Deep learning emulation and compression of an atmospheric chemical system using a chained training regime

Makoto Kelp[1], Nathan Kutz[2], Julian Marshall[3], Christopher Tessum[4]
Harvard University — [1]Dept. of Earth and Planetary Sciences
University of Washington — [2]Dept. of Applied Mathematics, [3]Dept. of Civil and Environmental Engineering
Illinois at Urbana-Champagne — [4]Dept. of Civil and Environmental Engineering
Contact: mkelp@g.harvard.edu

CACES

Motivation

Modeling atmospheric chemistry is vital to major environmental problems including air pollution and climate change. These models are **computationally expensive**, largely because of the high cost of solving systems of coupled differential equations. Previous studies have shown that machine-learned chemical mechanisms can be orders-of-magnitude less computationally expensive than traditional methods but tend to **suffer from exponential error accumulation** over longer simulations [1, 2]. Here, we present a modeling framework that **reduces error accumulation compared to previous work while maintaining computational efficiency.**

Methods

- We create an encoder-operator-decoder neural network (**Fig. 1**) to emulate the CBM-Z/MOSAIC mechanism. Error function prioritizes ozone (O_3).
- **I/O:** CBM-Z/MOSAIC predictions of changes in concentrations of **101 gas- and aerosol-phase chemical species** over 24h, given a range of pseudo-randomized chemical and meteorological initial conditions.

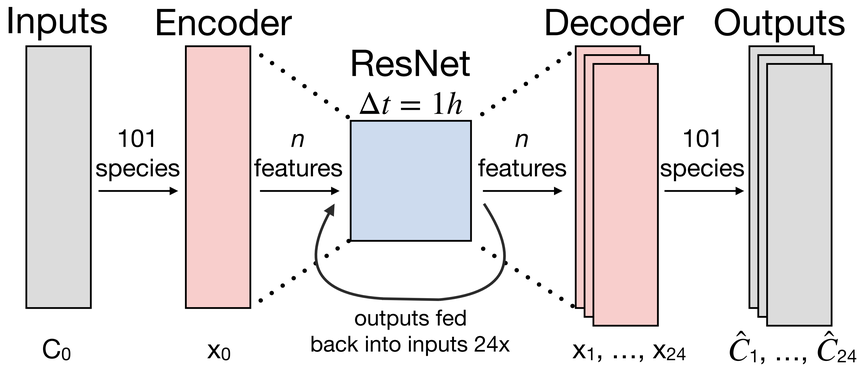

Fig. 1. Encoder-operator-decoder residual neural network architecture

Conclusions and Future Work

- The recurrent training regime results in extended simulations **without exponential error accumulation**.
- The neural network can reversibly **compress the number chemical species by >90%**, leading to a smaller memory footprint.
- We observe a **~260× reduction in computation time** compared to the reference mechanism (~1900× on a GPU).
- For O_3, our model predictions match those of CBM-Z/MOSAIC with **~ 6% mean error across 99% of all randomly initialized simulations**. The top 1% of simulation error can be significantly higher, but these results are qualitatively similar to the reference model.
- We show that these models **may be extended to a week without exponential error**.
- Future work: reduce top 1% simulation error, characterize embedding groupings, and take steps to implement this framework into a chemical transport model.

Results

- Recurrent training regime curtails exponential error amplification compared to single-step training (**Fig. 2**).
- For O_3 with training dataset concentrations ranging from 0–500 ppb, neural network predictions differ from CBM-Z/MOSAIC predictions over 24h by <8.2 ppb in 90% of all simulations (mean rel. error: 5.5%), 27 ppb in 99% of all simulations (mean rel. error: 6.2%). However, the top 1% of simulations predictions differs by up to 133 ppb (mean rel. error: 16.8%) (**Fig. 2, 3**)
- The encoder can faithfully compress 101 species into 4 features when noise is added during training (**Fig. 4**)

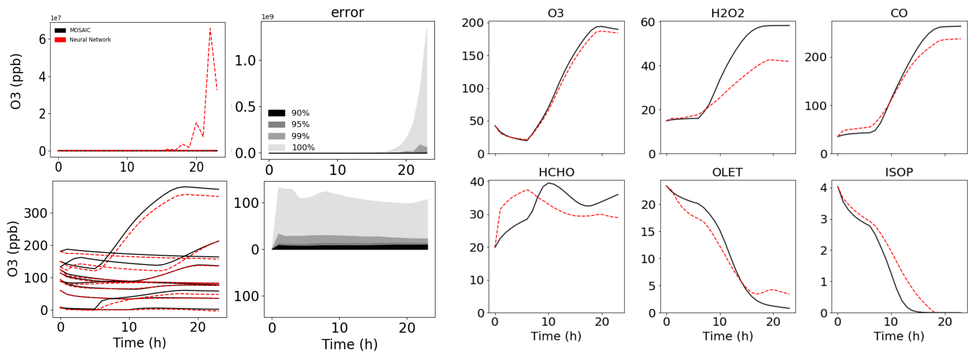

Fig 2. O_3 sample performance over 24h for 10 simulations (left), with absolute error (right) metrics for one million simulations for a neural network trained on single steps (top) and on a chained training regime (bottom).

Fig 3. Example time evolution of O_3 and other dominant species in CBM-Z/MOSAIC (black) compared to the neural network (red). H_2O_2 and VOC species (CO, HCHO, OLET, ISOP) help control O_3 formation in the CBM-Z/MOSAIC model.

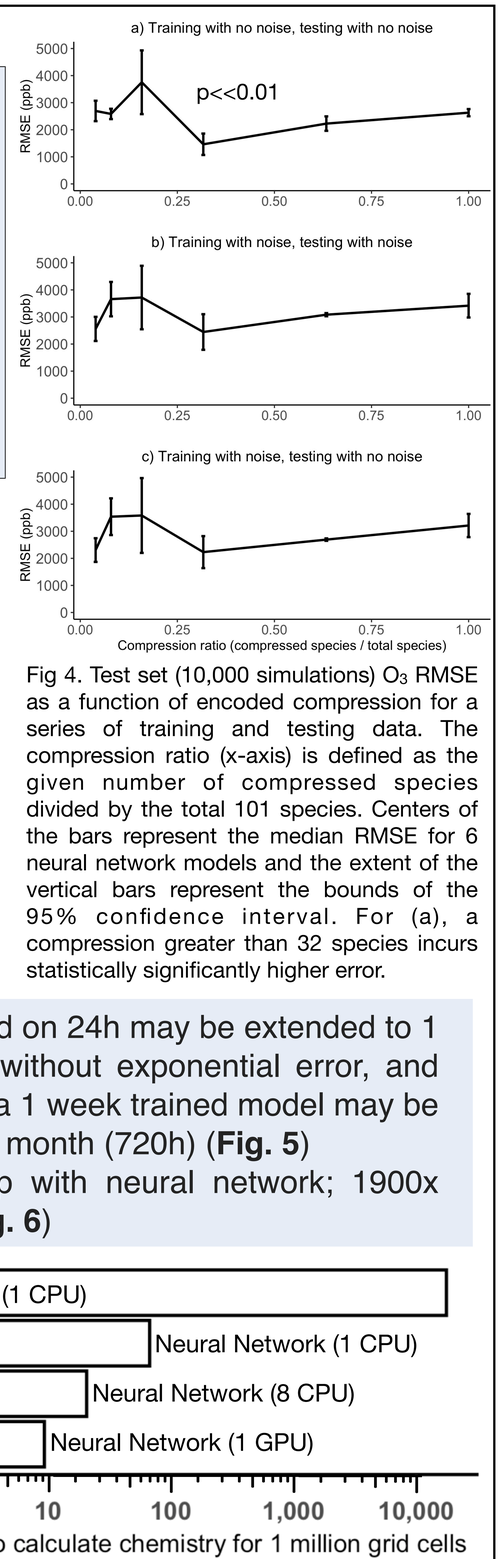

Fig 4. Test set (10,000 simulations) O_3 RMSE as a function of encoded compression for a series of training and testing data. The compression ratio (x-axis) is defined as the given number of compressed species divided by the total 101 species. Centers of the bars represent the median RMSE for 6 neural network models and the extent of the vertical bars represent the bounds of the 95% confidence interval. For (a), a compression greater than 32 species incurs statistically significantly higher error.

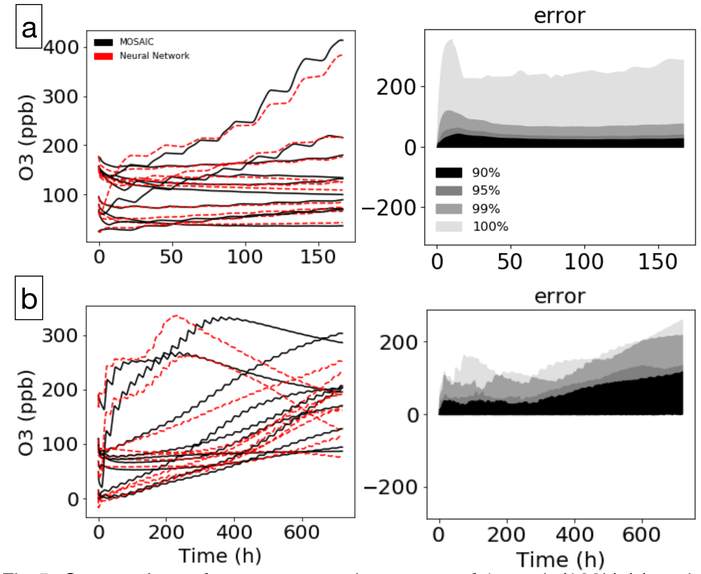

Fig 5. O_3 sample performance over the course of 1-week (168h) (a) and 1-month (using 1 week model, 720 hours) (b) for 10 simulations (left), with absolute error (right). All simulations have random initial conditions. Percent in the right panels represent the error percentile for the entire testing dataset.

- Models trained on 24h may be extended to 1 week (168h) without exponential error, and similarly with a 1 week trained model may be extended to 1 month (720h) (**Fig. 5**)
- 260x speedup with neural network; 1900x with GPU (**Fig. 6**)

Fig 6. Time required for one million independent simulations.

References: [1] Kelp, et al. (2018), arXiv:1808.03874., [2] Keller and Evans (2019), *Geosci. Model Dev.*, *12*(3), 1209-1225.
Funding: This poster was developed under Assistance Agreement No. RD83587301 awarded by the U.S. Environmental Protection Agency. It has not been formally reviewed by EPA. The views expressed in this poster are solely those of the authors and do not necessarily reflect those of the Agency. EPA does not endorse any products or commercial services mentioned in this poster.