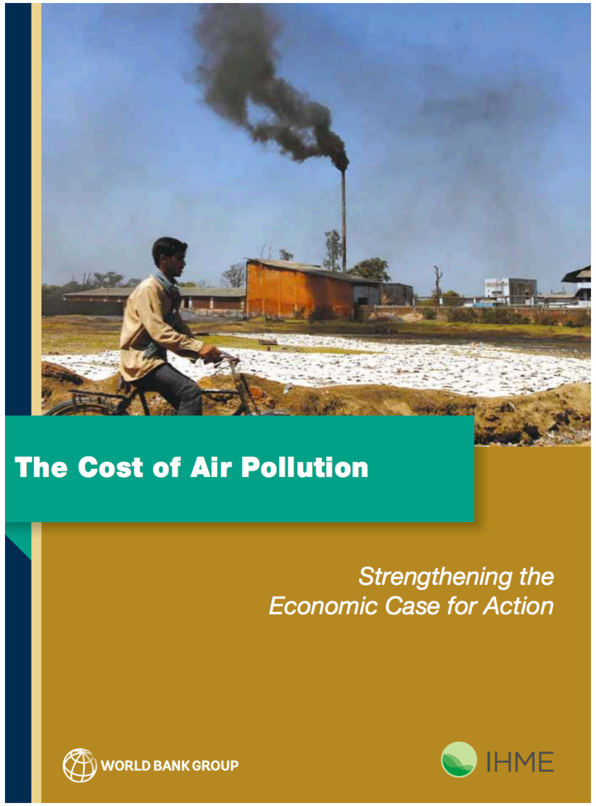
Atmospheric Chemistry Modeling and Air Quality Forecasting using Machine Learning

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

1st NOAA Workshop on Leveraging AI
23-25 April 2019

Air pollution is a global problem



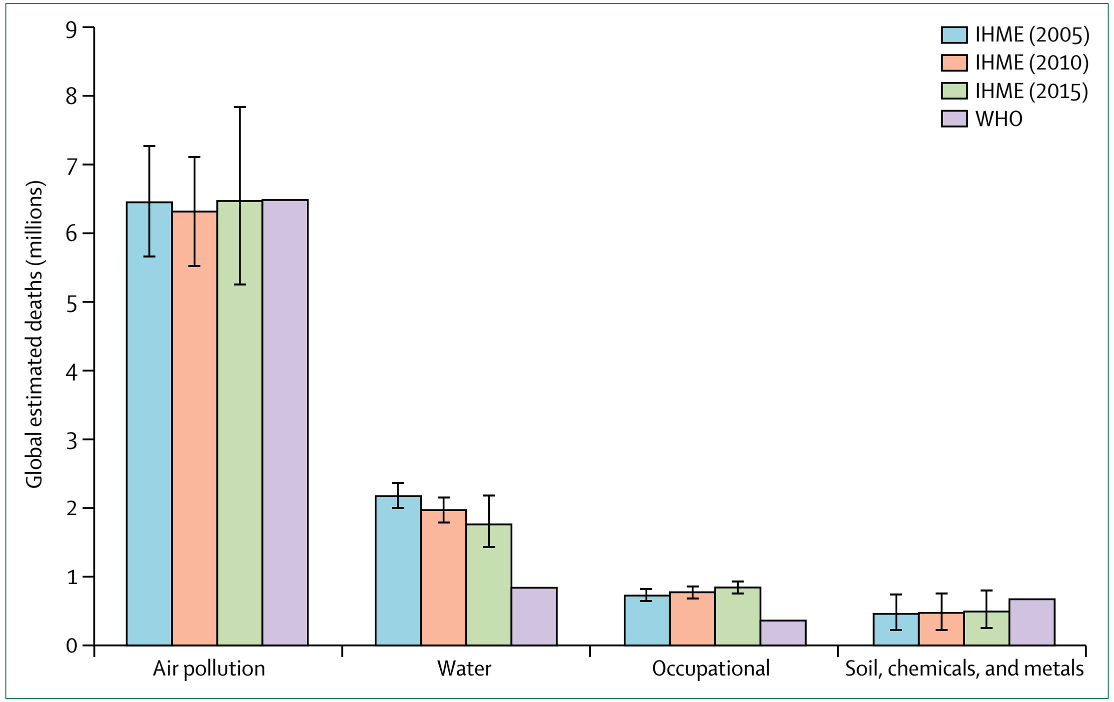
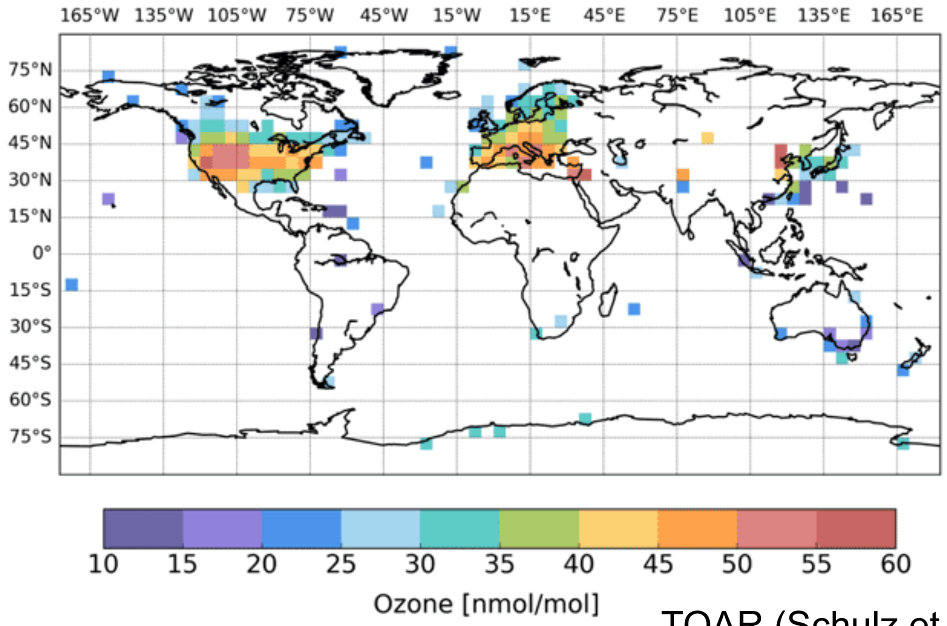


Figure 4: **Global estimated deaths (millions) by pollution risk factor, 2005–15**
Using data from the GBD study[42] and WHO.[99] IHME=Institute for Health Metrics and Evaluation.

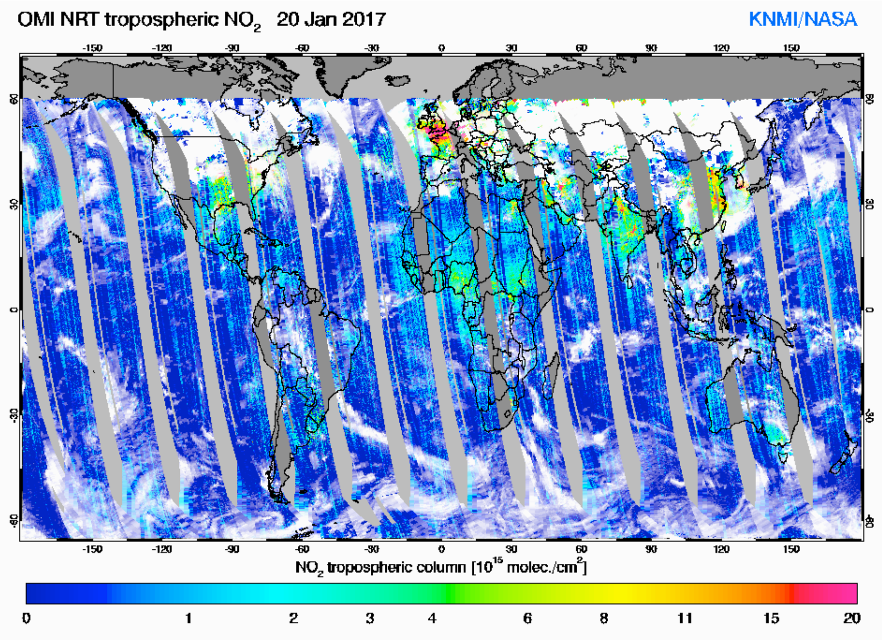
World Bank: ~$5 trillion in welfare losses in 2013

The Lancet (2017): Air pollution is responsible for 6-7 millions death / year

Need models to fill gaps in observations



TOAR (Schulz et al., 2017)

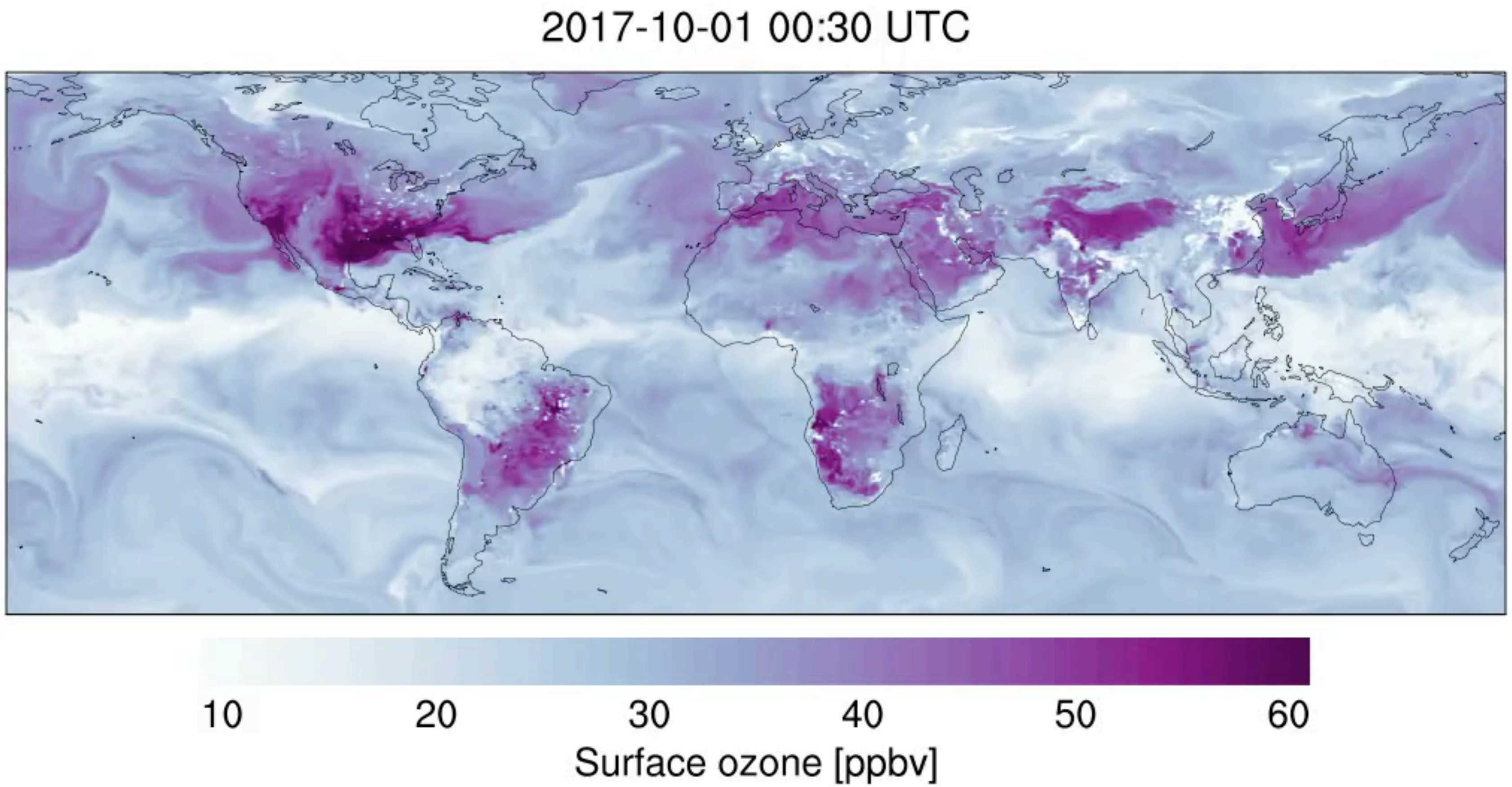


Surface observations are not global

Satellite observations are also discontinuous

christoph.a.keller@nasa.gov

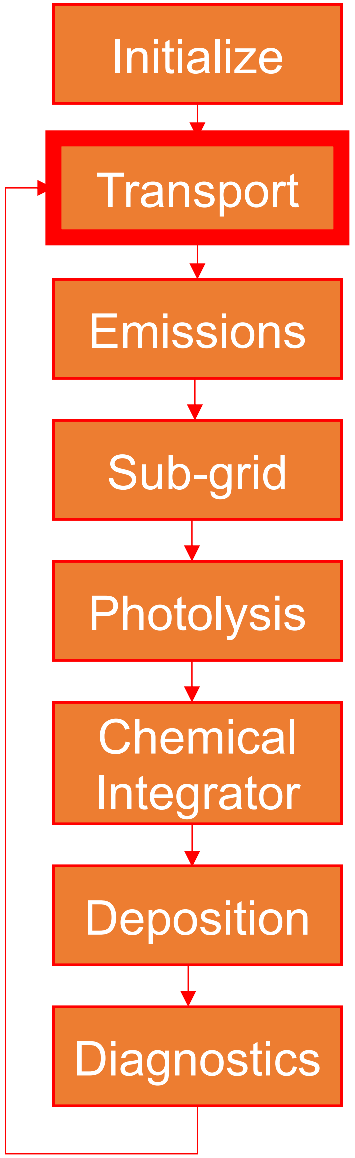
Numerical simulation of atmospheric chemistry



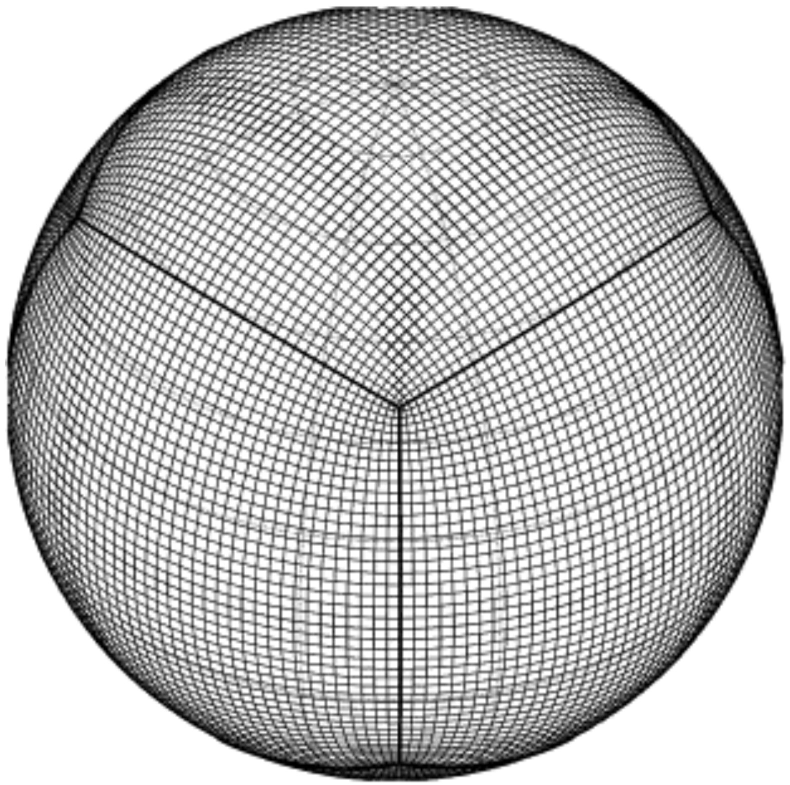
- 0.25° resolution (~ 25km), 72 levels, 250 chemical species

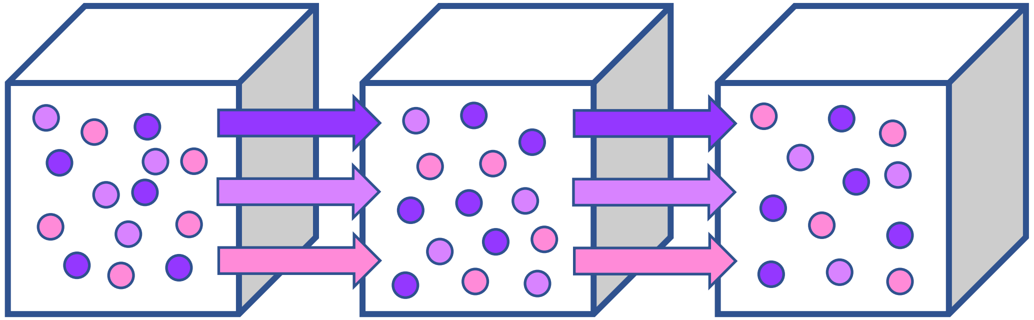
christoph.a.keller@nasa.gov

Numerical simulation of atmospheric chemistry



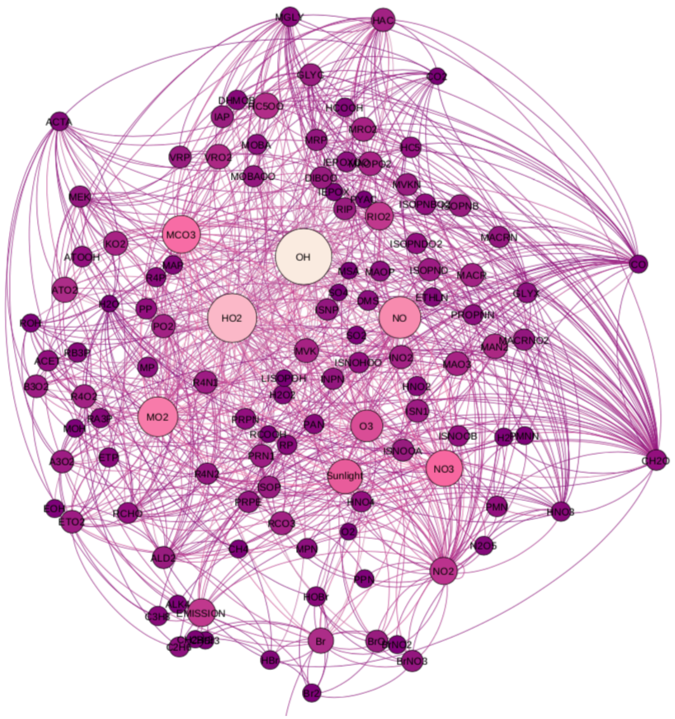
Transport process: Move chemicals across grid boxes





christoph.a.keller@nasa.gov

Numerical simulation of atmospheric chemistry



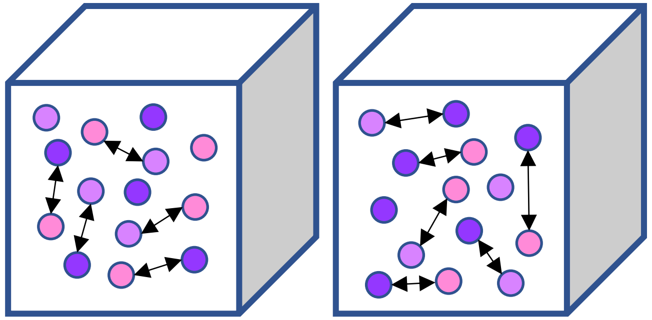
Transport process: Move chemicals across grid boxes

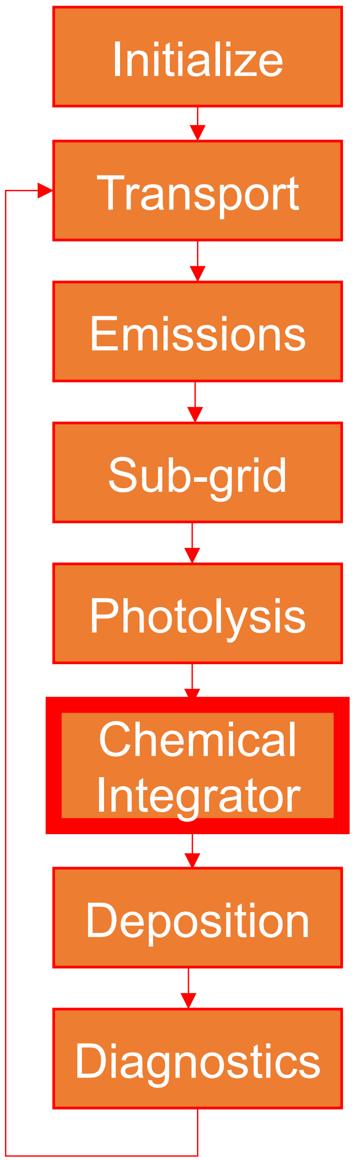
Chemistry process: In each grid box, solve chemical reactions, i.e. solve stiff ordinary differential equations (ODEs)

$$A + B \rightarrow C + D$$

its rate is calculated as

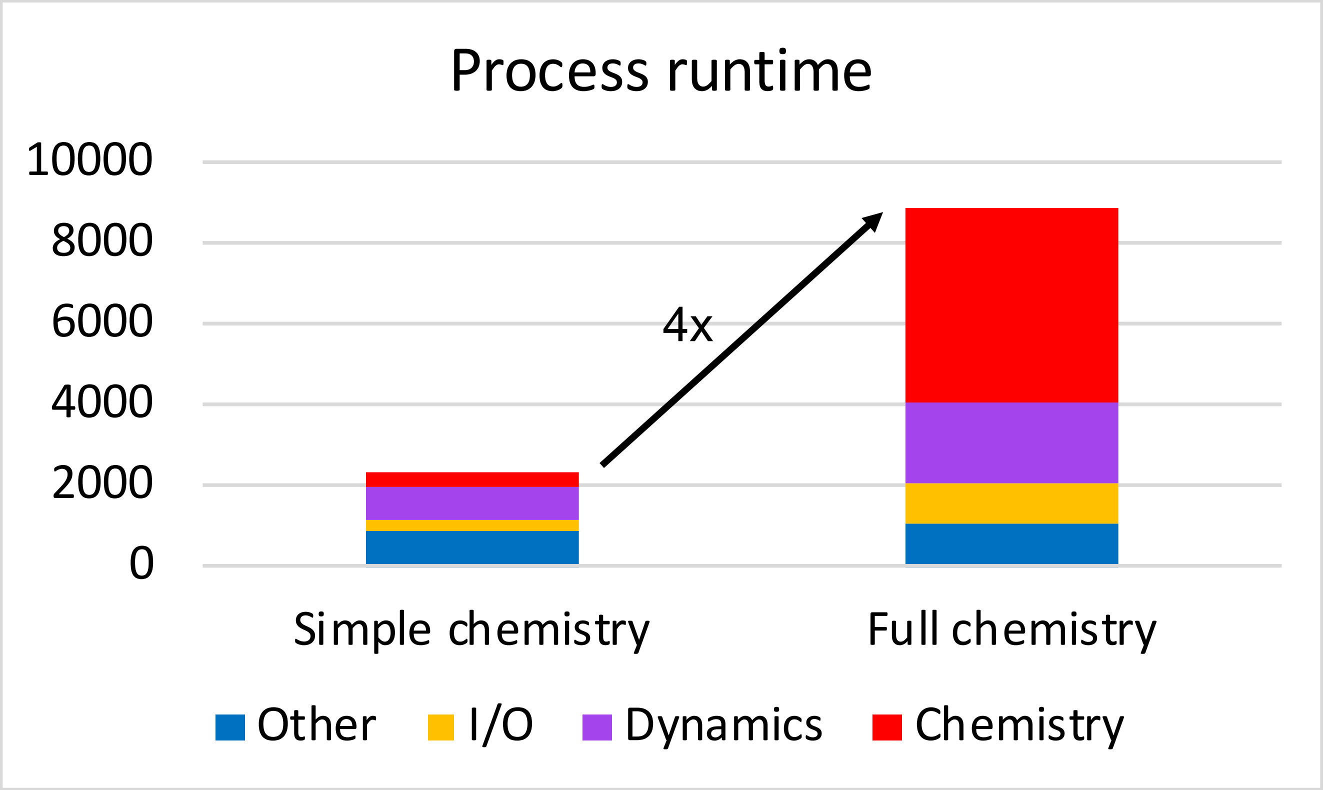
$$-\frac{d}{dt}[A] = -\frac{d}{dt}[B] = \frac{d}{dt}[C] = \frac{d}{dt}[D] = k[A][B]$$





christoph.a.keller@nasa.gov

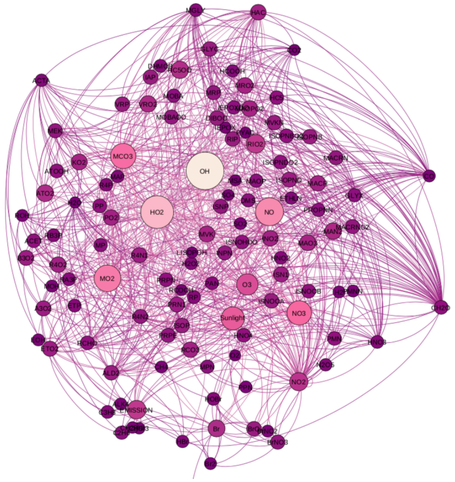
Atmospheric chemistry models are computationally expensive



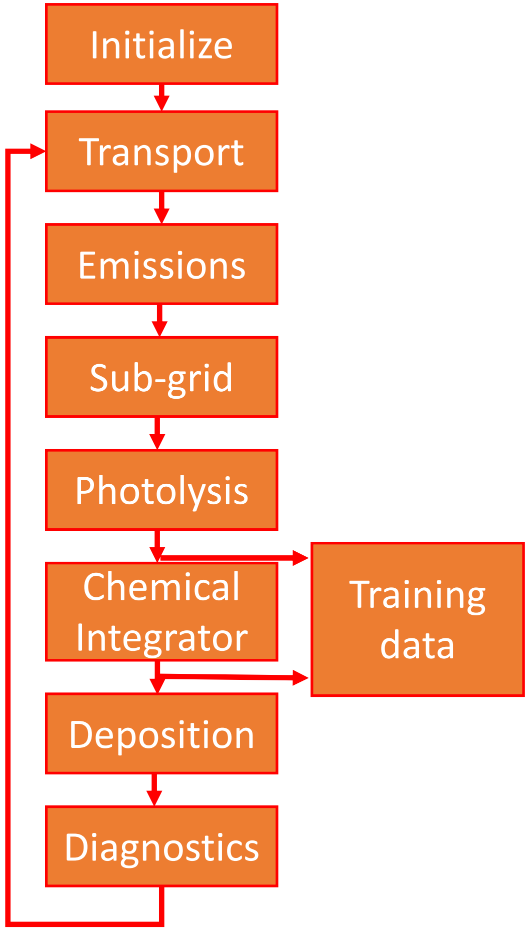
- High-resolution chemistry simulation requires >1000 CPU’s
- Throughput: approx. 20 days in 24 hours

christoph.a.keller@nasa.gov

Replace chemical integrator with machine learning model

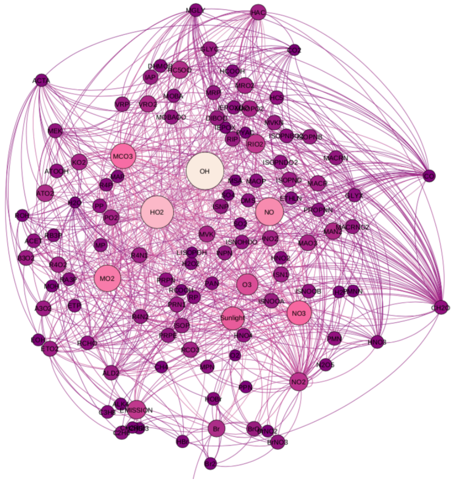


Numerical model

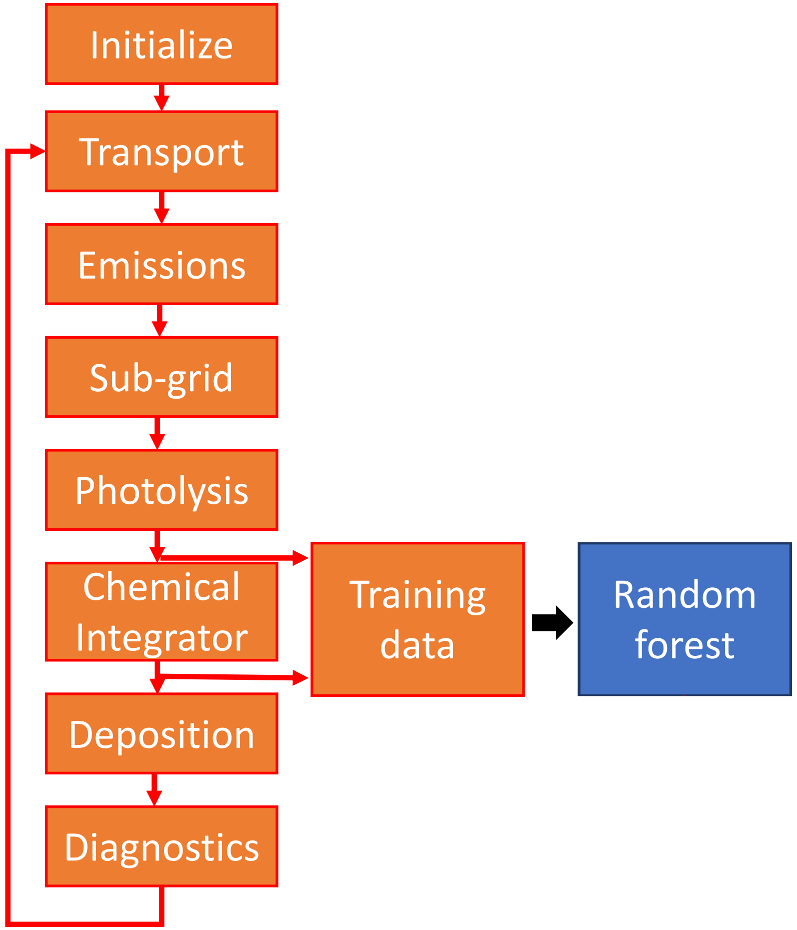


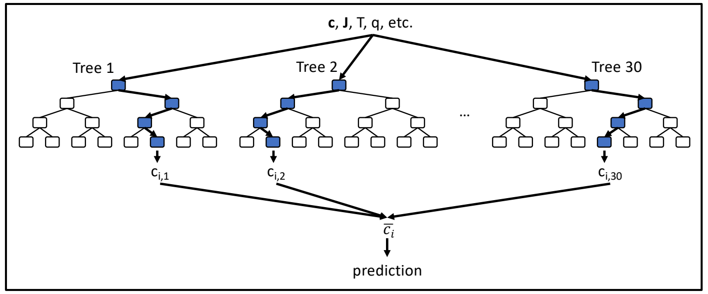
christoph.a.keller@nasa.gov

Replace chemical integrator with machine learning model



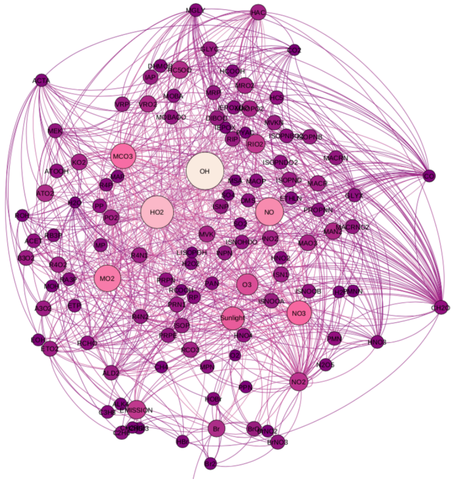
Numerical model

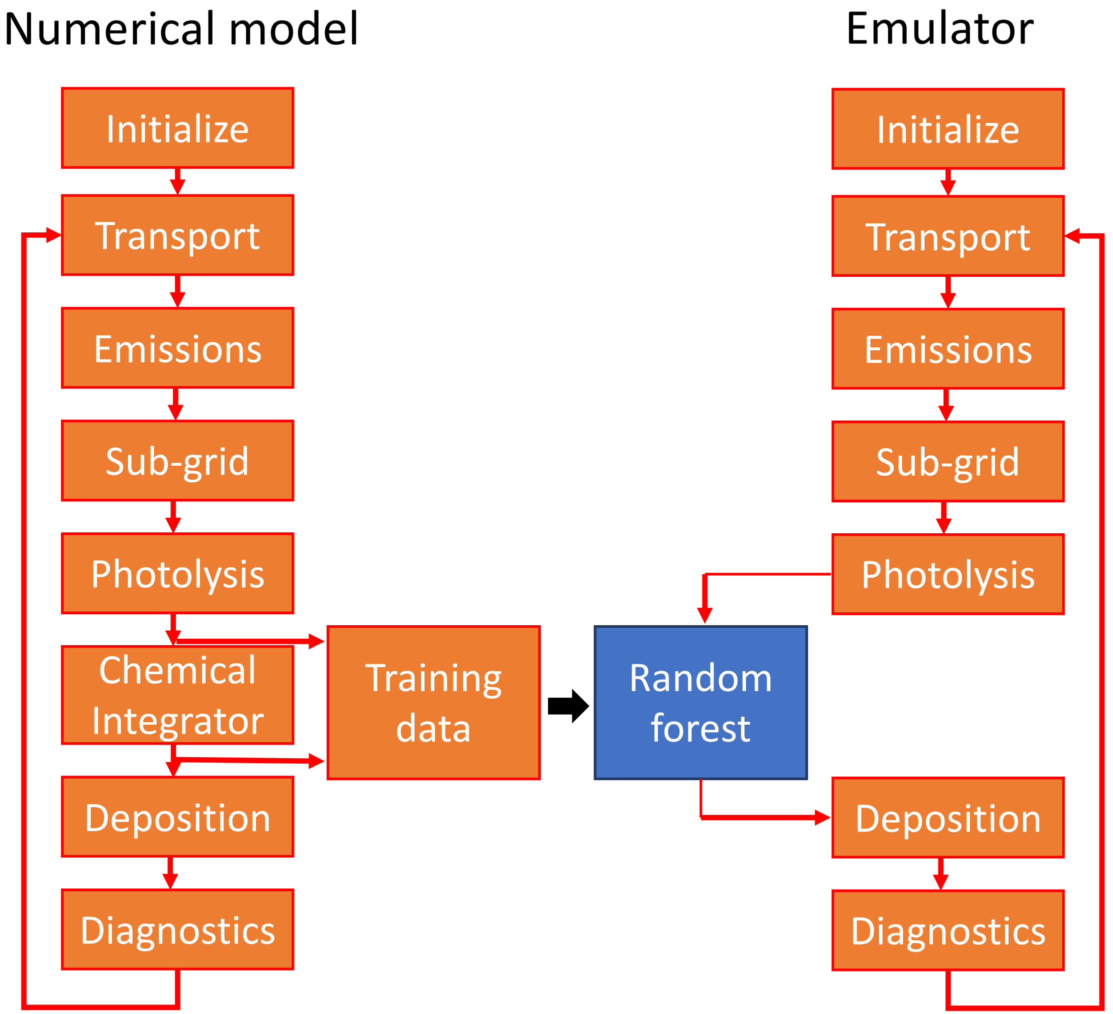


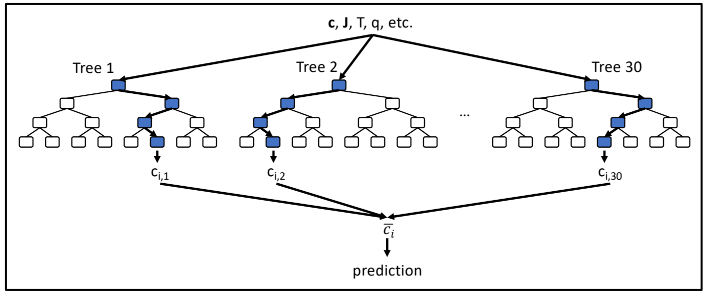


christoph.a.keller@nasa.gov

Replace chemical integrator with machine learning model







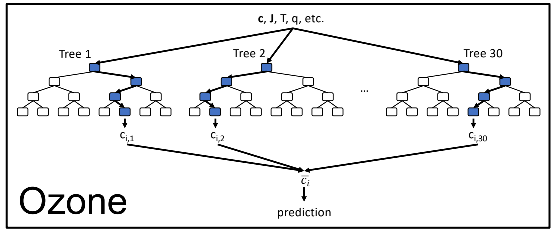
christoph.a.keller@nasa.gov

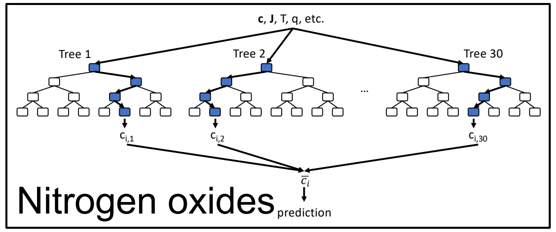
Machine learning for atmospheric chemistry modeling

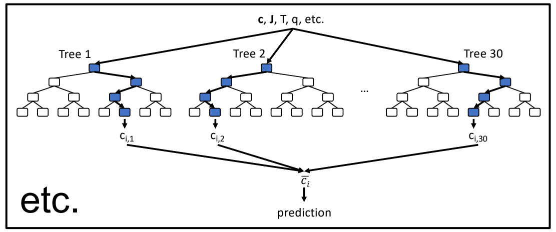
- 143 chemical species
- 91 photolysis rates
- Temperature
- Pressure
- Rel. humidity
- Solar zenith angle

Concentrations after chemistry

Separate model for each species



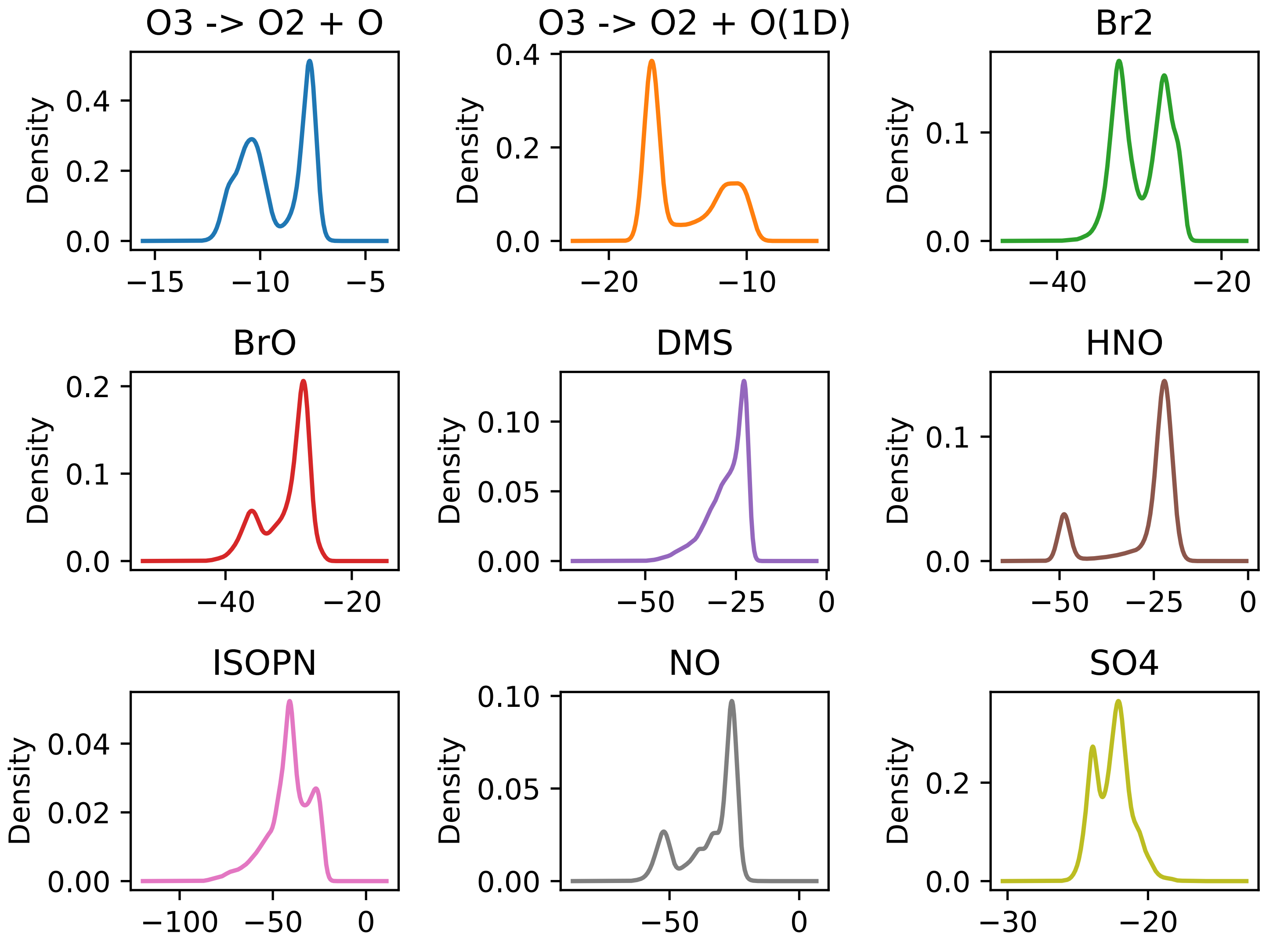




- ➢ Training data set: 2.7 billion data points (44 GB)
- ➢ Tested: (neural network), random forest and XGBoost

christoph.a.keller@nasa.gov

Many input features have multiple modes



christoph.a.keller@nasa.gov

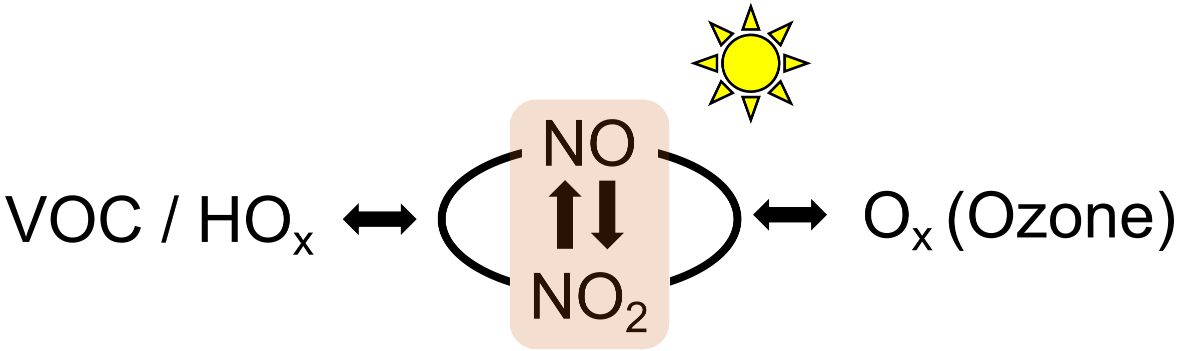
Impose chemical constraints on ML model to improve (long-term) accuracy

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

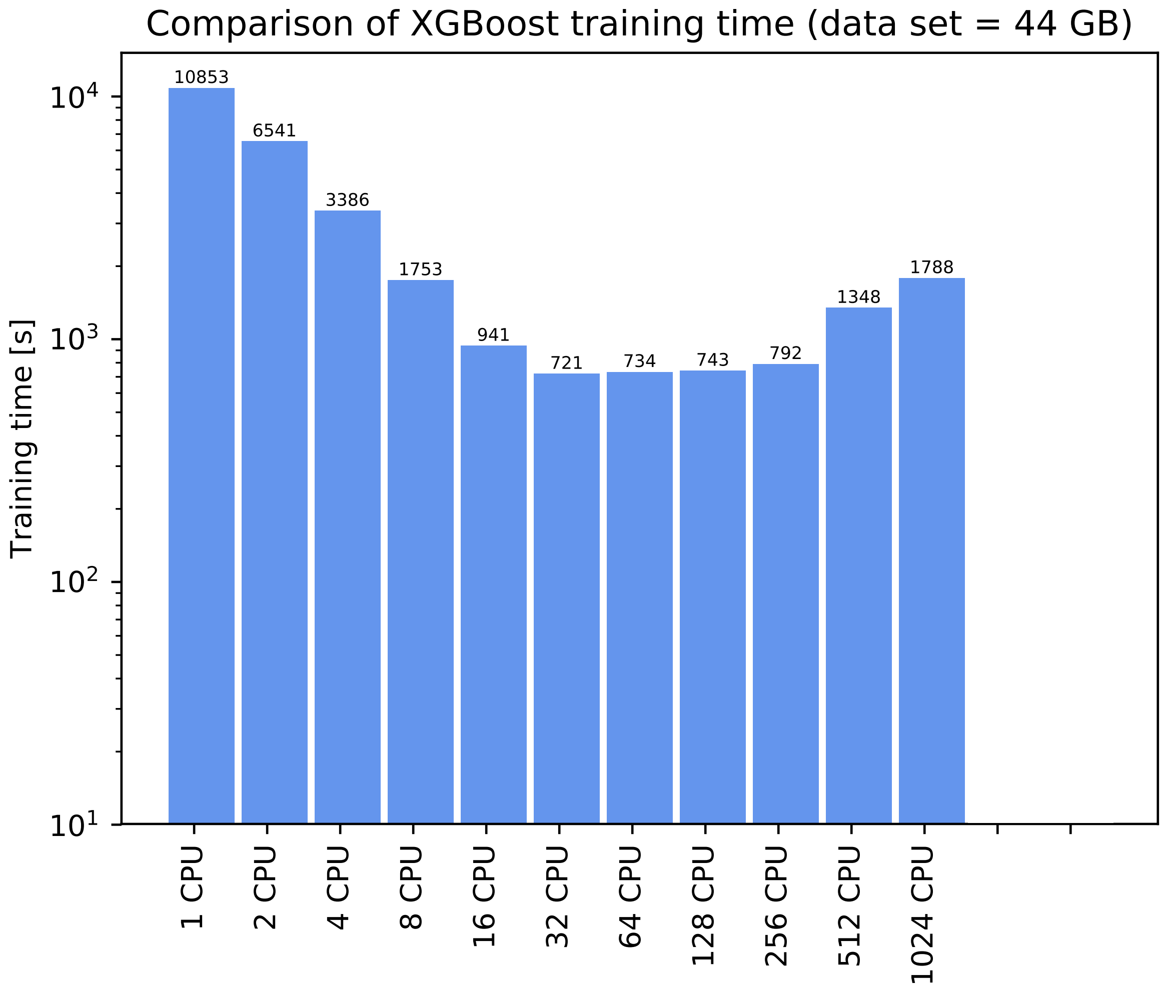
Long-lived (tendencies): $[X_i]_{T+\Delta T} = [X_i]_T + f(\,\mathbf{k}, \mathbf{J}, [\mathbf{X}]\,)$

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

2. Predict NO + NO_2 combined (NOx family approach)

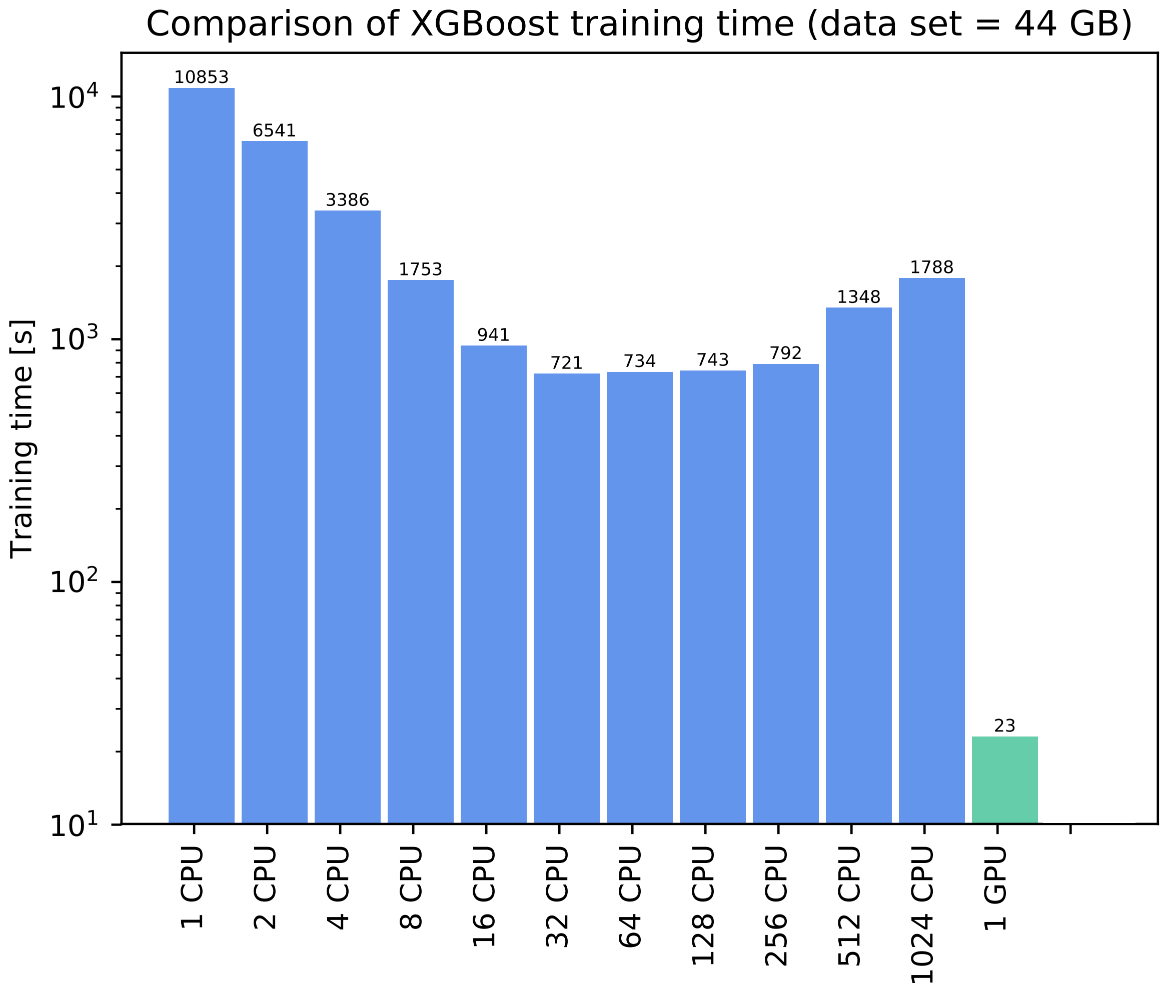


Random forest / XGBoost training benchmarks



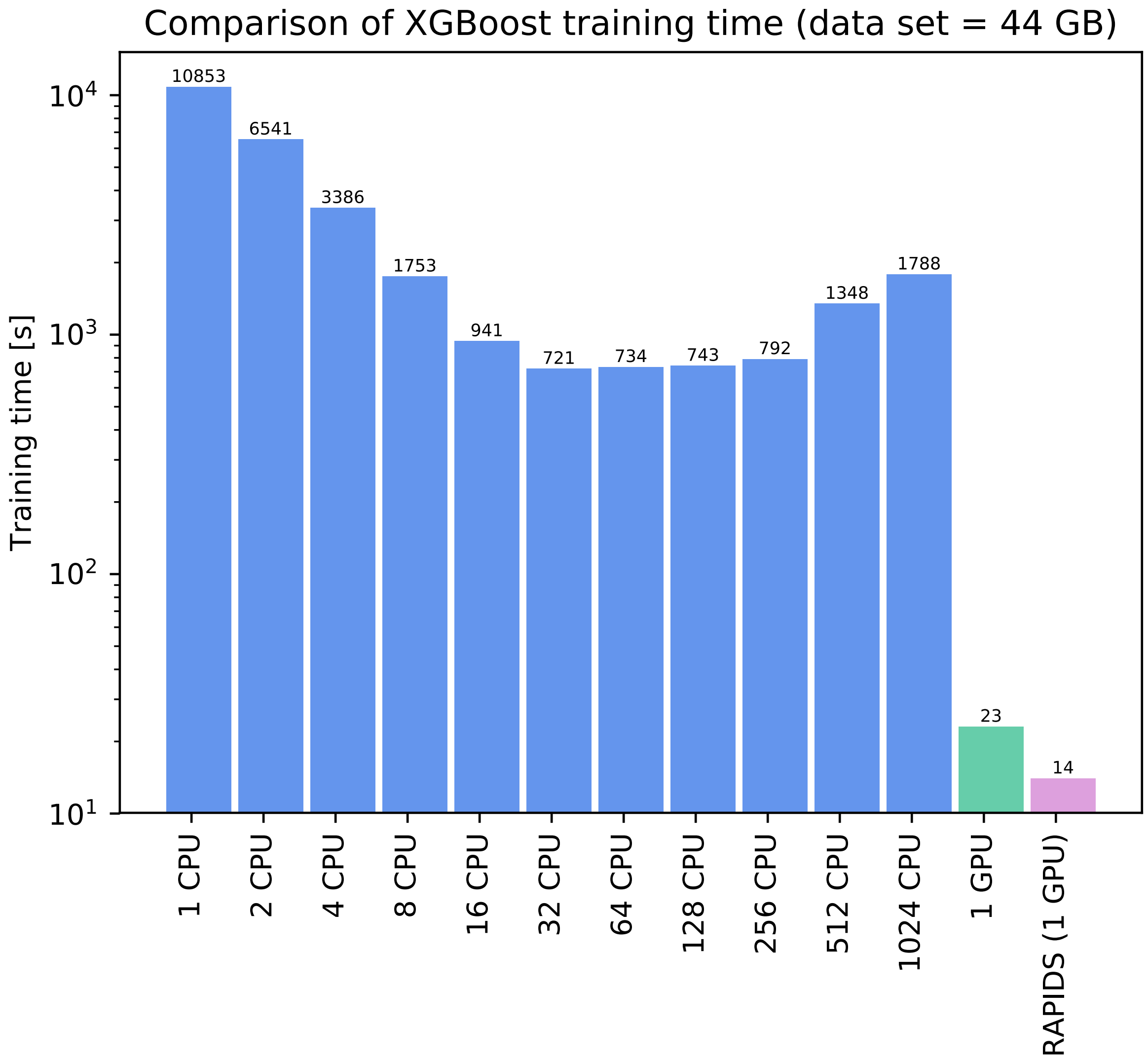
christoph.a.keller@nasa.gov

Random forest / XGBoost training benchmarks



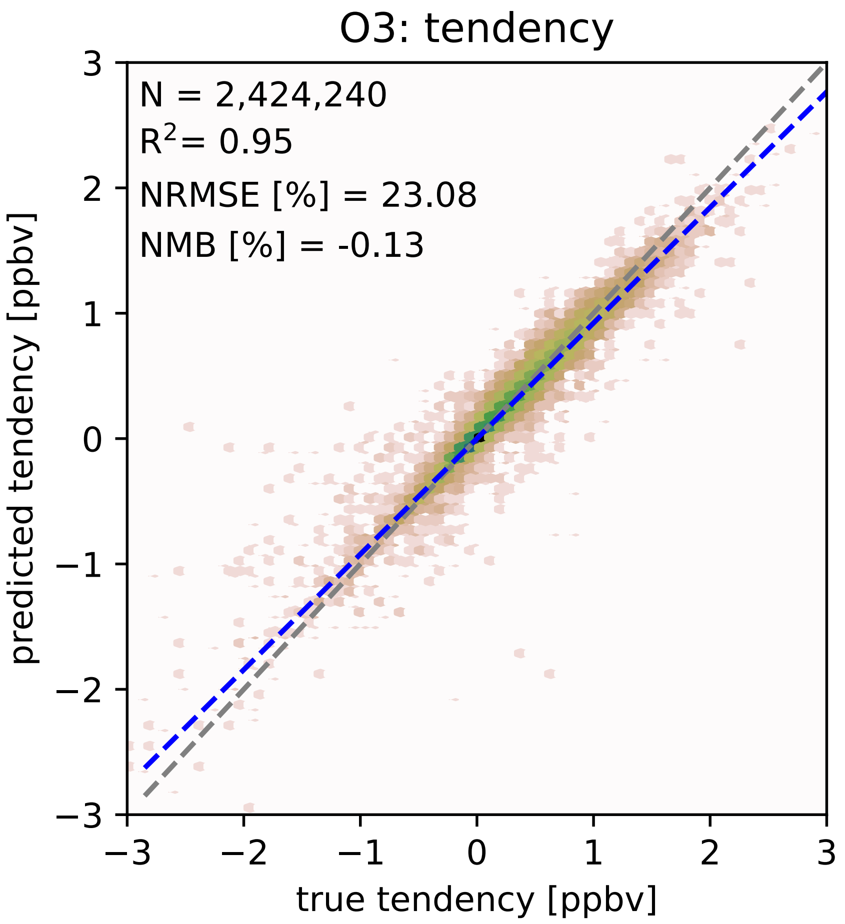
christoph.a.keller@nasa.gov

Random forest / XGBoost training benchmarks



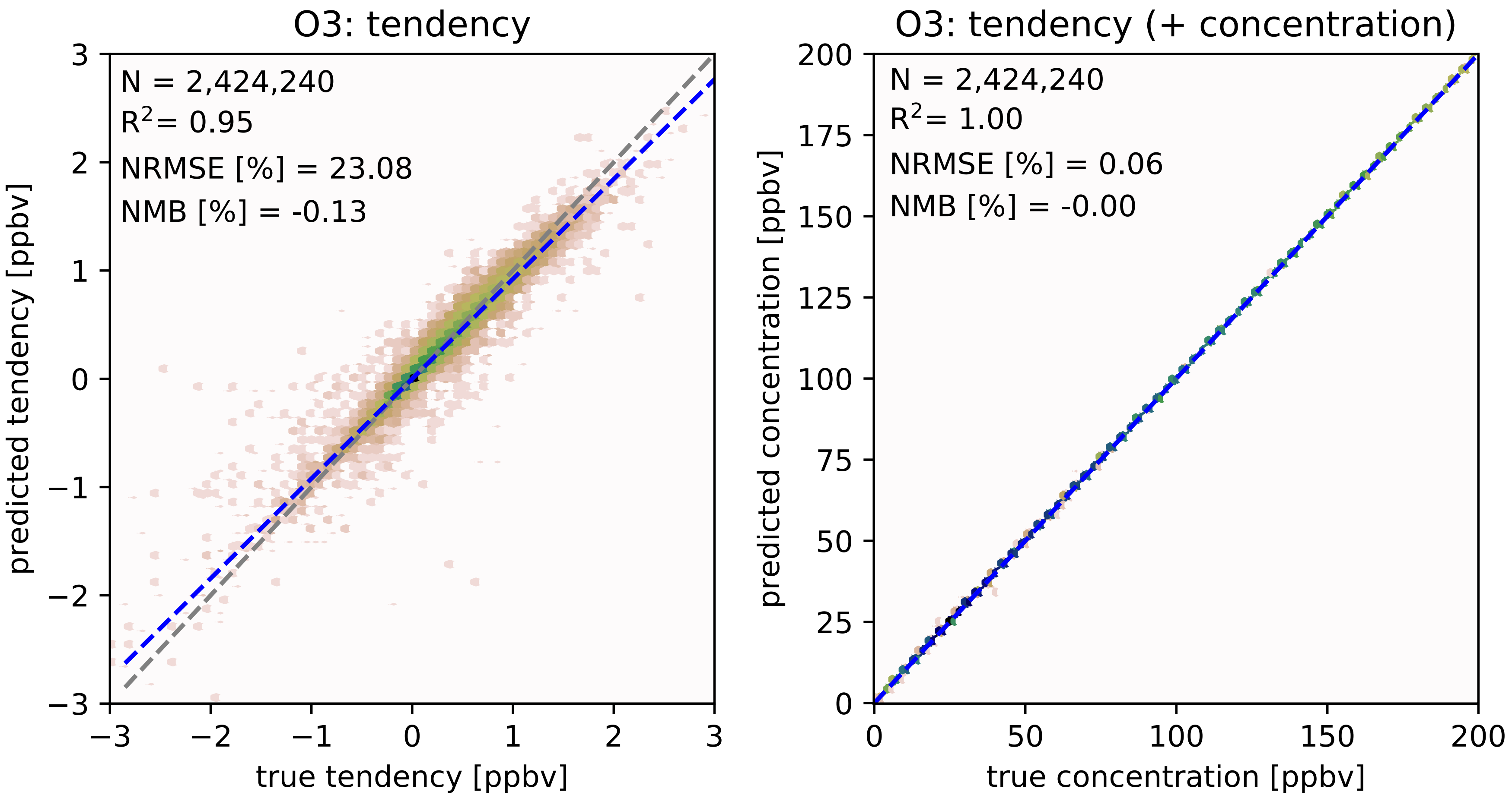
christoph.a.keller@nasa.gov

Random forest / XGBoost reproduce target concentrations well (single-step prediction)



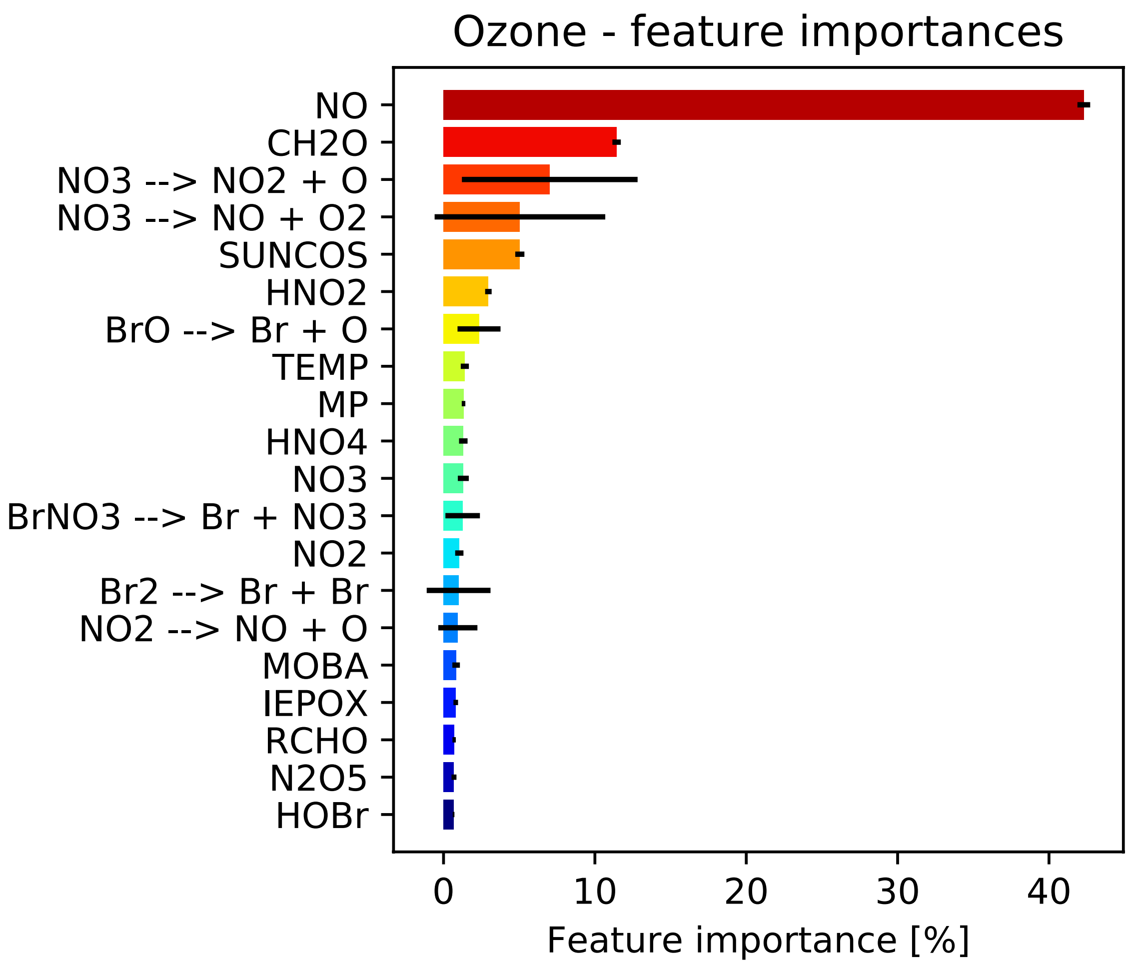
christoph.a.keller@nasa.gov

Random forest / XGBoost reproduce target concentrations well (single-step prediction)



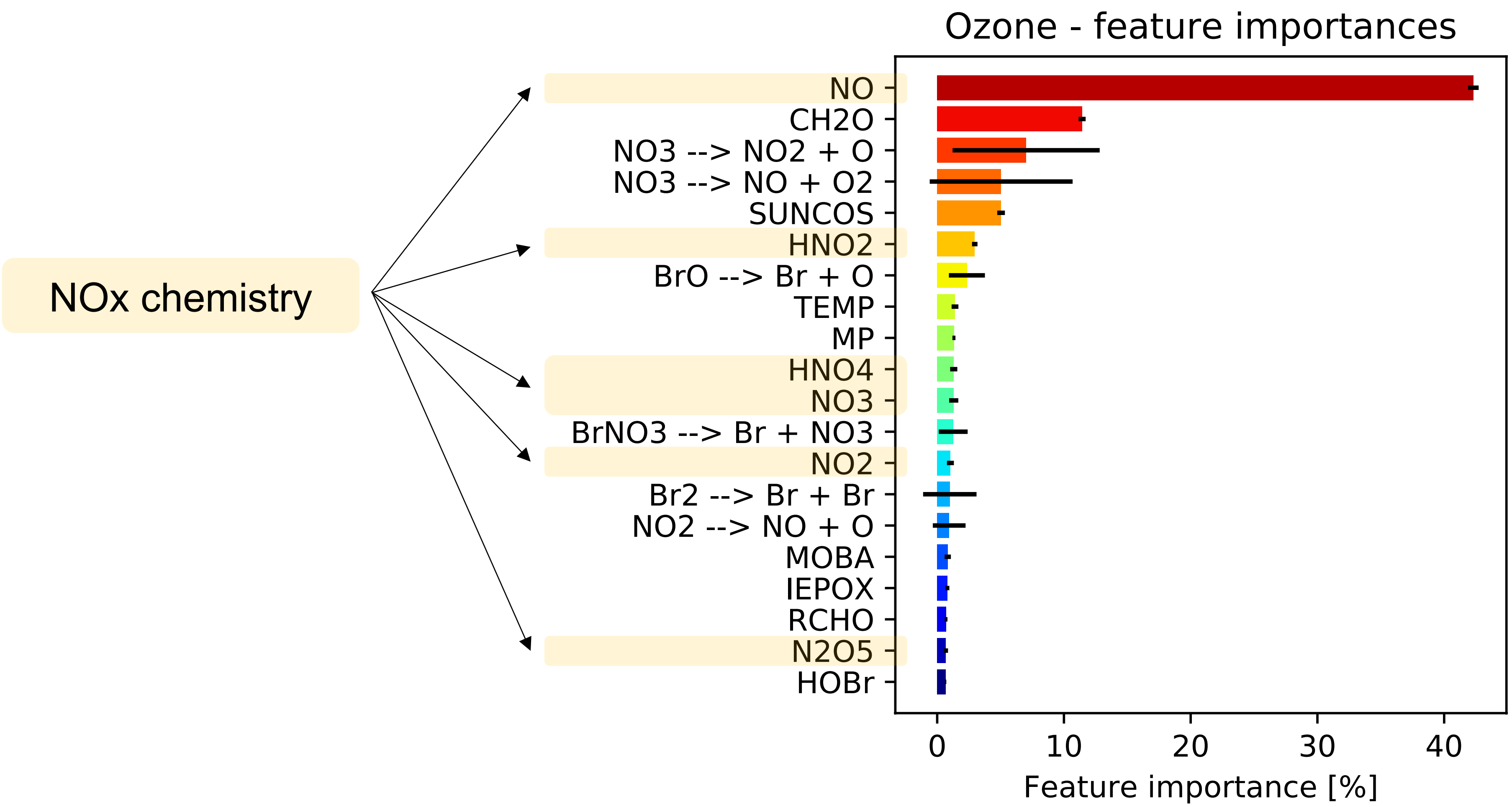
christoph.a.keller@nasa.gov

Random forest / XGBoost solutions reflect known features of chemical kinetics



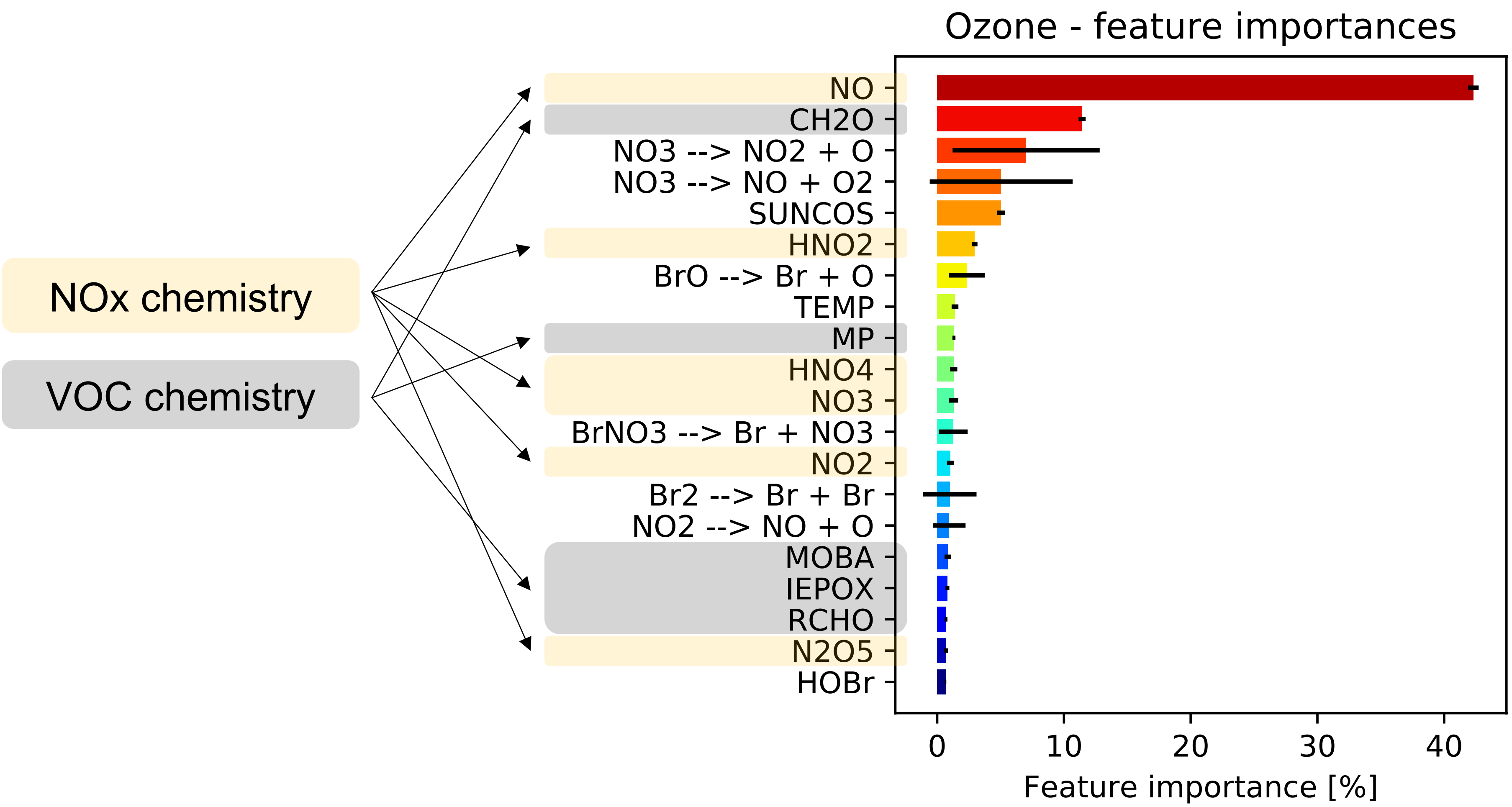
christoph.a.keller@nasa.gov

Random forest / XGBoost solutions reflect known features of chemical kinetics



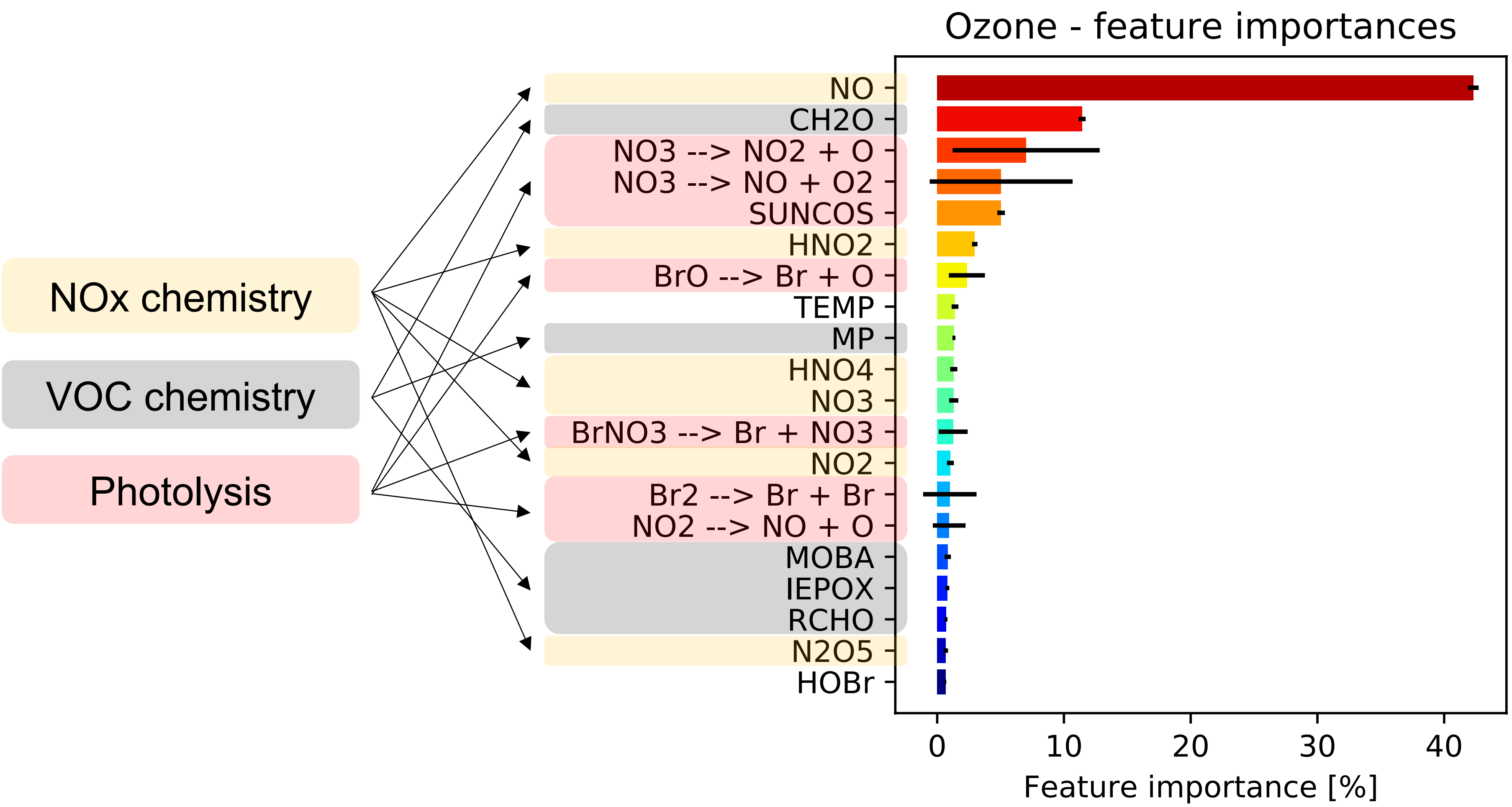
christoph.a.keller@nasa.gov

Random forest / XGBoost solutions reflect known features of chemical kinetics



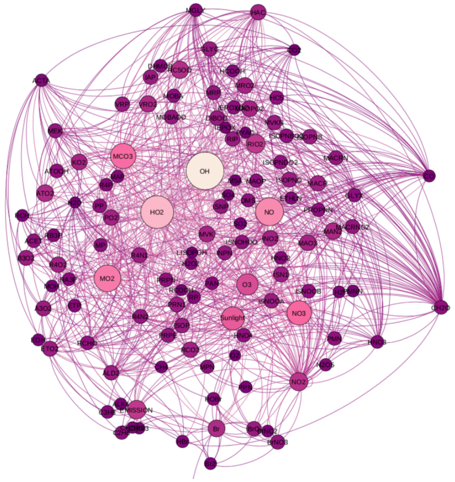
christoph.a.keller@nasa.gov

Random forest / XGBoost solutions reflect known features of chemical kinetics



christoph.a.keller@nasa.gov

1-month simulation with random forest emulator



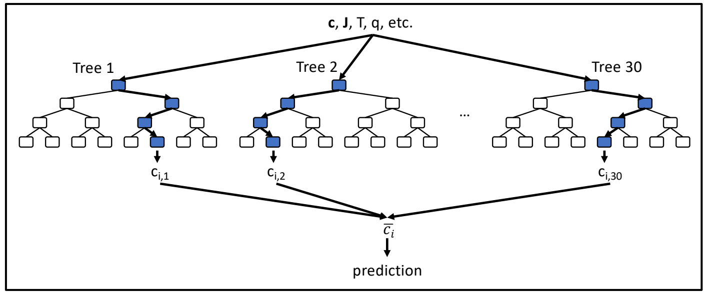
Numerical model

Initialize → Transport → Emissions → Sub-grid → Photolysis → Chemical Integrator → Deposition → Diagnostics

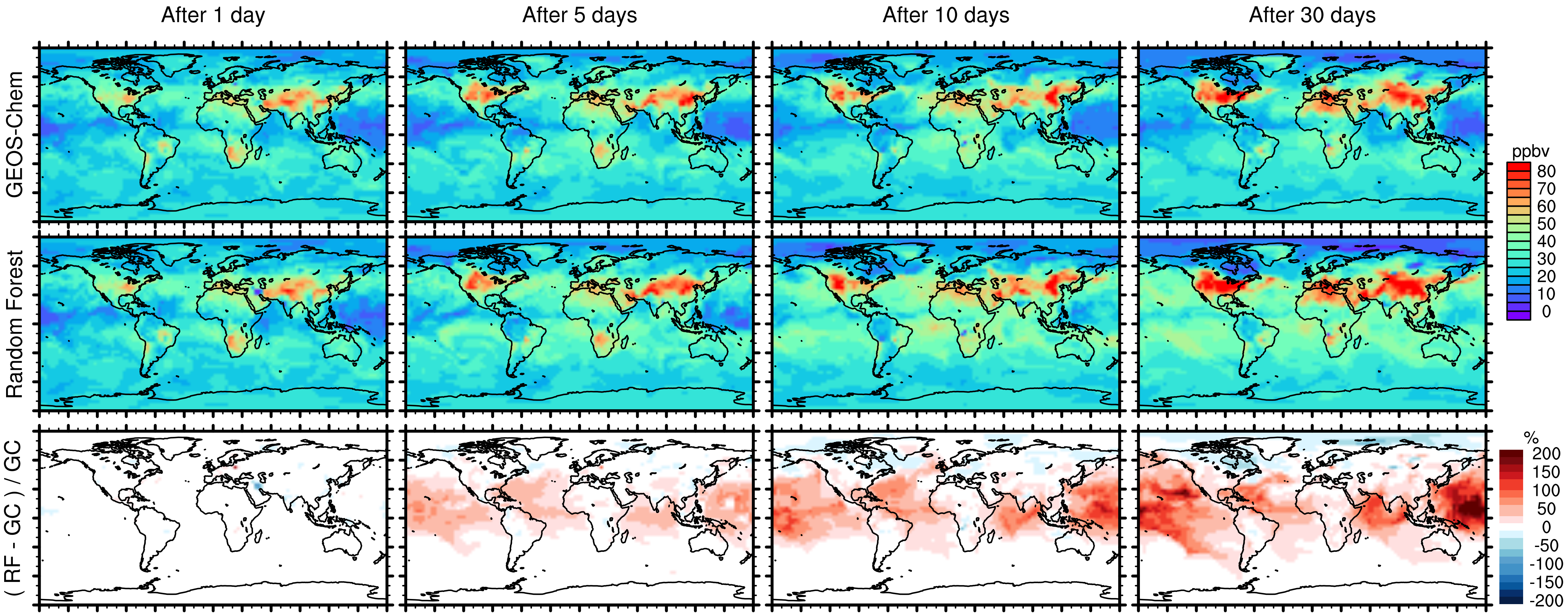
Training data → Random forest

Emulator

Initialize → Transport → Emissions → Sub-grid → Photolysis → Random forest → Deposition → Diagnostics



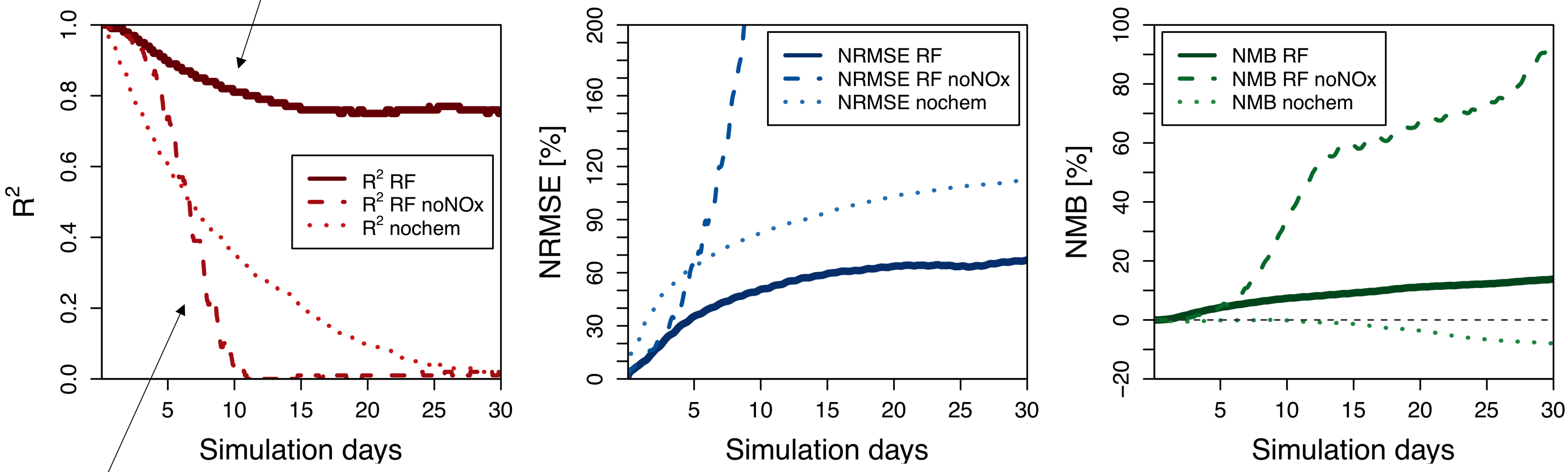
Random forest overestimates ozone surface concentrations over remote regions



christoph.a.keller@nasa.gov

Machine learning model remains stable over the long-term (but only if NOx is predicted as a family)

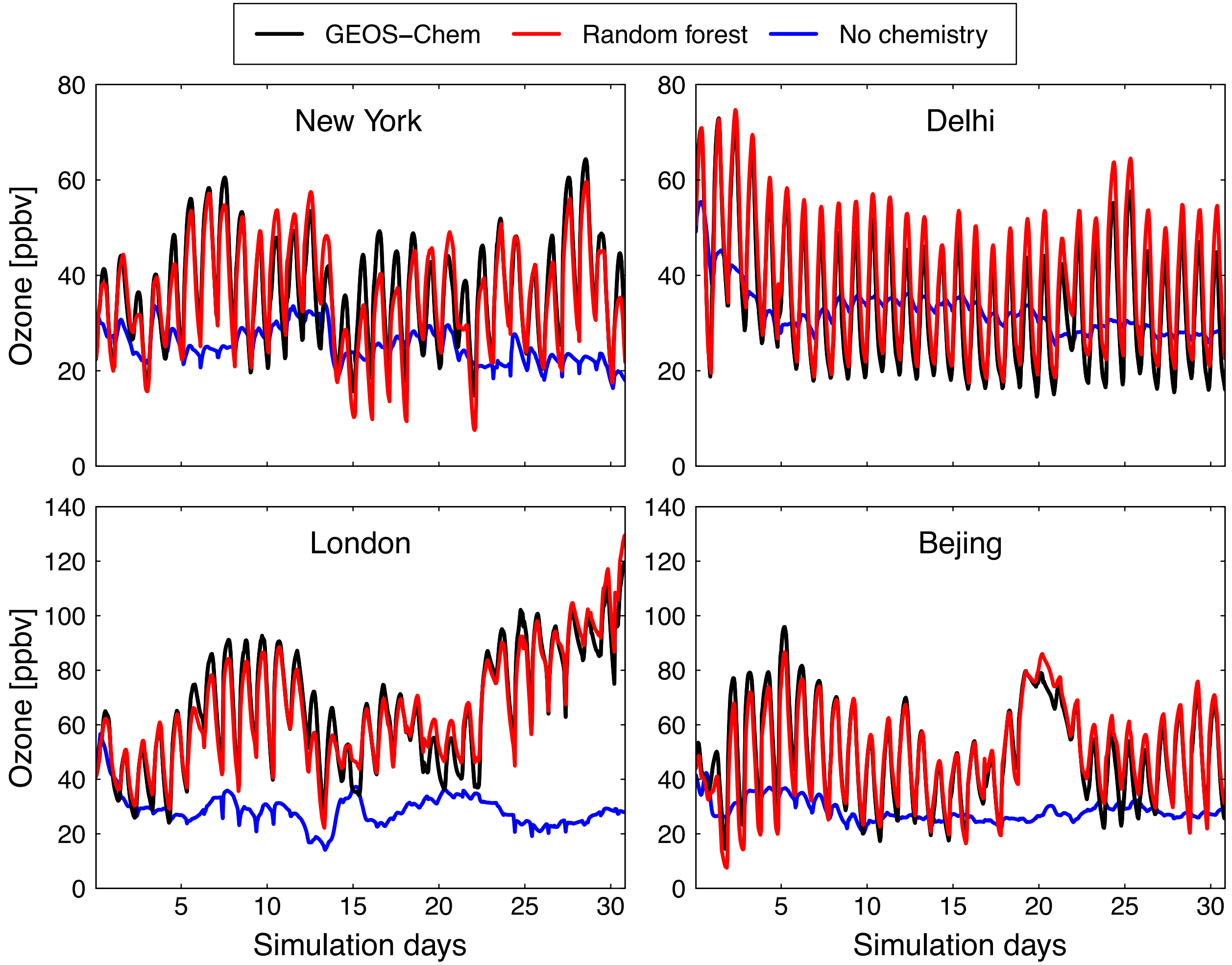
Model with NOx family prediction



Model without NOx family prediction

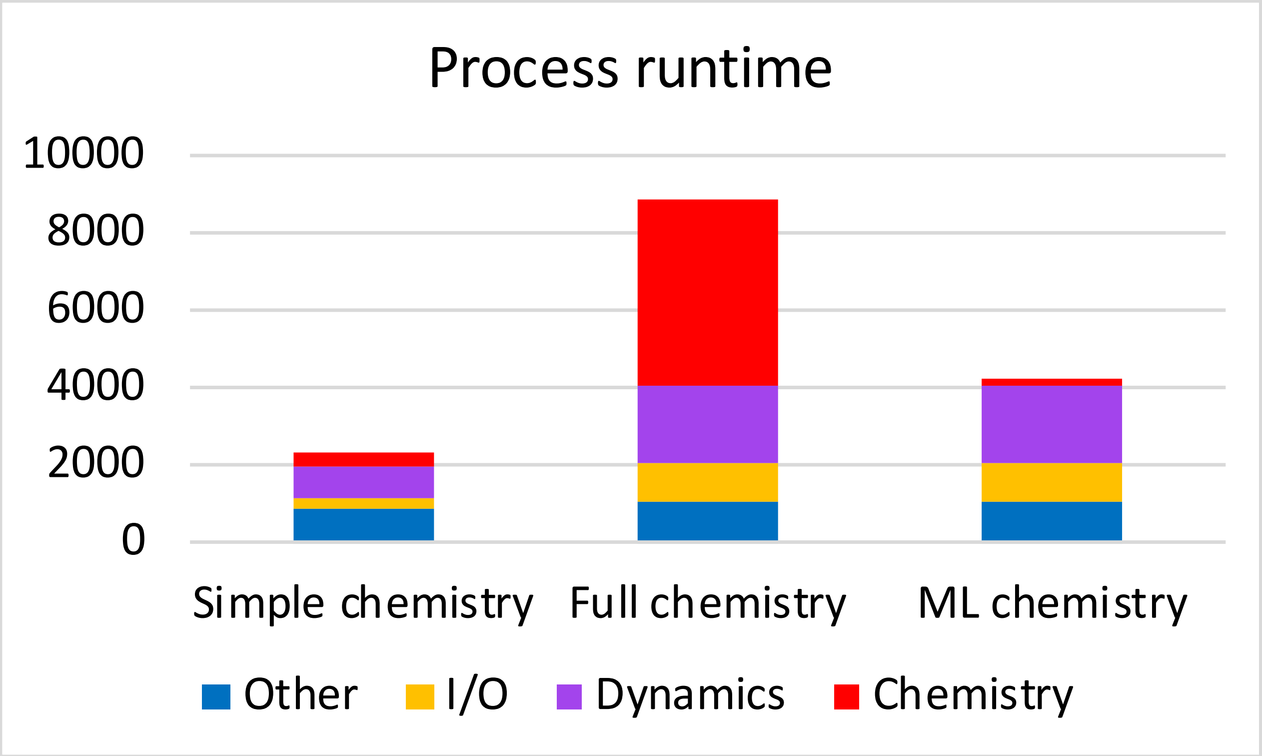
christoph.a.keller@nasa.gov

Surface concentrations over polluted regions are well reproduced by ML model



christoph.a.keller@nasa.gov

Speedup potential



- Offline evaluation of single forest is 1000x faster than numerical integration
- Current implementation is very inefficient (2x slower than full chemistry)
- Currently working on seamless integration of XGBoost

christoph.a.keller@nasa.gov

Summary

- Tree models do a decent job at simulating atmospheric chemistry
- Adding constraints (e.g., chemical families) to the machine learning model is critical
- Potential applications:
  - Chemical data assimilation
  - (Short-term) 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.

christoph.a.keller@nasa.gov