Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2019-85-RC1, 2019

Atmospheric Chemistry and Physics
Discussions
Open Access

Interactive comment on “Technical Note: Deep Learning for Creating Surrogate Models of Precipitation in Earth System Models” *by* Theodore Weber et al.

Nathan Urban (Referee)

nurban@lanl.gov

Received and published: 9 May 2019

This is an interesting paper on the under-studied field of deep learning for climate forecasting. I am having some trouble evaluating the skill of the predictions, and have concerns about how the skill changes (or rather, doesn’t) with lead time.

Major comments:

I am not sure the best skill comparisons are being made. Persistence forecasting is a low bar to meet. It would be better to compare to some time series forecasting method that has memory, like a damped persistence forecast from an AR(1) model fit to each

grid cell, or more generally, a multi-lag ARIMA model or something with automatic model order selection, or Holt-Winters exponential smoothing. (I acknowledge that this isn’t always done in other papers, but I feel it’s a stronger comparison, and not necessarily that hard to do.)

Also, the data don’t appear to be deseasonalized, which leads to odd-looking oscillations in comparisons to persistence, making it harder to compare.

I would expect the forecast model skill to asymptotically approach persistence as dependence on the initial state is lost, but this doesn’t appear to happen. Why not? Is it an artifact of the way persistence is being defined? Is it because there is some skill in the non-stationary forced response (these are transient runs, not control) that the neural network learns? In general, this particular point requires more careful treatment, since a lot of skill at decadal timescales can be due to the climate signal, not actual initial-state predictability. I would expect this to be less true for precipitation, which doesn’t exhibit very strong decadal trends, but the point should still be discussed. It may have been better to use an unforced control run as training data, unless the point of the paper is to evaluate predictability coming from the forced response.

Even more worryingly, the skill of the forecast doesn’t seem to really change with lead time, for the 18-layer ResNet. Shouldn’t it be more skillful in the near term? Is there any initial-value predictability here at all? If not, where is the supposed skill improvement over persistence coming from? This continues to make me concerned that there is something problematic with how skill is being defined relative to persistence.

Other comments:

There should be more reference to the decadal prediction literature. Yeager et al., "Predicting near term changes ...", BAMS (2018) is one recent paper on initialized numerical model forecasts. They give useful diagnostics in addition to RMS error, like anomaly correlation coefficients. This manuscript should compare the predictability found by the neural network method to the predictability found by other decadal forecasts of precipitation, as in Yeager et al. (e.g. Fig. 5), or other papers — this is just one of many.

There are also prior examples of using neural networks (such as recurrent networks) for climate forecasting, for example McDermott and Wikle, Enivronmentrics (2018); Ouyang and Lu, Water Resources Management (2018).

I was confused by some of the exposition. It took a long time to understand that the forecasts are being made on the basis of 5 years (60 months) of precipitation data ... some details of the setup are introduced too late.

The scheduled sampling exposition was particularly confusing. At first I didn't understand why you wouldn't have all of the ground truth data available for forecasts. Then I realized that a fixed-window forecast (as opposed to a recurrent forecast) wouldn't have the full 60 months of data available; it would be using some model forecasts. This should be made more explicit, and also clarify a few details like what "ground truth" data is here (perfect-model data at lags before the forecast). Also, it would help to point out that "inference" is what climate scientists call "prediction". (To statisticians, "inference" often means "training" to a computer scientist.)

I'm not sure what Figure 2 is showing. The average of many 1-month-ahead forecasts take over 21 years of initial conditions? I don't know if this is a useful comparison. I would expect the average of bunch of very short-term forecasts will resemble the average of the actual data. Perhaps I am misunderstanding.

How many neural network parameters are there in total using this architecture, compared to the size of the data? Are there generalization error plots to assess overfitting?

Interactive comment on Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2019-85, 2019.