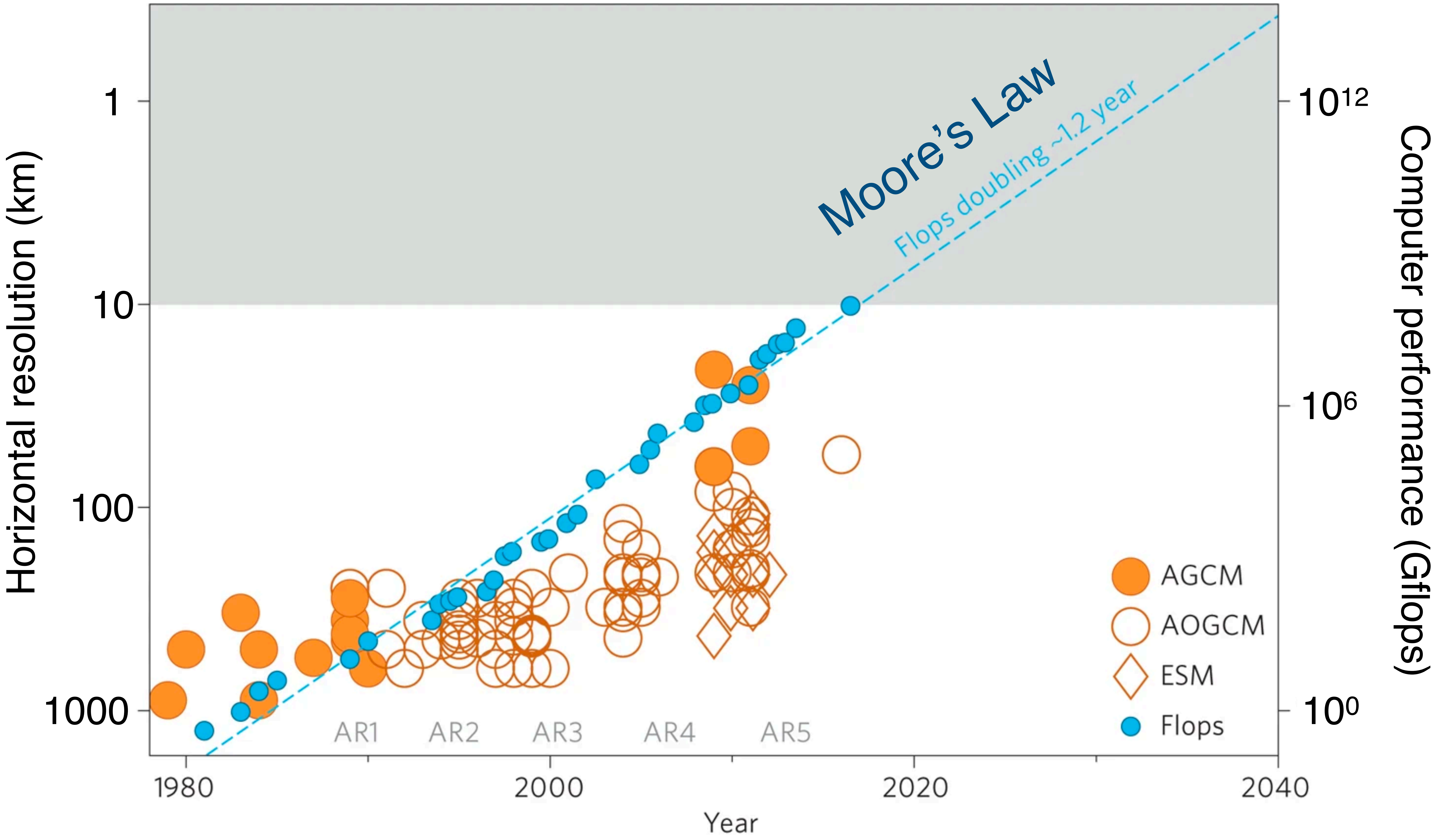
An Online-Learned Neural Network Chemical Solver for Stable Long-Term Global Simulations of Atmospheric Chemistry

Makoto Kelp

ACM Conference 20221208

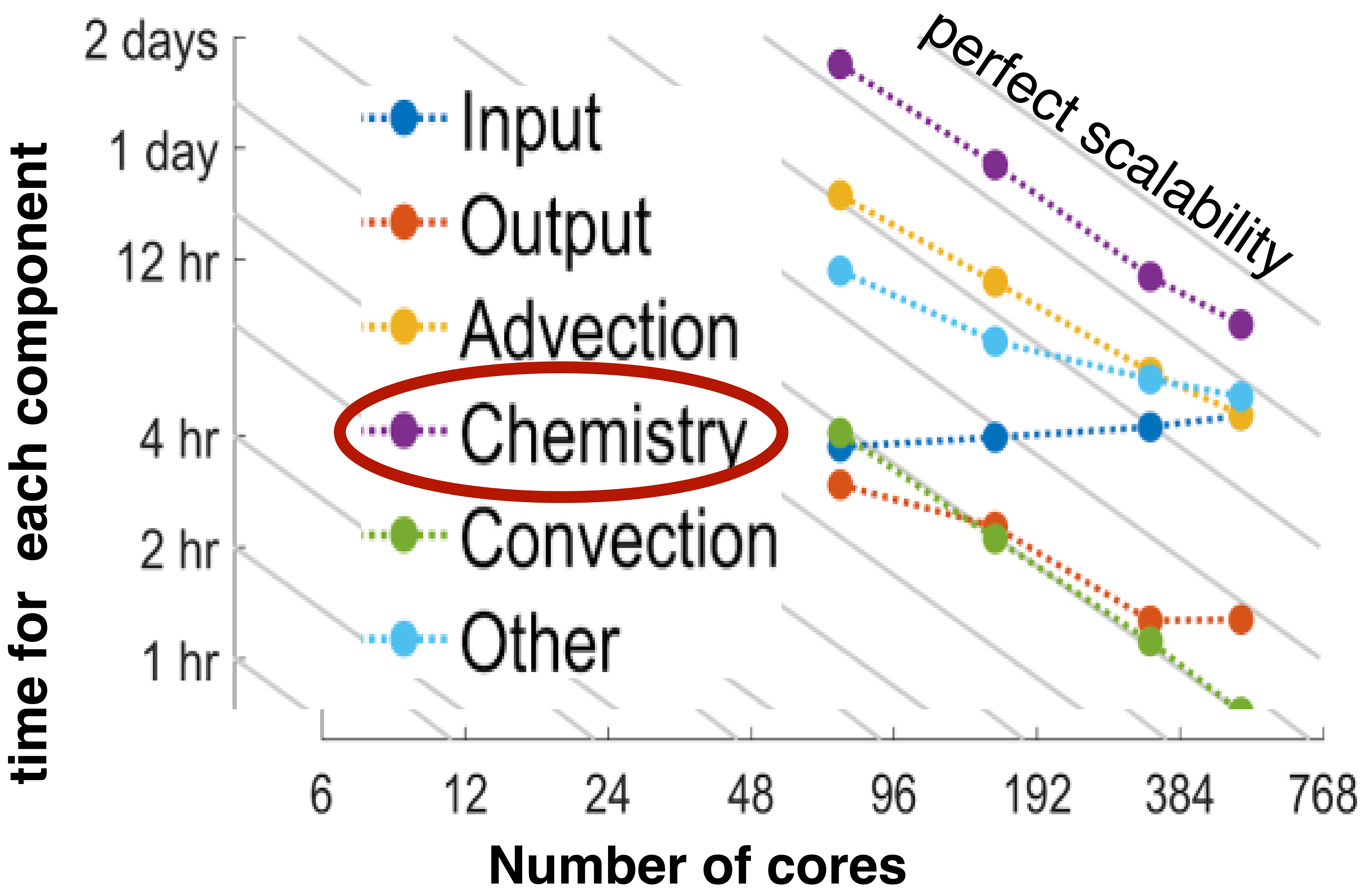
Scientific advancements outpace computational progress



Schneider et al., 2017 Nature Climate Change

Global modeling of atmospheric chemistry is a grand computational challenge

GCHP 1-month benchmark at 0.5° (C180) resolution



-Chemistry **dominates** the cost of a simulation (**~40%**) even though ideally scales

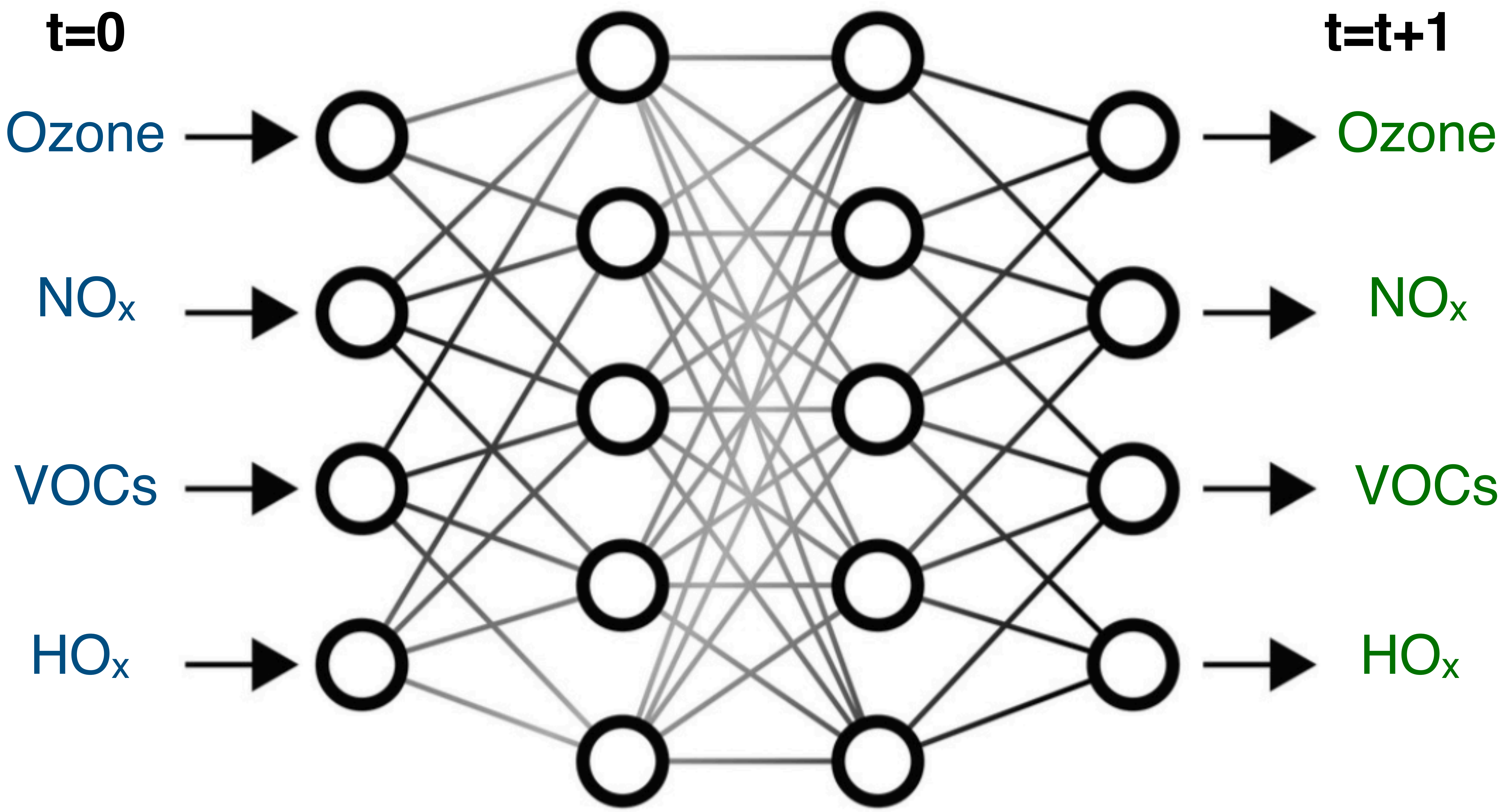
-Weather and climate models typically have **~4 variables**

-Chemistry models have **hundreds** of evolving species

Bottom Line: Adding chemistry into an Earth system model becomes computationally infeasible

Eastham et al., 2018 GMD

Machine learning (ML) methods can provide a solution to this problem

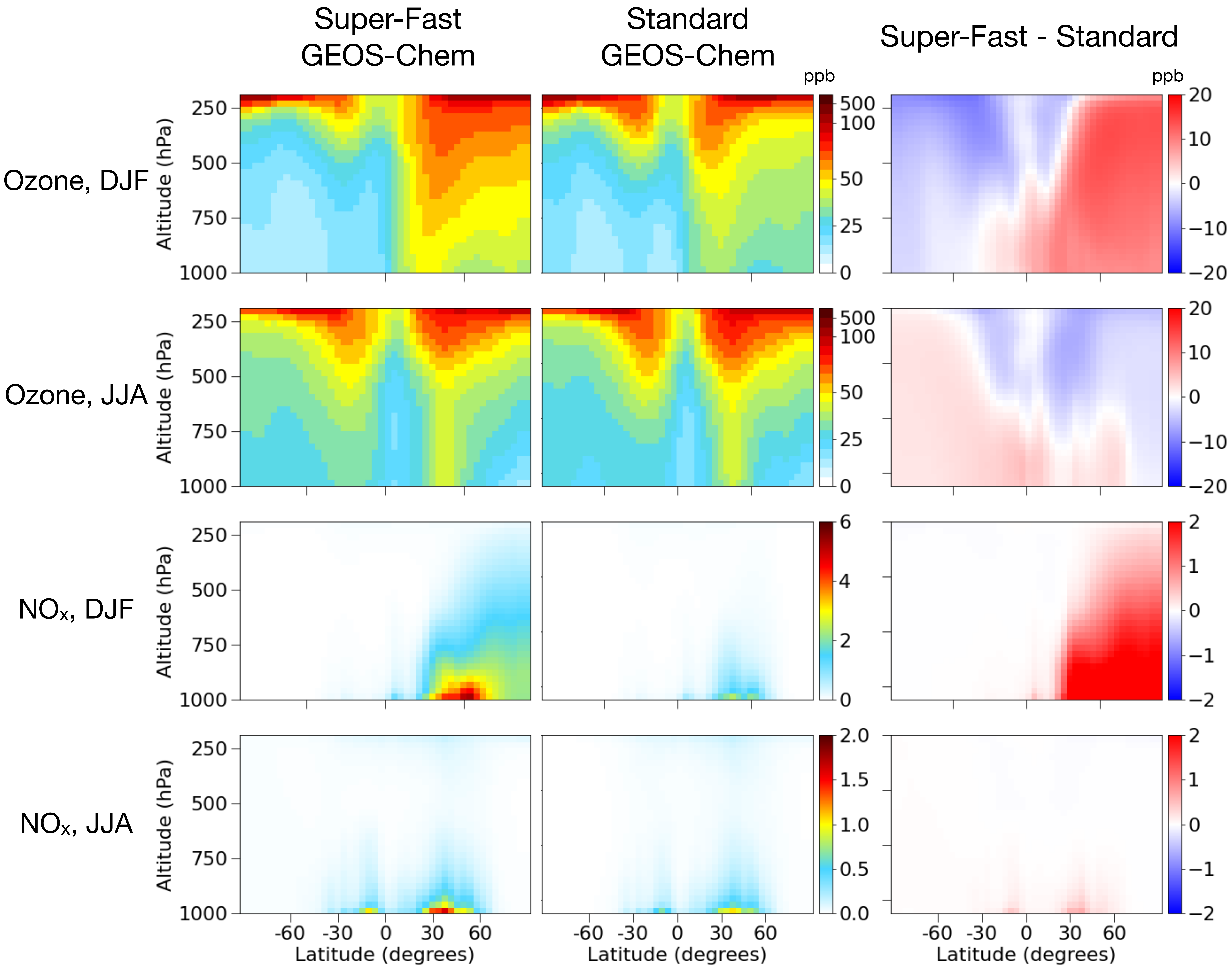


1. Nonparametric, **universal** function approximators
2. Learn to predict based on large dataset of **repeated** patterns
3. Proven to **speed up** solving ODEs at orders of magnitude (Malek and Shekari, 2006)

Application of machine learning to chemical solvers

Why	Why not
Chemical calculation is expensive	High dimensionality
Highly repetitive	Lack physics constraints
Fully deterministic	Error growth

The ‘Super Fast’ chemical mechanism will allow us to better define ML methods and understand limitations in a full 3-D global modeling framework



- Global mechanism with 12 species [Brown-Steiner et al., 2018]
- Benchmarked in GEOS-Chem v12.0.0
- 4x5° resolution

1-hour chemical time step output

20 variables:

2 physical var: T, air density

6 photolysis frequencies

12 gas-phase species

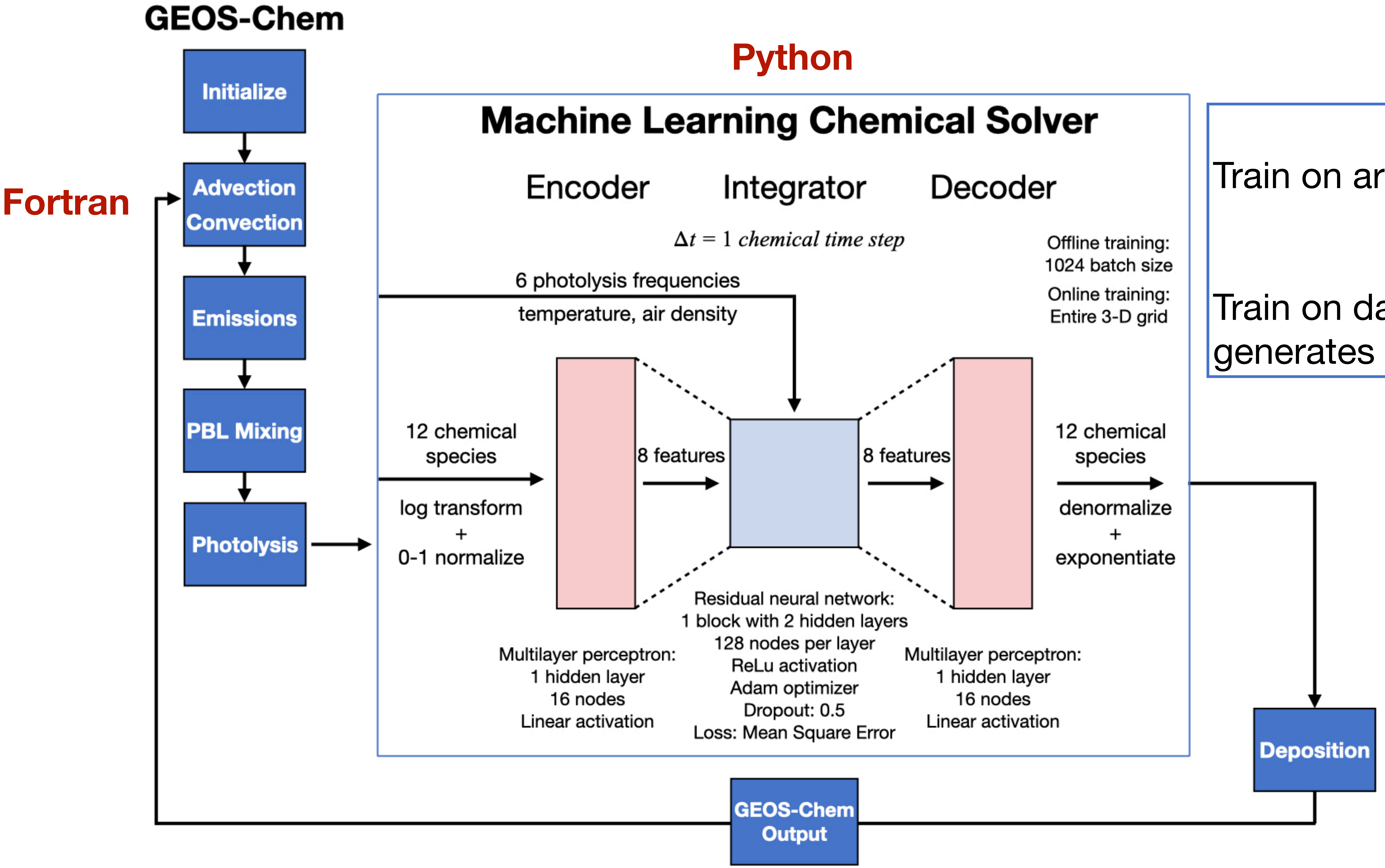
1 month dataset would contain:

lon×lat×lev×days×hours = 46×72×~25×31×24 ->~62 million samples

Training: 2016, Test: 2017

Kelp et al., 2022

We train ML solvers online synchronously with a GEOS-Chem simulation to train on representative conditions in sequence



GEOS-Chem

Fortran

Initialize

Advection Convection

Emissions

PBL Mixing

Photolysis

Python

Machine Learning Chemical Solver

Encoder

Integrator

Decoder

$\Delta t = 1$ *chemical time step*

6 photolysis frequencies

temperature, air density

12 chemical species

log transform + 0-1 normalize

8 features

8 features

12 chemical species

denormalize + exponentiate

Offline training: 1024 batch size

Online training: Entire 3-D grid

Multilayer perceptron:
1 hidden layer
16 nodes
Linear activation

Residual neural network:
1 block with 2 hidden layers
128 nodes per layer
ReLu activation
Adam optimizer
Dropout: 0.5
Loss: Mean Square Error

Multilayer perceptron:
1 hidden layer
16 nodes
Linear activation

Deposition

GEOS-Chem Output

Offline:

Train on archived dataset

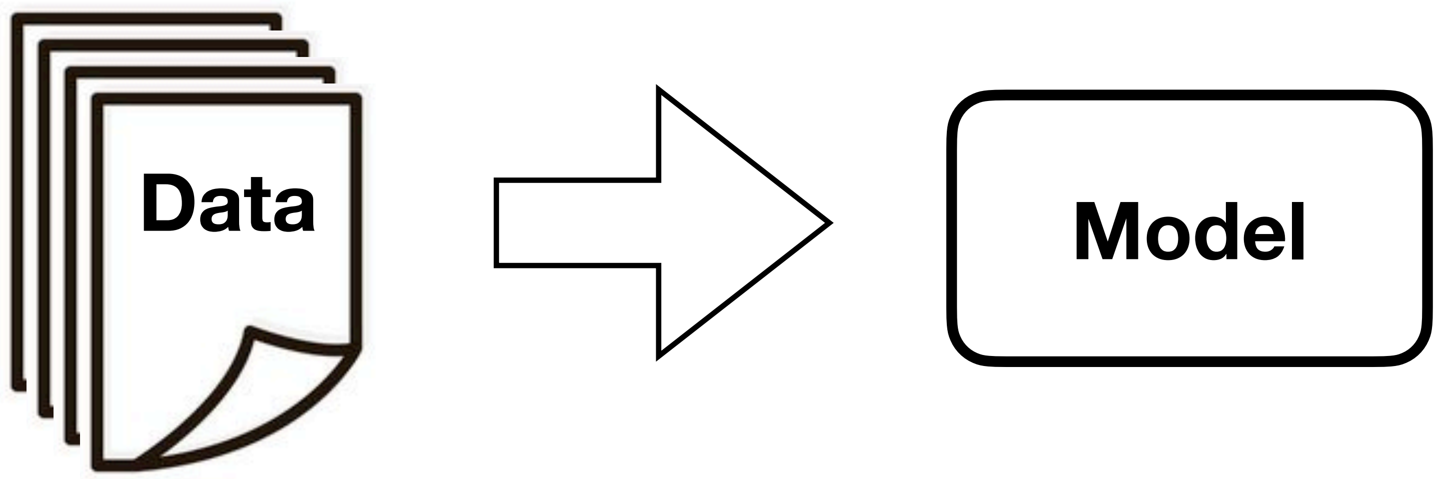
Online:

Train on data as GEOS-Chem generates it

Kelp et al., 2022

Online learning prevents overfitting to training data

Offline (batch) learning



Pros:

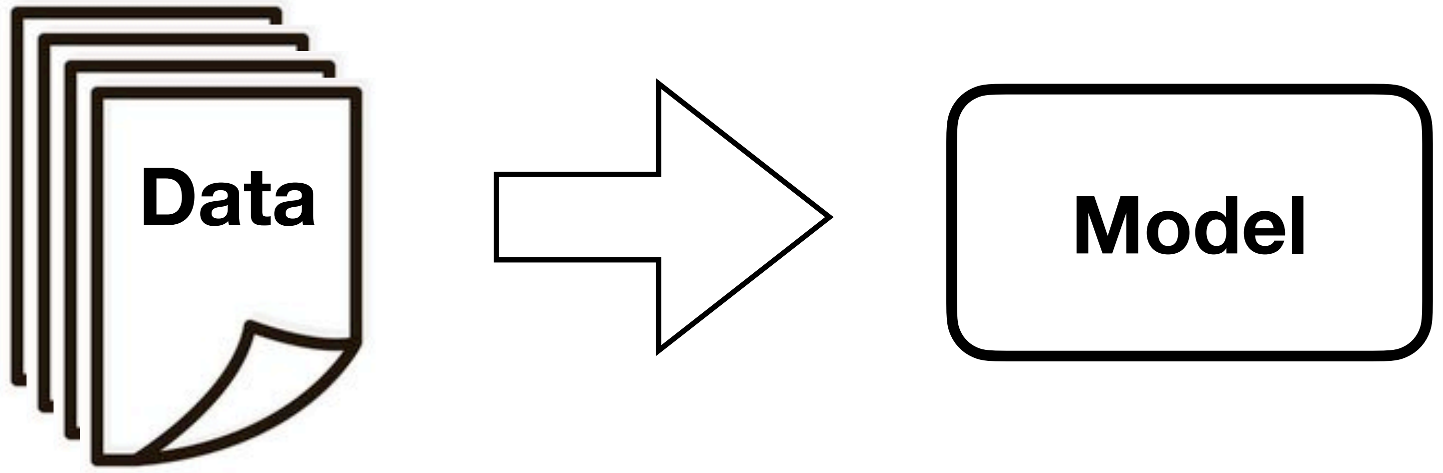
Simple to code
Fast, easy to train + manipulate data (recursive train)

Cons:

Overfitting (overly reliant on training data)
Generate massive data archives
Under/oversample chemical environments

Online learning prevents overfitting to training data

Offline (batch) learning



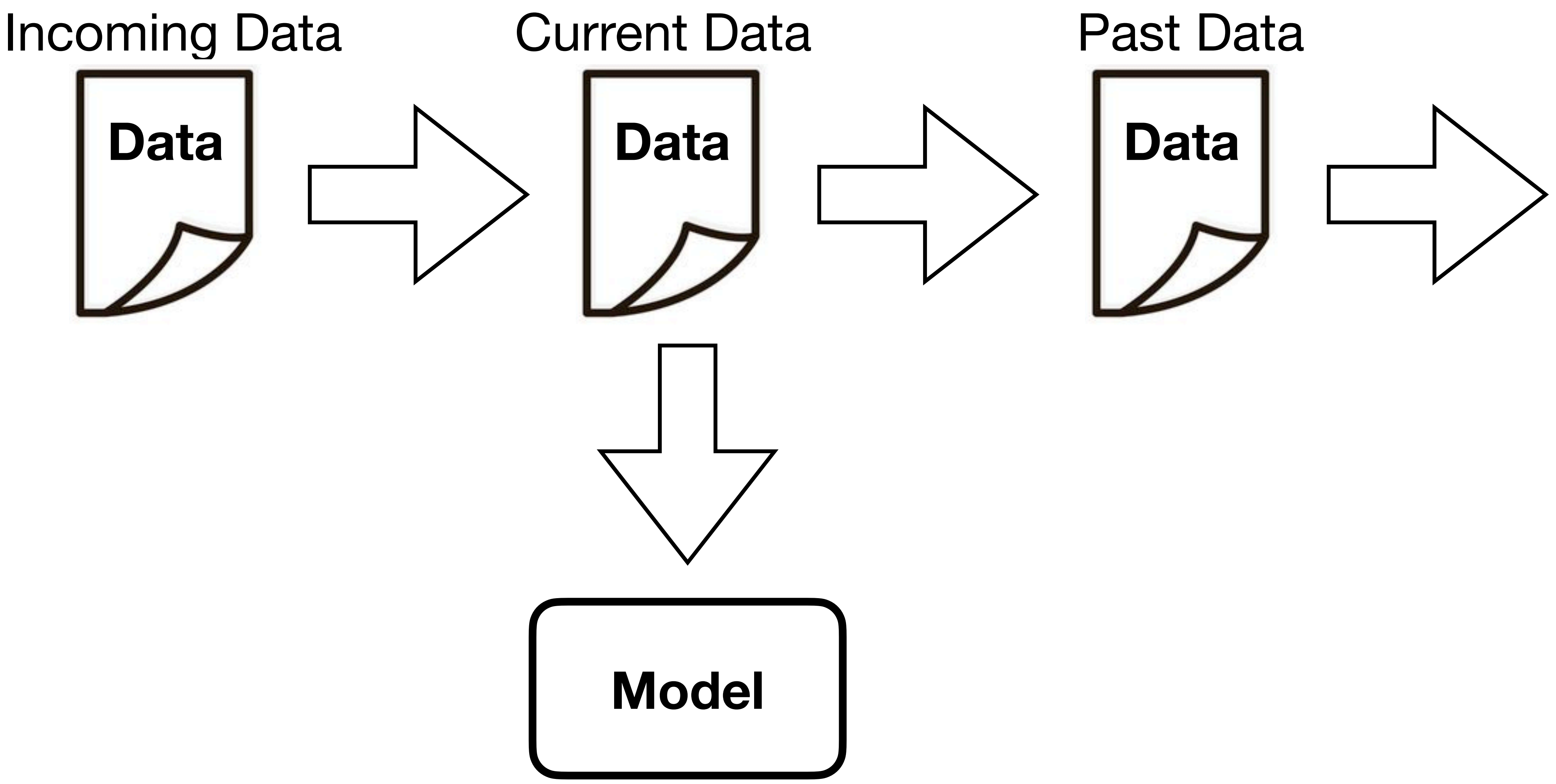
Pros:

Simple to code
Fast, easy to train + manipulate data (recursive train)

Cons:

Overfitting (overly reliant on training data)
Generate massive data archives
Under/oversample chemical environments

Online (sequential) learning



Pros:

Cannot overfit: each data point used once
Representative realizations
No need to generate data archive

Cons:

Hard to implement
Very expensive training! (Each GC time step)
Limited observational window
"Catastrophic forgetting problem" [McCloskey and Cohen, 1989]

Online training improves accuracy and stability over offline training

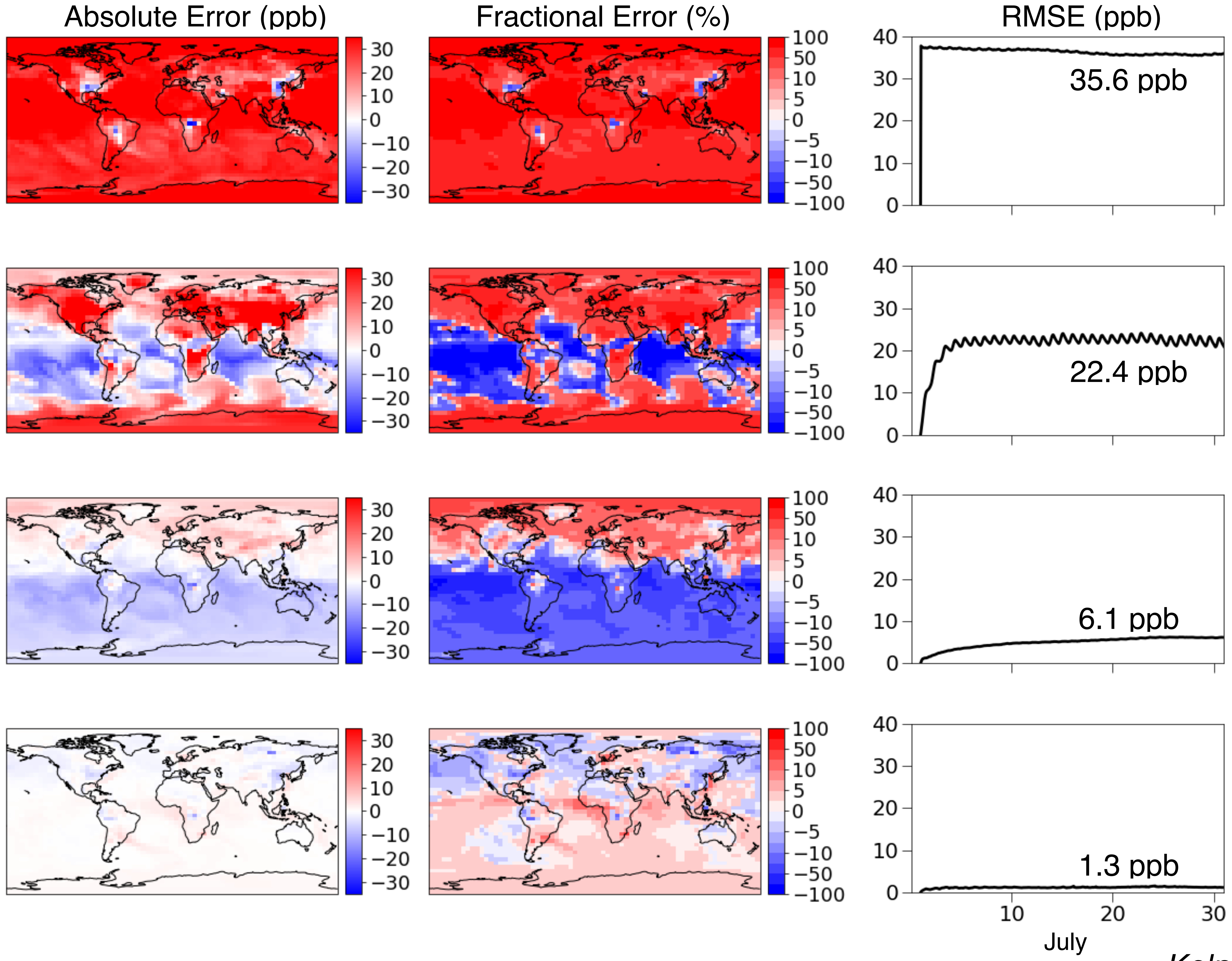
Ozone

Offline
t to t+1

Offline
24h recursive

Offline retrained
to online

Online



Train:
JJA 2016

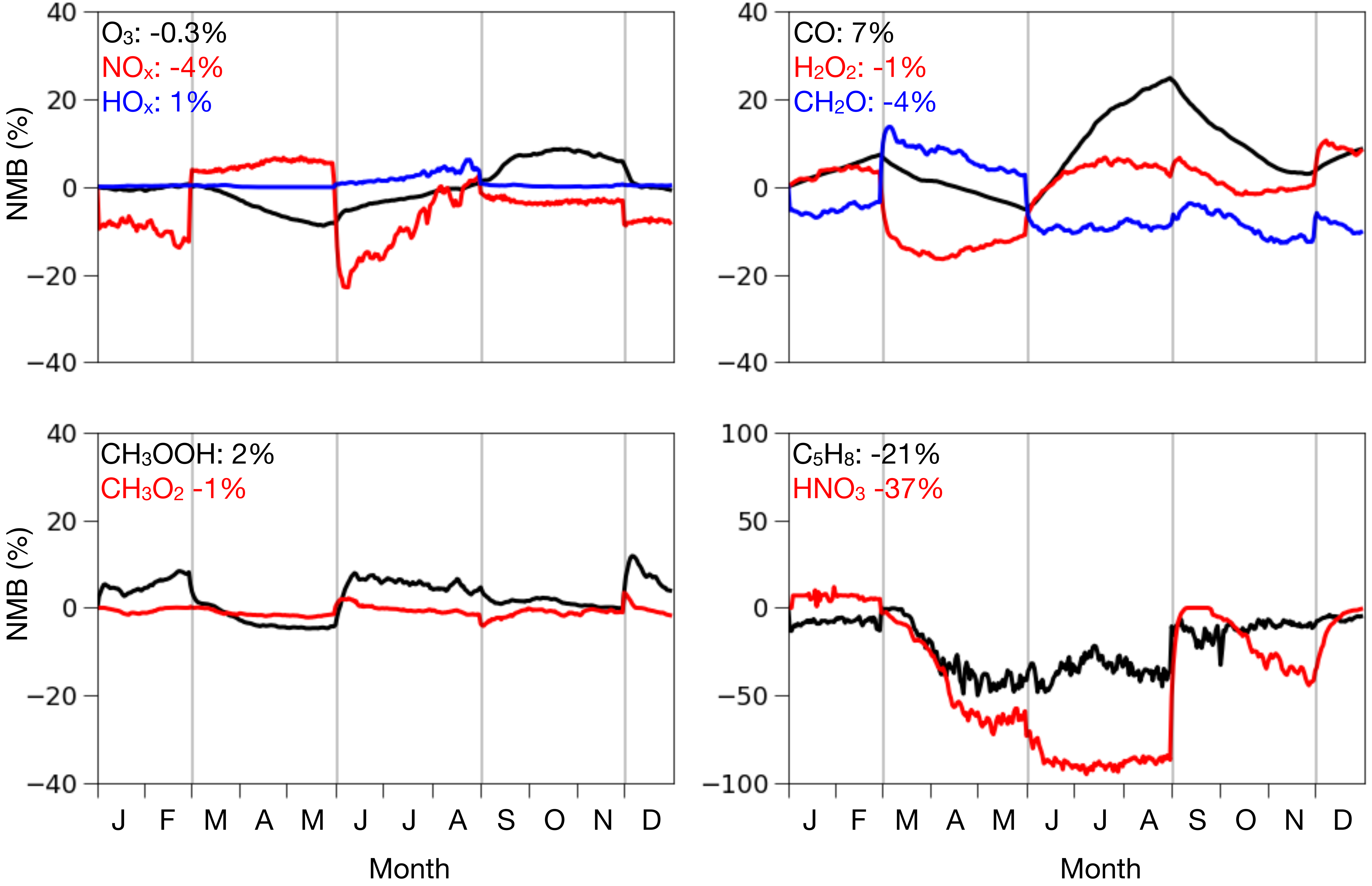
Test:
July 2017

Kelp et al., 2022

Simulation are stable and average errors for the year are <10%, but ML solvers have different seasonal fits of accuracy

Separate ML solvers for:

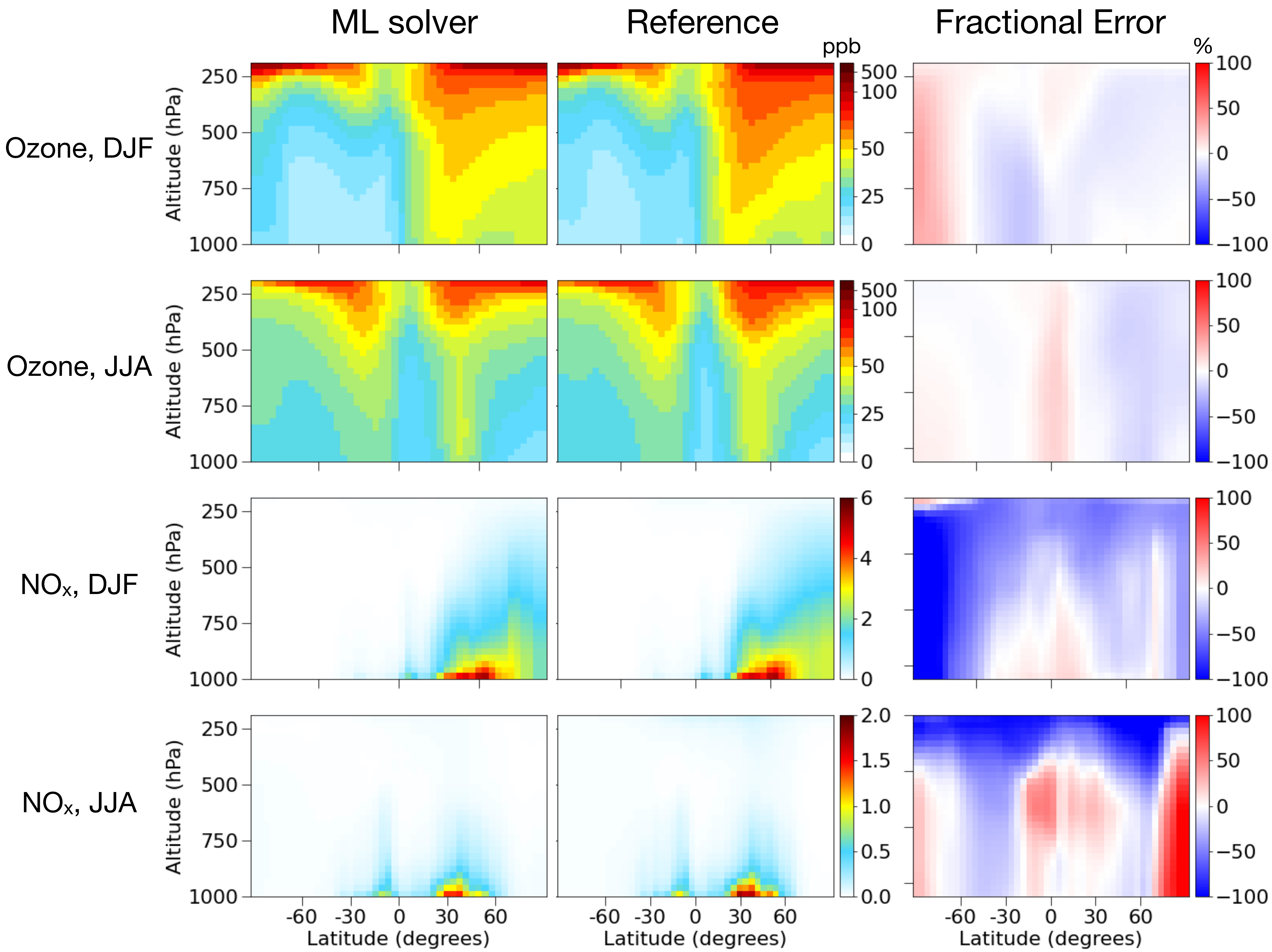
- Species
- Season



Error does not accumulate over the course of a year

Kelp et al., 2022

Errors are largest at remote latitudes and high altitudes due to chemical error accumulation as air ages



Kelp et al., 2022

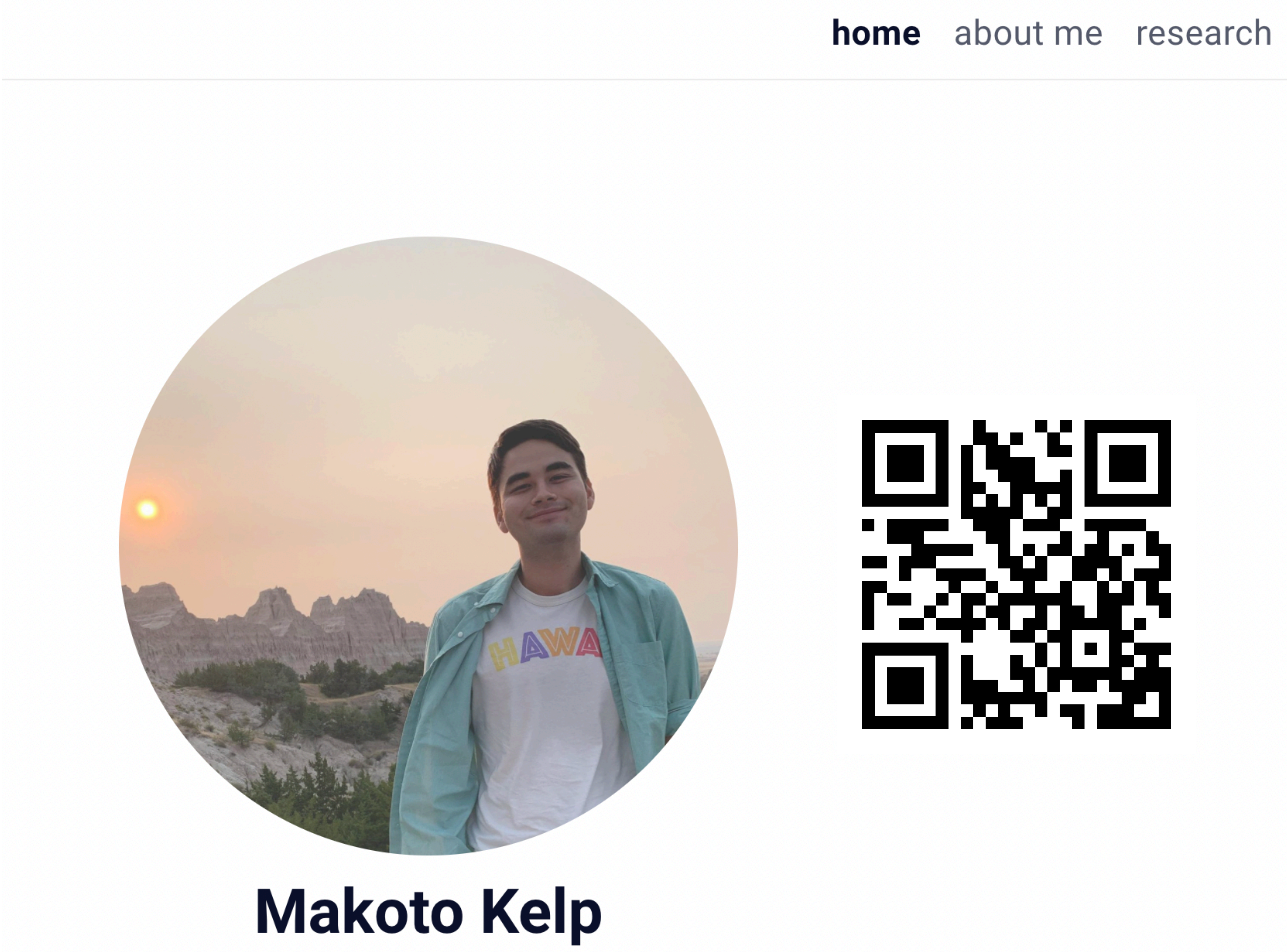
Takeaways

- Application of ML chemical solver in global 3-D atmospheric chemistry models may require online training.

- **Stable** year-long global simulation of chemistry can be achieved with a ML solver applied to the Super-Fast mechanism in GEOS-Chem.

- Computational speedup is five-fold relative to the reference Rosenbrock solver in GEOS-Chem.

- Large regional biases for ozone and NO_x under remote conditions where chemical aging leads to error accumulation.
  - Regional biases remain a major limitation for practical application, and ML emulation would be more difficult in a more complex mechanism.

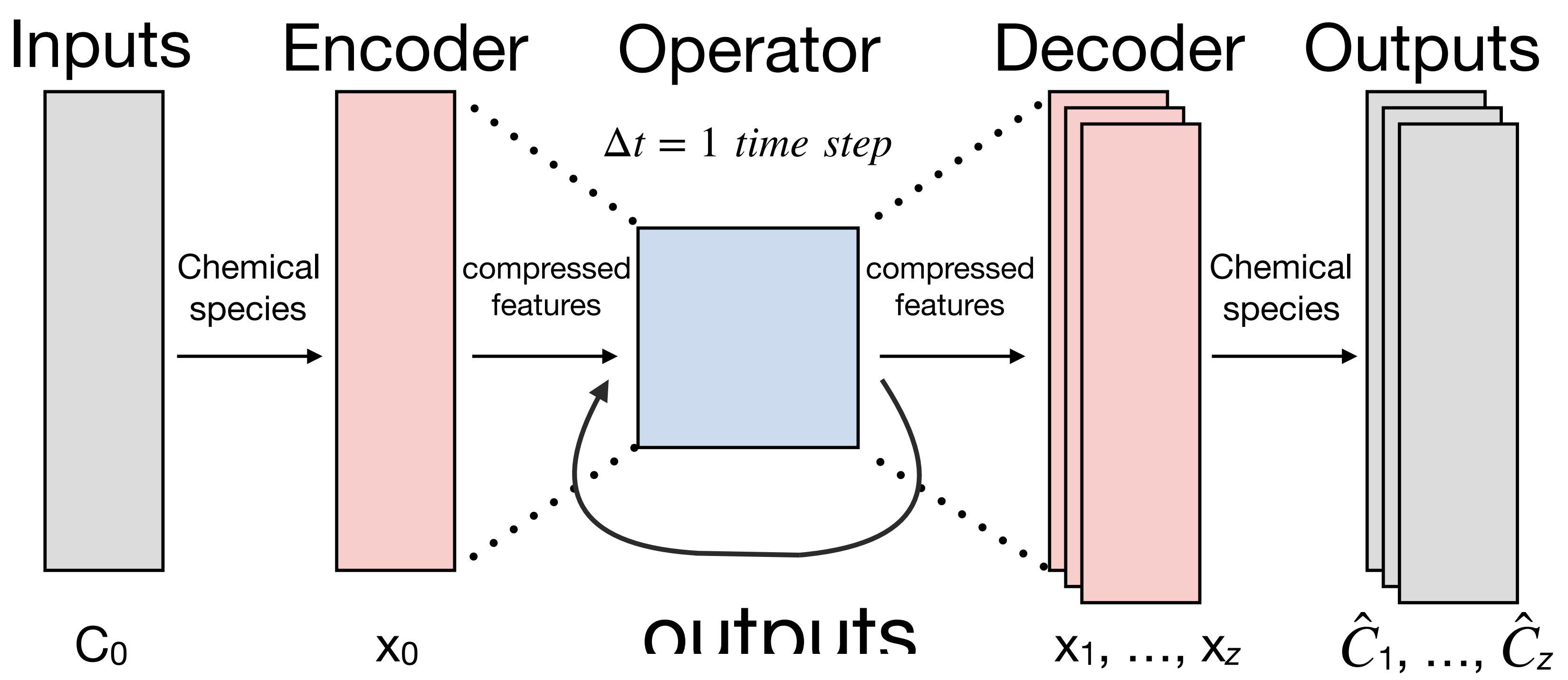


Makoto Kelp

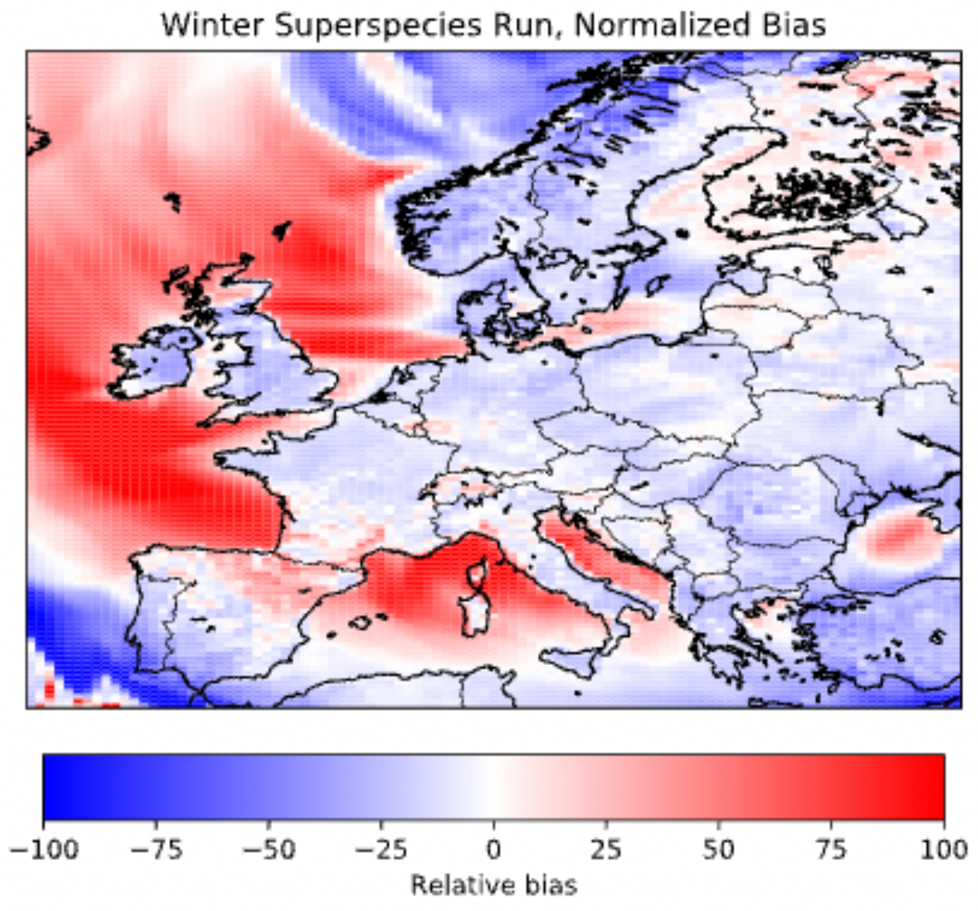
What does this mean for ML and CTMs?

Promising

Compression

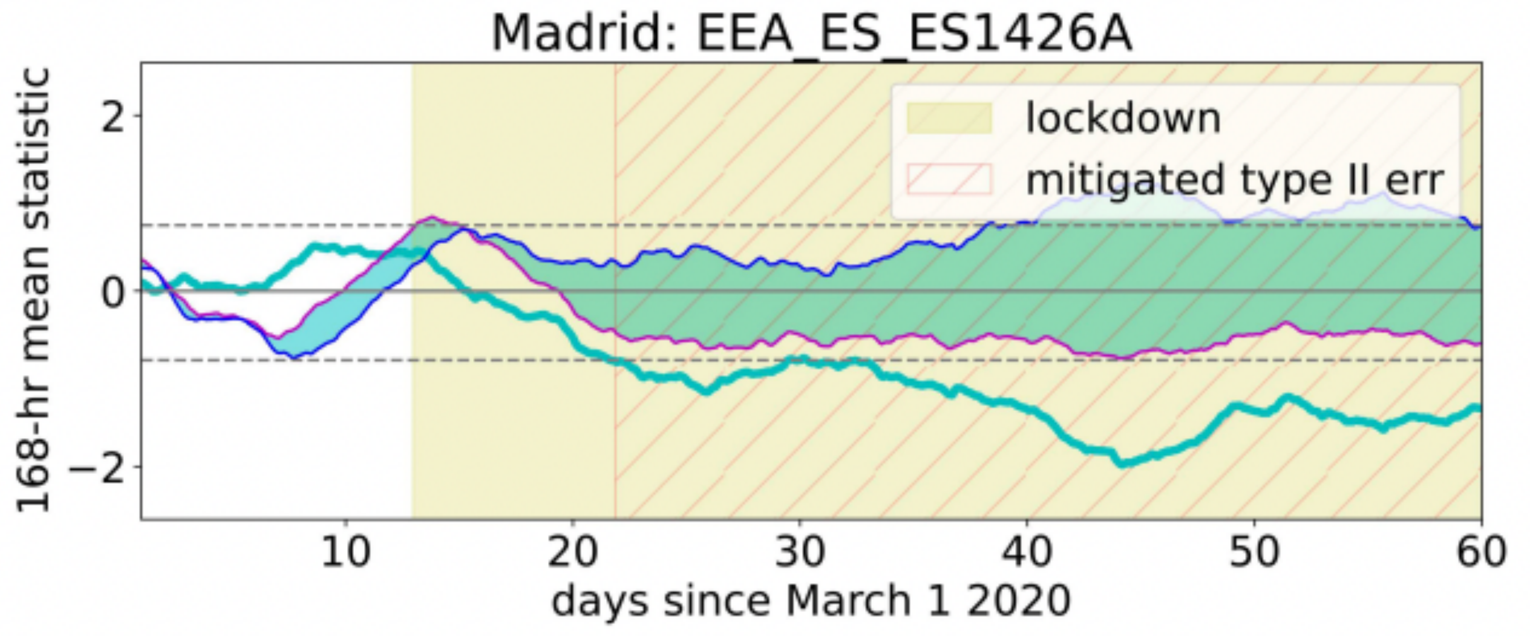


Mechanism compression
Kelp et al. 2020

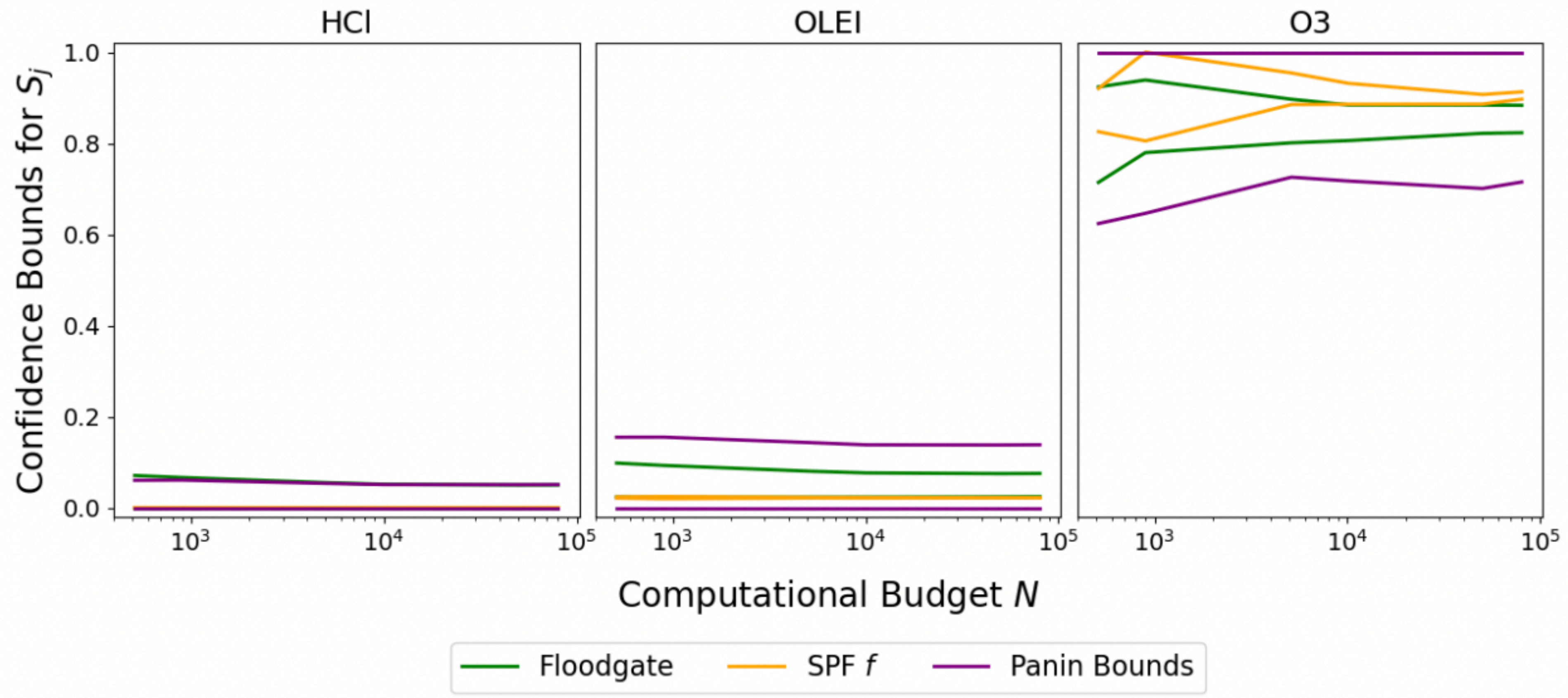


Advecting superspecies
Sturm et al. 2022 (in-review)

Anomaly detection

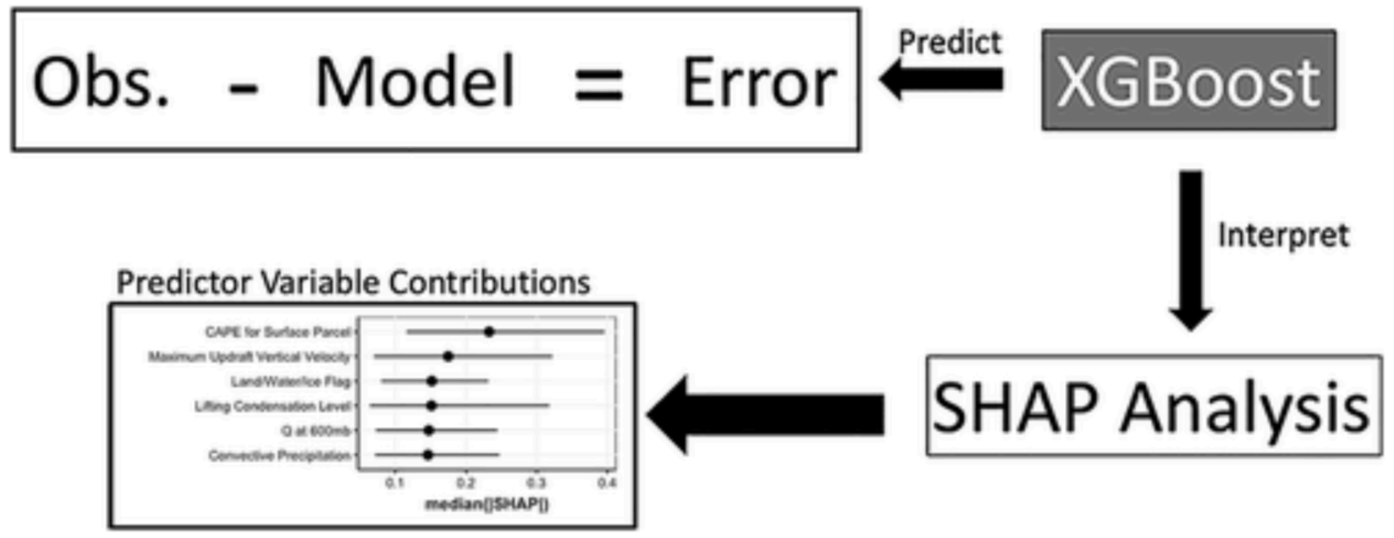


Mallen et al. 2022 (arXiv)



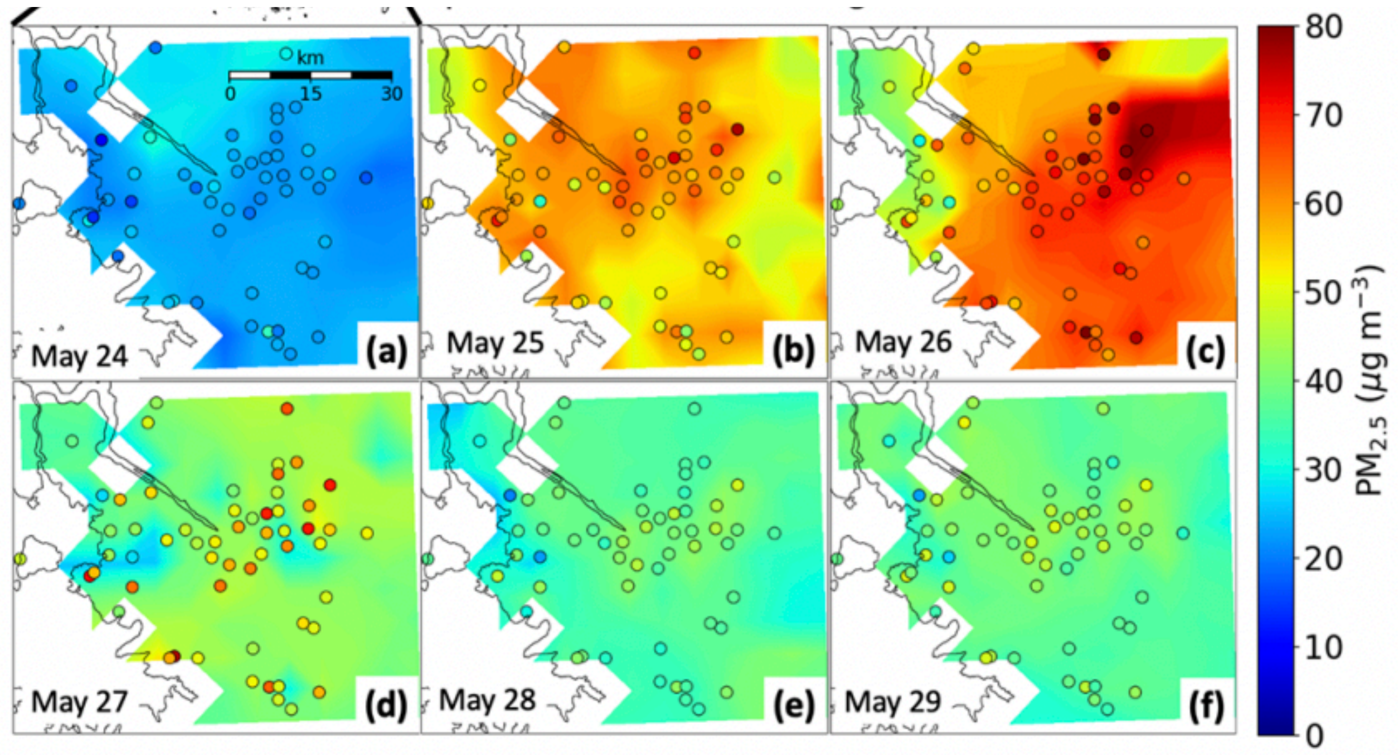
Aufiero and Janson, 2022 (arXiv)

Parametrization



Silva et al. 2022

Dataset curation

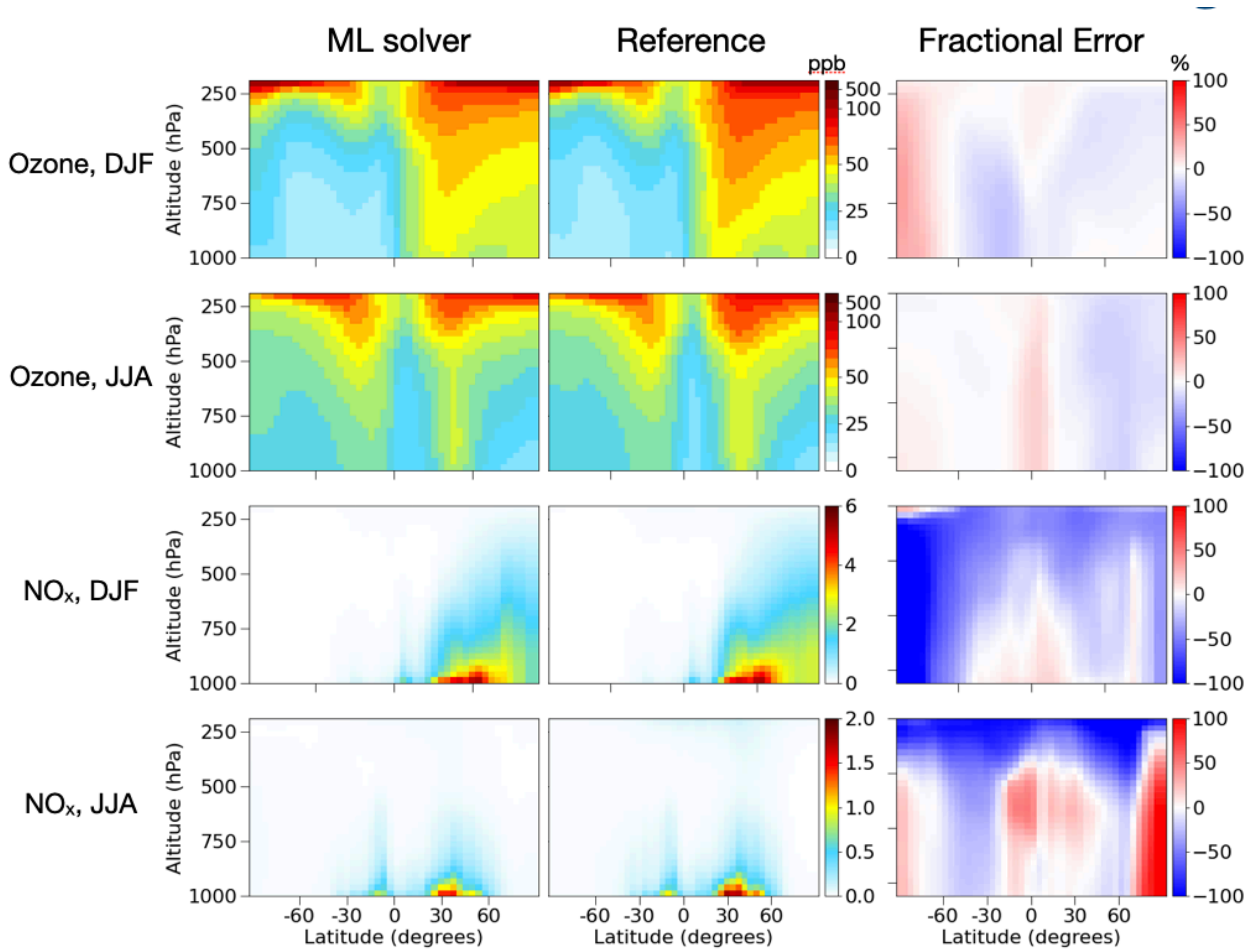


Pendergrass et al., 2022

What does this mean for ML and CTMs?

Not so promising?

Chemical Solvers



Kelp et al., (2022)

-Physical constraints don't help?

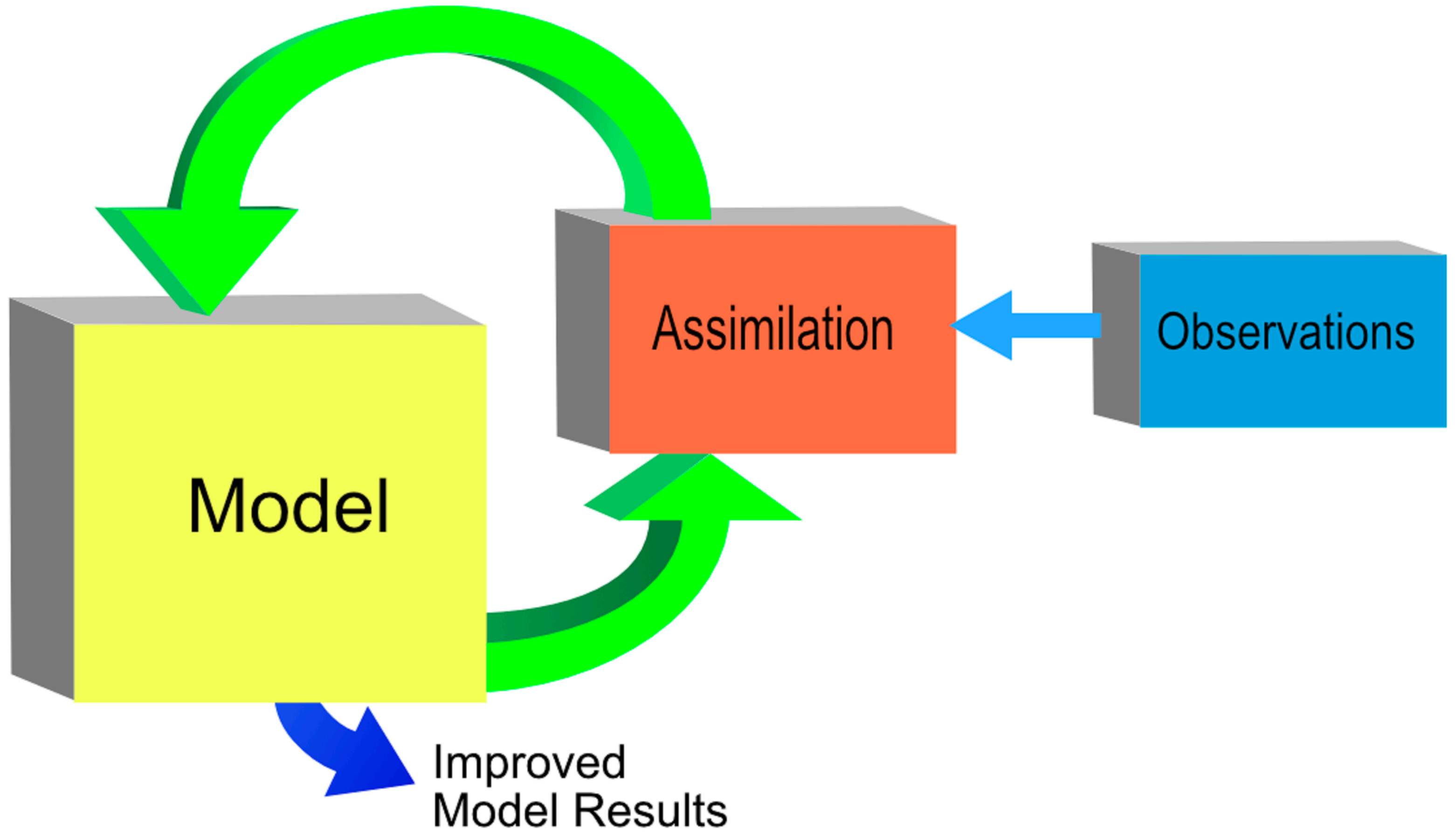
- Ground truth is a ODE solver not bulk statistics

-Demand for specificity/accuracy incredibly high:

 -'Eye test' won't pass

-Too early to see if 'drop-in' CTM ML components useful

Data Assimilation



-DA with ML redundant

-DA and adjoint modeling is basically gradient descent (already been around for a long time with efficient architectures)

-DA also better at ingesting sparse and noisy data as opposed to a lot of ML DA applications