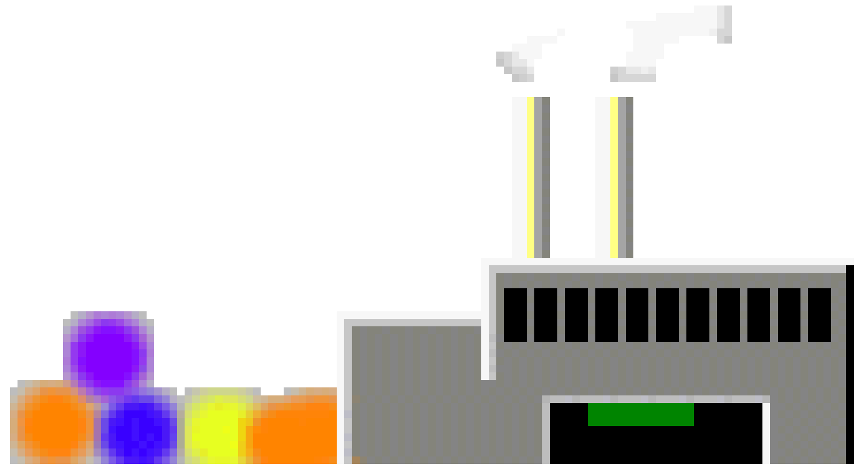
An Online-Learned Neural Network Chemical Solver for Stable Long-Term Global Simulations of Atmospheric Chemistry

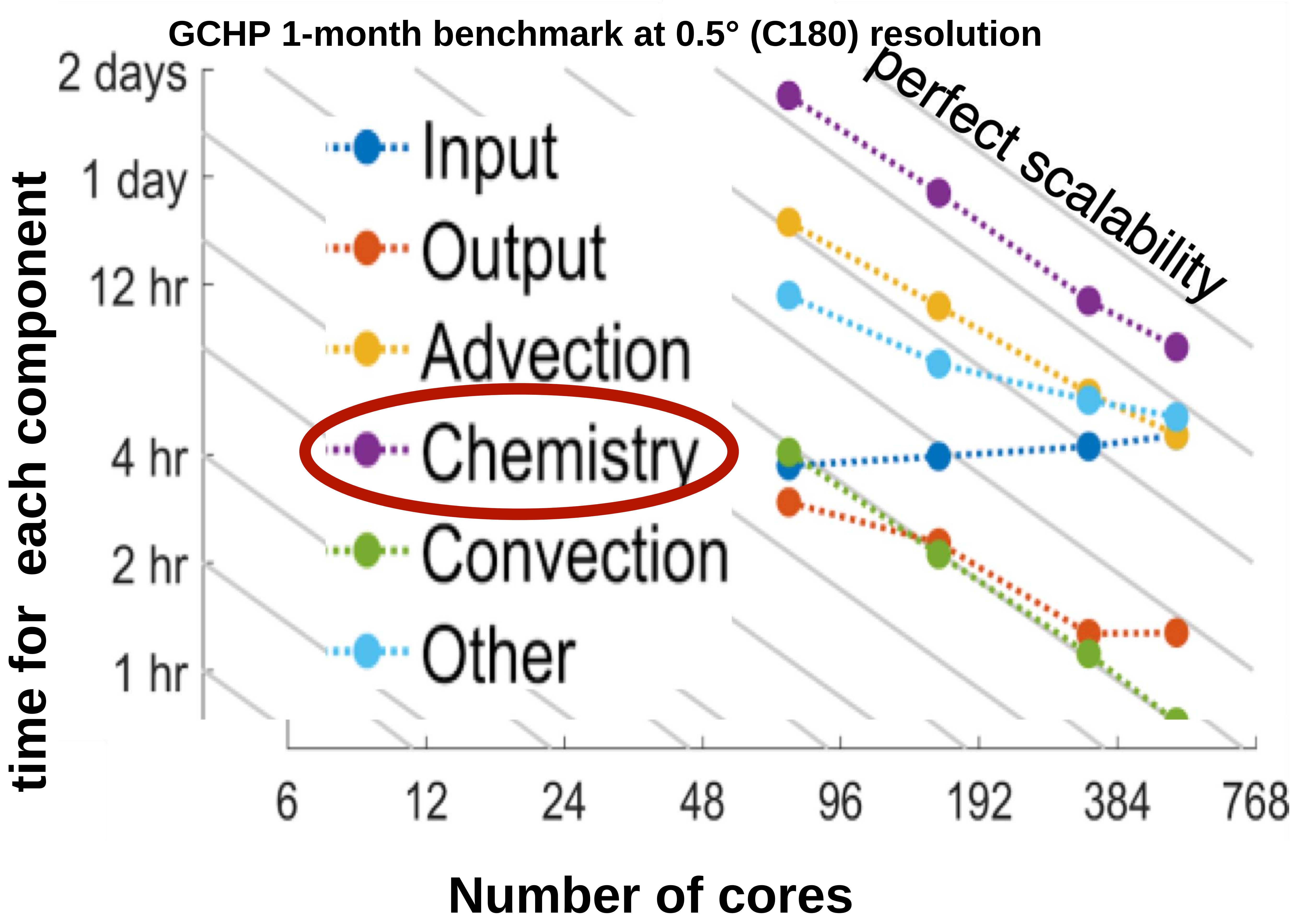
Makoto Kelp

with Daniel Jacob, Haipeng Lin, Melissa Sulprizio

ECMWF Machine Learning Workshop 20220329



Global modeling of atmospheric chemistry is a grand computational challenge



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

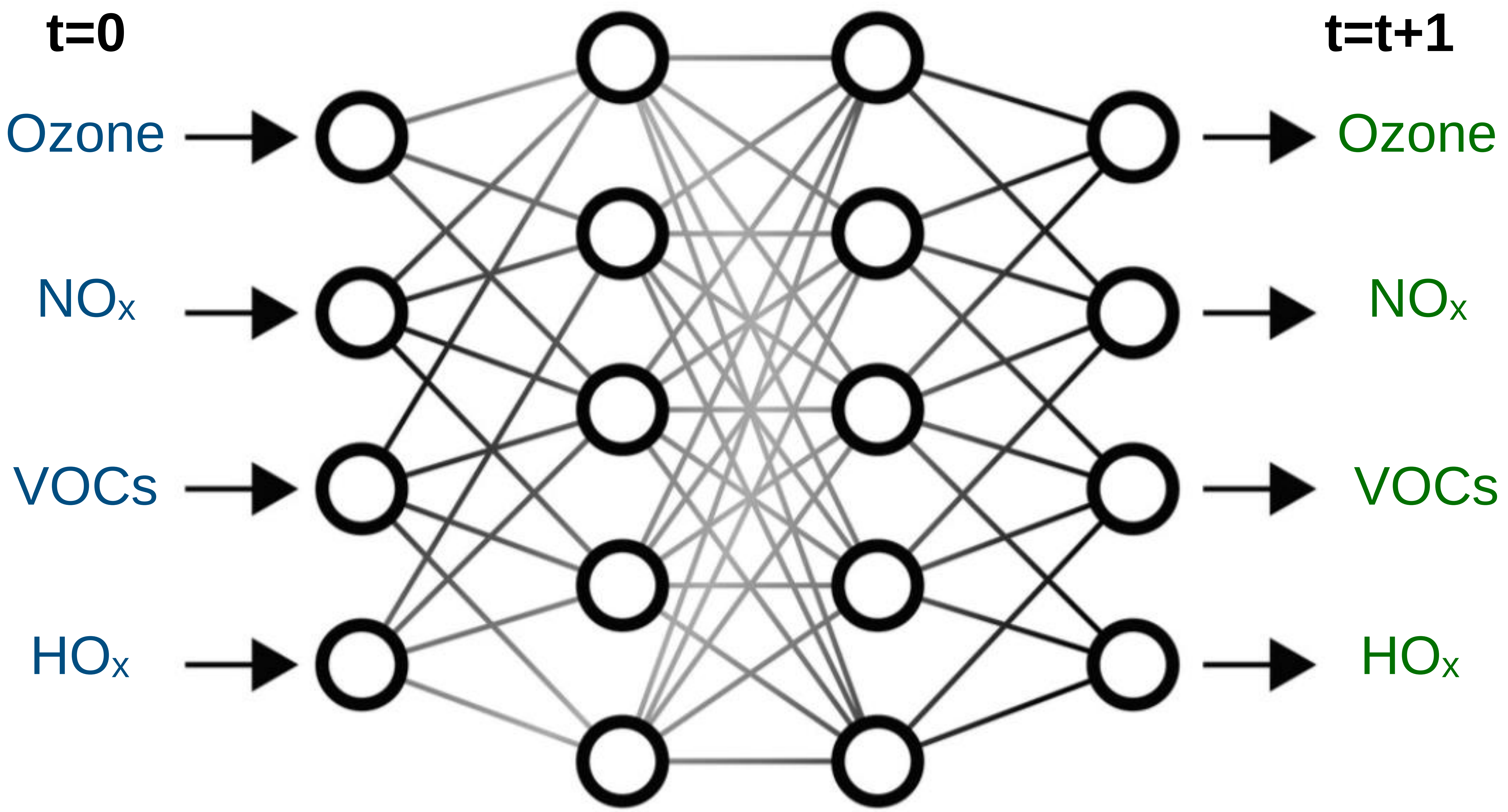
-Weather and climate models typically have **~4 variables**

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

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

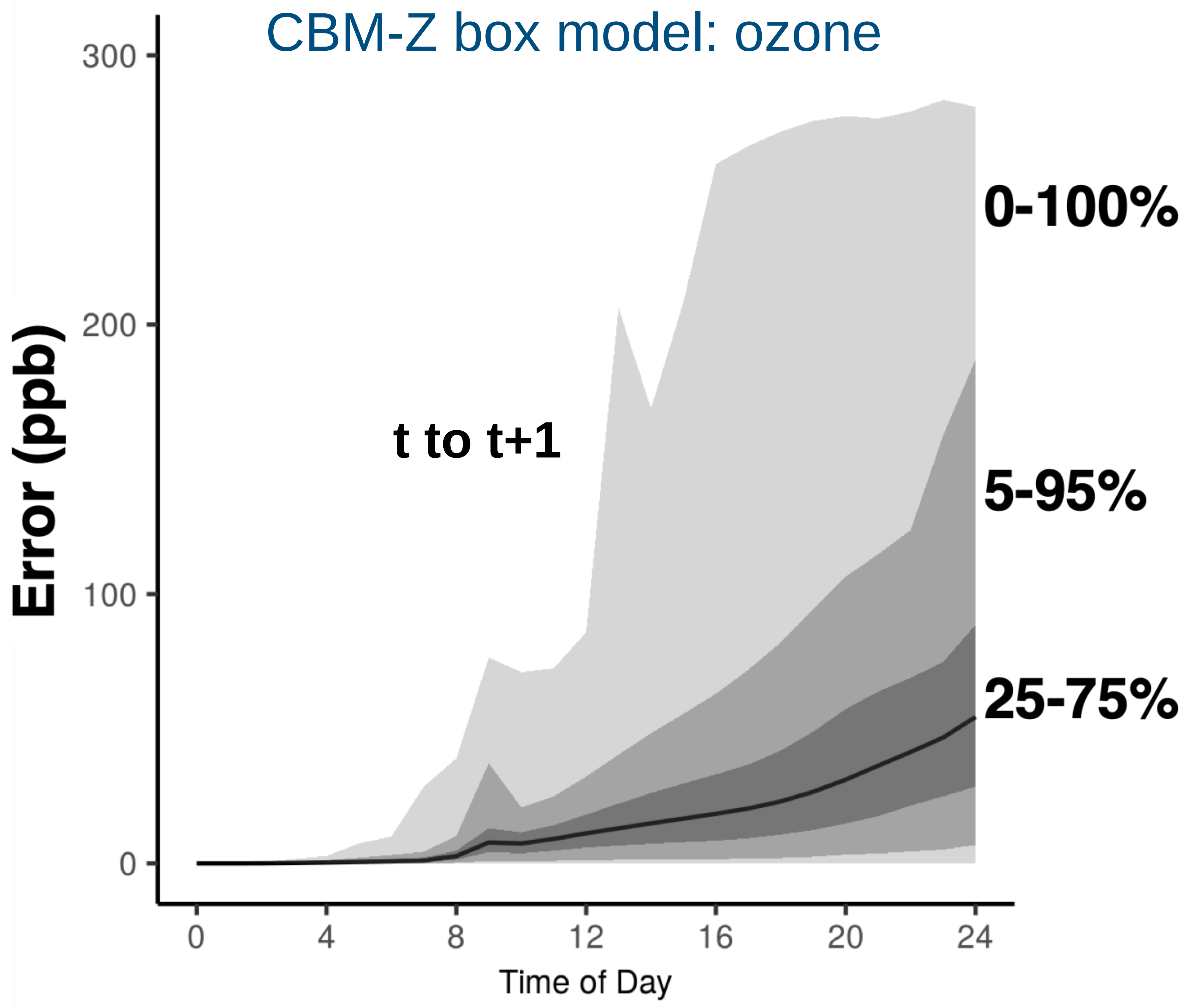
Eastham et al., 2018 GMD

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

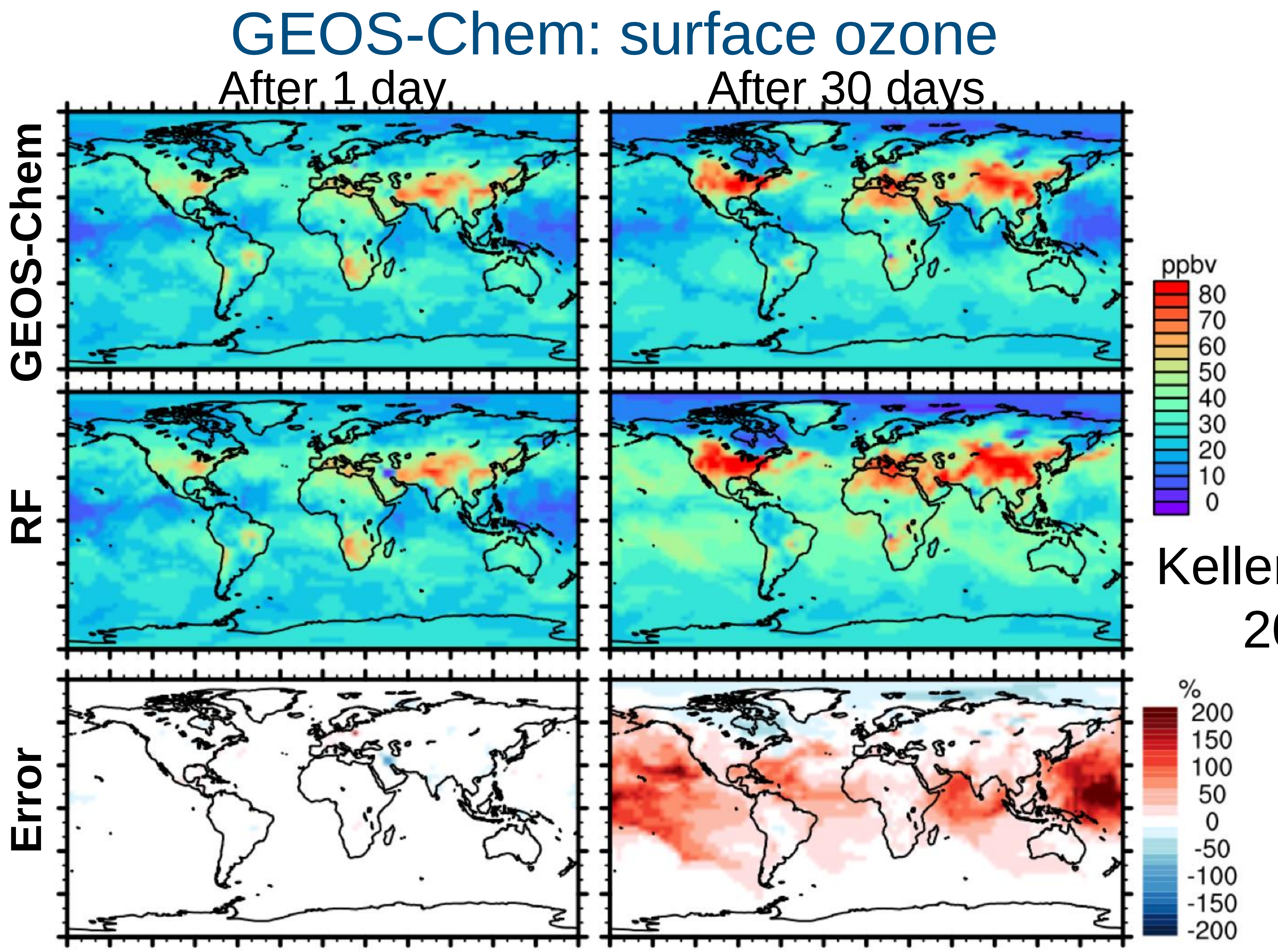


1. Nonparametric, **universal** function approximators
2. Learn to predict based on large dataset of **repeated** patterns
3. Proven to **speed up** solving ODEs at orders of magnitude (Malek and Shekari, 2006)

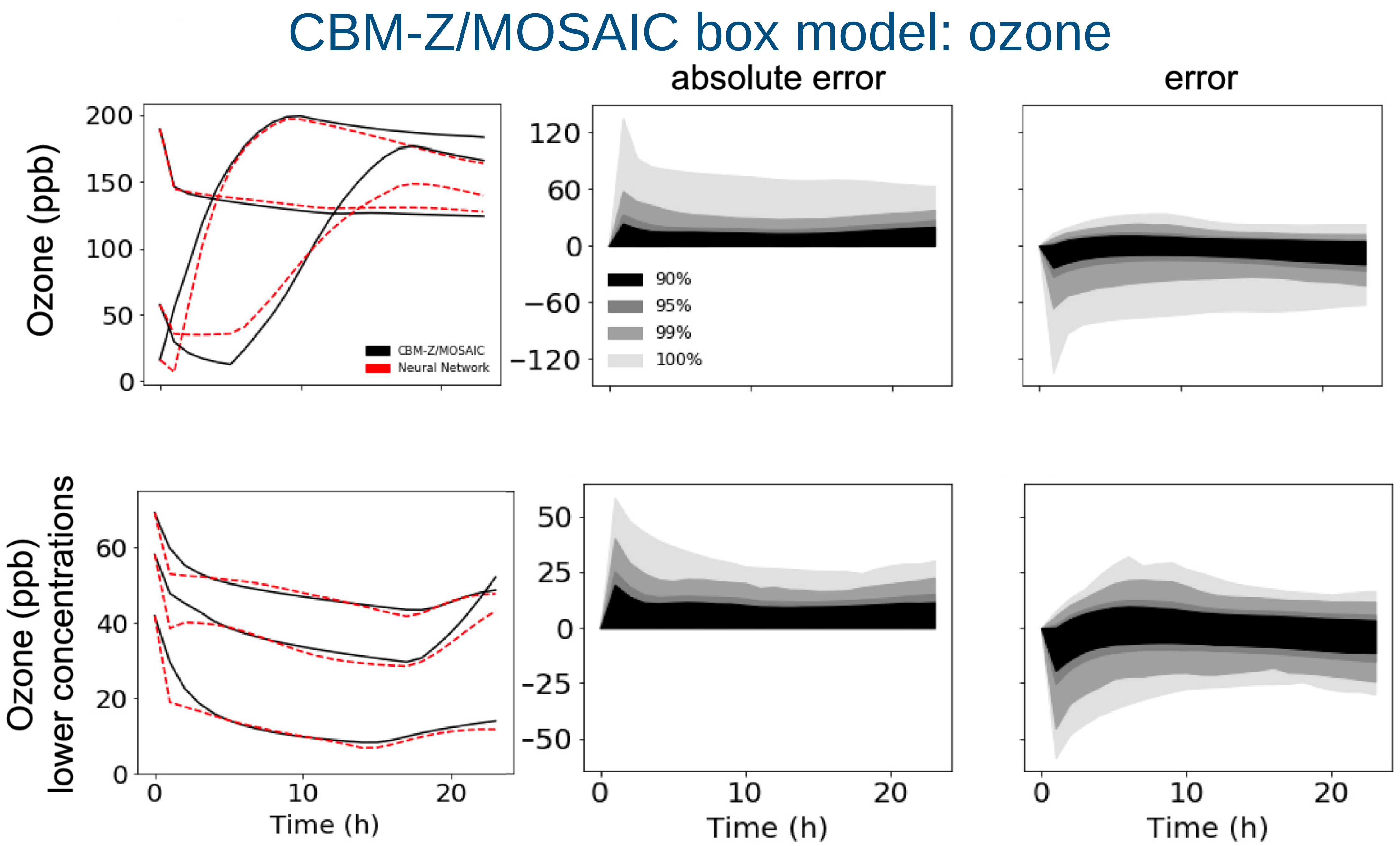
Past ML chemical solver attempts have encountered runaway error growth and have been limited to box model approaches



Kelp et al. 2018 ArXiv

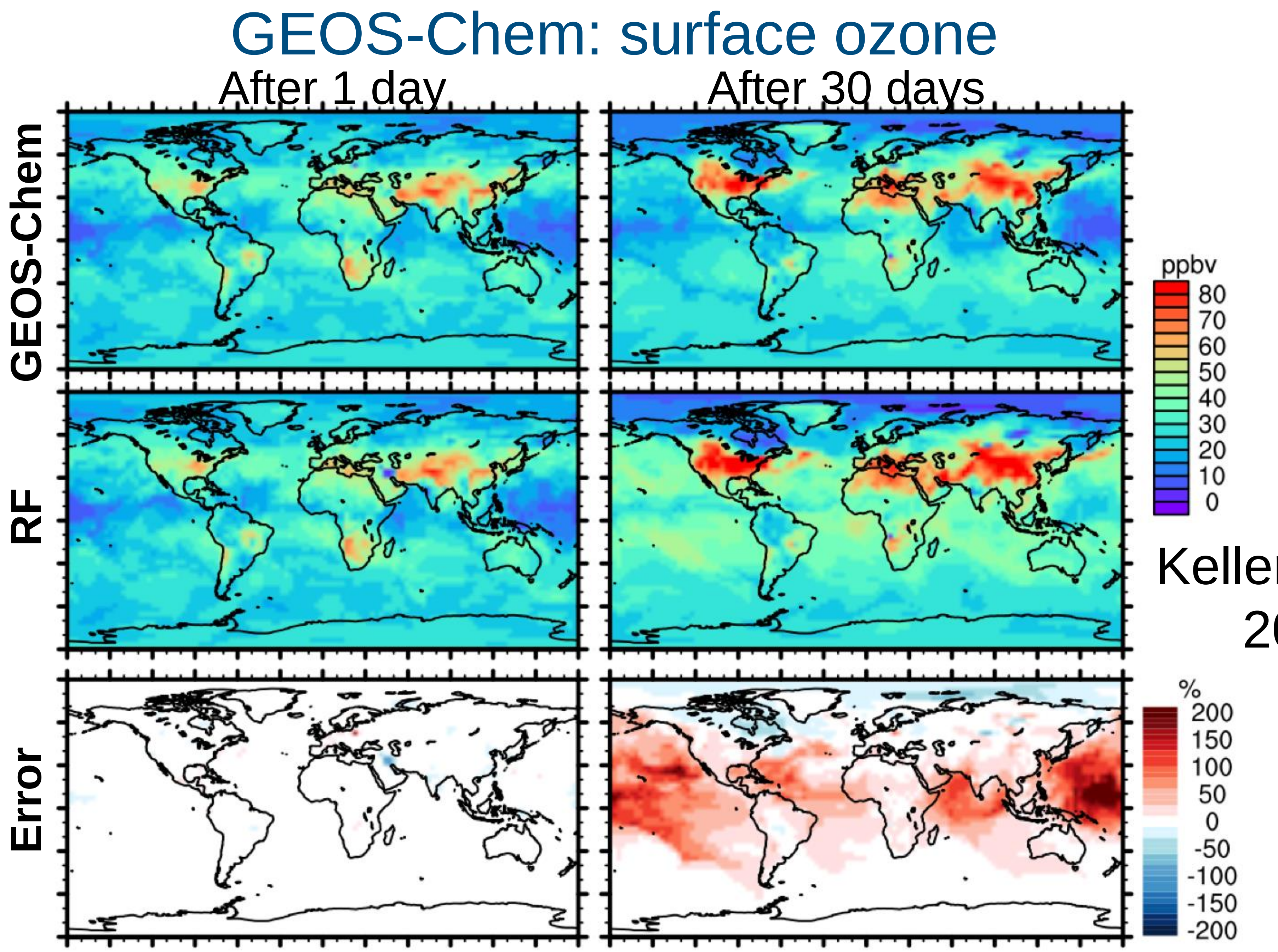


Keller and Evans, 2019 GMD

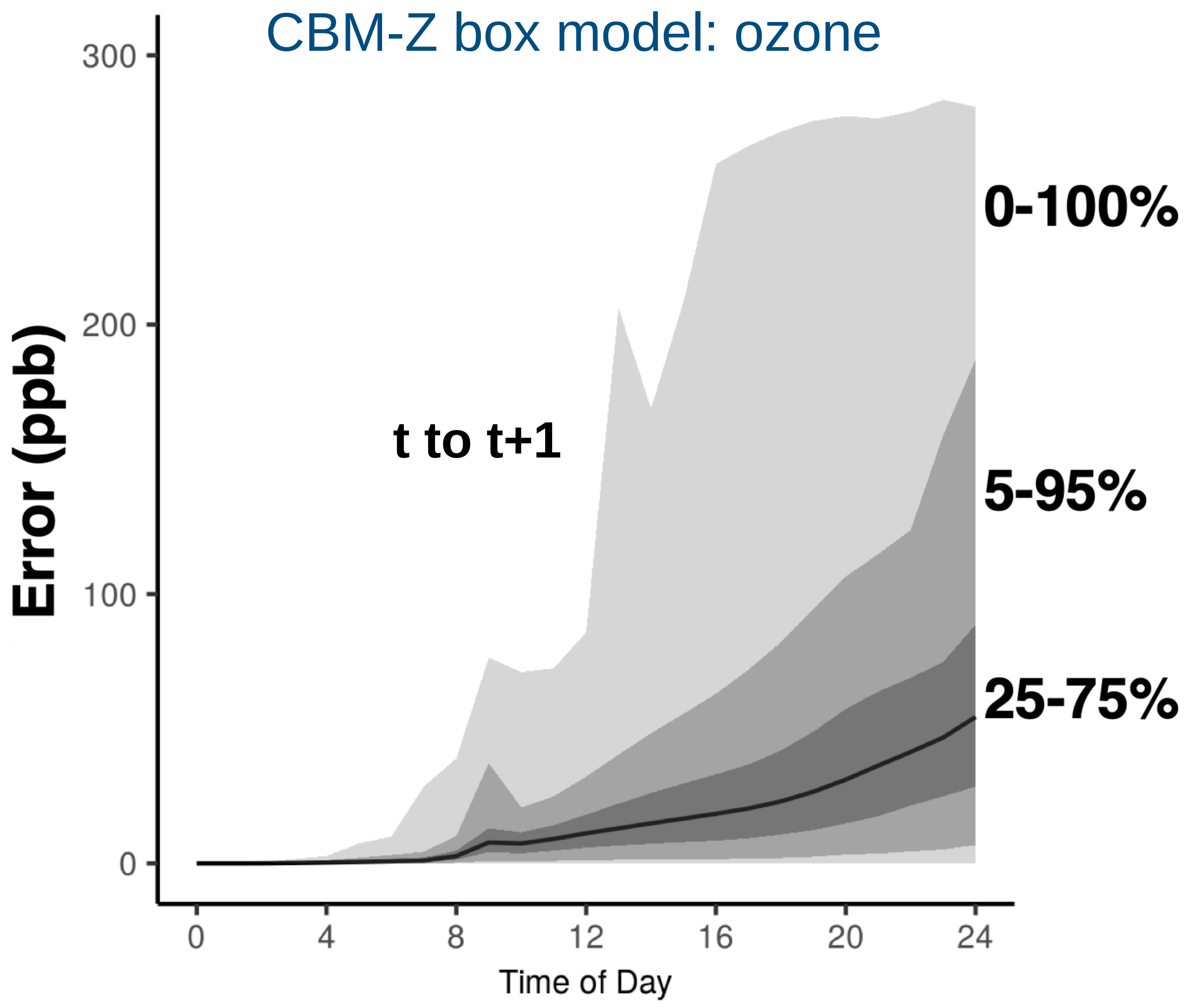


Kelp et al. 2020 JGR

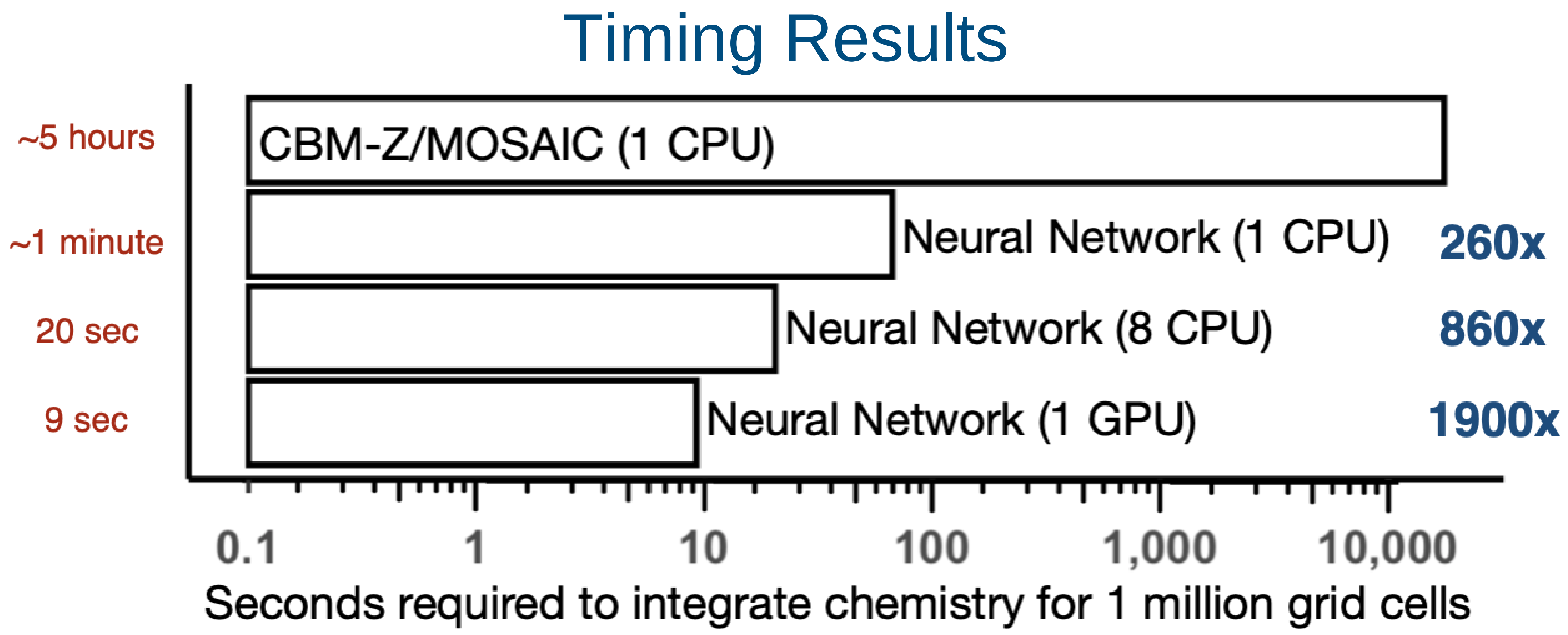
Past ML chemical solver attempts have encountered runaway error growth and have been limited to box model approaches



Keller and Evans, 2019 GMD

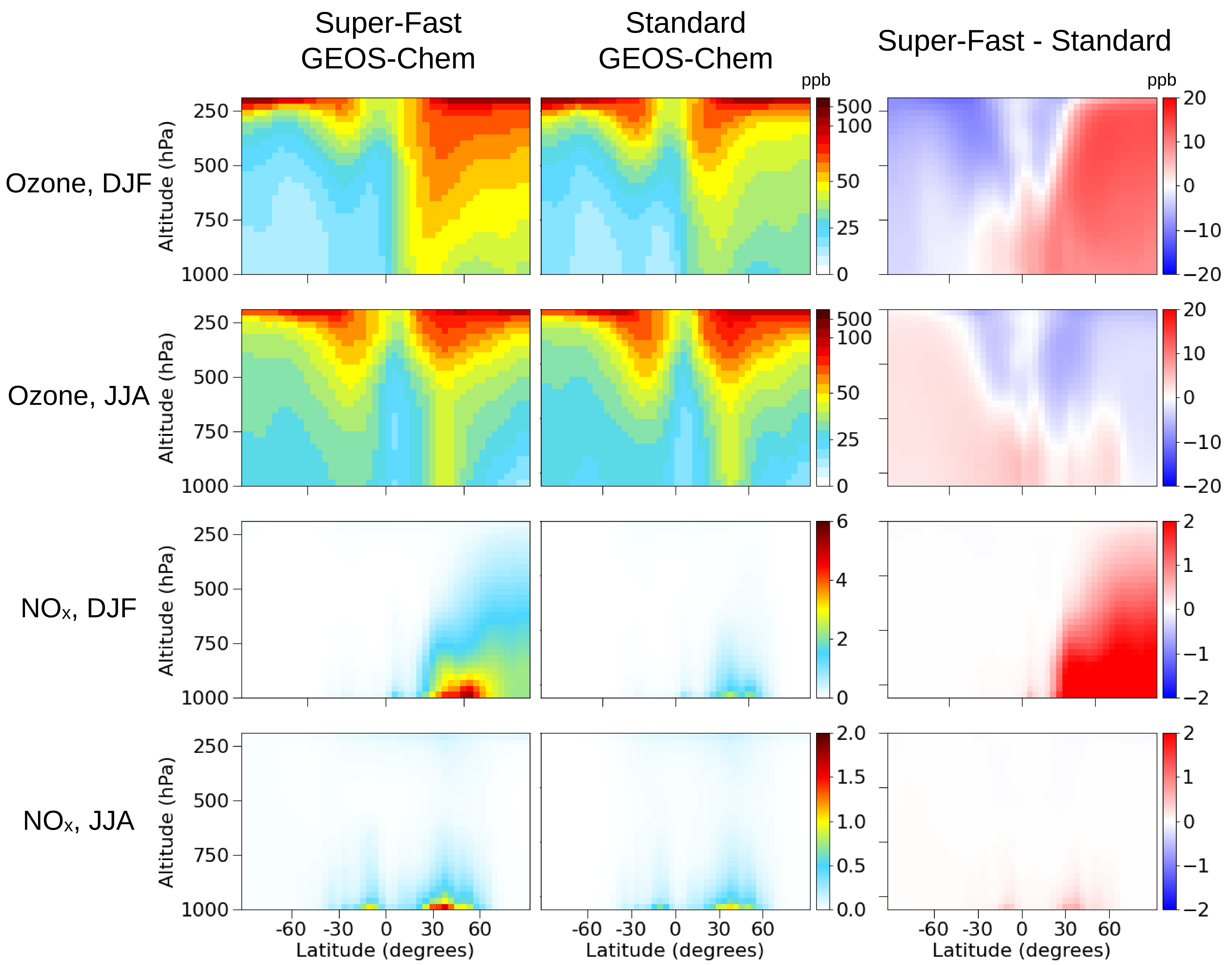


Kelp et al. 2018 ArXiv



Kelp et al. 2020 JGR

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

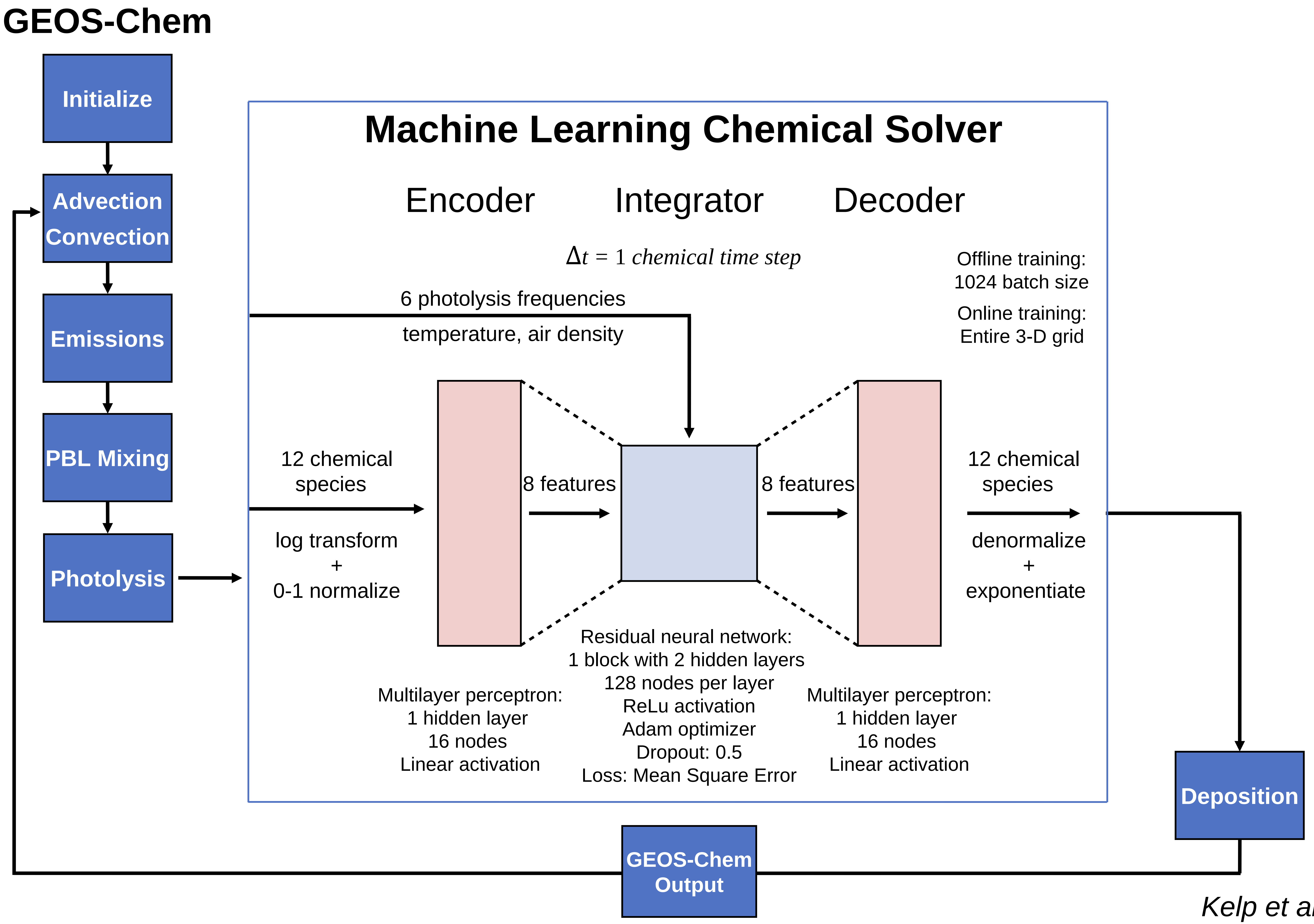
1 month dataset would contain:

lon×lat×lev×days×hours =

46×72×~25×31×24 **->**~62 million samples

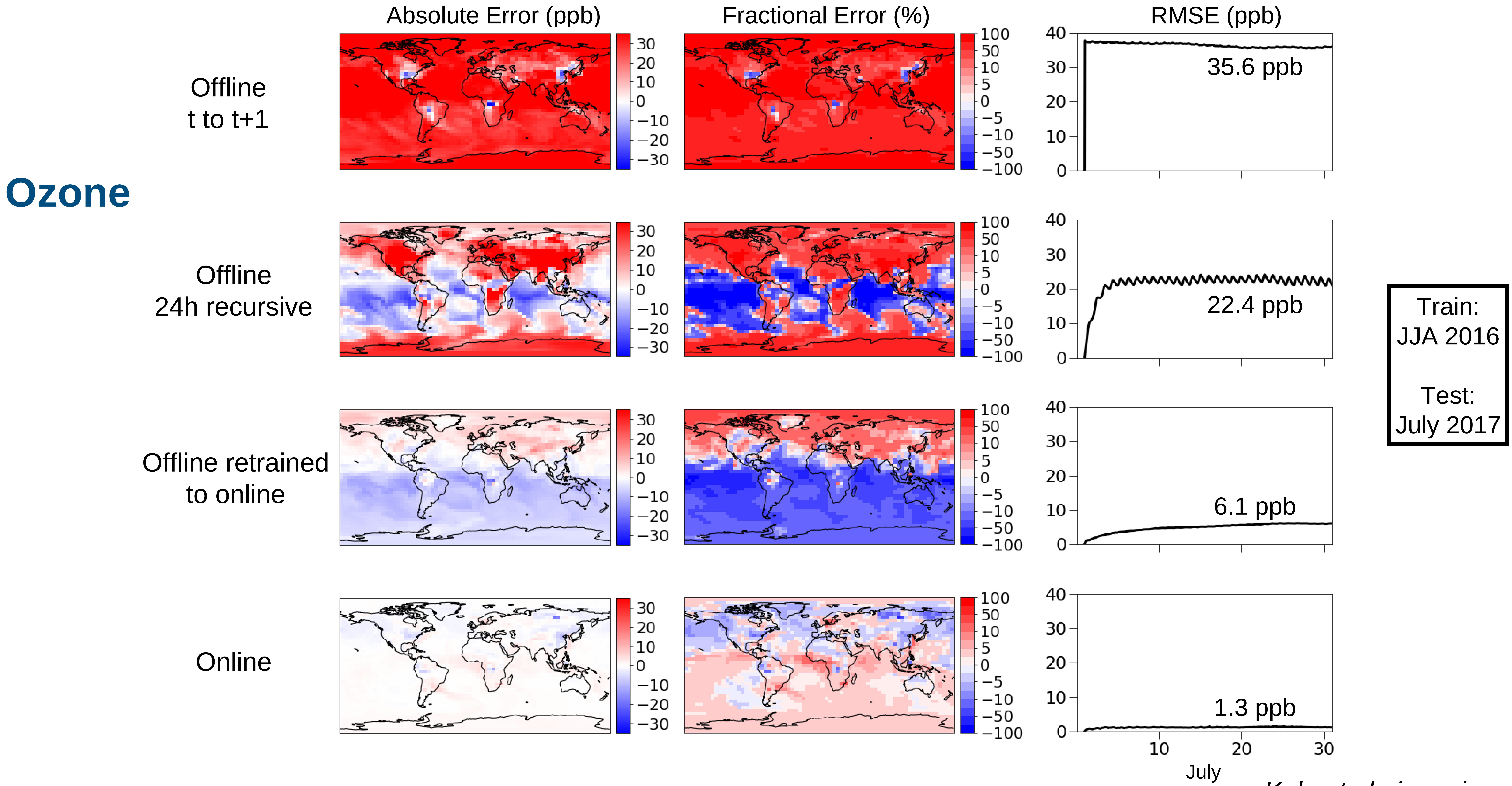
Training: 2016, Test: 2017

Kelp et al., in review



Kelp et al., in review

Online training improves accuracy and stability over offline training

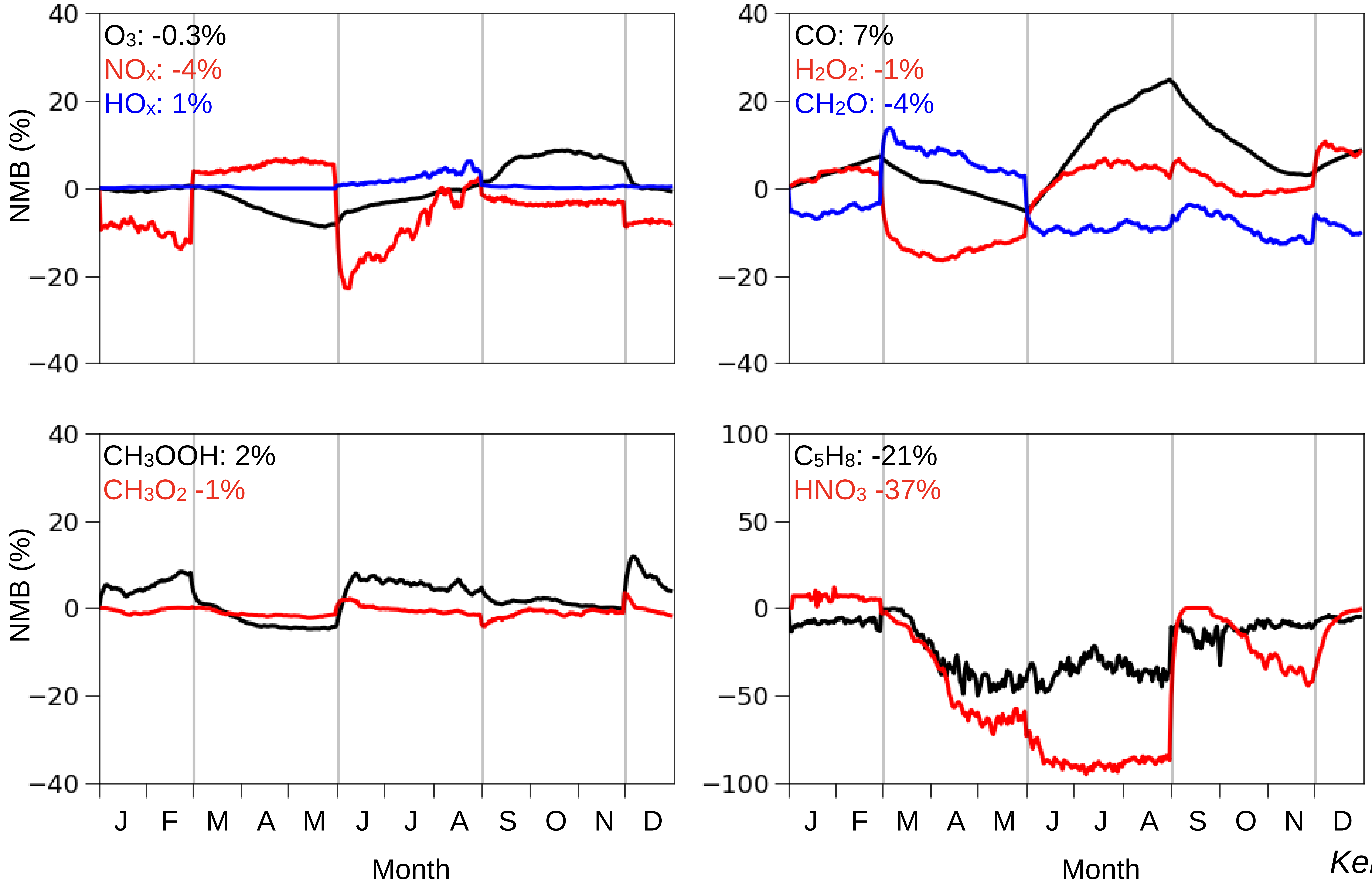


Kelp et al., in review

ML solvers have different seasonal fits of accuracy

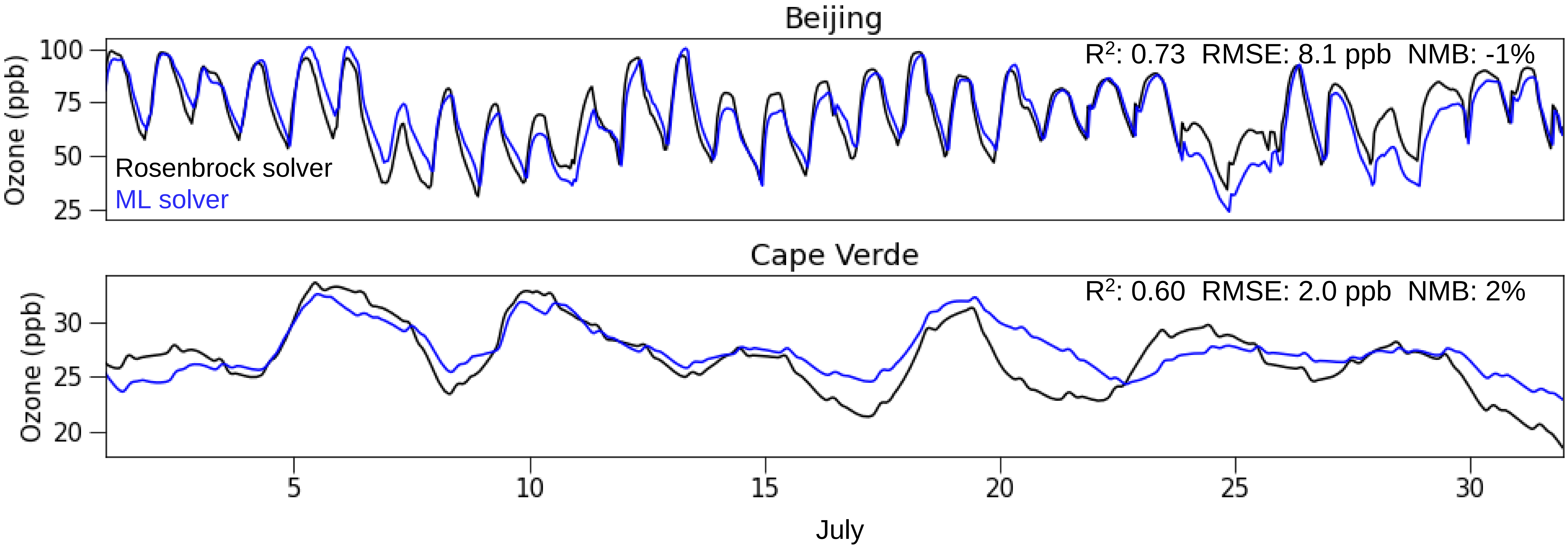
Separate ML solvers for:

- Species
- Season



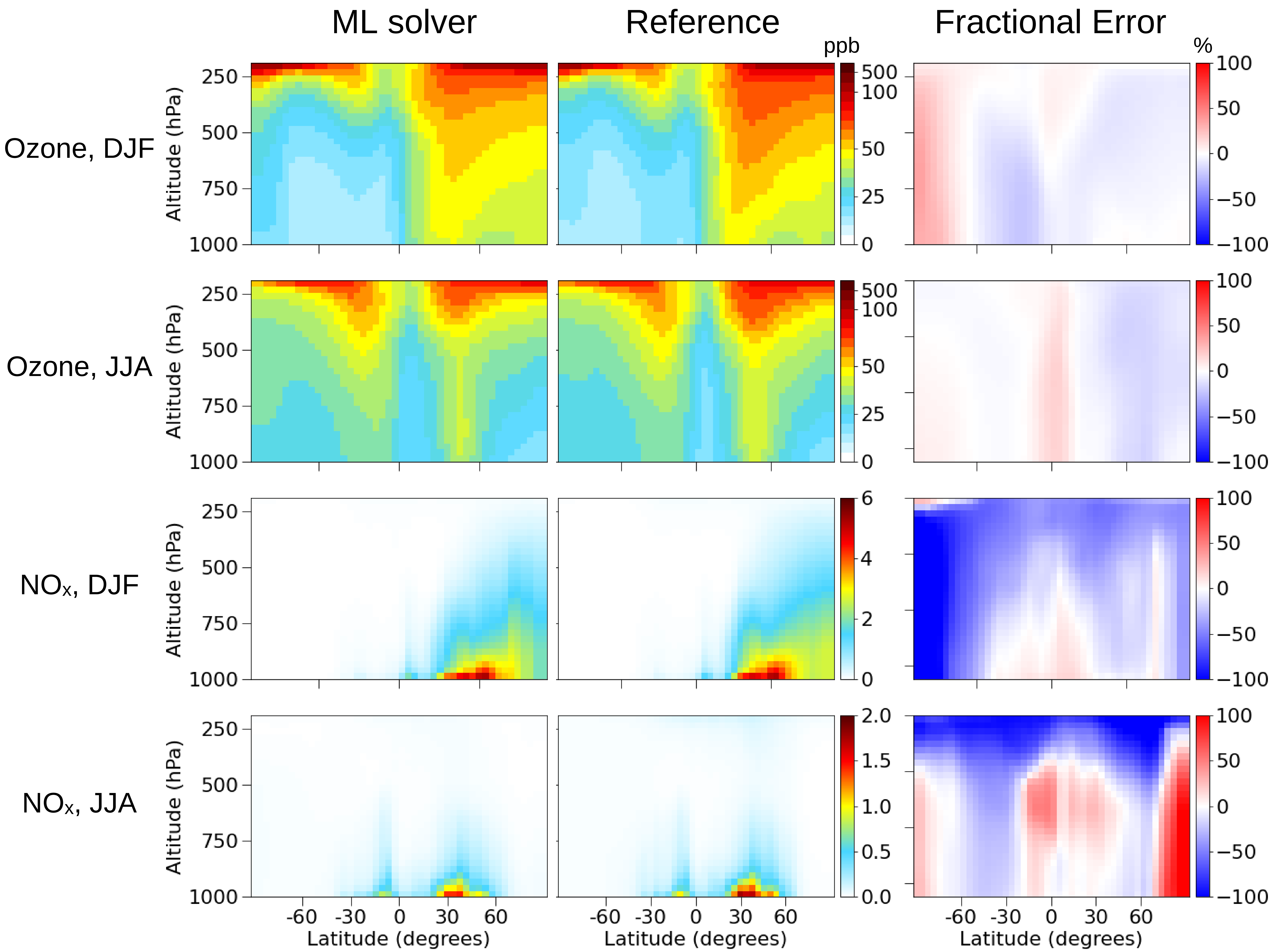
Kelp et al., in review

ML solver able to capture the diurnal and synoptic variability of ozone at polluted and clean sites



Kelp et al., in review

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



Kelp et al., in review

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.

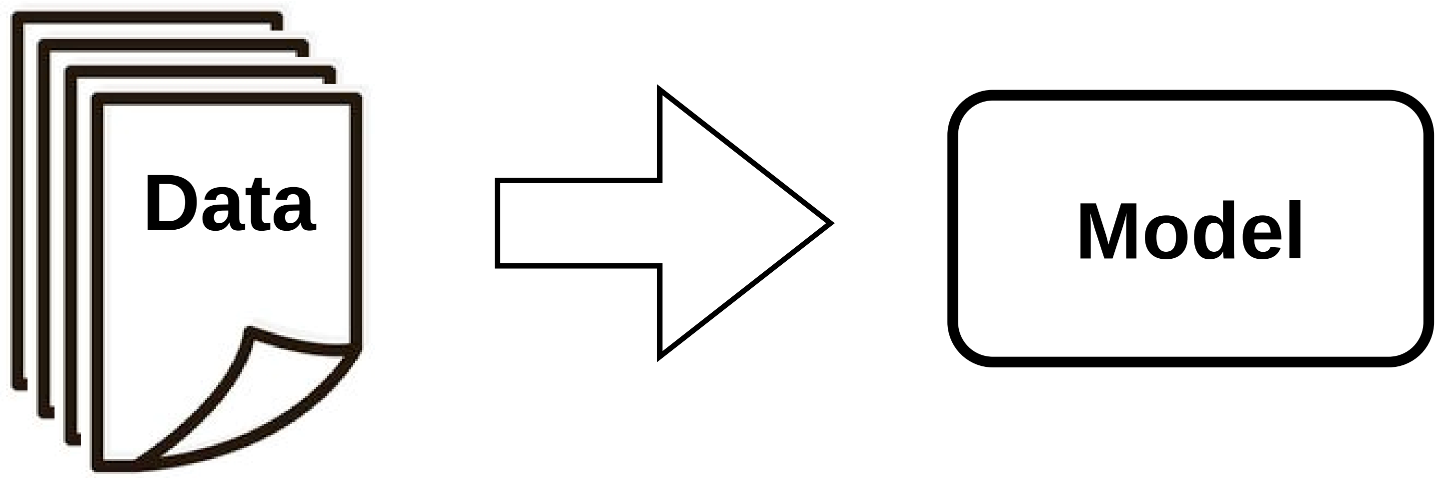
- Large regional biases for ozone and NO_x under remote conditions where chemical aging leads to error accumulation.
  - Regional biases remain a major limitation for practical application, and ML emulation would be more difficult in a more complex mechanism.

home about me research

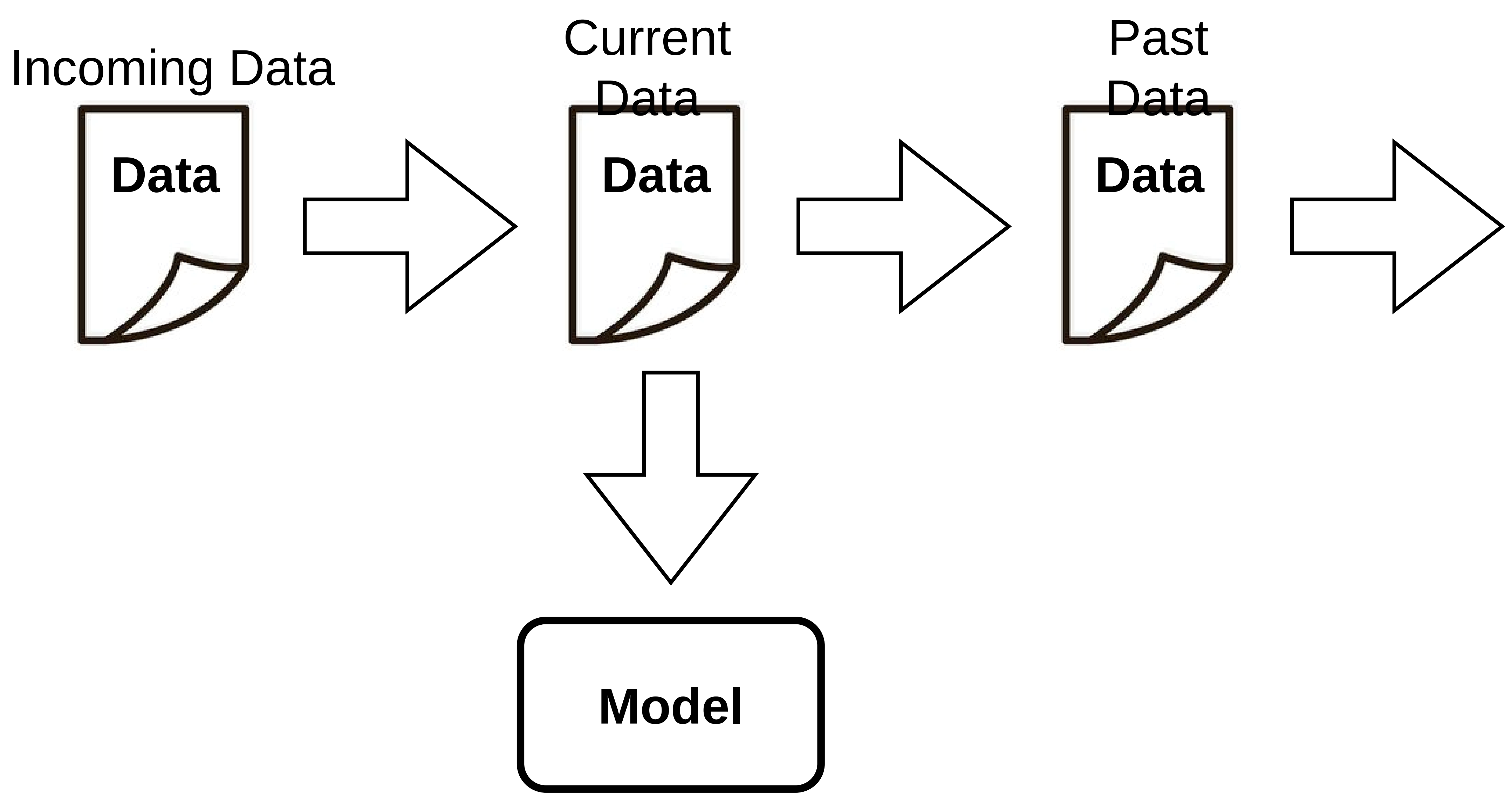
Makoto Kelp

Online learning prevents overfitting to training data

Offline (batch) learning



Online (sequential) 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

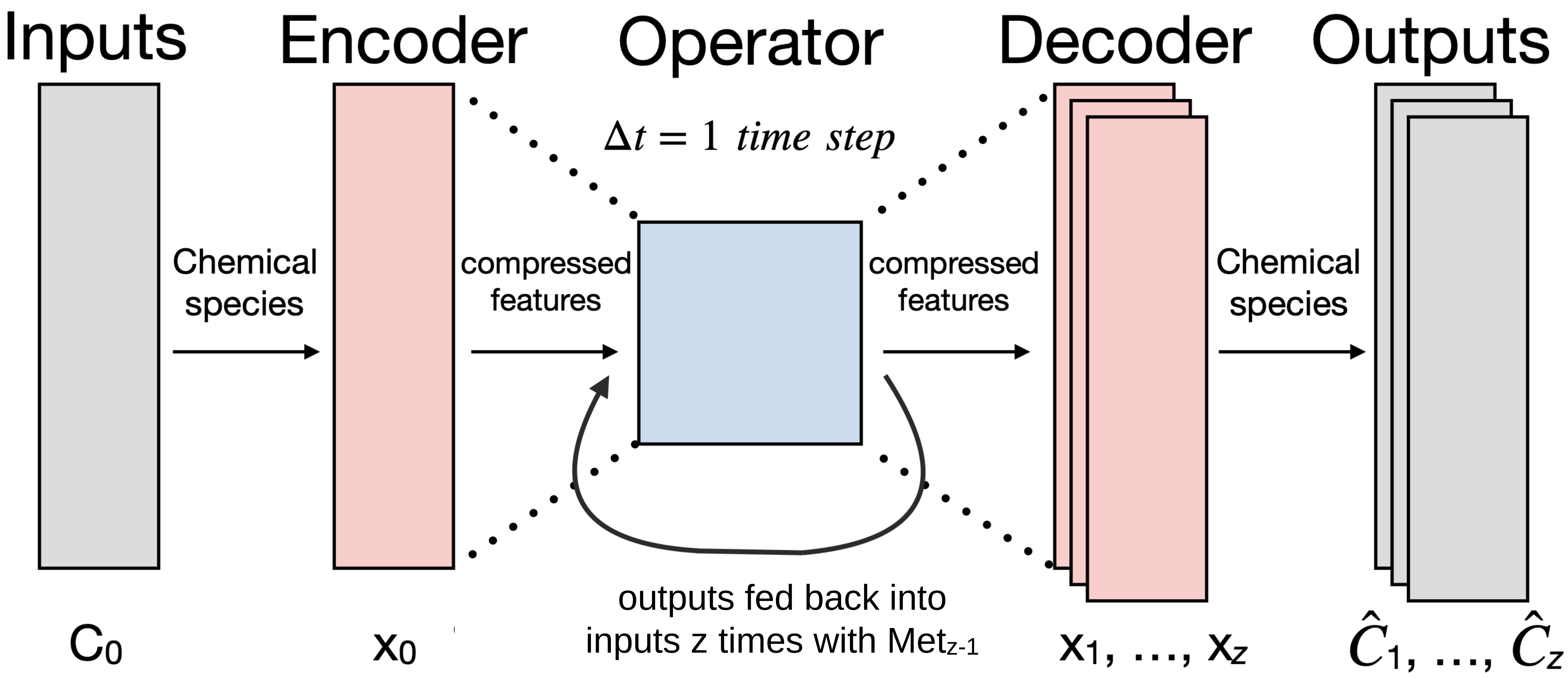
Pros:

Cannot overfit: each data point u
Representative realizations
No need to generate data archiv

Cons:

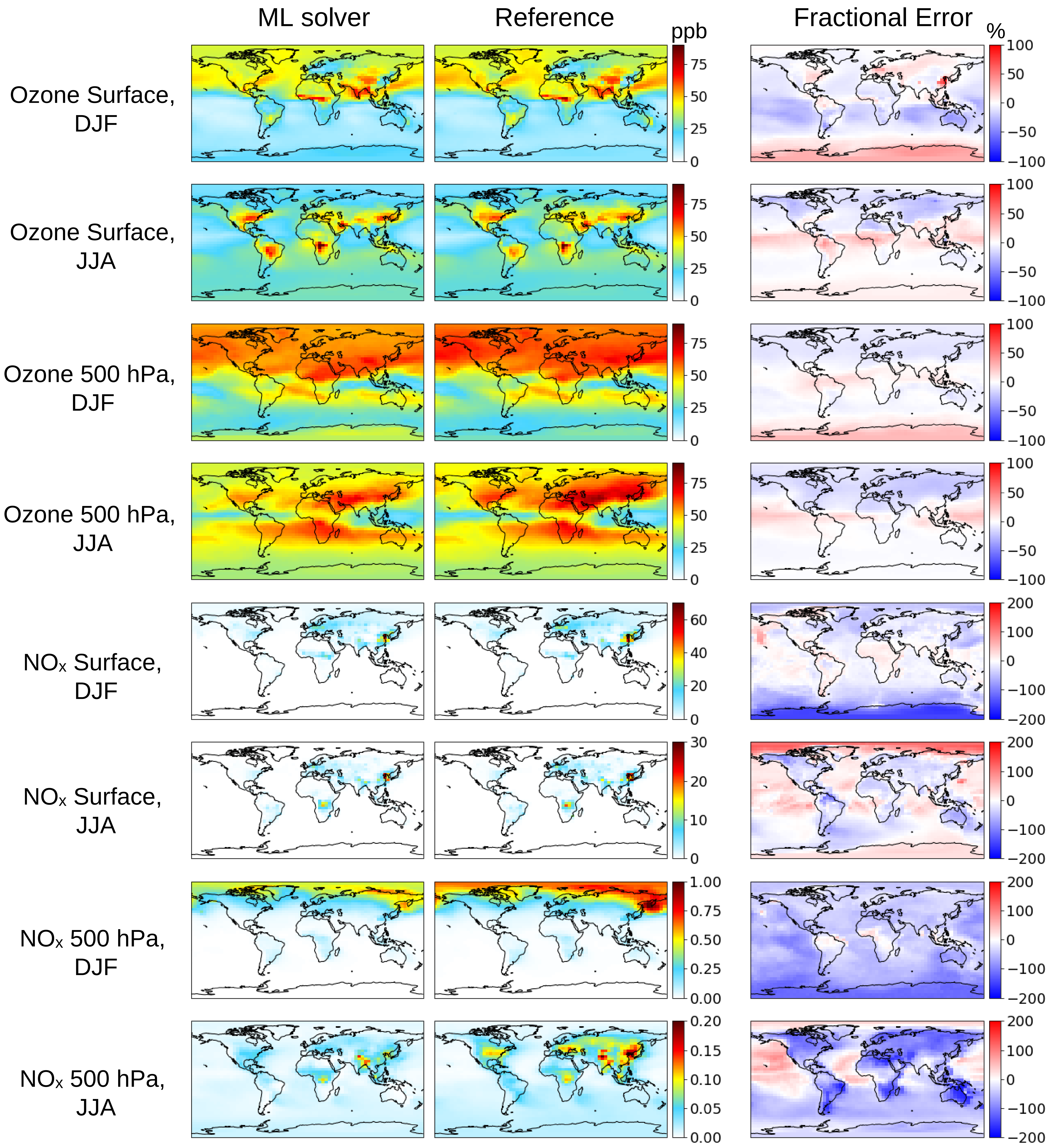
Hard to implement
Very expensive training! (Each C
Limited observational window
“Catastrophic forgetting problem

New model framework 1) compresses dimensionality and 2) captures slower chemical modes during ***training***



Mechanism: CBM-Z/MOSAIC Box model
101 species: 77 gas, 24 aerosol
4 meterological variables: T, P, RH, Solar angle

Kelp et al., 2020 JGR: Atmospheres

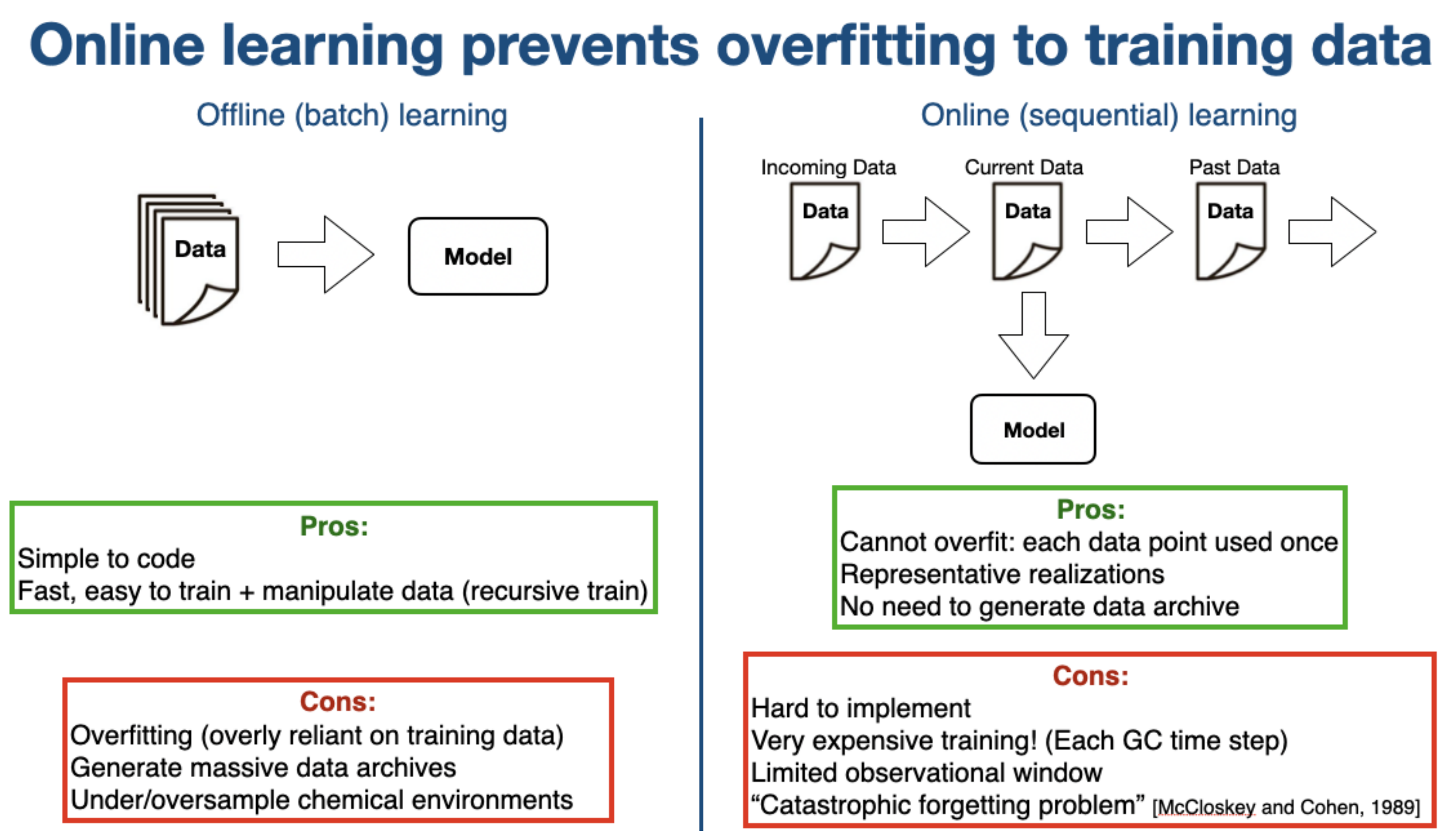


Largest errors are in polar sunlit conditions where the effect of chemical aging during long-range transport is particularly important

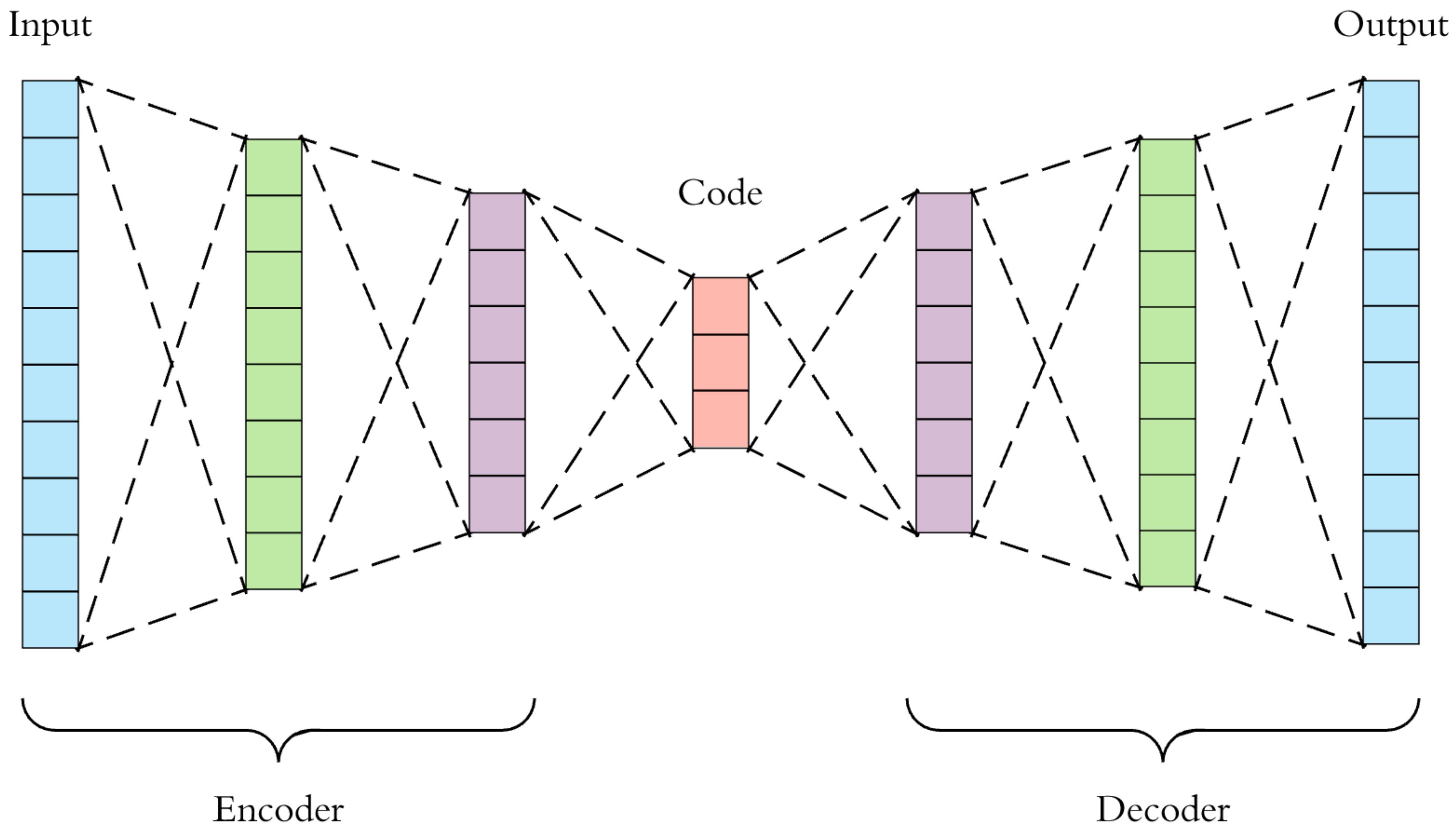
Kelp et al., in review

Steps moving forward

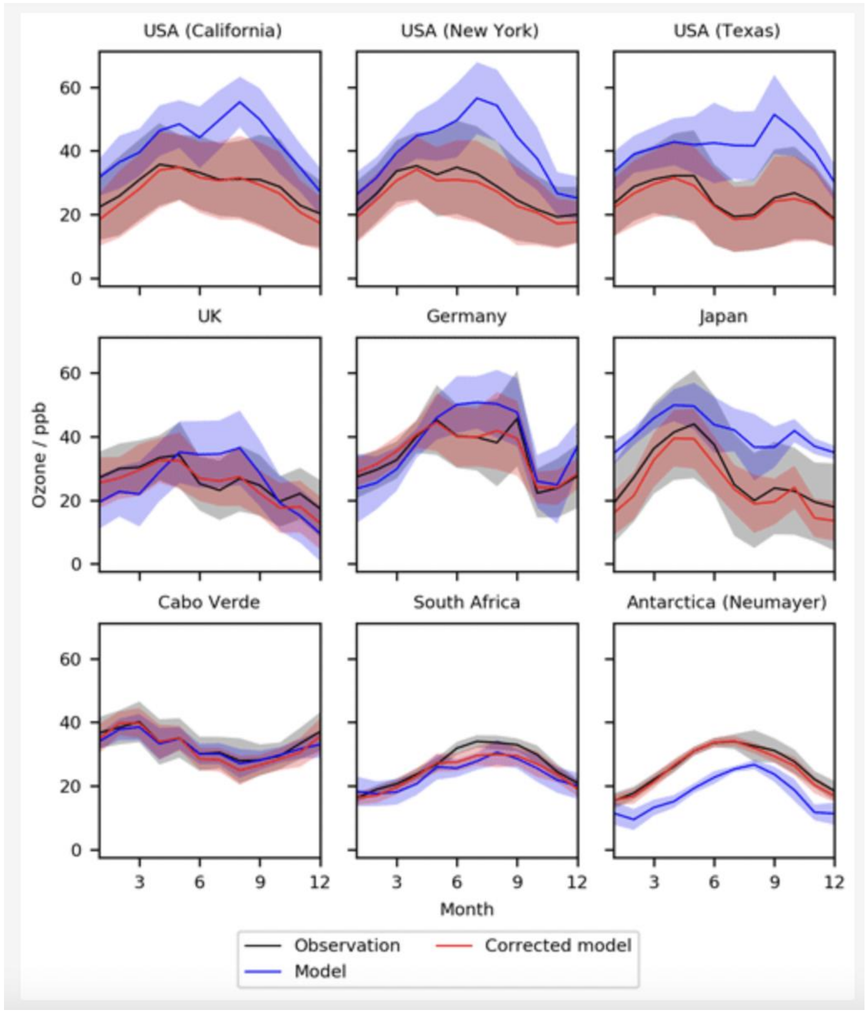
1. Mimic and improve online training offline



2. Use Encoder/Decoder to reduce dimensionality of standard mechanism



3. Train a ML bias corrector to nudge ML toward Rosenbrock



Ivatt and Evans (2020)

4. Train a GAN for failure prediction and call Rosenbrock

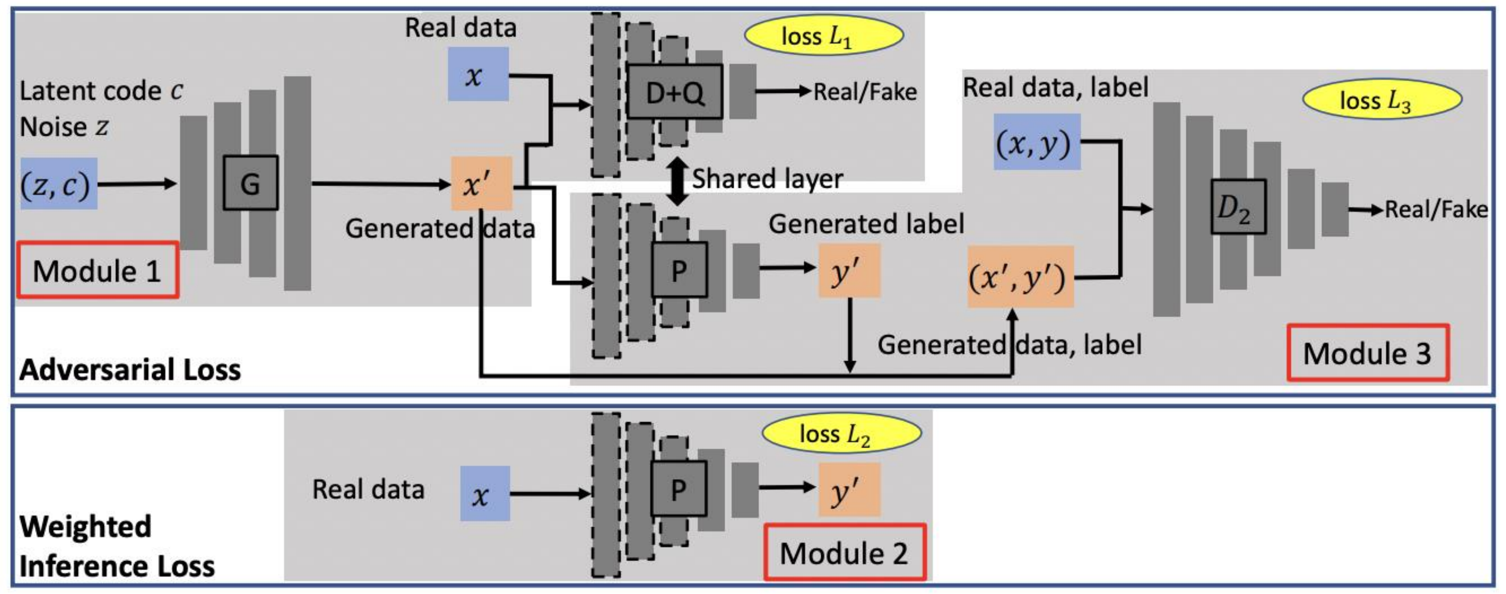


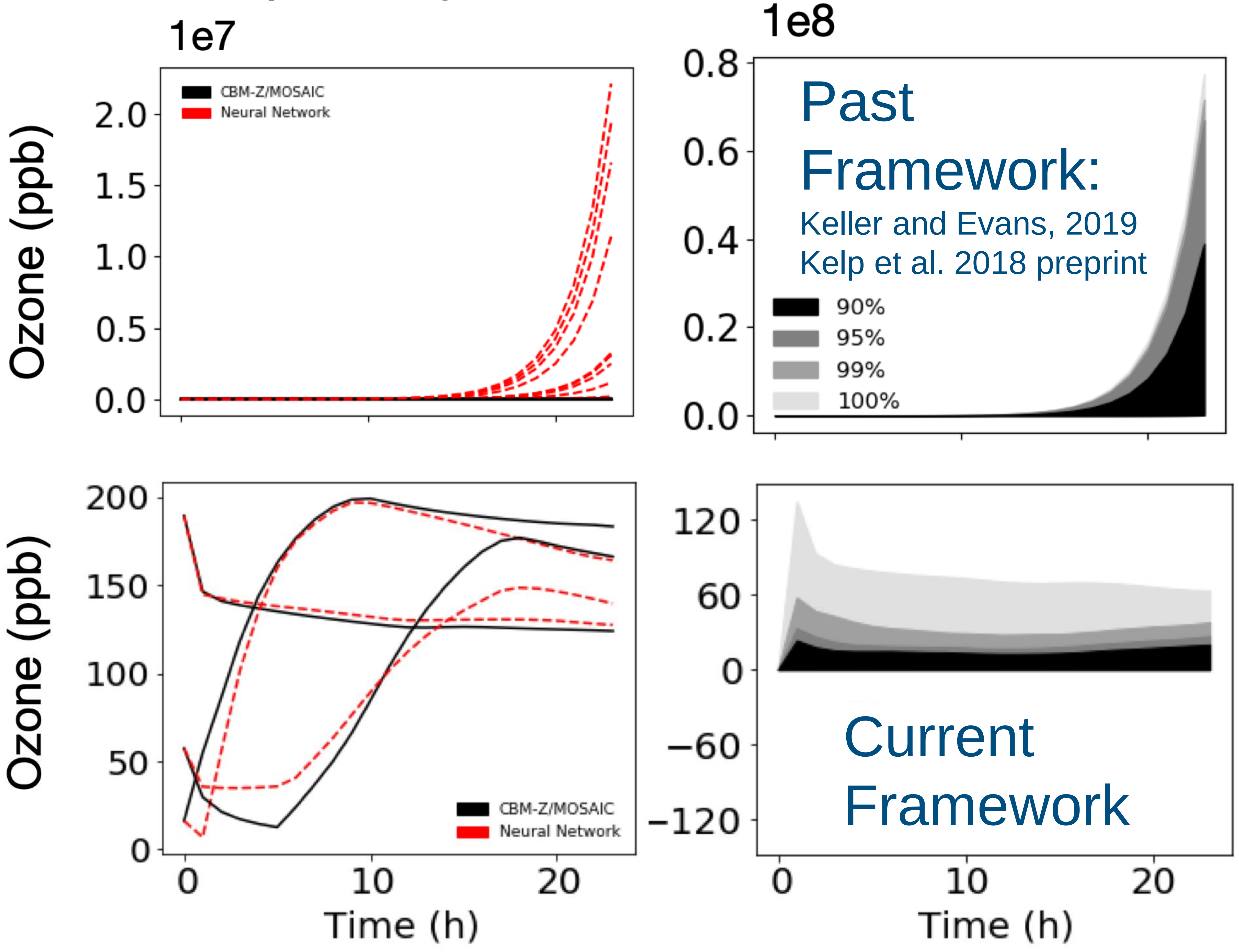
Fig. 3: GAN-FP architecture: there are 3 modules. Module 1 (network G, D and Q) is used to generate failure and non-failure samples using adversarial loss L_1 (Eq.(3)). Module 2 (network P) is an inference module with weighted loss L_2 (Eq.(5)), which trains a deep neural network using real data and label. Module 3 (network P and D_2) is a modified CGAN module with adversarial loss L_3 (Eq.(6)), where network D_2 takes data-label pair as input and tries to distinguish whether the pair comes from real data label $(\mathbf{x}, y)$ or from generated data label $(\mathbf{x}', y')$.

Zheng et al. (2019)

New ML model able to prevent error accumulation over time horizon of interest + achieve orders-of-magnitude speedups

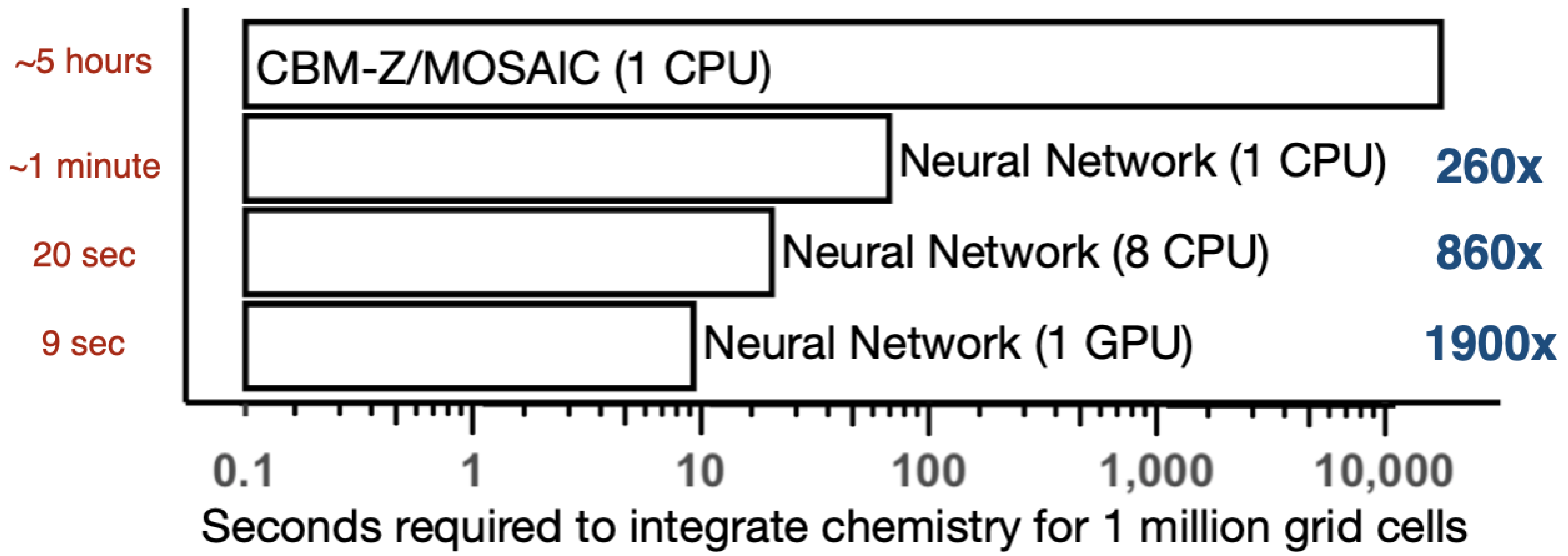
Sample Trajectories

Absolute Error



1 million test cases

Timing Results



Kelp et al., 2020 JGR: Atmospheres