Machine Learning Data Commons Web Portal

Developing a ML tutorial using a GECKO-A dataset

Omar Chaarawi,
CU Boulder & SIParCS

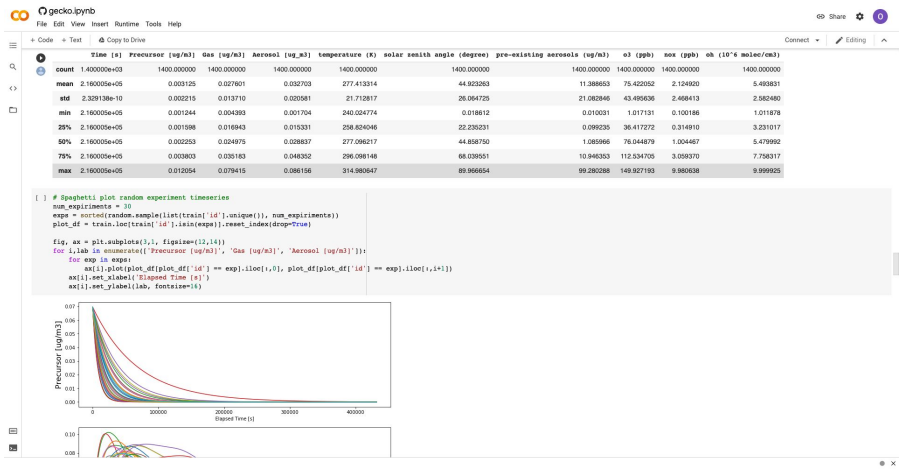
July 27, 2021

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

Motivation

Previous tutorials developed:

- Artificial Intelligence for Earth Systems Science (AI4ESS) Summer School
- AMS ML Short Course



AI4ESS Hackathon made use of GECKO-A data, but is not introductory level course

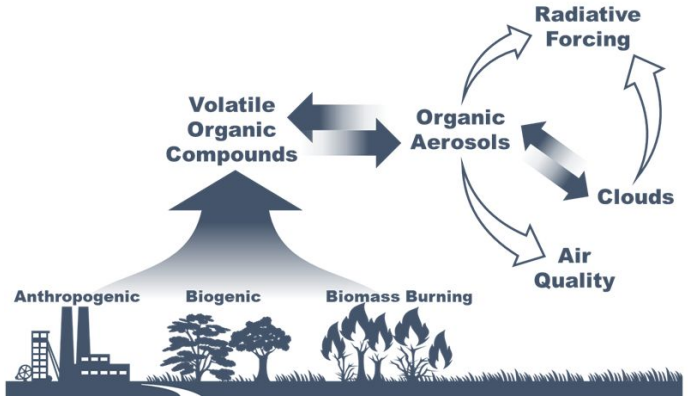
Goal: Combine principles of hackathon notebook and AMS ML Short course to develop introductory ML courses that utilize GECKO-A data

GECKO-A Data

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

Suggested Layout for ML Tutorials

How would a ML scientist approach the GECKO-A emulator problem?

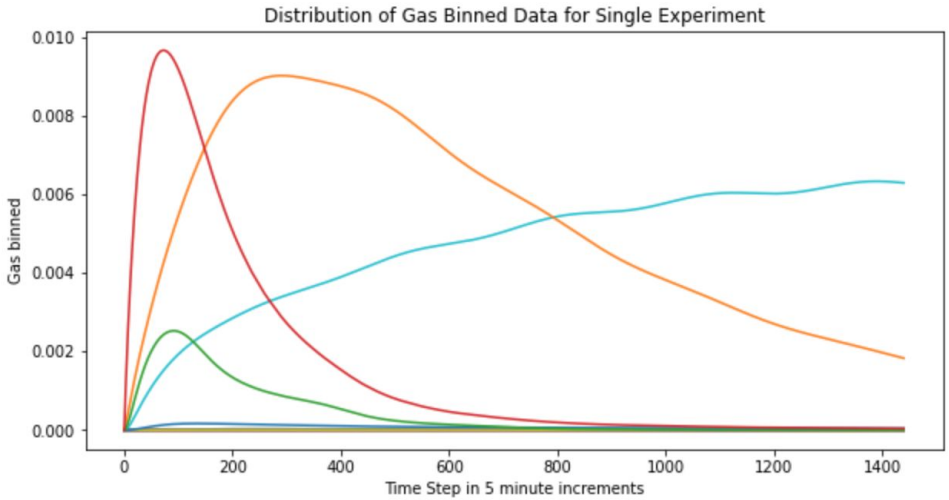
Notebook 1:
- Gathering Data
- Preparing Data

Notebook 2:
- Choosing a model
- Training that model using our data

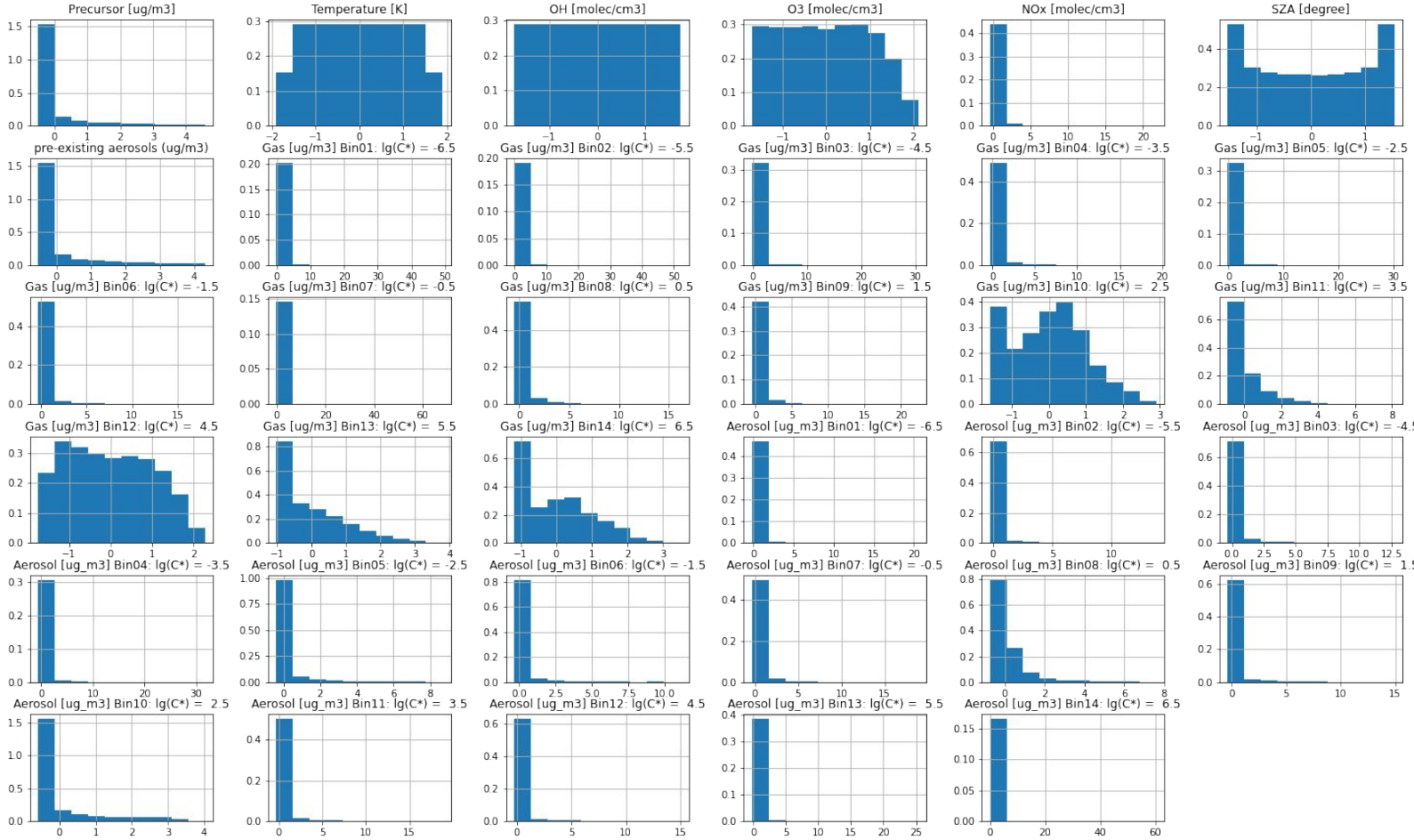
Answering initial questions like: how are the data for each feature distributed?

Notebook 3:
- Evaluating that model
- Tuning parameters

Notebook 4:
- Predict

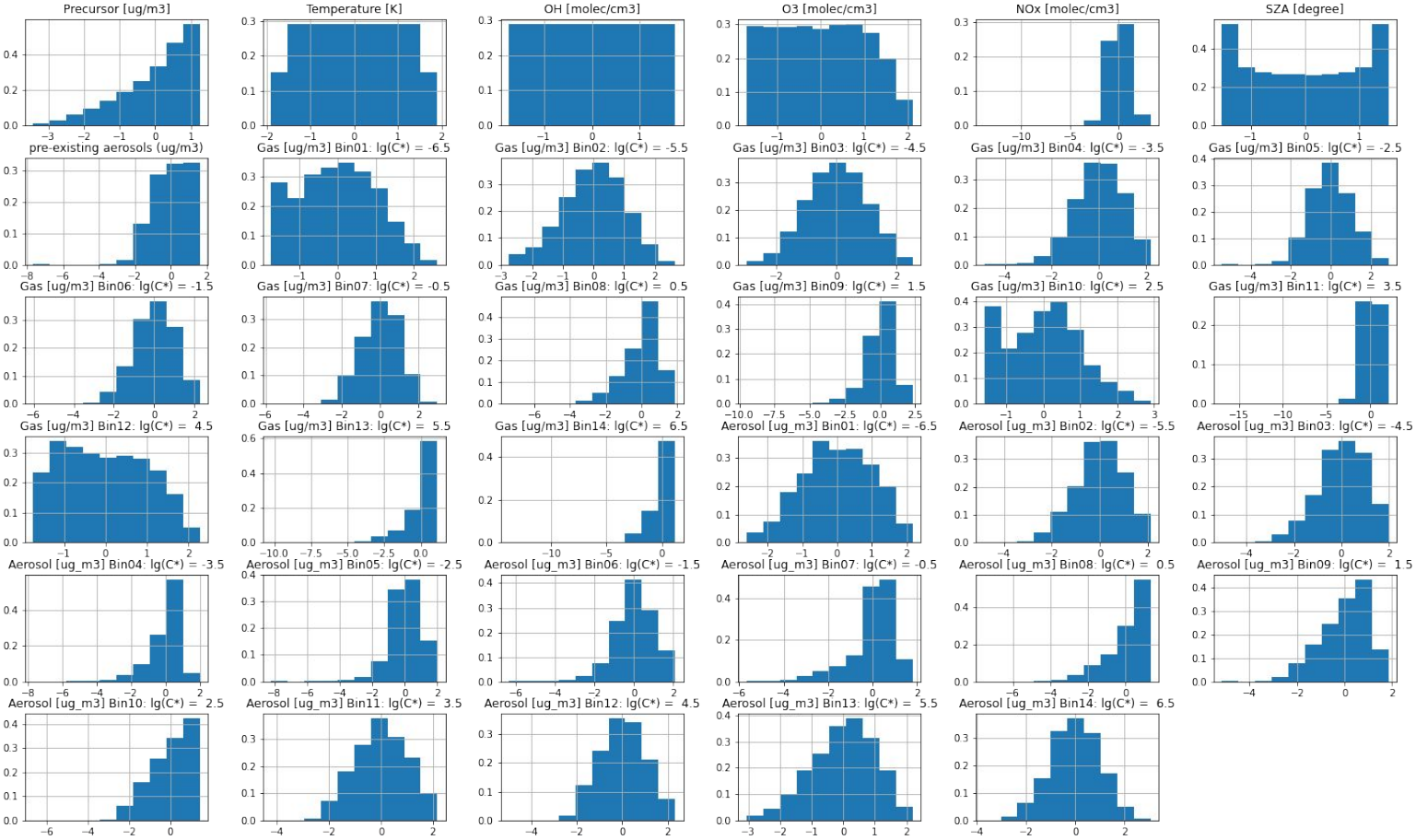


Data Distributions



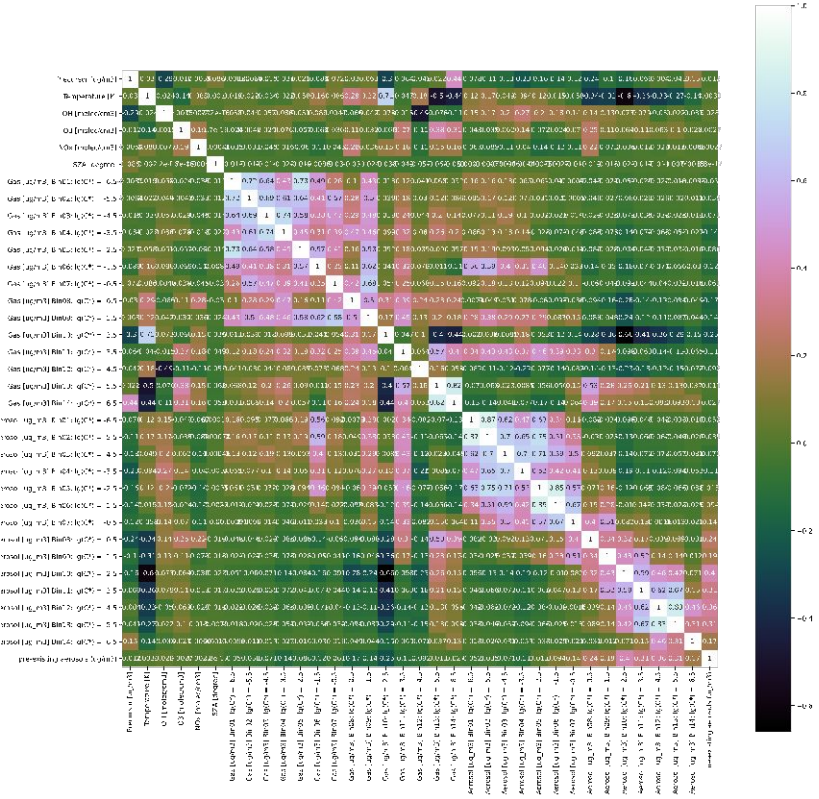
Tools of the trade are used to better distribute the data

Data Distributions

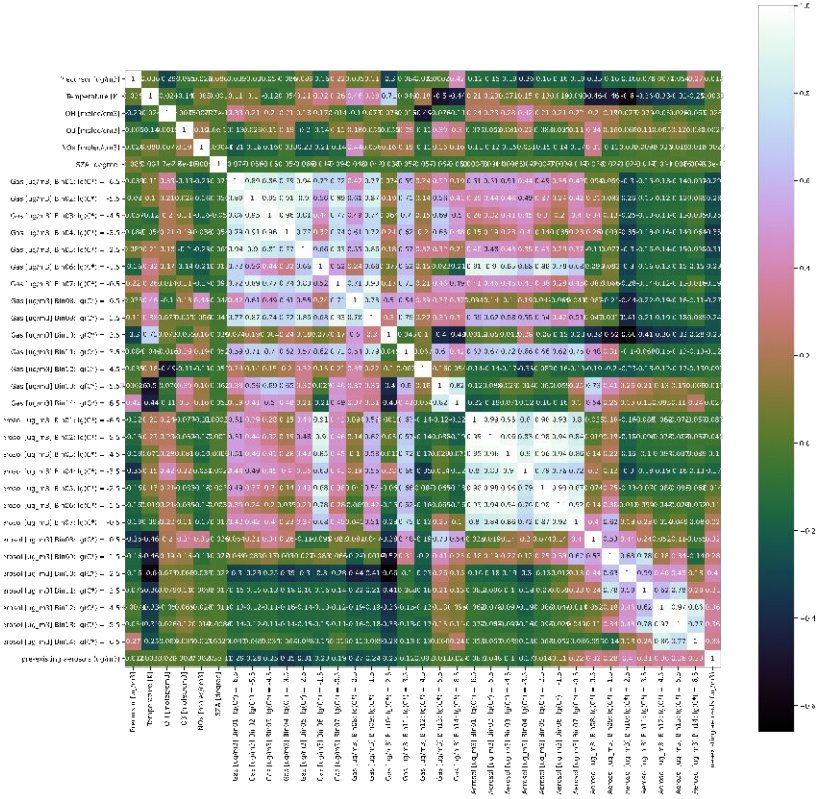


Log transform followed by a z-score standardization allows for a more normal distribution

PCA Visualization and Exploration



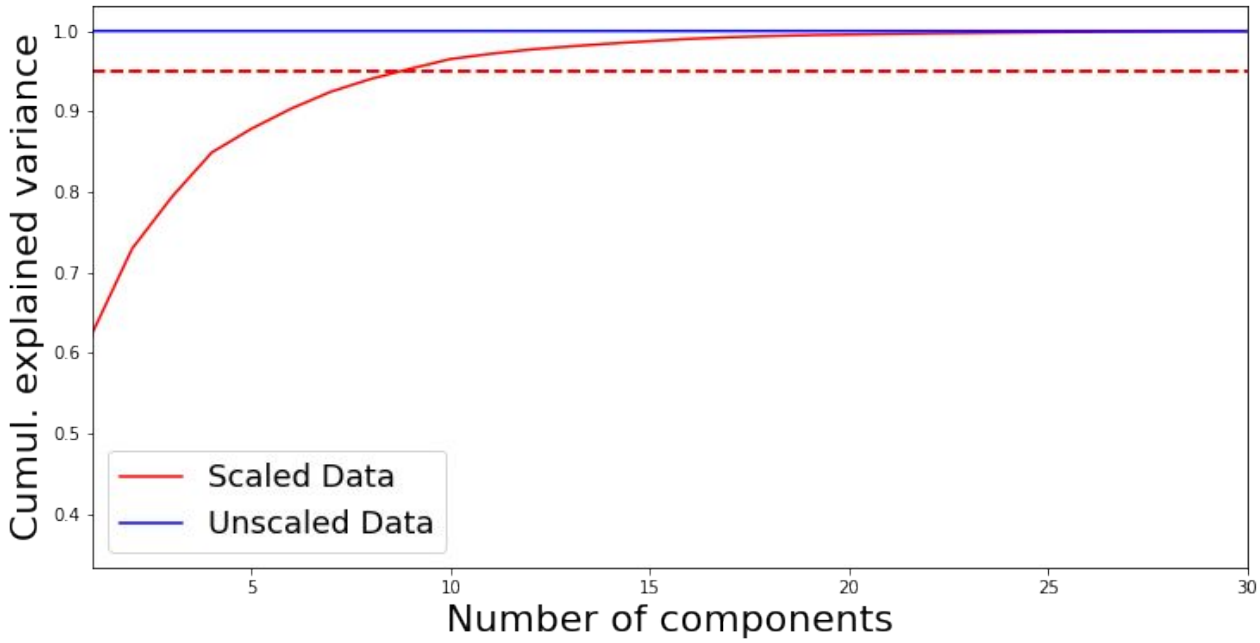
Correlation Matrix before PCA



Correlation Matrix of inverse transform after PCA

Shows the possibility for aggregation of some data (reduce dimension)

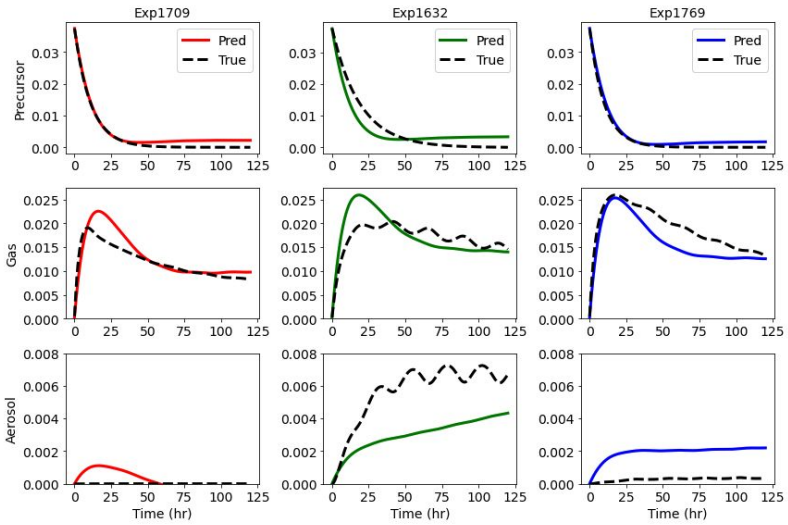
PCA Visualization and Exploration



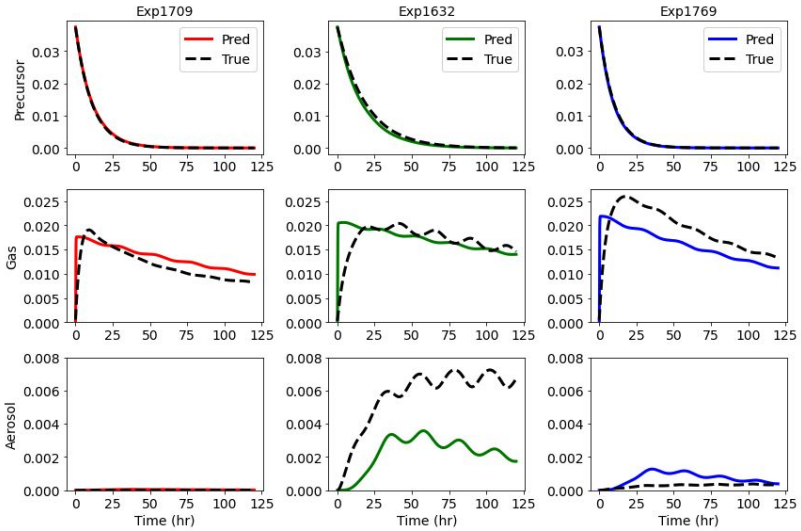
Much of the information can be explained using only 10 principal components

Reasonable to aggregate some of the data. Summing was used in practice, but option to further explore PCA

Choosing a model



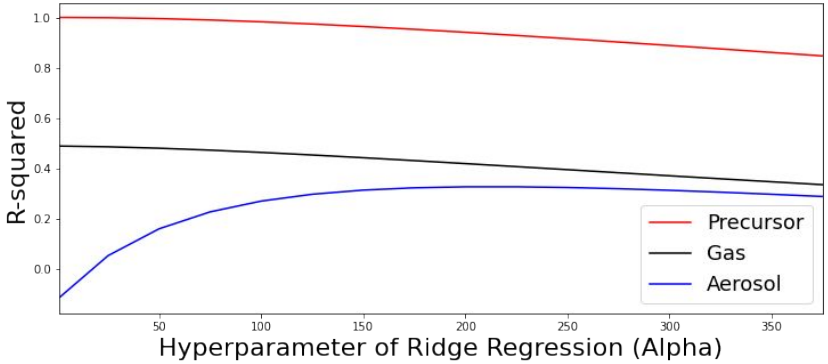
Linear Regression on unscaled data



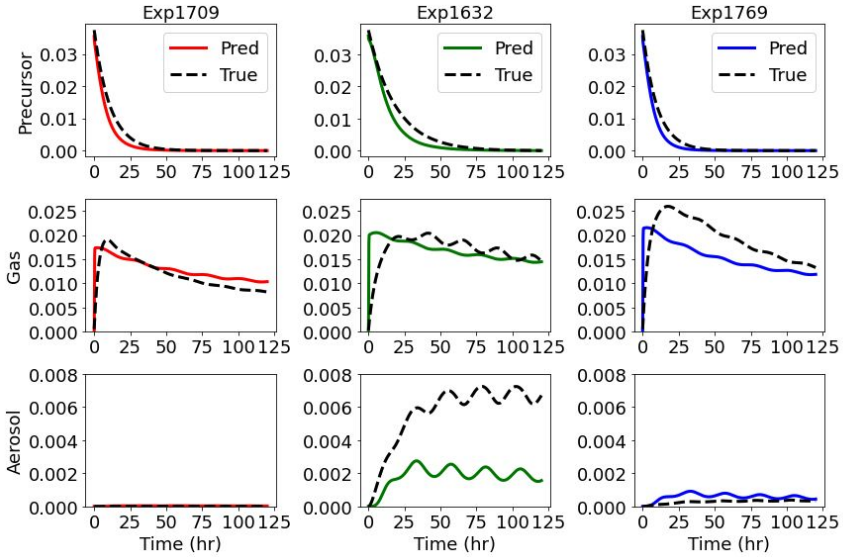
Linear Regression on scaled data

- Different methods for scaling are discussed and their effect on building a model reflected through comparison
- Even minor improvements draw out major data incites
- Performing a log transform on some features allowed us to capture the diurnal signal of the data, which is not an easy task

Choosing a model



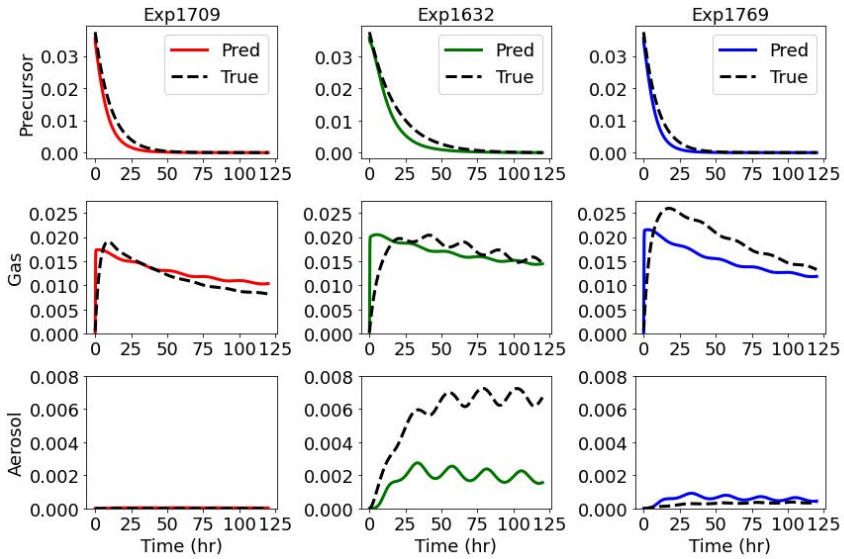
- Though linear models are often the easiest start with, they have limits
- May not be the best model for the problem at hand
- How do you show that you have exhausted a model

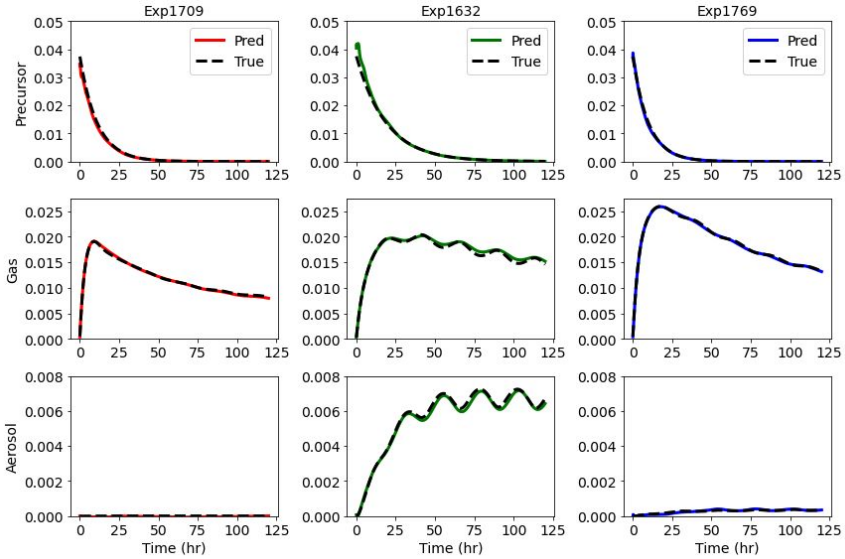


- Ridge regression at alpha=200
- Shows improvement in prediction of aerosol with only small cost in accuracy for other components

We have essentially exhausted a linear model and can move on to others

Choosing a model





Best current model

Goal

The goal of the tutorial is to lead the user to a model that performs as well as the model developed by the AIML group

There is room for further exploration. The PCA route is still underdeveloped

Could a user improve the linear models using PCA to aggregate features, rather than summing the bins?

Future Direction

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

David John Gagne

John Schreck

The AIML group

SIParCS program

NCAR, UCAR