The development of a machine learning chemistry emulation package for use with GEOS-Chem

Peter Ivatt[1,2]*, Julie Nicely[1,2], Christoph Keller[3], Melanie Follette-Cook[1,4] and Daniel Tong[4]

*peter.d.Ivatt@nasa.gov (primary email)

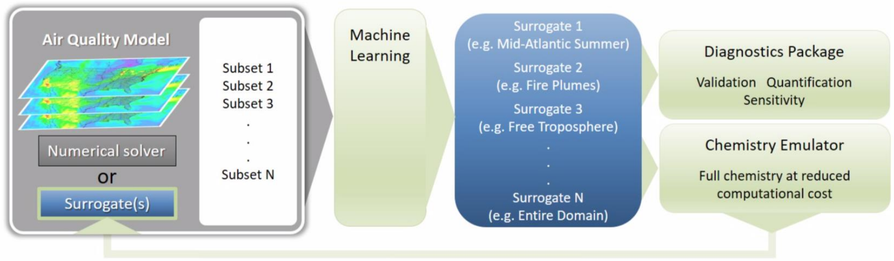
1. NASA Goddard
2. University of Maryland
3. Morgan State University
4. George Mason University

Introduction

The chemical solver in chemical transport models (CTM)s is often the most time-consuming component. Emulation via machine learning offers a computationally inexpensive alternative method to simulate atmospheric chemistry.

Here we are developing a "model agnostic" user friendly machine learning package to generate chemical mechanism emulators (e.g., Carbon Bond 6, GEOS-Chem) using gradient boosted regression for use in CTMs.

Surrogate Models

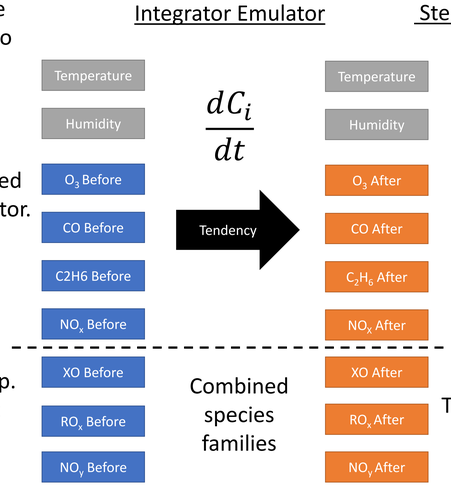


Framework

- A reduced variable input (compared to using all species output), results in less error being carried forward.
- Less data is required to train the emulator.

Species included in this stage are:

- Chemical species with lifetimes greater than the chemistry timestep.
- Species important for deposition.
- Families of quick cycling species.



- Time resolution independent.
- Can be trained using a box model to go beyond GEOS-Chem's chemical resolution.
- Only turn on emulators for output that is required.

dmlc XGBoost

Training was performed on 80 days of 1° resolution hourly chemistry output for all species using standard chemistry over the Eastern US. Testing was on 10 days.

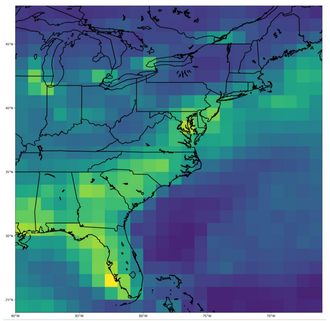
- The CTM can then be programmed to use a different scheme depending on the chemistry present. This will initially be done with simple geofencing but will advance to identifying regimes based on chemical fingerprints or secondary machine learning algorithms.
- Once samples are allocated into regions. These subsets can be trained on different chemistry schemes depending on regime present. Allowing more complex region-specific "surrogate" schemes to be learnt.

- The mosaic of complex schemes could allow the CTM to go above and beyond the performance using conventional integration. Which could be particularly useful for operational models and ensemble runs.

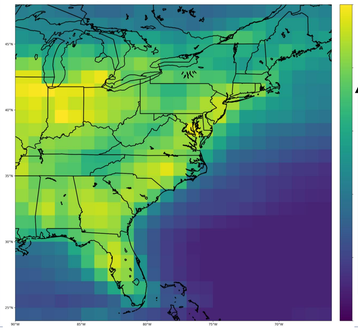
Information Entropy Sampling

- Full training dataset is extremely large (131 Tb per year at 1° resolution and thus requires subsampling.
- As there is spacial variance in the degree of chemical complexity a way of quantify complexity can improve the training dataset.

Acetone surface entropy



O_3 surface entropy



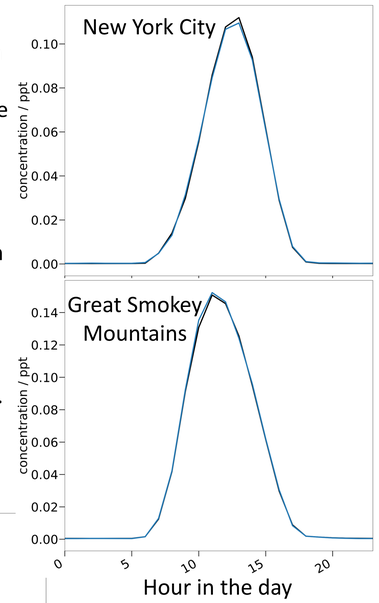
Here sample entropy is used as a metric to quantify "surprise" in the time series in each grid box. High surprise = more complexity.

Initial results find a 10 - 15% RMSE decrease in some species.

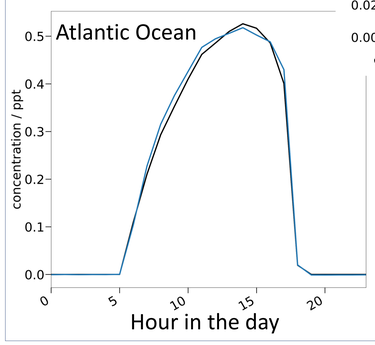
OH Steady state emulator

- The steady state emulator is very effective at predicting the OH concentration within the model.
- Across the whole domain the NRMSE of OH is approximately 3%.
- As the emulator is using the integrator results of the long-lived species to train on as well as physical parameters. A box model could be used to generate OH predictions using more complex chemistry schemes. Such as forest fire specific schemes.

OH emulation



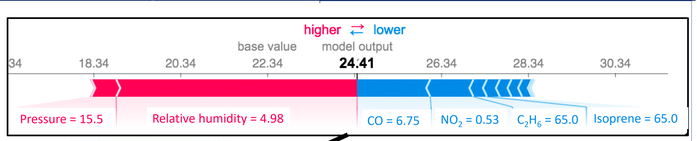
BrO emulation



Halogen species benefit from the use of steady state emulation. The BrX family was used in the integrator phase with Br, Br_2, and BrO predicted in the steady state phase.

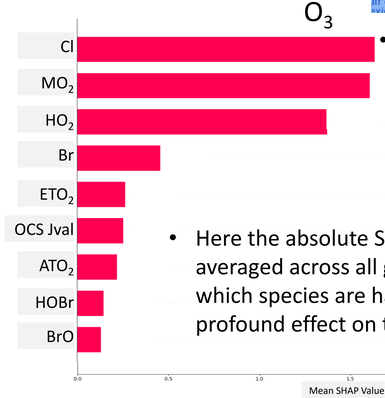
Diagnostics

- Shapley additive explanations (SHAP values) allow insight into how the algorithm is predicting a result. XGBoost produces a base value (mean value) the trees then increase or decrease this value based on the input features.



- Individual grid boxes can be selected to identify what governs the prediction at specific locations.

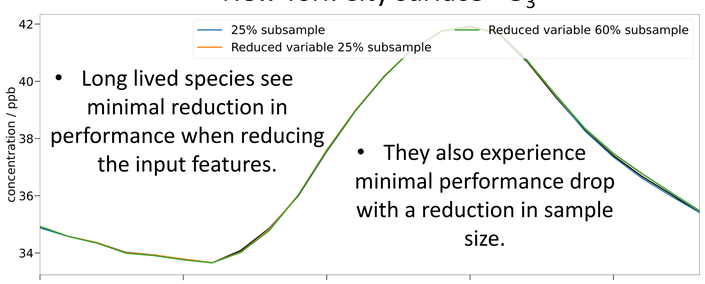
O_3



- Using these values will help identify what species are required for the integrator stage of the prediction.
- Here the absolute SHAP values is averaged across all grid boxes to identify which species are having the most profound effect on the O_3 prediction.

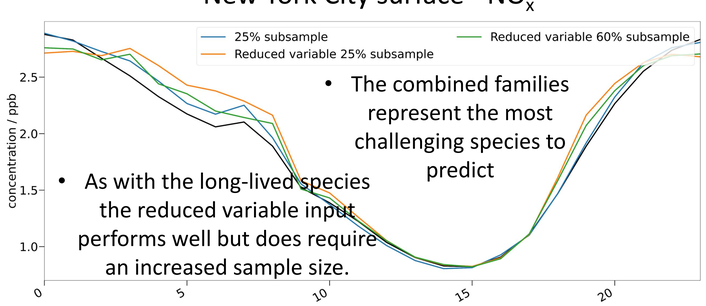
Integrator Emulator

New York City surface - O_3



- Long lived species see minimal reduction in performance when reducing the input features.
- They also experience minimal performance drop with a reduction in sample size.

New York City surface - NO_x



- The combined families represent the most challenging species to predict
- As with the long-lived species the reduced variable input performs well but does require an increased sample size.

- Long lived species have very low error in concentration.

Species	NMRSE (%)
O_3	0.13
CO	0.01
Acetone	0.07
C_2H_6	0.02
Benzene	0.05

We acknowledge funding from the EPA STAR Program via Grant #84001101.