* with over 10,000 citations in google scholar.

Dr. Joshua S. Fu is the Chancellor's Professor, John D. Tickle Professor, and James. G. Gibson Professor in the Department of Civil and Environmental Engineering and the Inaugural Professor of the UT-ORNL Bredesen Center for Interdisciplinary Research and Graduate Education, Joint Appointment Professor in the integrated Computational Earth Sciences Group in Computational Sciences and Engineering Division at Oak Ridge National Laboratory. Fu has served as vice-chair of the Measurement-Model Fusion for the Global Total Atmospheric Deposition (MMF-GTAD) of the new initiative and Task Team on Artificial Intelligence (AI) for Weather Forecast (AI4WX) in the World Meteorological Organization (WMO), contributed as a co-author of the Final Report of the Hemispheric Transport of Air Pollution for the United Nations Economic Commission for Europe (UN Task Force Hemispheric Transport of Air Pollution) and reviewing committee member for air quality status in East Asia for the EANET, a governmental consortium in East Asia and located in Japan. Fu has received numerous awards from national and international associations and is a recipient of the Fellow of AAAS and A&WMA and Board-Certified Environmental Engineering Member from AAEES and other distinguished and endowed professorship awards. Fu has published more than 200 referred journal articles including PNAS and Nature journals, and 120 peer-reviewed conference proceedings. He has been interviewed or reported on the Conversation, Wall Street Journal, New York Times, Science Daily, @EurekAlert!, Red Orbit, and more than 30 media to discuss issues of international and national importance on weather/extreme events, (e.g. wildfires) and air quality on human health and agriculture. Fu is developing the integration of Agriculture, Earth Systems Science, Business, Engineering, and Human Health with AI.