Response to Reviewer 1

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.

Authors' response: We thank the Reviewer for the positive feedback and constructive suggestions. In this revision, we have carefully addressed your comments, which are detailed in our point-by-point responses below. We hope that the revisions adequately resolve your concerns and further improve the clarity and contribution of our work.

1.1. 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.

Authors' response: We appreciate Reviewer 1 for highlighting the limitations of threshold-dependent metrics under class imbalance, for which we fully agree. Following your comments, we have extended the evaluation framework in this revision to include two new threshold-independent metrics, i.e., ROC–AUC and PR-AUC, in all of our analyses. As shown in our revised figures 3, 4, 5, 8, and 9, the results show that ROC–AUC remains consistently high across forecast lead times in both labeling methods (approximately 0.76–0.77 in the Past Domain and 0.79–0.87 in the Dynamic Domain method), indicating stable discriminative ability. Of note, PR–AUC, which is more sensitive to class imbalance, demonstrates consistent performance and confirms that improvements in our ResNet-18 design are not artifacts of a specific threshold choice. For the Past Domain method, PR–AUC increases progressively with lead time (from 0.06 at 6 hours to 0.27 at 48 hours), while it remains lower but consistent with the higher prediction difficulty for the Dynamic Domain method. These additional metrics thus reiterate that our ResNet-18 model maintains strong ranking capability and reliable positive-event detection under imbalance. We have revised the manuscript accordingly with all of these additional ROC and PR curves and analyses.

1.2. 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.

Authors' response: Your point is well taken. In response to your suggestion of evaluating robustness against the negative-sampling strategy, we have revised our analyses for which the test set negatives are generated using Random Under-Sampling (RUS) with the same sampling ratio as used in training, thereby emulating a randomly-sampled negative set and explicitly probing potential dependence on a particular negative-selection scheme (see revised Figures 8-9). Note that for these revised analyses, we use multiple RUS ratios (1:4, 1:10, 1:20, 1:30) and two loss-weighting configurations (balanced versus dynamic weights) for both the Past and Dynamic Domains, reporting precision, recall, F1, ROC–AUC, and PR–AUC across all forecast horizons (6h–48h). The results show that TCG-Net remains stable and consistently competitive under these randomly-sampled negatives: For the Past Domain, performance trends are preserved across lead times, and the best configurations (notably RUS 1:4 with

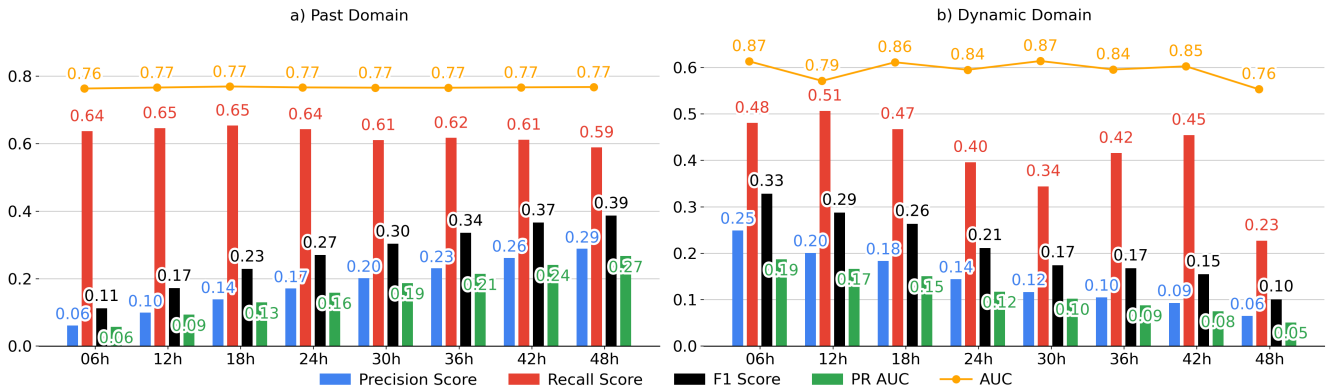

Figure 3: Overall performance of TCG-Net in terms of Precision (blue), Recall (red), F1 score (black), precision–recall area under the curve (PR-AUC, green) and area-under-the curve ROC (AUC-ROC, yellow) for the TCG prediction on a) Past Domain, and b) Dynamic Domain.

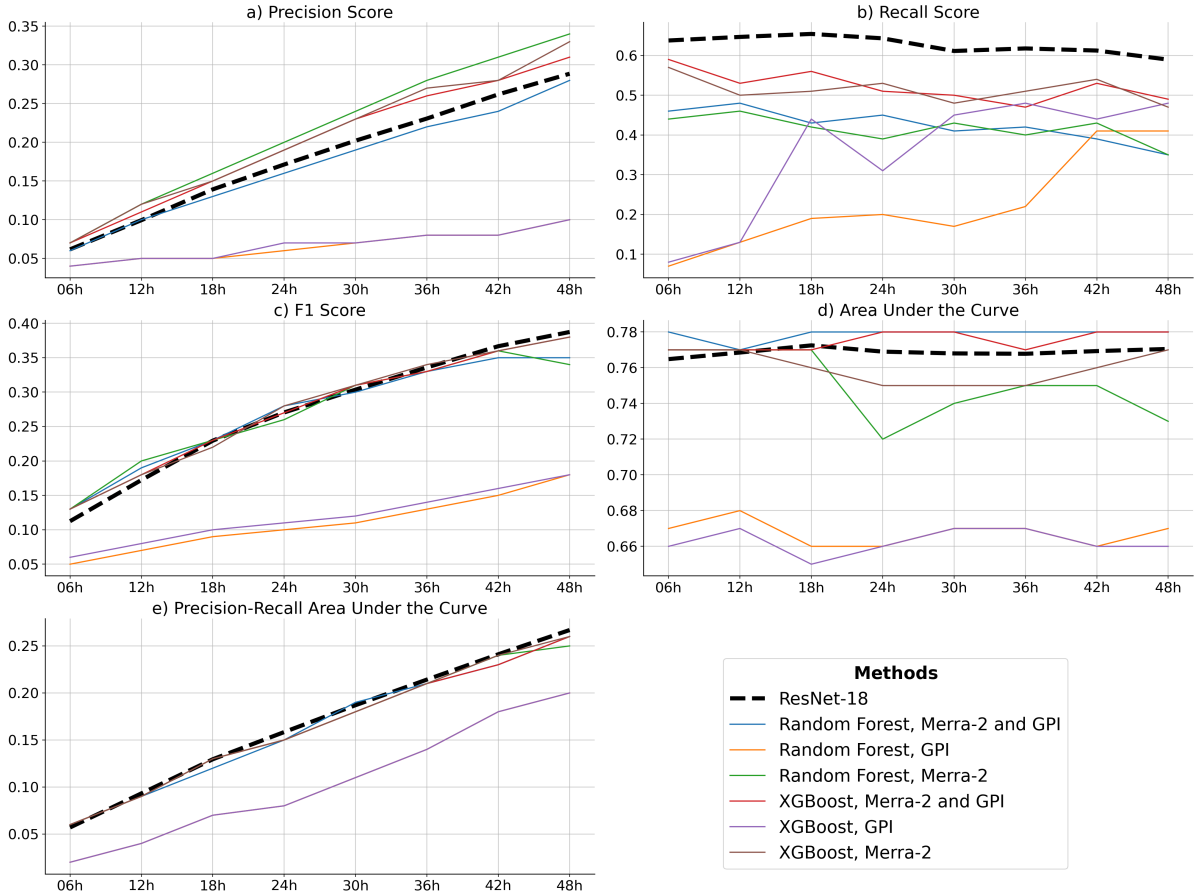

Figure 4: Comparison of the model performance for the **Past Domain** task between TCG-Net and traditional classification models, i.e., XGboost and random forest, which are enhanced by including the climatological Genesis Potential Index (GPI). The thick dashed line denotes the performance of TCG-Net with ResNet-18 backbone.

dynamic weighting) maintain strong precision/F1 together with high ROC–AUC (0.78) and PR–AUC (0.85). For the Dynamic Domain method for which the task is intrinsically harder and non-stationarity is stronger, the model performance degrades with increasing lead time for all configurations, yet the relative ranking of methods and the overall behavior remain consistent. This indicates that TCG-Net's skill is not an artifact of a specific negative set construction. These updated results are now included in the revised manuscript, strengthening the evidence that our proposed DL approach could generalize under alternative and randomized negative-sampling conditions.

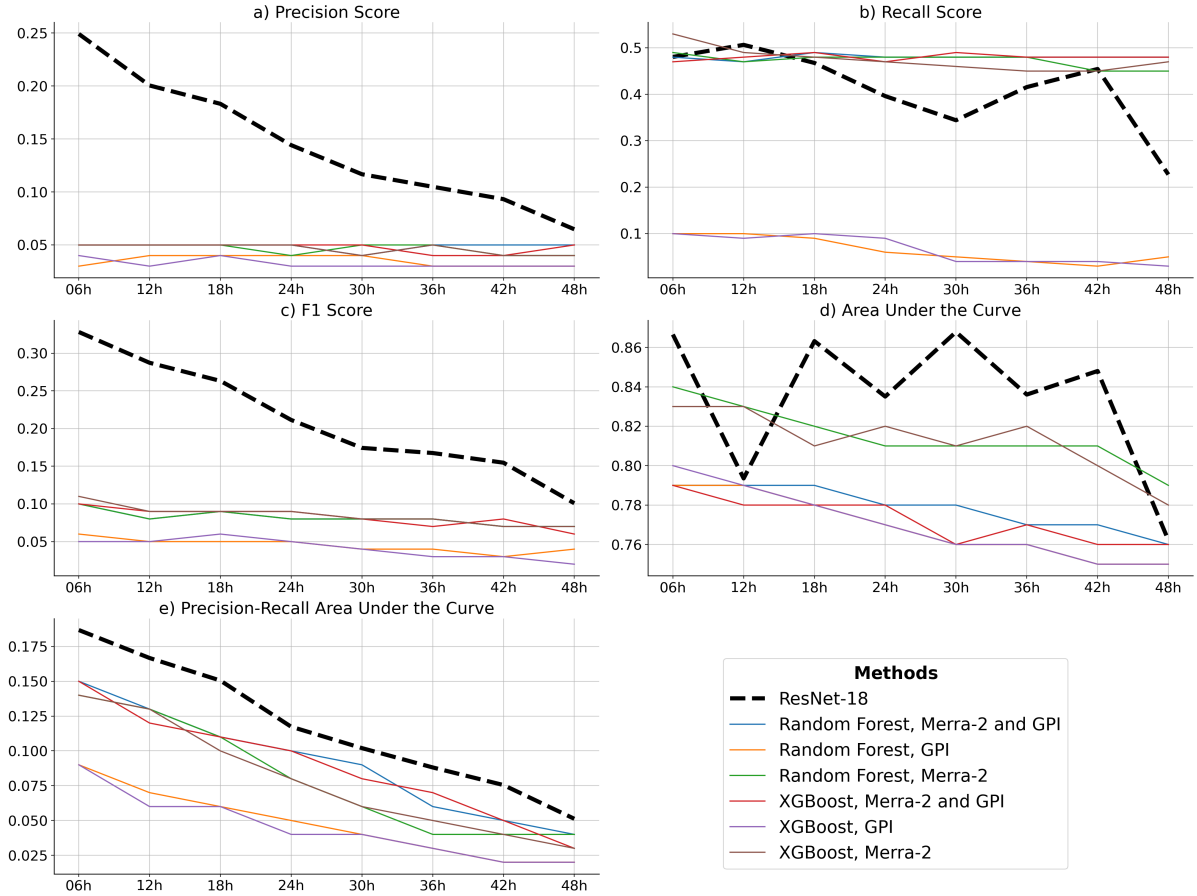

Figure 5: Comparison of the model performance for the **Dynamic Domain** task between TCG-Net and traditional classification models, i.e., XGboost and random forest, which are enhanced by including the climatological Genesis Potential Index (GPI). The thick dashed line denotes the performance of TCG-Net with ResNet-18 backbone.

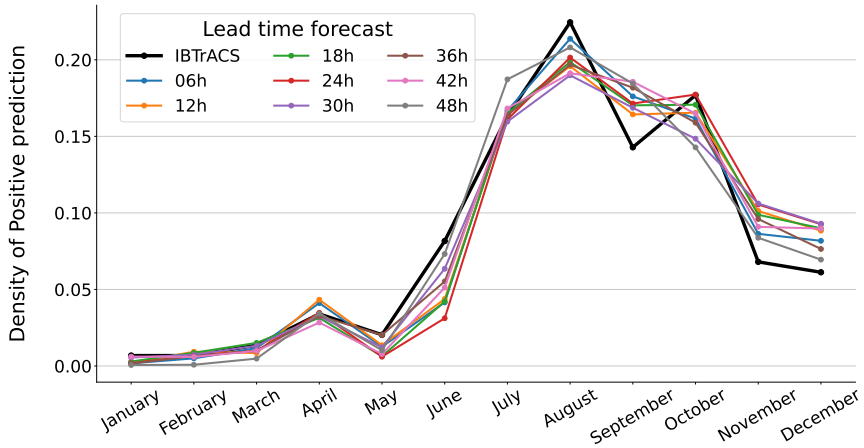

Figure 6: Monthly distribution of TCG frequency detected in the WNP basin from the test data (2017-2022), using the best-tuned ResNet-18 model for the DD strategy with data enrichment windows from 6 to 48 hours. The black solid curve denotes the TCG frequency obtained from the best track during the same time period.

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

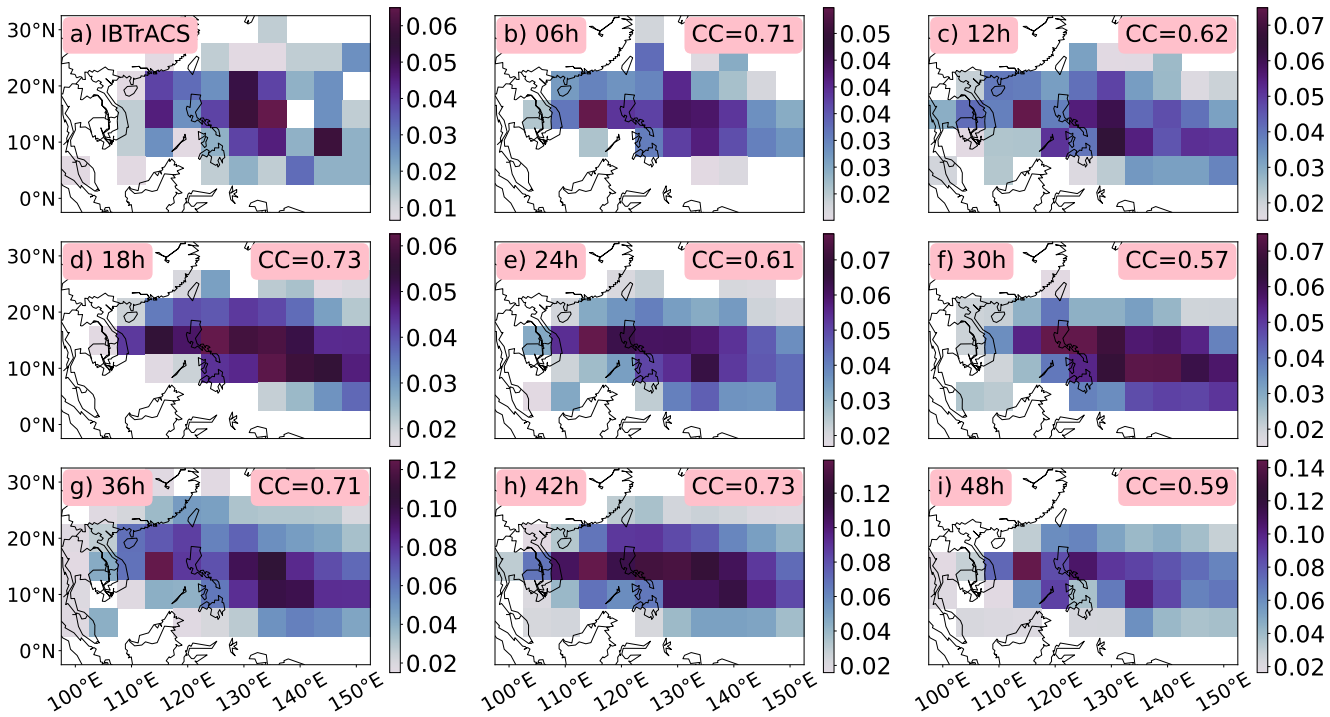

Figure 7: a) The spatial distribution of the observed TCG density (shaded) during the 2017-2022 period as obtained from the best track; (b)-(i) 5-year average of TCG probability prediction that is obtained from the ResNet-18 model with different data enrichment windows from 6-48 hours during the same test period as in (a). Note that different shading scales are used for different data enrichment windows so that one can better see the contrast between the areas of maximum probability for TCG predicted by the ResNet-18 model.

Genesis Potential Index (GPI), which is not even suitable for 6-hour interval prediction.

Authors' response: Thank you very much. Given your suggestion regarding benchmarking against traditional approaches, we have conducted in this revision additional experiments comparing TCG-Net with several widely used machine learning baselines including Random Forest and XGBoost, along with different combinations of MERRA-2 only, GPI only, and GPI as an additional feature of TCG-Net (revised Figures 4-5). Note that we also explicitly included the traditional GPI–based predictors to assess the value of this physically-derived index. The new results show that while tree-based models achieve reasonable recall at short lead times, their precision, F1 score, and especially PR–AUC degrade substantially with increasing forecast lead time, particularly for the Dynamic Domain method. In contrast, TCG-Net with ResNet-18 consistently maintains higher F1 and PR–AUC across all lead times (6–48 hours), indicating stronger robustness under class imbalance and temporal variability. Moreover, GPI-only models exhibit limited predictive skill, especially for 6-hour interval forecasting, confirming that purely GPI-based approaches are not sufficient for TCG prediction. Overall, the extended benchmarking demonstrates that the performance gains of TCG-Net are systematic rather than incidental, and that its training complexity yields measurable improvements over both classical machine learning and index-based methods.

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.

Authors' response: Your comment directly addresses a central challenge in TCG climatology reconstruction, for which we fully agree with. In fact, the timing and location of a TCG event derived from best-track datasets can vary substantially in both time and space in some cases, especially during the early INVEST stage when it is often

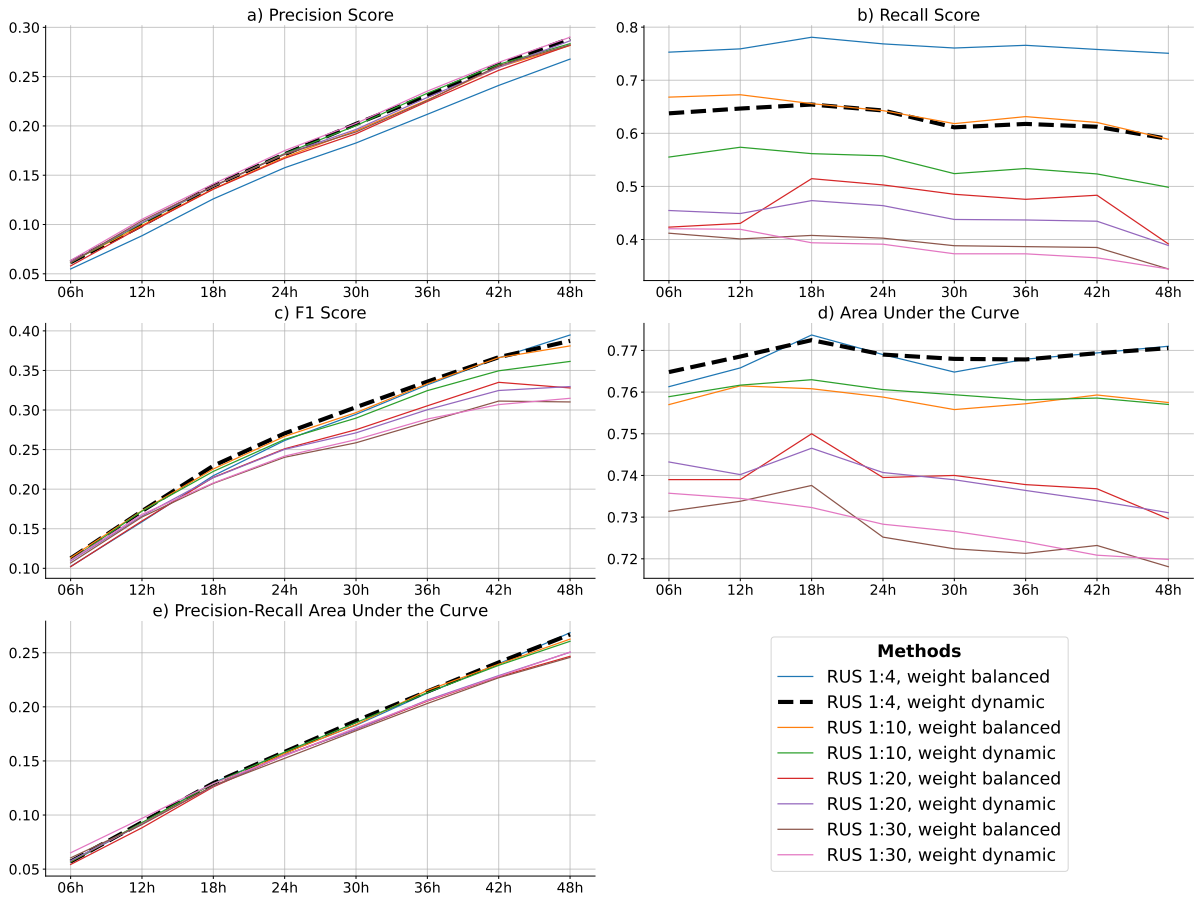

Figure 8: (a) The precision score P for the TCG prediction of TCG-Net using a range of the RUS ratio and class weight (solid colors) for the **Past Domain** task and the same sampling ratio for both the training and test sets; (b)-(c) similar to (a) but for the R and F1 scores, respectively. The dashed black line denotes the reference obtained from our best-tuned model. Note that *weight balanced* assigns fixed importance to each class based on frequency whilst *weight dynamics* adaptively adjusts sample or class importance.

difficult to identify the exact onset time of TCG. This uncertainty is exactly why we introduce a data-enrichment strategy that extends beyond the single time stamp obtained from the best-track record. Specifically, as described in the Methods and Results sections, we label positive TCG events using a retrospective temporal window of up to 48 hours prior to the best-track TCG timing in all of our analyses. In addition, spatial uncertainty in TCG location is explicitly considered by defining a relatively large positive-label domain ($\sim 18 \times 18$ degrees), thus making sure that the favorable environment for a TCG event is fully captured. These inherent temporal and spatial uncertainties are also the reasons why we use sliding-window approaches when reconstructing TCG climatology, as illustrated in Figures 6–7.

In response to your comment, we have expanded the discussion of these issues in the revised manuscript and hope this more clearly conveys the rationale behind our methodological choices and the interpretation of the resulting climatological patterns.

> *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.*

Authors' response: Agree. Gradient-based XAI methods are indeed very useful for analyzing the role of different variables in probability maps for specific case studies. In our previous work (Nguyen and Kieu, 2024), for example,

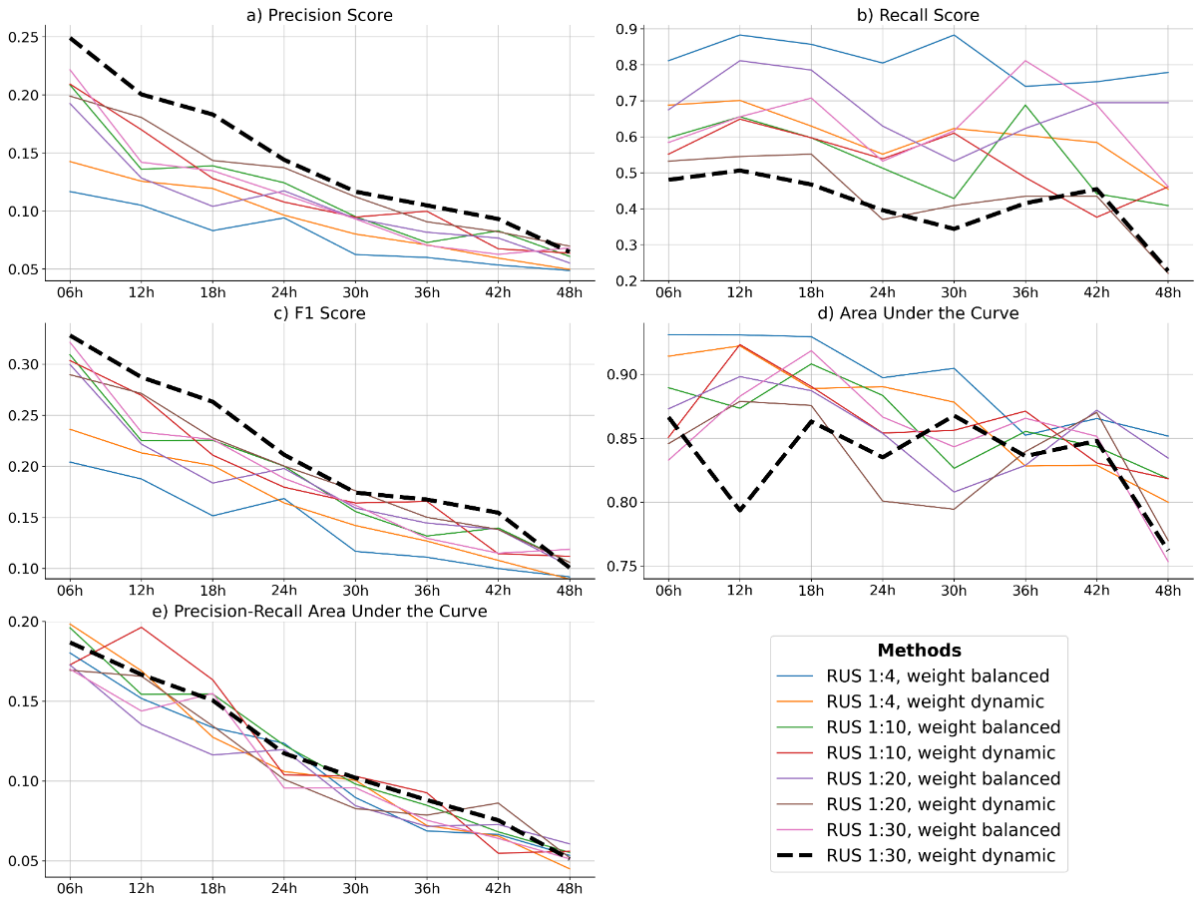

Figure 9: (a) The precision score P for the TCG prediction of TCG-Net using a range of the RUS ratio and class weight (solid colors) for the **Dynamic Domain** task and the same sampling ratio for both the training and test sets; (b)-(c) similar to (a) but for the R and F1 scores, respectively. The dashed black line denotes the reference obtained from our best-tuned model. Note that *weight balanced* assigns fixed importance to each class based on frequency whilst *weight dynamics* adaptively adjusts sample or class importance.

the integrated-gradients method was applied to identify where information contributing to a predicted TCG event was drawn from. In this study, our focus focuses however more on reconstructing TCG climatology rather than analyzing individual events. Applying a gradient-based XAI approach for each case during the reconstruction would therefore be overwhelming. Our main aim in this study is to design a DL model that can represent the climate mean robustly, rather than individual realizations as typically emphasized in weather-forecast applications. As such, we adopt in this study an XAI framework that emphasizes sensitivity analyses with respect to different input channels, model parameters, and seasonality shown in Figures 6, 12–15, and Table 4, which can be naturally interpreted from a climatological perspective. These sensitivity analyses provide insight into how variations in each input channel influence the reconstructed TCG climatology as a whole, and so they are a form of XAI by nature.

In response to your comment, we have now included additional statistics on the ranking of each input variable based on its weights, which is averaged from the first CNN block of the RestNet-18 model (Figures 10–12). This ranking offers further information on the relative contribution of each channel to the overall TCG climatology reconstruction as expected.

Minor Comments
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 reanal-

Table 4: List of features selected using the feature engineering and feature ranking filter approach as obtained for each labeling strategy during the training period.

ID	Name of Features	Past Domain		Dynamic Domain	
		Feature Engineering	Feature Ranking	Feature Engineering	Feature Ranking
1	QL		400, 700, 825, 900, 950		100, 1000. 150, 200, 300, 400, 500, 600, 700, 800, 875, 900, 950, 975
2	H	500	200, 925	500	100, 550, 950
3	QI		250, 450, 600, 800, 900, 925, 950, 1000		100, 1000, 150, 500, 600, 700, 900
4	OMEGA	500	450, 875	500	100, 150, 250, 600, 925, 1000
5	T	500, 900	725	500, 900	150, 200, 900
6	U	200, 800	825, 1000	200, 800	1000, 550, 200
7	V	200, 800	150, 550	200, 800	1000, 600, 400 150, 100
8	RH	750	950	750	100, 200, 400, 700, 825, 875, 925, 1000
9	QV				100, 150, 900
10	VOR	200, 700, 900		200, 700, 900	
11	DIV	200		200	

ysis. 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.

Authors' response: Thank you. We wish to take this opportunity to clarify that our use of MERRA-2 at 0.5-degree spatial resolution is primarily motivated by its similar resolution to most of the global climate projection products from current CMIP5/CMIP6. Training the deep learning (DL) model at this resolution, thus, facilitates subsequent fine-tuning with other climate datasets, which is our further aim after this study. In fact, our ongoing extension of this work that aims at reconstructing TCG during the pre-satellite era follows this exact strategy. That is, our DL model is first trained using MERRA-2 data and then further fine-tuned with ERA5. A similar approach could be adopted using other reanalysis datasets, such as JMA, CFS, or NCEP-FNL, all of which are available at 0.5-degree resolution. To the best of our knowledge, there is currently no comprehensive study comparing TCG climatology across these reanalysis products. As such, any of these datasets is equally valuable for DL model development. This consideration underlines our decision to present a DL model development using the MERRA-2 dataset in this study. These discussions have been included in this revision per your comment.

Regarding the Zarr format, we could see that its use is indeed becoming increasingly popular within the atmospheric community. At present, most existing climate datasets are distributed in NetCDF or GRIB1/2 formats. For this reason, we chose to output our preprocessed data in the same original NetCDF format to maintain a consistent and unified framework across all I/O stages (and to minimize use of external libraries). Incorporating additional data formats would require only minimal modifications to our workflow, as users can readily extend the preprocessing component to support alternative formats and generate a common intermediate data representation.

For the final note on TC tracking, we would like to mention that our main aim of this study is for TCG, which concerns only the very first location when a TC forms, instead of the entire TC track. Thus, we do not have any analyses for TC movement, which will require a new component for TC vortex tracking beyond the scope of this work.

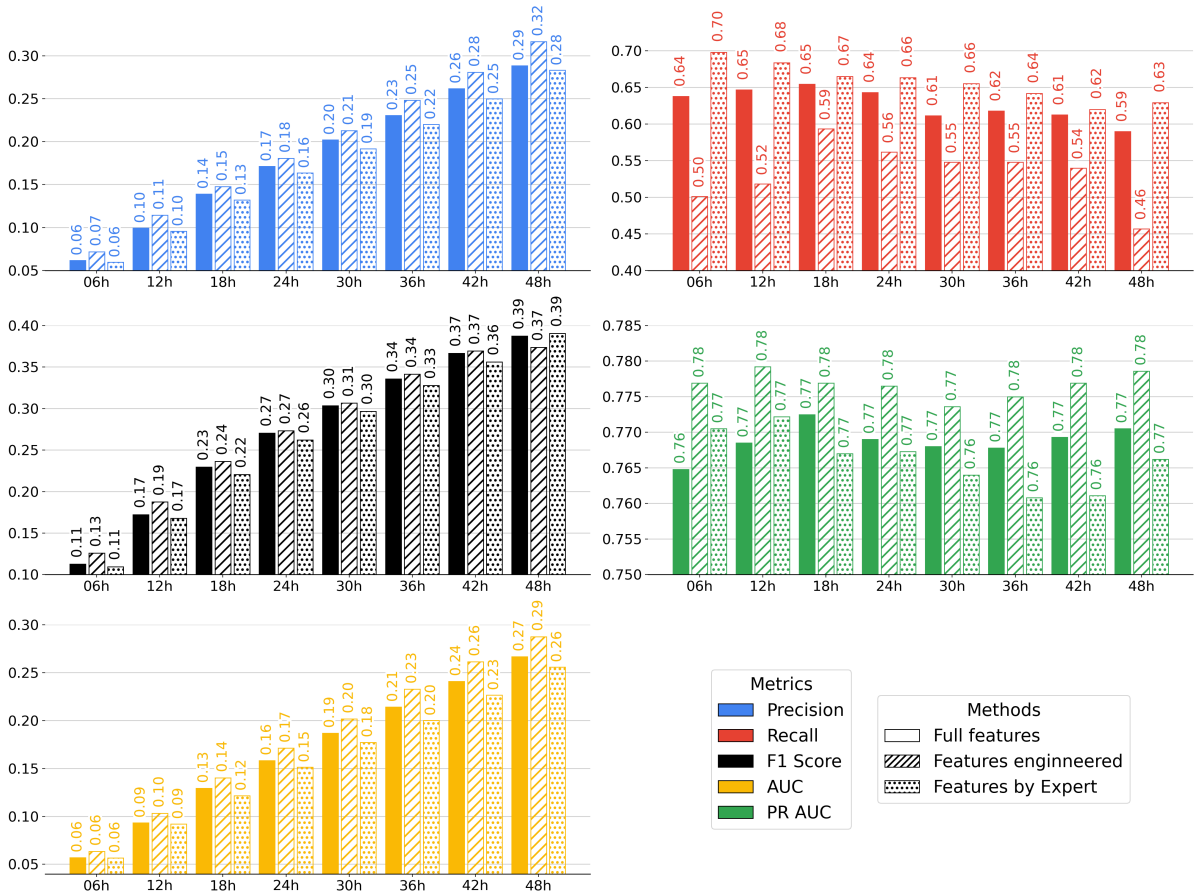

Figure 12: Evaluation of the model performance for several different feature selection methods including all features (solid color columns), 13 selected features based on feature engineering in (Nguyen and Kieu, 2024) (striped columns), and feature ranking of top 10% (dotted columns) using the **Past Domain** task.

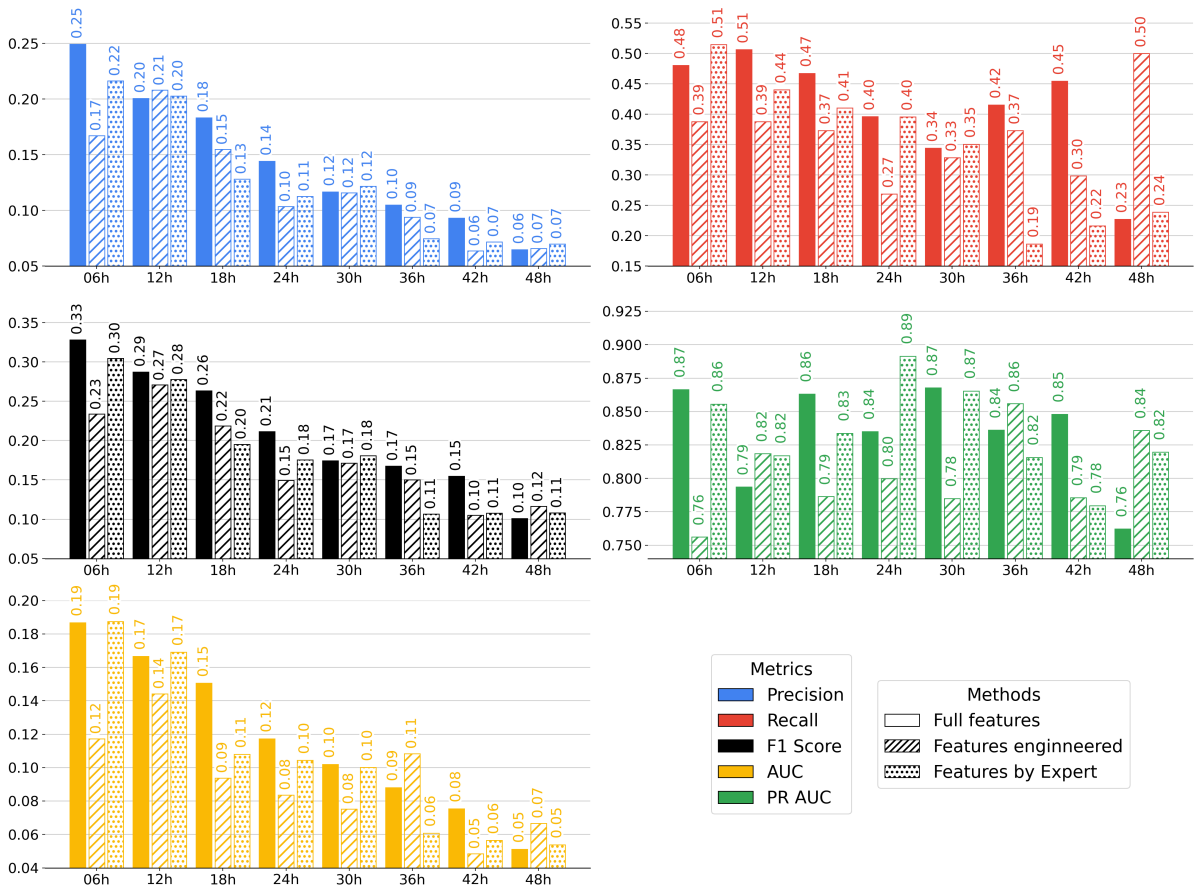

Figure 13: Evaluation of the model performance for several different feature selection methods including all features (solid color columns), 13 selected features based on feature engineering in (Nguyen and Kieu, 2024) (striped columns), and feature ranking of top 10% (dotted columns) using the **Dynamic Domain** task.

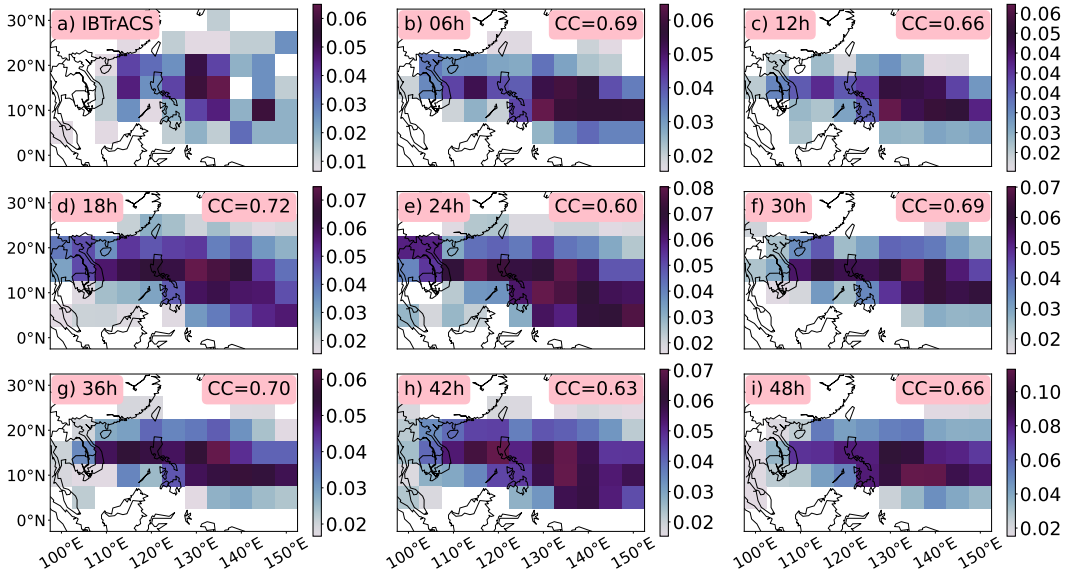

Figure 14: Similar to Fig. 7 but for the feature engineering approach.

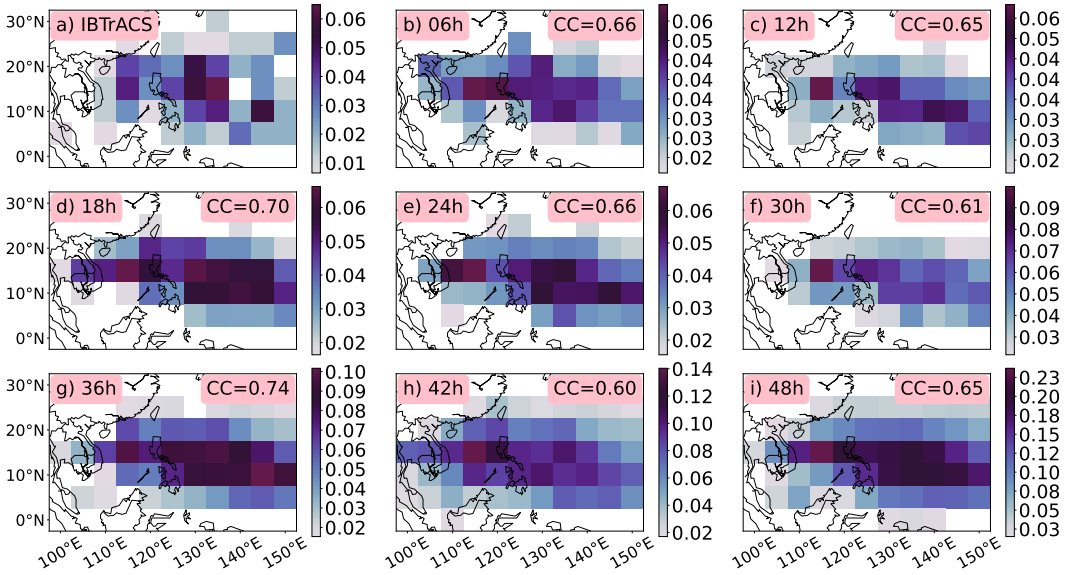

Figure 15: Similar to Fig. 7 but for the automatic feature ranking approach.