"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

The authors addressed a number of my previous comments in their responses, as well as edits made to the manuscript. I also thank the authors for taking the time to edit and proofread the manuscript for grammatical issues.

Remaining Comments

✳*Comment– 1. Please be consistent with your use of acronyms. For example, you use "TCs" in Line 122 to refer to Tropical Cyclones, yet "TC" in Line 127 also refers to Tropical Cyclones. However, the "TC" acronym is defined on line 117 as "Tropical Cyclones" – so should the acronym be TC or TCs in this case ? Please edit the text as you see appropriate to have consistent plural/singular use of this (and other) acronyms.*

RESPONSE– Agreed, and modified throughout the paper.

✳*Comment– For the record, standardisation is not necessary for random forest applications (lines 153-154).*

RESPONSE– Thank you for that comment with which we agree. The standardisation has been done to anticipate tracking TCs in climate models with biases compared to ERA5. The standardisation removes part of the mean and variance biases of the climate models and potentially eases the transferability to the tracker to climate models without recalibration. This has been clarified in the paper in lines 138-141 of the revised manuscript as :

CHANGES– Note that standardisation is not necessary for the current application of random forests. However, it has been made here to anticipate tracking TCs in climate models with biases compared to ERA5. The standardisation removes part of the mean and variance biases of the climate models and potentially eases the transferability to the tracker to climate models without recalibration.

✳*Comment– 3. Lines 201-204 : As long as testing is done for a different basin, i.e., model trained on ENP for the full period and "tested" on NATL for the same period, this approach should be valid. However, if an*

ENP model is trained on a period and subsequently tested on that same period, I would have significant concerns about misrepresenting skill.

RESPONSE– As mentioned in the manuscript, three types of experiments are set. Assuming that the reviewer refers to the calibration experiment in the second part of this comment [*However ...*] : The calibration experiment is only performed as a first-order evaluation of the tracker. In machine learning, the first step is to evaluate the ability of the model to reproduce the training data. The calibration experiment has that sole purpose. As mentioned in the paper, the major part of the evaluation is performed through the validation and test experiments. It has been clarified in lines 179-180 of the revised manuscript as :

CHANGES– It is only performed as a first-order evaluation of the tracker and its ability to reproduce the training data.

✳*Comment– 4. Line 193 : Why are there 100 Random Forests used for each subsampling test ? The Random Forest by nature is an ensemble of Decision Trees – why do you need 100 RFs ?*

RESPONSE– The subset of zeros provided to each RF is different. The effect of subsampling zeros on the track reconstruction is evaluated with 100 RFs. We showed the effect of the subsample to be marginal in Figure 3 of the manuscript. It has been clarified in line 176 of the revised manuscript as :

CHANGES– with a different subset of zeros provided to each RF

✳*Comment– 5. Lines 210-214 and previous review comment : Both I and another reviewer asked to see forecast skill as a function of probability, e.g., reliability diagrams. I'm not convinced that the 50% threshold used here is not more-or-less arbitrary. Since you have developed a probabilistic prediction system, probabilistic skill metrics should be computed to evaluate skill properly and robustly.*

RESPONSE– The 50% threshold choice has not been chosen arbitrarily, but was based upon a key practical criterion : having the ability to track TC (high POD) while having a low FAR. As mentioned in the answer to your previous comments, (i) the higher the threshold, the lower the POD and (ii) the lower the threshold, the higher the FAR. This behaviour was quasi-linear, so we chose the middle 50%. We also looked at BSS or AUC, but they did not help us choose a threshold : they were extremely similar for different thresholds due to the unbalanced nature of the classification. We also looked into ROC curves to determine the optimal threshold. For both basins, the optimal threshold is around 5%. With that threshold, we reach a POD of 100% associated with a 70% FAR, which is obviously undesired. Similarly, we are not convinced that a reliability diagram is adapted to define a TC threshold, especially for such an unbalanced problem. Figure R1 shows these diagrams for ERF under the validation experiment and both basins. From these figures, we are not able to define a threshold. The 50% threshold seems a good practical compromise.

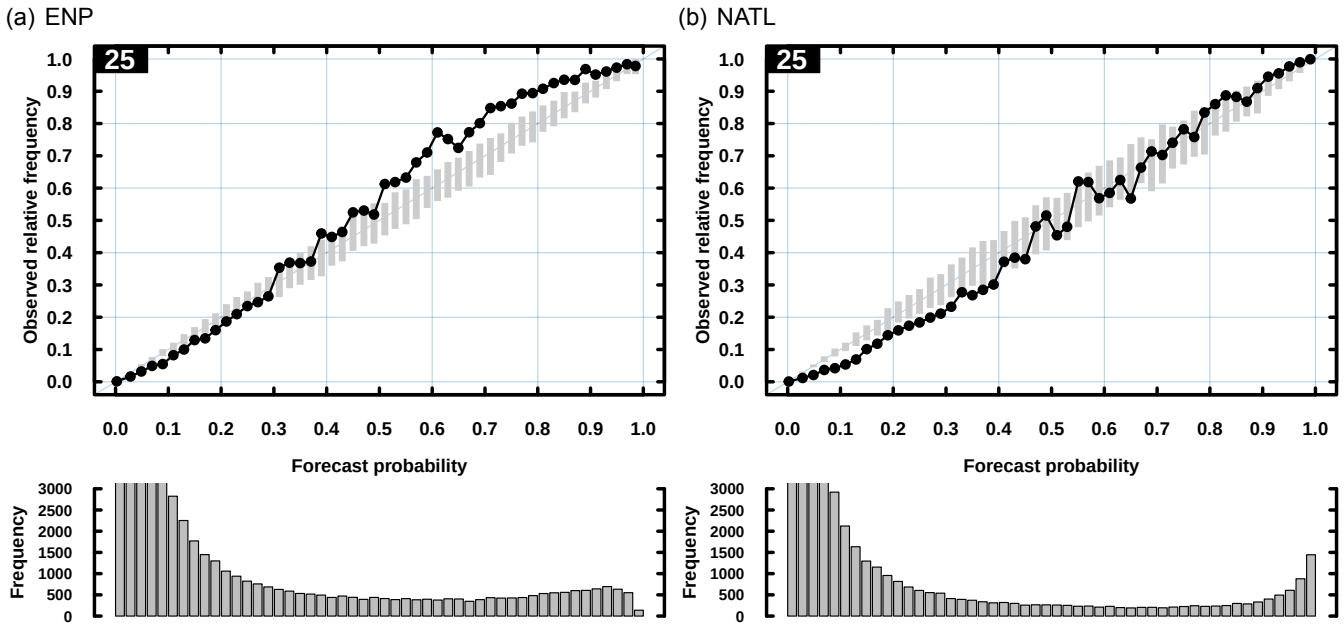

Figure R1 – Reliability diagramme curves for both basin and the histogram associated with the prediction (bottom). The histograms are cut to 3000 ; low probability values reach maximum frequencies above 5×10^5.