Introducing CREDIT-chem: A Hybrid Physics–Machine Learning Framework for Modeling Atmospheric Aerosol Chemistry

James "LJ" Dunphy[1,2]

Mentors: David John "DJ" Gagne[2], Charlie Becker[2], Alma Roux[3], Ben Gaubert[3], John Schreck[2]

[1]Florida State University, [2]NCAR Computational and Information Systems Lab, [3]NCAR Atmospheric Chemistry Observations & Modeling Lab

Background

The Importance of Atmospheric Chemistry

Atmospheric aerosols are **central** to both **regional air quality** and **global climate forcings**. To create accurate earth systems projections, **models** must **simulate their life cycles** and **chemical reactions**, including radiative effects, cloud microphysics, and aerosol aging.

The Challenges & CREDIT-chem's advancements

- **Physical Solvers:** Trapped between **either high accuracy** (extremely expensive) **or fast approximations** (physical oversimplifications)
- **Existing Machine Learning (ML):** **Very fast**, but **not physically conservative** and **doesn't scale well** with more chemical species or higher resolution
- **CREDIT-chem:** Our framework aims to use **ML's speed advantages**, while maintaining **accuracy, scalability, and physical conservation**

Input and Target Data Setup

CAM6-chem Target Dataset

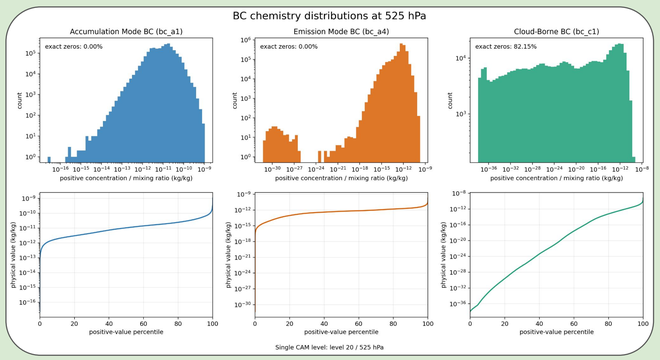
- **Data Resolution:** ~1° global resolution, 32 vertical levels, 1-hour time-step
- **Target Species:** **Emphasis on Black Carbon (BC)**; includes simulations for primary organic matter **(POM)**, secondary organic aerosols **(SOA)**, and **sulfates**; with given **meteorological inputs**

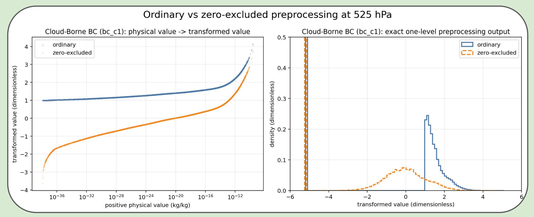
Quantile Transform – Zero Exclusion method

Chemical aerosol concentrations are **heavily non-normal**, **span multiple orders of magnitude**, and have **heavy zero-inflation**. Standard ML models **require normally distributed inputs** centered near 0 ($\mu \approx 0$, $\sigma^2 \approx 1$). To combat this, we developed the following transform method:

1. Empirically **fit a CDF** to exclusively all **non-zero aerosol concentrations**
2. Apply a **Logit transform to the CDF**, bounded to [-5, 5], to transform from a uniform distribution to a **normal distribution**
3. Exact **0 inputs** are **encoded as -5.25**; **outputs <-5.0** are **decoded to 0**

Methods - Data Distributions





Methods – Model Architecture

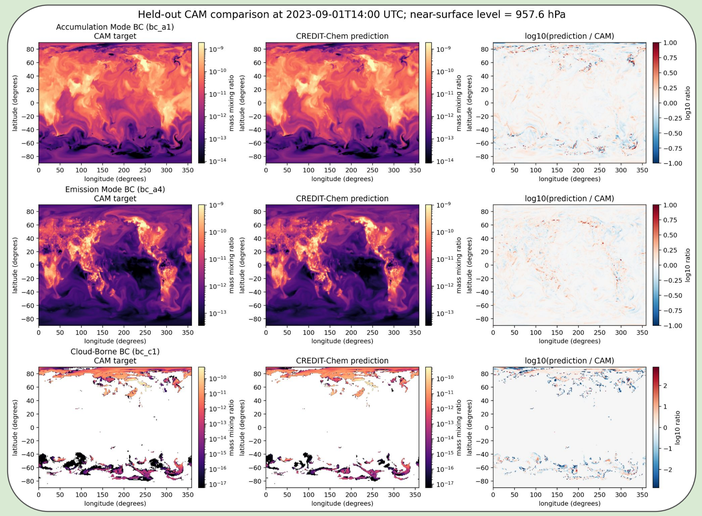
Hybrid Physical Transport–Columnar ML Prediction Scheme

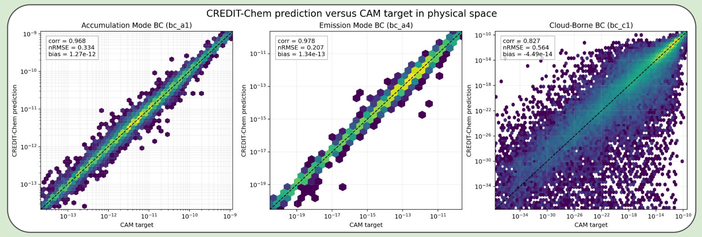
- **State Variables Update:** Initial **aerosol concentrations** and **meteorological variables** are **processed by a meteorological model** to yield updated state variables for the next time step
- **Physical Transport:** **Aerosol tracers** are **physically transported** across the global grid using a advection/convection scheme
- **Columnar ML Emulation:** The transported aerosols and updated state variables are **passed into the ML model,** which evaluates **each vertical column independently and in parallel** to predict local chemistry.

Why Columnar Predictions

- **Physical Constraints:** **Decoupling** aerosol transport with chemical processes allows for **explicit physical constraints**
- **Scalability:** Increasing grid resolution scales **inference time linearly**, while the **per column compute remains constant**
- **Model Capacity:** Processing a single column at a time allows for scaling to **much higher capacity models** while staying **fast at inference time**
- **Sample Richness:** Treating each column as a separate forecast allows for **much better generalization** compared to global models

Results





Conclusions

CREDIT-chem's Key Advantages

- **High Accuracy & Speed:** **Matches CAM6-chem outputs** with high correlation while **accelerating inference time** by multiple orders of magnitude.
- **Long-Term Stability:** Preliminary multi-step rollouts demonstrate **strong stability over extended time horizons**, validating its use for both short-term forecasting and long-term climate runs.

Limitations/Future Work

- **Comparative Performance:** Fully end-to-end ML models (e.g. CREDIT-chem cubed) **outperform the hybrid setup** under specific conditions, though they remain **bound by spatial resolution** scaling constraints.
- **High-Resolution Testing:** While the column-wise design is **theoretically built to scale**, testing accuracy and performance at sub-degree global resolutions **remains ongoing work.**
- **Transport Bottlenecks:** Framework runtime remains tightly **bounded by the efficiency of the physical transport** solver, especially at higher resolutions

Acknowledgments

I would like to first thank all my mentors: DJ, Charlie, Alma, Ben, and John, for their invaluable guidance and support throughout the development of CREDIT-chem. I would also like to thank NCAR and the SIParCS program for giving me the opportunity to conduct this research, as well as providing the computational resources essential to completing this work. Lastly, I would like to thank everyone who provided feedback and insight throughout this project, shaping both this framework and my trajectory as a researcher.

This material is based upon work supported by the U.S. National Science Foundation under Grant No. RISE-2019758. "Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. National Science Foundation."