Review for JAMES manuscript 2020MS002324 by Benjamin Toms, PhD

The authors apply spherical convolutions and hemispheric-sharing weights to data-driven weather forecasting to assess whether these modifications to conventional CNNs can help improve forecast accuracy. The authors state that the intent of the paper is not to maximize accuracy, but rather to assess the relative accuracy using their various proposed methods.

The paper overall does present novel potential improvements to conventional CNN-driven weather forecasts. Based on the authors' findings, these suggested changes to conventional CNN approaches should be communicated to the geoscientific community because there is broad potential benefit. However, I think there are some discussions within the manuscript that could be changed to improve the manuscript's scientific clarity.

I wish the authors all the best, and I hope that my comments are helpful. I would like to be sure they know that my major and minor comments are intentionally only critiques of their paper, and that I do think their paper overall presents novel, useful advances to data-driven weather forecasting. I look forward to reading the final product.

Major comments:

1. The authors use a fully convolutional architecture, which enables the architecture to intake reanalysis data of arbitrary resolution. However, because the size of the convolutional kernels is not changed between the low-resolution and high-resolution datasets, the kernels are extracting information from spatial patterns of different dimensions. For example, a 3x3 kernel for the high-resolution data examines a 4.2-degree by 4.2-degree box and for the low-resolution data an approximate 8.4-degree by 8.4-degree box. This may have significant implications for the accuracy of each model. With that said, it is likely still worthwhile to study the impacts of spherical convolutions and hemispheric sharing of weights for each resolution separately.

2. The statement in Lines 353 through 357 is not conclusive based on the findings of this study. Because the purely data-driven model presented in this manuscript does not include the full extent of physics included in dynamical models, the errors within the forecasts could be simply caused by missing statistical representations of physics within the CNNs. This would lead to the point in parentheses being the primary driver of error – that the presented method is insufficient for accurately predicting the atmosphere. Because this study does not optimize the CNNs to be as accurate as possible, the predictability of the atmosphere itself cannot be assessed. Only the fidelity of the modeling strategy can be assessed.

3. The argument presented within the conclusion that the present paper improves over other studies is not justified, since the intents of each paper is different. This paper does offer additional options for improving data-driven models built for weather prediction but does not improve upon the accuracy of predictions presented by Weyn et al. (2020),

for example. These claims could be made if the same architecture used by Weyn et al. (2020) was used in this paper, and then the suggested modifications to the convolutional kernel structure and hemispheric weight sharing were applied. It is not clear from this paper whether the re-gridding approach of Weyn et al. (2020) or the spherical convolution approach is superior for forecast accuracy. The authors state this themselves multiple times, such as in their final statement of the conclusions, and so as a reader, I would have felt more confident in the author's findings if they had stuck to this tone.

4. I would have appreciated a more thorough discussion of why the spherical convolution + hemispheric-shared weights approach is the most accurate within the first few days,but is then surpassed by the non-spherical convolutions with hemispheric-shared weights. Do the authors think that these different types of kernels are extracting different meteorological patterns? Or is this a numerical stability issue, whereby the spherical convolutions capture the physical patterns more accurately but numerical error leads to a cascading increase of total forecast error?

5. I don't question the authors' knowledge on the topic, but their analysis is very targeted towards one particular implementation of spherical CNNs. There are a few other papers that discuss the usage of spherical CNNs that may provide more contect for the reader of the usage of such methods in geoscience and computer science should they be discussed in this manuscript. Two examples of such papers are Jiang et al. (2019) and Cohen et al. (2018). Links here: https://arxiv.org/abs/1901.02039; https://arxiv.org/abs/1801.10130

Minor comments:

1. There are quite a few typos throughout the manuscript that need to be cleaned up. I haven't commented on all of them here, since the authors can edit these for themselves.

2. I understand that saying "low-resolution" and "high-resolution" can sound repetitive, but I suggest the authors change all instances of "lres" and "hres" to their longer counterparts. This would improve the readability of the manuscript, in my opinion. This also applies to the "hod" and "doy" abbreviations.

3. It would be helpful if the temporal resolution of the data were more clearly stated. I can infer it is 6-hourly from the paragraph spanning lines 185 through 191, but a clear statement would be helpful.

4. Line 120: It would be helpful to clarify this sentence: “However, we use data on a regular grid, and allow only multiples of the gridpoint distance at the equator as points in the latitude direction.”

5. Line 186: This sentence could be clarified: “The input of the networks is comprised of two timesteps (2 variables 187 at 2 pressure levels each)”