Response to Reviewer 1

We thank Reviewer 1 for your thoughtful comments and valuable suggestions, with which we fully agree with. In this revision, we have made several major changes to address your concerns, as detailed in our point-by-point responses below. These revisions are also shown in the tracked-changes version of the manuscript submitted with this response.

This study presents a convolutional neural network (CNN) framework—TCNN v1.0—for retrieving key tropical cyclone (TC) intensity and structure metrics, such as maximum sustained wind speed (VMAX), minimum central pressure (PMIN), and radius of maximum wind (RMW), from gridded climate data. A major strength of this framework lies in its ability to infer realistic TC intensity characteristics from relatively coarse-resolution reanalysis (MERRA-2), addressing a long-standing challenge in global climate models where TC structures are typically under-resolved. The authors argue that this approach has the potential to improve TC intensity estimation from both current numerical weather predictions and future climate model outputs.

Your evaluation and positive comments on our work are appreciated. We wish to take this opportunity to re-emphasize two key points that we have not been able to fully convey in our previous version:

- Climate datasets like MERRA-2 contain some meaningful environmental information about TC intensity and structure that deep learning (DL) models can effectively learn, even at coarse spatial resolutions. This environmental information allows DL models to retrieve TC intensity and structure much better than the traditional vortex tracking methods that directly calculate the maximum wind speed or the minimum central pressure on the climate data grid.
- Despite this promising capability, we also highlight that the MERRA-2 dataset has inherent limitations in representing TC fine-scale features for DL models to learn as with any other climate reanalysis products. Due to the lack of these fine-scale processes, our DL model performance appears to reach an upper limit beyond which further improvements are unlikely, regardless of the DL model tuning or architectures. The results presented in this study may thus represent the maximum capacity of DL-based TC intensity/structure retrieval from coarse-resolution datasets that we wish to present.

These insights are important, as they have not been discussed in previous literature, particularly in the context of TC climate downscaling from gridded climate data. We are currently working on a follow-up study focused on the second bullet point above, as it has significant implications for future machine learning applications in TC research. In this study, we acknowledge that this second point is more like a hypothesis, as it is only partially supported in the present work and remains to be fully demonstrated. We hope these clarifications help place our study in a broader context and convey its potential implications more clearly.

The study includes a thorough analysis of model sensitivity to input variables, domain configuration, and especially data sampling strategies. The results underscore the importance of proper train-test data partitioning, as the model's performance degrades substantially when tested on unseen TCs using a chronological split. This finding is important and well-motivated. However, if the generalization issue is one of the study's key conclusions, the decision to report the model's primary performance metrics based on random sampling (where samples from the same TC may appear in both training and test sets) needs further justification. Specifically, while

the reported RMSE for VMAX prediction (7.11 kt) appears to outperform previous methods, this result may overestimate the model's actual predictive capability, as the RMSE increases to 19.2 kt under a more realistic chronological split.

We agree with your comments. Our previous abstract was indeed unclear and incomplete, as we meant to report both the best possible performance with our TCNN model in retrieving TC intensity that we can achieve with the random sampling and the other performance with sampling by year. It is not our intention to claim that our TCNN model is better than previous retrieval methods based on satellites, because our focus here is on retrieving TC intensity from gridded climate dataset, which differs from previous studies that focused on retrieval from satellite images or radar data. In this revision, we have revised the abstract, the result discussion extensively as well as our conclusions to avoid the misleading information as in our previous version.

Furthermore, the authors cite existing studies such as Chen et al. (2019), which also employ CNN-based approaches to retrieve TC intensity from satellite data. Since Chen et al. used a chronological split in their validation, a more direct and critical comparison would be appropriate, even if the architectures and input data sources differ, especially given the common goal of improving TC intensity retrieval.

Thank you for pointing this out, which is again related to the unclear discussion of our results as we responded to your comment just above. We wish to mention again that our main focus here is quite different from what presented in Chen et al. (2019) in the sense that we want to retrieve TC intensity/structure from coarse-resolution gridded climate data for climate downscaling purposes, while Chen et al. (2019) applied DL to satellite imagery for operational forecast. This distinction was not clearly highlighted in our previous work, but it is in fact a key part of this study because gridded climate data contains much different information about TC intensity from satellite imagery. In particular, gridded data like MERRA-2 does not contain full TC structure that can be matched with an observed TC intensity. Thus, there is a limit on how much we can retrieve TC intensity from gridded climate data. Chen et al. (2019) on the other hand presented the TC intensity retrieval from a different perspective, with satellite images as an input. Thus, their results are more applicable to real-time forecast, while our results are more applicable to climate research such as downscaling future projection. We hope this revision could make this point clearer.

These issues also call into question the core assumption of the study—that ambient environmental conditions at 0.5° resolution contain sufficient information to estimate TC intensity. If the model struggles to generalize to new TCs, this may suggest that it is learning TC-specific patterns rather than robust physical relationships. As this assumption is foundational to the study's broader claims, especially regarding the potential application to future climate projections, further justification or clarification is needed.

This is in fact one of the two key points we wish to emphasize. For reference, we now include in this revision (Figures 2, 3, and 4) the direct calculation of TC intensity/structure from MERRA-2 grid data based on vortex tracking methods commonly used in previous studies for downscaling TC intensity climatology. When compared to observed TC intensity, it is evident that these

directly-computed intensities significantly underestimate actual TC intensity, particularly for storms reaching Category 1 or higher.

In contrast, our TCNN model, which assumes that environmental conditions contain useful signals for intensity estimation, demonstrates markedly improved performance in retrieving TC intensity and structure even with chronological sampling (as shown in Figures 2, 3, and 4). While our model operates at a relatively coarse resolution of 0.5°, its ability to extract meaningful environmental signals represents a significant scientific finding. This result supports the premise that accurate representation of storm-scale environmental features is critical for predicting TC intensity changes. We have accordingly revised Figures 2, 3, and 4 and added new Figure 5 to better illustrate the relative performance of intensity retrieval from our DL model versus traditional vortex tracking methods, which underscores the role of environmental information in improving TC intensity prediction as we want to present in this study.

The sensitivity test on domain size (Section 3.2.1) is informative, and the conclusion that a 25°×25° input domain yields the best performance is reasonable. Still, more discussion linking the domain size results with those from model architecture and convolutional kernel experiments would strengthen the study. This would also help clarify how spatial context is encoded and used by the CNN. Similarly, the reported seasonal variation in TCNN performance deserves more physical interpretation, particularly regarding how environmental influences on TC intensity may vary by season.

For the domain size sensitivity, we note that the domain of 25x25 degree would correspond to an area of radius 1600 km around a TC center. This domain is sufficiently large to include all TC basic structure including the far-field outflow and related subsidence, which can account for TC-environment interaction and explain for the better performance of our TCNN model as obtained in this study. However, this domain size comes with an issue that the sample size is now significantly reduced after we pre-process the training data, thus making it less robust. We have included this discussion in this revision to provide readers with more information.

For the seasonal variability, we note that TC seasons generally peak from May-October with an average of 3-4 storms per month, while the off-peak winter months has an annual average < 1 storm on average. With such a lack of TC statistics for the off-peak winter period, our DL model cannot provide a reliable result, consistent with the larger error bars seen in the previous Figure 10. Any systematic evaluations of environmental conditions during these winter months will suffer from the lack of statistical robustness. In this revision, to address your concerns, we have revised our previous Figure 9 (which is Figure 10 in this revision) with channel importance derived for 2 periods (January-April) and (May-November). These results could help answer which environmental factors plays a more important role during the off-peak and the peak TC seasons as you commented. We hope this new result could address your concern.

In summary, while this study presents an innovative and potentially valuable approach for estimating TC intensity and structure from gridded climate data, the current manuscript does not yet provide sufficient justification for its core claims. The reliance on a data sampling strategy that inflates performance metrics, coupled with limited generalization to unseen TCs, raises concerns about the framework's robustness and applicability, particularly for future climate

projections, which inherently involve unseen conditions. Furthermore, the key physical assumptions underlying the model are not adequately supported by the results, and the sensitivity analyses, while informative, could be more cohesively interpreted to strengthen the physical insights.

With the major revisions outlined above including i) providing a clearer physical interpretation of our results, ii) adding new results, iii) better clarification of the significance of our work, and iv) correcting several misleading discussions that you pointed out, we hope this revised version meets your expectations. We thank you again for your insightful comments and suggestions, which have greatly enhanced the quality of our manuscript.