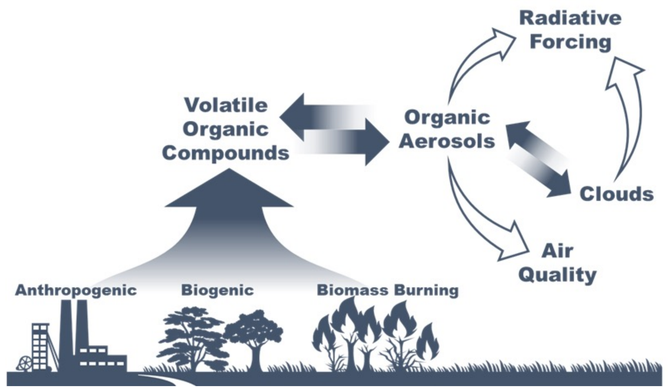
Machine Learning Data Commons Web Portal Tutorials

Omar Chaarawi, David John Gagne, John Schreck
SIParCS, NCAR, Boulder, CO 80305

Motivation

- ❑ Previous machine learning tutorials have been developed
- ❑ No introductory level tutorials that make use of GECKO-A data
- ❑ GECKO represents a real-world problem that could be solved using ML
- ❑ Climate scientist what to use ML techniques, but may be hindered or discouraged by a knowledge gap
- ❑ Tutorials fall in line with NCAR's mission to support the scientific community

GECKO-A



- ❑ GECKO-A is a hyper-explicit mechanism for determining quantities of chemical precursor present in the atmosphere
- ❑ Quantities of chemical precursors play a role in climate through both direct and indirect radiation effects
- ❑ Accurate predictions of aerosolized quantities of VOCs might improve climate prediction models, though GECKO-A is far too computationally expensive
- ❑ ML Techniques could be used to create an emulator that would generate data close enough to explicit models that it would be useful in developing climate models

Goal: Combine principles of hackathon notebook and AMS ML Short course to develop introductory ML courses that utilize GECKO-A data

Notebook Layout

- ❑ Notebook 1: Gathering Data | Preparing Data
- ❑ Notebook 2: Choosing a model | Training that model using our data
- ❑ Notebook 3: Evaluating that model | Tuning parameters
- ❑ Notebook 4: Predict
- ❑ Data hosted on AWS

Concepts explained

Visualization and PCA

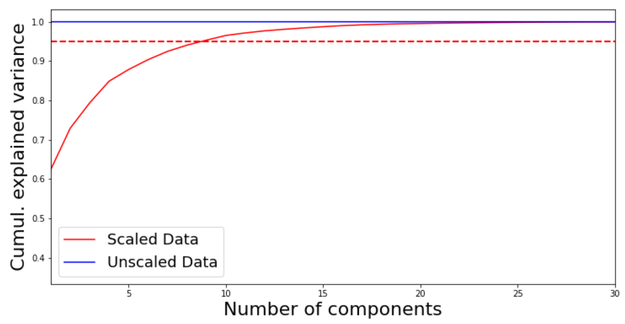


Figure 1. One solution to reducing the number of bins used for gas and aerosol is PCA. We can explain about 95% of the cumulative explained variance ratio using only about 10 of the 35 components. This has not been tried yet and serves as a source of inspiration

Limits of linear regression

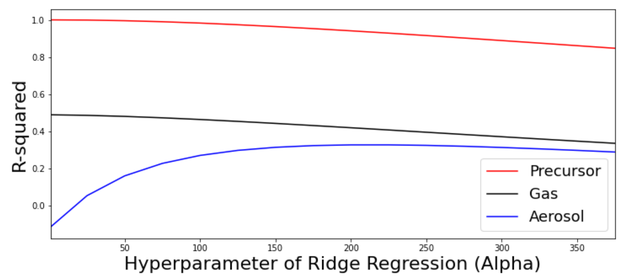


Figure 2. We can increase the accuracy of our predictions by varying the hyperparameter, but a linear model is limited and may not be the best model for this data.

Concepts explained cont.

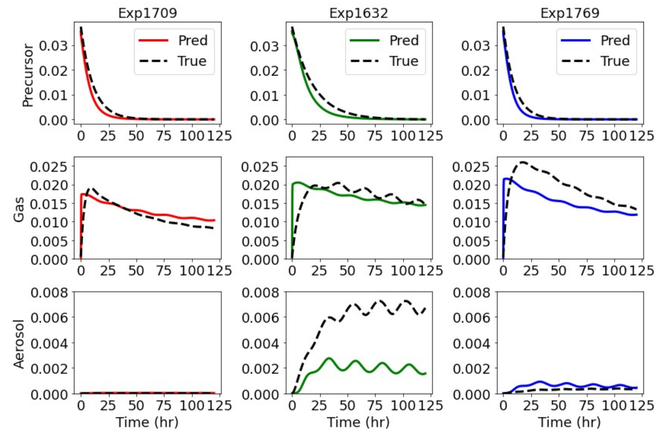


Figure 3. The prediction for aerosol data is improved at a small cost to precursor and gas and the tutorial illustrates the limitations of linear models

Future Directions

Finish and release tutorial
Tutorials will be released as a self-guided set of Jupyter Notebook tutorials with publicly available GECKO-A data.

Google Collab
Tutorials developed in Jupyter will be migrated to Google Collab and tested in that environment.

Test in an off-premises environment
The tutorials have so far been developed in NCAR's Casper supercomputer. Testing needs to be done to insure the tutorials remain accessible in a local environment.

Discontinuation of the "Iris" Dataset
The popular dataset is problematic because of historical reasons and does not make for an appropriate dataset for a tutorial. Examples using this dataset will be removed and replaced.

Acknowledgements

I would like to thank my mentors, David John Gagne and John Schreck, for their guidance throughout this project. Thank you Charlie Becker for being there to help answer my several questions. Finally I would like to thank NCAR for the opportunity to work on this project as a SIParCS intern.