EGUsphere, referee comment RC2
https://doi.org/10.5194/egusphere-2022-147-RC2, 2022

Comment on egusphere-2022-147

Anonymous Referee #2

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-RC2, 2022

This paper applies a CNN—based on Liu *et al.* (2016) —to classify tropical cyclones (TCs) based on two reanalysis datasets (ERA5 and MERRA2). Although the application of CNN in image classification is nothing new, the way to generate training datasets presented in this study is interesting and noteworthy. Another noteworthy aspect of this paper is that the authors compared the performance of their TC classification models across datasets (*e.g.*, ERA5 vs. MERRA2) and horizontal resolutions (native vs. interpolated).

This paper does not have any major flaws and presents another example of the useful application of machine learning in the field of geoscience; however, I had difficulties appreciating this work due to the vaguely written motivation and the lack of comparison of the performance of their CNN algorithm against other state-of-art algorithms. For these reasons (more details in my comments below), I do not think the current form of manuscript is ready to be published yet as it needs revision. I look forward to reading the revised manuscript.

General comments

[1] Vague motivation:

The motivation of this study is not clear. I think the main purpose of this study is not simply showcasing the performance of CNN in TC detection—but examining the sensitivity of CNN-based TC detection algorithm to input and training datasets (as written in L49-50 and L53-54) and/or presenting a new way to preparing the training datasets for TC

detections (as shown in figures 2 and 3). However, this point is not effectively highlighted in the introduction; in fact, it is not mentioned at all. My suggestion is to restructure the introduction to make clear the purpose of this study and also state the main purpose of this study clearly in the abstract.

Besides, I have the following comments for Section 1:

- (L27-29; L347-389) The authors claim this work contributes towards "automatic detection of TC in climate simulations without the need to retrain the CNN for each new climate model or climate model resolution." However, this point needs to be further justified. First, the cross examination (*e*., training on dataset A and evaluating on dataset B) would give more consistent results for two same-generation reanalyses than for two different climate models. Reanalysis datasets are at least grounded by observations; but two different climate models can wildly differ (*e.g.*, dynamical core, subgrid schemes, and coupling between components). Second, the current study examined only 0.25° and 0.5° resolutions which is much finer than the majority of CMIP-class climate models (~1°). The authors test interpolation from a finer to a coarse resolution, and accordingly, some high-resolution information is still carried over to interpolated low-resolution fields. However, the native resolution of common climate models is already too coarse (~1°) and their output variables calculated on a coarse grid are likely to miss finer dynamics/physics. In this regard, the author's CNN algorithm may not work well for climate model applications.
- (L30-40) The physical algorithms require preset thresholds. It provides intrinsic weakness because the performance of a given algorithm would depend on those thresholds. However, CNN classification (and, generally, any ML approaches) also suffer from similar issues. The performance of CNN is sensitive to the choice of hyperparameters. Unlike the thresholds used in physical algorithms (which can be physically interpreted), we do not know how each of these hyperparameters affects the performance of neural networks. As the thresholds in a physical algorithm is tuned, the hyperparamters of neural networks are tuned. In this view, the ML approach does not solve the problems related to thresholds in a physical mechanism. Besides, the authors point out that physical algorithms are *usually* applied in a limited domain. I wonder if it is due to the incapability of such algorithms to be applied in a wider domain—or if it is simply that people have not tried to apply it for a wider domain. If the latter is the case, the ML approach does not really address the issue of limited domains.
- (L41-54) Several ML studies has been cited in this paragraph. However, the current manuscript does not clearly present the weakness (or strength) of these previous studies over this study. Since this study is extending the ML approach side of TC detection algorithms, the previous studies (in terms of their algorithms and performance) need to be introduced with more details such as which aspect of these studies are limitations and how this study improves on those points.

[2] Insufficient comparison of the CNN model performance:

Section 4 presents the performance of the author's ML detection algorithm. However, it is hard to tell if the authors' algorithm is good or not because the benchmark comparison with other state-of-art algorithms is not provided. Only benchmark comparison provided is "simple metrics" in Figure 7; even then, the details about "simple metrics" are not included in the manuscript. Regardless, *simple* metrics are not a fair benchmark for the current study. A better (and more useful) benchmark should be state-of-art TC detection algorithms—including both physical and ML algorithms—that has been presented in recent years (*e.g.*, some of previous studies listed in L31 and L42). In particular, Liu *et al*. (2016) which this study is based on, would provide an objective baseline. The suggested comparison will further enhance the merit of the authors' work.

[3] Lack of details on the methodology of statistical tests and metrics:

I do appreciate the authors' efforts to use more robust statistical metrics and tests. I believe the following metrics and tests are less common (unlike 'accuracy'): AUC, AUPRC, Shapiro-Wilks test, and Kruskal-Wallis test. Including brief mathematical descriptions for those metrics and tests—maybe in the appendix—would be a great service to readers (including myself) who are not familiar with them.

[4] Insufficient details of the authors' ML methodology:

I suggest adding more details about how the authors implemented in the *main* text.

- Move Tables A2 and A3 to the main text. These are key information for ML implementations. Since this paper evolves around ML, these bits of information should be included in the main body of the manuscript.
- Table 1: Add information about stride and padding for convolution and pooling layers. Readers can infer them from the values in the table, but it is better explicitly shown. Also, specify pooling was done using either 'max' or 'averaging'.
- Table A3: Add a column that shows the search space for hyperparameter search.
- The hyperparameter optimization done in this study does not include some of the important hyperparameters, *g*., the number of layers, the number of neurons per layer, and the number of filters (convolution). It is likely that these hyperparameters would have impact on the performance of the CNN model more strongly than those listed in table A3. Provide justification on why these hyperparameters were omitted during the hyperparameter search process.
- In Section 3.1 (or somewhere in Section 3), explicitly write how the input vectors were

normalized and/or transformed. This information is presented vaguely later in the manuscript (L241-243), but it should be written more clearly, considering input vector normalization is one of the critical factors that determines the performance of ML algorithms.

Specific comments / typos

- In abstract, clarify that this work is based on TCs in the North Atlantic only.
- L45: This later work ☐ This latter work
- L49: "lack sufficient details": elaborate "details"
- L75-76: Explain why tropical depressions are not included as TCs. Is the performance of the current CNN algorithm sensitive to the inclusion of tropical depressions as TCs?
- Section 2.2: The difference and similarity between ERA5 and MERRA2 need to be presented in more detail. The atmospheric models and assimilation schemes are surely different, but the observation dataset they use might have good overlaps. How different these two datasets are is important to interpret the results of cross examination. That is, if they are only slightly different, the robust results from cross examination could be rather trivial.
- L83-84: Specify if 0.25x0.25 and 0.5x0.6 is either native or interpolated resolutions. If it is the latter, specify which interpolation scheme was used to regrid.
- L130-131: Rewrite the sentence: "thought", "those", and "they" are unclear.
- L134: Provide a proper citation for UML2.
- L176: Specify which task was for the 135-min CPU time.
- L203: Specify N. Is it 10 (Table A3)?
- Figures 4, 5, and 10: include colorbars.
- L216: What was the main criticisms, offered by Provost *et al*. (1997) and Ling *et al.* (2003), to the accuracy metric?
- L226-240: Rewrite. The authors' choice (iterative cross-validation) was introduced in a confusing way (after hold-out and k-fold). Consider starting the paragraph with mentioning iterative cross-validation and then explaining its benefit over k-fold.
- L256: Provide the usefulness of Youden's index, as done for AUC and AUPRC in L217-219
- Figure 6: Specify the y-axis uses a logarithmic scale in the figure legend.
- Figure7: Define the x-axis tick labels in the figure legend, as done in Figure 9
- L288-290: The authors' conclusion seems premature. There are other possibilities, for example, the current CNN architecture might be somehow more suitable for ERA5 dataset.
- Figure 8: The convention of box plots is standard, but I suggest including a brief explanation about what whiskers, boxes, and diamond markers stand for in the figure legend.
- Figure 9: 'purple' looks more similar to red.
- L324: Figure number is missing.