Enhancing Air Quality Forecast Capabilities with Interpretable Machine Learning for Model Acceleration, Ensemble Prediction, and Satellite Data Assimilation

PI: Christopher Tessum, Department of Civil and Environmental Engineering, University of Illinois Urbana-Champaign

To be considered for submission to NOAA FY2025 Weather Program Office Research Programs 11.459: Weather and Air Quality Research, Air Quality Research and Forecasting, in response to priorities AQRF-4 (model ensembles), AQRF-5 (remote sensing), and AQRF-7 (computational efficiency).

Planned products/outputs

The proposed project will result in the following three main products:

1. Machine-learned operators for gas-phase chemistry and aerosol processes based on multiple reference models, implemented and validated in the CMAQ chemical transport model (CTM).
2. Ensembles of the above machine-learned mechanisms implemented in CMAQ to allow probabilistic air quality forecasts.
3. Demonstrations of TEMPO satellite data assimilation using the above CMAQ ensemble models.

(We could use a different CTM base for implementation, such as UFS-Chem or MUSICA, if preferred by NOAA.)

Planned impacts/benefits/outcomes

The proposed project will initially result in demonstrations in CMAQ of machine-learned gas-phase chemical mechanisms which are orders-of-magnitude faster than the reference mechanisms with a factor of ~10 fewer state variables, resulting in an expected initial ~2–5× speed-up of the overall CMAQ model, with similar surrogate models for aerosol processes subsequently developed and implemented to yield further speedups. These fast, simple individual models will allow ensemble simulations to generate probabilistic forecasts in CMAQ, and these ensemble simulations will in turn allow data assimilation through ensemble Kalman filtering and inversion. Together, these advancements will result in a substantial advancement in NOAA air quality forecasting capabilities.

Planned methodology and timelines

Project activities will be organized into three Thrusts, each leading to one of the Products above.

Thrust 1. In preliminary work we have developed a method for the data-driven generation reduced-order chemical mechanisms which combines a linear autoencoder for "lumping" the chemical species with a novel method we term SIMADy (Sparse Identification of

Mass Action Dynamics) which identifies chemical reactions between the encoded chemical species to reproduce the dynamics found in the training data. Our method is unique among similar approaches in that it offers provable guarantees of numerical stability and scalar positivity. We have successfully used this method to emulate MCM (the Master Chemical Mechanism) and the GEOS-Chem chemical mechanism, and have implemented it online in GEOS-Chem. In this thrust we will also apply this method to CRACMM, implement the resulting mechanisms in CMAQ, and extensively evaluate the result. We will implement a similar process for emulating aerosol operators including ISORROPIA, MOSAIC, CAMP, VBS, with modal and sectional size schemes. [Timeline: Year 1: Gas-phase chemistry surrogate implemented in CMAQ; Year 2: Aerosol models developed; Year 3: Aerosol models implemented in CMAQ].

Thrust 2. In preliminary work (https://arxiv.org/abs/2407.09757) we have demonstrated the generation of ensemble predictions using machine-learned chemical mechanisms. In this Thrust we will apply similar methods to the mechanisms developed in Thrust 1, using the multiple reference models as well as Stochastic Gradient Langevin Dynamics or bootstrapping as sources of variability among the ensemble members. We will evaluate the resulting ensemble predictions in CMAQ in terms of accuracy, uncertainty calibration, and performance-accuracy tradeoffs. [Timeline: Year 1: CMAQ ensembles for CONUS demonstrated at 36km resolution; Year 2: Ensembles demonstrated at 12km resolution; Year 3: Ensembles demonstrated at 4km resolution].

Thrust 3. We will leverage the ensemble simulations developed in Thrust 2 to assimilate NASA TEMPO satellite data using Ensemble Kalman Filter and Ensemble Kalman Inversion techniques, both individually and combined in a joint state-parameter estimation framework. [Timeline: Year 1: Data assimilation for chemical state; Year 2: Data assimilation for updating model parameters; Year 3: Joint state-parameter updates].

During the above activities we will leverage the EarthSciML geoscientific modeling framework (https://earthsci.dev/) for rapid prototyping and development, which is being developed by my research group with funding from NASA and NSF. The EarthSciML framework is currently too experimental for operational use but offers next-generation capabilities including equations-to-code compilation and automatic differentiation which allow a rapid-experimentation workflow that we have found to substantially accelerate the development of machine-learned model components. After performing initial development and testing using EarthSciML, we will port each model component to CMAQ (or another model framework if preferred by NOAA) for further evaluation and eventual operational deployment.

The proposed starting and ending Readiness Levels (RLs), any proposed use of NOAA Testbeds, HPC resources, and whether a NOAA Transition Plan has been developed for earlier work on this topic

The machine learned chemical mechanism framework (SIMADy) described in Thrust 1 is at Readiness Level 6: we have demonstrated its use in the GEOS-Chem CTM. The ensemble model generation system described in Thrust 2 is at a Readiness Level 5: we have demonstrated its use in a box-model format. We expect that the data assimilation framework described in Thrust 3 will be at a Readiness Level of 6 before December 2: we have all the necessary parts ready and are currently working on producing an Ensemble Kalman Inversion demonstration using TEMPO data and the EarthSciML model framework. At the end of the project, all components will be at least at a Readiness Level of 7: implemented in CMAQ with performance evaluated. We will also work with NOAA staff to achieve levels 8 and 9 with operational use in forecast products. No use of NOAA testbeds or HPC resources is planned, and no NOAA transition plan exists.

Potential operational, commercial, or other end-user adopter(s) of the project outputs

The main target user is the NOAA National Air Quality Forecast Capability (NAQFC). However, the capabilities we propose to develop are general, and could also be used by the NASA GEOS-CF composition forecast as well as by the regulatory and research communities in CTMs such as CMAQ, GEOS-Chem, and WRF-Chem.

Budget

The proposed project will support three graduate student research assistants (one per Thrust), 2–4 months of effort per year of a Research Scientist to provide software and computational support, and one month per year of summer salary for the PI, plus miscellaneous expenses.

	Year 1	**Year 2**	**Year 3**
Grad Student Salary	91,672	105,698	109,397
Research Scientist	29,187	16,068	13,970
PI Summer Salary	13,717	14,197	14,694
Fringe	29,387	24,977	24,617
Grad Student Tuition	58,670	67,647	70,014
Travel / Equipment / Misc.	19,095	16,965	13,256
Indirect Costs	107,272	104,252	103,097
Total	349,000	349,804	349,045