"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 #2

Summary :

This study uses an ensemble of random forests (ERFs) to identify and track tropical cyclones (TCs) within ERA5 in the North Atlantic and East Pacific basins. The identified TCs and tracks are compared with observations taken from IBTrACS, and the ERF performance was compared with the Tempest Extreme tracking algorithm. Overall, the authors demonstrate that the ERF performs well in identifying and tracking observed TCs (high probability of detection and low false alarm ratio).

Beyond simply demonstrating that the ERF "works", the authors also nicely examined the characteristics of false alarms and misses. The authors found that missed TCs and false alarms were generally associated with short-duration storms that were of marginal tropical storm intensity. In addition, the authors examined which of the chosen predictors from ERA5 had the largest Gini-based feature importance and the contribution of each to the outcome of the random forest prediction using SHAP values.

I personally found the manuscript an interesting and useful application of ERFs. I particularly appreciated the authors' discussion on the misses, false alarms, and predictors that most informed the random forest outcome. I also believe the manuscript can be further improved through both the comments below and a more careful editing of the spelling and grammar within the text. I am specifically interested in encouraging the authors to more carefully consider the probability provided by the ERFs using traditional ensemble verification methods such as Brier Skill Score and ROC diagrams. It would also be of interest to better understand if the characteristics of the misses and false alarms from the ERF and Tempest Extreme exhibit any noteworthy differences in location, intensity, duration, or environmental conditions. Overall, I believe this is a study worthwhile of publication after addressing the below comments.

Specific Comments :

✳*Comment– One of the main benefits of the ERFs is the probabilities provided. I wish the authors examined this in more detail. I recommend that the authors reconsider the use of a strict threshold, greater than 50% probability, as defining a TC event. There is no requirement for this to be the cutoff, and the authors may wish to explore alternative thresholds. Furthermore, the authors may wish to examine the reliability of the ERFs by assessing whether the spread correctly represents the forecast uncertainty by examining the spread-error ratio. On average, the ensemble spread should be equal to the error. In addition, I suggest the authors examine reliability diagrams, which compare the forecasted probability with the observed frequency, ROC diagrams, and Brier skill score. Each of these analyses will help determine the benefits of the probability provided by the ERFs and may reveal weaknesses of the ensemble design.*

RESPONSE– Thank you for that comment. We tested different thresholds below and above 50%. The effect was that the higher the threshold, the lower the POD, while the lower the threshold, the higher the FAR. This behaviour was quasi-linear, so we chose the middle 50%. In addition, when looking at BSS or AUC, they were extremely similar for different thresholds and did not help us to make a choice. We also looked into ROC curves to determine the optimal threshold. Figure R1 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. Concerning the spread-to-error ratio, we are not sure if it is pertinent for a binary (or any discrete) variable. Indeed, we want the probability to be a close as possible to 0 or 1, and the spread around 0 or 1 does not make sense. Though we computed it for ENP (0.044) and NATL (0.047). 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.

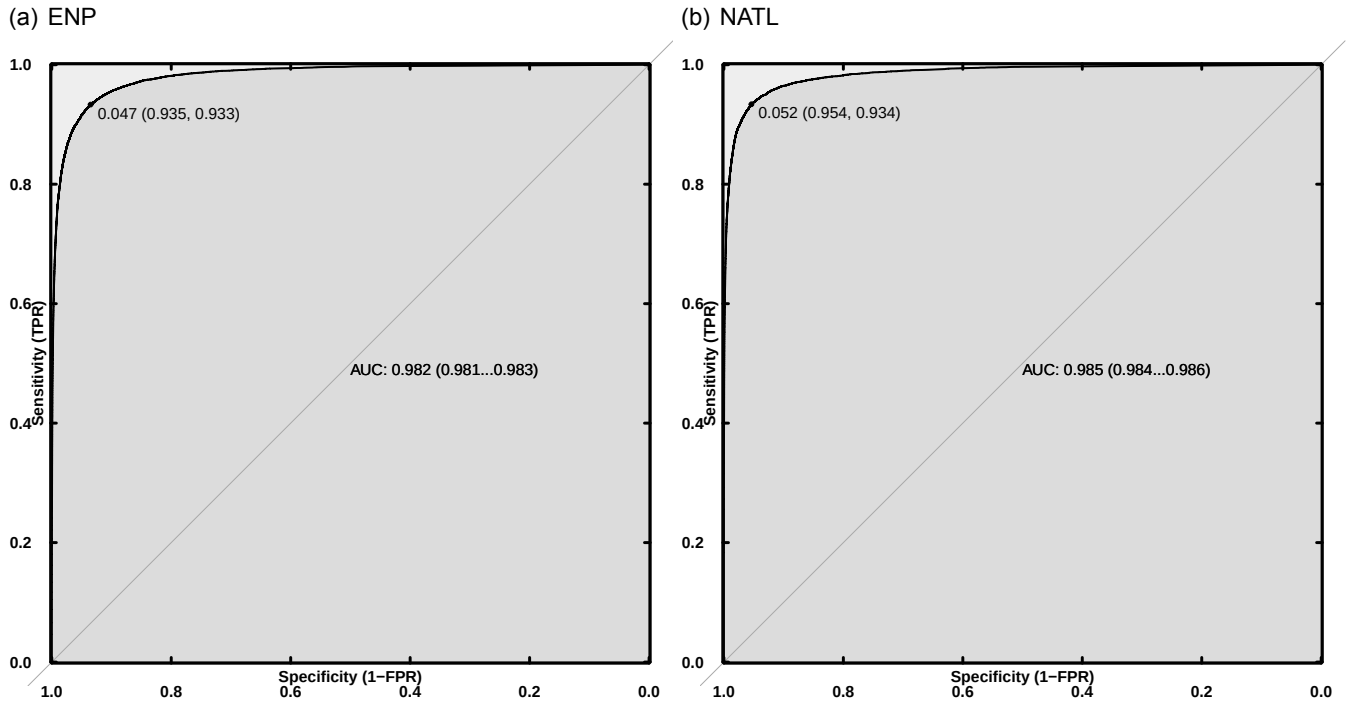

Figure R1 – ROC curves.

✳*Comment– I struggled to fully understand the details of the calibration, validation, and test experiments (L174-182). I am still a bit confused by the overlap between the calibration, validation, and testing periods. It appears from point 3 that the whole 1980-2021 period is used for testing. This is not a fair testing dataset, as the ERF was also tested using much of this same period. I believe the authors should perform testing using an entirely new period that was not used during training.*

RESPONSE– Eventhough the whole period is used for the evaluation, the validation experiment with the 6-fold cross-validation does not use the same data for training and validation, and the test experiment uses the whole period data from one basin for the training and data from the other

basin for the reconstruction, so data is never used in the training. 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– The authors also mentioned a potential change in the quality of IBTrACS with time in motivating their choice of validation data (every 6 years). I am curious if the authors tested how the performance of the ERF would change if they trained on an earlier period and then tested on a more recent period. This would be interesting for several reasons, including serving as an "easier" initial test for the potential application to future climate simulations, as the authors mention in the summary section.*

RESPONSE– Thank you for that comment. The cross-validation scheme chosen in the study was only one possibility among others. We did a 6-fold test in which validation years were stacked (1980-1986,···,2015-2021). In the following table, we show the POD and FAR for tracks reconstructed with the probability obtained from one RF and the validation experiment using the "stacked" cross-validation : (i) over the whole 1980-2021 period (6-fold) and (ii) for the last fold (2015-2021 validation years). We see that the results for the whole period are similar to those of the validation experiment in Table 1 of the manuscript. The results are even better when the last fold, with the latest validation year, is considered (except the larger FAR for NATL, 10.7%).

Validation	ENP		NATL	
years	POD	FAR	POD	FAR
1980-2021	78.1	8.4	77.5	8.5
2015-2021	85.6	4.7	87.0	10.7

This is now mentioned in lines 278-282 of the revised manuscript as :

CHANGES– Note that the choice of validation data (one year every 6 years) in the cross-validation scheme in the study was only one possibility among others. A test (not shown), with a 6-fold cross-validation scheme in which validation years are stacked (1980-1986,···,2015-2021), yielded similar results for the validation experiment and, results are even better when only the last fold, with the latest validation years (2015-2021), is considered.

✳*Comment– 300 km (L193) appears to be a generous threshold for the distance between an observed and identified TC to be considered a hit. This value is still probably small enough that it is identifying the same storm, but large enough that the center location may be off by the approximate size of the TC. Why was this value chosen and how sensitive are the results to this threshold ?*

RESPONSE– In Bourdin et al., 2022, a sensitivity analysis was conducted on this specific parameter in appendix D. In a nutshell, it was shown that this is not a sensitive parameter : choosing values between 200 and 400km did not make a difference in the final results. It was shown that the matching distance is usually below 100km for the SLP-based trackers (UZ & CNRM) – i.e. a couple of grid cells – and in more than 99% of the cases below 200km. Visual and statistical examination showed that, even with a large matching distance, these tracks are actual matches for at least a day. Taking values above 200km allows keeping the very few tracks where IBTrACS and the tracker disagree on part of the track, for example, the genesis in cases where two systems merged. For this reason, 300 km was selected as a reasonable value for the mean distance between the two tracks over the period for which they both exist. This has been clarified lines 197-199 of the revised manuscript as :

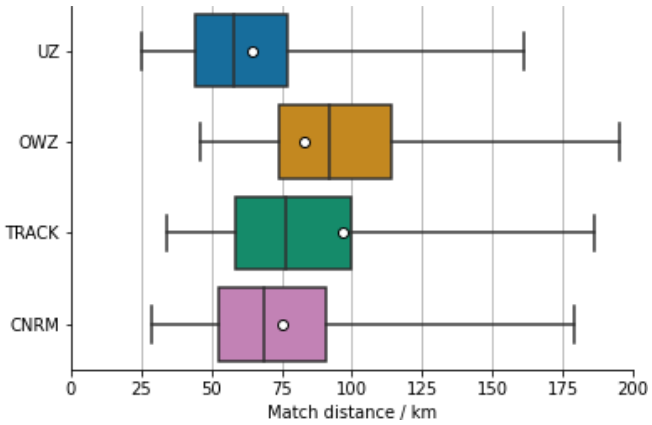

Figure D3 Distribution of distance between matching detected and observed tracks. Whiskers display the 1st and 99th percentiles, and white points show the mean of the distributions. Outliers are not shown.

CHANGES– In Bourdin et al., 2022, a sensitivity analysis was conducted on the 300 km distance limit in appendix D. In a nutshell, it was shown that results are not sensitive to this limit, and 300 km was selected as a reasonable value.

✳*Comment– It would be helpful to readers to define each of the predictors from ERA5 within a table in the supplementary material.*

RESPONSE– The description of the variable is added as Table S1 of the supplementary material and is mentioned in line 118 of the revised manuscript as :

CHANGES– These variables are described in Table S1 of the supplementary material.

Table 1 – Variables description extracted from the Climate Data Store website. https://cds.climate.copernicus.eu/datasets/

Variable	Abbreviation	Unit	Description
Mean sea level pressure	MSLP	Pa	This parameter is the pressure (force per unit area) of the atmosphere at the surface of the Earth, adjusted to the height of mean sea level. It is a measure of the weight that all the air in a column vertically above a point on the Earth's surface would have if the point were located at mean sea level. It is calculated over all surfaces - land, sea and inland water.
10-m wind intensity	UV10	m s^{-1}	This parameter can be calculated by combining the eastward (U) and northward (V) components of the 10m wind. U is the horizontal speed of air moving towards the east, and V is the horizontal speed of air moving towards the north, at a height of ten metres above the surface of the Earth.
Total column water vapour	TCWV	kg m^{-2}	This parameter is the total amount of water vapour in a column extending from the surface of the Earth to the top of the atmosphere. This parameter represents the area-averaged value for a grid box.
Relative Vorticity at 850 hPa pressure level	RV850	s^{-1}	This parameter is a measure of the rotation of air in the horizontal, around a vertical axis, relative to a fixed point on the surface of the Earth. On the scale of weather systems, troughs (weather features that can include rain) are associated with anticlockwise rotation (in the Northern Hemisphere), and ridges (weather features that bring light or still winds) are associated with clockwise rotation. It is extracted at the 850 hPa pressure level.
Geopotential thickness between 500 and 300hPa pressure levels	THZ300_Z500	m	This parameter is obtained from the gravitational potential energy of a unit mass, at a particular location, relative to mean sea level. It is also the amount of work that would have to be done, against the force of gravity, to lift a unit mass to that location from mean sea level. The geopotential height can be calculated by dividing the geopotential by the Earth's gravitational acceleration, g (=9.80665 m s^{-2}), and the thickness is the difference in height between two levels of pressure.

✳*Comment– I am interested in understanding if the characteristics of the false alarms and misses are similar to the ERFs and Tempest Extremes. I suggest the authors recreate Figures 5, 6, and 7 for Tempest Extreme within the supplemental figures. This analysis may help identify the strengths and weaknesses of each approach.*

RESPONSE– Figures 5 to 7 have been reproduced for UZ and both basins. They have been added to the supplementary material as Figures S7 to S12. Panel b of Figure 5 and the two top panels of Figure 7, showing the probability associated with the tracks, are irrelevant for UZ. For both basins, the track properties obtained from UZ are similar to those obtained with ERF (see Figs. S7, S8, S10 and S11). The main difference resides in the properties of missed tracks. These tracks are similarly short but more frequent and slightly more intense (higher maximum wind and lower SLP). Properties of FAs are similar for UZ and ERF in the NATL basin. In the ENP basin, FAs from UZ are significantly more frequent and intense than the FAs generated by ERF. The spatial distribution of Miss and FA of the UZ tracker is also similar to that of ERF (Fig. S9 and S12). Miss tracks are distributed over the entire domain with low intensity, and FA tracks are mostly located in the primary development region and coastal areas, where TCs are typically weaker.

These comments have been added in lines 363 to 371 of the revised manuscript as :

CHANGES– Figures 5 to 7 have been reproduced for UZ and both basins. They have been added to the supplementary material as Figures S7 to S12. Panel b of Figure 5 and the two top panels of Figure 7, showing the probability associated with the tracks, are irrelevant for UZ. For both basins, the track properties obtained from UZ are similar to those obtained with ERF (see Figs. S7, S8, S10 and S11). The main difference resides in the properties of missed tracks. These tracks are similarly short but more frequent and slightly more intense (higher maximum wind and lower

SLP). Properties of FAs are similar for UZ and ERF in the NATL basin. In the ENP basin, FAs from UZ are significantly more frequent and intense than the FAs generated by ERF. The spatial distribution of Miss and FA of the UZ tracker is also similar to that of ERF (Fig. S9 and S12). Miss tracks are distributed over the entire domain with low intensity, and FA tracks are mostly located in the primary development region and coastal areas, where TCs are typically weaker.

Technical edits :

✳*Comment– L2 : change "evanesce" to "weaken"*

RESPONSE– Done

✳*Comment– L37-38 : Another tracking algorithm the authors may be interested in referencing here is TRACK (Hodges, 1994). This algorithm differs from others in that it is more general and tracks all vorticity maximums and only later filters out TCs using a warm core threshold. Hodges, K. I., 1994 : A General Method for Tracking Analysis and Its Application to Meteorological Data. Mon. Wea. Rev., 122, 2573–2586, https ://doi.org/10.1175/1520-0493(1994)122<2573 :AGMFTA>2.0.CO ;2.*

RESPONSE– We are aware of the TRACK algorithm. A comparison of TRACK, UZ (TempestExtremes) and two other tracking algorithms was performed previously in [Bourdin et al., 2022]. For this study, we wanted to keep it simple and compare to only one "traditional" tracker. We chose TempestExtremes/UZ for two reasons : the RF tracker was targeted towards IBTrACS as a ground truth, and the comparison showed that TempestExtremes/UZ had the lowest False Alarm Rates with respect to IBTrACS, and TempestExtremes is faster and easier to run than TRACK.

We agree that TRACK is designed in a way that allows it to capture a wide range of features. Unfortunately, that is also the reason why it has a large number of false alarms. In our opinion, TRACK's strength lies in its ability to capture the early and late stages of TCs. In the case of the present study, we are focusing on the mature stage of TCs as captured within IBTrACS. It should also be noted that, in a similar way to TRACK, TempestExtremes is a software that can be used to implement many algorithms, and the warm core criterion can be removed to detect all depressions (or vortices if vorticity is used as a variable).

✳*Comment– L144 : The authors should more carefully describe what is meant by "standardized". This is important for reproducibility.*

RESPONSE– This has been clarified in lines 137-138 of the revised manuscript as :

CHANGES– standardised (*i.e.* centred and divided by the standard deviation).

✳*Comment– L252 : I suggest the authors replace "different subsampling of zeros" with language more physically intuitive.*

RESPONSE– Line 251-253 of the original manuscript was removed since it is a repetition of the caption of figure 3. "different subsampling of zeros" has been replaced by "different subsampling of non-TC situations (*i.e.* zeros)"

✳*Comment– Figure 6 : The layout of the figure panels in Figure 6 are a bit confusing. I was repeatedly confusing panels (a) and (b). I suggest revising the layout to avoid this.*

RESPONSE– We invert panel labels a) to b).

✳*Comment– L311 : remove "basin"*

RESPONSE– Done, now in line 320 of the revised manuscript.

✳*Comment– L314-315 : What is the basis for this hypothesis ?*

RESPONSE– IBTrACS record relies on the human interpretation of satellite images from different meteorological centres. Since these TCs have intensities around the tropical storm threshold ($P_{\min} \leq 1005$ hPa and $u_{10} \geq 16 \mathrm{ms}^{-1}$), some of them can be recorded as not having reaching the tropical storm intensity and were therefore excluded from our study.

✳*Comment– L323 : Change "with" to "which".*

RESPONSE– Done, now in line 331 of the revised manuscript.

✳*Comment– L325 : Change "since they are associated with the strong surface winds and the location of the cyclone eye, respectively".*

RESPONSE– Done, now in line 332 of the revised manuscript.

✳*Comment– L332-333 : A transition would be helpful here to emphasize the different information provided by each of these analyses.*

RESPONSE– Agreed. Some clarifications have been made in lines 340-341 of the revised manuscript as :

CHANGES– Indeed, we want to evaluate the ability of RFs to learn the relevance of each predictor to drive each TC/non-TC prediction.

✳*Comment– L362-362 : I suggest splitting this into two sentences. Ending the first sentence after "literature".*

RESPONSE– Done, now line 380-381 of the revised manuscript.

✳*Comment– L390-391 : The end of this sentence, "indicating us to be…" should be revised.*

RESPONSE– Done, now in line 409 of the revised manuscript and modified as :

CHANGES– indicating ~~us~~ that we ought to be cautious when removing predictors.

References

Bourdin, S., Fromang, S., Dulac, W., Cattiaux, J. & Chauvin, F. (**2022**). “Intercomparison of four algorithms for detecting tropical cyclones using ERA5”. Geoscientific Model Development. doi : 10.5194/gmd-15-6759-2022.