Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC2
https://doi.org/10.5194/gmd-2022-146-RC2, 2022

Comment on gmd-2022-146

Omar Jamil (Referee)

Referee comment on "Data-driven Global Subseasonal Forecast Model (GSFM v1.0) for intraseasonal oscillation components" by Chuhan Lu et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2022-146-RC2, 2022

General comments:
It is good to see more work being done in machine learning applications for weather and climate. This work aims to build on the work done by Rasp et al (2020) by applying SE-Resnet instead of Resnet. The focus here is on learning the intra-seasonal oscillations components. The physical model being compared to is CFSv2.

General scientific comments:
I have some concerns about the methodology being employed in this paper. One of the first things lacking in the paper is a detailed description of data processing. With data-driven methods it is really important to be transparent about how the data were processed and applied. The reader is referred to WeatherBench paper, but the authors of this paper must outline what steps they took ensure their machine learning models were trained properly e.g. training-validation split, data standardisation/normalisation, what were the inputs for the neural network, how were the inputs structured? All these details are not only important for understanding the performance of the model, but also reproducibility.

As the field of machine learning is an empirical science, it would be good to see the reasons behind why certain architectures were chosen e.g. number of residual blocks was set at 17 with two convolutional blocks -- why is this the best setup?

I am also concerned by point-to-point prediction setup with 30 models for different lead times. This suggests the models are not generalising and being optimised for a very specific subset of the problem. I would suggest asking the question, what is the real-world application of this? Training a separate model for each day of lead time is not efficient and suggests the models are not learning any of the underlying physical processes and therefore require retraining every time there is an improvements in the training model data.

There are extensive comparisons between CFSv2, Resnet, SE-Resnet, Climatology error and accuracy, which could benefit from having a table to make it easier to interpret, but some of the differences are so small that I would question its statistical significance. The main objective of the paper appears to be to show how deep learning can improve on the physical model's sub-seasonal forecast. However, for that comparison to be fair, it is important to show how well CFSv2 does under different setups and against other physical models. My question would be why choose CFSv2 for this comparison? Is that best model there is for these lead times? Also, the comparisons between SE-Resnet and Resnet do not appear to be statistically significant. In order to understand how different parameters impact ML model's performance, a Pareto front type analysis is required. For understanding the impact of random noise on the ML model, cross-validation should be used to show the difference between two different architectures is consistent and statistically significant.

Specific comments:

Figures 2, 7, A1 need to be bigger because in their current size it is almost impossible to see the details being discussed. Perhaps this was just formatting issue for the pre-print.

Line 246: 1.01% lower RMSE is a very small difference. It would be good to see statistical significance. Please see above my suggestions of Pareto front and cross-validation.

Lines 252-254: There is some odd phrasing which talks about 50% of deep learning models forecast being below climatology, but for CFSv2 50% of forecasts are above climatology. I am sure the authors did not mean to suggest this, but it currently reads as if CFSv2 and deep learning models are the same, but phrased in favour of deep learning models.

Line 265: 0.75% RMSE improvement. Statistical significance?

Line 272: It would be good to see the RMSE values for other models quoted too. A table of comparison would be really useful for at-a-glance interpretation.

Line 288, 291: SE-Resnet and Resnet are so close that there appears to be no difference between them. What is the main benefit of SE-Resnet for this application?

Line 315: SE-Resnet and CFSv2 are very close, so what are the improvements being brought about by SE-Resnet?