Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC1
https://doi.org/10.5194/gmd-2022-163-RC1, 2022

Comment on gmd-2022-163

Anonymous Referee #1

Referee comment on "Deep learning for stochastic precipitation generation – deep SPG v1.0" by Leroy J. Bird et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2022-163-RC1, 2022

This manuscript presents a regression-style deep neural network that serves as a single-site stochastic precipitation generator. The neural network relates daily (or hourly) precipitation to previous days (hours) precipitation and sinusoidal terms to capture seasonality. The output of the neural network is a mixture distribution with four components. The manuscript shows that the generator is able to match the statistical characteristics of observed precipitation series well. The manuscript then goes on to explore how to represent non-stationarity from climate change, assessing whether the quantiles of precipitation robustly scale with temperature and concluding that the estimated relationships may not be robust. Finally the manuscript scales the generated time series from the stationary model using the relationship with temperature derived from weather@home simulations as an exploration of a possible methodological approach.

I am a statistician with experience with climate statistics and so will concentrate on some of the statistical/machine learning aspects of the work. In general, I thought the work was a useful exploration of one way to use machine learning methods for precipitation simulation. It was interesting to see the use of a regression-based neural network rather than a network specifically designed to reproduce temporal structure such as an LSTM. The simulator seems to do a good job of capturing precipitation variation over time. However, I did have a number of questions about the methodological approach and thought there were some areas of lack of clarity in terms of the methods.

1.) The neural network modeling framework is complicated when one considers all the choices made in terms of the network structure (Fig. 1), choice of predictors (Section 3.1), the mixture distribution (a four-component mixture), and the optimization parameters (lines 219-221, 225-228). The authors do not describe at all how any of these choices were made, yet these choices are fundamental to the work and represent a potentially important contribution to the literature. Were these choices made empirically, based on trying different approaches on data (which could lead to overfitting, depending on how this was done)? Do the authors have any idea whether other choices would give similar performance? We get a little bit of information from the comparison to the regression, but

the black box nature of the model choice is troubling and the lack of discussion of how the methods were arrived at feels like a big omission.

On a minor note, I didn't really understand the rationale for the extra fully-connected layer (lines 180-181). In a standard neural network, one would map directly from the dimensionality of the inputs to the desired dimensionality of the first hidden layer. What does the more complicated structure at the top of this network achieve?

2.) I'm having some trouble understanding the loss results, summarized in Fig. 2 and Section 3.6.

- How can it be that the validation loss for the neural network is approximately constant over the epochs? (Perhaps the scale of the y-axis makes things hard to see?) Shouldn't the optimization result in a decrease in validation for the initial epochs? Given this, it's hard to understand the statement (line 265) about 10 epochs, as the validation loss for the daily model seems to have a minimum at the first epoch.
- I'd also like to understand how the validation was done. Was this basically one-step ahead prediction, using the actual observations as the inputs for each of the 1000 (or 10000) validation observations.
- The losses seem to be the average loss per observation. If one considers the total loss, one might be able to think more about whether the difference between the linear and neural network models is important. For example if one has nested statistical models, one can use a likelihood ratio test to assess whether the additional parameters in a more complicated model are warranted given how much better the loss is compared to a simpler model. One can't do that exactly here, but with 1000 observations, I think that the negative log-likelihood for the neural network is 926 compared to 932 for the linear model. That doesn't actually seem to be a very big difference (considering the usual chi-square statistics used in likelihood ratio test), so I'm not fully convinced that (particularly for the daily case) a neural network is really doing that much better than a simpler model.
- If the validation data were used for early stopping, then they can't really be used to compare the neural network and regression models. One would presumably need a three-way split of the data, with some test data fully held out of the fitting process, including the early stopping determination.

3.) The assessment of the simulated series against observations explores various aspects of the pattern of precipitation and is generally (though not entirely) reassuring in terms of the simulations matching the statistics of observed precipitation. However, I think the comparison could be improved by simulation more time series, since the whole point of a generator is the ability to generate as many time series as desired. For example, with many simulated series, the authors could show predictive uncertainty bounds in the quantile-quantile plots (e.g., Fig. 3) to better understand if the divergence between the simulated and observed series at high quantiles might be explained by stochasticity to some degree. Similarly, the autocorrelation plots show a lot of wiggliness (particularly Fig. 5) because of the limited number of data points. This is unavoidable for the observations, but if one simply simulated more series, one would probably get autocorrelation estimates that vary smoothly with time lag for the generator.

4) I found the sections on non-stationarity somewhat hard to follow.

- Line 344 says that within each percentile the data were averaged over each calendar year. I'm not sure what this means. If we take an observed time series and calculate, say, the 60th percentile within each year, that is a single number for each year. So I don't know what numbers are being averaged in each year.
- Related to this, I don't know where the uncertainty bands in Figs. 11-13 come from.
- I generally found Section 5.3 hard to follow. In particular, I can't tell if the text in lines 387-391 is meant to summarize the text that follows, or if it describes prepatory steps that precede the steps described in the steps that follow. I.e., are the 'correlation' and 'scaling' discussed here what is described in more detail in lines 406-409?

5) In addition, I had some concerns about the formulation of non-stationarity.
- It would be helpful to see some evidence (e.g., based on scatterplots of input data overlaid with the fit) of whether the functional form relating precipitation to temperature (line 347) fits observed/modeled data well.
- Particularly for the observations, could there be other factors (in particular aerosols) that affect precipitation and might be correlated with temperature?
- Are the relationships in Fig. 16 (particularly for Christchurch) believable?
- The analysis of non-stationarity and the scaling relationships don't seem to account for seasonality. Would the scaling relationships be expected to be the same for different seasons?
- The discussion highlights the potential of the generator for use in climate change assessment -- this seems to depend critically on being able to estimate the scaling relationship with temperature, which the work casts some doubt on.