



Ensemble Random Forest for Tropical Cyclone Tracking

Pradeebane Vaithinada Ayar¹, Stella Bourdin², Davide Faranda¹, and Mathieu Vrac¹

¹Laboratoire des Sciences du Climat et de l'Environnement, UMR 8212 CEA-CNRS-UVSQ, Université Paris-Saclay & IPSL, CEA Saclay, 91191, Gif-sur-Yvette, France

²Atmospheric, Oceanic and Planetary Physics, Department of Physics, University of Oxford, Oxford, United Kingdom

Correspondence: Pradeebane Vaithinada Ayar (pradeebane@laposte.net)

Abstract. Even though tropical cyclones (TCs) are well documented during the intense part of their lifecycle until they start to evanesce, many physical and statistical properties governing them are not well captured by gridded reanalysis or simulated by earth system models. Thus, the tracking of TCs remains a matter of interest for the investigation of observed and simulated tropical cyclones. Two types of cyclone tracking schemes are available. On the one hand, there are trackers that rely on physical and dynamical properties of the TCs and user-prescribed thresholds, which make them rigid. They need numerous variables that are not always available in the models. On the other hand, there are trackers leaning on deep learning which, by nature, need large amounts of data and computing power. Besides, given the number of physical variables needed for the tracking, they can be prone to overfitting, which hinders their transferability to climate models. In this study, the ability of a *Random Forest* (RF) approach to track TCs with a limited number of aggregated variables is explored. Hence, the tracking is considered as a binary supervised classification problem of TC-free (zero) and TC (one) situations. Our analysis focuses on the Eastern North Pacific and North Atlantic basins, for which, respectively, 514 and 431 observed tropical cyclone track records are available from the IBTrACS database during the 1980-2021 period. For each 6-hourly time step, RF associates TC occurrence or absence (1 or 0) to atmospheric situations described by predictors extracted from the ERA5 reanalysis. Then situations with TC occurrences are joined for reconstructing TC trajectories. Results show the ability and performance of this method for tracking tropical cyclones over both basins, and good temporal and spatial generalization as well. RF has a similar TC detection rate as trackers based on TCs' properties and significantly lower false alarm rate. RF allows us to detect TC situations for a range of predictor combinations, which brings more flexibility than threshold based trackers. Last but not least, this study shed light on the most relevant variables allowing to detect tropical cyclone.

1 Introduction

Tropical cyclones (TCs) are among the most devastating extreme events in terms of casualties and economic losses (Centre for Research on the Epidemiology of Disasters, 2021; UN Office for Disaster Risk Reduction, 2020). Several conditions are necessary for their formation. TC genesis requires warm sea surface temperatures to draw its energy from, low wind shear and ample humidity to ensure a stable vortex while maintaining the convection and adequate influence from the Coriolis force combined with a pre-existing low-pressure disturbance in the atmosphere. Thus, a vortex is created around the depression with strengthening winds, and fuelled by ascending humid air (Emanuel, 2003; Weaver and Garner, 2023). It becomes a self-



sufficient system that continuously draws energy from the ocean until reaching an unfavourable environment or land (the landfall). Then, the TC loses its energy, which causes its rapid dissipation (Kepert, 2010; Rüttgers et al., 2019).

Understanding how human-induced climate change influences TC activity remains a challenging scientific issue (Knutson et al., 2010; Walsh et al., 2016, 2019). Given the incomplete theoretical understanding of TCs and the limited observation
 30 timespan, the study of the changes in their properties relies on model simulations (Knutson et al., 2019, 2020). Despite the tremendous effort made to increase the resolution of the Earth System Models (around 100 km for the last generation of models), it is still too low to simulate realistic TCs (Camargo and Wing, 2016; Roberts et al., 2020). Leveraging the recent advances in computational resources, a handful of global simulations with atmospheric spatial resolutions between 25 and 50 km are now available and reveal a clear improvement in simulating TCs (Murakami et al., 2015; Walsh et al., 2015; Roberts
 35 et al., 2020; Bourdin et al., 2024).

To study TCs simulated by global climate models, we need algorithms to objectively detect them. Such algorithms, known as TCs trackers, are traditionally based on physical and dynamical properties of cyclones (see Zarzycki and Ullrich, 2017; Bourdin et al., 2022, and the reference therein for details about different trackers). These algorithms identify tropical cyclone points and connect them to reconstruct TC tracks employing thresholds applied to variables. Depending on the variables involved in the
 40 tracking process, Bourdin et al. (2022) defined two categories of trackers: physics-based and dynamics-based. Physics-based trackers rely on thermodynamic properties of a tropical cyclone, such as the local minimum sea-level pressure combined with a warm-core (temperature anomaly or a geopotential thickness). Dynamics-based trackers rely on dynamical variables such as vorticity or other derivatives of the velocity. Both usually include an intensity criterion to discriminate the systems.

The thresholds used in these trackers that are tracking-scheme specific and subjective, and may also depend on the particular
 45 TC formation basin as well as on the TC categories (Camargo and Zebiak, 2002; Befort et al., 2020). This may lead to a potential inability of tracking schemes to generalize to other domains or data from sources other than those used to calibrate the thresholds (Raavi and Walsh, 2020). In order to avoid subjective choice of thresholds and make the tracker more flexible in identifying cyclonic situations, this study focuses on data-driven algorithms, the so-called machine learning algorithms. Indeed, these types of algorithms rely on data to identify cyclones based on different combinations of variables, independent
 50 of user prescribed thresholds.

For instance, the detection skills of three machine learning approaches (Decision Trees, Random Forest, Support Vector Machines) and a model based on Linear Discriminant Analysis have been compared for satellite data in Kim et al. (2019a). Machine learning approaches showed better skill hit rates, while the linear approach showed lower false alarm rates. Among machine learning approaches, neural network based deep learning approaches have lately gained attention for TC detection.
 55 They have been implemented through segmentation, edge detection, circle fitting, and comprehensive decision for satellite images (Kumler-Bonfanti et al., 2020; Wang et al., 2020; Nair et al., 2022). Kim et al. (2019b) leveraged a Convolutional Long Short-Term Memory network to detect and forecast hurricane trajectories on Community Atmospheric Model v5 simulation data.

However, these approaches use satellite and model data that can be biased and are mainly focused on shape detection in
 60 images. As such, they are not comparable to the physics-based trackers previously mentioned, which have been developed



from reanalysis and evaluated with respect to observed data and focus on TC-related physical processes. For instance, in Bourdin et al. (2022), trackers were applied to the fifth generation of ECMWF Reanalysis (ERA5, Hersbach et al., 2020) and evaluated with respect to the observational record of the International Best Track Archive for Climate Stewardship (IBTrACS, Knapp et al., 2010). In that context, Gardoll and Boucher (2022) and Accarino et al. (2023) have developed convolutional
 65 neural network (CNN) architectures to detect cyclones. They respectively used eight and six 6-hourly variables extracted from ERA5 in their CNN. The added value of such approaches is the ability to constrain the detection with cyclone record provided by IBTrACS, by associating each set of 6-hourly data field to the occurrence (absence or presence) of a cyclone (called labelling in machine learning). This makes tracking TCs a supervised classification problem.

The issue with the use of the latter type of algorithms in the case of TC detection is that the number of TC-related atmospheric
 70 situations is very low compared to TC-free situations. These algorithms trained with such data will learn from a larger diversity of TC-free situations and thus will be more accurate and inclined to rule for TC-free situations and, therefore, biased towards them. In addition, neural network based algorithms need, by essence, huge amounts of data, which can sometimes be qualified as data greedy. This calls for strategies to equilibrate TC/TC-free ratio while keeping enough data to obtain a robust tuning of the CNN. Gardoll and Boucher (2022) reduced variable fields in the North Atlantic to $8^{\circ} \times 8^{\circ}$ windows around the eye of
 75 the cyclone for every time step with a TC and sampled for each one of these windows two TC-free which drastically reduce the data sample (28 521 images). This potentially leads to overfitting and limits the generalizability of the tracker by reducing the diversity of TC-free situations and the spatial variability of the potential TC location due to the choice windows around past TC locations. Only binary properties (TC/TC-free) of the tracker were evaluated in Gardoll and Boucher (2022). Accarino et al. (2023) considered non-overlapping $10^{\circ} \times 10^{\circ}$ windows over the whole joint North Pacific and Atlantic basins and rather
 80 opted for a data augmentation procedure of TC situations to reach a 50/50 ratio (425 358 images). Their CNN-based tracker produced comparable performance as the physics-based trackers in terms of TC track detection but generated larger numbers of false alarms, which is not desirable. Finally, this type of method processes large amounts of data, which calls for huge computing power (typically GPU in Gardoll and Boucher (2022) and High Performance Computing infrastructure in Accarino et al. (2023)). Both studies did not provide physical interpretation of the performance of the tracker.

85 In this study, the objective is to leverage and adapt a well-known and proven method, the Random Forest (RF, Breiman, 2001), to a TC tracking problem by associating a given atmospherical situation described by a limited set of predictors to the presence or the absence of TCs. This choice has been made by considering computing cost, the need for several meteorological variables, and the ultimate goal of such a tracker being the tracking of TC in future climate simulations. Indeed, having many variables implies potential overfitting and impeded interpretation of the results and lower transferability to future climate
 90 simulations. Random forest provides interpretation means and lower computational costs. Higher data frugality will be achieved by considering simple variable statistics instead of entire variable fields, which will potentially improve the transferability of the tracking to climate simulations.

The paper is organized as follows: Section 2.1 describes the datasets used and its pre-processing. The RF-based tracker, the experimental setup and the evaluation metrics are described in Section 2.2. Section 3 presents the results of the tracker



95 developed in this paper. Sensitivity tests are provided in Section 4 and Section 5 gives a summary of the main conclusions from
 this work, while also sketching relevant perspectives.

2 Material and Method

2.1 Data

This section describes the data sets on which this study is based.

100 2.1.1 International Best Track Archive for Climate Stewardship, IBTrACS

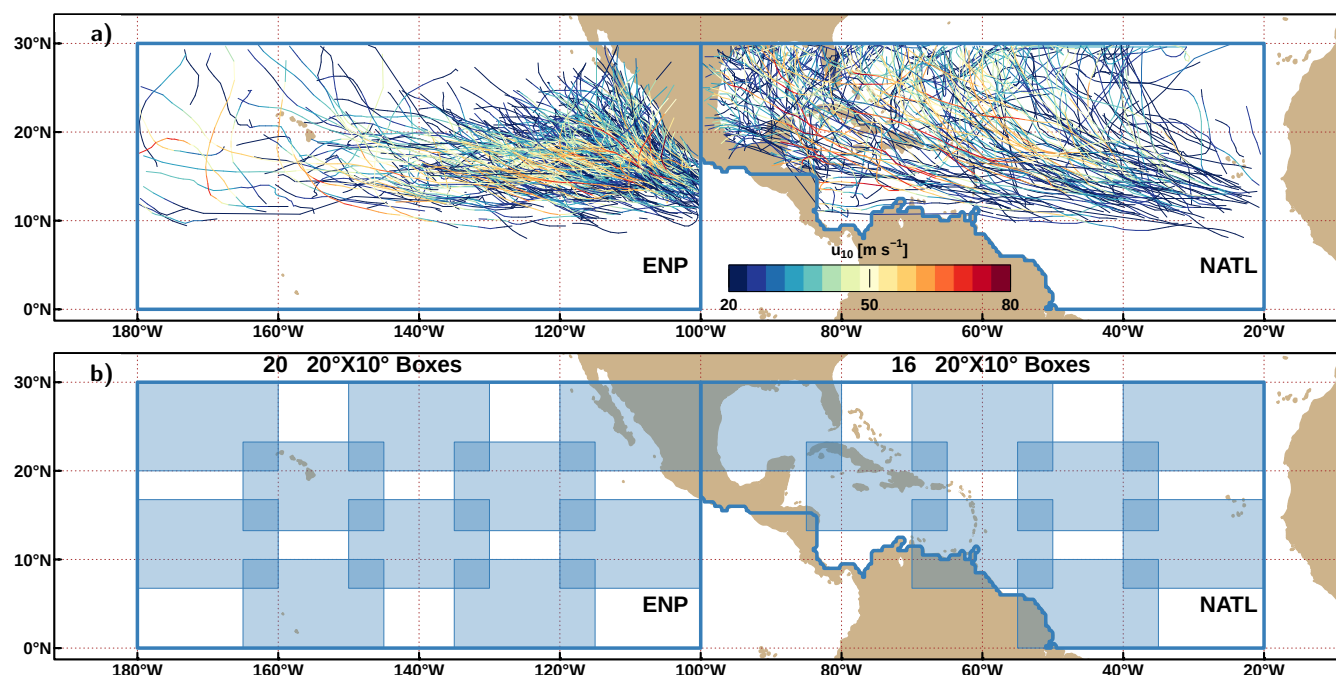


Figure 1. a) Eastern North Pacific (ENP) and the North Atlantic (NATL) basins, with the TCs tracks and the associated wind intensity (in m s^{-1}) used in this study. b) Boxes patching both basin. Only one out of every two box are pictured for clarity reasons.

In this study, the IBTrACS “since 1980” set (Knapp et al., 2018) was retrieved. In the following, two basins are going to be considered: the eastern North Pacific (ENP) and the North Atlantic (NATL) (cf. blue contours in Fig. 1 a)). In these basins, the U.S. National Hurricane Center (NHC) is responsible for reporting tropical cyclones’ best tracks. In total, 42 cyclonic seasons between 1980 and 2021 (from June to November in the Northern Hemisphere) are considered. At the time of this study, tracks in 2022 and later are removed, since some of them are still labelled provisional. Those labelled “spur” are also removed. Track records that are not providing maximum wind and minimum pressure are also removed. Finally, tracks that do not reach the Tropical Storm (TS) stage are also removed. This is decided according to the storm category given by the



values of the minimum sea level pressure $P_{\min}$ and the 10 min near-surface sustained wind u_{10} . Based on Table 2 of Bourdin et al. (2022), TS stage is reached when $P_{\min} \leq 1005$ hPa and $u_{10} \geq 16\text{ms}^{-1}$. Tropical Cyclones (TC) stage is reached when
110 $P_{\min} \leq 990$ hPa and $u_{10} \geq 29\text{ms}^{-1}$. Once processed, the ENP and NATL basins respectively contain 514 and 431 tracks at a 6-hourly timestep.

2.1.2 ERA5

Our main objective is to associate climate variables and determine the main drivers that contribute to the formation and strengthening of TCs during their lifetime. Hourly estimates of atmospheric variables are provided by ERA5 at $0.25^\circ \times 0.25^\circ$ from 1979
115 to the present day. While having similar performances as JRA-55 or NCEP-CFSR for a range of metrics (Zarzycki et al., 2021; Roberts et al., 2020), ERA5 does not perform any specific assimilation for TCs (Zarzycki et al., 2021). This motivated our choice to use ERA5 to evaluate the tracker developed in this paper. 6-hourly data from 1980 to 2021 are extracted, consistent with the period of the IBTrACS data. The choice of using 6-hourly data stems from the overall objective to use the tracker on climate model simulations, whose output is very rarely provided at higher temporal resolution. Five variables have been
120 extracted from ERA5:

- the mean sea level pressure, MSLP (in hPa),
- the 10-m wind intensity, UV10 (in m s^{-1}),
- the total column water vapour, TCWV (in kg m^{-2}),
- the relative vorticity at 850 hPa pressure level, RV850 (s^{-1}),
- 125 – the geopotential thickness between 300 and 500 hPa pressure level, THZ300_Z500 (in m).

These variables have been selected based on their ability to characterize specific physical properties of TCs and on their wide availability in climate model simulations output. In particular, TCs have a warm core, with the most intense winds found close to the surface. TCs are structured with an eye at the centre, an eyewall, and spiral convective rainbands around them. TCs are driven by diabatic processes, meaning that their energy comes from extracting oceanic moisture that releases latent heat once
130 condensed in the upper troposphere. Considering this, MSLP characterizes the spatially coherent low-pressure structure (the eye and the eyewall), UV10 the strong surface wind, TCWV the moisture and the potential for rain, RV850 the TC vortex and THZ300_Z500 the upper-level warm core associated with the local depression in the TCs.

2.1.3 Data-set preparation

Several steps are followed to prepare the data. First, both basins are patched by $20^\circ \times 10^\circ$ overlapping boxes (see shaded blue
135 boxes in Fig. 1b)), totalling 20 and 16 boxes respectively for ENP and NATL. This is done in order to deal with cases where two or more TCs occur at the same time in a given basin. Then, for every box, a vector of zeros and ones is constructed: every



timestep, 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. Thus, the TC tracking problem is handled as a binary classification problem.

Then, ERA5 predictors associated to these binary vectors are built as follows: instead of considering the whole variable field within a box, only four statistics of that field are considered: minimum, mean, maximum and standard deviation. Thus, for a given timestep, the atmospheric situation within a box is described by a set of 20 predictors (5 climate variables $\times$ 4 statistics). Those predictors are labelled with the physical variable name attached to the statistic corresponding suffix (min, mean, max, sd). For instance, MSLPmin, MSLPmean, MSLPmax and MSLPsd are obtained for MSLP. Finally, for a given basin, the binary vector and the associated set of predictors of every box are concatenated and standardized. A table, with about 600 000 and 490 000 rows, are respectively obtained for ENP and NATL.

2.2 Methods

In this section, the description of (i) the random forest based TC tracker and the experimental set-up of the calibration/validation/test, (ii) the reconstruction of TC tracks, (iii) the evaluation metrics and (iv) the tracker used as benchmark are given.

2.2.1 Ensemble Random Forest for unbalanced data and experimental set-up

Random Forest (RF, Breiman, 2001; Hastie et al., 2009) is a supervised machine-learning algorithm based on generating an ensemble ("forest") of decision trees grown in parallel, referred to as bagging in machine learning. Each decision tree in the forest separates the target variable into homogeneous groups according to a sequence of *if-else* decision rules applied to the predictors. In our binary classification framework, each new separation according to a decision rule between the nodes have been performed via maximal impurity reduction, using the Gini index as an impurity function (Breiman et al., 1984). A random subset of data is provided for each tree (the in-the-bag dataset), and a random subset of covariates is tried at each node in each tree, bringing robustness to the classification. In this paper, such an implementation of RF is provided by the R package "ranger" (Wright and Ziegler, 2017) and follows the approach developed in Malley et al. (2012) to obtain the probabilities of a diagnosis of diabetes or appendicitis given sets of medical tests. Each classification tree gives a probability on the 0/1 class of a datum by taking the majority vote in a terminal node. The average of the tree probability estimate gives the forest probability estimate for class occurrence for each datum. All the hyperparameters are set to default values.

In the case of TC tracking, an imbalanced data problem presents itself. Indeed, the class "presence of TC" is under-represented with only 2.1% (resp. 2.6%) of the data for NATL (resp. ENP). This results in low-performing RFs due to two phenomena: (i) successive partitioning of the data when growing the decision trees causes them to 'see' fewer and fewer of the rarer class, thus fitting more and more to the majority class ("absence of TC"); and (ii) interactions between covariates can go unlearned by the decision trees due to the sparseness of the data induced by partitioning (He and Garcia, 2009). Kuhn (2013) discussed resampling methods that can resolve class imbalances, but there is little consensus on the best approach. Siders et al. (2020) compared different approaches and showed that combining the subsampling of the majority class with an ensemble of random forest (ERF), *i.e.*, the use of multiple random forest with different subsampling of that majority class gave the best performance.



170 In this study, the ERF approach is leveraged to tackle the class imbalance issue. The subsampling of the majority class is performed by setting the number of zeros as n times the number of ones. Several setups are tested with $n \in \{10, 15, 20, 25, 30, 35\}$ and referred to as 'n-times' setup and one setup is referred to as 'FULL' without subsampling. In order to evaluate the effect of the subsampling, for each n , 100 RFs are performed. Three experiments are set for each basin:

1. Calibration experiment: one calibration of the ERF is made using the whole data during the 1980-2021 period and all
 175 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 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 the effect of any potential trend and interannual variability in the TC statistics (frequency, intensity) and the changes in IBTrACS
 180 data quality. Most of the ERF evaluations will rely on this experiment.
3. Test experiment: from the calibration performed over the whole time period for a given basin, the TC tracks over the other basin are reconstructed. This is done to evaluate the generalizability of ERF.

Depending on the experiment, setup and basin, the training of one RF took between 1 and 10 minutes when performed on a laptop with an 11th Gen Intel®Core™ i7-1165G7 @ 2.80GHz with height cores and 16 Go RAM and between 30 seconds
 185 and less than 3 and a half minutes when performed on a computing node Intel®Xeon®CPU E5-2650 v2 @ 2.60GHz with 16 cores and 65 Go of RAM (8 Go would be sufficient).

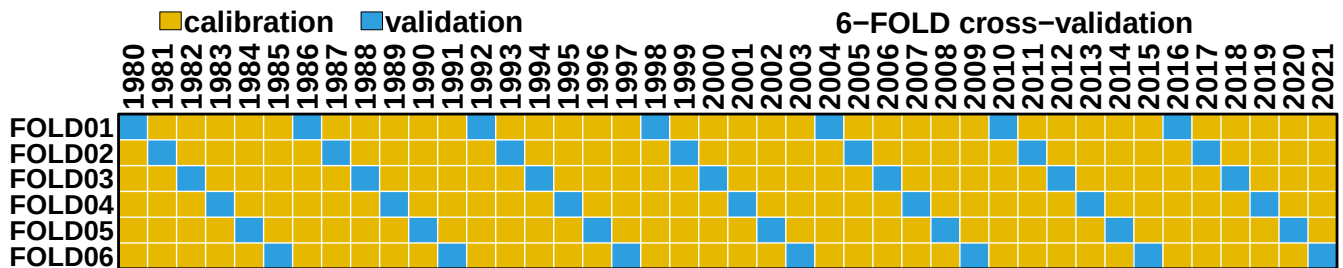


Figure 2. Scheme of the 'Validation experiment': 6-fold cross-validation scheme over the 1980-2021. Yellow years are used for the calibration and blue for validation. One out of six years are used for validation, making seven out of 42 per fold.

2.2.2 Track reconstruction and matching

In a given box, if RF gives probability of TC above 0.5, the location of the TCs is estimated by the position of the minimum of MSLP in that box. From there, tracks are reconstructed from one TC location to the next. 24 hours gaps within a radius of 450
 190 km are allowed during the reconstruction. A track is kept only if it lasts at least 24h.

The track-matching procedure used in this study is similar to the one in (Bourdin et al., 2022). Let us consider, at time t_i , a point d_i of a detected track D . It is associated to the closest points of a given observed track O at each time t_i that is located



closer than 300 km (with a possibility that such a point does not exist). Points of O that have been associated with any point of track D is denoted as O_D -paired. It is composed of NOD elements. There are four possibilities:

1. $NOD=0$: None of the points of D has been paired to a point in O and D is considered to be a false alarm (FA),
2. $NOD>0$ and all the points in O_D -paired belong to the same observed track O : D is a match for O and considered a hit (Hit),
3. $NOD>0$ and all the points in O_D -paired belong to multiple observed tracks: D is a match for the observed track having the largest number of paired points and considered a hit (Hit),
4. None of the points of a given O has been matched: O is a miss ($Miss$).

To complete the matching, a final treatment is performed: if an observed track is paired with two or more detected tracks, these detected tracks are merged into a single track. It happens when parts of the same observed tracks are detected separately due to the filtering consisting in coding 0 every timesteps in the observation IBTrACS that do not reach TS intensity.

2.2.3 Evaluation metrics and analysis

The first evaluated aspect is the performance of ERF in terms of binary classification. For that, the Matthews correlation coefficient (MCC, Matthews, 1975) is used as a measure of the quality of binary (two-class) classifications. It has been introduced by Yule (1912) and its values range from -1 to +1. A score of 1 represents a perfect prediction, 0 an average random prediction, and -1 an inverse prediction. The MCC is particularly useful when the classes are imbalanced, as it accounts for the imbalance in the calculation. It is similar to the Pearson correlation coefficient in its interpretation. The MCC is more informative than other metrics in evaluating binary classification because it takes into account the balance ratios of the four categories of the contingency (or confusion) matrix: true positives (TP), true negatives (TN), false positives (FP), false negatives (FN) (Chicco and Jurman, 2020). The MCC is computed from the confusion matrix (see Table A1):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

The second aspect evaluated is the ability of ERF to reproduce observed TC tracks. Once all tracks are labelled Hit , $Miss$, and FA two detection skills metrics are defined, the Probability of Detection (POD, sometimes referred to as “Hit Rate”) and the False Alarm Rate (FAR): $POD = \frac{Hit}{Hit + Miss}$; $FAR = \frac{FA}{FA + Hit}$. POD and FAR are expressed in %, and good performance is achieved when POD is high and FAR is low.

Another aim of this paper is to provide some physical interpretation for the presence or absence of a TC given a certain atmospherical situation. Breiman (2001) proposed to evaluate the importance of a predictor variable (or feature) X_j for predicting Y (here the probability) by adding up the weighted impurity decreases $p_t \Delta_i(s_t, t)$ for all nodes t where X_j is used, averaged over all trees ϕ_m (for $m = 1, \dots, M$) in the forest:

$$Importance(X_j) = \frac{1}{M} \sum_{m=1}^M \sum_{t \in \phi_m} \mathbb{1}(j_t = j) [p(t) \Delta_i(s_t, t)],$$



where $p(t)$ is the proportion $\frac{N_t}{N}$ of samples reaching node t , j_t denotes the identifier of the predictor used for splitting node t and $\Delta_i(s_t, t)$ is the impurity decrease at split s_t . For each one of the 20 predictors, the feature importance is then the
225 contribution in % of each variable to the total reduction of impurity.

Then the idea is to determine the importance of each predictor in the prediction of every single outcome (all zeros and ones) by RF. This is performed by computing the SHapley Additive exPlanation (or SHAP) values with the method proposed by Lundberg et al. (2020) with an implementation for tree-based algorithms provided in the R package “treeshap” (Komisarczyk et al., 2023). The general idea of SHAP values is to explain each outcome of RF as a sum of effect φ_i of each predictor X_i . The
230 SHAP value is φ_i , which stems from a concept introduced in cooperative game theory (Shapley, 1951). The idea is to determine the average contribution of a predictor to a specific prediction (here a probability) to every combination of predictors. This can be written as follows:

$$\varphi_i = \frac{1}{\# \text{ predictors}} \times \sum_{\substack{\text{combinations} \\ \text{excluding } X_i}} \frac{\text{marginal contribution of } X_i \text{ to combination}}{\# \text{ combinations excluding } X_i \text{ of this size}}.$$

Once the SHAP values φ_i for all predictors X_i and for every outcome of the RF forest are computed, SHAP-based partial
235 dependence plots are obtained by plotting φ_i against X_i . These plots will help to interpret the presence of TCs given an atmospherical situation described by a set of predictors X_i and explore the evolutions of TC probability according to the evolutions of predictor X_i .

2.2.4 UZ algorithm

For comparison purposes, we use the UZ algorithm, which is a physics-based detection scheme developed in Zarzycki and
240 Ullrich (2017) and implemented in TempestExtremes (Ullrich et al., 2021). It was shown in Bourdin et al. (2022) to have good detection scores, and in particular a low False Alarm Rate. The UZ scheme is based on a 2-step procedure. The first step is the detection step, where candidate TC points are identified. These candidates are MSLP local minima associated with an upper-level warm core, which is measured by the geopotential thickness between 300 and 500 hPa pressure level. The second step is the stitching step, where candidates are stitched together into tracks. The tracks must be associated with a maximum
245 wind speed of at least 10 m/s over at least 54h. For more technical details, the reader is redirected to (Zarzycki and Ullrich, 2017; Ullrich et al., 2021; Bourdin et al., 2022).

3 Results

3.1 Zero class subsampling choice

As mentioned in the method section 2.2.1, random forest is subject to bias toward majority (here zero) class when applied to
250 unbalanced data. In this section, the results of ERF for different subsampling over the NATL basin are used to select the best one. Figure 3 shows the Validation experiment boxplots of MCC, POD and FAR metrics of the tracks obtained for the 100 RFs of ERF for the validation years for different subsampling of zeros. The purple markers are the metric values for tracks obtained

for the average of the probabilities given by the 100 RFs. The MCCs, given in the top panel of Figure 3, are quite similar for the different sub-samplings. It ranges from a little below 0.74 for the FULL setting to a little above 0.75 for the 20-times setting, with very little difference between the 15-times and 25-times setting.

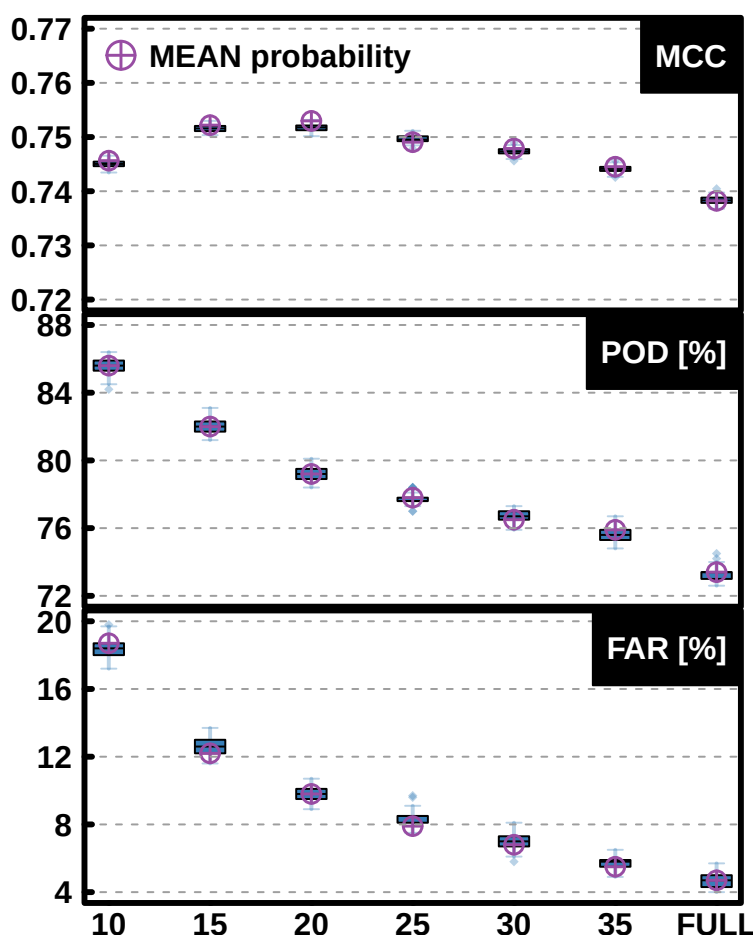


Figure 3. Validation experiment boxplots of MCC (*top*), POD (*middle*) and FAR (*bottom*) obtained over validation years obtain for the 100 RFs of different ERF with different subsampling of zeros, for NATL basin. Violet symbol represent the metrics for the tracks obtained from the average of the probabilities given by the 100 RFs of ERF.

POD and FAR metrics are respectively given in the middle and bottom panels of Figure 3. The POD decreases almost linearly from 85% to 73% from the '10-times' to the 'FULL' setting. Similarly, the FAR also decreases from 18% to 5%. This indicates that a good ability of detecting TCs goes along with a high level of generating false alarm. This also explains the similar MCC metrics for the different settings, indicating some sort of compensation between the four categories of the confusion matrix. The subsampling '25-times' setup has the *medium* performance, with POD around 78% and FAR around 8% (see Fig. 3), is chosen.



The other result drawn from Figure 3 it that the effect of the sampling of zeros given an n-times setup on MCC, POD and FAR is very marginal considering the very narrow boxplots. It means that even though the tracks reconstructed from the average probability obtained from the 100 RFs are used in the following of this study, a lower number of RFs would be sufficient.

265 3.2 ERF detection analysis

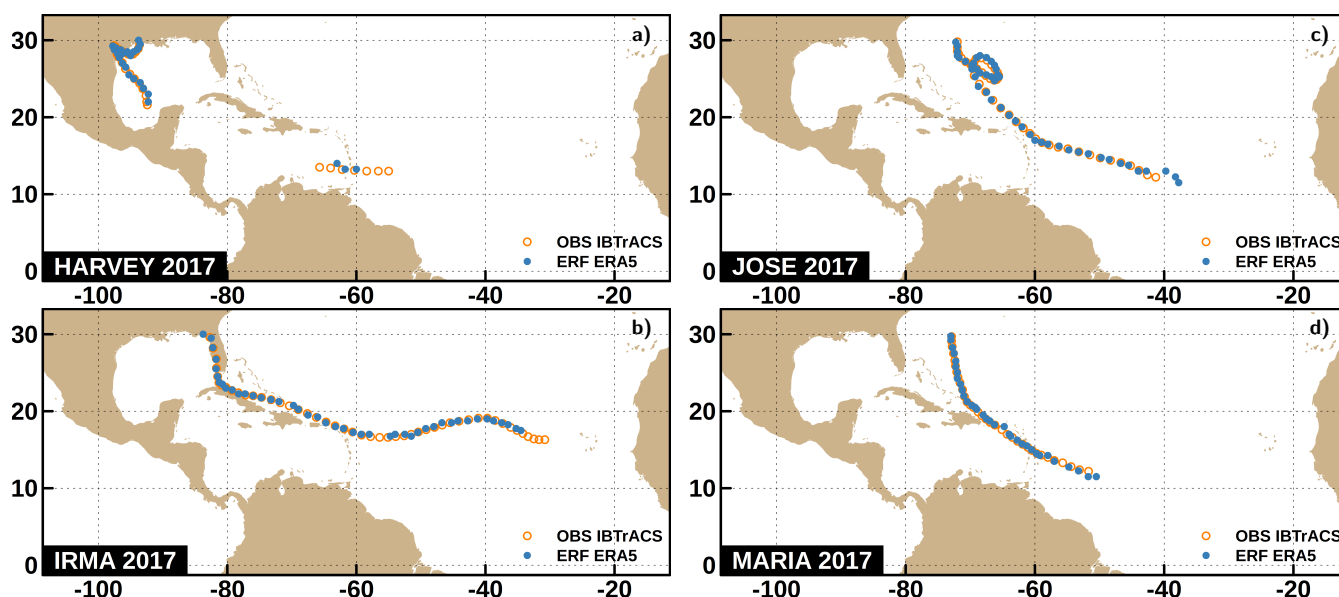


Figure 4. ERF-based TC tracks reconstructed over validation year 2017 for the NATL for the '25-times' setup. a) Harvey, b) Irma, c) Jose and d) Maria.

Figure 4 shows four example of TC tracks reconstructed over the NATL basin for the validation year 2017, from the average probability obtained from the 100 RFs with the '25-times' setup. Similar reconstructed tracks for the ENP basin with similar ERF setup are shown in Figure S1 of the supplementary material. TC tracks reconstructed from ERA5 with ERF are very close to the observed tracks from IBTrACS even though the trajectories have very different shapes. Note the long gap in the Harvey cyclone (Fig. 4a)) is due to the filtering consisting in selecting only the time steps reaching TS intensities (see Sect. 2.1.1). Table 1 gives the POD and FAR metrics for the tracks reconstructed from average probability from ERF for calibration, validation and test experiments for the '25-times' setup for both basins. For the validation experiment, POD are respectively of 77.5% and 77.8% for ENP and NATL basins. FAR are respectively of 8.7% and 7.9% for ENP and NATL basins. Note that for calibration experiments, POD are above 90% and FAR is around 2% for both basins.

275 In the following, the statistical and physical properties of the detected tracks are investigated. Figure 5 a) shows the track duration histograms for the observed, ERF-detected tracks (Hit and FA) and missed tracks. ERF-based Hits have duration distribution quite similar to IBTrACS tracks, but with substantial differences for short duration tracks (1 to 3 days). These short duration tracks have typically short lifespan and are of lower intensity. This discrepancy is supported by the Miss tracks duration



Table 1. POD and FAR for tracks reconstructed from average probability from ERF for calibration, validation and test experiments for the '25-times' setup and for UZ in %. Multi-basin refers to POD and FAR from ERF applied to both ENP and NATL basins under '25-times' setup discussed in Sect. 4.1. The right part of the table referred to as 'Ablation experiments' gives the POD and FAR of ERF experiments conducted with a reduced number of predictors discussed in Sect. 4.2

	Main experiments				Ablation experiments			
	ENP		NATL		ENP		NATL	
	POD	FAR	POD	FAR	POD	FAR	POD	FAR
Calibration	91.1	2.0	93.6	2.3	89.7	4.1	89.5	3.6
Validation	77.5	8.7	77.8	7.9	77.2	11.9	76.7	13.2
UZ	76.4	24.1	78.4	15.0	-	-	-	-
Test	NATL (calib.)		ENP (calib.)		NATL (calib.)		ENP (calib.)	
	76.6	15	68.4	7.8	73.5	15.5	69.3	15.5
Multi-basin								
Calibration	91.8	3.1	90.9	2.4	90.5	5.5	87.8	3.1
Validation	79.2	9	74.8	5.3	78.1	13.7	70.1	8.3

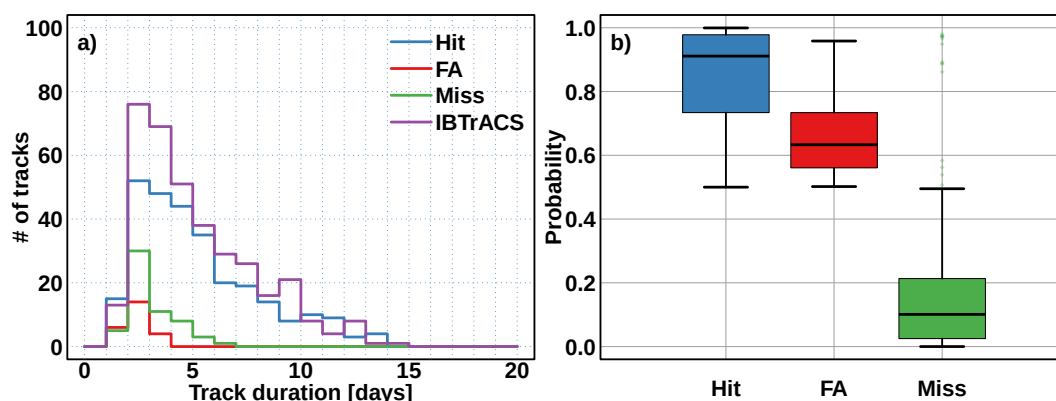


Figure 5. Statistical properties of IBTrACS (purple), ERF-detected TC tracks: Hit (blue) and FA (red) and ERF-missed tracks: Miss (green). a) TC tacks duration histograms, b) Boxplot of ERF-based average probabilities for associated to each time steps of Hit, FA and Miss tracks.

distribution, which are mainly short durations tracks (the majority of them last between 2 and 4 days). False alarm are also of the same durations. Differences in probabilities of TCs given by ERF associated with every time step of Hit, FA and Miss of tracks are then investigated. Figure 5 b) show that probabilities associated to Hit tracks (median above 0.9) are substantially different compared to those associated to FA (median 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.

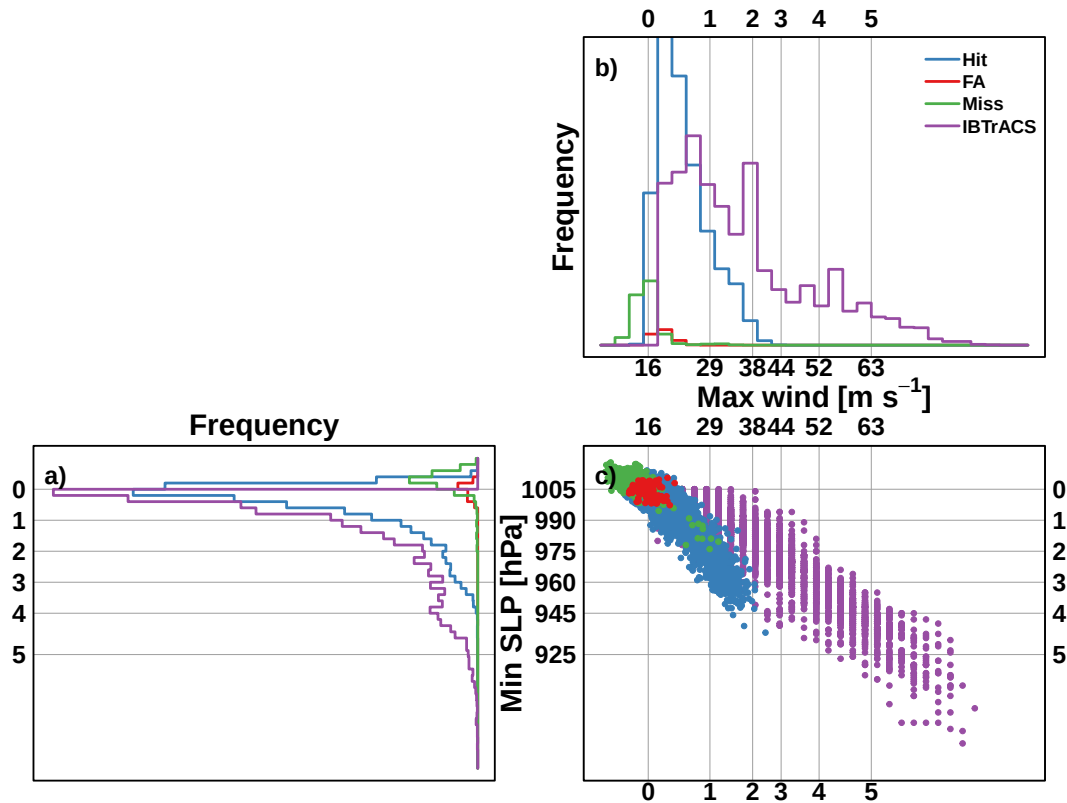


Figure 6. Physical properties of IBTrACS (purple), ERF-detected TC tracks: Hit (blue) and FA (red) and ERF-missed tracks: Miss (green). a) histograms of minimum sea level pressure [in hPa], b) histograms of maximum surface wind [in m s^{-1}] and c) the scatterplot of maximum surface wind against minimum sea level pressure. Vertical lines indicate the TC intensity classification Saffir-Simpson Hurricane Scale thresholds of 10 min sustained wind. Horizontal lines indicate pressure thresholds based on Klotzbach et al. (2020).

285 In order to investigate how these different tracks diverge in nature, the maximum wind and minimum sea level pressure associated to these different types of tracks are considered. Figure 6 c) shows the scatterplot of maximum surface wind against minimum sea level pressure associated to every timestep of observed, detected (Hit and FA) and missed tracks. Figure 6 a) and b) respectively give the associated sea level pressure and maximum wind histograms. In general, and as already pointed out in Bourdin et al. (2022) and Dulac et al. (2024), the wind-pressure relationship in ERA5 is different from the one in the observations (purple dots versus the rest). Detected TCs are weaker than observed ones. In particular, Hit tracks barely reach category 4 when considering ERA5 minimum sea level pressure, and it is even worse when considering ERA5 maximum wind Hit tracks barely reach category 3. In addition, these figures also provide insight on FA and Miss tracks. Miss tracks are, for the majority of them, associated in ERA5 with minimum pressure above 1005 hPa and maximum wind below 16 m s^{-1} which are the TS intensity threshold. It means that these tracks are missed because ERA5 is failing to represent these TCs correctly.

290 Concerning FA tracks when examining Figure 6 a-c), the maximum winds and minimum pressure are located around 16 m s^{-1}

295



and 1005 hPa pressure which are again the threshold for TS intensity. Thus, these FA tracks may be related to the uncertainty of ERF which associates an atmospheric situation to a TC even though none has been observed.

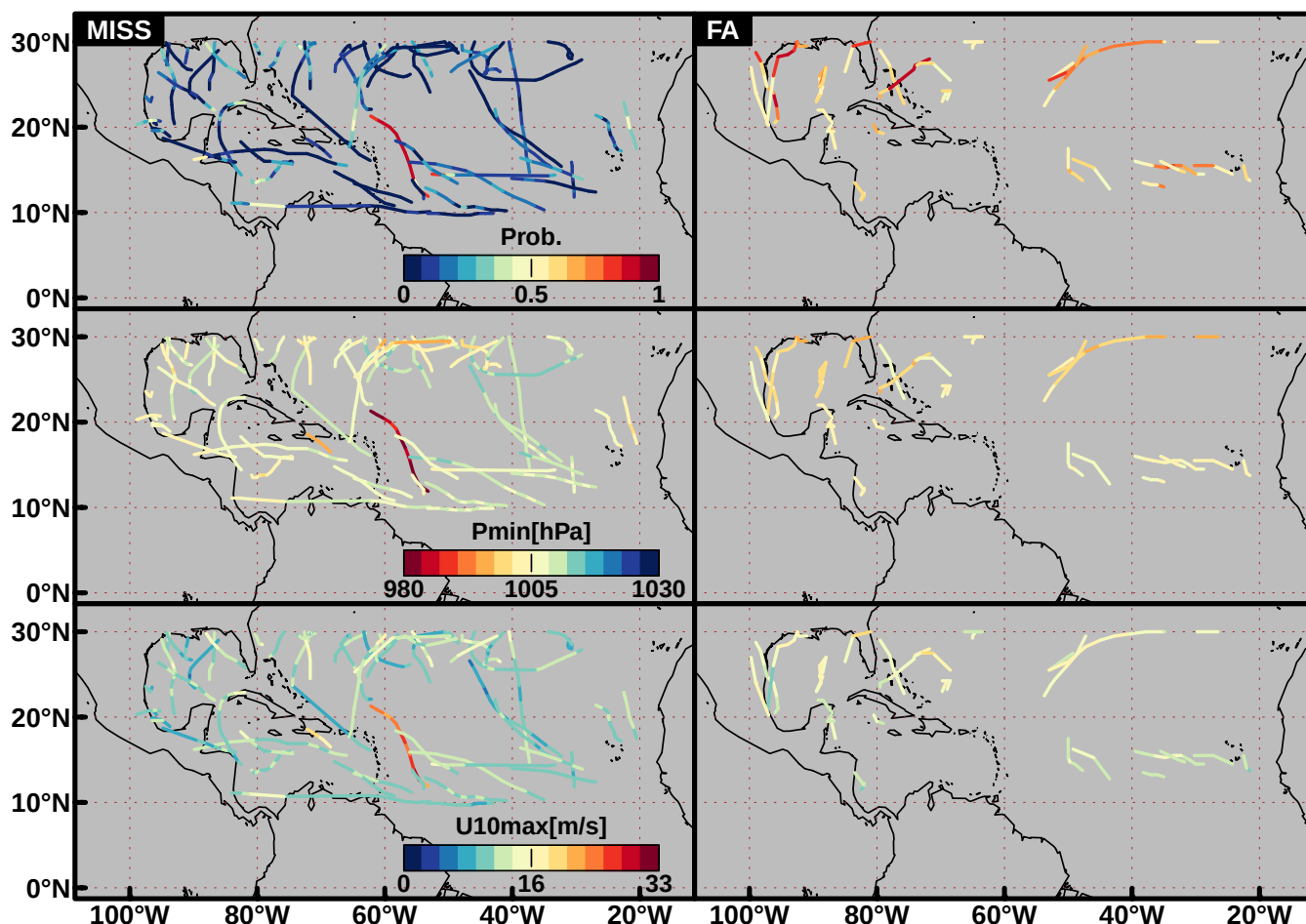


Figure 7. Average probability (top row), minimum sea level pressure, P_{min} [in hPa] and maximum wind U_{10max} [in $m s^{-1}$] for Miss tracks (left column) and FA tracks (right column). Colours are saturated for P_{min} and U_{10max} .

Figure 7 shows Miss and FA tracks distributed over NATL basin and the associated ERF-based average probability, the minimum pressure and the maximum wind. Miss tracks are distributed over the entire domain and confirm the results of Figures 5 and 6 in terms of low probability and intensity. However, one track shows high probability, intense wind and low pressure and is pictured in reddish colours in the three left-hand side panels of Fig. 7. This track illustrates one small drawback of dividing the basin into $20^\circ \times 10^\circ$ overlapping boxes: ERF is only able to detect one TC at a time within a box. However, it is possible that two TCs happen at the same time within one box. Figure 8 shows the IBTrACS track of the TC IRIS spotted in Figure 7 and the stronger TC HUGO occurring at the same time. The probability, the pressure and the wind associated to the missed TC IRIS in Figure 7 are actually those of the strong TC HUGO. The FA tracks are rather distributed at the edge of the



domain. In particular, they are located in areas where TCs are typically weaker: the main development region (eastern part of the domain between 10°N and 20°N) where TC are developing, and coastal areas where they disappear. This can be related to the uncertain aspect of these tracks that are yielded by lower probabilities.

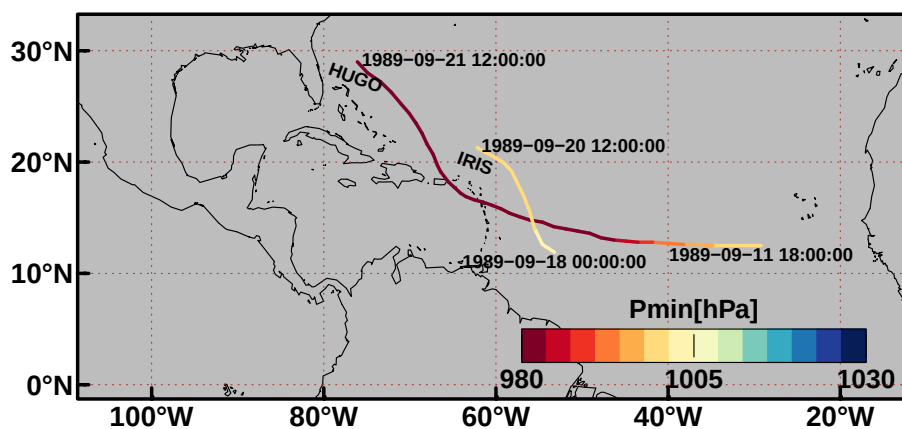


Figure 8. Example of TC IRIS that has been missed by ERF due to the presence of a stronger TC HUGO and the associated Pmin [in hPa].

Similar figures for the ENP basin are given in Figures S2 to S5 of the supplementary material and give similar conclusions as for the NATL basin. The major difference being that the distinction in terms of intensity between Hit and FA is less obvious based on ERA5. The wind-pressure relationship basin in ERA5 compared to the observation is even worse for ENP where TCs barely reach category 2 intensity for wind and pressure scale. The median probability of Hit and FA tracks are closer (0.8 vs 0.65) and yield a higher FAR ratio. Even though the majority of FA tracks are associated with wind and pressure around 16 m s⁻¹ and 1005 hPa, some of them present more intense values. One hypothesis may be that these tracks have not been recorded in IBTrACS.

3.3 Physical interpretation

In this section, the contribution of the different predictors to the detection of TCs is analysed to provide physical insights into the presence or absence of a TC. Figure 9 represents the boxplot of the gini-based feature importances from the 100 RFs of the ERF for the calibration experiment and the '25-times' setup for both basins.

The figure shows that, for both basins, the six variables with the largest feature importance are the same: RV850sd, MSLPmin, UV10max, RV850max, THz300_z500max and TCWmax. These predictors are physically well-founded in explaining the presence of a cyclone. RV850sd characterizes the singularity of the vortex within a box: the higher it is, the more the TC vorticity stands out from the vorticity of the rest of the area within a box. It is more important than the RV850max, which is the fourth most important variable. Then, UV10max and MSLPmin are the following most important variable. This makes sense, since they are respectively associated to the strong surface winds and the location of the cyclone eye. The following variables are the TCWVmax and THz300_z500max. The former reflects the potential for rain and the moisture lifted by the TC, the latter

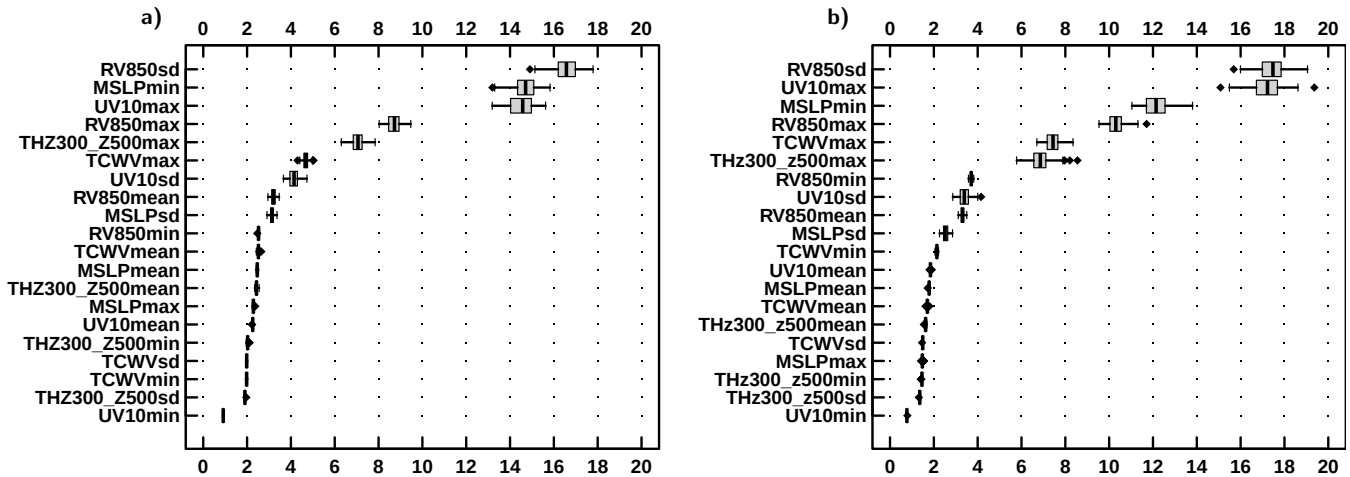


Figure 9. Boxplot of gini-based feature importances of ERF for the calibration experiment and the '25-times' setup for a) ENP and b) NATL.

characterizes the upper-level warm core associated to the TC. Note that the order of importance is slightly different between the basins. For instance, maximum wind is more important than sea level pressure for NATL basin, while it is the opposite for ENP basin. It may be the result of the different wind-pressure relationships between both basins (see Fig. 6 c) and Fig. S3 c) of the supplementary.). TCWVmax is less important in explaining the presence of TC situation for ENP basin.

Feature importance is interesting to quantify the average contribution of a given predictor in discriminating TC from non-TC situations. However, it would be interesting to determine the contribution for each predictor over each outcome of an RF. This is provided by the SHAP-based partial dependency plots shown in Figure 10 for NATL basin. This figure pictures the relationship between the six top predictors (according to feature importance) and their respective SHAP values. Note that given the large computing time necessary for SHAP values, there are only computed for one RF among the 100 RFs of the calibration experiment and the '25-times' setup. These partial dependence plots are probably very similar for the 100 RFs given the small dispersion of the MCC, POD and FAR performances' metrics (see Fig. 3) and feature importance (see Fig. 9). Let us consider the partial dependency plot in panel a) of Figure 10. On the abscissa is given the physical range of RV850sd and the associated SHAP values on the ordinate. It shows the contribution of RV850sd given its value to the probability value of TC occurrence.

The partial dependence show a clear distinction between the "zero" and the "one" populations with marginal overlap. For the zeros, the SHAP values are always very close to 0 while for the ones the SHAP values always steeply increase when the associated predictor increases (except MSLPmin, SHAP values increase when it decreases). For these six predictors, SHAP values tend to reach at cap value after the predictors reach a certain level, and they even decrease for MSLPmin and UV10max. This means that the contribution of these predictors in discriminating TC from non-TC situations does not change when a certain intense value is reached. This figure also shows us that TC situations can occur for a very large range of values and, therefore, potentially very diverse combinations of these predictor values. This advocates for TC tracking approaches that bring more flexibility than threshold-based approaches. Similar results can be drawn based on the Figure S6 of the supplementary

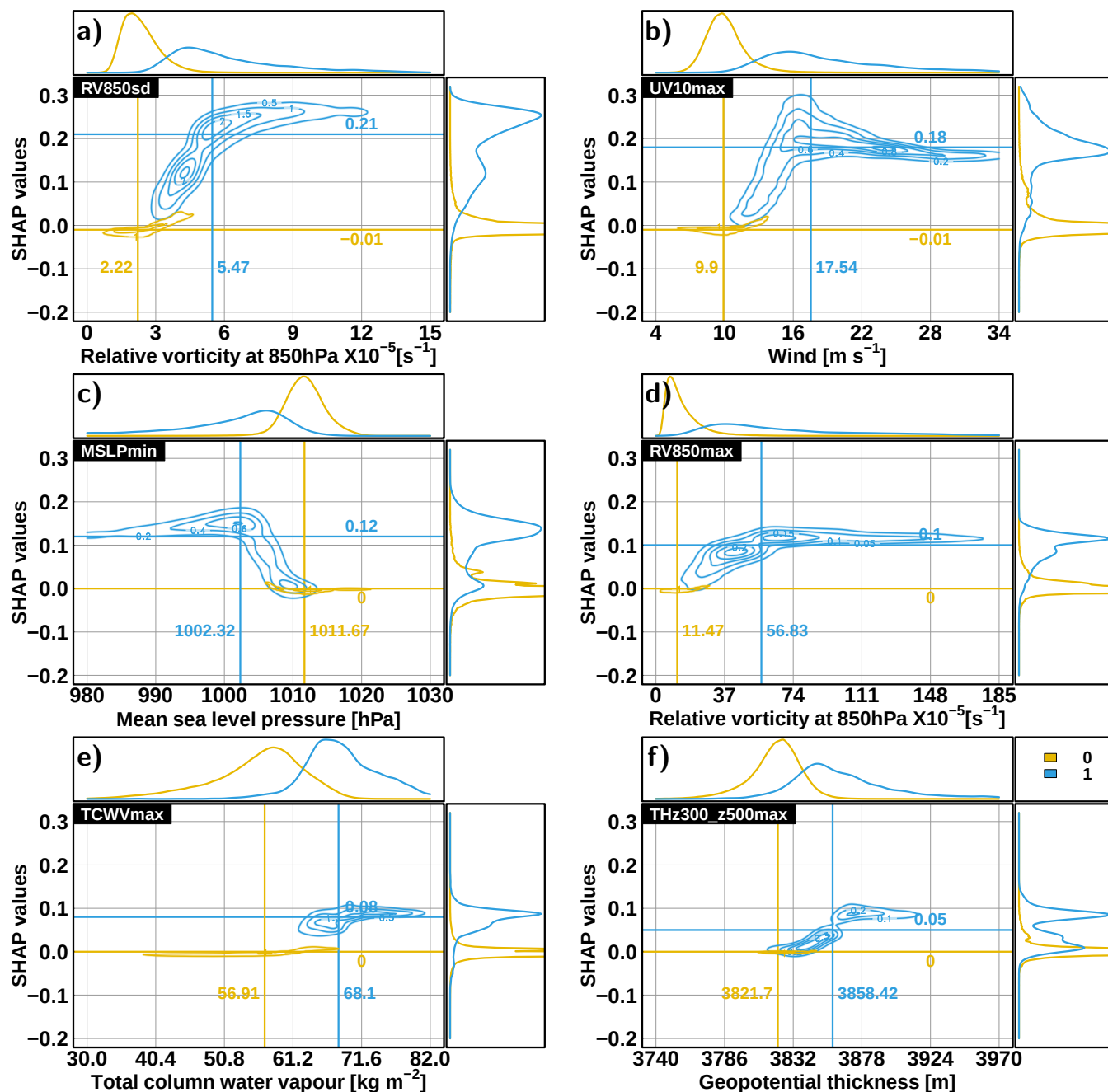


Figure 10. Partial dependence plot for top six predictors a)-f) obtained for one of the 100 RFs of the calibration experiment and the '25-times' setup for NATL basin. Contour lines indicate the density of the scatter plot between one predictor and the associated SHAP values. Yellow and blue characterize respectively the density of the zeros (probability<0.5) and the ones (probability>0.5) population. Vertical and horizontal lines respectively indicate the median of the predictors and the associated SHAP values for both population. The distributions of the predictors and SHAP values are also given conditionally to both population.



material for the ENP basin. This analysis highlights the advantages of using Random Forest (RF) over traditional trackers that rely on sequential thresholding, as most variables exhibit nonlinear interdependencies.

3.4 Comparison with UZ

POD and FAR for UZ in both basins are reported in Table 1. POD are quite close to those of ERF (Validation experiment) with 1% lower POD for ENP (76.4%) and less than 1% higher (78.4%) for NATL. However, FAR, which reach 24.1% in the ENP and 15% in the NATL amount to almost respectively three and two times the ERF scores, which not desirable. In order to explore the similarities and differences between UZ and ERF, Figure 11 shows the number of tracks that are common for observations (IBTrACS, IB) and tracks that are detected by ERF (Validation experiment and '25-times' setup) and UZ. A large portion of tracks of IBTrACS are detected by both UZ and ERF (335/252 for ENP/NATL basins) and some tracks are only detected by UZ (28/31 for ENP/NATL basins) or ERF (27/22 for ENP/NATL basins). In total, Hit numbers are similar in both methods. UZ produces a larger number of FA (115/50 for ENP/NATL basins) than ERF (41/31 for ENP/NATL basins). Finally, there are no common FA between ERF and UZ. This means the FA appear for different reasons in the two methods.

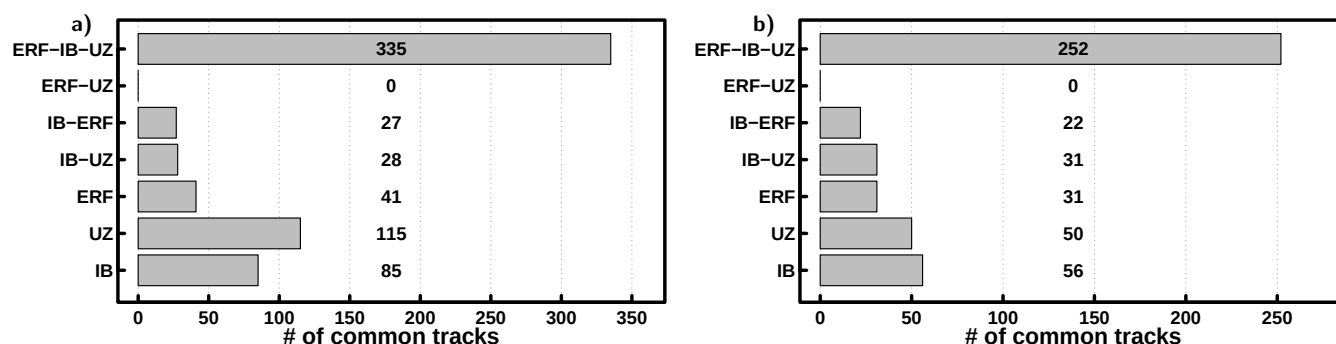


Figure 11. Number of tracks in common between observed tracks (IBTrACS, IB) and detected tracks from ERF and UZ for the following case: all three datasets, two out of the three and tracks specific to a dataset for the Validation experiment and '25-times' setup. a) ENP and b) NATL basins.

4 Sensitivity test

4.1 Need for regional tracker

In this study, ERF has been applied separately for each basin, which is uncommon in the literature in which, usually, the tracking algorithm is applied over multiple basins (Bourdin et al., 2022; Accarino et al., 2023). This has been done in order to test the spatial generalizability of the ERF approach. This ability of ERF is based on the test experiment, which consists in reconstructing the tracks of the ENP basin using the ERF fitted for the calibration experiment ('25-times') setup for the NATL basin and vice versa (see sect. 2.2.1). Table 1 reports the FAR and POD for these test experiments. POD and FAR are



respectively 76.6% and 15% for ENP basin and 68.4% and 7.8% for NATL. In the case of ENP, the test POD is similar to the validation one, while FAR is degraded. It is the opposite for NATL. This shows a certain specificity of the TC tracking according to the basin, which may stem from different factors. For instance, the latitudinal distribution of the cyclones are quite different between both basin: for ENP basin, TC tracks are mostly located between 10°N and 20°N (see Fig. 1) while they are distributed all over the basin for NATL. Thus, this may involve different processes between TC in NATL and ENP. Differences of feature importance in Figure 9 between both basin may be an illustration of that. This can also be due to the differences of the quality of TC representation between both basin in ERA5. For example, differences in the wind-pressure relationship between both basin illustrate this quality difference (see Fig. 6 c) and Fig. S3 c)).

In order to highlight the need for regional tracker, ERF has been carried out under the '25-times' setup for the data pooled from both basins, referred to as 'Multi-basin'. Table 1 gives the POD and FAR for calibration and validation experiments. Compared to basin-specific experiments, the performances are close, with slightly better POD and slightly worse FAR for ENP basin. In contrast, it is the opposite for NATL basin with larger differences. The better performance for the ENP basin can be explained by the higher weight it holds on the ERF training, given its larger data size (see Sect. 2.1.3) and the small total number of TCs. The results remain nevertheless better than UZ (see Tab. 1). It is therefore up to the user to decide if either one ERF for all basins or one specific ERF for each basin is necessary by considering if the loss of performance of Multi-basin ERF compared to the regional ERF is acceptable or not.

4.2 Ablation experiment

An ablation experiment is conducted in order to get a more parsimonious ERF by reducing the number of predictors. Based on the feature importance (Fig. 9), the top six predictors are kept (see Sect. 3.3). Such a model is expected to generalize better, have a smoother behaviour when looking at partial dependencies, and be potentially more intrinsically interpretable.

In this section, the Calibration and Validation (for regional and Multi-basin) and Test (for regional) experiments have been performed under the '25-times' setup. The right part of Tab. 1 gives the POD and FAR of all experiment performed with a reduced number of predictors. For all experiments, POD are only slightly reduced but remain very close to the POD without the ablation. Interestingly, FAR are more strongly degraded (sometimes doubled) with the ablation. This means that predictors with lower feature importance control FA, indicating us to be cautious when removing predictors. SHAP-based partial dependency plot for validation experiment are given in Figure S7 and S8 for the Validation experiment and both basins. In general, these figures are similar to the one of ERF performed with the full set of predictors. The only difference is the better distinction between "zero" and "one" populations, which can result from the higher FA rate.

5 Summary and perspectives

In this study, we used random forest for tracking tropical cyclones in the eastern North Pacific and North Atlantic basins over the 1980-2021 period by associating atmospheric situations described by five climatic predictors extracted from ERA5 to observational IBTrACS records of tropical cyclones. More precisely, the tracking problem in this paper was equivalent to



performing binary classification over imbalanced data containing substantially more TC-free situations than TC. This imbalance problem was addressed by combining an ensemble of random forests with the subsampling of TC-free situations. Before applying this method, the amount of data fed to it was reduced by considering four statistics of each predictor instead of its whole field (minimum, mean, maximum and standard deviation). In addition, basins were patched by overlapping boxes. In this way, our approach was able to learn the characteristics related to the presence of the TC inside a box, regardless of its position, and allowed us to detect cyclones occurring at the same time within one basin or in both basins (for the multi-basin experiments).

Our data-driven ERF tracker showed good performances for detecting TC tracks: In validation, POD/FAR of 77.5%/8.7% and 77.8%/7.9% were obtained for the ENP and the NATL basins respectively. Compared to the physics-based UZ tracker, used as benchmark in this study, ERF showed similar POD but better (*i.e.*, lower) FAR. UZ was chosen because it was the most accurate among the physics-based trackers (Bourdin et al., 2022), and it was also better than other data-driven trackers (*e.g.*, the deep learning approach in Accarino et al., 2023). ERF has the advantage to require low computing power (see Sect. 2.2.1). Tracks detected by ERF have similar duration frequencies to IBTrACS tracks, except for short (2 to 4 days) and lower intensity tracks (see Figs. 5 and 6). Missed and false alarm tracks are mainly short tracks (1 to 3 days). Detected TCs have weaker intensity in ERA5 than in IBTrACS, due to ERA5 systematically under-estimating TC intensity. So much so that it is likely that some cyclones are missing because they were reanalysed too weak to be detected. False alarm tracks are very close to the tropical storm intensity thresholds and thus illustrate the uncertainties of ERF. These tracks are located in developing and landfall areas of cyclones, where their signal is less clear and more uncertain.

For both basins, the six most important variables for detecting the presence of TCs are the same and characterize the main physical and thermodynamic properties of TC. The order of importance differs between the basins, highlighting potential specificities in the TC patterns and processes depending on the basin. Relying on the SHAP-based partial dependency plots, we showed that TC can be detected through potentially very diverse combinations of predictor values, which brings more flexibility than physics-based approaches that need user-prescribed thresholds.

Two aspects of our ERF trackers were then tested: the spatial generalizability of ERF, and the possibility to reduce the number of predictors. When the tracking was performed in one basin based on an ERF calibration performed on the other basin (resp. on both basins): (i) for ENP the POD is similar (resp. improved) and the FAR is degraded (resp. similar) and (ii) for NATL the POD is degraded (resp. degraded) and the FAR is similar (resp. improved). This shows an overall good ability for spatial generalizability of ERF while showing potential need for regional tracking that can stem from the specificities of TC tracks and the differences of ERA5 quality between both basins. The ablation experiment showed that reducing the number of predictors according to their feature importance does not change (or only very marginally) the POD, but strongly degrades the FAR for all experiments. This showed a certain control on false alarm of the removed predictors.

As future work, the next focus will be on extending the tracking for all the basins at global scale. Some tests will also be necessary to choose the minimum number of RF (lower than 100) and to determine which predictors control FAs and which predictors to remove without degrading the performance of the tracker. From there, the major challenge will be to apply ERF to Earth System Models from the Coupled Model Intercomparison Project phase 6 (CMIP6, Eyring et al., 2016) and, in particular,



to the subset of CMIP6 simulations from the High Resolution Model Intercomparison Project (HighResMIP, Haarsma et al.,
435 2016). Indeed, HighResMIP simulations are better in simulating TCs and their tracking has already been done with other
physics-based trackers (Roberts et al., 2020).

The primary target will be to apply the ERF tracker calibrated from ERA5 directly to HighResMIP simulations without
any new ERF calibration. This raises the question of the transferability of ERF calibrated from ERA5 to the models. This
issue is two-fold: first, is there a need for bias correction? And second, will ERF be transferable to future climate projections
440 given the climate change signal? The first question will deal with the possibility of the mismatch between the models and
ERA5, preventing ERF to detect cyclones. The second one will address the possibility that the climate change will induce
non-stationarity strong enough to prevent ERF from detecting cyclones.

Multiple applications can be foreseen. For instance, we will be able to study the statistical and physical properties of TCs
detected under climate change. More precisely, we will be able to compare them to the cyclones detected by physical-based
445 trackers and evaluate the complementary added-value brought by the flexibility of ERF for detecting cyclones. Furthermore,
thanks to the partial dependency plots, the differences in the relation between the predictors and the TC presence probability
between the models and ERA5 will be evaluated. Another application can be dedicated to climate change attribution studies
by comparing the properties of TCs in simulations realized under controlled emission scenarios and future climate scenarios.

Ultimately, we would like to make our method widely available. Hence, efforts will be made to make it easy-to-use through
450 an open-source software.

Data availability. ERA5 data are available on the Copernicus Climate Change Service Climate Data Store (CDS, <https://cds.climate.copernicus.eu/datasets/reanalysis-era5-pressure-levels?tab=download>, last access: April 2024). The IBTrACS database is provided by NOAA, National
Centers for Environmental Information, <https://www.ncei.noaa.gov/products/international-best-track-archive> (last access: April 2024)

Appendix A: Tables

Table A1. Confusion Matrix

		Prediction	
		0	1
Observations	0	TN	FP
	1	FN	TP

true positives (TP), true negatives (TN),
false positives (FP), false negatives (FN)



455 *Author contributions.* PV conceived the study, prepared the data and figures, conducted the analysis and wrote the original manuscript. SB contributed to the comparison with the UZ approach. MV discussed the experimental setup. All authors discussed, commented and edited the manuscript.

Competing interests. The authors have no competing interests to declare.

Acknowledgements. Pradeebane Vaittinada Ayar thanks Tom Beucler of the University of Lausanne for his comments. All computations and
460 figures are made using the R free software (R Core Team, 2024). This work has mainly received support from the European Union's Horizon 2020 research and innovation program within the project 'XAIDA: Extreme Events – Artificial Intelligence for Detection and Attribution' (Grant agreement No. 101003469).

This work also benefited from state aid managed by the National Research Agency under France 2030 bearing the reference ANR-22-EXTR-0005 (TRACCS-PC4-EXTENDING project). The authors acknowledge the support of the INSU-CNRS-LEFE-MANU grants
465 (projects CROIRE and COESION), and of the Institut Pascal at Université Paris-Saclay with the support of the program TROPICANA, under the reference ANR-11-IDEX-0003-01.

SB received financial support from the NERC-NSF research grant n° NE/W009587/1 (NERC) & AGS-2244917 (NSF) Hurricane Risk Amplification and Changing North Atlantic Natural disasters (Huracan), and from the EUR IPSL-Climate Graduate School through the ICOCYCLONES2 project, managed by the ANR under the "Investissements d'avenir" programme with the reference 37 ANR-11-IDEX-
470 0004 - 17-EURE-0006.



References

- Accarino, G., Donno, D., Immorlano, F., Elia, D., and Aloisio, G.: An Ensemble Machine Learning Approach for Tropical Cyclone Localization and Tracking From ERA5 Reanalysis Data, *Earth and Space Science*, 10, e2023EA003106, <https://doi.org/10.1029/2023EA003106>, 2023.
- 475 Befort, D. J., Kruschke, T., and Leckebusch, G. C.: Objective identification of potentially damaging tropical cyclones over the Western North Pacific, *Environmental Research Communications*, 2, 031005, <https://doi.org/10.1088/2515-7620/ab7b35>, 2020.
- Bourdin, S., Fromang, S., Dulac, W., Cattiaux, J., and Chauvin, F.: Intercomparison of four algorithms for detecting tropical cyclones using ERA5, *Geoscientific Model Development*, 15, 6759–6786, <https://doi.org/10.5194/gmd-15-6759-2022>, 2022.
- Bourdin, S., Fromang, S., Caubel, A., Ghattas, J., Meurdesoif, Y., and Dubos, T.: Tropical cyclones in global high-resolution simulations
 480 using the IPSL model, *Climate Dynamics*, 62, 4343 – 4368, <https://doi.org/10.1007/s00382-024-07138-w>, 2024.
- Breiman, L.: Random forests, *Machine learning*, 45, 5–32, <https://doi.org/10.1023/A:1010933404324>, 2001.
- Breiman, L., Friedman, J., Olshen, R., and Stone, C.: *Classification and Regression Trees* (1st ed.), Chapman and Hall/CRC, <https://doi.org/10.1201/9781315139470>, 1984.
- Camargo, S. J. and Wing, A. A.: Tropical cyclones in climate models, *WIREs Climate Change*, 7, 211–237, <https://doi.org/10.1002/wcc.373>,
 485 2016.
- Camargo, S. J. and Zebiak, S. E.: Improving the Detection and Tracking of Tropical Cyclones in Atmospheric General Circulation Models, *Weather and Forecasting*, 17, 1152 – 1162, [https://doi.org/10.1175/1520-0434\(2002\)017<1152:ITDATO>2.0.CO;2](https://doi.org/10.1175/1520-0434(2002)017<1152:ITDATO>2.0.CO;2), 2002.
- Centre for Research on the Epidemiology of Disasters, C.: Disaster year in review 2020: Global trends and perspectives, Cred Crunch, 62, <https://www.cred.be/sites/default/files/CredCrunch62.pdf>, 2021.
- 490 Chicco, D. and Jurman, G.: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation., *BMC Genomics*, 21, 6, <https://doi.org/10.1186/s12864-019-6413-7>, 2020.
- Dulac, W., Cattiaux, J., Chauvin, F., Bourdin, S., and Fromang, S.: Assessing the representation of tropical cyclones in ERA5 with the CNRM tracker, *Climate Dynamics*, 62, 223–238, 2024.
- Emanuel, K.: Tropical Cyclones, *Annual Review of Earth and Planetary Sciences*, 31, 75–104,
 495 <https://doi.org/https://doi.org/10.1146/annurev.earth.31.100901.141259>, 2003.
- Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, *Geoscientific Model Development*, 9, 1937–1958, <https://doi.org/10.5194/gmd-9-1937-2016>, 2016.
- Gardoll, S. and Boucher, O.: Classification of tropical cyclone containing images using a convolutional neural network: performance and
 500 sensitivity to the learning dataset, *Geoscientific Model Development*, 15, 7051–7073, <https://doi.org/10.5194/gmd-15-7051-2022>, 2022.
- Haarsma, R. J., Roberts, M. J., Vidale, P. L., Senior, C. A., Bellucci, A., Bao, Q., Chang, P., Corti, S., Fučkar, N. S., Guemas, V., von Hardenberg, J., Hazeleger, W., Kodama, C., Koenigk, T., Leung, L. R., Lu, J., Luo, J.-J., Mao, J., Mizielinski, M. S., Mizuta, R., Nobre, P., Satoh, M., Scoccimarro, E., Semmler, T., Small, J., and von Storch, J.-S.: High Resolution Model Intercomparison Project (HighResMIP v1.0) for CMIP6, *Geoscientific Model Development*, 9, 4185–4208, <https://doi.org/10.5194/gmd-9-4185-2016>, 2016.
- 505 Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H.: *The elements of statistical learning: data mining, inference, and prediction*, Springer, <https://doi.org/10.1007/978-0-387-84858-7>, 2009.



- He, H. and Garcia, E. A.: Learning from Imbalanced Data, *IEEE Transactions on Knowledge and Data Engineering*, 21, 1263–1284, <https://doi.org/10.1109/TKDE.2008.239>, 2009.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.: The ERA5 global reanalysis, *Quarterly Journal of the Royal Meteorological Society*, 146, 1999–2049, <https://doi.org/10.1002/qj.3803>, 2020.
- 510 Keper, J. D.: Tropical Cyclone Structure and Dynamics, pp. 3–53, WORLD SCIENTIFIC, https://doi.org/10.1142/9789814293488_0001, 2010.
- Kim, M., Park, M.-S., Im, J., Park, S., and Lee, M.-I.: Machine Learning Approaches for Detecting Tropical Cyclone Formation Using Satellite Data, *Remote Sensing*, 11, <https://doi.org/10.3390/rs11101195>, 2019a.
- 515 Kim, S., Kim, H., Lee, J., Yoon, S., Kahou, S. E., Kashinath, K., and Prabhat, M.: Deep-Hurricane-Tracker: Tracking and Forecasting Extreme Climate Events, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1761–1769, <https://doi.org/10.1109/WACV.2019.00192>, 2019b.
- Klotzbach, P. J., Bell, M. M., Bowen, S. G., Gibney, E. J., Knapp, K. R., and Schreck, C. J.: Surface Pressure a More Skillful Predictor of Normalized Hurricane Damage than Maximum Sustained Wind, *Bulletin of the American Meteorological Society*, 101, E830 – E846, <https://doi.org/10.1175/BAMS-D-19-0062.1>, 2020.
- 520 Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J., and Neumann, C. J.: The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying Tropical Cyclone Data, *Bulletin of the American Meteorological Society*, 91, 363 – 376, <https://doi.org/10.1175/2009BAMS2755.1>, 2010.
- Knapp, K. R., Diamond, H. J., Kossin, J. P., and Kruk, M. C. and Schreck, C.: International Best Track Archive for Climate Stewardship (IB-TrACS) Project, version 4r00. [subset: since1980], NOAA National Centers for Environmental Information, <https://doi.org/10.25921/82ty-9e16> [last access April 2024], 2018.
- 525 Knutson, T., Camargo, S. J., Chan, J. C. L., Emanuel, K., Ho, C.-H., Kossin, J., Mohapatra, M., Satoh, M., Sugi, M., Walsh, K., and Wu, L.: Tropical Cyclones and Climate Change Assessment: Part I: Detection and Attribution, *Bulletin of the American Meteorological Society*, 100, 1987 – 2007, <https://doi.org/10.1175/BAMS-D-18-0189.1>, 2019.
- 530 Knutson, T., Camargo, S. J., Chan, J. C. L., Emanuel, K., Ho, C.-H., Kossin, J., Mohapatra, M., Satoh, M., Sugi, M., Walsh, K., and Wu, L.: Tropical Cyclones and Climate Change Assessment: Part II: Projected Response to Anthropogenic Warming, *Bulletin of the American Meteorological Society*, 101, E303 – E322, <https://doi.org/10.1175/BAMS-D-18-0194.1>, 2020.
- Knutson, T. R., McBride, J. L., Chan, J., Emanuel, K., Holland, G., Landsea, C., Held, I., Kossin, J. P., Srivastava, A., and Sugi, M.: Tropical cyclones and climate change, *Nature geoscience*, 3, 157–163, <https://doi.org/10.1038/ngeo779>, 2010.
- 535 Komisarczyk, K., Kozminski, P., Maksymiuk, S., and Biecek, P.: treeshap: Compute SHAP Values for Your Tree-Based Models Using the 'TreeSHAP' Algorithm, <https://CRAN.R-project.org/package=treeshap>, r package version 0.3.0, 2023.
- Kuhn, M.: Applied predictive modeling, <https://doi.org/10.1007/978-1-4614-6849-3>, 2013.
- Kumler-Bonfanti, C., Stewart, J., Hall, D., and Govett, M.: Tropical and Extratropical Cyclone Detection Using Deep Learning, *Journal of Applied Meteorology and Climatology*, 59, 1971 – 1985, <https://doi.org/10.1175/JAMC-D-20-0117.1>, 2020.
- 540 Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I.: From local explanations to global understanding with explainable AI for trees, *Nature machine intelligence*, 2, 56–67, <https://doi.org/10.1038/s42256-019-0138-9>, 2020.
- Malley, J. D., Kruppa, J., Dasgupta, A., Malley, K. G., and Ziegler, A.: Probability machines, *Methods of information in medicine*, 51, 74–81, <https://doi.org/10.3414/ME00-01-0052>, 2012.



- 545 Matthews, B.: Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405, 442–451, [https://doi.org/10.1016/0005-2795\(75\)90109-9](https://doi.org/10.1016/0005-2795(75)90109-9), 1975.
- Murakami, H., Vecchi, G. A., Underwood, S., Delworth, T. L., Wittenberg, A. T., Anderson, W. G., Chen, J.-H., Gudgel, R. G., Harris, L. M., Lin, S.-J., and Zeng, F.: Simulation and Prediction of Category 4 and 5 Hurricanes in the High-Resolution GFDL HiFLOR Coupled Climate Model, *Journal of Climate*, 28, 9058 – 9079, <https://doi.org/10.1175/JCLI-D-15-0216.1>, 2015.
- 550 Nair, A., Srujan, K. S. S. S., Kulkarni, S. R., Alwadhi, K., Jain, N., Kodamana, H., Sandeep, S., and John, V. O.: A Deep Learning Framework for the Detection of Tropical Cyclones From Satellite Images, *IEEE Geoscience and Remote Sensing Letters*, 19, 1–5, <https://doi.org/10.1109/LGRS.2021.3131638>, 2022.
- R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, <https://www.R-project.org/>, 2024.
- 555 Raavi, P. H. and Walsh, K.: Sensitivity of Tropical Cyclone Formation to Resolution-Dependent and Independent Tracking Schemes in High-Resolution Climate Model Simulations, *Earth and Space Science*, 7, e2019EA000906, <https://doi.org/https://doi.org/10.1029/2019EA000906>, 2020.
- Roberts, M. J., Camp, J., Seddon, J., Vidale, P. L., Hodges, K., Vanniere, B., Mecking, J., Haarsma, R., Bellucci, A., Scoccimarro, E., Caron, L.-P., Chauvin, F., Terray, L., Valcke, S., Moine, M.-P., Putrasahan, D., Roberts, C., Senan, R., Zarzycki, C., and Ullrich, P.: Impact of
- 560 Model Resolution on Tropical Cyclone Simulation Using the HighResMIP–PRIMAVERA Multimodel Ensemble, *Journal of Climate*, 33, 2557 – 2583, <https://doi.org/10.1175/JCLI-D-19-0639.1>, 2020.
- Rüttgers, M., Lee, S., Jeon, S., and You, D.: Prediction of a typhoon track using a generative adversarial network and satellite images, *Scientific reports*, 9, 6057, <https://doi.org/10.1038/s41598-019-42339-y>, 2019.
- Shapley, L. S.: Notes on the n-Person Game – II: The Value of an n-Person Game, Rand Corporation, https://www.rand.org/content/dam/rand/pubs/research_memoranda/2008/RM670.pdf, 1951.
- 565 Siders, Z. A., Ducharme-Barth, N. D., Carvalho, F., Kobayashi, D., Martin, S., Raynor, J., Jones, T. T., and Ahrens, R. N.: Ensemble random forests as a tool for modeling rare occurrences, *Endangered Species Research*, 43, 183–197, <https://doi.org/10.3354/esr01060>, 2020.
- Ullrich, P. A., Zarzycki, C. M., McClenny, E. E., Pinheiro, M. C., Stansfield, A. M., and Reed, K. A.: TempestExtremes v2. 1: A community framework for feature detection, tracking, and analysis in large datasets, *Geoscientific Model Development*, 14, 5023–5048, <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016GL071606>, 2021.
- 570 UN Office for Disaster Risk Reduction, U.: The Human Cost of Disasters: An Overview of the Last 20 Years (2000–2019), <https://www.undrr.org/media/48008/download>, 2020.
- Walsh, K., Camargo, S., Knutson, T., Kossin, J., Lee, T.-C., Murakami, H., and Patricola, C.: Tropical cyclones and climate change, *Tropical Cyclone Research and Review*, 8, 240–250, <https://www.sciencedirect.com/science/article/pii/S2225603220300047>, 2019.
- 575 Walsh, K. J., McBride, J. L., Klotzbach, P. J., Balachandran, S., Camargo, S. J., Holland, G., Knutson, T. R., Kossin, J. P., Lee, T.-c., Sobel, A., et al.: Tropical cyclones and climate change, *Wiley Interdisciplinary Reviews: Climate Change*, 7, 65–89, <https://doi.org/10.1002/wcc.371>, 2016.
- Walsh, K. J. E., Camargo, S. J., Vecchi, G. A., Daloz, A. S., Elsner, J., Emanuel, K., Horn, M., Lim, Y.-K., Roberts, M., Patricola, C., Scoccimarro, E., Sobel, A. H., Strazzo, S., Villarini, G., Wehner, M., Zhao, M., Kossin, J. P., LaRow, T., Oouchi, K., Schubert, S., Wang, H., Bacmeister, J., Chang, P., Chauvin, F., Jablonowski, C., Kumar, A., Murakami, H., Ose, T., Reed, K. A., Saravanan, R., Yamada, Y., Zarzycki, C. M., Vidale, P. L., Jonas, J. A., and Henderson, N.: Hurricanes and Climate: The U.S. CLIVAR Working Group on Hurricanes, *Bulletin of the American Meteorological Society*, 96, 997 – 1017, <https://doi.org/10.1175/BAMS-D-13-00242.1>, 2015.
- 580



- Wang, P., Wang, P., Wang, C., Yuan, Y., and Wang, D.: A Center Location Algorithm for Tropical Cyclone in Satellite Infrared Images, IEEE
Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 2161–2172, [https://dx.doi.org/10.1109/JSTARS.2020.](https://dx.doi.org/10.1109/JSTARS.2020.2995158)
585 2995158, 2020.
- Weaver, M. M. and Garner, A. J.: Varying genesis and landfall locations for North Atlantic tropical cyclones in a warmer climate, Scientific
Reports, 13, 5482, <https://doi.org/10.1038/s41598-023-31545-4>, 2023.
- Wright, M. N. and Ziegler, A.: ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R, Journal of
Statistical Software, 77, 1–17, <https://doi.org/10.18637/jss.v077.i01>, 2017.
- 590 Yule, G. U.: On the Methods of Measuring Association Between Two Attributes, Journal of the Royal Statistical Society, 75, 579–652,
<https://doi.org/10.2307/2340126>, 1912.
- Zarzycki, C. M. and Ullrich, P. A.: Assessing sensitivities in algorithmic detection of tropical cyclones in climate data, Geophysical Research
Letters, 44, 1141–1149, <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016GL071606>, 2017.
- Zarzycki, C. M., Ullrich, P. A., and Reed, K. A.: Metrics for Evaluating Tropical Cyclones in Climate Data, Journal of Applied Meteorology
595 and Climatology, 60, 643 – 660, <https://doi.org/10.1175/JAMC-D-20-0149.1>, 2021.