Overall, this manuscript develops a deep-learning model (TCG-Net) for tropical cyclone genesis (TCG) prediction using hourly reanalysis data. The topic has potential practical value, and the paper reflects substantial effort in data preparation and model design. However, the main limitations are that the evaluation framework and baseline comparisons are not sufficiently comprehensive, making the benefits and trade-offs of the proposed approach relative to traditional methods unclear. In addition, the treatment of uncertainty in the "genesis time/location" definition from best-track data is insufficient and may be inconsistent with an hourly prediction setting. The manuscript also lacks case-level analyses and more direct XAI diagnostics to support physical consistency and conclusions about variable contributions. Given these issues, I recommend that the authors strengthen the work by adding more systematic evaluations of metrics, baseline comparisons, sensitivity to genesis definition, case studies, and explainability analyses. Detailed comments are provided below.

Major Comments

1. Evaluation method

The authors use three metrics—precision, recall, and F1—to evaluate their models. Precision and recall can be strongly affected by class imbalance, and F1, as a function of precision and recall, inherits similar limitations. The authors should include additional evaluation metrics, such as the ROC curve (and AUC) and the precision–recall (PR) curve (and PR-AUC), which assess performance across thresholds and are more informative under imbalanced settings. In addition, it would be useful to test the model with randomly sampled negative examples to examine whether TCG-Net remains robust and whether its performance depends on the specific negative-sampling strategy.

2. Comparison with Traditional Method

This study lacks a comparison with traditional approaches, so it remains unclear what the benefits and trade-offs of training TCG-Net actually are. For example, how would a Random Forest/LightGBM model perform? Alternatively, the authors should at least compare against a traditional index-based method, such as the classic Genesis Potential Index (GPI), which is not even suitable for 6-hour interval prediction.

3. The uncertainty of TC track data

In this study, the authors define the first reported position in the TC best-track dataset as the genesis location. It would be important to conduct a sensitivity test to assess how the model's performance changes if the genesis time/location is shifted slightly (e.g., by $\pm 1 \sim 3$ hours or ± 1 degree). Since this study applies hourly reanalysis dataset rather than monthly dataset to construct DL models, such analysis is necessary.

4. Lack of single case

Given that this study uses hourly reanalysis data and reports skill at a 6-hour lead time, it is also necessary to evaluate the model's performance for individual TC cases. How does TCG-Net generate and present the genesis probability for a specific TC (e.g., as a time series leading up to genesis, and/or a spatial probability map around the eventual genesis location), and how should this be compared against the best-track genesis time and location?

5. XAI method

In Section 4.3, the authors conduct sensitivity analyses by comparing different sets of input reanalysis variables. However, for modern deep-learning models it is straightforward to include gradient-based XAI diagnostics. For example, the authors could compute saliency maps (or related methods such as Integrated Gradients/Grad-CAM) to visualize which regions and which variables most influence the genesis probability, and then quantify the relative importance of each variable based on these attributions. This would provide a more direct and model-consistent assessment of variable importance than input-subset sensitivity tests alone.

Minor Comments

1. P4 L95: The rationale for selecting MERRA-2 needs to be strengthened. First, TC climatology reconstruction does not necessarily require very high resolution, since the ML models in this study are not tracking individual TCs. Do the authors analyze how fast TC moves when it is generated, and how uncertain the first location of TC best track represents TCG? Second, in most ML (especially DL) workflows, the original data are typically converted into standardized analysis-ready formats (e.g., zarr), so data format alone is not a compelling reason to prefer a particular reanalysis. Please provide additional, science-based justification—for example, whether MERRA-2 has demonstrated advantages over other reanalyses in representing key genesis-relevant environmental fields such as vertical wind shear and low- to mid-level humidity.