Atmospheric Chemistry Modeling using Machine Learning

Christoph A. Keller
NASA Global Modeling and Assimilation Office (GMAO)
Universities Space Research Association (USRA)

Mat J. Evans
Wolfson Atmospheric Chemistry Laboratories, University of York
National Centre for Atmospheric Sciences, University of York

EGU Annual Meeting
11 April 2019
Numerical simulation of atmospheric chemistry

- 0.25° resolution (~ 25km), 72 levels, 250 chemical species
Numerical simulation of atmospheric chemistry

**Transport process:** Move chemicals across grid boxes

**Chemistry process:** In each grid box, solve chemical reactions, i.e. solve stiff ordinary differential equations (ODEs)

\[ A + B \rightarrow C + D \]

its rate is calculated as

\[ -\frac{d}{dt}[A] = -\frac{d}{dt}[B] = \frac{d}{dt}[C] = \frac{d}{dt}[D] = k[A][B] \]
The current solution: wait, wait, wait

- High-resolution chemistry simulation requires 3416 CPU’s
- Can simulate approx. 20 days in 24 hours
- Outputting the full chemical state is 1.5 TB / simulation day

Courtesy of W. Putman, NASA GMAO
Replace chemical integrator with machine learning model
Machine learning for atmospheric chemistry modeling

- 143 chemical species
- 91 photolysis rates
- Temperature
- Pressure
- Rel. humidity
- Solar zenith angle

- Training data set has 2.7 billion data points (44 GB)
- Tested: (neural network), random forest and XGBoost

Separate model for each species

Concentrations after chemistry
Impose chemical constraints on ML model to improve (long-term) accuracy

1. Distinguish between short-term vs. long-term species

   Long-lived (tendencies): \[ [X_i]_{T+\Delta T} = [X_i]_T + f(k, J, [X]) \]

   Short-lived (steady state): \[ [X_i]_{T+\Delta T} = f(k, J, [X]) \]

2. Predict NO + NO\(_2\) combined (NOx family approach)

VOC / HO\(_x\) ↔ NO ↔ NO\(_2\) ↔ O\(_x\) (Ozone)
Random forest / XGBoost training benchmarks

Comparison of XGBoost training time (data set = 44 GB)
Random forest / XGBoost training benchmarks

Comparison of XGBoost training time (data set = 44 GB)
Random forest / XGBoost training benchmarks

Comparison of XGBoost training time (data set = 44 GB)
Random forest / XGBoost training benchmarks

Comparison of XGBoost training time (data set = 44 GB)

XGBoost training time

- 1 CPU
- 8 CPU
- 64 CPU
- 1 GPU
- RAPIDS (1 GPU)

Training time [s]

10^1 10^2 10^3 10^4

Training sample size

10^4 10^5 10^6 10^7 10^8 10^9
Random forest / XGBoost can reproduce target concentrations almost perfectly (single-step prediction)

O3: tendency

N = 2,424,240
R² = 0.95
NRMSE [%] = 23.08
NMB [%] = -0.13

O3: tendency (+ concentration)

N = 2,424,240
R² = 1.00
NRMSE [%] = 0.06
NMB [%] = -0.00
Random forest / XGBoost solutions reflect known features of chemical kinetics
1-month simulation with random forest emulator
Surface concentrations over polluted regions are well reproduced by ML model.
Machine learning model remains stable over the long-term (but only if NOx is predicted as a family)
- Offline evaluation of one forest is 1000x faster than numerical integration
- Current implementation is very inefficient (2x slower than full chemistry)
- Currently working on seamless integration of XGBoost
Summary

- Tree models do a good job at simulating atmospheric chemistry
- Adding constraints (e.g., chemical families) to the machine learning model is critical
- Potential applications:
  - Chemical data assimilation
  - Air quality forecasting
- Issues:
  - Train on very large data sets (>1 TB)
  - Dynamics for >200 chemical species is still slow

Keller and Evans: Application of random forest regression to the calculation of gas-phase chemistry within the GEOS-Chem chemistry model v10, GMD, 2019.
Prediction of NOx

**NOx - feature importances**

- NO2
- ALD2
- NO
- HNO2
- SUNCO
- SO2
- CH2O
- CO
- PRPE
- N2O5
- NO3
- NO3 -> NO2 + O
- C3H8
- BrO -> Br + O
- C2H6
- HNO4
- TEMP
- O3
- HNO3
- ACET

**NOx: tendency**

- N = 2,424,240
- R² = 0.96
- NRMSE [%] = 21.10
- NMB [%] = 0.28

**NOx: tendency (+ concentration)**

- N = 2,424,240
- R² = 1.00
- NRMSE [%] = 0.52
- NMB [%] = -0.01
Surface concentrations over polluted regions are well reproduced by ML model.