Response to Reviewer 2

This manuscript presents TCG-Net, a deep-learning-based framework for the reconstruction of the tropical cyclone genesis (TCG) distribution over the western North Pacific (WNP). The authors introduce two task-specific labelling strategies, combined with temporal feature enrichment and imbalance-aware training, to extract both seasonal and spatial characteristics of TCG directly from MERRA-2 reanalysis data. The effort put forth by the authors is commendable, but the manuscript requires significant revision. In particular, clearer clarification is needed on the applicability of TCG-Net, the definition and climatological representation of TCG, the claimed physical novelty. Before considering this paper for publication, I have several concerns and suggestions outlined below.

Authors' response: We are grateful to Reviewer 2 for your thorough evaluation and detailed suggestions. In this revision, we have carefully considered all of your comments and have made substantial changes to address the concerns you raised, as detailed in our point-by-point responses below. All revisions are indicated in the tracked-changes version for your convenience. We hope that the revised manuscript meets your expectations and will be acceptable to you.

1. TCG-Net is an application-oriented framework, yet its practical strengths and limitations relative to existing approaches are not sufficiently clarified. The authors claim that vortex tracking methods face challenges with coarse-resolution climate models ($> 0.5°$) and TCG-Net therefore serve as a complementary tool. In fact, several traditional detection algorithms has been developed for coarse-resolution datasets, such as OWZP and TRACK (Tory et al., 2013; Hodges et al., 2017), which exhibit reliable performance in reproducing the climatological distribution of TCG in both reanalysis datasets and climate models (Bell et al., 2019; Bourdin et al., 2022). A comparison between TCG-Net and traditional detection algorithms would be valuable.

Authors' response: Thank you for this comment. After conducting this type of comparison, we feel that such a comparison would be challenging to implement. The primary difficulty arises from differences in spatial resolution, aim, and study designs. Recall that our DL reconstruction is applied to a 0.5° resolution, whereas previous studies that you suggested (e.g., Tory et al.) derived TC climatology at coarser resolutions ($1° \times 1°$ or $1.5° \times 1.5°$). Applying their methods directly to our dataset would require a complete re-tuning of the vortex-tracking parameters in the OWZP framework to accommodate the higher resolution. Such re-tuning would introduce an additional layer of uncertainty and make a direct comparison between their established algorithm and our DL-based approach subjective and potentially less meaningful. Similarly, the study by Hodges et al. focused on detecting TCs from reanalysis datasets rather than constructing TCG climatology. As a result, their vortex-tracking methodology is not directly comparable to the TCG climatology framework used in our study.

Beyond the few you suggested, we have been exploring other alternative vortex-tracking methods, such as TempestExtremes (Ullrich and Zarzycki 2017), but have not yet obtained stable results due to some technical issues. Figure 1 in this response shows an example of the TCG distribution when we apply vortex detection directly on MERRA-2 data that we obtain from TempestExtreme for a test period from 2017-2022. The result turns out to be very sensitive to several parameters such as TC lifetime, vortex strength, or the surface wind threshold, which all affect the first moment that a TC is recorded for the TCG reconstruction.

At present, we still do not have a systematic way to adjust these parameters for every test period or data enrichment window for this analysis. For these reasons, we have decided to restrict our analysis to a direct comparison between the DL-reconstructed TCG climatology and that derived from best-track data to minimize introducing more subjective analyses. We hope Reviewer 2 will agree with us on this approach.

2. TCG-Net was not developed based on objective TC structural characteristics, but rather on large-scale environmental factors in present-day climates, similar to GPIs. As a result, the applicability of TCG-Net to future climate projections is uncertain, as the relative importance of critical environmental factors may change under warming conditions (Murakami and Wang, 2022).

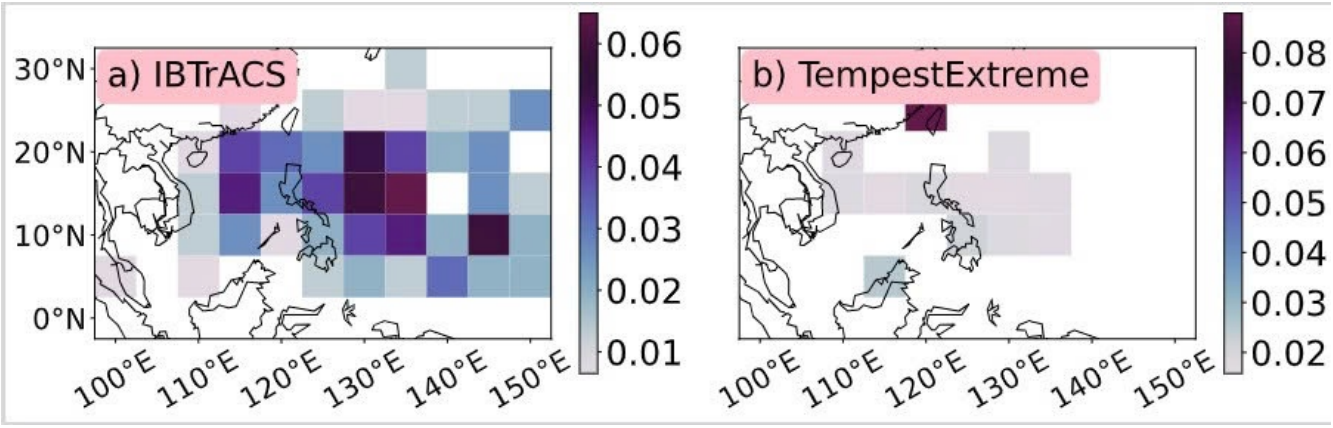

Figure 1: Spatial distribution of TCG density as obtained from a) best-track, and b) tempestExtreme vortex tracker for a test period 2017-2022, using the set of vortex tracking thresholds for the NCEP 0.5-degree resolution.

Authors' response: This point is indeed central to our study. We would like to emphasize that the traditional approach to TCG climatology based on indices such as the Genesis Potential Index (GPI) and other well-established variables such as SST, CAPE, and vertical wind shear relies largely on empirical relationships or approximate theoretical frameworks. These commonly-cited environmental conditions for TCG have been derived from previous observations, field campaigns, and modeling studies. While they are statistically robust and physically meaningful, they are not universally sufficient to guarantee TCG across all basins. Each basin has distinct regional characteristics that may enhance, modulate, or suppress particular processes.

In this study, we introduce a DL framework that systematically evaluates a broad range of environmental variables and identifies the dominant controlling factors within a unified modeling architecture. As presented in Section 3, our results confirm several key variables that are consistent with existing physical understanding, while also revealing additional factors that appear to be specific to the WNP basin. These newly identified factors provide further insight into TCG behavior in the WNP and may offer added value for assessing potential future changes in TCG under a changing climate. Importantly, our approach is systematic and readily extendable to other basins, datasets, or climate regimes to identify the most relevant controlling mechanisms in each context.

In response to your comment, we have incorporated new analyses in which GPI is (1) used as a proxy for TCG and (2) included as an additional input channel to the DL model. The results (Figs. 4–5) show that GPI alone does not perform as effectively as DL reconstruction. On the other hand, including GPI as an independent input channel contributes little to overall model performance, likely because the environmental variables used to compute GPI are already included among our existing input channels. In this revision, we have updated Section 3 to provide additional discussion regarding the potential role and contribution of GPI, as you suggested.

> *3. The authors use the TCG distribution derived from 2017–2022 to represent "TCG climatology," which may not be appropriate. Given the strong interannual variability of TCG, a five-year period is generally insufficient to characterize climatological conditions.*

Authors' response: Your point is well taken. We acknowledge that this is a key limitation of our study, as a 5-year period cannot be considered fully representative of TCG climatology. The primary constraint preventing the use of a longer evaluation period is the availability of the MERRA-2 dataset (or any reanalysis datasets in general), which are most reliable from 1980 onward. With roughly 40 years of data in total, allocating 5 years for testing already represents a substantial fraction of the data record, leaving about 35 years for training. Extending the testing period further would reduce the amount of training data, potentially limiting the model's ability to generalize across different climate modes. Our strategy therefore prioritizes maximizing the training dataset so that the model can robustly learn the environmental conditions conducive to TCG. While alternative approaches such as multiple 5-year leave-one-out experiments are possible, they still do not fully resolve this issue because each test climatology would still be constrained to a limited 5-year window.

We have added a discussion of this limitation in the revised manuscript. This challenge is indeed inherent to any DL-based approaches applied to relatively short climate records, and we believe it is important to clearly inform

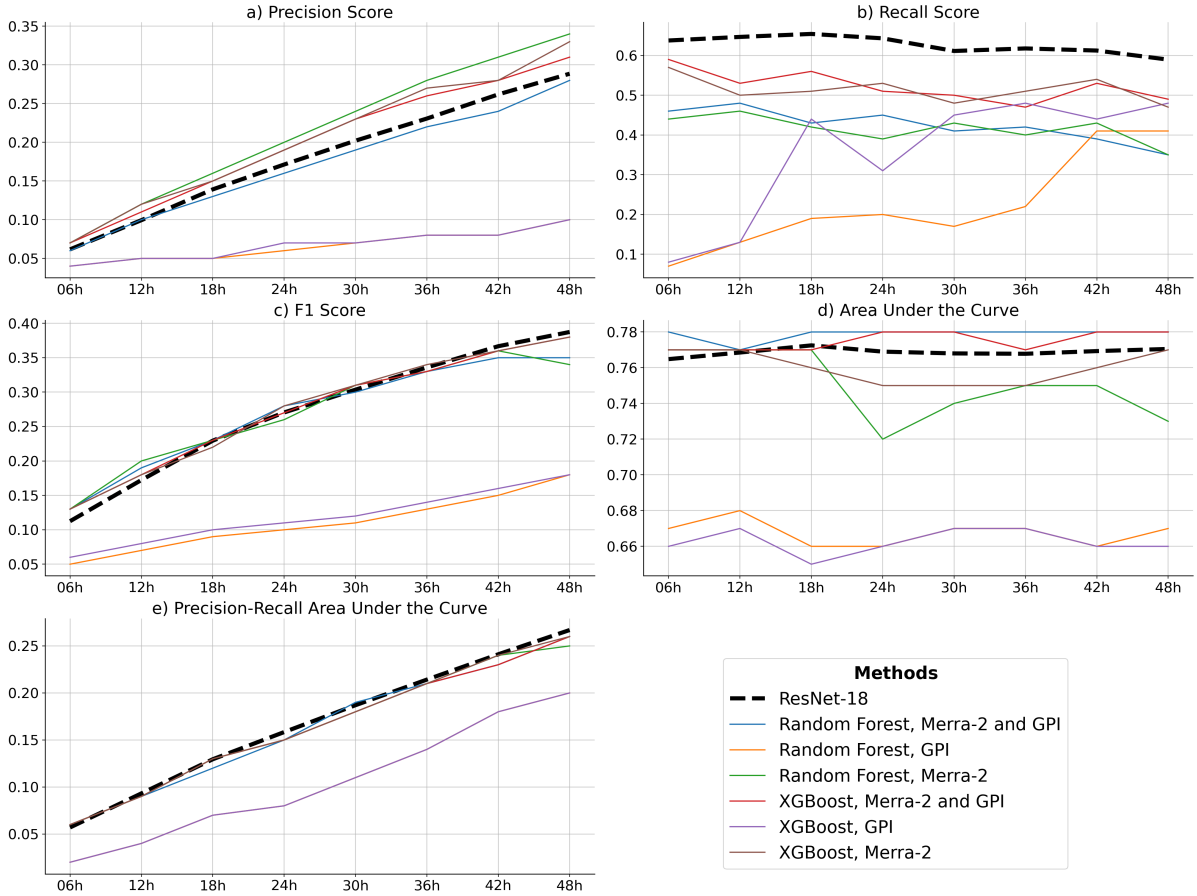

Figure 4: Comparison of the model performance for the **Past Domain** task between TCG-Net and traditional classification models, i.e., XGboost and random forest, which are enhanced by including the climatological Genesis Potential Index (GPI). The thick dashed line denotes the performance of TCG-Net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

readers of this potential limitation when interpreting the results as you commented.

> *4. In this study, TCG was defined as the first time that a TC was recorded in observations, which can be reasonable in a weather prediction context because it allows earlier detection of cyclogenesis. However, in climatological studies, TCG is more commonly defined when the storm intensity first reaches 35 kt in order to exclude weak or short-lived vortices (Klotzbach et al., 2022; Lai and Toumi, 2023). How sensitive the performance of TCG-Net may be to the definition of TCG?*

Authors' response: Yes, you are correct that the timing of a TCG event derived from best-track datasets can vary substantially, especially during the early INVEST stage or weak systems for which it is often 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 provided by the best-track record in all of our analyses (i.e., the time window from 6-48 hrs in Figs 2-15). 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 record. This way, the inherent temporal uncertainties can be captured by our use of multiple data-enrichment windows when reconstructing TCG climatology. As illustrated in, e.g., Figures 6–7, the performance of our TCG-Net is quite robust for the enrichment up to 36 hr before the genesis timing recorded in the best track. Taking further data beyond 36-hr will reduce the performance of our TCG-net model. In response to your comment, we have expanded the discussion of these issues in the revised manuscript and hope this could better address the uncertainty issue that you raised.

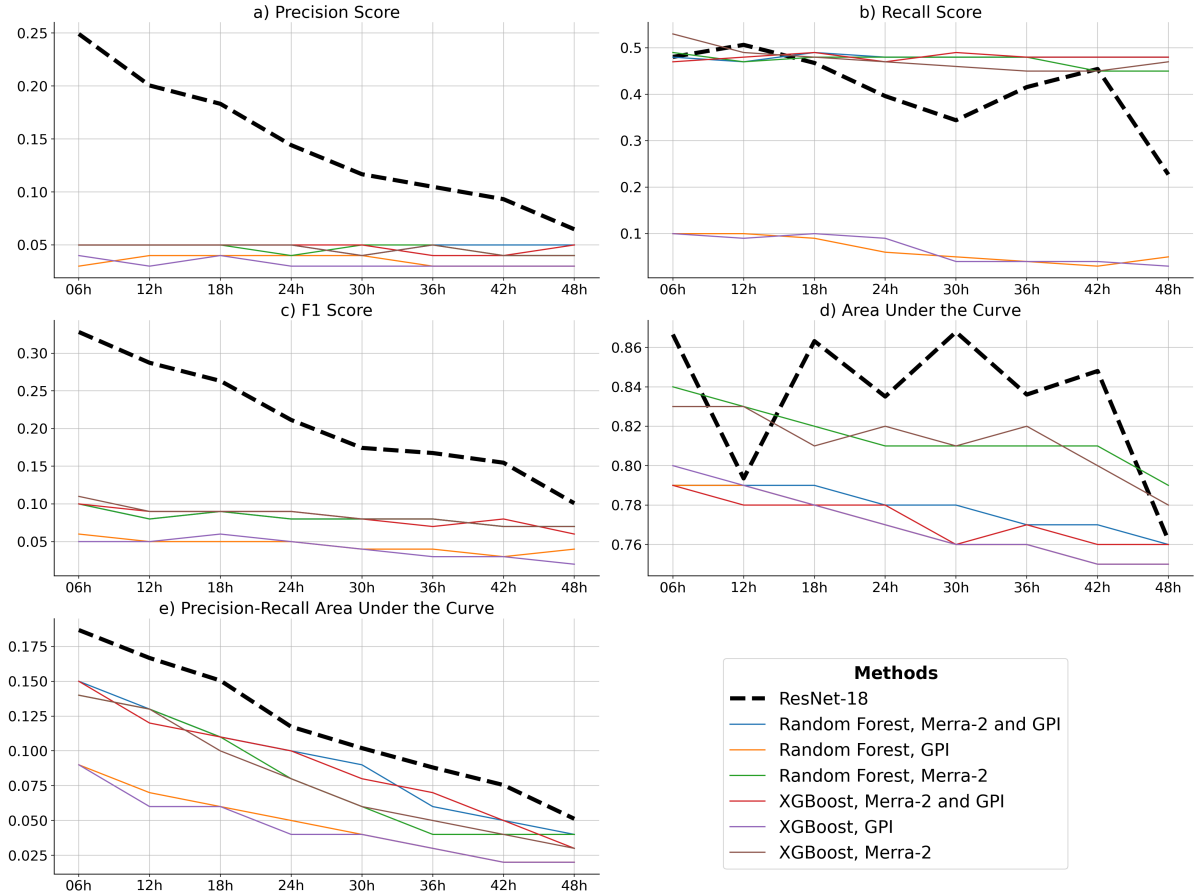

Figure 5: Comparison of the model performance for the **Dynamic Domain** task between TCG-Net and traditional classification models, i.e., XGboost and random forest, which are enhanced by including the climatological Genesis Potential Index (GPI). The thick dashed line denotes the performance of TCG-Net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

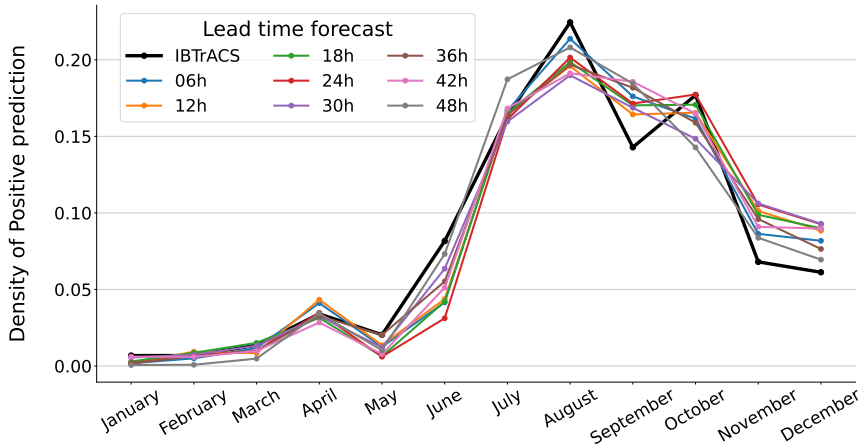

Figure 6: Monthly distribution of TCG frequency detected in the WNP basin from the test data (2017-2022), using the best-tuned ResNet-18 model for the DD strategy with data enrichment windows from 6 to 48 hours. The black solid curve denotes the TCG frequency obtained from the best track during the same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

5. While the authors emphasizes novelty in terms of large-scale environmental drivers of TCG, the selected large-scale factors based on feature ranking are largely consistent with previous studies (Emanuel, 2010; Wang and Murakami, 2020). In this regard, the results appear to largely reproduce established findings rather than provide genuinely new physical insights, and the claimed level of innovation in this aspect may be overstated.

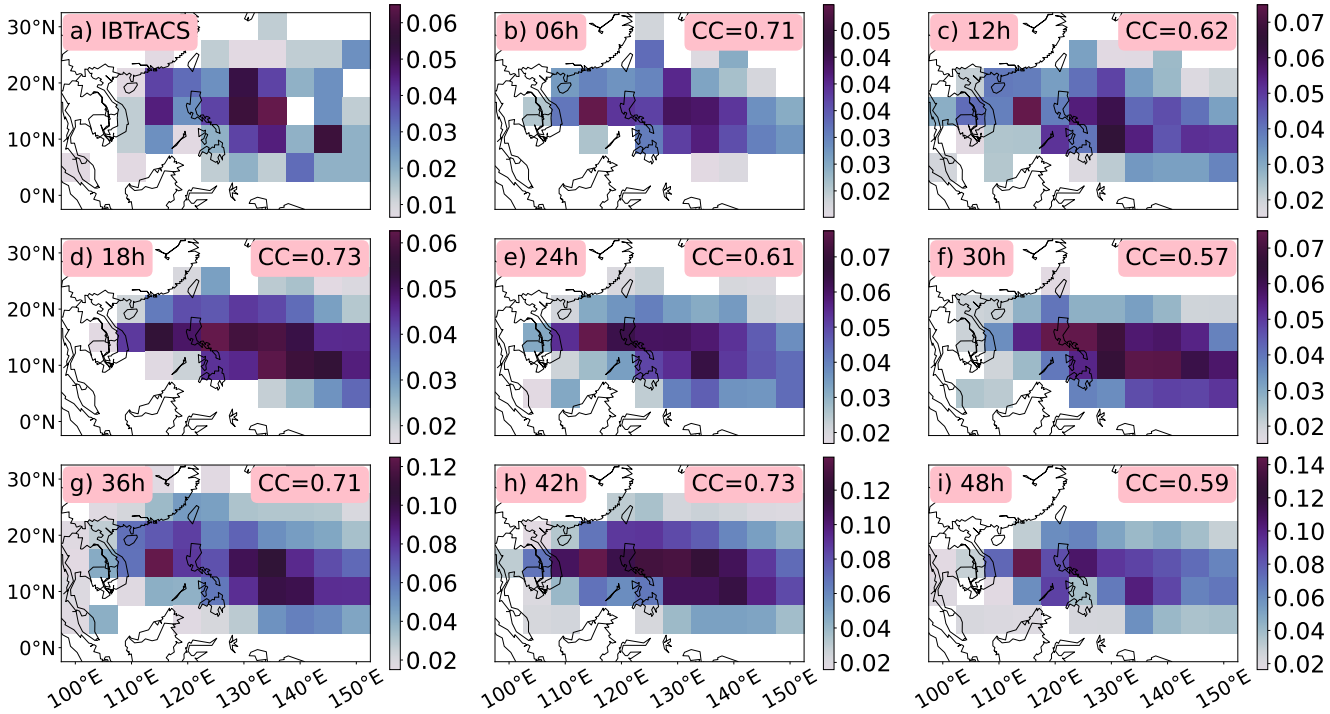

Figure 7: a) The spatial distribution of the observed TCG density (shaded) during the 2017-2022 period as obtained from the best track; (b)-(i) 5-year average of TCG probability prediction that is obtained from the ResNet-18 model with different data enrichment windows from 6-48 hours during the same 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.

Authors' response: This comment is somewhat related to your earlier comment (#2) that we responded above. One of the major points of our DL approach is its ability to identify and evaluate the role of various factors, which are further highlighted in the new analyses of factor ranking in this revision (Figs 12-13). As shown in these results, we not only confirm several key factors for TCG as identified in previous studies, but also uncovering several new factors relevant to our TCG reconstruction such as the cloud water content (QI/QL) or lower troposphere temperature, depending on the sampling strategies (see also Table 4). The fact that our DL could recover the previous factors gives us confidence that these new factors are indeed significant for TCG in the WNP basin. Note that these additional factors are basin dependent, and so there may emerge other factors in different basins. Regardless of the basin, our DL approach can help uncover within a consistent framework that we wish to present in this study. In this revision, we have further revised the feature importance section to better highlight the significance of our approach. We hope that this addition will offer readers deeper insights into the TCG process in the WNP basin, extending beyond the climatology-based reconstruction in previous studies.

6. I was a little confused about domain chosen in this study. While a positive TCG label was defined as the square box of size $18^\circ \times 18^\circ$ *centered on the first recorded TCG location, the ResNet-18 model was applied on each* 5×5 *box. It is therefore unclear how the labeling domain and the prediction domain were reconciled during training and evaluation.*

Authors' response: Thank you for your comment. Our previous discussion was unclear here. To clarify, the ResNet domain is $18^\circ \times 18^\circ$ degrees, but we shift the domain center by 5 degrees each time when sliding the domain so we can cover the entire WNP basin. The TCG probability obtained from each domain is then assigned to the center of that domain, allowing us to construct a map with a resolution of 5 degrees. We have revised this description to ensure that our approach is clearly presented, as per your suggestion.

Minor comments
1. L95: Another advantage of choosing MERRA-2 is that TC-related information has been assimilated into MERRA-2, whereas ERA5 does not include this (Gelaro et al., 2017).

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	

2. L114: IBTrACS database compiles global TC tracks information from multiple agencies. Which agency's observations were utilized in this study?

3. L179: Change "DM" to "DD"

4. L362-363: The discrepancy between the DL model results and observations may not solely reflect limitations in DL model optimization, but could also arise from deficiencies in the ability of large-scale environmental factors to reproduce the seasonal variability of TCG (Menkes et al., 2012; Tippett et al., 2011).

5. L375-381: The pattern correlation coefficients with the observations are encouraged to quantify the performance of the DL model.

6. L432: The pressure levels described here is inconsistent with the Table 4

Authors' response:

1. Thank you for pointing this out the benefit of MERRA-2 that we were not aware of. The work by Gelaro et al. (2017) has now been included in this revision.

2. We use the JTWC agency in the IBTrACS data, which is now mentioned explicitly in this revision.

3. This paragraph has been revised.

4. Totally agree. This session has been revised accordingly.

5. Pattern correlations for all maps have been now included in this revision.

6. This inconsistency has now been corrected.

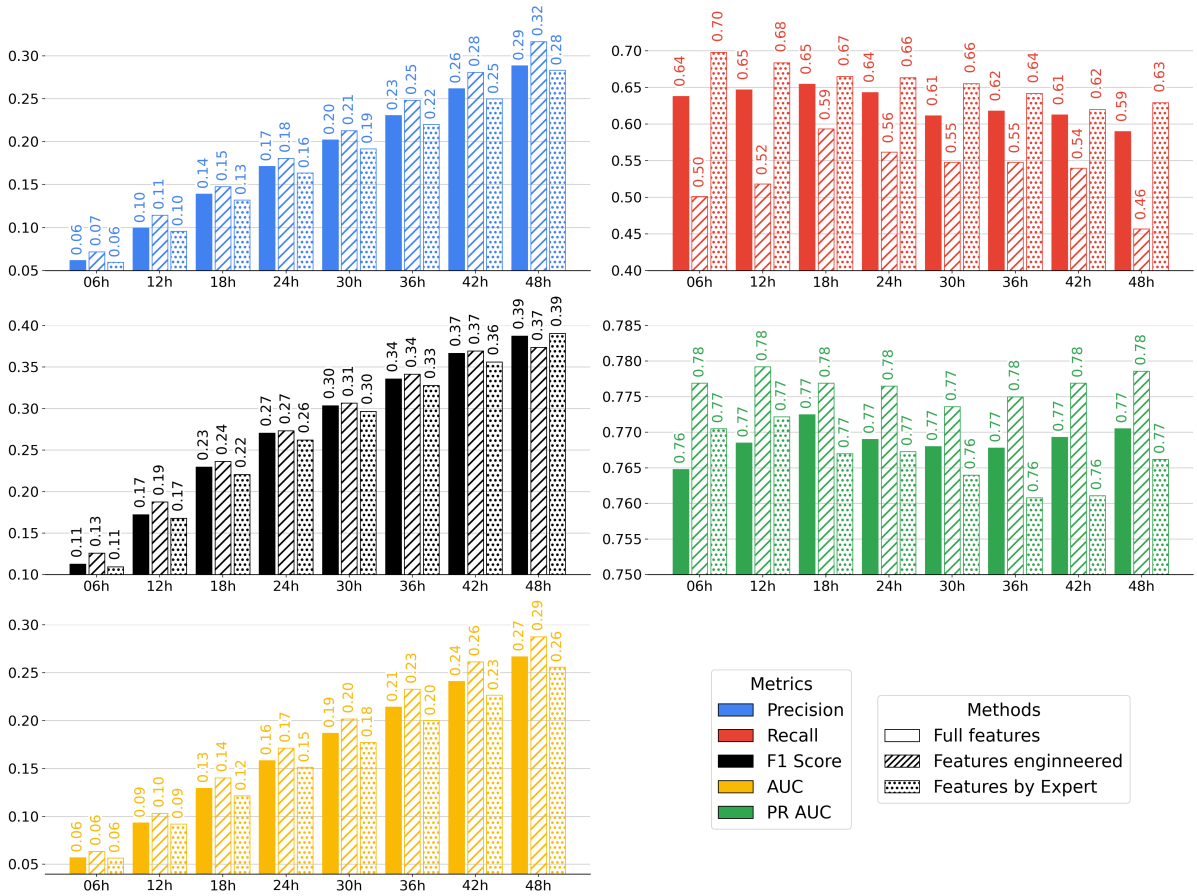

Figure 12: Evaluation of the model performance for several different feature selection methods including all features (solid color columns), 13 selected features based on feature engineering in [**?**] (striped columns), and feature ranking of top 10% (dotted columns) using the **Past Domain** task.

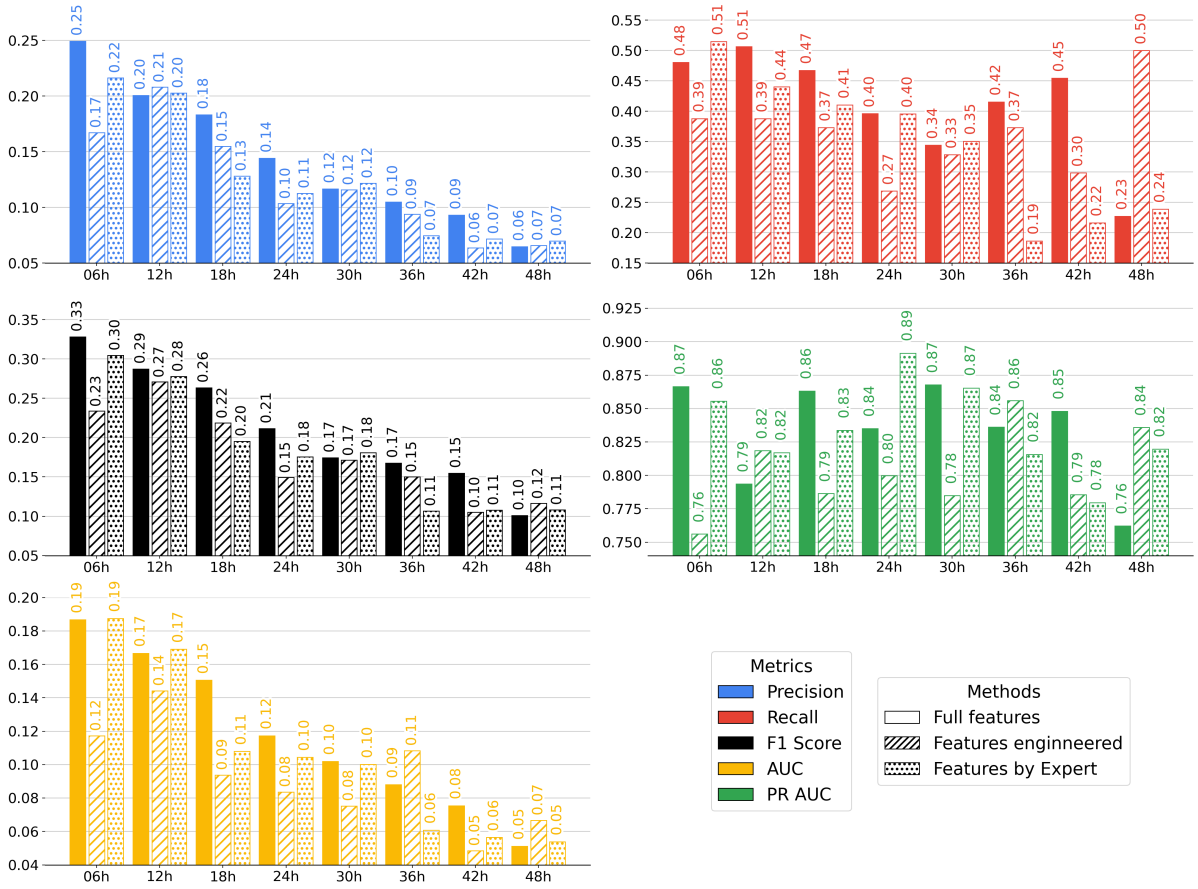

Figure 13: Evaluation of the model performance for several different feature selection methods including all features (solid color columns), 13 selected features based on feature engineering in [**?**] (striped columns), and feature ranking of top 10% (dotted columns) using the **Dynamic Domain** task.