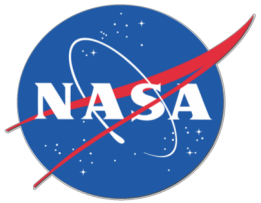
Accelerated simulation of air pollution using NVIDIA RAPIDS

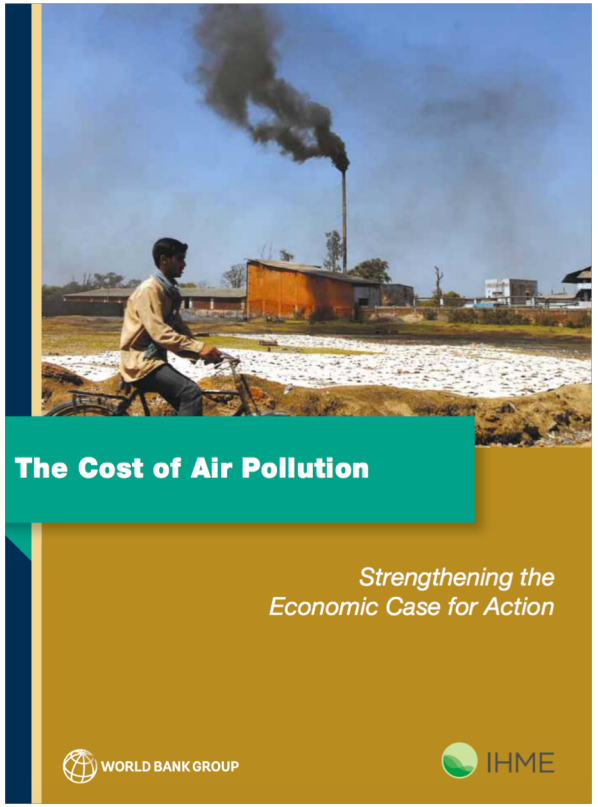
Christoph A. Keller[1,2], Thomas L. Clune[1], Matthew A. Thompson[1,3], Matthew A. Stroud[1,4], Mat J. Evans[5], Zahra Ronaghi[6]

[1]NASA Global Modeling and Assimilation Office (GMAO)
[2]Universities Space Research Association (USRA)
[3]Science Systems and Applications, Inc (SSAI)
[4]ASRC Federal Inuteq, [5]University of York, [6]NVIDIA

GPU Technology Conference :: 4-6 November 2019 :: Washington, DC



Air pollution is a global problem - mitigating it is a big opportunity



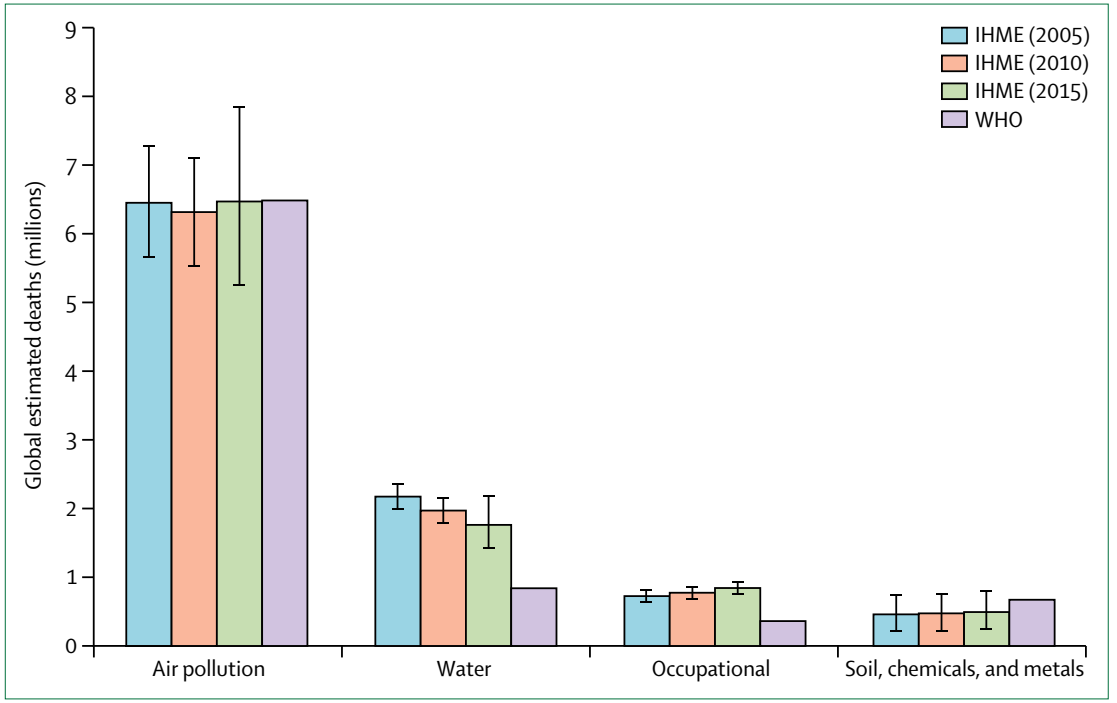


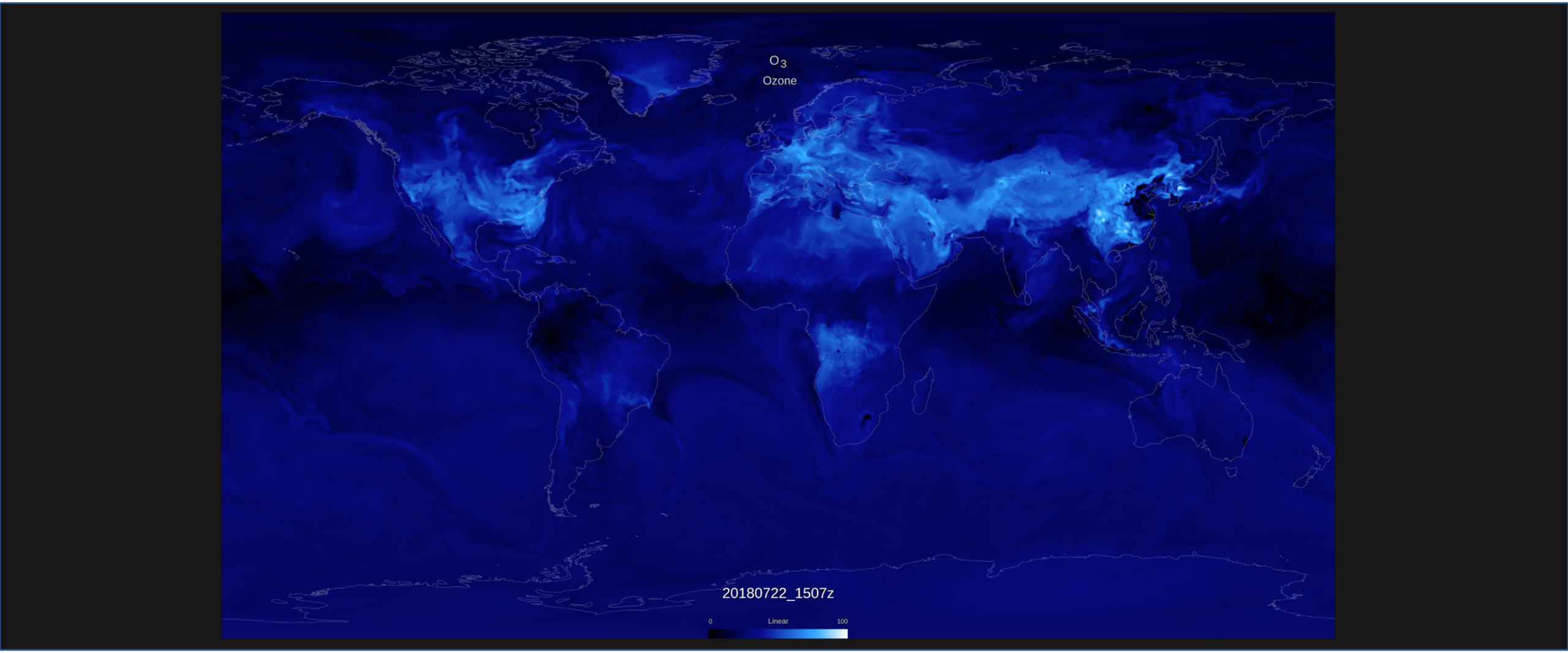
Figure 4: **Global estimated deaths (millions) by pollution risk factor, 2005–15**
Using data from the GBD study[42] and WHO.[99] IHME=Institute for Health Metrics and Evaluation.

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

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

christoph.a.keller@nasa.gov

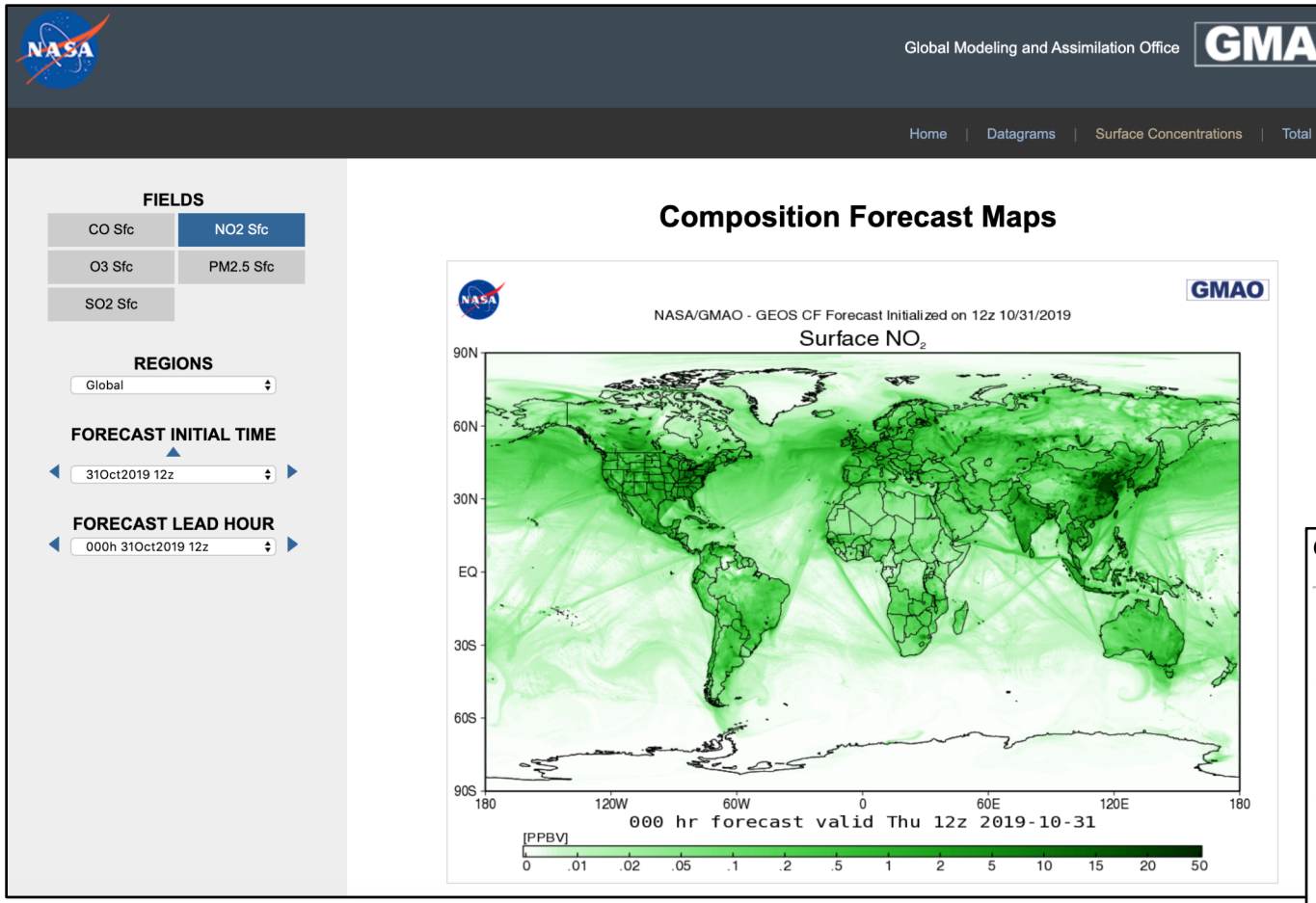
Numerical simulation of atmospheric chemistry



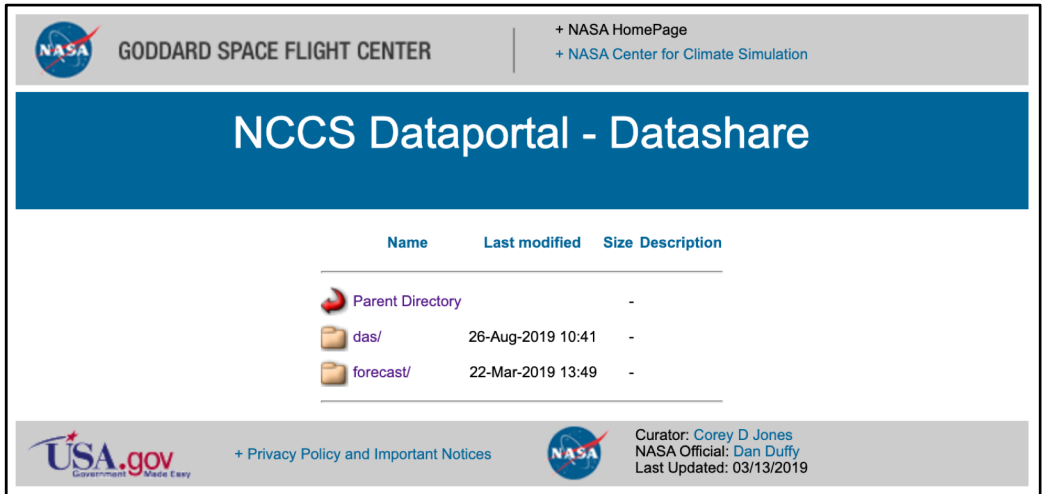
- 56 million grid cells (25x25 km^2, 72 levels), 250 chemical species

christoph.a.keller@nasa.gov

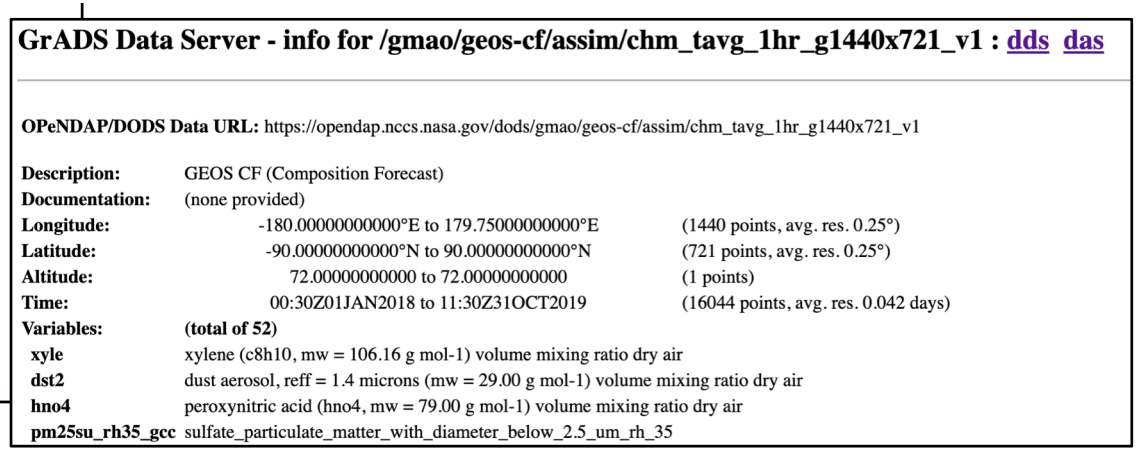
NASA's GEOS composition forecast (GEOS-CF) model conducts global air quality simulations in near real-time



https://fluid.nccs.nasa.gov/cf/



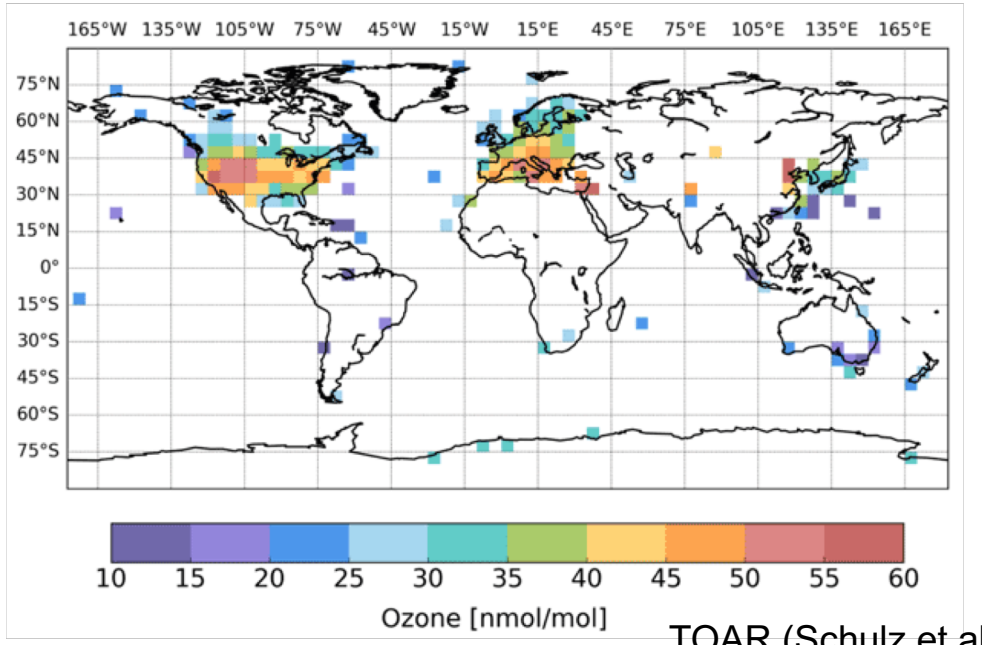
https://portal.nccs.nasa.gov/datashare/gmao/geos-cf/v1/



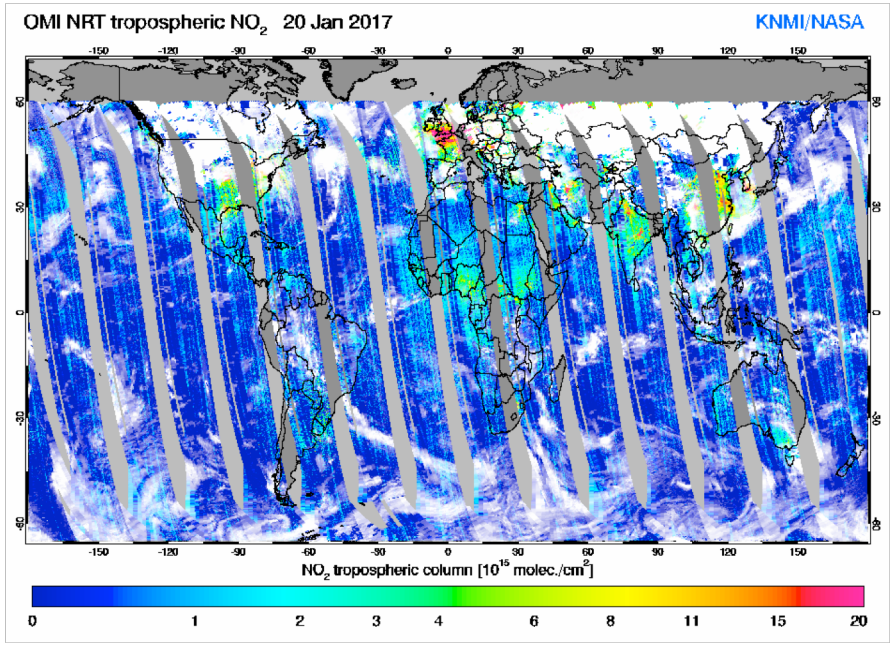
https://opendap.nccs.nasa.gov/dods/gmao/geos-cf/

christoph.a.keller@nasa.gov

Need models to fill temporal and spatial gaps in observations



TOAR (Schulz et al., 2017)

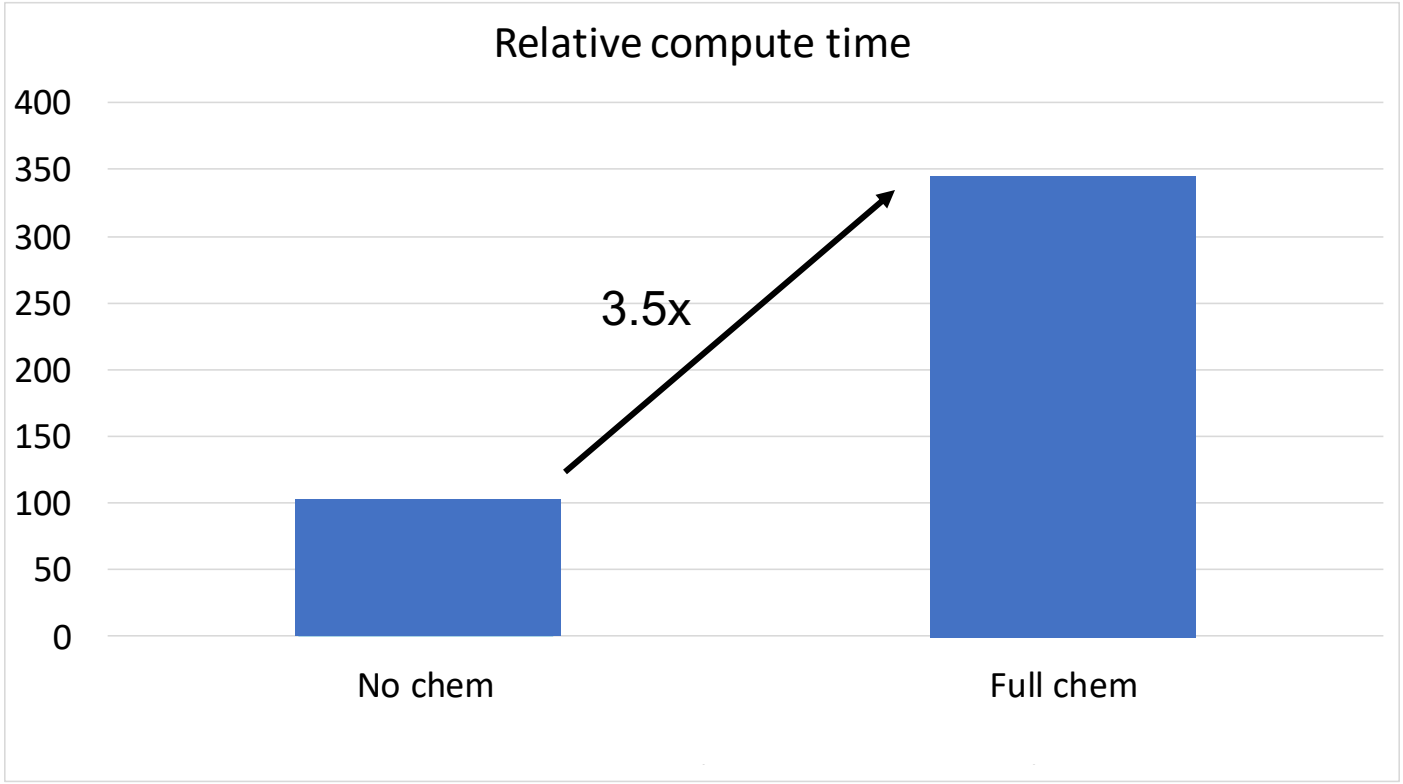


Surface observations are not global

Satellite observations are also discontinuous

christoph.a.keller@nasa.gov

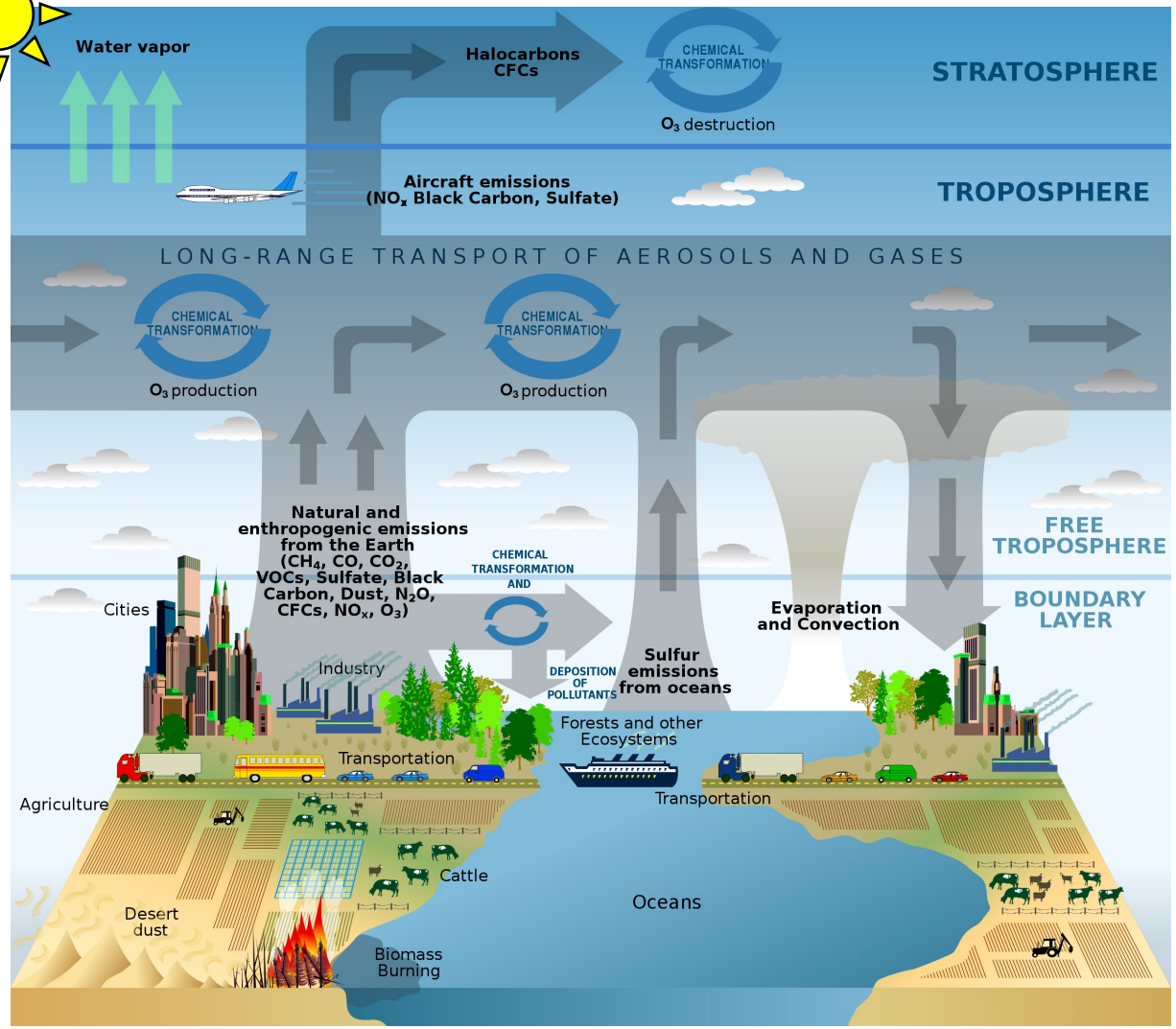
Atmospheric chemistry models are computationally expensive

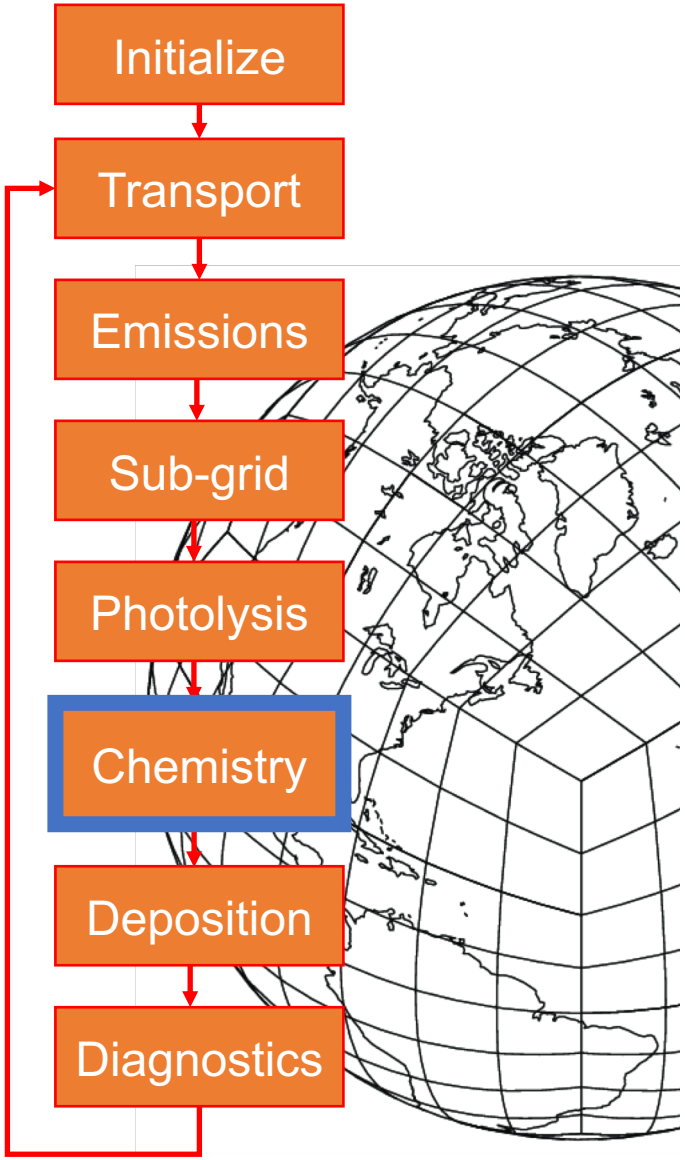


- High-resolution chemistry simulation requires >1000 CPU's
- Throughput: approx. 20 simulation days in 24 hours
- Outputting the full chemical state: ~1.5 TB / simulation day

christoph.a.keller@nasa.gov

Numerical simulation of atmospheric composition

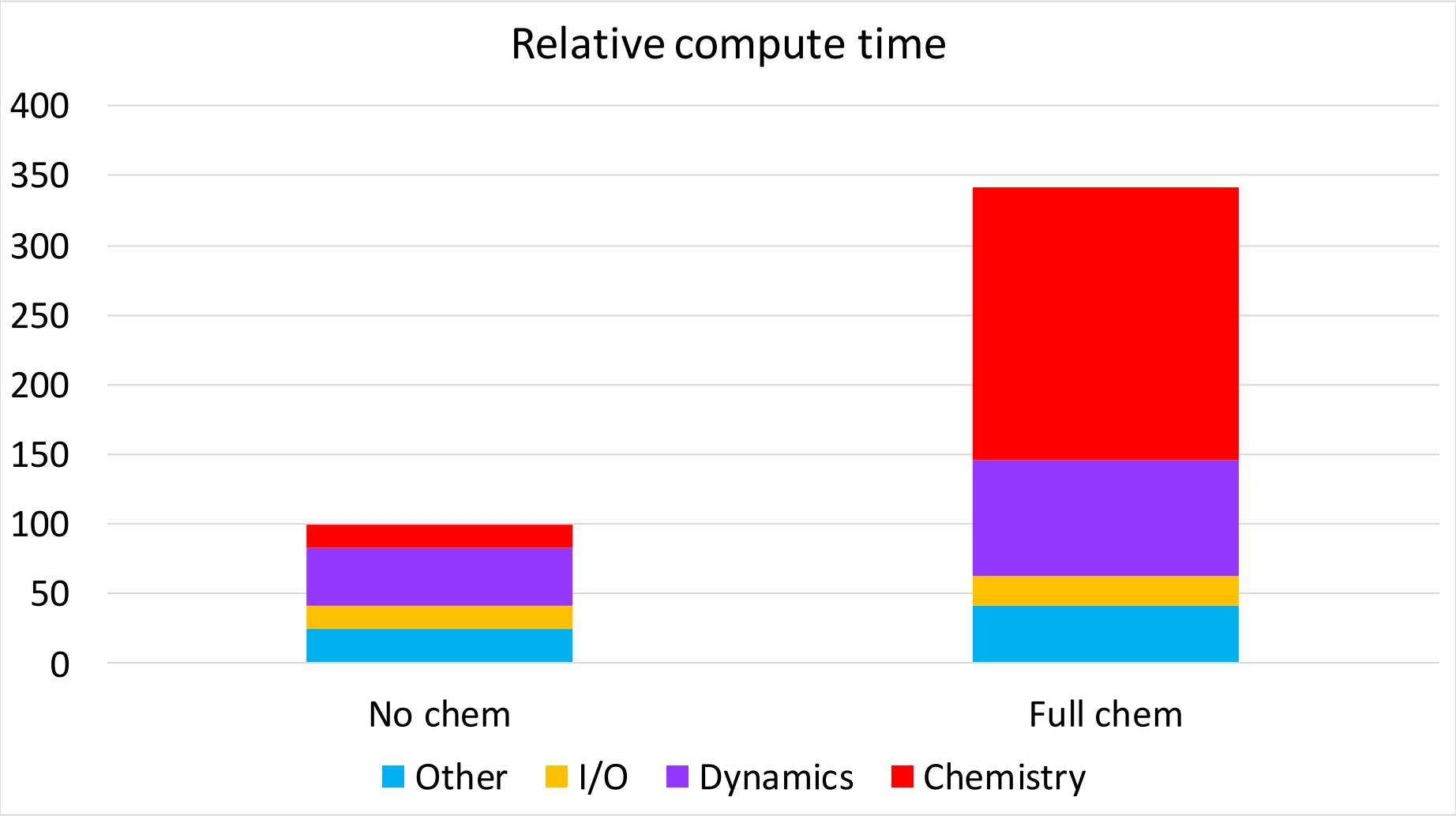




https://digital.library.unt.edu/ark:/67531/metadc11954/m1/37/

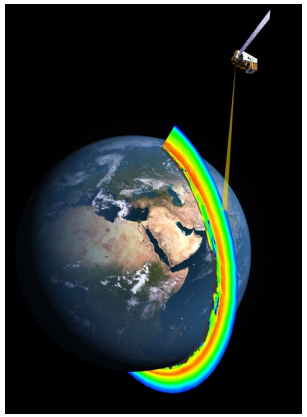
christoph.a.keller@nasa.gov

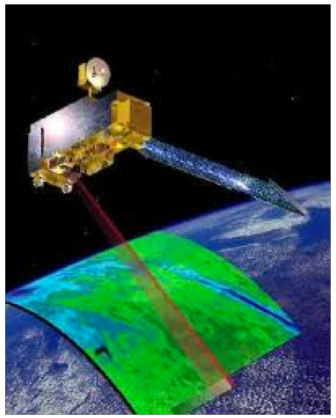
Chemistry accounts for more than 50% of compute time



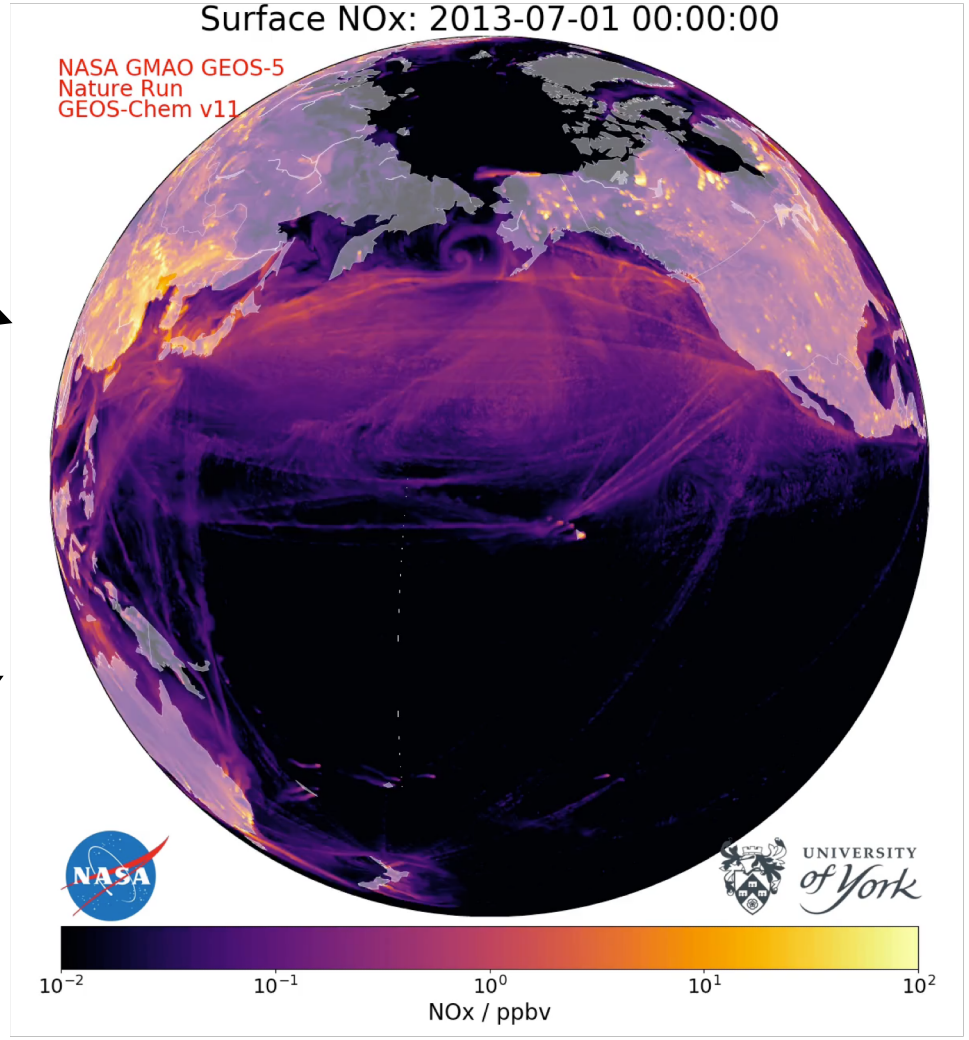
christoph.a.keller@nasa.gov

High computational cost of chemistry currently prevents optimal use of observations

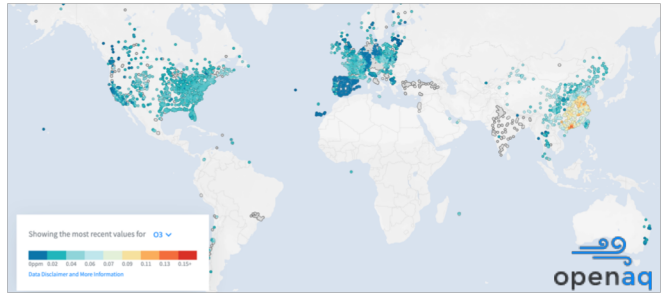




www.nasa.gov



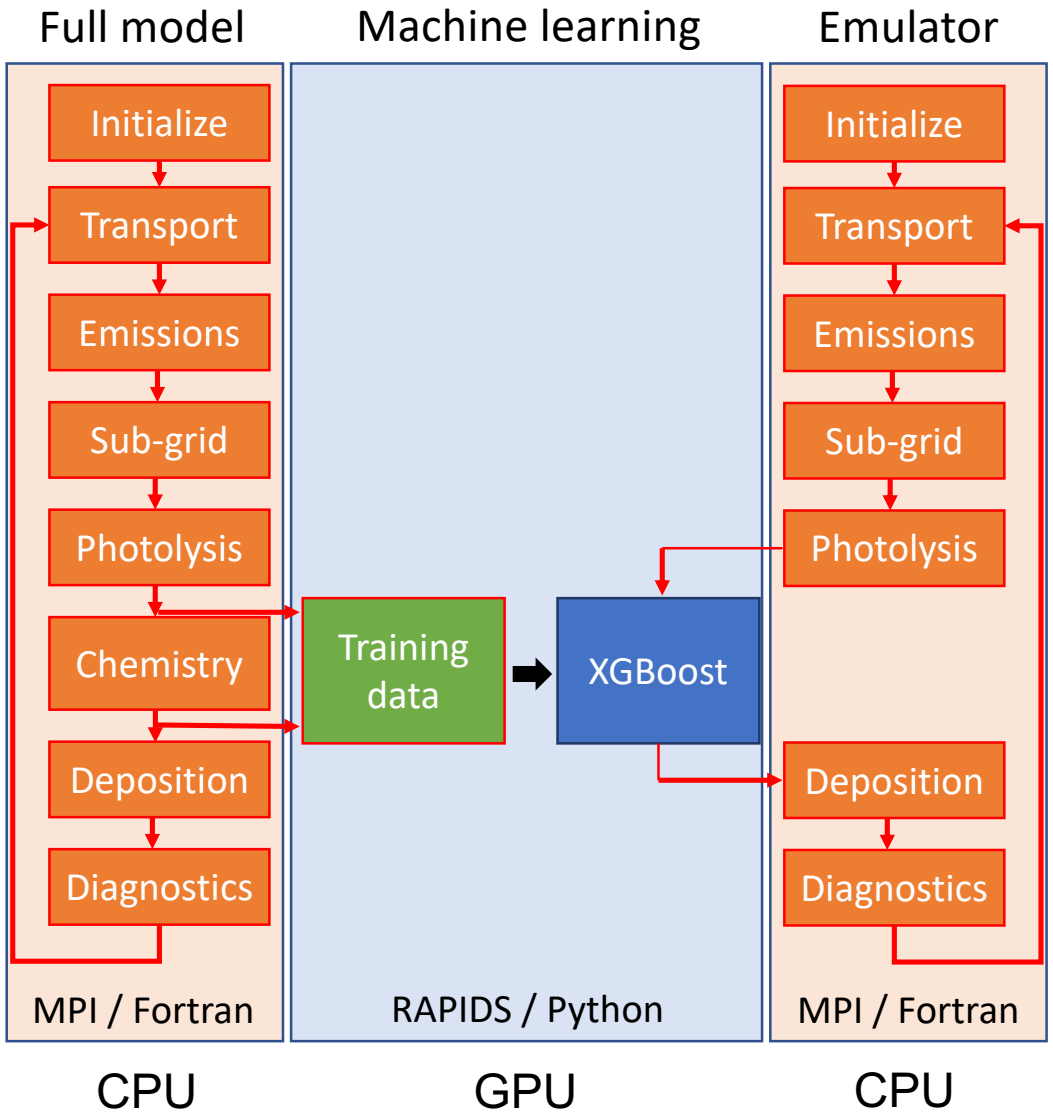
purpleair.com



openaq.org

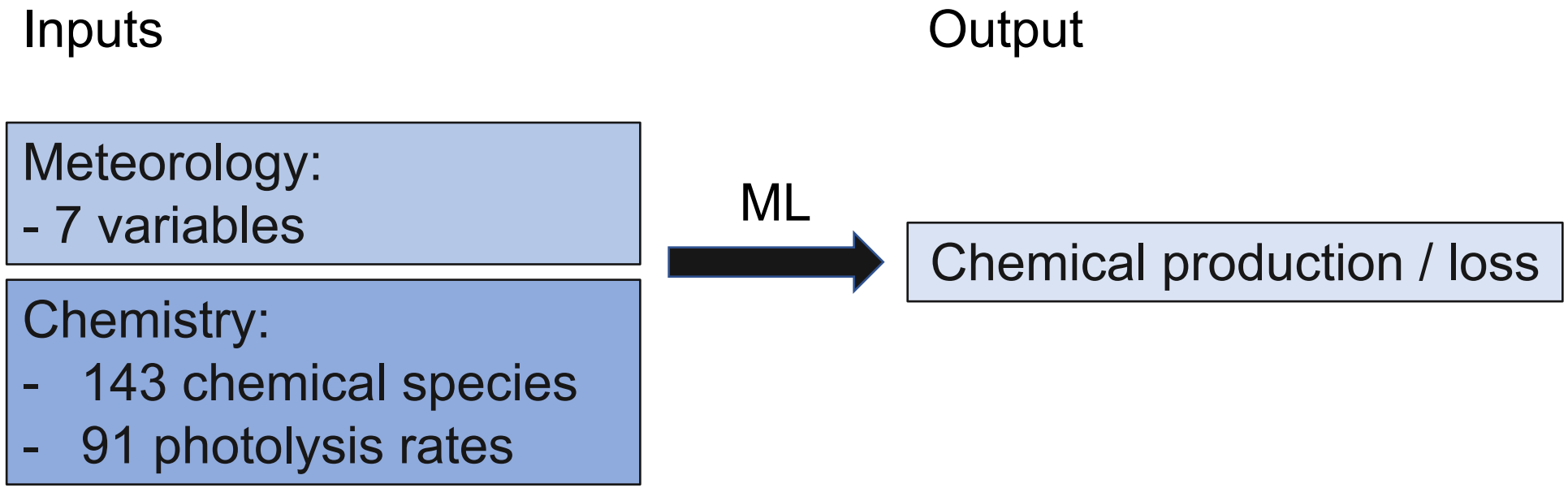
christoph.a.keller@nasa.gov

Replace slow chemical integrator with machine learning model



christoph.a.keller@nasa.gov

Use machine learning to emulate chemical transformations in the atmosphere

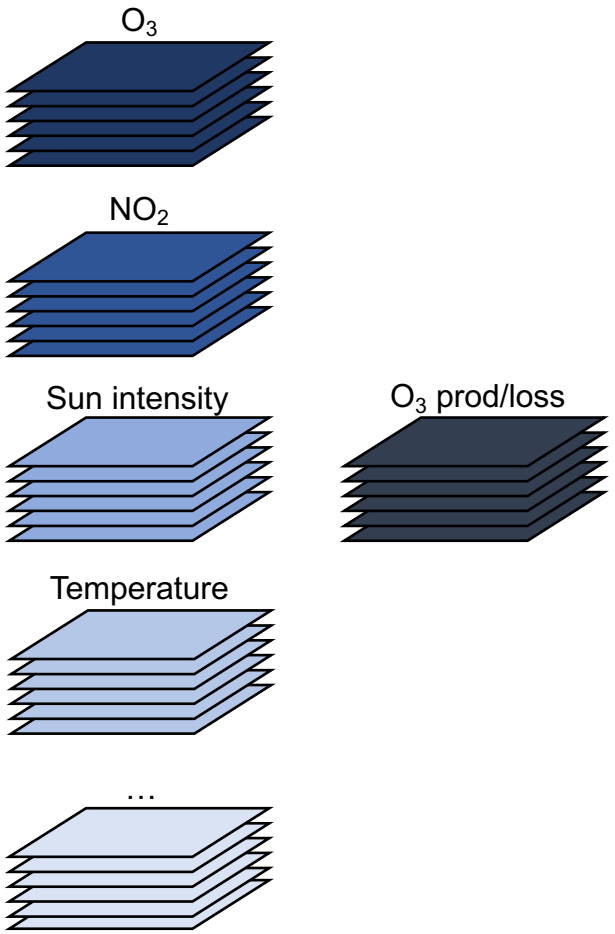


- Algorithm: extreme gradient boosted decision trees (XGBoost)
- Train separate algorithm for each species

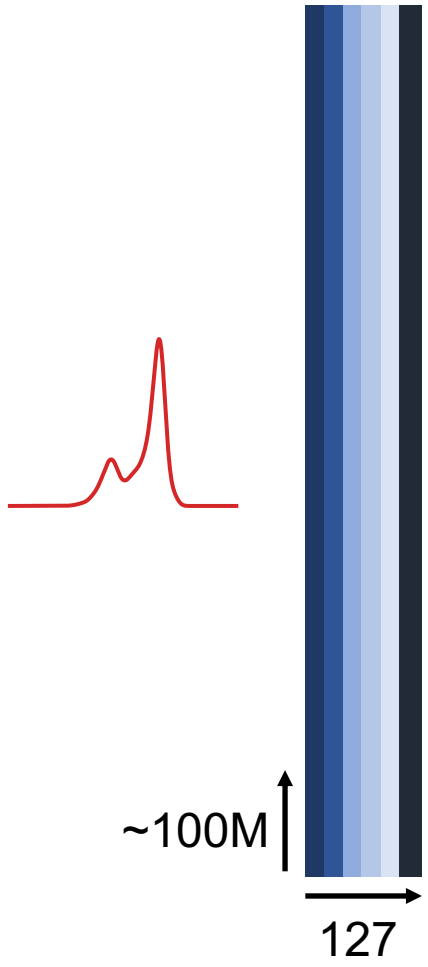
christoph.a.keller@nasa.gov

Machine learning workflow

Training data (netCDF format)



Subsample & flatten, write to csv (xarray)



Train (XGBoost):

- Read csv, convert to DMatrix
- Train

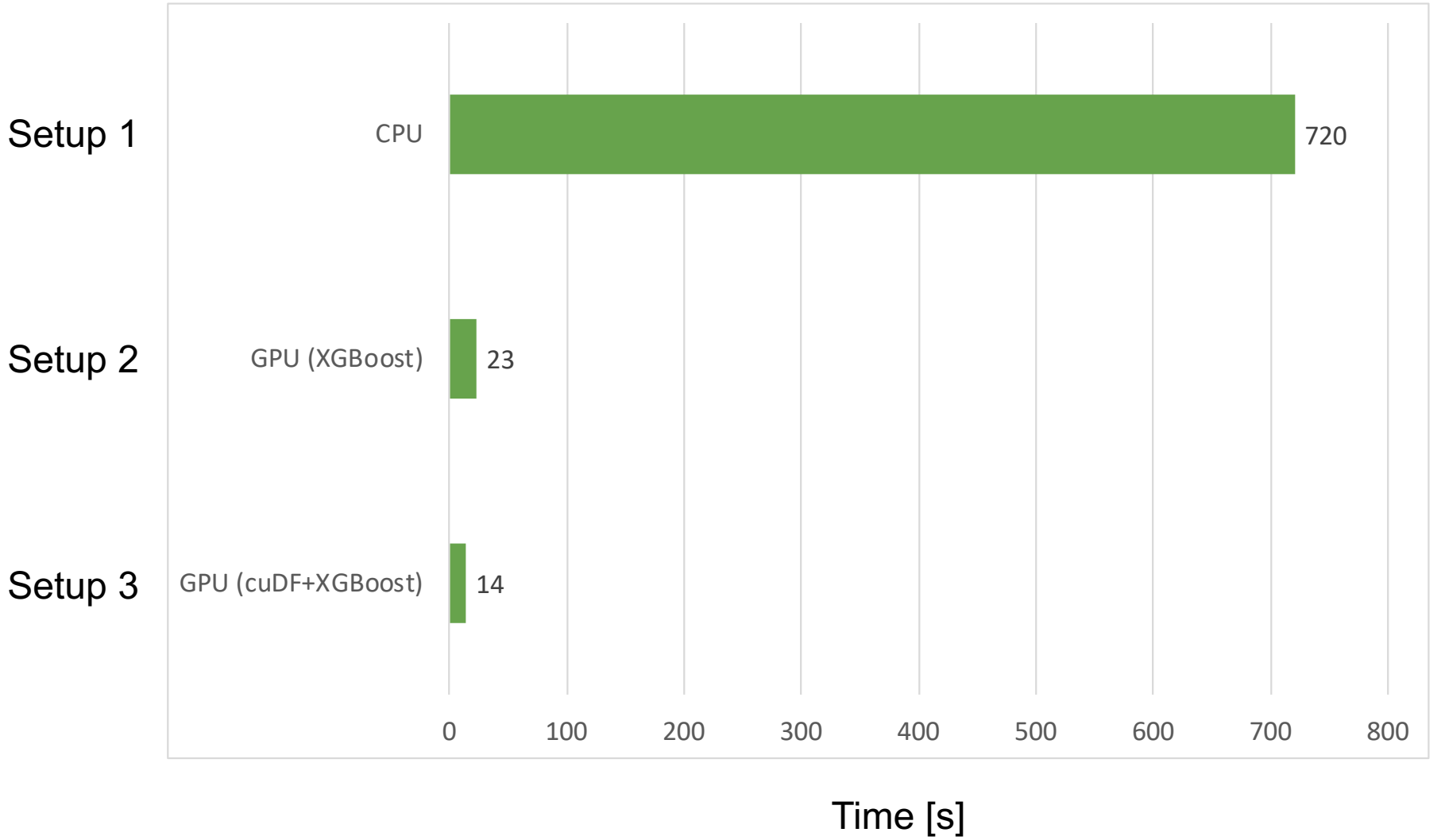
Setup 1
Read on CPU (Intel Haswell)
Train on CPU

Setup 2:
Read on CPU
Train on GPU (V100)

Setup 3:
Read on GPU (cuDF/cuIO)
Train on GPU (dask-XGBoost)

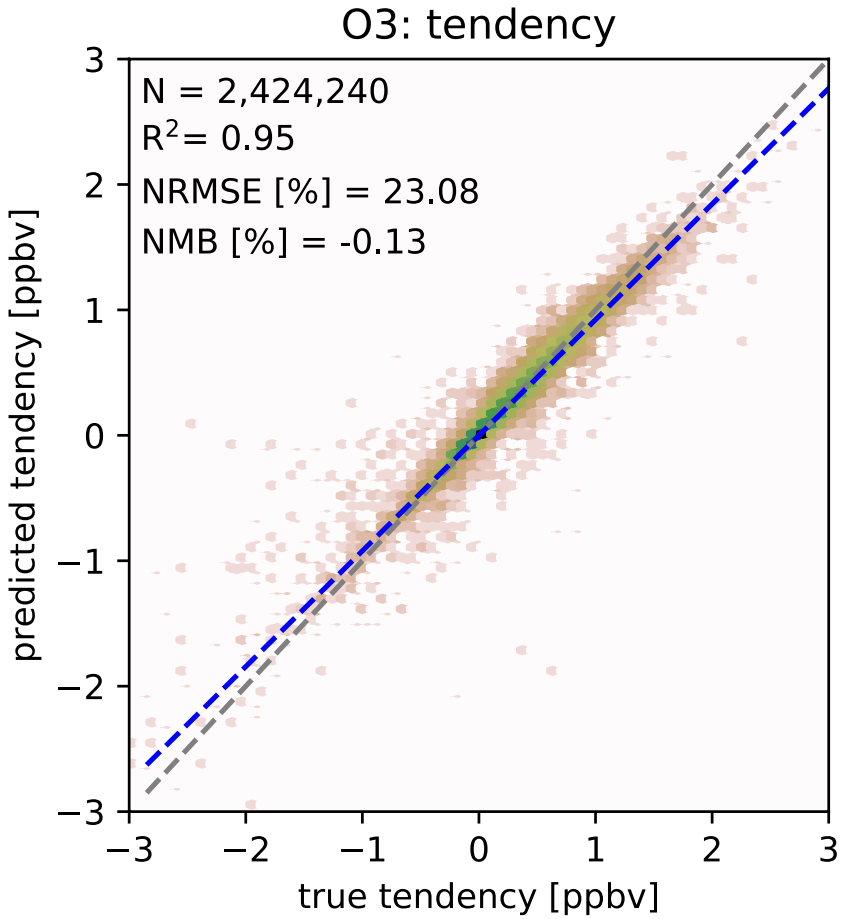
christoph.a.keller@nasa.gov

XGBoost training benchmarks



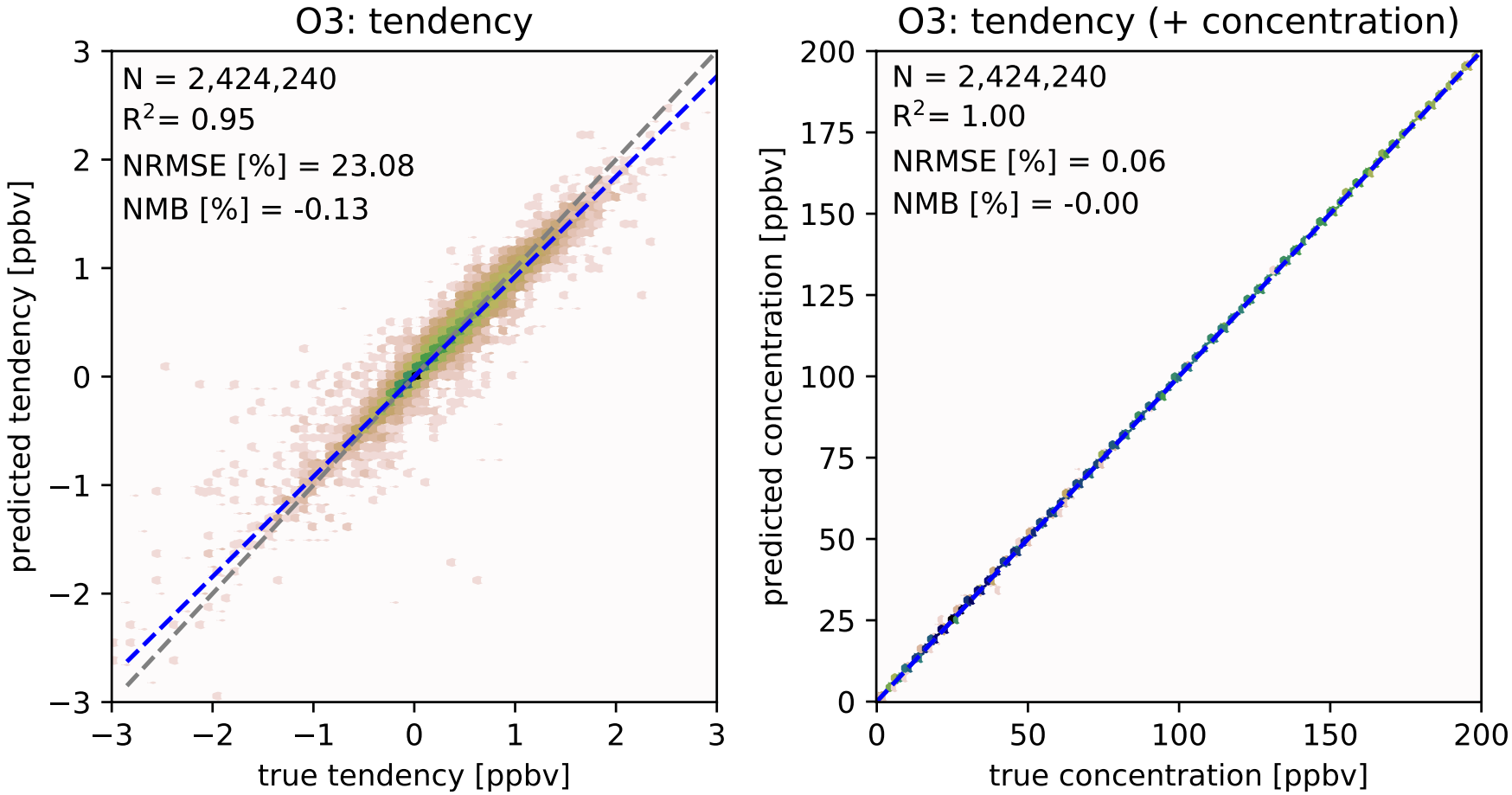
christoph.a.keller@nasa.gov

XGBoost reproduces target concentrations well (single-step prediction)



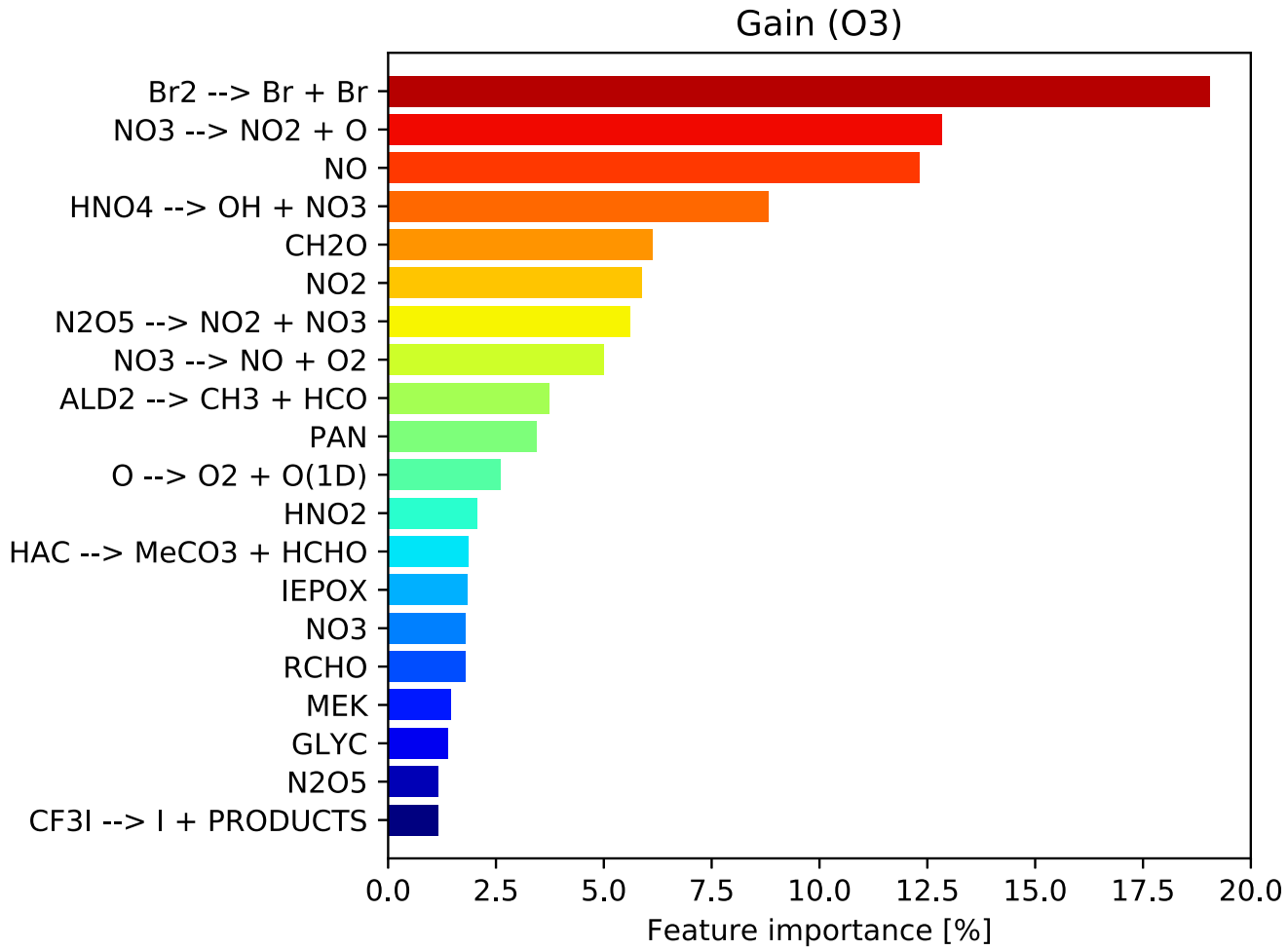
christoph.a.keller@nasa.gov

XGBoost reproduces target concentrations well (single-step prediction)



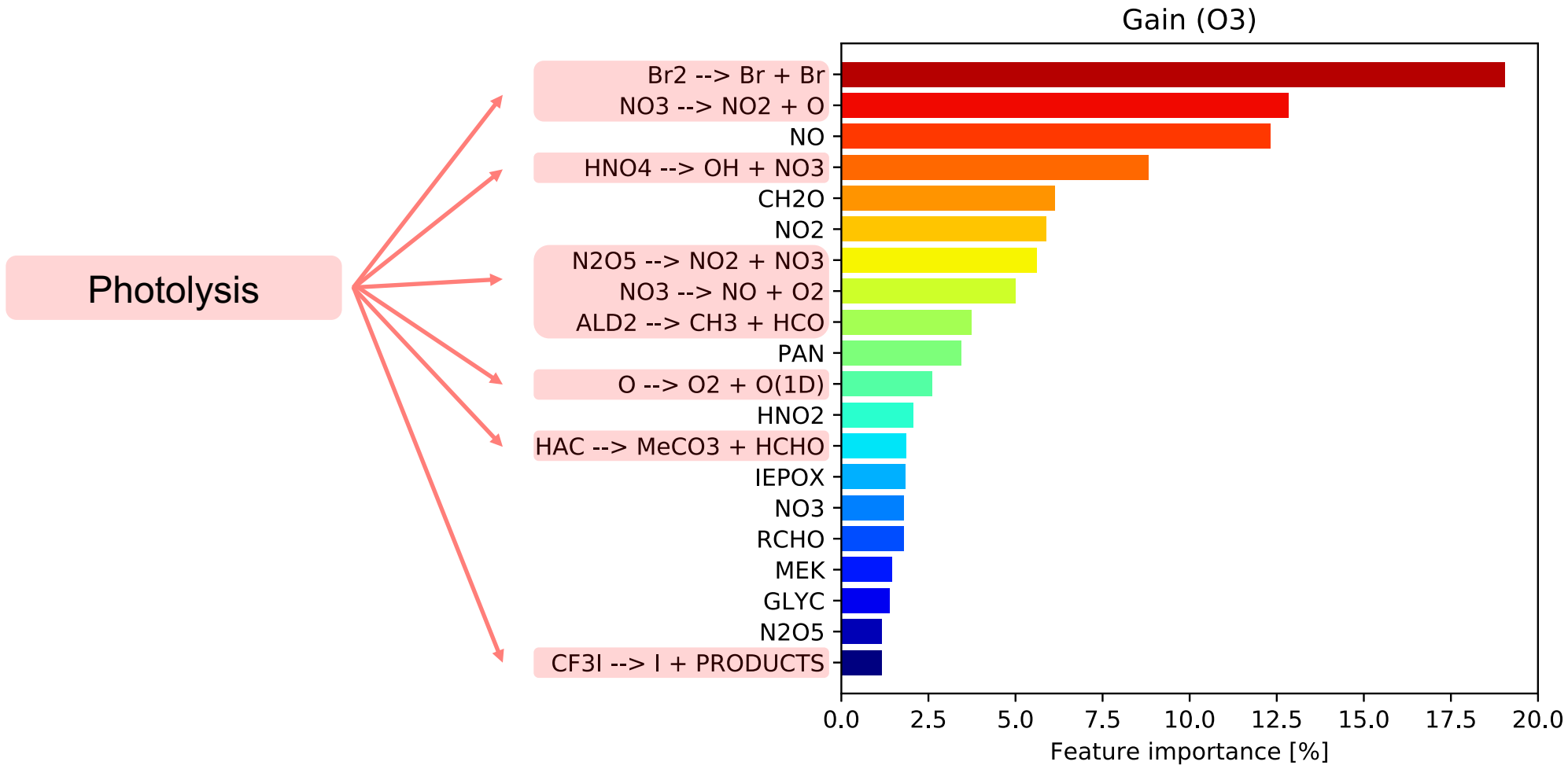
christoph.a.keller@nasa.gov

XGBoost solution reflects known features of chemical kinetics



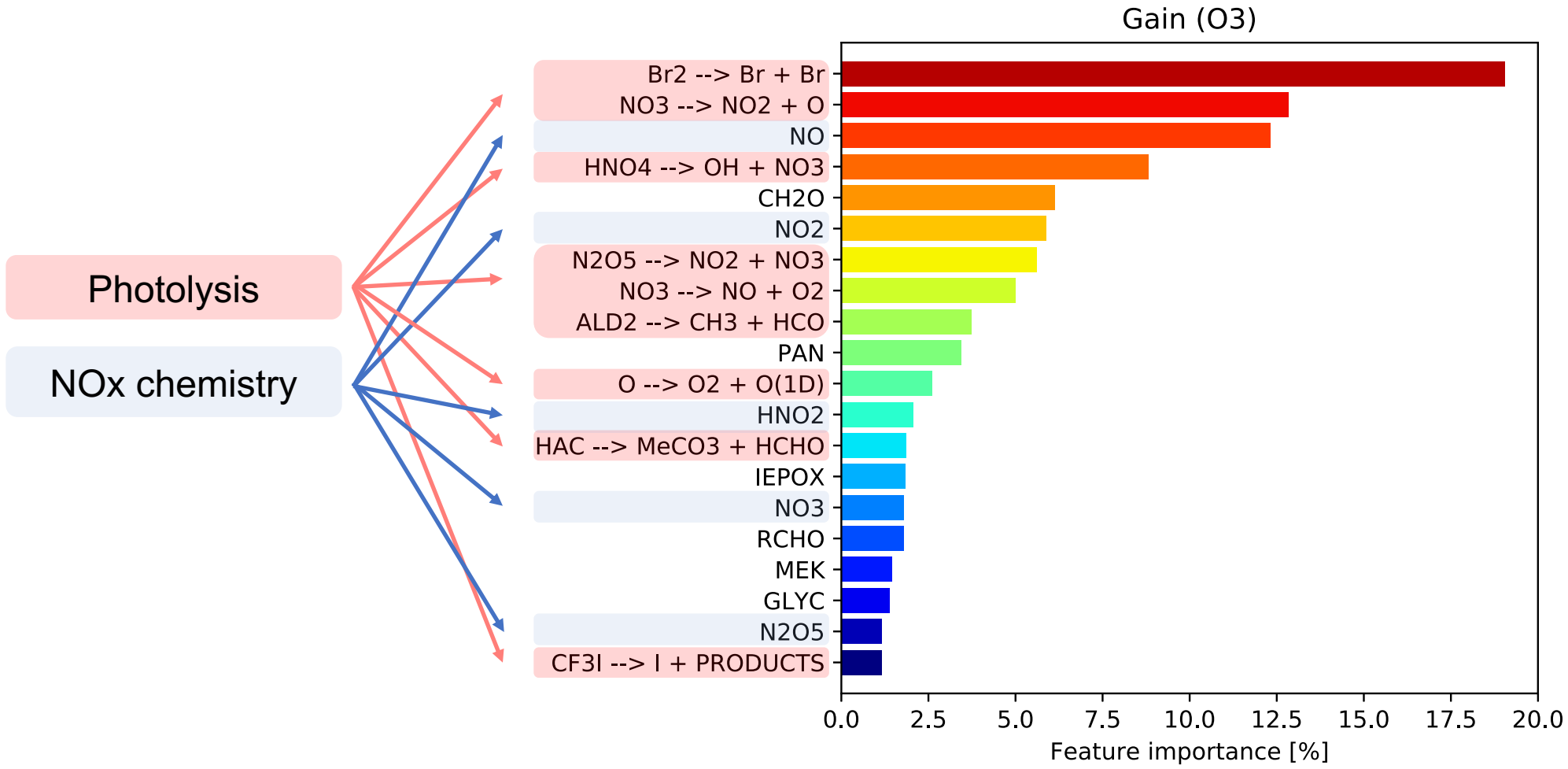
christoph.a.keller@nasa.gov

XGBoost solution reflects known features of chemical kinetics



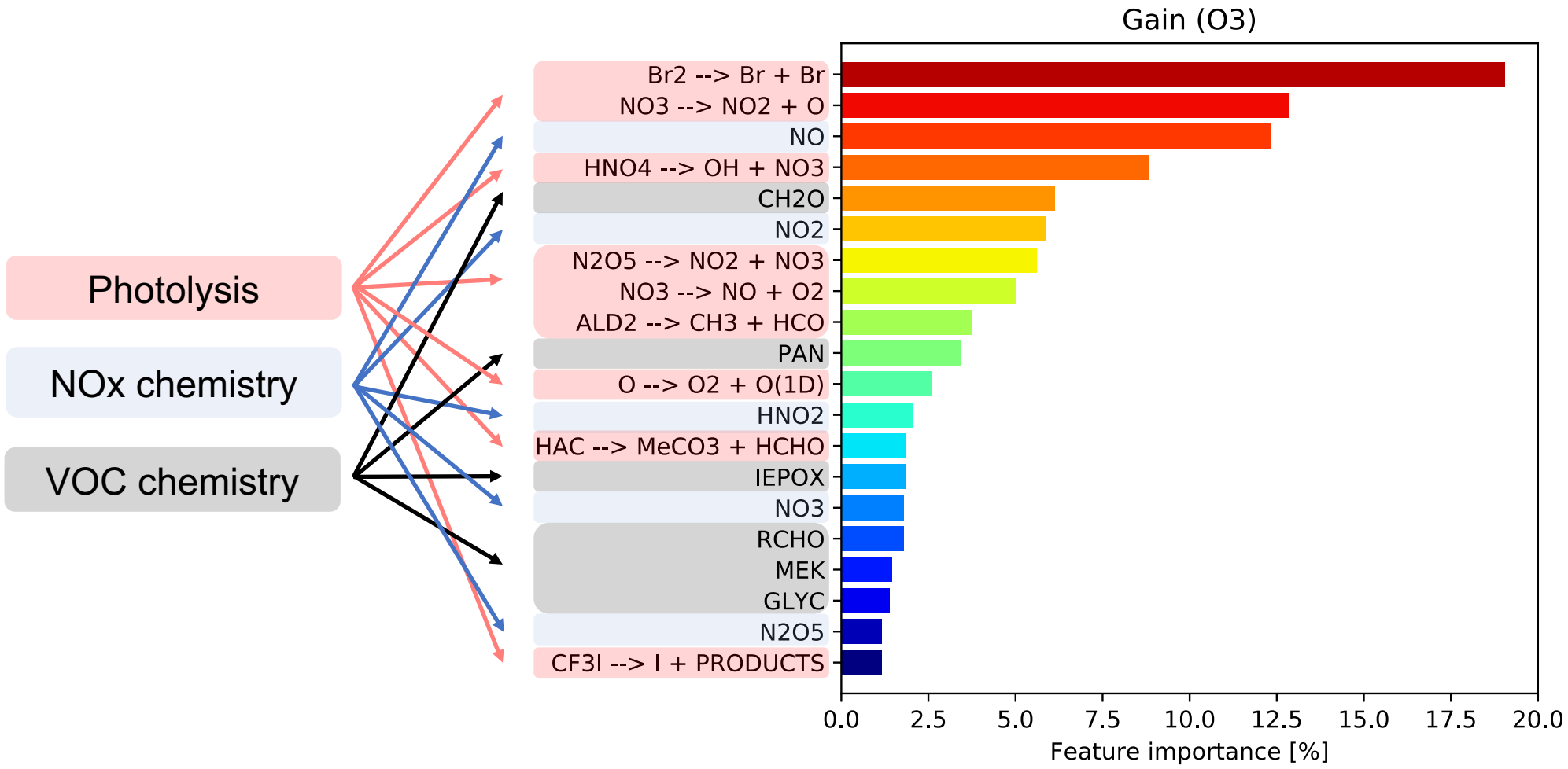
christoph.a.keller@nasa.gov

XGBoost solution reflects known features of chemical kinetics



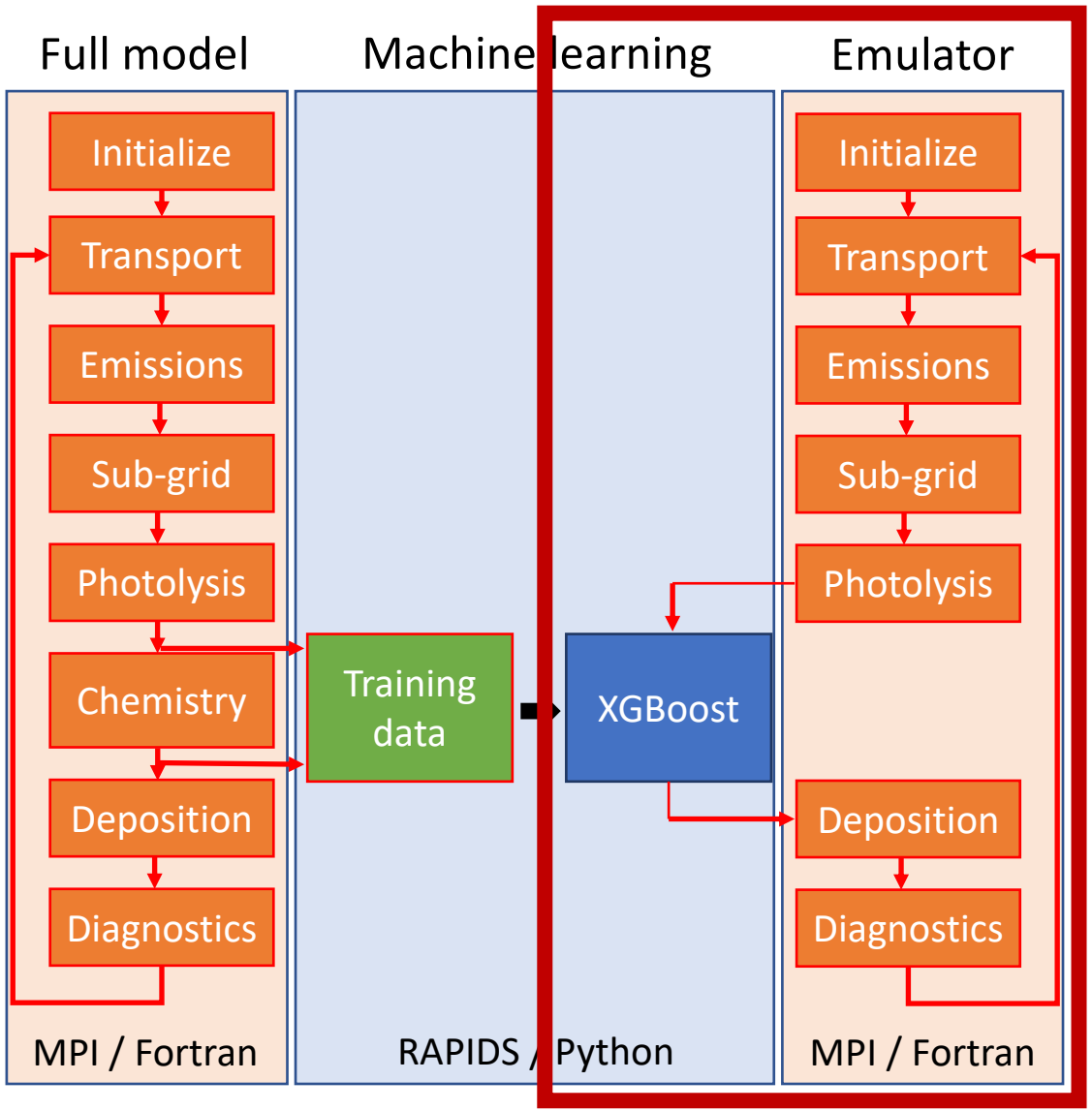
christoph.a.keller@nasa.gov

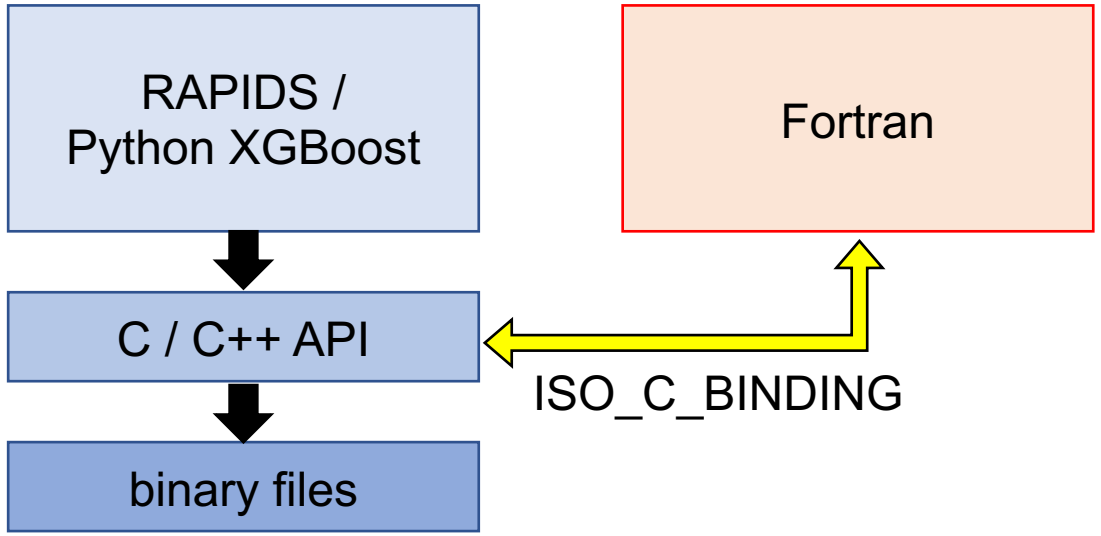
XGBoost solution reflects known features of chemical kinetics



christoph.a.keller@nasa.gov

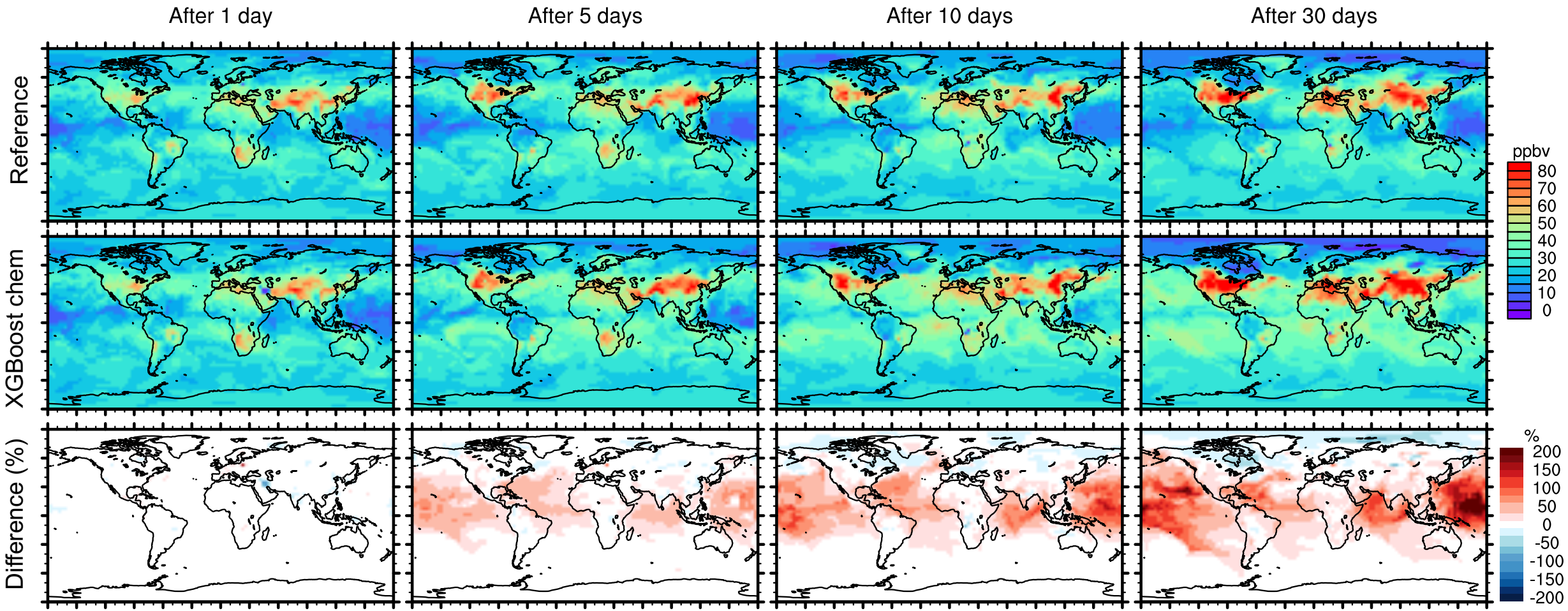
1-month simulation with XGBoost emulator





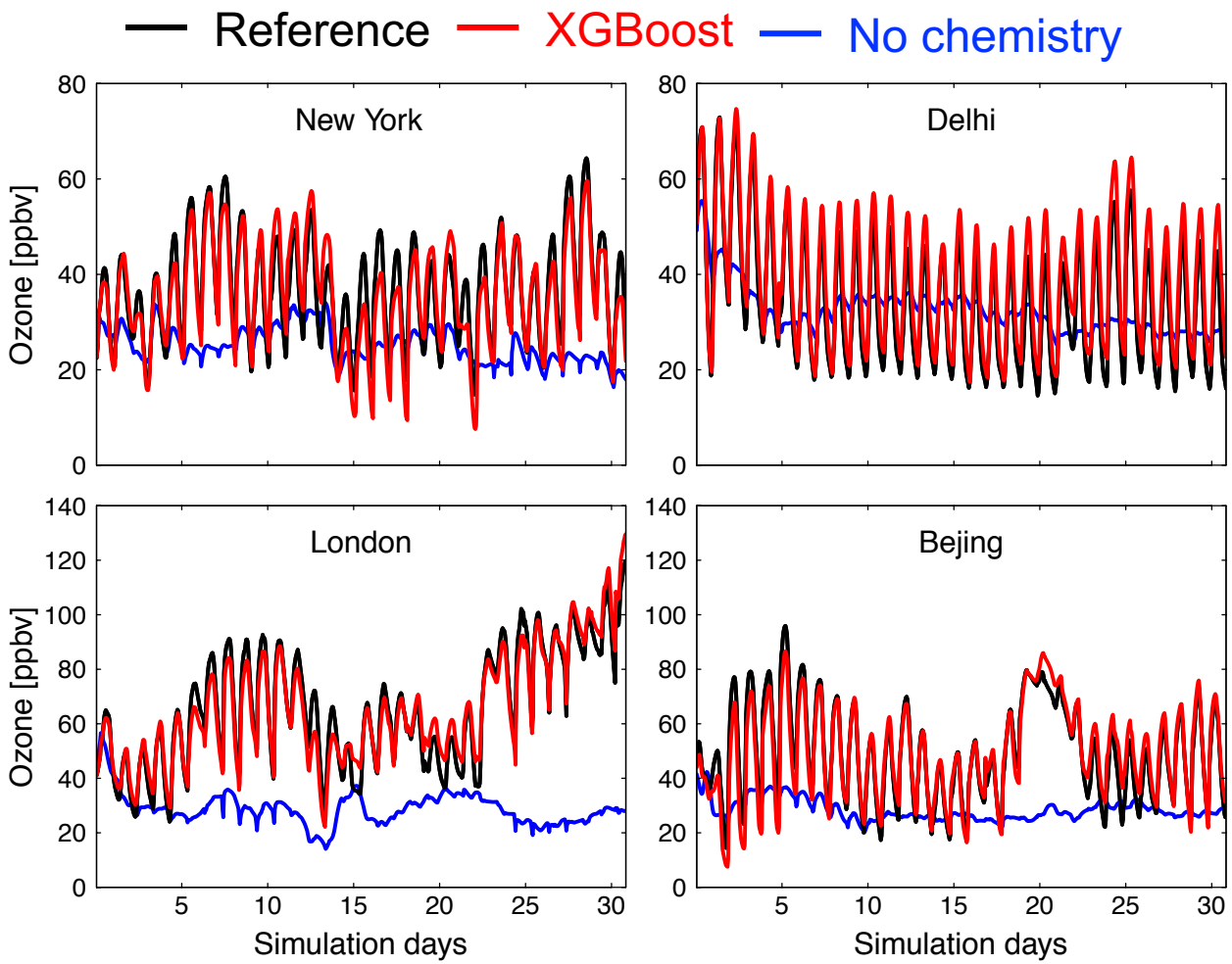
christoph.a.keller@nasa.gov

Emulator model is generally accurate, but overestimates ozone concentrations over remote regions



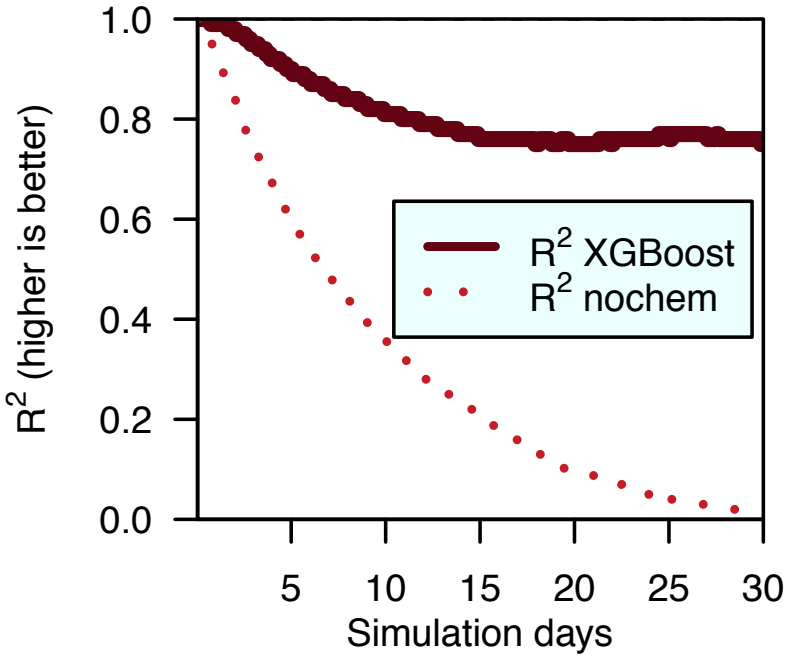
christoph.a.keller@nasa.gov

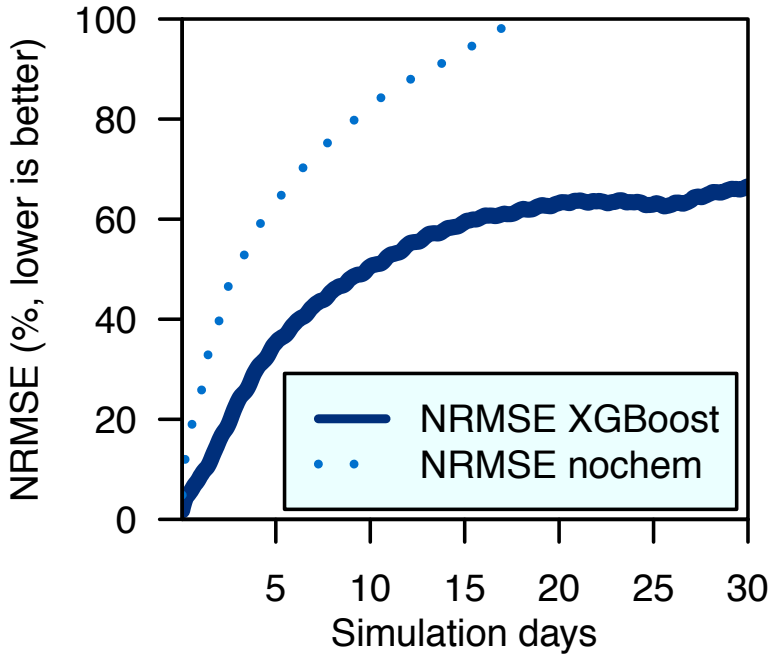
Surface concentrations over polluted regions are well reproduced by ML model

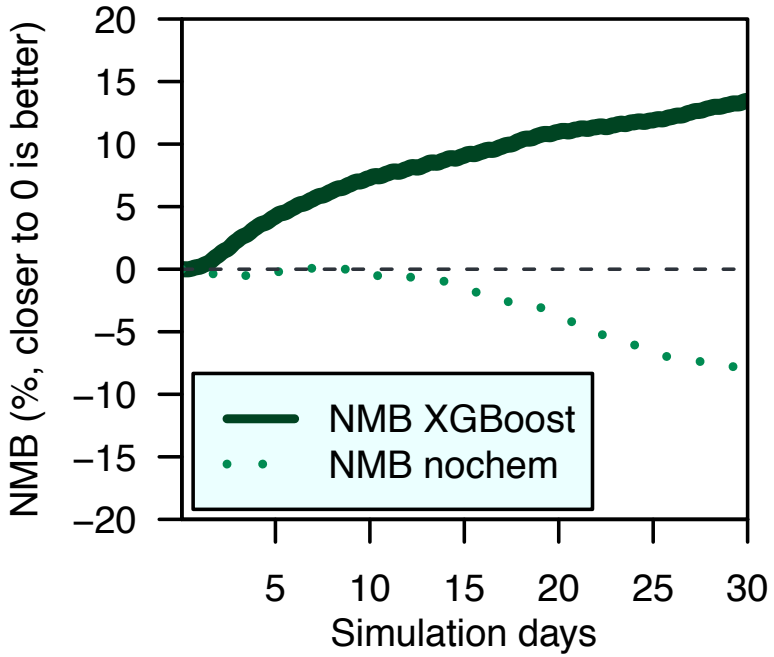


christoph.a.keller@nasa.gov

Machine learning model remains stable over the long-term

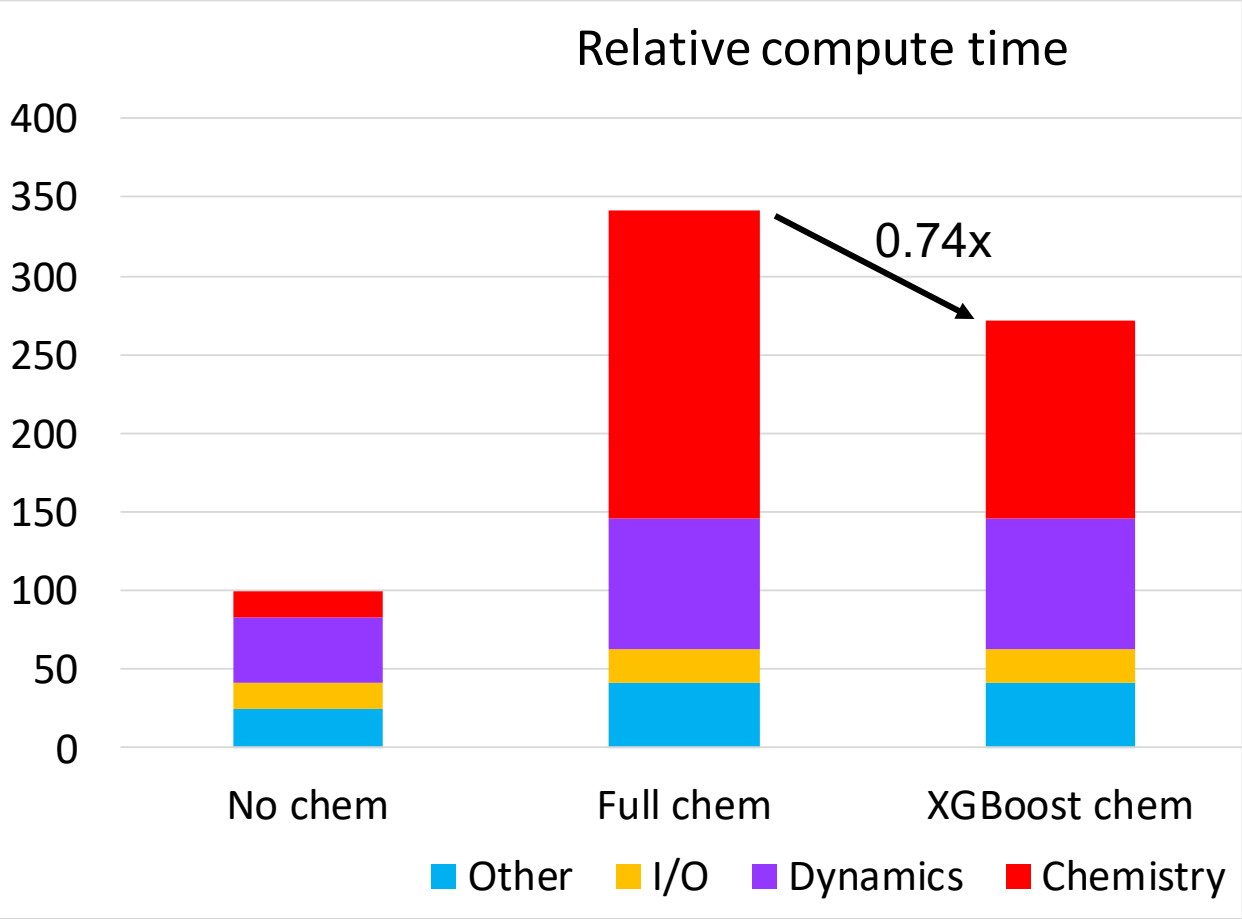


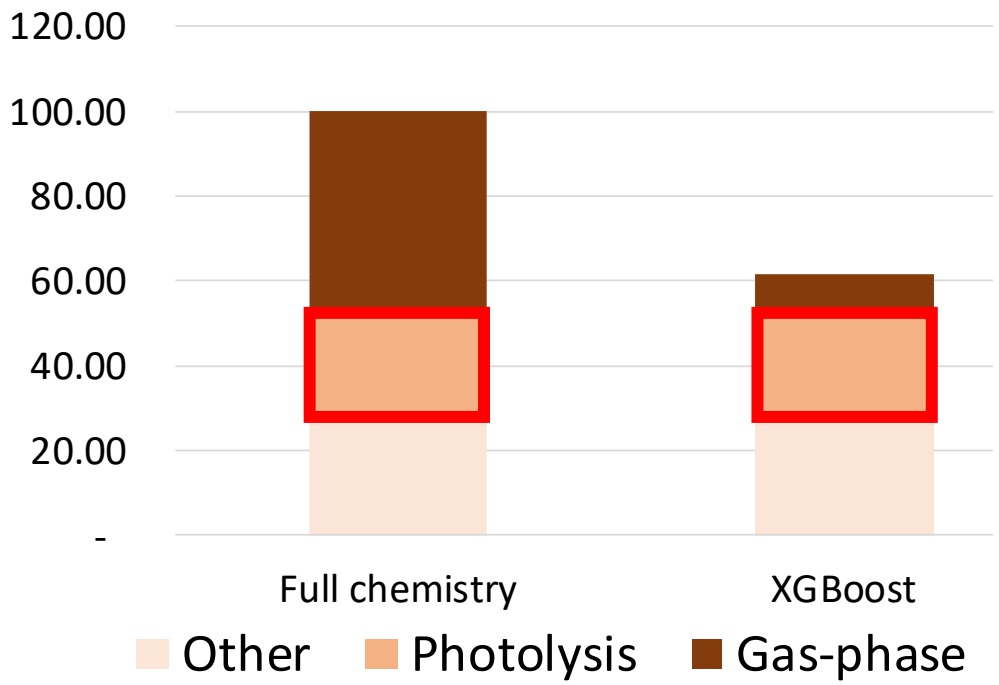




christoph.a.keller@nasa.gov

Model speedup

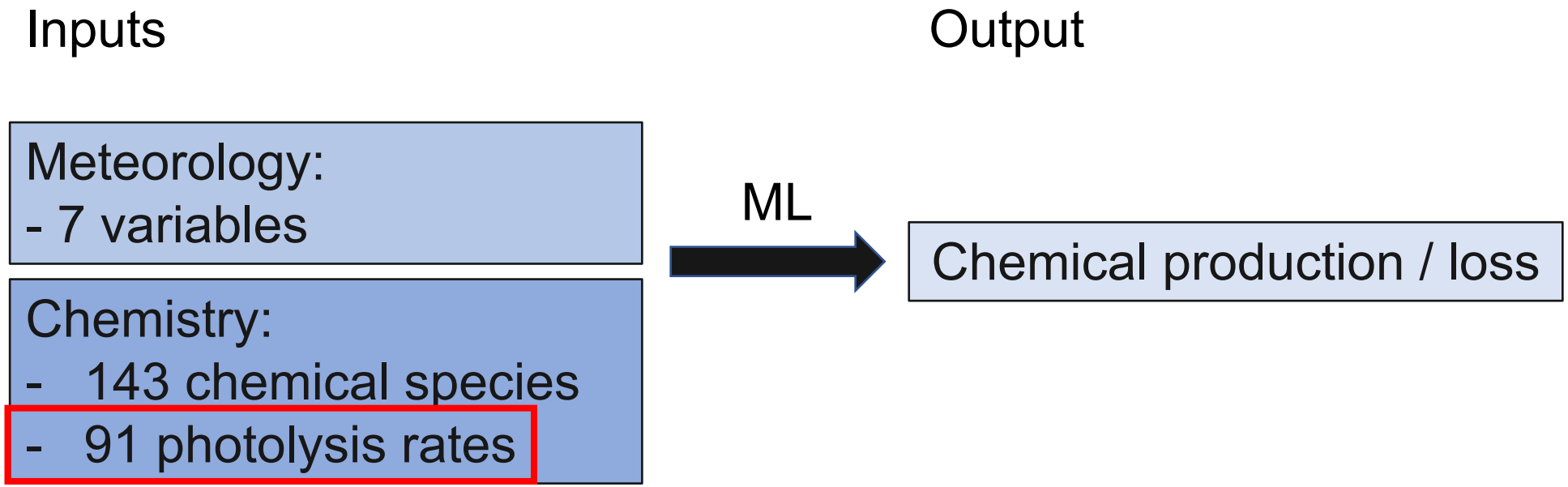




- XGBoost model is ~25% faster than reference model
- Chemistry is still slowest part of the model

christoph.a.keller@nasa.gov

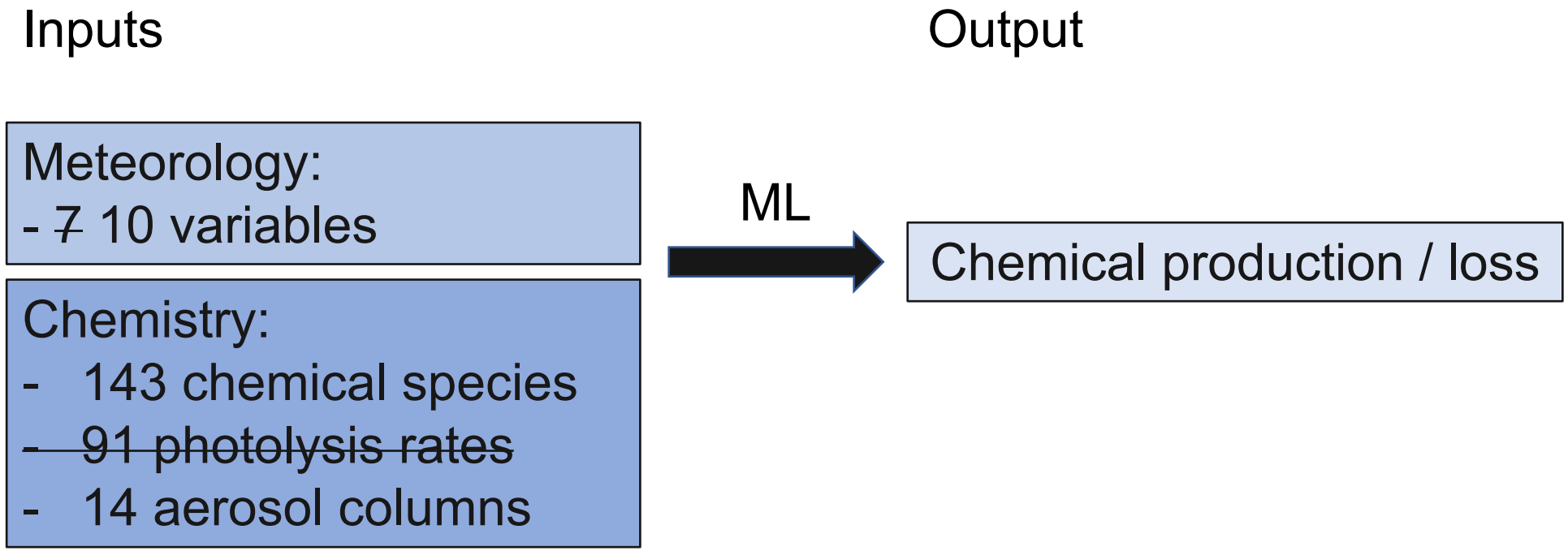
Incorporating photolysis calculation into the ML algorithm



- Original ML algorithm uses as input 91 photolysis rates

christoph.a.keller@nasa.gov

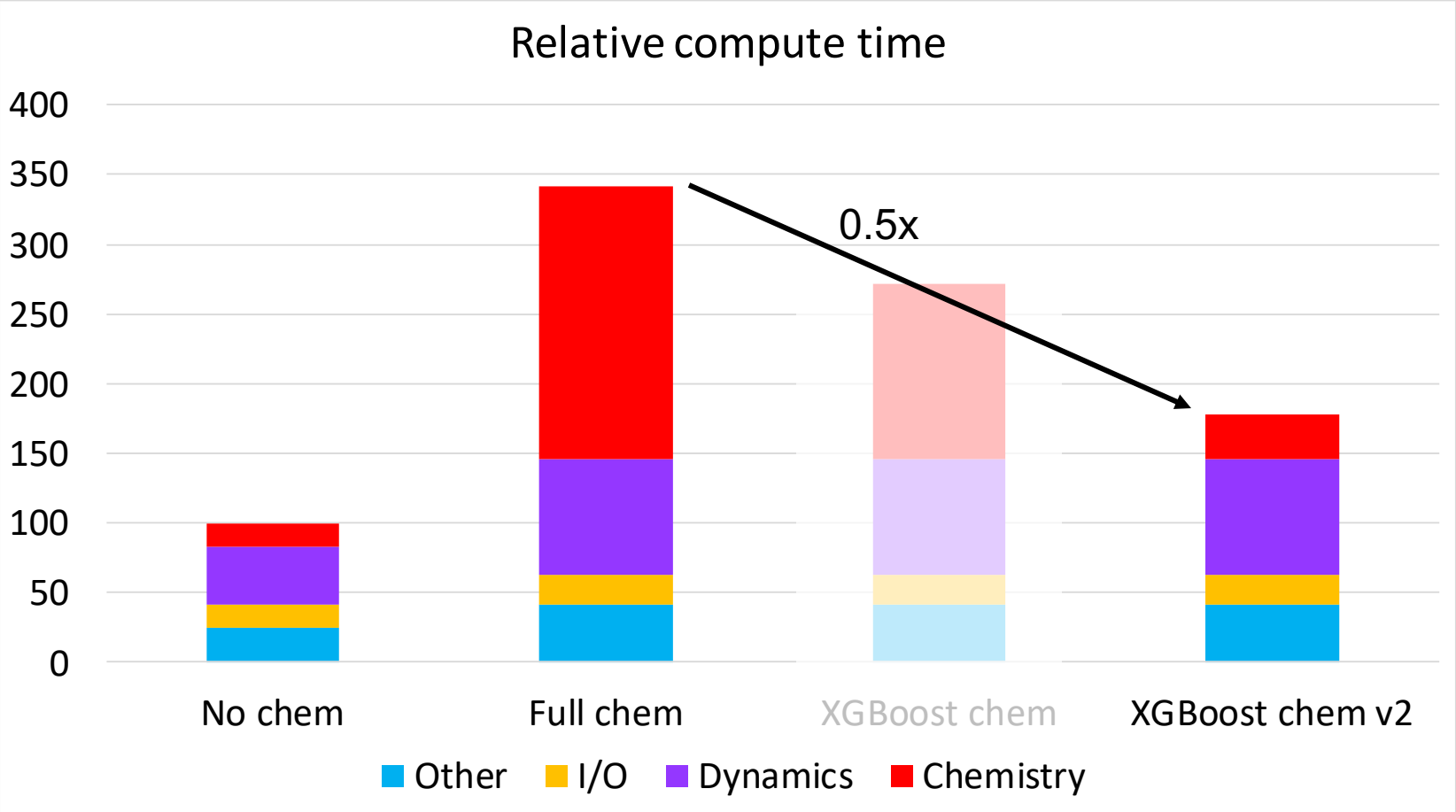
Incorporating photolysis calculation into the ML algorithm



- Original ML algorithm uses as input 91 photolysis rates
- Replace photolysis rates with quantities needed to compute photolysis

christoph.a.keller@nasa.gov

Model speedup with optimized XGBoost model



- XGBoost chemistry model is now ~2 times faster than reference model
- Chemistry >6x faster than before, dynamics becomes bottleneck

christoph.a.keller@nasa.gov

Summary

- Machine learning can help speed up air quality models by at least 2-5x
- Benefits:
  - Better use of satellite observations
  - Improve (short to medium-term) air quality forecasts
- Ongoing work:
  - Train on very large data sets (>1 TB)
  - Better coupling between CPU and GPUs (model side)
  - 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

christoph.a.keller@nasa.gov