EGUsphere, referee comment RC1
https://doi.org/10.5194/egusphere-2022-147-RC1, 2022

Comment on egusphere-2022-147

Anonymous Referee #1

Referee comment on "Classification of tropical cyclone containing images using a convolutional neural network: performance and sensitivity to the learning dataset" by Sébastien Gardoll and Olivier Boucher, EGUsphere, https://doi.org/10.5194/egusphere-2022-147-RC1, 2022

Review of: "Classification of tropical cyclone containing images using a convolutional neural network: performance and sensitivity to the learning dataset"

Summary:

This work involves training convolutional neural networks to detect instances of tropical cyclones in two reanalysis datasets, MERRA-2 and ERA5, using the HURDAT2 database as ground truth. The authors demonstrate a simple but effective CNN that performs well on both reanalysis datasets. They also show that CNNs trained on one dataset can be successfully applied to another and demonstrate that their classifier trained on ERA5 data has better performance than the classifier trained on MERRA-2 data.

Recommendation:

This is a useful study, with in-depth analysis of CNN performance across the two reanalysis datasets. Demonstrating that CNNs trained on one simulation can successfully detect key features in another (even at differing resolutions) is an important prerequisite to using them for online analysis of climate simulations. I do, however, have some concerns that are listed below. While I have made some suggestions for additional analysis that I think would significantly enhance the paper, it is not strictly necessary, and all of the comments should be addressable through edits to the manuscript. I recommend publication pending several revisions.

Comments:

This study uses a very simple CNN design with two convolutional layers based on a paper from 2016. CNN research has progressed dramatically since then and more sophisticated architectures with much better performance for image classification now exist. Many of these have already found their way into atmospheric studies. While there is nothing wrong with using a simple model (and the simple model performs quite well in this case) the last 8-years of CNN research should probably be acknowledged in the background section.

Lack of large training data and potential difficulties applying ML models to new datasets are both concerns/points mentioned in this study. Data augmentation schemes and model regularization techniques like dropout and batch normalization are proven ways to improve the robustness of CNN models when applied to new datasets and in the face of limited training data. These have become so ubiquitous that almost all CNN-based image recognition studies now use at least one of them. I think not using or at the very least acknowledging the existence of these methods is a major omission, since they are both very common in the literature and very relevant to this specific problem.

I remain skeptical of the claim in the conclusions section that the lower performance of the CNN on MERRA-2 implies that MERRA has lower information content for TC identification. I suppose one can argue that ERA5 has more information content simply because it is higher resolution, but that additional information may not improve cyclone detection. Only one CNN was tested, and perhaps a larger CNN, different input fields, different training procedure, data augmentation, etc. would yield better results on the MERRA data. Are all the storms present in HURDAT2 present in MERRA? If so, one could likely design a model better suited for the MERRA dataset. I think a more reasonable claim is made at the end of Section 4.2.2: "Thus we can conclude that the ERA5 dataset is more information rich than the MERRA-2 dataset for the classification of cyclone images *using our CNN."*

Section 3.2.3: It seems that a great deal of effort has been put into identifying non-cyclone training samples that thoroughly cover the region where cyclones have been observed. However, this appears to exclude any negative samples from other parts of the North Atlantic basin. Including negative samples from the Eastern Atlantic or the extratropical storm track region is important if the algorithm will ultimately be applied to the entire basin to ensure there is not unexpected behavior in these areas. The background section mentioned the potential of TCs moving farther north in the Atlantic in a warming climate.

Section 3.3 and 3.4: I was left confused about exactly how the training was performed, and I think more details should be included to improve reproducibility. Here are some notes:

- Information like the loss function and optimizer used are more important to include in the main text than details about the hardware used. This information is key to

reproducibility, but the hardware is not.

- Is the early stopping based on training loss (line 203) or test set loss (table A3 line 2)? Is there evidence of overfitting without early stopping?
- Which dataset was used for the hyperparameter optimization (HPO)?
- The exact range of parameters tested during HPO is not provided
- What kind of train/test/validation split was used for the HPO process? Was a separate test set held out during hyperparameter tuning to avoid an overfit?

Section 3.4: This section would benefit from a re-write. Currently, it introduces how k-fold cross validation is typically performed only to then say you used a different procedure, then adds that you actually train 20 models with $1/10^{th}$ sized test sets. Perhaps open by describing precisely what you did and follow with the motivations for some of your decisions (dealing with temporal autocorrelation and increasing the number of samples). Also, k-folds is a common ML technique. You could just cite a reference here to avoid some of the clutter of talking about typical k-folds strategies that weren't used in this paper, e.g. Bishop, C. M., 2006, "Pattern Recognition and Machine Learning", Ch. 1.

The authors justify use of the 2-convolution layer model by saying it will help prevent overfitting, but no analysis is done to back up this argument. Is overfitting actually a problem?

Maybe I missed it, but I don't think the "simple models" shown in Figure 7 are described anywhere. Also, the legend calls them "metrics" is it supposed to say "models"?

The use of the term "images" to refer to reanalysis data may be a bit confusing, particularly for those not familiar with reanalysis but perhaps familiar with CNNs which are often applied to photos. I think it is fine to use but it should be addressed early in the manuscript that "images" here does not refer to photographs, optical data, or data from any type of imager for that matter. Instead, these are chunks of regularly gridded data extracted from a numerical simulation.

There is some discussion of ensuring that input data has been interpolated to the correct grid resolution for consistent inputs to the CNN, but is there any concern about the non-uniform grid spacing on a lat lon grid? Samples closer to the poles will cover a smaller area.

Line 10: specify spatial interpolation

Line 85: Only 5 variables are listed. You should mention that you are using winds and temps at two levels which gets you to 8 input variables.

Line 85: I think these input variables are very reasonable choices, but no motivation was given for why these specific ones were used.

Section 3.2.6: I assume we are using bi-linear interpolation here?

Line 241: It would be helpful to move this note about standardizing the input variables into the "data" section. Or Section 3.2 "image preparation"

Figures 4, 5, 10, and 11 are missing units. If they are dimensional you could add colorbars, if they are all standardized you could mention it in the caption.

Technical comments:

I have noted several in the technical comments, but this paper contains quite a few grammatical mistakes and misused words and could use thorough editing.

Line 7: "similar locations and times to the TC containing images"

Line 8: "accuracy, but"

Line 9: "activity, but"

Line 10: this sentence sounds incorrect

Line 19: "causes"

Line 24: "Indeed,"

Line 46: "output from the Nonhydrostatic"

Line 93: why “usually”?

Line 94: “relevant” is an odd word choice, maybe “appropriate”?

Line 98: “are able to capture better” --> “better capture”

Line 101: strike “So”

Line 107: add a comma after “indeed”

Line 148: “a third” --> “one third”

Line 276: Figure number is missing

Line 287: change “experienced”

Line 287: strike “well”

Line 241: “image” --> “images”