“Ensemble Random Forest for Tropical Cyclone Tracking” by

P. Vaittinada Ayar et al.

We first would like to thank the anonymous reviewer for her/his thorough reading and very positive and constructive comments. We tried to take them into account as much as possible. A detailed point-by-point reply to these comments is provided below. Changes in the manuscript are indicated in **blue**.

Answer to Referee #1

Overview

This work applies Random Forest (RF) models to track tropical cyclones using environmental variables from a global reanalysis (ERA5) with an eventual goal of using the RF tracker in long-running climate simulations. The Eastern Pacific and Northern Atlantic TC basics were chosen for investigation. Random Forests were trained by categorising localised boxed regions in each basin as either containing a TC or not (TC-free) and associating statistics of environmental variables in each box from ERA5 to the binary events. Variables of mean sea level pressure, relative vorticity, column water vapour, and thickness were used as they represented different facets of physical mechanisms and TCs. Statistics are computed for these variables and included as inputs during RF training.

Training is conducted with 6-fold cross-validation to generate a range of RF solutions that are then used to compute MCC, POD, and FAR over a series of subsampling experiments – the authors note a significant proportion of their samples are TC-free compared to TC samples. Generally, a ratio of 25-1 is seen as reasonable with POD and FAR tradeoffs as the ratio is increased/decreased. Detection skill is notably better than the baseline UZ method in both basins. Further investigation of skill suggests the model primarily misses TCs at low intensity and low duration. The authors also devise analyses to interpret physical meaning, although I have some comments on this aspect of the analysis below.

Overall, the authors have employed RFs in a very unique and potentially innovative application area to track TCs in global reanalyses. The manuscript could benefit from improved grammar and clarity in locations, along with consideration of additional analyses or methods to improve the scientific presentation. I look forward to seeing a revised manuscript after careful revision.

Comments

✳*Comment– Lines 119-125 : If one of the objectives of the manuscript is to determine the physical relationships that govern TC tracks, why not allow the ML model to do the variable selection for you ? Provide X number of variables and employ variable selection procedures like sequential forward selection or sequential backwards selection ? Or use explainable AI techniques to do variable selection ? If we are informing which variables the RF should learn from, aren’t we prescribing our own biases into the physical mechanisms ?*

RESPONSE– In our study, we provided 20 variables (five physical variables and the four associated statistics, see lines 139-145), that are preselected based on physical considerations mentioned in lines 126-132 and are based on past literature. Regardless of the number of predictors, we are not informing RF which predictors it should learn from. RF is expected to make the difference between the informative variables and those that are not for the considered problem, namely, the TC probability estimation. Therefore, even if we were to prescribe our own biases, RF would not consider variables that are not relevant. In addition, one of the objectives is to use modest computing power, which led us to choose random forest. A use of explainable AI for variable selection would have naturally extended to the tracking stage. The objective of this study is to provide some physical interpretation of the occurrence of a TC while having a frugal model (*i.e.,* few features). Last but not least, the number of variables is limited to those that are the most commonly available in climate models. This led us to use 20 variables, and we did not see the need to further reduce that number. Though we tested the sensitivity of the tracking by removing the least important variables, and showed in discussions that these are also important in controlling the false alarms.

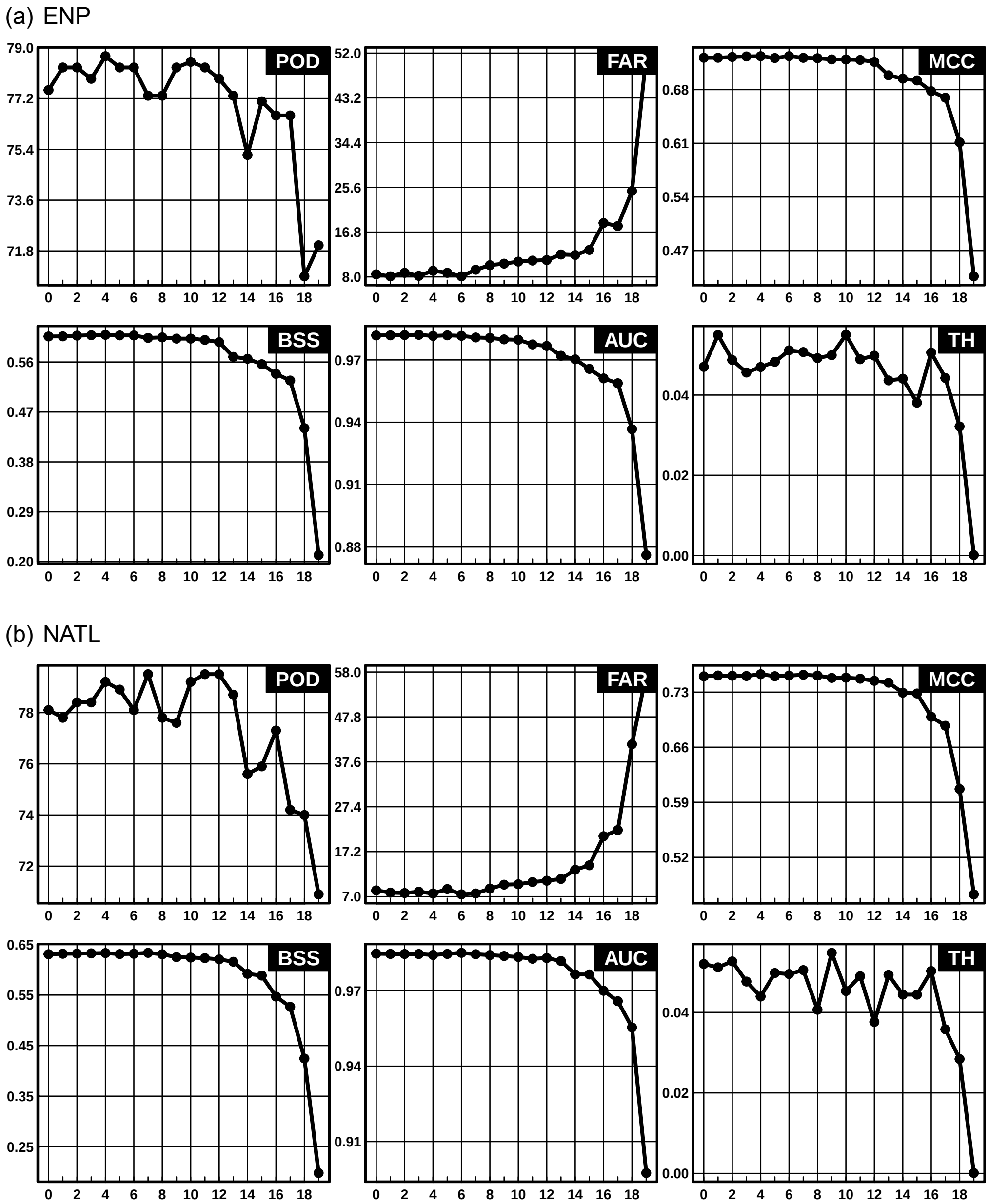

Figure R1 – Evolution with respect to the number eliminated of features in abscissa (0 to 19) of POD, FAR, MCC, brier skill score BSS, area under the ROC curve (AUC) and the optimal threshold obtained from the ROC curve (TH) for a) ENP and b) NATL basins.

To further answer the reviewers' comment, we performed a recursive feature elimination (RFE) on the 6-fold cross-validation set-up (validation experiment) for **one RF**. It consists of eliminating the least important variable at each iteration. Figure R1 represents, with respect to the number of features eliminated in the abscissa (0 to 19), the evolution of POD, FAR, MCC, Brier skill score BSS, area under the ROC curve (AUC) and the optimal threshold obtained from the ROC curve (TH) for both basins. We see that the performance of RF remains quite stable until we remove 12 to 14 variables, which roughly corresponds to the number of variables we kept for the sensitivity test. The exception resides in FAR, which shows a 50% increase when we remove more than 6 variables, confirming the role of the least important in controlling the false alarms. This has been mentioned in lines 413-418 of the revised manuscript as :

CHANGES– To explore how many variables can be removed without deteriorating the performance of the tracker, a recursive feature elimination on the 6-fold cross-validation set-up (validation experiment) has been performed. It consists of eliminating the least important variable at each iteration. The results (not shown here) show that the performance of ERF remains stable until we remove 12 to 14 variables, which roughly corresponds to the number of variables we kept for the sensitivity test. The exception resides in FAR, which shows a 50% increase when we remove no more than 6 variables, confirming the role of the least important in controlling the false alarms.

✳*Comment– Line 160 : Hyperparameter testing should be done for all RFs developed in this work and should be explicitly provided to readers for reproducibility. It is uncommon for RFs to be trained and validated without hyperparameter testing and tuning – I have to see any published works that used the default values given in whichever software package was being used. By testing and tuning the hyperparameters, the authors guarantee that the RFs are not "good by luck" and there is a repeatable process for future experimentation. Further, all hyperparameters should be listed in the manuscript/in a table.*

RESPONSE– We realised a hyper-parameter tuning by performing a grid-search on three key parameters : the number of trees, - NTREE = 100, 250, 500 (default) -, the random number of features (predictors) considered to perform the best split - split try = 1, 2, 4 (default), 6, 10 - and the minimal size of end nodes - End node size = 1, 5, 10 (default), 20. The parameters' tuning is performed on the 6-fold cross-validation set-up (validation experiment) for **one RF**. Figure R2 represents the POD, FAR, MCC, BSS, AUC and for both basins and the 60 combination of parameters. We can see that the impact of the hyperparameters appears quite minimal. There is not one configuration that is the best for every index. The choice of the default parameters seems a reasonable choice. This question has been clarified in lines 156-160 of the revised manuscript as :

CHANGES– A grid-search is performed on the three key parameters : (i) the number of trees, (ii) the random number of features considered to perform the best split to grow the trees and (iii) the minimal size of end nodes (not shown). Results showed that the impact of the hyperparameters is quite minimal, and no configuration of the hyperparameters yielded significantly better results. Therefore, the hyperparameters were set to the default values : 500 trees, 4 randomly chosen features to perform the best split and a minimal end node size of 10.

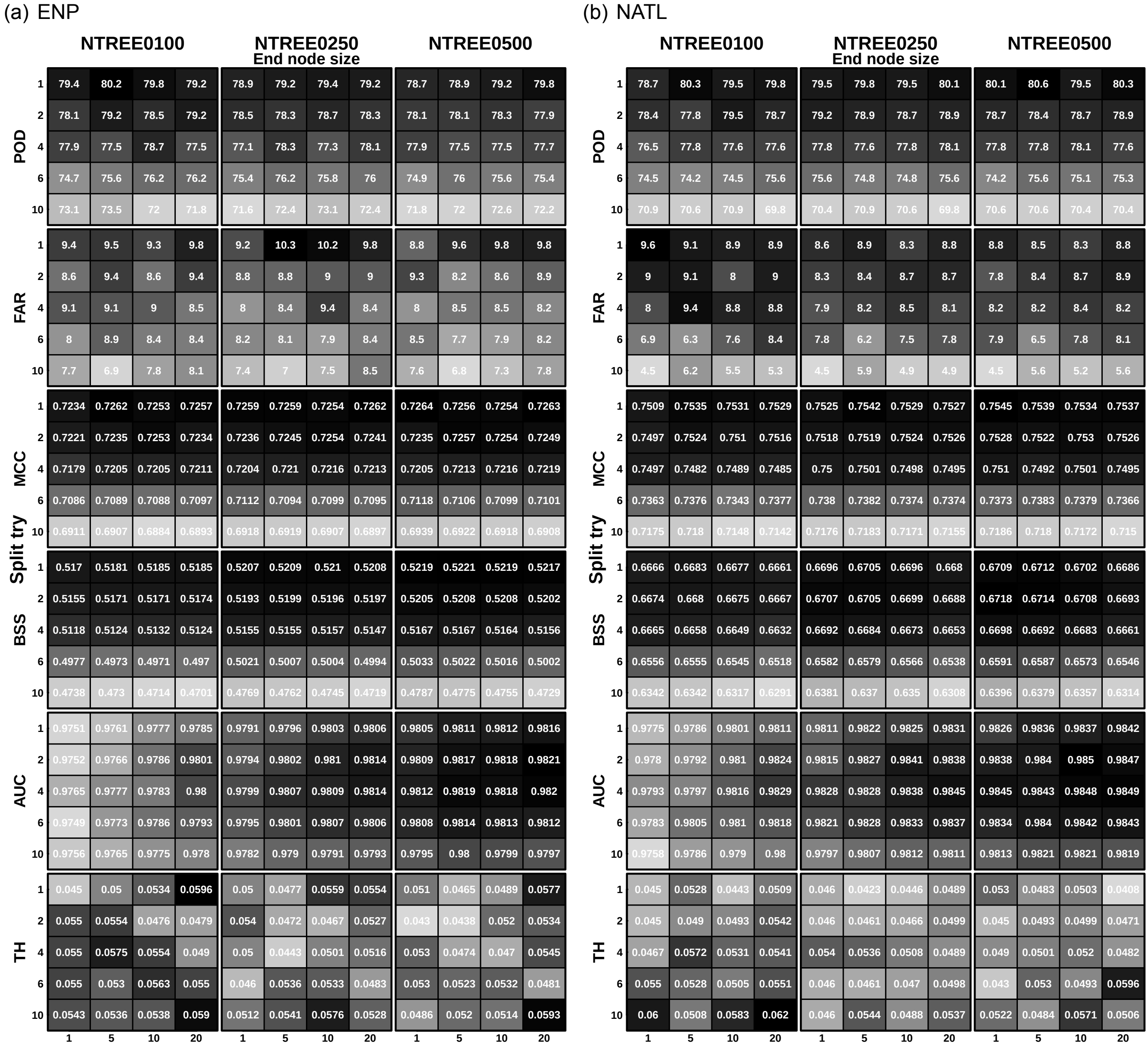

Figure R2 – Grid search results-In lines POD, FAR, MCC, Brier skill score BSS, area under the ROC curve (AUC) and the optimal threshold obtained from the ROC curve (TH) for a) ENP and b) NATL. In columns, are the results for different numbers of trees, and each 5 $\times$ 4 matrices are the results for a different random number of features considered to perform the best split (vertically) and different end node size (horizontally).

✳Comment– Line 225 : The authors should note some of the limitations of gini-based importance, namely the emphasis on input variables at the tops of the decision trees when proportions are largest. An alternative approach the authors could consider would be permutation importance, single-pass or multi-pass, which offers a more robust consideration of importance. See McGovern et al., 2019

RESPONSE– The permutation importance has been computed for **one RF** under the calibration experiment configuration (similar to Figure 9 of the manuscript). Figure R3 shows permutation importance for both basins. We see that some of the most important features are similar to those obtained from the Gini-based importance for both basins (RV850sd, MSLPmin). Otherwise, the order of the feature of importance between the two basins is very different. Besides, some variables like MSLPmax (ENP) or RV850min (NATL) are shown to be important, while UV10max is shown to be of lower importance (ENP). This makes the interpretation of feature importance unclear. Given these results, it seems that permutation-based importance is less consistent in our case.

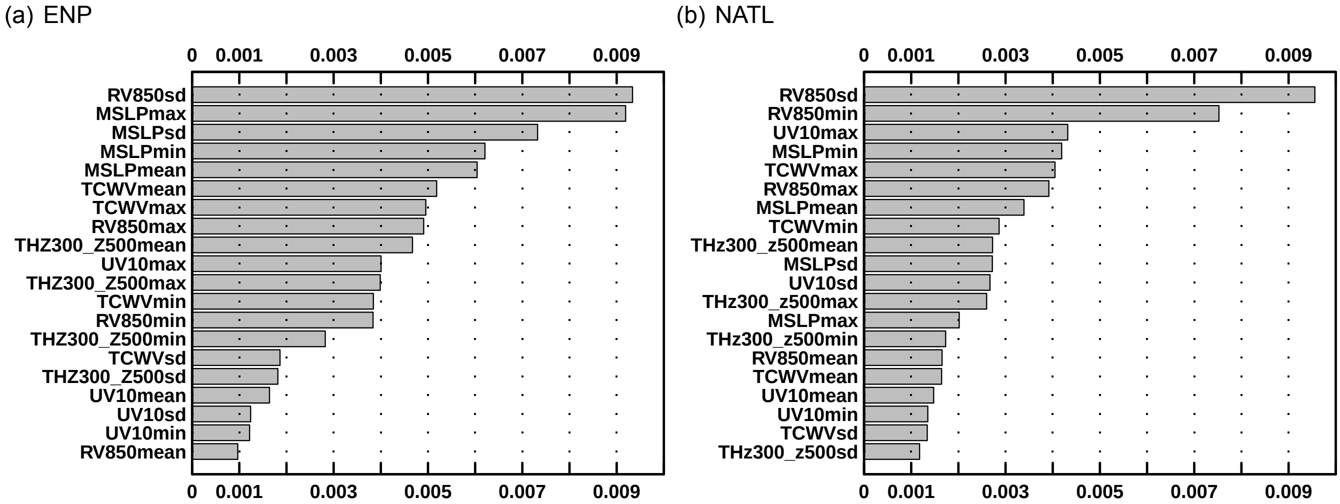

Figure R3 – Permutation-based importance.

✳*Comment– The authors use MCC, POD, and FAR based on a nebulous threshold of 50% to define TC track objects. Rather than using this approach, authors could leverage existing verification metrics that assess the probabilistic skill of the RF models (e.g., Brier Skill Score, Reliability diagrams). This approach would effectively assess the skill of the system at a range of probabilistic thresholds, which could also be leveraged for the area under the ROC statistics.*

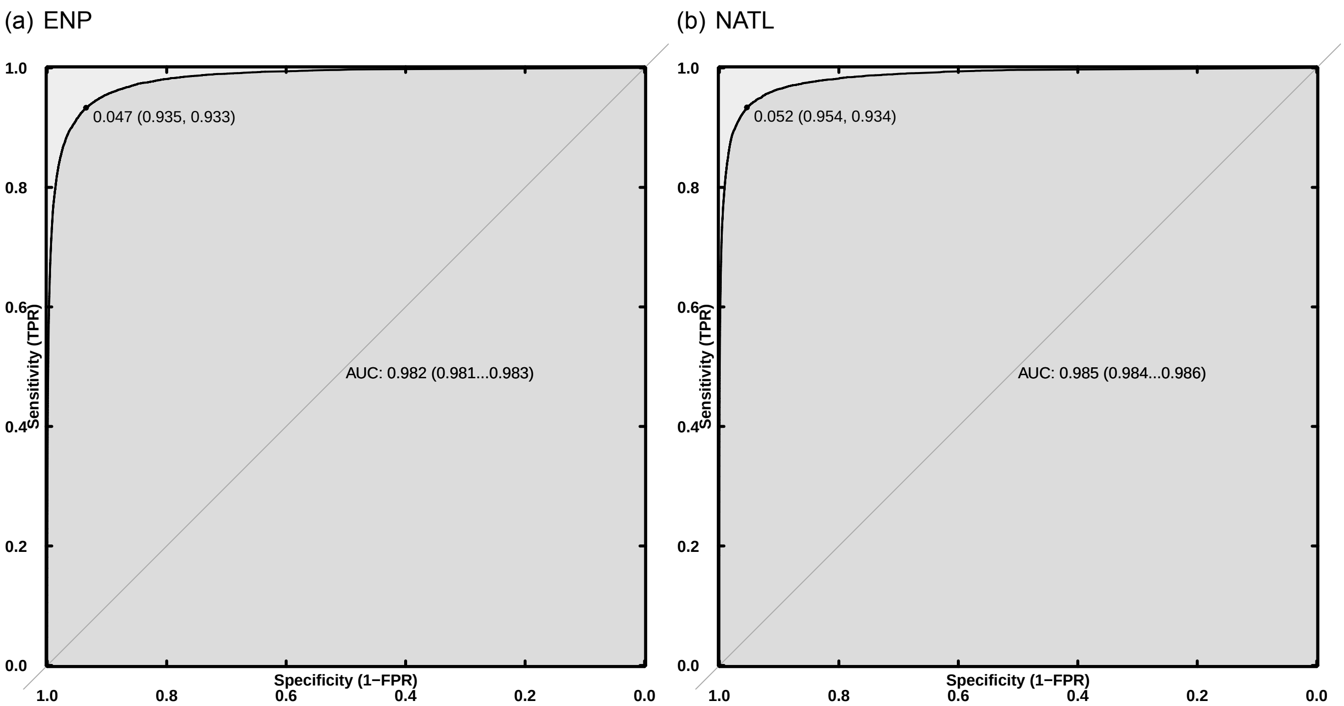

Figure R4 – ROC curves.

RESPONSE– Thank you for this comment. We tested different thresholds below and above 50%. The effect was (i) that the higher the threshold, the lower the POD and (ii) that the lower the threshold, the higher the FAR. This behaviour was quasi-linear, so we chose the middle 50%. One can adapt this level according to the desired applications. In addition, when looking at BSS or AUC, they were extremely similar for different thresholds (as in Figures R1 and R2) and did not help us to make a choice. We also looked into ROC curves to determine the optimal threshold. Figure R4 shows ROC curves for **one RF** under the validation experiment. For both basins, the optimal threshold is around 5%. With that threshold, we reach a POD of 100% but a FAR of 70%, which is fiercely undesired. This choice of 50% and the use of MCC, POD, and FAR was the best choice at hand to evaluate our tracker. The choice of threshold was clarified in lines 191-194 of the revised manuscript as :

CHANGES– Different thresholds below and above 0.5 have been tested (not shown). The result was (i) that the higher the threshold, the lower the ability to detect TC and (ii) that the lower the threshold, the higher the false alarms. This behaviour was quasi-linear, so we chose 0.5 to be performant to detect while having a low false alarm rate. One can adapt this level according to the desired applications.

Technical Edits and Questions

✳*Comment– Generally : the authors should spend a substantial amount of time proofreading the document for lingering grammar issues.*

RESPONSE– We tried as best as we could, we hope that the updated version of the manuscript is now satisfactory.

✳*Comment– Line 48 : Change to "this study focuses on data-driven algorithms using machine learning". Sometimes "so-called" can have a negative/inappropriate connotation, which I don't believe was your intent.*

RESPONSE– It has been removed and rewritten in line 47 of the revised manuscript

✳*Comment– Lines 93-96 : While I understand it is a long-held tradition to include a "table of contents paragraph" in this manner, you can remove this paragraph – it has no particular value for readers. The scientific structure of manuscripts has remained unchanged for decades, and every reader knows that methods will come next, results afterwards, and so on. If a reader is interested in a particular section, they can seek out the section header to know what is contained within.*

RESPONSE– Agreed and done

✳*Comment– Line 99 : Remove this single line*

RESPONSE– Done

✳*Comment– Line 99 : Remove "cyclonic" – seasons are not "cyclonic". Alternatively, can adjust to "cyclone seasons"*

RESPONSE– We modified accordingly in line 97 of the revised manuscript

✳*Comment– Line 106 : "Track records that do not provide"*

RESPONSE– Rewritten line 99-100 of the revised manuscript as :

CHANGES– Those labelled “spur”, not providing maximum wind and minimum pressure, and not reaching the Tropical Storm (TS) stage, are removed.

✳*Comment– Lines 106-107 : If a TC undergoes extratropical transition, how is the transition from TC to extratropical TC handled ? Also, how is the TC's demise to the depression stage handled ? Only the TC achievement is mentioned here (i.e., genesis).*

RESPONSE– The transition from TC to extratropical TC is handled by limiting our study basin to 30°N. Thus, only TCs are considered and transitions to ET cyclones are not considered. For TC crossing this northward boundary, only the portion lying below 30°N is kept. In this section, we explain which TCs are kept for the following of the study. If the tropical storm level is not reached, the whole track is not considered. We do not handle the demise to the depression stage. Clarification about ET cyclones is given in lines 95-97 of the revised manuscript as :

CHANGES– First, extratropical cyclones are not considered in this study. Our study basins are limited to 30°N. Thus, only TCs are considered, and transitions to extratropical cyclones are not. For TC crossing this northward boundary, only the portion lying below 30°N is kept.

✳*Comment– Line 131 : Moisture is misspelt*

RESPONSE– Corrected in line 125 of the revised manuscript

✳*Comment– Lines 136-138 : The description here appears to have two statements in conflict with one another. First, the text says that every box has a vector of ones and zeros constructed : is this for every grid point in the box ? The next sentence says the box is encoded as a 1 or 0. Some additional clarity and perhaps the rewording of these sentences is needed to clarify the approach. I suspect it is the latter, but the wording is a bit confusing.*

RESPONSE– A vector of zeros and ones is constructed for each 20°×10° overlapping box not for each grid-point in the box. Lines 136-138 of the original manuscript are clarified in lines 129-131 of the revised manuscript as :

CHANGES– Then, for every box, a vector of zeros and ones is constructed ~~: every timestep,~~ as follows : a box containing an IBTrACS point reaching TS intensity ($P_{\min} \leq 1005$ hPa and $u_{10} \geq 16\text{ms}^{-1}$) is coded 1, and 0 otherwise at every timestep.

✳*Comment– Line 134 : Why are the boxes not immediately adjacent to one another ? Could a TC be missed if it lies outside of the boxes in the white areas of Figure 1 ?*

RESPONSE– As indicated in line 134 of the original manuscript (line 130 in the revised version), the boxes overlap. Additionally, as stated in the caption of Figure 1, only every second box is shown to improve clarity. Plotting all overlapping boxes would make the figure unreadable.

✳*Comment– Lines 139-140 : What is the motivation for synthesising the ERA5 data in the boxes to single-statistic values ? Other works have used spatial regions to encode relevant spatial relationships into RFs [see* Hill *et al., 2020 ;* Hill & Schumacher*, 2021 ;* Hill *et al., 2023 ;* Hill *et al., 2024 ;* Schumacher *et al., 2021] and have had tremendous success, including deducing how those spatially oriented data contribute to forecast skill [*Mazurek *et al., 2025]. Others tackling severe weather hazards have taken a synthesising approach too [see* Clark & Loken*, 2022 ;* Loken *et al., 2022]. Were there any tests that also included the full box of ERA5 data to demonstrate that the single-value statistics were a better methodological choice ?*

RESPONSE– No formal test has been performed to demonstrate that the use of four single-value statistics instead of the whole field in the box was better. In this study, the TC tracking problem is handled as a binary classification problem. What we seek is the presence or absence of a TC within one box, given an atmospheric situation, regardless of its position. The position is deduced from the minimum of sea level pressure. Thus, four statistics summarising the spatial structure are preferred to describe the whole $20^\circ \times 10^\circ$ box. Furthermore, the ERA5 spatial resolution is 0.25°, resulting in 3200 grid-points per box for each physical variables. Using 16000 predictors to predict one probability does not seem reasonable. Last but not least, as already stated in the last paragraph of the introduction, the ultimate goal of such a tracker is the tracking of TC in future climate simulations. Indeed, having many variables implies potential overfitting, impeded interpretation of the results and lower transferability to future climate simulations, in particular due to their systematic biases. Higher data frugality achieved by considering simple variable statistics instead of entire variable fields potentially improves the transferability of the tracking to climate simulations. This has been clarified in lines 139-143 of the revised manuscript as :

CHANGES– No formal test has been performed to demonstrate that using these four single-value statistics instead of the whole field in the box was better. Since only the presence or absence of a TC within one box, regardless of its position, is sought, these four statistics summarising the spatial structure are preferred to describe the whole $20^\circ \times 10^\circ$ box. Furthermore, the ERA5 spatial resolution is 0.25°, resulting in 3200 grid-points per box for each physical variable. Using 16000 predictors to predict a single outcome does not seem reasonable.

✳*Comment– Lines 147-148 : This sentence is not needed – can be removed. All of this information is contained in the section headers.*

RESPONSE– Agreed, have been removed.

✳*Comment– Line 174-175 : To be consistent with both machine learning and atmospheric science literature, the "calibration" phase should be referred to as the "training" phase of the ERF. Then, you use cross-validation to validate the trained model on withheld periods – you don't use those withheld periods to "calibrate" the models.*

RESPONSE– As stated in lines 174-175 of the original manuscript : the calibration experiment consists in training ERF over the whole 1980-2021 period and validating over the whole period. Some clarifications have been made in lines 174-183 of the revised manuscript as :

CHANGES–

1. Calibration experiment : one ~~calibration~~ training of the ERF is made using the whole data during the 1980-2021 period and validated over the same period where all the tracks are sought to be reconstructed~~from it~~,

2. Validation experiment : a 6-fold cross-validation (see Fig. 2) where yellow years within each fold (35 years) are used to ~~calibrate~~ train the ERF. The validation is performed over tracks reconstructed for all the validation years (in blue) from the six folds, allowing to validate ERF over the whole 1980-2021 period. This cross-validation is chosen to ~~minimize~~ minimise the effect of any potential trend and interannual variability in the TC statistics (frequency, intensity) and the changes in IBTrACS data quality. Most of the ERF evaluations will rely on this experiment.

3. Test experiment : from the ~~calibration~~ training performed over the whole ~~time period for a given~~ period in the calibration experiment for ENP (resp. NATL) basin, the TC tracks over the ~~other basin are reconstructed~~ NATL (resp. ENP) are reconstructed over the same period. This is done to evaluate the generalizability of ERF.

⁕*Comment– Line 188 : Should RF actually be ERF ?*

RESPONSE– Agreed and modified accordingly in lines 189 of the revised manuscript.

⁕*Comment– Line 188 : Did you consider alternative probability thresholds (beyond just 50%) to assignment detected tracks (D) ?*

RESPONSE– Thank you for that comment. We test different thresholds lower and above 50%. The effect was that the higher the threshold, the lower the POD and the lower the threshold, the higher the FAR. This behaviour was quasi-linear, so we chose the middle 50%. The choice of threshold was clarified in lines 191-194 of the revised manuscript as :

CHANGES– Different thresholds below and above 0.5 have been tested (not shown). The result was (i) that the higher the threshold, the lower the ability to detect TC and (ii) that the lower the threshold, the higher the false alarms. This behaviour was quasi-linear, so we chose 0.5 to be performant to detect while having a low false alarm rate. One can adapt this level according to the desired applications.

⁕*Comment– Lines 251-253 : This text is best reserved for the figure caption – please move it there if not already. This text is just describing the figure, not the science.*

RESPONSE– Agreed and removed, already described in the caption.

⁕*Comment– Figure 3 : It would be good to see the full distribution of MCC scores for the 100 RFs plotted as error bars, akin to a 95% confidence interval. Are the MCC values truly statistically indifferent ? (It is hard to tell, but maybe this detail is plotted as light blue lines ? If so, please try and make these lines clearer so they can be discerned and provide a description in the figure caption)*

RESPONSE– The boxplots for MCC have been made clearer, and we add a description of the boxplots in the caption of Figure 3 as follows :

CHANGES– The *top* and *bottom* fences are situated at 1.5 times the interquartile range from the box, and the dots are the values beyond these fences. The orange line represents the median value.

✳*Comment– Lines 273-274 : What is meant by “calibration experiments”? Are you just evaluating the model’s ability to detect storms over the testing period for which it was trained? It is to be expected that POD will be high and FAR low.*

RESPONSE– Yes, this is what we described as “calibration experiments”. We hope the definitions of the different experiments in the revised manuscript, lines 174-183, make things clearer.

✳*Comment– Line 283-284 : Isn’t a missed track by definition lower probability? Aren’t hits/misses defined by probabilities greater than or less than 50%? These box plots in Figure 5b are being more or less constrained by the methods used, and don’t necessarily provide much scientific reasoning for “FAs are less likely to happen than hits”. The authors should reconsider the usefulness of this analysis regarding their methodological choices.*

RESPONSE– Thanks for these comments.

- Low probability means absence of TC. Missed tracks are tracks that were not detected in ERA5, although they were observed in IBTrACS. We explained in the manuscript that it is because they are absent in ERA5 and not a deficiency of ERF. This is something noteworthy to us.
- The 50 % threshold fixes if a TC is detected or not. Hit/FA/Miss/ assignment depends on whether a detected track was in IBTrACS or not (Hit/FA) and if an observed track was not detected (Miss). The method estimates the probabilities. These are evaluated conditionally on the Hit/FA/Miss/ assignment, which makes Fig. 5b interesting.

This has been clarified in lines 289-293 of the revised manuscript as :

CHANGES– Figure 5b ~~) show that probabilities associated to~~ shows TC probabilities conditionally on its labelling as Hit, Miss, or FA. Probabilities associated with Hit tracks (median above 0.9) are substantially different compared to those associated ~~to~~ with FA (median a little above 0.6). This means that even if FA tracks are detected (probability >0.5) by ERF, FA are less likely to happen than Hits. Miss tracks are associated with very low probabilities, meaning they are completely missed by ERF while having been recorded in IBTrACS.

✳*Comment– Lines 320-322 : As mentioned earlier, they are also prescribed by the authors, so these results are not extremely surprising. See major comment above.*

RESPONSE– As stated in the answer to your first comment, we are surely prescribing the predictors, but we are not informing RF from which predictor it should learn to estimate the probability. So, analysing which are the most important predictors in the estimation of that probability is interesting.

✳*Comment– Lines 348-349 : This information is once again best reserved for the figure caption.*

RESPONSE– Assuming it refers to lines 318-319 (Figure 9), Agreed and removed, already described in the caption.

✳*Comment– Figure 10 : This is an excellent figure that clearly demonstrates how the RFs are learning the relevance of each predictor to drive the yes/no predictions.*
RESPONSE– Thanks for this comment.

References

Clark, Adam J. & Loken, Eric D. (**2022**). "Machine Learning–Derived Severe Weather Probabilities from a Warn-on-Forecast System". Weather and Forecasting. doi : 10.1175/WAF-D-22-0056.1.

Hill, Aaron J., Herman, Gregory R. & Schumacher, Russ S. (**2020**). "Forecasting Severe Weather with Random Forests". Monthly Weather Review. doi : 10.1175/MWR-D-19-0344.1.

Hill, Aaron J. & Schumacher, Russ S. (**2021**). "Forecasting Excessive Rainfall with Random Forests and a Deterministic Convection-Allowing Model". Weather and Forecasting. doi : 10.1175/WAF-D-21-0026.1.

Hill, Aaron J., Schumacher, Russ S. & Green, Mitchell L. (**2024**). "Observation Definitions and Their Implications in Machine Learning–Based Predictions of Excessive Rainfall". Weather and Forecasting. doi : 10.1175/WAF-D-24-0033.1.

Hill, Aaron J., Schumacher, Russ S. & Jirak, Israel L. (**2023**). "A New Paradigm for Medium-Range Severe Weather Forecasts : Probabilistic Random Forest–Based Predictions". Weather and Forecasting. doi : 10.1175/WAF-D-22-0143.1.

Loken, Eric D., Clark, Adam J. & McGovern, Amy (**2022**). "Comparing and Interpreting Differently Designed Random Forests for Next-Day Severe Weather Hazard Prediction". Weather and Forecasting. doi : 10.1175/WAF-D-21-0138.1.

Mazurek, Alexandra C., Hill, Aaron J., Schumacher, Russ S. & McDaniel, Hanna J. (**2025**). "Can Ingredients-Based Forecasting Be Learned ? Disentangling a Random Forest's Severe Weather Predictions". Weather and Forecasting. doi : 10.1175/WAF-D-23-0193.1.

McGovern, Amy, Lagerquist, Ryan, Gagne, David John, Jergensen, G. Eli, Elmore, Kimberly L., Homeyer, Cameron R. & Smith, Travis (**2019**). "Making the Black Box More Transparent : Understanding the Physical Implications of Machine Learning". Bulletin of the American Meteorological Society. doi : 10.1175/BAMS-D-18-0195.1.

Schumacher, Russ S., Hill, Aaron J., Klein, Mark, Nelson, James A., Erickson, Michael J., Trojniak, Sarah M. & Herman, Gregory R. (**2021**). "From Random Forests to Flood Forecasts : A Research to Operations Success Story". Bulletin of the American Meteorological Society. doi : 10.1175/BAMS-D-20-0186.1.