Earth and Space Science

RESEARCH ARTICLE
10.1029/2023EA003106

Gabriele Accarino and Davide Donno contributed equally to this work.

Key Points:

- An Ensemble Machine Learning approach for localizing Tropical Cyclone (TC) center using reanalysis and observation data is proposed
- The ensemble approach was able to improve TC localization performance with respect to single model estimates
- The data-driven localization approach was integrated with a deterministic tracking scheme and compared against other trackers in literature

Correspondence to:

G. Aloisio,
giovanni.aloisio@cmcc.it

Citation:

Accarino, G., Donno, D., Immorlano, F., Elia, D., & Aloisio, G. (2023). 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

Received 14 JUN 2023
Accepted 10 OCT 2023

Author Contributions:

Conceptualization: Gabriele Accarino, Donatello Elia, Giovanni Aloisio
Data curation: Francesco Immorlano
Formal analysis: Gabriele Accarino, Davide Donno
Funding acquisition: Giovanni Aloisio
Investigation: Gabriele Accarino, Davide Donno, Francesco Immorlano
Methodology: Gabriele Accarino, Francesco Immorlano, Donatello Elia
Project Administration: Giovanni Aloisio
Resources: Donatello Elia
Software: Davide Donno
Supervision: Giovanni Aloisio

An Ensemble Machine Learning Approach for Tropical Cyclone Localization and Tracking From ERA5 Reanalysis Data

Gabriele Accarino[1], Davide Donno[1], Francesco Immorlano[1,2], Donatello Elia[1], and Giovanni Aloisio[1,2]

[1]Advanced Scientific Computing Division, Centro Euro-Mediterraneo sui Cambiamenti Climatici, Lecce, Italy, [2]Department of Innovation Engineering, University of Salento, Lecce, Italy

Abstract Tropical Cyclones (TCs) are counted among the most destructive phenomena that can be found in nature. Every year, globally an average of 90 TCs occur over tropical waters, and global warming is making them stronger and more destructive. The accurate localization and tracking of such phenomena have become a relevant and interesting area of research in weather and climate science. Traditionally, TCs have been identified in large climate data sets through the use of deterministic tracking schemes that rely on subjective thresholds. This study presents a Machine Learning (ML) ensemble approach for locating TCs center coordinates. The ensemble combines TCs center estimates of different ML models that agree about the presence of a TC in input data. ERA5 reanalysis data was used for model training and testing jointly with the International Best Track Archive for Climate Stewardship (IBTrACS) records. Compared to single models estimates, the ML ensemble approach was able to improve TCs localization in terms of Euclidean Distance with respect to the observed TCs locations from IBTrACS. Moreover, a hybrid tracking scheme was defined: starting from the individual TC center locations detected by the ML ensemble approach, a deterministic tracking algorithm was used for reconstructing TC trajectories. The hybrid tracking scheme was then compared with four deterministic trackers reported in literature, achieving a Probability of Detection and a False Alarm Rate of 71.49% and 23%, respectively, over 40 years of reanalysis data.

Plain Language Summary Every year an average of 90 Tropical Cyclones (TCs) occur globally, and this number is expected to rise due to global warming, which is also increasing the frequency and the intensity of such extremes. The localization and tracking of TCs have traditionally been addressed by means of deterministic tracking schemes. The present study introduces an ensemble approach based on Machine Learning (ML) that locates the TC center coordinates. Basically, the idea is to rely on several ML models accomplishing the same task—each with different training configurations—to integrate their results. The climate variables, used as predictor for training and testing of the models, were gathered from ERA5 reanalysis, while the historical TCs center positions, used as target, were retrieved from the International Best Track Archive for Climate Stewardship data set. The results showed the effectiveness of the proposed approach against the use of a single ML model. Moreover, starting from the individual TC center locations detected by the ML ensemble approach, a deterministic tracking scheme was used from literature to reconstruct TC trajectories. The proposed hybrid tracking algorithm was then compared with four deterministic trackers reported in literature, showing comparable skills.

1. Introduction

Tropical Cyclones (TCs), also known as hurricanes or typhoons, are counted among the most fascinating and destructive phenomena that can be found in nature (Emanuel, 2003). Several conditions are at the basis of TC formation. As described in (W. M. Gray, 1975; Weaver & Garner, 2023), "TC genesis requires warm sea surface temperatures, low wind shear, ample humidity, adequate influence from the Coriolis force, and a pre-existing low-pressure disturbance in the atmosphere." Besides the aforementioned conditions, the cyclone center (i.e., the eye) is typically located in a low-pressure region surrounded by strong winds and deep cumulonimbus. As the TC travels, it becomes a self-sufficient system that continuously gathers energy from the ocean. If the TC moves toward land (i.e., the so-called landfall), the TC loses its energy, which causes rapid dissipation (Kepert, 2010; MetOffice, 2023; Rüttgers et al., 2019).

Validation: Gabriele Accarino, Davide Donno, Francesco Immorlano
Visualization: Davide Donno, Francesco Immorlano
Writing – original draft: Gabriele Accarino
Writing – review & editing: Donatello Elia, Giovanni Aloisio

The geographical areas that lead to the formation of TCs are called *cyclone formation basins*. There are seven basins around the world, each with specific water depth and sea surface temperature, which translates to a different number of TCs per year and varied seasons in which they develop (Roy & Kovordányi, 2012). Every year, globally an average of 90 TCs occur over tropical waters (Emanuel & Nolan, 2004) and global warming is making them stronger, larger and more destructive, as found out by Elsner et al. (2008), Mendelsohn et al. (2012), and Sun et al. (2017). As reported by the World Meteorological Organization, 1,942 disasters have been attributed to TCs, which caused US $ 1,407.6 billion in economic losses and almost 8 million casualties over the past 50 years (World Meteorological Organization, 2023), thus making TCs impact quite significant on different sectors, such as infrastructures, economy, human health, but also in terms of social unrest.

The accurate detection and tracking of these phenomena have become a relevant and interesting area of research in weather and climate science (Dabhade et al., 2021; Scoccimarro et al., 2014). Traditionally, TCs have been identified in large climate data sets through the use of deterministic tracking schemes, also known as TCs trackers (Horn et al., 2014). These algorithms are capable of identifying—by means of thresholds applied on variables significant to the cyclogenesis—patterns related to a warm core in gridded data sets and connecting them along the TC trajectory (Bourdin et al., 2022). Depending on the particular variables involved in the tracking process, two main categories of schemes exist: physics-based (see Camargo and Zebiak, 2002; Chauvin et al., 2006; Horn et al., 2014; Murakami, 2014; Zarzycki and Ullrich, 2017; Zhao et al., 2009) and dynamics-based that include the TRACK method (Hodges et al., 2017; Strachan et al., 2013) and the Okubo-Weiss-Zet. algorithm (Tory et al., 2013b). In particular, physics-based trackers rely on thermo-dynamical variables, "they are based on the detection of a local minimum sea-level pressure (SLP) combined with a warm-core criterion—usually expressed as a temperature anomaly or a geopotential thickness—on top of which discriminating intensity criteria are applied based on surface winds or vorticity." (Bourdin et al., 2022), whereas dynamics-based trackers rely on "dynamical variables such as vorticity or other derivatives of the velocity" (Bourdin et al., 2022), and are referred to be resolution-independent (Tory et al., 2013a).

The aforementioned thresholds are set by the author of the scheme, therefore they are subjective and mainly rely on the human expertise about the phenomena under investigation (Dabhade et al., 2021; Enz et al., 2022). Moreover, thresholds may depend on the particular geographical region of study and the related formation basin, as well as on the TC categories (Befort et al., 2020; Bloemendaal et al., 2021). However, manual threshold tuning may lead to subjective bias, including the potential inability of tracking schemes to generalize to other domains or data from sources other than those used to calibrate the thresholds.

The state of different climate variables, which the tracking schemes are applied on, is simulated by physics-based Earth System Models (ESMs) that provide large amounts of data at different spatio-temporal resolutions. In addition to ESM data, ground-based in situ observations and satellite retrievals contribute to further increase the data volume. Such large-scale data introduces issues in terms of how scientific data can be effectively managed and processed to make the best out of it (J. Gray et al., 2005). Indeed, climate scientists, meteorological agencies and policy decision makers need to process and extract meaningful information from these huge data sets in a cost-effective manner and in a reasonable amount of time (Sebestyén et al., 2021). In this context, High Performance Data analytics systems can address some of the issues and provide support for descriptive/statistical analysis of this large-scale data (Elia et al., 2021). Nevertheless, in the last few years Machine Learning (ML) and Deep Learning (DL) algorithms became popular as data-driven paradigms for supporting feature extraction from the vast amounts of scientific data currently available (Hey et al., 2020). ML and DL algorithms can actually go beyond what can be extracted with traditional descriptive and deterministic methodologies.

To this extent, several research efforts can be found in scientific literature toward the development of cutting-edge TC detection approaches beyond the existing deterministic tracking schemes. For example, many studies have been focusing on the use of satellite data and DL approaches for accurately locating the TC center (Carmo et al., 2021).

Several works, such as Haque et al. (2022), Lam et al. (2023), Pang et al. (2021), and Shakya et al. (2020), framed the identification of the TC center as an object detection task for which the You Only Look Once v3 DL object detection model was adopted. Similarly, a DL-based object detection approach was proposed in Wang et al. (2020) with the aim of retrieving the TC center through segmentation, edge detection, circle fitting, and comprehensive decision of satellite infrared images. Segmentation of the shape and size of the detected TC in high-resolution satellite images was also provided by Nair et al. (2022). To this extent, a pipeline consisting

of a Mask Region-Convolutional Neural Network (R-CNN) detector, a wind speed filter and a CNN classifier was adopted to accurately detect TCs. In Xie et al. (2022), a Feature Pyramid Network was proposed as a feature extractor and region proposal network that searches for the potential areas of cyclones along with a Faster Region-based CNN to calibrate the locations of TC regions. A faster Region-based CNN was also used by Xie et al. (2020) to classify the presence of TCs in Mean Wind Field-Advanced Scatterometer satellite data. The authors of S. Kim et al. (2019) exploited a Convolutional Long Short-Term Memory to detect, track and predict hurricane trajectories on Community Atmospheric Model v5 simulation data. With the aim of capturing both temporal dynamics and spatial distribution, trajectories were modeled as time-sequential density maps. The detection of tropical and extratropical cyclones was addressed as an image segmentation task by Kumler-Bonfanti et al. (2020). They used the Global Forecasting System and Geostationary Operational Environmental Satellite to compare four state-of-the-art U-Net-based models designed for the detection task. In Carmo et al. (2021), data from the Sentinel-1 C-band satellite was used to provide a DL-based detector of the TC center, also providing estimates of the related category according to sea surface wind and rain-related topological patterns. Authors further provided explainability through the analysis of key patterns highlighted by the Gradient-based Class Activation Map method. M. Kim et al. (2019) used eight predictors gathered from the WindSat satellite to frame a TC detection task. Then, they compared the detection skills of three ML algorithms, namely Decision Trees, Random Forest, Support Vector Machines and a model based on Linear Discriminant Analysis.

This work proposes a TC center localization approach based on ML and applied on the joint North Pacific and Atlantic TC formation basins. Although this task is similar to other studies from the state of the art, there are some important differences in the algorithmic approach used.

From a methodological perspective, a ML ensemble approach is proposed to accurately locate the TC center coordinates. Exploiting a single ML model for locating the TC center would have resulted in unreliable results because of the inherent complexity of the TC center localization task. The ensemble, instead, allows combining TC center estimates of different ML models that are in agreement about the presence of a TC in input data. In this way, each model can learn different spatial characteristics of the TC structure and the ensemble allows more accurate TC center estimates. With respect to other approaches available in literature for uncertainty quantification in Artificial Neural Networks (ANNs) predictions, our approach proposes an extension of the multi-model ensemble method, as reported in Haynes et al. (2023). The multi-model approach consists of several ANNs trained on the same data and hyperparameters, but with different initial conditions, whereas our approach considers several ANNs architectures trained on the same data but with different hyperparameters and initial conditions. Moreover, a deterministic tracking algorithm was used to reconstruct trajectories from the detected TCs centers. As a consequence, these models can easily complement deterministic tracking schemes for the TC detection task. This results in a hybrid tracking scheme combining data-driven detection for selecting the TC center candidates with a deterministic tracking algorithm.

In contrast to other studies, a total of six ERA5 reanalysis TCs predictors (i.e., Mean Sea Level Pressure (MSLP), 10 m wind gust since previous post-processing, instantaneous 10 m wind gust, relative vorticity at 850 mb, and temperature at 300 and 500 mb) were used in place of satellite data. Reanalysis data combines model simulations and observations to provide the best representation of climate variables in the past (ECMWF, 2020a). The TC center geographical coordinates were retrieved from the International Best Track Archive for Climate Stewardship (IBTrACS) data set, the most complete global collection of historical TC occurrences (National Oceanic and Atmospheric Administration, 2023).

The rest of the paper is organized as follows: Section 2 describes data sources and the processing steps required to build a suitable data set for ML training. Moreover, the experimental setup is described, along with Deep Neural Network (NN) models architectures, the ensemble procedure and the hybrid tracking scheme adopted. Section 3 presents the results of (a) the ML ensemble approach for localizing TC centers, (b) the comparison of the hybrid tracking scheme with four deterministic TC trackers from literature, and (c) the outcomes of the hybrid tracker on two test cases. Section 4 discusses the obtained results, highlighting strengths and limitations of the proposed approach, and draws the main conclusions from this work while also pointing out some relevant future activities.

2. Materials and Methods

2.1. Data Sources

This subsection provides the description of the two data sources used to build the data set for ML setup.

2.1.1. The International Best Tracks Archive for Climate Stewardship

The IBTrACS presented by Knapp et al. (2010) is an institutional, open access and centralized archive that provides the most complete set of historical TC best track data on a global level. It integrates historical records with observations retrieved from 12 different meteorological agencies. The main aim of IBTrACS is fostering research in the context of such events by keeping track of their geographical position, frequency and intensity worldwide. IBTrACS reports global TCs occurrences at 0.1° (∼10 km) of spatial resolution from 1841 to present with a 3-hourly temporal frequency. However, in this study, only TC records between 1980 and 2019 were selected from IBTrACS v4 (Knapp et al., 2018) at a temporal frequency of 6 hr. Although IBTrACS provides TC records from 1841 to date, 1980 is considered the beginning of the Modern Era, characterized by the extensive use of geostationary satellite imagery on a global scale. On the other hand, more recent TC information is subject to frequent reanalysis by the different meteorological agencies contributing to IBTrACS, and, for these reasons, TC selection was limited to 1980–2019. Furthermore, 6-hourly data provides additional information about the TC characteristic, such as the Maximum Sustained Wind (MSW), contrary to 3-hourly data (Knapp et al., 2010).

Concerning the geographical domain, this study mainly targets the North Pacific formation basin, which is widely recognized as a particularly active region where most TCs occur every year (Roy & Kovordányi, 2012). Since a substantial number of TC events cross both the North Pacific and North Atlantic regions, thus reaching up to 320°E of longitude, the final domain of interest is 100–320°E, 0–70°N (i.e., joint North Atlantic and North Pacific).

2.1.2. ERA5 Reanalysis

Climate variables are the main drivers that contribute to the formation and strengthening of TCs during their lifetime, and they were retrieved from the Copernicus Climate Change Service ERA5 reanalysis data sets. ERA5 reanalysis combines global numerical weather predictions with newly available observations in an optimal way to produce consistent estimates of the state of the atmosphere (ECMWF, 2020b). In this study, MSLP [Pa] (*msl*), 10 m wind gust since previous post-processing [ms^{-1}] (*fg10*) and the instantaneous 10 m wind gust [ms^{-1}] (*i10fg*) were gathered from the ERA5 reanalysis on single levels (Hersbach et al., 2023b), whereas the relative vorticity at 850 mb [s^{-1}] (*vo850*) and the temperature at 300 and 500 mb [K] (*t300* and *t500*, respectively) were collected from the ERA5 reanalysis on the pressure levels data set (Hersbach et al., 2023a). The six variables considered in this study to characterize the storm structure and its intensity have been selected according to other related studies (Scoccimarro et al., 2017; Zhao et al., 2009). Each of the aforementioned climatic variables was provided on a regular grid of 0.25° × 0.25° (∼27 × 27 km) of spatial resolution, targeting the geographical domain previously described, and it was managed as a 2-dimensional map of 280 × 880 pixels size. Moreover, data was collected with a 6-hourly temporal resolution (i.e., 00.00, 06.00, 12.00 and 18.00 time steps) for the period 1980–2019, thus matching TCs records selected from IBTrACS, except for *fg10* that was originally collected with an hourly temporal resolution. In particular, the ERA5 *fg10* variable reports the maximum wind gust in the preceding hour. Therefore, to match the 6-hourly temporal resolution of this study, the maximum over the previous 6 hr was computed for each time step.

2.2. Data Processing

2.2.1. IBTrACS Filtering and Selection

Starting from trajectories belonging to the joint North Atlantic and North Pacific geographical domain (100–320°E, 0–70°N), only IBTrACS records were considered with *track_type* field flagged as *main*. Therefore, *provisional*, *spur* and *provisional-spur* tracks were implicitly discarded as they are characterized by a higher level of uncertainty (IBTrACS Science Team, 2019). It is noteworthy that data from recent years is typically provided as *provisional* or *spur*, meaning that the corresponding values have not been reanalyzed yet and therefore are of lower quality compared to *main* tracks. This can happen because some variables—such as the intensity, position and storm categories—are subject to change based on posterior reanalysis by the

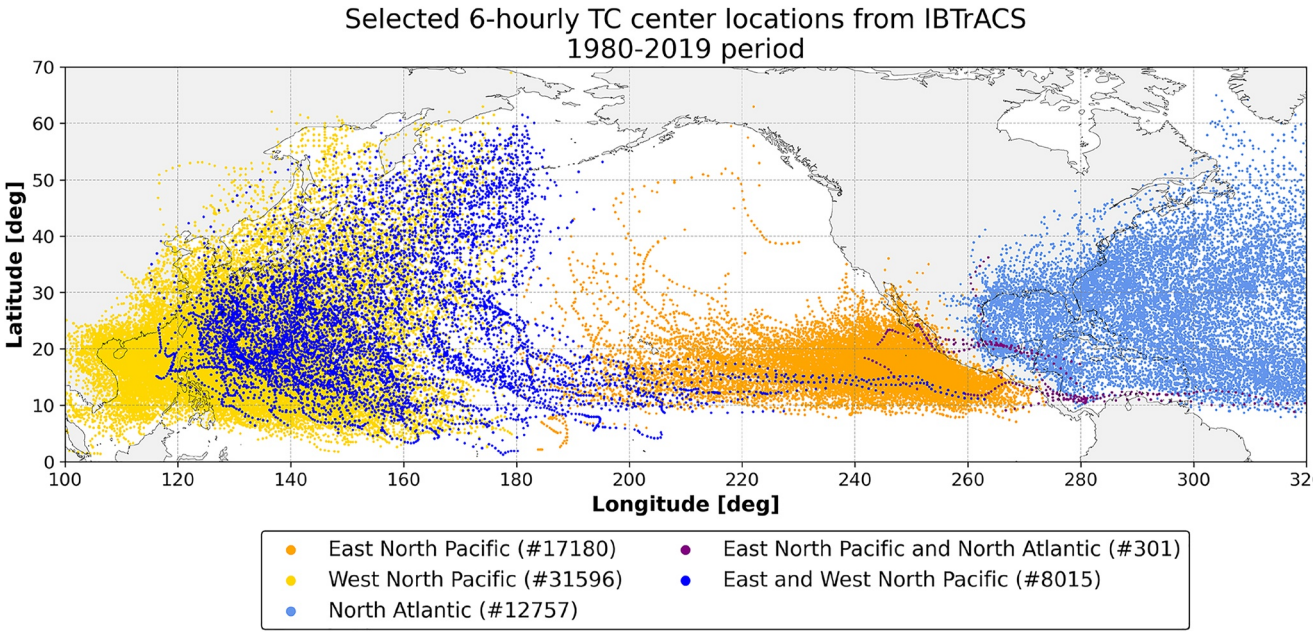

Figure 1. Visualization of 6-hourly Tropical Cyclone (TC) center locations within the 1980–2019 period in the joint East and West North Pacific and North Atlantic basin (100–320°E, 0–70°N). TC locations in each sub-basin are highlighted by a different color, along with the relative number of occurrences. Only International Best Track Archive for Climate Stewardship (IBTrACS) records whose nature is Tropical, Subtropical and Extra Tropical Storm are shown in the picture.

meteorological agencies. Moreover, uncertainties in the observing system may result in contradictory opinions by different agencies about the storm location, leading to *spur* tracks. This is mainly due to difficulties in localizing the center of circulation or in the case of storms merging (i.e., Fujiwhara effect) (IBTrACS Science Team, 2019). As an additional selection step, tracks were filtered out based on the nature field, specifically discarding those trajectories marked as: (a) *Not Reported* (*NR*) whose nature is unknown, (b) *Disturbance Storms* (*DS*) that correspond to not-well-formed storms characterized by a MSW less than 34 knots, and (c) *Mixture* (*MX*) that correspond to tracks that received contradicting reports about the nature of the observing system from different agencies. At the end of this filtering and selection process, only *Tropical Storms* (*TS*), *Extra Tropical* (*ET*) and *Subtropical Storms* (*SS*) main tracks were considered at a 6-hourly temporal resolution.

Figure 1 shows the observed 6-hourly TC center locations occurring in the domain of study in the 1980–2019 period. The TC center locations were further divided into non-overlapping groups according to the basins involved during their lifecycle. The figure also reports the number of occurrences in each group. The locations of TC center in the North Atlantic (light blue), West North Pacific (yellow) and East North Pacific (orange) remain confined to such basins (i.e., they originate and dissipate in the same basin), whereas East and West North Pacific locations (blue) as well as East North Pacific and North Atlantic ones (purple) involve different basins during their lifecycle.

2.2.2. Patches Generation and Labeling

For each of the six climatic drivers, ERA5 maps (280 × 880 pixels) were evenly tiled into 7 × 22 non-overlapping patches of 40 × 40 pixels size each (see Figure 2). The TC center can occur in every pixel of the patch, not necessarily at the center, thus non cyclone-centric patches were generated. Then, drivers were stacked together, resulting in data of 40 × 40 × 6 dimension, hereafter defined as an input patch. In order to associate patches containing a TC (from now on referred to as *positive patches*) with its center position (i.e., the TC eye), the latitude and longitude geographical coordinates extracted from IBTrACS were rounded to match the resolution of the ERA5 grid (0.25° × 0.25°). Subsequently, rounded coordinates were further converted into local-patch positions in terms of (x,y) index pairs (considering the 40 × 40 patch as a matrix).

Different TC phenomena may simultaneously occur in the domain of interest at a particular time, thus multiple *positive patches* can be retrieved from a single ERA5 map. The patches that do not contain a TC (from now on

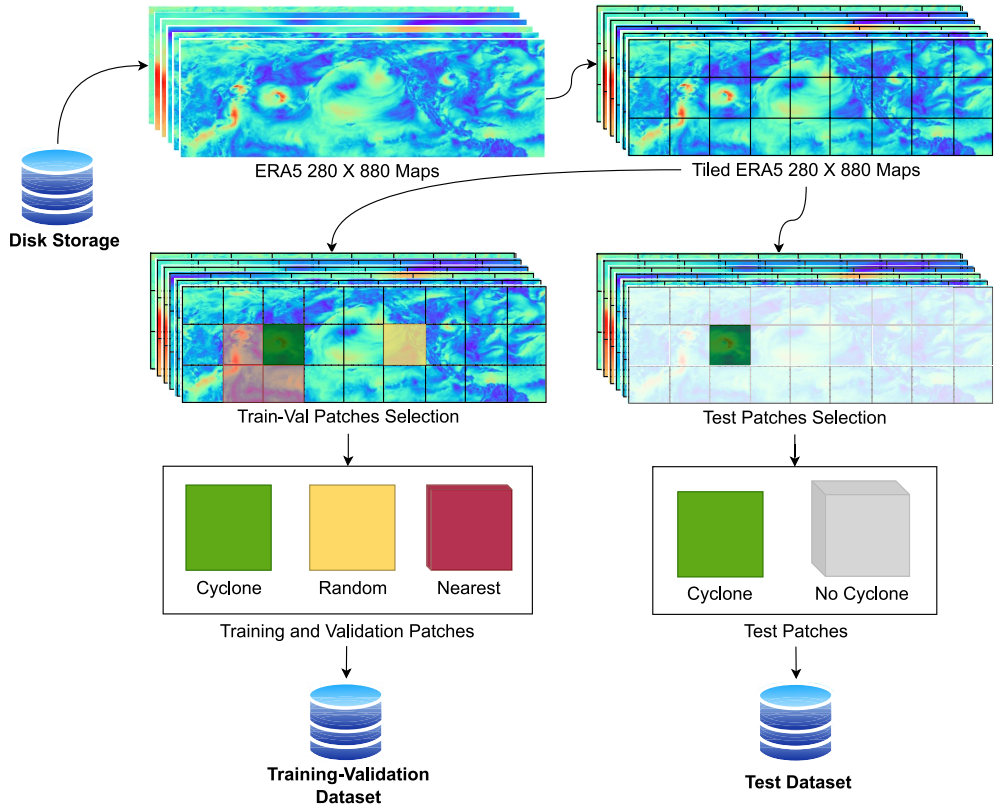

Figure 2. Overview of the data set building pipeline. ERA5 maps are tiled into non-overlapping patches of 40 × 40 grid points in size. To create the training and validation subsets, for each tiled ERA5 map, the patch containing the Tropical Cyclone center is considered along with nearest and random patches, discarding the remaining ones. Concerning the test subset, all the patches within ERA5 maps are considered.

referred to as *negative patches*) were labeled with a negative (x,y) coordinate (i.e., $(-1, -1)$), indicating the TC absence. In this way, retrieved local patch (x,y) pairs were used as the target of the detection task.

2.3. Experimental Setup

2.3.1. Data Set Creation

Selected patches in the considered period (1980–2019) were split into training and test sets as follows: patches belonging to two consecutive years for each decade were selected for testing (1983, 1984, 1993, 1994, 2003, 2004, 2013, 2014, respectively), whereas patches in the remaining years were used for training. In this way, the training set comprises patches that span the whole time period, enabling ML models to capture and learn potential climate change patterns that may affect the input atmospheric drivers (World Meteorological Organization, 2022).

In order to build the training set, *negative patches* were carefully selected to enhance the variance of the data set, as well as to improve the predictive skills of ML models. Among the edge patches surrounding a *positive* one, the three corner patches closest to the storm center were considered as *negative* (referred to as *nearest patches*, see purple patches in Figure 2). Despite nearest patches are labeled as *negative*, they may contain residual structures (e.g., spiral wind gust tails, minimum regions of MSLP, etc.) of the TC located in the central patch. Therefore, including such patches can actually benefit model training.

Additionally, for each *positive patch* a further *negative sample* was randomly selected among the 7 × 22 patches of the map excluding the edge ones previously mentioned, thus ensuring that no major TC phenomena occur in the randomly selected patch. By construction, the training set is imbalanced toward *negative samples* (i.e., 55,639 *positive patches* and 212,679 *negative* ones, yielding 20% of samples containing a TC). To address the imbalance ratio, and also to increase the variance of the training set, a selective data augmentation procedure was used to

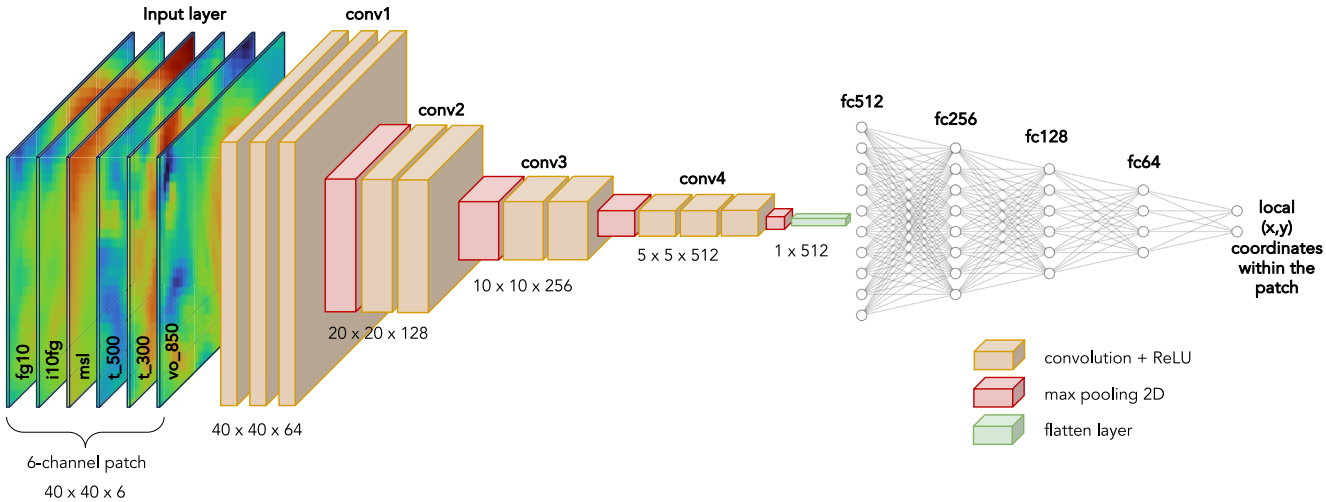

Figure 3. Basic representation of a Visual Geometry Group-like architecture for the Tropical Cyclone (TC) detection task. Patches related to the six input drivers are stacked together and the local (x,y) coordinates of the TC center are used as the target. The proposed architectures differ in their complexity.

reach a 50/50 ratio. For each *positive sample*, three transformations were applied: *left-right flip*, *up-down flip* and 180° *rotation* (Shorten & Khoshgoftaar, 2019).

Conversely, all 7 × 22 patches belonging to each map of the test set years were selected to assess the actual performance of ML models on out-of-sample data. The resulting test set consists of 967,513 *negative patches* and 14,149 *positive* ones, ending up in a strongly imbalanced test set (i.e., only 1.46% of positive samples).

2.3.2. Neural Network Architectures

For each input patch that comprises the six climatic drivers, the proposed TC center localization task predicts the local (x,y) coordinates of the TC center location within the input patch, if present. If both the (x,y) coordinates fall within the admissible range of the patch (0–39, 0–39), the patch is classified as *positive* (i.e., the TC center is identified), otherwise the patch is classified as *negative* (i.e., no TC center is identified). In case of *positive patches*, the local coordinates are converted to the global (lat,lon) coordinates.

To this extent, several Visual Geometry Group (VGG)-like NN architectures (Simonyan & Zisserman, 2014) were developed and trained, each differing in terms of number of layers, filters and kernel sizes.

Figure 3 depicts a basic representation of a VGG-like architecture. Input patches are processed by a series of convolutional and max-pooling layers that encode the input volume, thus progressively decreasing height and width dimensions, while increasing the depth of the activation volume at the same time. The convolutions squeeze the processed information resulting in a lower dimensional representation of the patch content (Barrera, 2022). After the convolutional blocks, the resulting output is flattened and processed through a series of dense layers that gradually reconstructs the information and links it with the target (x,y) coordinates of the TC center within the patch. In this sense, the VGG-like network is trained to learn the mapping between input climatic drivers content and the two output coordinates.

A total of four different architectures have been assessed for the TC center localization task, called VGG V1, VGG V2, VGG V3, and VGG V4, which differ in the convolutional block complexity (i.e., composition and number of parameters).

Starting from the aforementioned four VGG architectures, a total of 13 different ML models were trained for the TC center localization task. Each model differs in terms of the hyperparameters configuration (e.g., loss function, kernel size) used in the training stage.

Two different loss functions were implemented and used. The first one is the *Mean Absolute Error* (MAE) between real and predicted coordinates. The second one is a custom loss defined by the authors for this study, called *Cyclone Classification Localization* (CCL) loss, which is a linear combination of the MAE, the *Binary Cross Entropy* (BCE) loss, and the Euclidean Distance (ED) between real and predicted coordinates (L2). CCL

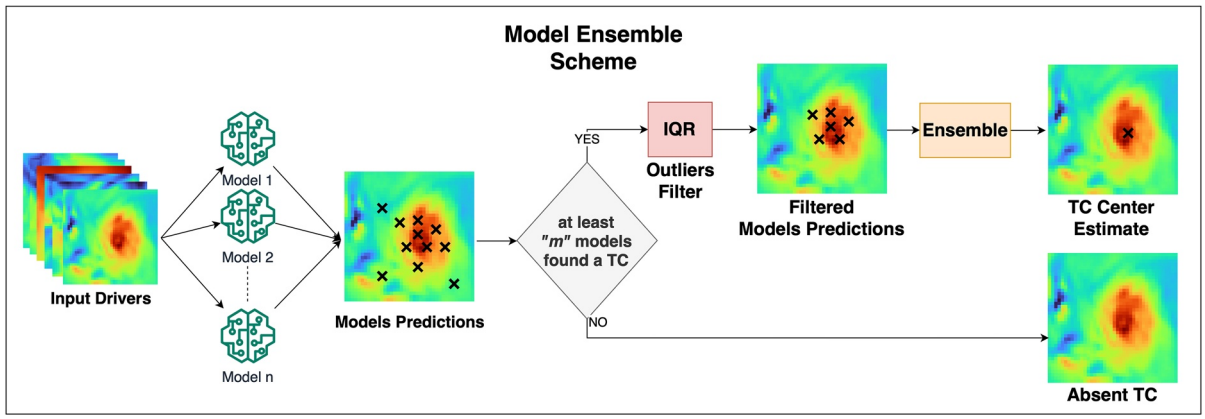

Figure 4. Diagram representing the models ensemble approach. All the *n* pre-trained models are fed with the same patches, yielding *n* (x,y) couples. If less than *m* (given) models localize the Tropical Cyclone (TC) center in the patch, the TC is considered as absent and the patch is labeled with negative coordinates. Otherwise, the Interquartile Range (IQR) algorithm is applied on the output of the models localizing the TC center; the final estimated location of the TC center is computed by averaging the values of the predictions not filtered by IQR.

tries to achieve two goals at once: (a) minimizing the classification error (through BCE) and (b) minimizing the localization error (through MAE and L2 terms).

For each of the 13 models, 25% of the training patches were used for validation purposes, whereas the remaining ones served the actual training. The six input drivers were normalized in the [0,1] range using min-max normalization, and augmented according to the data augmentation procedure presented in Section 2.3.1.

More details on the VGG architectures proposed in this work and the ML training are reported in Appendix A.

2.3.3. Metrics for Evaluating the ML TC Center Localization

The test set was used to evaluate the generalization capabilities of the trained ML models on out-of-sample data. Hereafter we define as True Positives (TPs) the TC center occurrences in IBTrACS correctly identified by the ML models; False Negatives (FNs) are the TC center occurrences in IBTrACS that the ML models are unable to identify, and False Positives (FPs) as the TC centers incorrectly identified by the ML models. For TC centers correctly classified as positive by the ML models, the *ED* is evaluated. Moreover, to understand the skill of the ML models in identifying TC centers, the following two metrics are computed: *Hit Rate* (see Equation 1), representing the rate of the actually identified TC centers with respect to the observed TC center occurrences in IBTrACS. Additionally, the F_2-score (see Equation 2) was computed in order to weight more missed TC center identifications (i.e., FNs), rather than those incorrectly identified (i.e., FPs).

$$Hit\ \ Rate = \frac{TP}{TP + FN} \tag{1}$$

$$F_2 - score = \frac{5TP}{5TP + 4FN + FP} \tag{2}$$

2.3.4. Consensus and Models Ensemble

Since the 13 models are trained with a different set of hyperparameters and/or layers configuration, each of them learns different characteristics and high-level features in the training set patches. Therefore, an ensemble approach (Ganaie et al., 2022) has been assessed in this study to combine the predictions made by different models with the aim of improving the overall accuracy skills (see Figure 4).

As depicted in Figure 4, for each patch of the test set, the approach consists in evaluating first how many models agree on classifying it as positive (i.e., the TC center is identified). An additional hyperparameter—*m* in Figure 4—was introduced to define the minimum number of models that need to be in agreement about the presence of a TC center in the input patch. Therefore, *m* represents the required level of consensus among the 13 models.

After a trial and error procedure on the validation set, this parameter was set to 7 in order to maximize the Hit Rate, given that a lower number of FNs is preferable for the task of predicting the occurrence of such extreme events. Each of the 13 models can potentially provide very different estimates about the location of the TC center for the same input patch. Therefore, the Interquartile Range (IQR) method was adopted as a further filtering step to keep only the estimates closer to their median value. In particular, the method consists in considering as outliers those TC center estimates (x) that satisfy the following inequality:

$$x < Q_1 - 1.5 * IQR \vee x > Q_3 + 1.5 * IQR \quad (3)$$

Indeed, the IQR is computed as the difference between the third (Q_3) and the first (Q_1) quartile, providing information about the spread of the data around the median value. Finally, the localization of the TC center is performed as the ensemble average of the (x,y) estimates of inliers.

According to the proposed ML ensemble approach based on the consensus procedure, the probability of observing a TC center given an input patch depends on the number of ML models that reached the minimum level of consensus (i.e., $m = 7$). This means, for example, that if 10 models out of 13 participating in the ML ensemble are in agreement in a patch ($10 > m = 7$), the probability of observing a TC center is $10/13 = 77\%$, which is greater than $7/13 = 54\%$. Clearly, the probabilities of observing a TC center could be better assessed when the number of involved models in the ensemble increases.

2.3.5. Hybrid Tracking Scheme

Starting from the individual TC center locations identified by the ML ensemble approach, a deterministic tracking algorithm based on the work by Scoccimarro et al. (2017) and Zhao et al. (2009) was used for reconstructing TC trajectories. The combination of the proposed data-driven TC center localization approach (see Section 2.3.4) with the aforementioned deterministic tracking algorithm resulted in a hybrid tracking scheme.

The deterministic tracking algorithm from Scoccimarro et al. (2017) and Zhao et al. (2009) implements the following steps:

- For each TC center location identified by the ML ensemble, the algorithm checks the presence of other identified TC centers after 6 hr at a distance of less than 400 km. The trajectory is considered complete if no TC center is found. Otherwise, if multiple TC centers are found in the next 6 hr, the closest one is considered as belonging to the same trajectory. This procedure is iteratively repeated until the track is complete.
- To mark the reconstructed trajectory as potentially valid, two conditions need to be verified: the trajectory (a) should last 3 days or more and (b) have a maximum surface wind speed larger than 17 ms^{-1} during at least 3 days (not necessarily consecutive) over an $8° \times 8°$ region centered on the middle of the TC.

2.3.6. Validation of the Hybrid Tracking Scheme

Reconstructed TC tracks are then associated with observations provided by IBTrACS, according to the procedure described in Bourdin et al. (2022). In particular, a detected track D is composed of n points (d_1, d_2, …, d_n) defined at times (t_1, t_2, …, t_n). Similarly, an observed track O from IBTrACS consists of a collection of points at given times. The tracks matching algorithm associates each point $d_i(t_i)$ of track D to those points of O at time t_i that are closer than 300 km from point d_i. It is worth noting that such points might not have a match in the set O of observed tracks. According to the formalism in Bourdin et al. (2022), the subset of points of O that have been associated with any point in D is denoted as $O_{D-paired}$, and its cardinality is indicated to $|O_{D-paired}|$. Three cases can be distinguished:

1. $|O_{D-paired}| = 0$: if no points in the reconstructed track D has a correspondence with a point in O, D is considered to be a False Alarm (FA);
2. $|O_{D-paired}| > 0$ and all the points in $O_{D-paired}$ belong to the same observed track in O, then this observed track is the best match of D;
3. $|O_{D-paired}| > 0$ and the points in $O_{D-paired}$ belong to multiple observed tracks in O, then the longest observed track is considered the best match of D.

Moreover, a further refinement is performed when an observed track is paired with two or more detected tracks. In this case, the detected tracks are merged into a single track. This allows taking into account situations in which the TC temporarily weakened before strengthening again (Bourdin et al., 2022).

Table 1
Average Metrics Over the Test Set for Each of the 13 Models

#	Model type	Loss	Kernel size	Euclidean distance (km)	F_2-score	Hit rate (%)
1	VGG V1	mae	3	128.94	0.50	89.69
2	VGG V2	mae	3	145.13	0.37	91.70
3	VGG V3	mae	3	151.84	0.34	91.31
4	VGG V4	mae	3	115.70	0.55	80.27
5	VGG V1	ccl	3	125.81	0.52	87.95
6	VGG V2	ccl	3	152.03	0.40	90.93
7	VGG V3	ccl	3	163.62	0.37	91.48
8	VGG V4	ccl	3	122.44	0.47	83.94
9	VGG V4	mae	5	116.41	0.38	82.96
10	VGG V4	mae	7	120.05	0.52	80.91
11	VGG V4	mae	9	123.47	0.43	86.98
12	VGG V4	mae	11	131.28	0.41	90.17
13	VGG V4	mae	13	149.19	0.40	90.95
–	ML ensemble	–	–	117.06	0.53	88.91

Note. The ML ensemble skills for the TC center localization task are also reported for comparison.

The Probability of Detection (POD) and False Alarm Rate (FAR) are used for assessing the track detection skills. These metrics are computed considering the TC tracks reconstructed by the proposed hybrid tracking scheme. The POD and FAR metrics are defined as follows:

$$POD = \frac{H}{H + M} \tag{4}$$

$$FAR = \frac{FA}{H + FA} \tag{5}$$

where H (Hits) refers to tracks detected from ERA5 data and actually present in IBTrACS; M (Misses) are tracks not detected from ERA5 data but present in IBTrACS, whereas FAs are incorrectly detected tracks that do not have a counterpart in IBTrACS.

Section 3.2 reports the POD and FAR metrics provided by the hybrid tracking scheme for comparison with four deterministic trackers (UZ, OWZ, TRACK and CNRM) reported in Bourdin et al. (2022).

3. Results

The following subsections present (a) the results achieved by the ML ensemble approach for localizing TC centers, (b) the comparison of the hybrid tracking scheme with four deterministic TC trackers from literature, and (c) the outcomes of the hybrid tracker on two test cases.

3.1. TC Center Localization Through the ML Ensemble Approach

Table 1 summarizes the averaged results produced by the 13 models on the test set, according to the evaluation metrics presented in Section 2.3.3. These 13 models are involved in the ensemble approach described in Section 2.3.4.

From the results reported in Table 1, it can be inferred that increasing the complexity of VGG architectures (models #1 to #3) resulted in an increased ED between the observed and estimated TC center locations. However, the increase in such localization error corresponds to an increase in the Hit Rate which is beneficial (i.e., higher is better). On the contrary, model #4, which corresponds to a VGG V4 architecture, has the lowest ED compared to models #1 to #3 but at the cost of a lower Hit Rate. Moreover, model #4 shows the highest value of F_2-score. Nevertheless, this value is not very high due to the test set being strongly imbalanced toward negative patches (as described in Section 2.3.1).

The lowest ED resulting from the VGG V4 architecture can be attributed to its different convolutional blocks composition, such as Batch Normalization and Dropout layers, that were not adopted in the design of the other VGG architectures. When the CCL loss is used in place of the MAE (i.e., Models #5 to #8) these results still hold. All the models from #1 to #8 were trained with a kernel size of 3.

Focusing on the VGG V4 architecture, different kernel sizes were also assessed to understand if the Hit Rate could be improved while keeping the ED low. As it can be noticed from Table 1, as the kernel size increases, the Hit Rate also increases at the cost of a higher ED. Therefore, it is clear that the localization accuracy and the Hit Rate are in trade-off and reaching the right compromise is difficult when a single ML model is used. To this end, the ML ensemble approach provided the best compromise in terms of Hit Rate (88.91%) and ED (117.06 km), while also achieving one of the highest F_2-score (0.53).

Figure 5 shows the ML ensemble approach applied on three different patches during the evolution of John TC (11 August to 13 September 1994), overlaid on the *msl* variable. Each row in Figure 5 refers to a specific time step of John's lifetime, whereas each column describes a particular step of the ML ensemble approach for locating TC centers. Starting from the top row in the Figure, Panel (a) reports the TC center estimates provided by all 13 models (red squares). In this case, the minimum consensus of 7 is reached (see Section 2.3.4). In

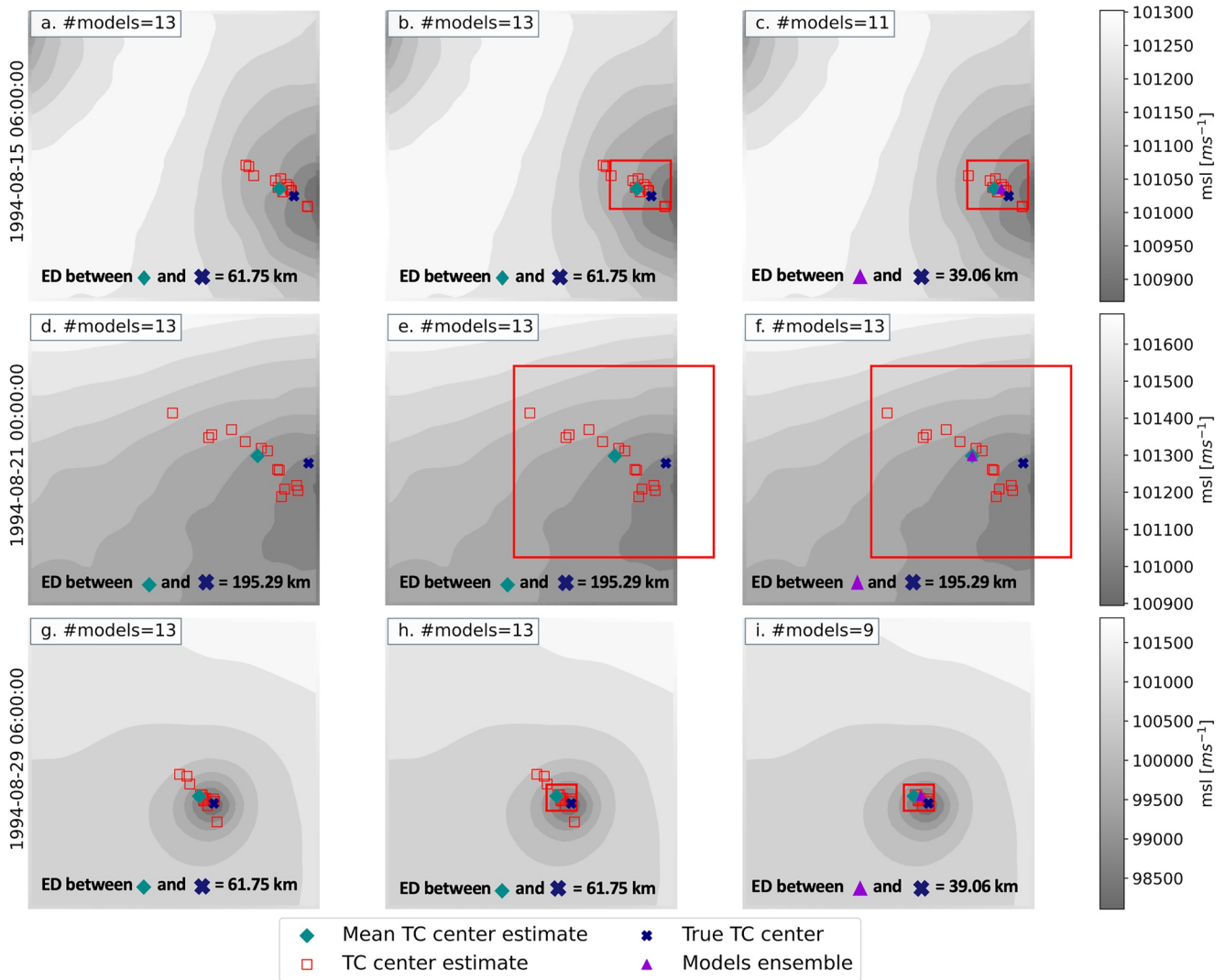

Figure 5. Machine Learning (ML) ensemble approach applied on three different time steps of John Tropical Cyclone (TC) lifetime (rows), overlaid on the mean sea level pressure (*msl*) variable. In each row, panels represent a particular stage of the proposed procedure. The number of models involved is reported in each panel and their TC center estimates are depicted as red squares, while the actual TC center is represented as a dark blue cross. Their average is reported as a green diamond. In the center panels (b), (e), and (h), the Interquartile Range is applied to detect outliers among models' TC center estimates. In the right panels, the model ensemble average (purple triangle) is computed only considering inlier values. The Euclidean Distance (ED) between the mean TC center and the true TC center is reported for panels (a), (b), (d), (e), (g), and (h), while the ED between the model ensemble and the true TC center is reported for panels (c), (f), and (i).

terms of localization error, the mean TC center estimate of the 13 models (green diamond) is 61.75 km far from the observed TC center (dark blue cross). In Panel (b), the IQR method (see Section 2.3.4) allows detecting 2 outliers out of 13. The outliers are depicted as the red squares outside the red box, which represents the Q1 and Q3 bounds of the IQR method. Therefore, Panel (c) reports only 11 remaining inliers (red squares inside the red box) along with the TC center estimate provided by the remaining 11 models in the ML ensemble (purple triangle). Then, by filtering out the outliers, the IQR method allowed reducing the distance between the ML ensemble TC center estimate and the observed TC center to 39.06 km, resulting in a 37% improvement with respect to the initial mean estimate. The same procedure also holds for the examples reported in Panels (g)–(i), where 9 models out of 13 were detected inliers by the IQR method, yielding a localization improvement of 36%. A different situation is represented in Panels (d)–(f), where the IQR method did not detect any outlier (i.e., all the red squares are inside the red box). Therefore, all the 13 models are involved in the ML ensemble to estimate the TC center location. The overlay of the results with the MSLP (*msl*) variable allows explaining why TC center estimates are spread, as in Panels (d)–(f). When the spatial

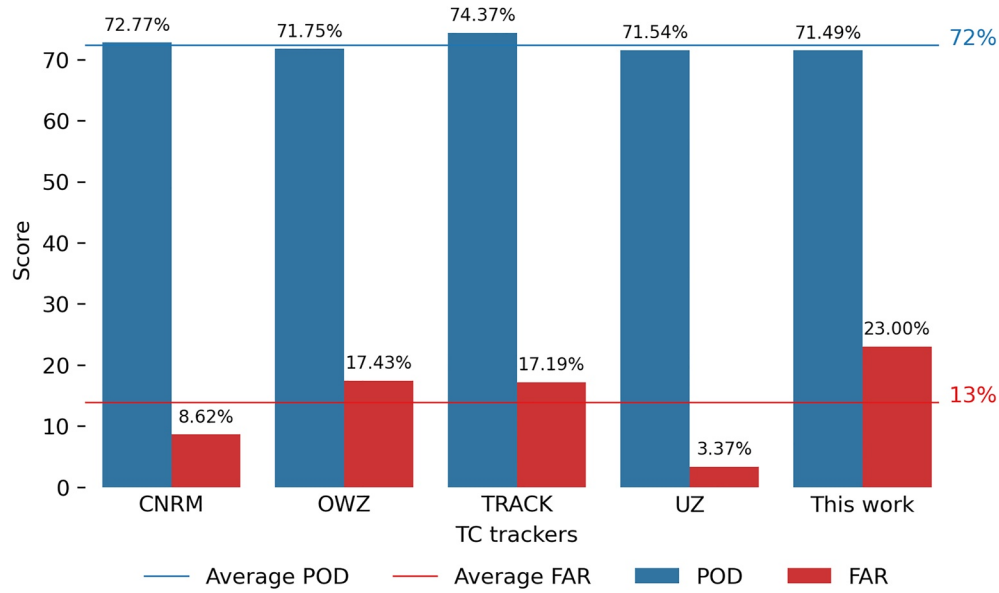

Figure 6. Intercomparison of Probability of Detection (POD) and False Alarm Rate (FAR) metrics after post treatment among four deterministic trackers (adapted from Bourdin et al., 2022) and the hybrid tracker proposed in this work. The horizontal lines show the average values among all the trackers. POD (higher is better) and FAR (lower is better) were computed over the 1980–2019 period for the joint West North Pacific, East North Pacific and North Atlantic basin covered in this study.

patterns of the *msl* variable are clearly defined (e.g., Panels (a)–(c) and Panels (g)–(i)), the ML models were able to accurately locate the TC center. In the other cases (e.g., Panels (d)–(f)), where instead spatial patterns of the *msl* are not clearly defined, although all the ML models are still able to detect the presence of a TC, their accuracy is lower, which results in a wider spread in the TC location estimates. The same behavior also applies to the other variables used for the model training (i.e., relative vorticity at 850 mb, 10 m wind gust since previous post-processing, instantaneous 10 m wind gust, MSLP, temperature at 300 and 500 mb), as reported in Appendix B.

To further investigate such behavior, a qualitative analysis of the ML ensemble approach was conducted during various phases of a different TC lifecycle. In particular, the Chantal TC (September, 10–15 September 1983) was analyzed during three different stages of its evolution characterized by varying intensities of MSW (as registered in IBTrACS). As explained for the John TC, also in this case, when the spatial circular patterns of the *vo850* become more evident around the TC center, the localization error between predicted and observed TC center locations is lower. Detailed results are reported in Appendix C.

3.2. Hybrid Tracking Scheme Skills

The results obtained by the proposed hybrid tracking scheme (Section 2.3.5) are here presented in terms of POD and FAR (Section 2.3.6), defined in Equations 4 and 5, respectively. A POD of 71.49% and a FAR of 23% were obtained by comparing the trajectories provided by the hybrid tracker with those reported in IBTrACS. The evaluation was performed on the joint Western North Pacific, East North Pacific and North Atlantic basin over 40 years of data (1980–2019 period). Specifically, the joint basin has been selected because it includes the highest number of TC tracks with respect to all the other basins (as reported in Bourdin et al., 2022).

Figure 6 compares the POD and FAR metrics of the hybrid approach with those achieved by four deterministic trackers (UZ, OWZ, TRACK and CNRM) from Bourdin et al. (2022). The results of POD and FAR related to the four deterministic trackers have been recomputed over the joint basin considered in this work. To provide a fair comparison with the deterministic TC trackers, the same IBTrACS data set including ET cyclones was used. This required using the Subtropical Jet Cut-off post-treatment method described in Bourdin et al. (2022) to filter out ET cyclones.

Table 2
Number of Hits, Misses and False Alarms (FA) for Each of the Four Deterministic Trackers and Our Hybrid Solution (Bold Values)

Tracker	Hit	Miss	FA	POD %	FAR %
CNRM	1,622	607	153	72.77	8.62
OWZ	1,648	649	348	71.75	17.43
TRACK	1,831	631	380	74.37	17.19
UZ	1,604	638	56	71.54	3.37
Hybrid (this work)	**1,630**	**650**	**487**	**71.49**	**23.00**

Note. Value of Probability of Detection (POD) and False Alarm Rate (FAR) are also reported.

We found that the POD value is almost identical to most of the deterministic trackers compared, while the FAR value is in-line (but slightly higher) with two of the trackers considered (OWZ and TRACK). It is important to remark that the focus of this work concerns the localization of TC centers, as shown by the high number of hits achieved by the ML ensemble approach (see Table 2). Moreover, the number of misses is aligned with the other trackers. Thus, the integration of the ML ensemble solution with the deterministic tracking scheme shows already promising results. Nevertheless, the implementation of a full data-driven tracking scheme is foreseen as future work in order to improve the tracker skills, in particular to reduce the number of FAs.

3.3. Test Cases: Keoni and Julio Tropical Cyclones

The Keoni and Julio TCs were selected as test cases because (a) they occurred in the domain of interest, and (b) they were long-lasting cyclones. Moreover, the MSW was available for all the IBTrACS records of these two cyclones. In this section, we focus on analyzing the results of the hybrid tracker proposed in this work, whereas Appendix D provides more in-depth details on the TC center localization skills through the ML ensemble approach for the two selected cyclones.

The trajectories resulting from the hybrid tracking scheme and the observations from IBTrACS are reported in Figure 7 for the Keoni TC and in Figure 8 for the Julio TC. Both figures are organized as follows: the left side shows the two tracks overlaid, whereas the right side shows, in both panels, the values of the six variables considered in this work (*msl, i10fg, fg10, t300, t500, vo850*), when selecting two TC center locations at different stages of their development.

3.3.1. Keoni TC

The Keoni TC occurred from 9 August to 4 September 1993. During its lifecycle, the TC became a hurricane characterized by strong winds that reached 115 knots of speed on the 16 August before starting to lose intensity from the 19 August until early September (Tropical Cyclones 1993 NOOA report, 1993).

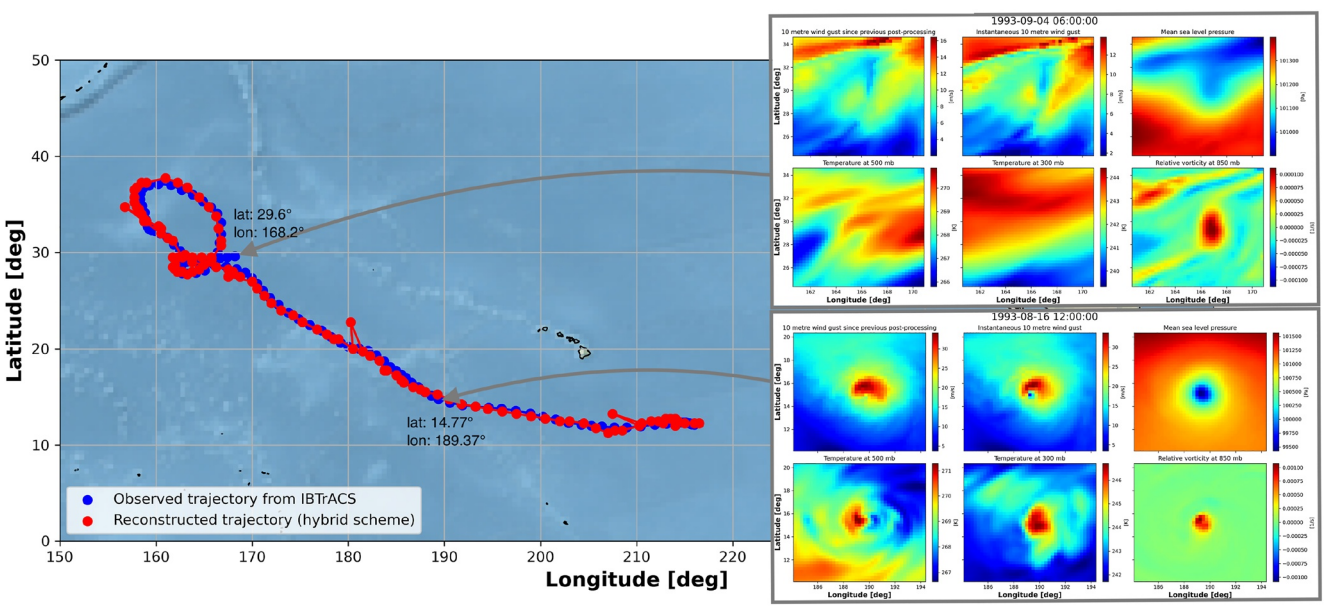

Figure 7. Comparison between the trajectories resulting from the hybrid tracking scheme proposed in this work and the observations from International Best Track Archive for Climate Stewardship (IBTrACS) for the Keoni Tropical Cyclone (TC). The left side shows the two tracks overlaid, while the right side shows, in both panels, the values of the six variables considered in this work (*msl, i10fg, fg10, t300, t500, vo850*). In particular, the upper and lower panels display the variables patterns for cyclones classified in IBTrACS as *Not Reported* TC and *Tropical Storm*, respectively.

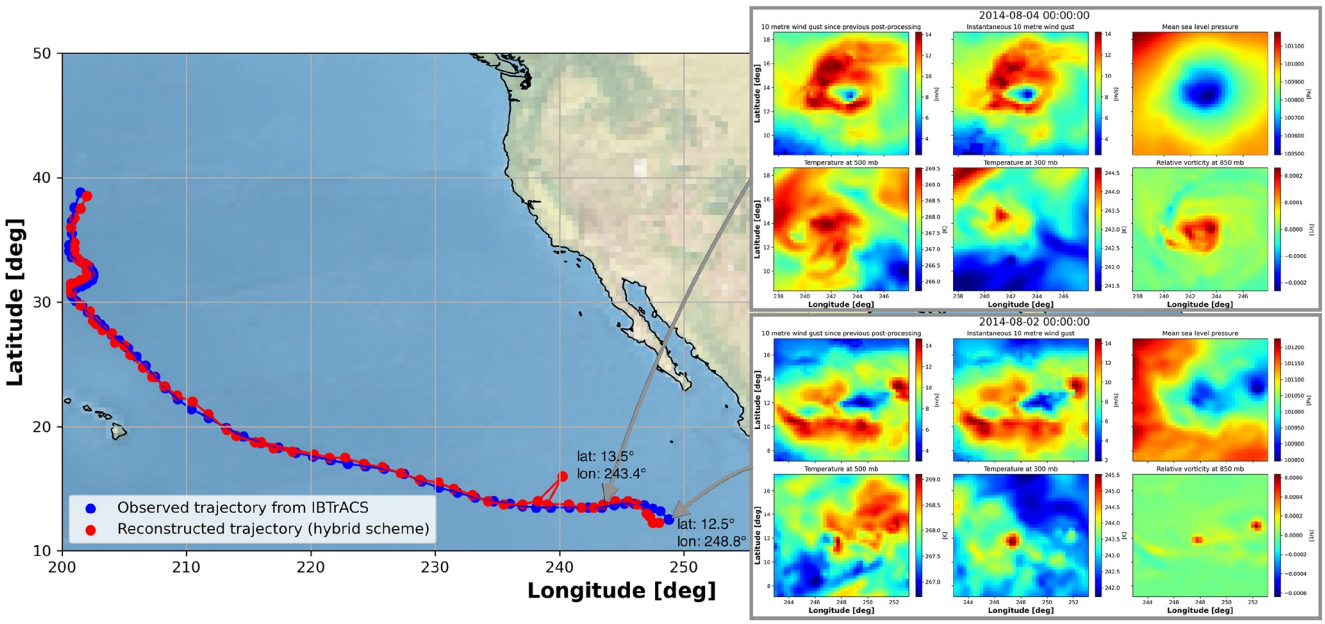

Figure 8. Comparison between the trajectories resulting from the hybrid tracking scheme proposed in this work and the observations from International Best Track Archive for Climate Stewardship (IBTrACS) for the Julio Tropical Cyclone. The left side shows the two tracks overlaid, while the right side shows, in both panels, the values of the six variables considered in this work (*msl, i10fg, fg10, t300, t500, vo850*). In particular, the upper and lower panels display the variables patterns for cyclones classified in IBTrACS as "Tropical Storm" and "Disturbance Storm," respectively.

Figure 7 shows that the hybrid tracker is able to closely follow the trajectory of the observed TC reported in IBTrACS. More in detail, looking at the single track points, when the TC is classified as *NR* in IBTrACS (point at lat 29.6° and lon 168.2°), the spatial patterns of the variables are not clearly defined and the accuracy of the TC center localization achieved by the ML ensemble is low: the estimate is located at lat 29.5° and lon 165.75°, 67.72 km away from the observation. On the other hand, when the cyclone gains more strength, and it is classified as *Tropical Storm* in IBTrACS (point at lat 14.77° and lon 189.37°), the spatial patterns of the input variables around the TC center become more evident and the localization accuracy achieved by the ML ensemble is higher: the estimate is located at lat 15.25° and lon 189.25°, 13.56 km away from the observation. Moreover, as it can be seen in Figure 7, the approach estimates some TC centers which are out-of-trajectory. The out-of-trajectory TC centers estimates were localized, for a specific time-step, in the adjacent patches. The reason for the poor TC centers localization can be attributed to tiling the domain into non overlapping patches, which, even though representing a limitation, does not have a major impact on the overall trajectory reconstruction. For example, being each patch of size 40 × 40 pixels (i.e., about 1,000 × 1,000 km, see Section 2.2.2), the rightmost out-of-trajectory TC center estimate in Figure 7 should have been localized in the patch with bounding box at 210°–220° longitude and 10°–20° latitude, but it was localized in the adjacent patch (with bounding box at 200°–210° longitude, 10°–20° latitude). Nevertheless, this out-of-trajectory TC center has been identified by the hybrid tracking scheme (see Section 2.3.5), as it satisfies the first step of the deterministic tracking algorithm used in this work (from Scoccimarro et al., 2017; Zhao et al., 2009), which states that for each TC center, the next one is considered as belonging to the same trajectory if located within 400 km after 6 hr, as it can potentially occur in a neighboring patch. For this specific case, the number of out-of-trajectory TC centers in the whole reconstructed trajectory (in red) is very low (just 3 points out of 111).

3.3.2. Julio TC

As second test case, the Julio TC was considered. It occurred from 2–18 August 2014. During its lifecycle, the TC became a "hurricane" characterized by strong winds that reached 105 knots of speed on the 8 August before starting to lose intensity. It was classified as "Disturbance Storm" from the 15 August (Stewart & Jacobson, 2016).

Similarly to the Keoni test case, Figure 8 shows that the hybrid tracker is able to closely follow the trajectory of the observed TC reported in IBTrACS. Also in this case the approach estimates a few out-of-trajectory TC centers (i.e., a single point out of 68) due to the patching procedure, as explained for Keoni TC. Looking at the single

track points, when the TC is classified as *Disturbance Storm* in IBTrACS (point at lat 12.5° and lon 248.8°), although the spatial patterns of the variables are not yet clearly defined, the accuracy of the TC center localization achieved by the ML ensemble is good: the estimate is located at lat 12.25° and lon 248.0°, 23.15 km away from the observation. This is an unexpected result since *DS* (as well as *NR* TCs—see Keoni test case) were filtered out from the training set (see Section 2.2). Such locations were not considered during training to avoid pushing ML models to learn these patterns as concepts. Their inclusion would have polarized the outcomes toward such samples accounting for about 15% of the entire data set. Anyway, we believe that the ability of the ML ensemble to also locate these phenomena can be considered as an added value of our approach.

When the cyclone gains more strength, and it is classified as *Tropical Storm* in IBTrACS (point at lat 13.5° and lon 243.4°), the spatial patterns of the input variables around the TC center become evident and the localization accuracy achieved by the ML ensemble is higher: the estimate is located at lat 13.5° and lon 242.75°, 17.95 km away from the observation. Appendix D provides additional details on the TC center localization skills.

4. Conclusion and Discussion

The present study proposed a ML ensemble approach aimed at localizing TC centers in terms of geographical coordinates. The ensemble combines TC center estimates of different ML models that agree about the presence of a TC in input data. Moreover, a hybrid tracking scheme was defined integrating the aforementioned ML ensemble approach with a deterministic tracking algorithm for reconstructing TC trajectories.

Given the inherent complexity of the TC centers localization, trusting the estimate of their position through a single ML model would have led to unreliable results. Therefore, an ensemble approach was proposed to integrate the knowledge learned by different ML models that are trained for the same localization task. The ensemble relies on 13 VGG-like architectures that are trained with distinct hyperparameters configurations on the same input-output pairs. This allows extracting different intrinsic patterns and features related to the TC evolution during its lifetime, as well as reducing the uncertainty associated with the estimate of the TC center position. The present approach is extendable either by adding new ML models to the ensemble or by fine-tuning the current ones to get better skills (i.e., ED with respect to observations and Hit Rate).

ERA5 reanalysis data concerning six input climatic drivers was jointly exploited with IBTrACS historical records to train and test the designed models. Reanalysis data combines model simulations with observations to provide the best state representation of different climatic variables in the past.

However, as recognized by Hodges et al. (2017), no assimilation of TCs is performed in ERA5, unlike other reanalyzes such as JRA-55 or NCEP-CFSR data sets. Nonetheless, the ensemble exhibits good accuracy in locating TC centers, specifically providing 88.91% of Hit Rate and an ED of 117.06 km with respect to IBTrACS observations (Section 3.1).

Roberts et al. (2020) and Zarzycki et al. (2021) evaluated a series of metrics on ERA5 with comparable performance to JRA-55 and NCEP-CFSR: the main reason can be found in the enhanced resolution of ERA5 with respect to the previous ERA-Interim product. This motivated the use of ERA5 reanalysis for the six input climate drivers in this study. Moreover, the presented processing methodology of ERA5 maps led to non-cyclone-centric input patches, that is, 40 × 40 images in which the TC center can occur in any position, not necessarily in its center. In this way, ML models were able to learn the drivers spatial patterns and characteristics related to the presence of the TC inside the patch, regardless of its position. As a result, beyond *Tropical*, *Subtropical* and *ET* storms, the ensemble was also capable of localizing the centers of *NR* and *DS* cyclones with a low error, even though they were not included in the training set, thus demonstrating the good generalization capabilities of the proposed approach. Furthermore, tiling ERA5 maps into non-overlapping patches of fixed size allowed ML models to detect multiple TCs that can simultaneously occur in the joint North Atlantic and Pacific geographical domain covered in this study. The application of the proposed approach to other formation basins was not assessed in this study and will be subject to future investigation.

Additionally, the proposed hybrid tracking scheme was compared with four trackers from literature (UZ, OWZ, TRACK, and CNRM) over 40 years of ERA5 reanalysis data (1989–2019), and considering the joint Western North Pacific, East North Pacific and North Atlantic basin. We found a POD value of 71.49%, that is almost identical to most of the deterministic trackers compared, and a FAR value of 23%, in-line (but slightly higher)

with two of the trackers considered (OWZ and TRACK). Therefore, the integration of the ML ensemble solution with the deterministic tracking scheme shows already promising results.

Concerning the limitations of the present research, it is important to take into account uncertainties related to IBTrACS and ERA5 data that may lead to biased TC center positioning. In particular, IBTrACS provides TC center geographical coordinates aligned on a $0.1° \times 0.1°$ grid and with an uncertainty that is inversely proportional to the storm intensity (IBTrACS Science Team, 2019). ERA5 maps, on the other hand, are provided on a $0.25° \times 0.25°$ grid. Therefore, TC centers were aligned on the ERA5 grid, as a preprocessing step (Section 2.2.2). As a result, all the sources of uncertainty implicitly affected both ML training and inference. As higher resolution reanalysis data will become available, the inherent uncertainty will also be reduced.

It is important to remark that the proposed study focused mainly on the use of an ensemble of ML models for localizing TC centers. The proposed hybrid tracker represents a first effort toward addressing the overall tracking process with a data-driven solution.

As future work, we plan to compare the hybrid tracking scheme on the whole global domain, as well as to extend the comparison with additional deterministic trackers such as the TStorm from the National Oceanic and Atmospheric Administration (https://www.gfdl.noaa.gov/tstorms/) and the GFDL-vortex from the Geophysical Fluid Dynamics Laboratory https://dtcenter.org/community-code/gfdl-vortex-tracker. The implementation of a full data-driven tracking scheme is also envisaged as future work. Moreover, the topic of uncertainty and probabilistic forecasting will also be covered, as well as the use of data processing techniques (such as using sliding tiles) to overcome the current limitation related to out-of-trajectory TC centers (see Section 3.3) and to further enhance the training data set with the aim of reducing FNs and FPs at the borders of the patches.

Finally, it is worth noting that the workflow for supporting TC tracking presented here is very complex, as it consists of heterogeneous data and software components. It requires large-scale data handling solutions, jointly with ML algorithms and access to High Performance Computing infrastructure. As next step, the authors aim to develop an integrated pipeline that can apply the pre-processing and ML model pipeline directly to the output of an ESM simulation. This effort is currently ongoing in the framework of the eFlows4HPC European project (https://eflows4hpc.eu/) (Ejarque et al., 2022) and interTwin project (https://www.intertwin.eu/). In the context of these two projects, we have been dealing with the design and development of efficient workflows for (near) real-time TC detection. In particular, we aim to support parallel execution of independent tasks (e.g., the ensemble model components) on large distributed data sets. Moreover, in the context of interTwin, Coupled Model Intercomparison Project experiments will be used with the aim of providing an indication of how climate change is going to affect TCs frequencies and locations in the future.

Appendix A: VGG-Like Architectures

This section provides additional details about the design of the VGG-like architectures considered in this study. Tables A1–A4 refer to VGG V1 up to V4 architectures, respectively, and report the characteristics of each convolutional block along with additional details, such as the number of filters, activation functions, the output shape and number of parameters. Starting from the VGG baseline architectures, a total of 13 different ML models were trained for the TC center localization task. Each model differs in terms of the hyperparameters configuration (e.g., loss function, kernel size) used in the training stage, as reported in Table 1. All the aforementioned models were trained for 500 epochs with a batch size of 8,192 patches, using the Adam optimizer (Kingma & Ba, 2014) with a learning rate of $1e^{-4}$.

Experiments were carried out exploiting the Juno hybrid cluster based on Central Processing Units (CPUs) and Graphics Processing Units (GPUs). Juno is the latest High Performance Computing systems available at the CMCC Supercomputing center. Juno delivers nearly 1.15 PetaFlops of peak performance and it is composed of 170 dual-processor nodes with a total of 12,240 cores and 87 TB of main memory. Each node is equipped with 2 Intel Xeon Platinum 8360Y 2.4 Ghz processors (36 cores each) and 512 GB of main memory. In particular, the cluster contains 10 dual-GPU nodes equipped with NVIDIA A100 GPUs. The storage I/O bandwidth is of 80 GB/s.

Regarding the software adopted for our experiments, data processing was performed through Pandas v1.5.3 (The Pandas Development Team, 2023) and Xarray v2022.6.0 (Hoyer & Hamman, 2017). The architecture and the training/test control flows were both written in Python v3.11.2 based on the Keras Application Programming Interface v2.12.0 (Chollet, 2015) and relying on the TensorFlow v2.12.0 (Abadi et al., 2015) back-end. Models training was performed in a distributed fashion by means of the TensorFlow distributed training and MirroredStrategy. The results of the present study were achieved by running the model on just one node of the Juno cluster exploiting the 2 GPUs available.

The training lasts about 2 hr and 40 min on average for all the models over 500 epochs with early stopping enabled. The inference time of the hybrid approach, consisting in the ML ensemble for TC center localization and the subsequent deterministic tracking scheme, is of few seconds up to a couple of minutes when tested over a single year of test data on a single CPU of the Juno supercomputer at CMCC. Clearly, since it is an embarrassingly parallel task, the execution scales almost linearly on multiple CPU cores when multiple input years/ML models are concurrently processed.

The source code for the ML TCs Detection and Tracking approach presented in this work is available at https://dx.doi.org/10.5281/zenodo.8321138 (Donno et al., 2023).

Table A1
Baseline VGG V1 Architecture

Block	Layer	# Filters	Activation	Output shape	# of parameters
0	Input	–	–	40 × 40 × 6	–
	Conv 3 × 3	64	ReLU	40 × 40 × 64	3,520
	Conv 3 × 3	64	ReLU	40 × 40 × 64	36,928
	Conv 3 × 3	64	ReLU	40 × 40 × 64	36,928
	MaxPool 2 × 2	–	–	20 × 20 × 64	–
1	Conv 3 × 3	128	ReLU	20 × 20 × 128	73,856
	Conv 3 × 3	128	ReLU	20 × 20 × 128	147,584
	MaxPool 2 × 2	–	–	10 × 10 × 128	–
2	Conv 3 × 3	256	ReLU	10 × 10 × 256	131,328
	Conv 3 × 3	256	ReLU	10 × 10 × 256	262,400
	MaxPool 2 × 2	–	–	5 × 5 × 256	–
3	Conv 2 × 2	512	ReLU	4 × 4 × 512	524,800
	Conv 2 × 2	512	ReLU	3 × 3 × 512	1,049,088
	Conv 2 × 2	512	ReLU	2 × 2 × 512	1,049,088
	MaxPool 2 × 2	–	–	1 × 1 × 512	–
4	Flatten	–	–	512	–
	Dense 512	–	ReLU	512	262,656
	Dense 256	–	ReLU	256	131,328
	Dense 128	–	ReLU	128	32,896
	Dense 64	–	ReLU	64	8,256
	Output	–	Linear	2	130
–	–	–	–	–	3,750,786

Table A2
Baseline VGG V2 Architecture

Block	Layer	# Filters	Activation	Output shape	# of parameters
0	Input	–	–	40 × 40 × 6	–
	Conv 3 × 3	32	ReLU	40 × 40 × 32	1,760
	Conv 3 × 3	32	ReLU	40 × 40 × 32	9,248
	Conv 3 × 3	32	ReLU	40 × 40 × 32	9,248
	MaxPool 2 × 2	–	–	20 × 20 × 32	–
1	Conv 3 × 3	64	ReLU	20 × 20 × 64	18,496
	Conv 3 × 3	64	ReLU	20 × 20 × 64	36,928
	Conv 3 × 3	64	ReLU	20 × 20 × 64	36,928
	MaxPool 2 × 2	–	–	10 × 10 × 64	–
2	Conv 3 × 3	128	ReLU	10 × 10 × 128	73,856
	Conv 3 × 3	128	ReLU	10 × 10 × 128	147,584
	Conv 3 × 3	128	ReLU	10 × 10 × 128	147,584
	MaxPool 2 × 2	–	–	5 × 5 × 128	–
3	Conv 2 × 2	256	ReLU	5 × 5 × 256	131,328
	Conv 2 × 2	256	ReLU	5 × 5 × 256	262,400
	Conv 2 × 2	256	ReLU	5 × 5 × 256	262,400
	Conv 2 × 2	512	ReLU	4 × 4 × 512	524,800
	Conv 2 × 2	512	ReLU	3 × 3 × 512	1,049,088
	Conv 2 × 2	512	ReLU	2 × 2 × 512	1,049,088
	Conv 2 × 2	512	ReLU	1 × 1 × 512	1,049,088
4	Flatten	–	–	512	–
	Dense 1,024	–	ReLU	1,024	525,312
	Dense 512	–	ReLU	512	524,800
	Dense 256	–	ReLU	256	131,328
	Dense 128	–	ReLU	128	32,896
	Output	–	Linear	2	258
–	–	–	–	–	6,024,418

Table A3
Baseline VGG V3 Architecture

Block	Layer	# Filters	Activation	Output shape	# of parameters
0	Input	–	–	40 × 40 × 6	–
	Conv 3 × 3	32	ReLU	40 × 40 × 32	1,760
	Conv 3 × 3	32	ReLU	40 × 40 × 32	9,248
	Conv 3 × 3	32	ReLU	40 × 40 × 32	9,248
	MaxPool 2 × 2	–	–	20 × 20 × 32	–
1	Conv 3 × 3	64	ReLU	20 × 20 × 64	18,496
	Conv 3 × 3	64	ReLU	20 × 20 × 64	36,928
	Conv 3 × 3	64	ReLU	20 × 20 × 64	36,928
	MaxPool 2 × 2	–	–	10 × 10 × 64	–

Table A3
Continued

Block	Layer	# Filters	Activation	Output shape	# of parameters
2	Conv 3 × 3	128	ReLU	10 × 10 × 128	73,856
	Conv 3 × 3	128	ReLU	10 × 10 × 128	147,584
	Conv 3 × 3	128	ReLU	10 × 10 × 128	147,584
	MaxPool 2 × 2	–	–	5 × 5 × 128	–
3	Conv 3 × 3	256	ReLU	5 × 5 × 256	295,168
	Conv 3 × 3	256	ReLU	5 × 5 × 256	590,080
	Conv 3 × 3	256	ReLU	5 × 5 × 256	590,080
	Conv 2 × 2	512	ReLU	4 × 4 × 512	524,800
	Conv 2 × 2	512	ReLU	3 × 3 × 512	1,049,088
	Conv 2 × 2	1,024	ReLU	2 × 2 × 1,024	2,098,176
	Conv 2 × 2	1,024	ReLU	1 × 1 × 1,024	4,195,328
4	Flatten	–	–	1,024	–
	Dense 1,024	–	ReLU	1,024	1,049,600
	Dense 512	–	ReLU	512	524,800
	Dense 512	–	ReLU	512	262,656
	Dense 256	–	ReLU	256	131,328
	Output	–	Linear	2	514
–	–	–	–	–	11,793,250

Table A4
Baseline VGG V4 Architecture

Block	Layer	# Filters	Activation	Output shape	# of parameters
0	Input	–	–	40 × 40 × 6	–
	Conv 3 × 3	32	–	20 × 20 × 32	1,728
	Gaussian Noise	–	–	20 × 20 × 32	–
	Batch Norm	–	LeakyReLU	20 × 20 × 32	128
1	Conv 3 × 3	64	LeakyReLU	10 × 10 × 64	18,432
2	Conv 3 × 3	128	–	5 × 5 × 128	73,728
	Dropout 0.5	–	LeakyReLU	5 × 5 × 128	–
3	Conv 3 × 3	256	–	3 × 3 × 256	294,912
	Gaussian Noise	–	LeakyReLU	3 × 3 × 256	–
4	Conv 3 × 3	512	LeakyReLU	2 × 2 × 512	1,179,648
0	Conv 3 × 3	1,024	–	1 × 1 × 1,024	4,718,592
	Batch Norm	–	–	1 × 1 × 1,024	4,096
	Dropout 0.5	–	LeakyReLU	1 × 1 × 1,024	–
4	Flatten	–	–	1,024	–
	Dense 1,024	–	ReLU	1,024	1,049,600
	Dense 512	–	ReLU	512	524,800
	Dense 256	–	ReLU	256	131,328
	Dense 128	–	ReLU	128	32,896
	Output	–	Linear	2	258
–	–	–	–	–	8,030,146

Appendix B: Additional Results of the ML Ensemble Approach

This section reports further results of the ML ensemble approach on the John TC described in Section 3.1, when overlaid on the other five variables considered in this study: 10 m wind gust since previous post-processing (*fg10*) (Figure B1), instantaneous 10 m wind gust (*i10fg*) (Figure B2), temperature at 300 mb (*t300*) (Figure B3), temperature at 500 mb (*t500*) (Figure B4), relative vorticity at 850 mb (*vo850*) (Figure B5). Similarly to the *msl* variable (see Section 3.1), when the spatial patterns of the variables are clearly defined, the ML models are able to accurately locate the TC center. In the other cases, although all the ML models are still able to detect the presence of a TC, their accuracy is lower, thus resulting in a wider spread in the TC locations.

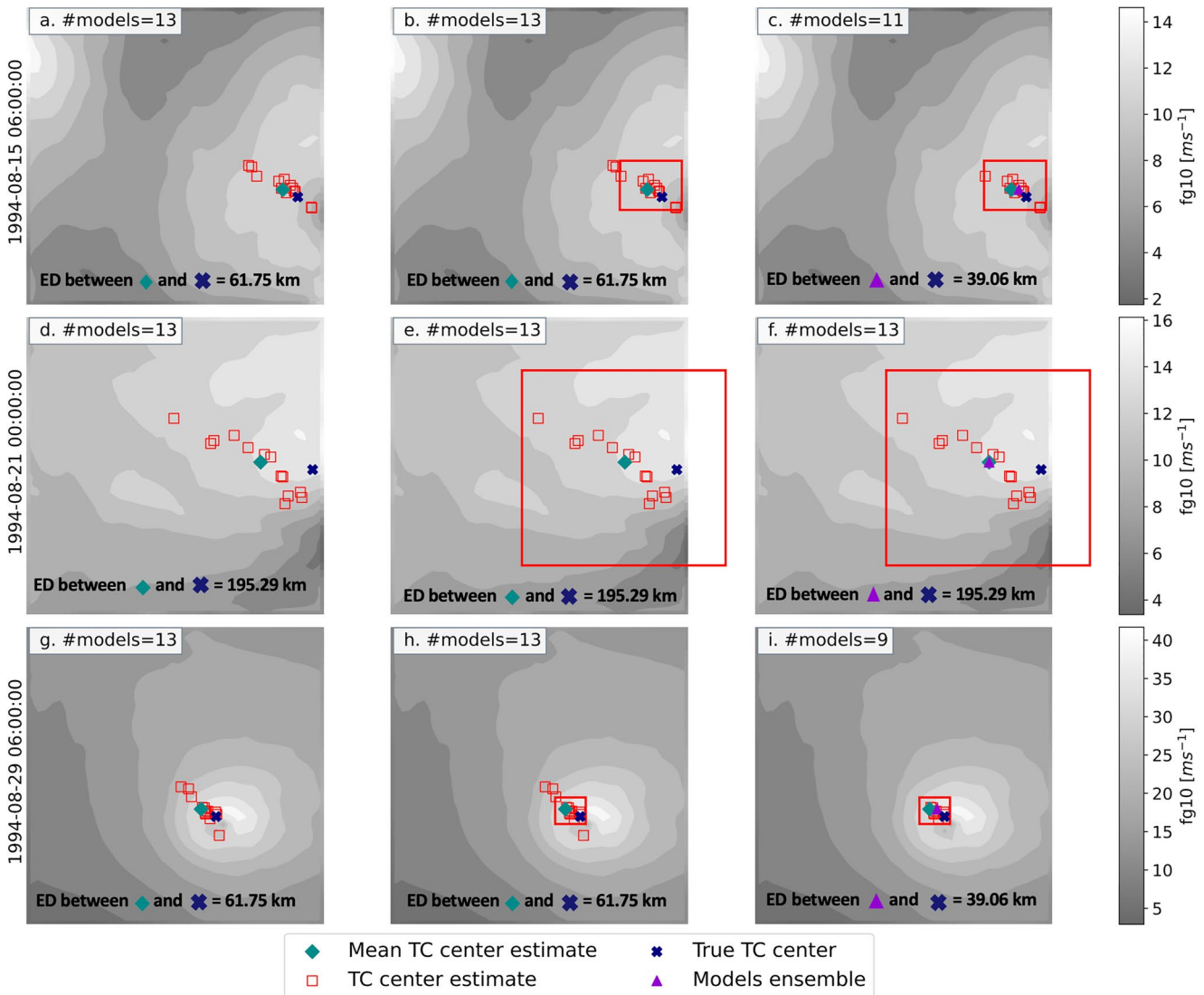

Figure B1. Machine Learning ensemble approach applied on three different time steps of John Tropical Cyclone (TC) lifetime (rows), overlaid on the *fg10* variable.

10.1029/2023EA003106

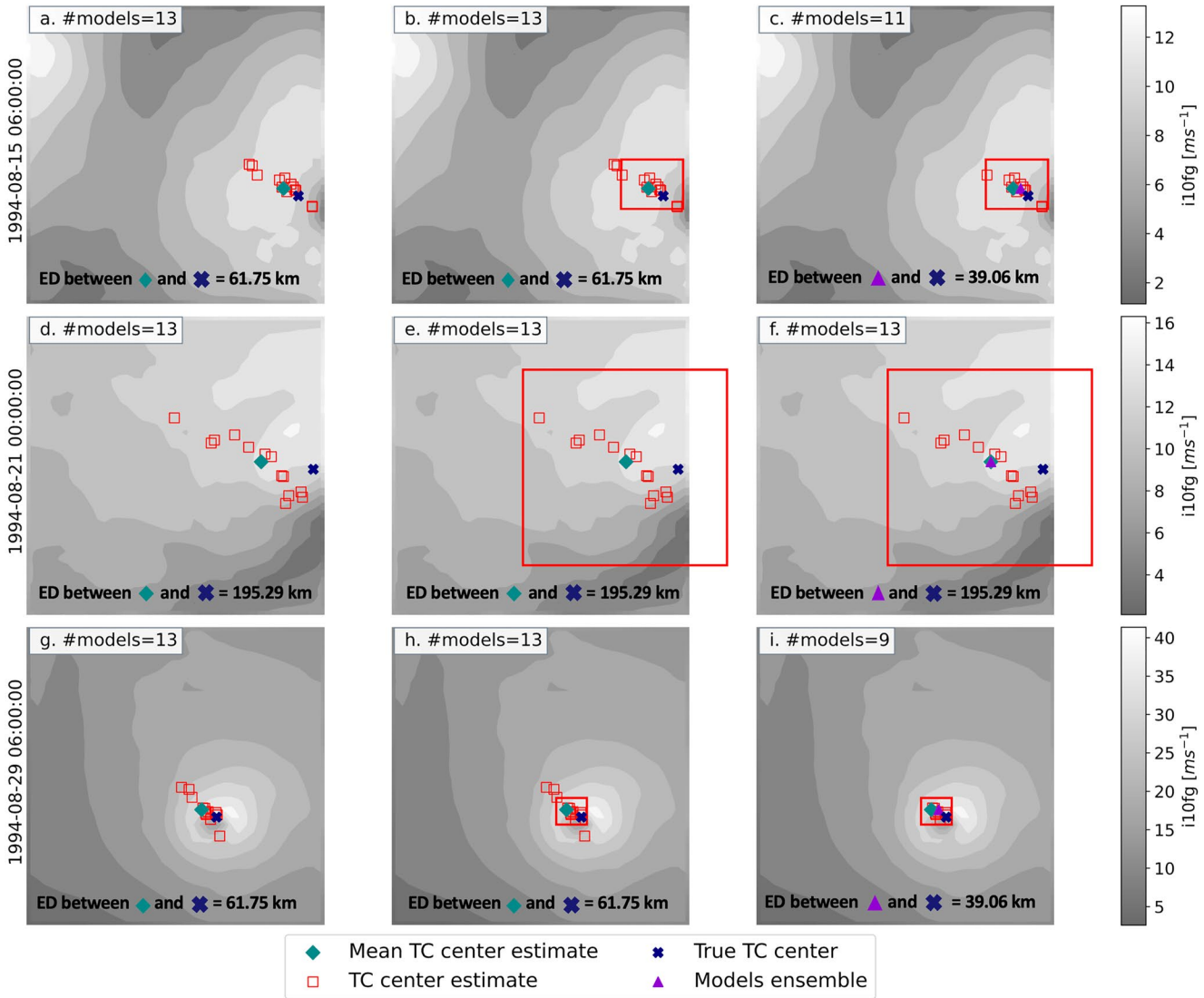

Figure B2. Machine Learning ensemble approach applied on three different time steps of John Tropical Cyclone (TC) lifetime (rows), overlaid on the *i10fg* variable.

10.1029/2023EA003106

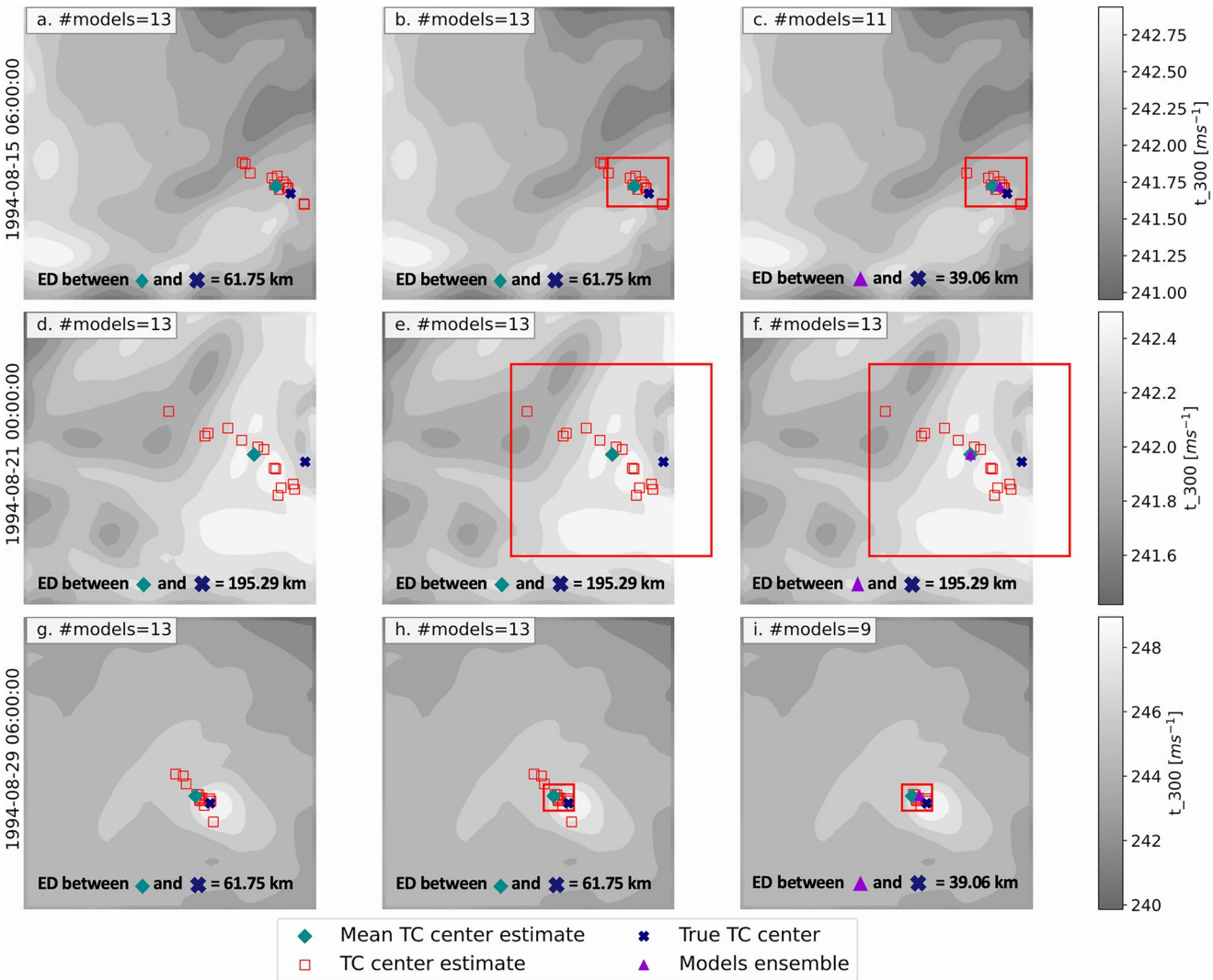

Figure B3. Machine Learning ensemble approach applied on three different time steps of Tropical Cyclone (TC) John lifetime (rows), overlaid on the *t300* variable.

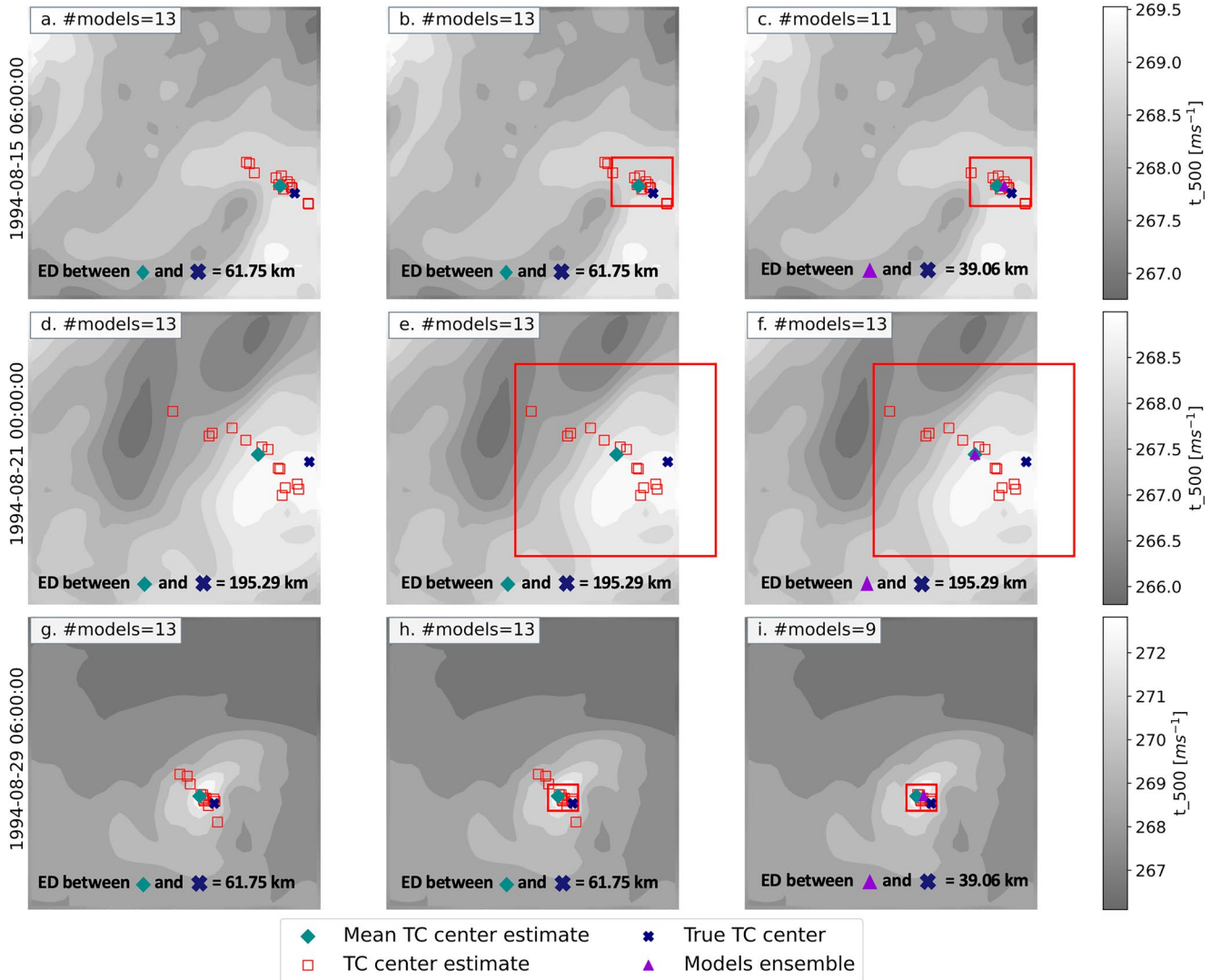

Figure B4. Machine Learning ensemble approach applied on three different time steps of John Tropical Cyclone (TC) lifetime (rows), overlaid on the *t500* variable.

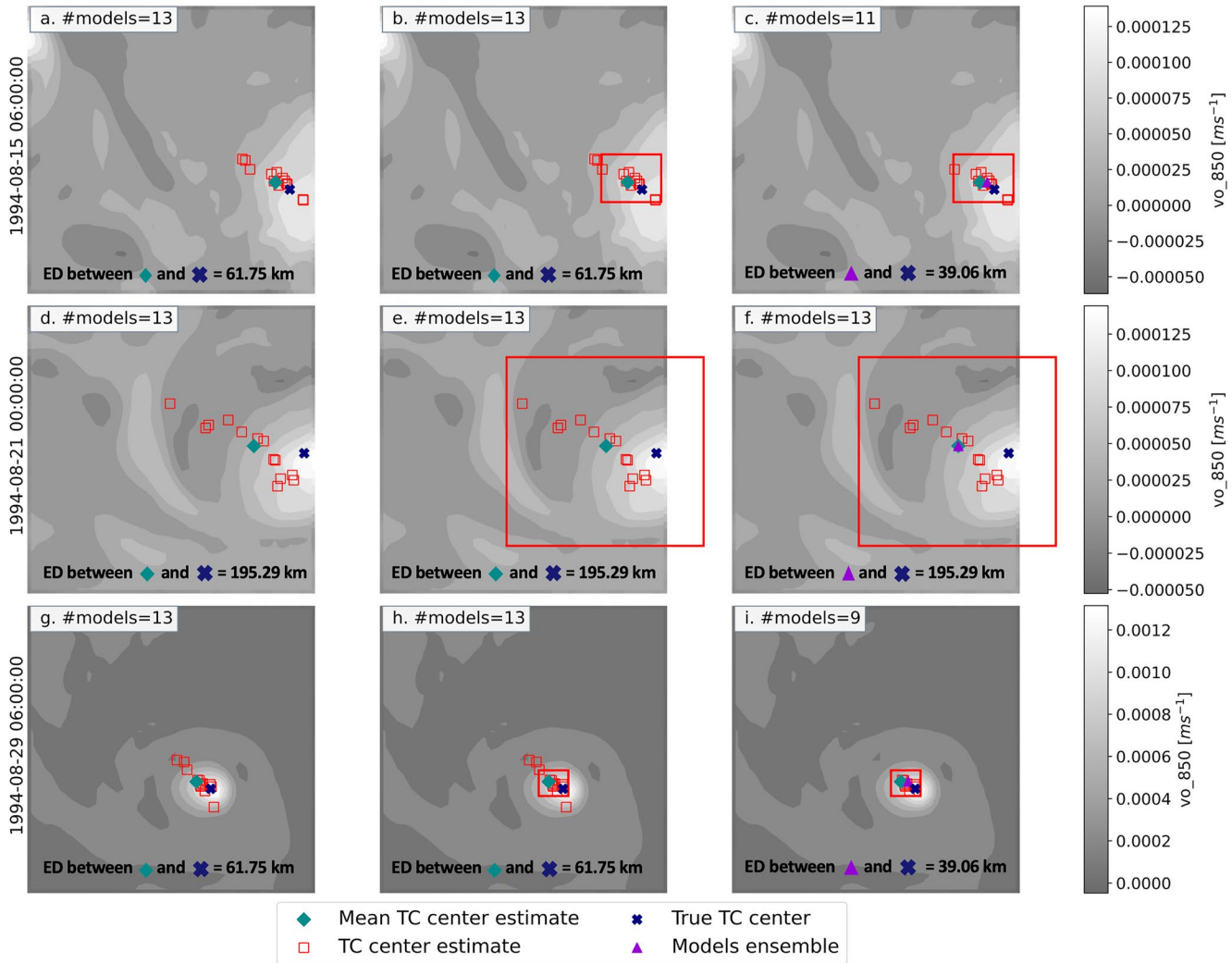

Figure B5. Machine Learning ensemble approach applied on three different time steps of John Tropical Cyclone (TC) lifetime (rows), overlaid on the *vo850* variable.

Appendix C: Qualitative Analysis of the Chantal TC Lifecycle

As an example, in Figure C1 models ensemble predictions are reported for the Chantal TC (10–15 September 1983), overlaid on the *vo850* input driver. In particular, three time steps over the Chantal lifecycle are shown, namely 11 September 1983 at 00.00, 12 September 1983 at 12.00 and 15 September 1983 at 06.00, respectively. In the early and final stages (i.e., (a) and (c) panels), the *vo850* variable does not show the typical circular spatial patterns surrounding the TC center and indeed the models ensemble struggles to accurately estimate the actual TC center (blue cross). This results in spread predictions and thus a higher standard deviation of the ML ensemble (red circle). To explain this situation, the MSW was retrieved from the IBTrACS data set for the corresponding timesteps. The early and final stages of the Chantal TC are characterized by MSW speeds of 35 and 30 knots (i.e., weak TCs in IBTrACS), respectively. On the contrary, during the middle stage of its evolution (Panel (b)), when the cyclone gains more strength (i.e., the MSW increases to 65 knots), spatial circular patterns of the *vo850* become more evident around the TC center. This leads to a lower localization error between predicted and actual TC center locations. Indeed, the models involved in the ensemble predict approximately the same position, leading to a lower standard deviation (i.e., circle radius in Panel (b) is smaller) from the ensemble TC center estimate.

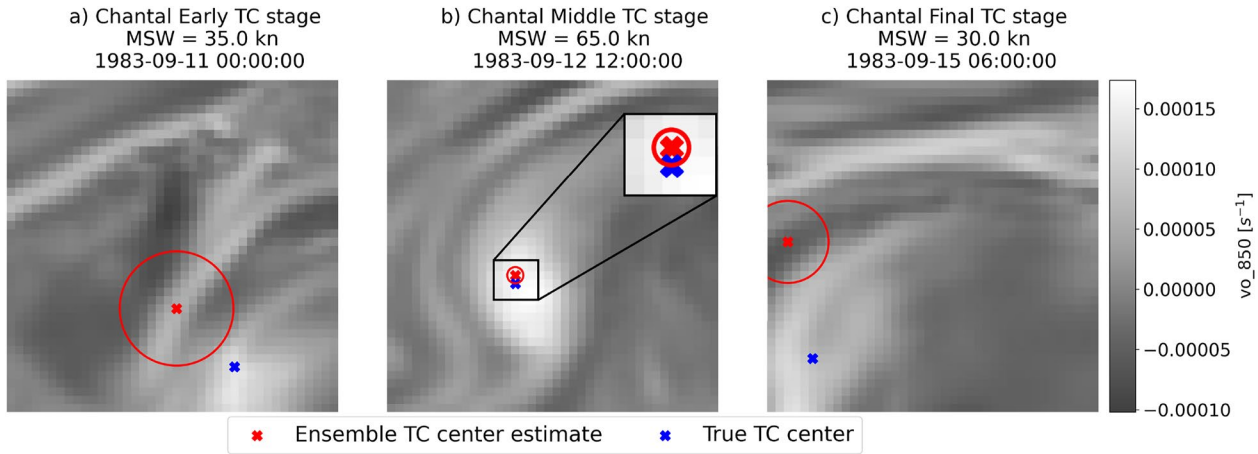

Figure C1. Chantal Tropical Cyclone (TC) *vo850* spatial patterns during early (left panel), middle (middle panel) and final (right panel) stages of its lifecycle. The ensemble TC center estimate (red cross) along with the true one (blue cross) is represented. The standard deviation of the ensemble TC center estimate is represented through the red circle.

Appendix D: Additional Results Related to Keoni and Julio Tropical Cyclones

This section provides more in-depth details about the TC center localization skills through the ML ensemble approach for Keoni and Julio TCs introduced in Section 3.3.

Both figures are organized as follows: the upper panel represents the MSW of the TC (in purple), expressed in knots. In addition, the MSLP was also available for the Julio TC from IBTrACS, and it was reported in the Figure (green line). The middle panel shows the ED (in red) between the observed TC center coordinates from IBTrACS and the estimated ones produced by the ML ensemble, along with the standard deviation among models in agreement, as resulting from the IQR method (light red area). Furthermore, the indication of the TC stages during its lifetime (i.e., *Tropical Storm* [*TS*], *Not Reported* [*NR*] and *Disturbance Storms* [*DS*]) is also reported for the upper and middle panels as vertical dashed lines. In the bottom panel, the observed TC center coordinates are depicted as blue points, whereas the ML ensemble estimates are reported as red circles.

D1. Keoni TC

Figure D1 shows that the early and final stages of the Keoni lifecycle are characterized by low MSW, and therefore the ML ensemble provided TC center estimates with a higher ED from the observed TC center coordinates. On the other hand, as the cyclone gains more strength, the spatial features of the input drivers around the TC center become more evident, thus making TC localization easier and the ED lower, as explained in Section 3.3.

The time steps in which the TC was not detected are hereafter referred to as discontinuities. This means that most of the models in the ML ensemble did not reach the minimum consensus (see Section 2.3.4) about the presence of a TC center in the corresponding time step.

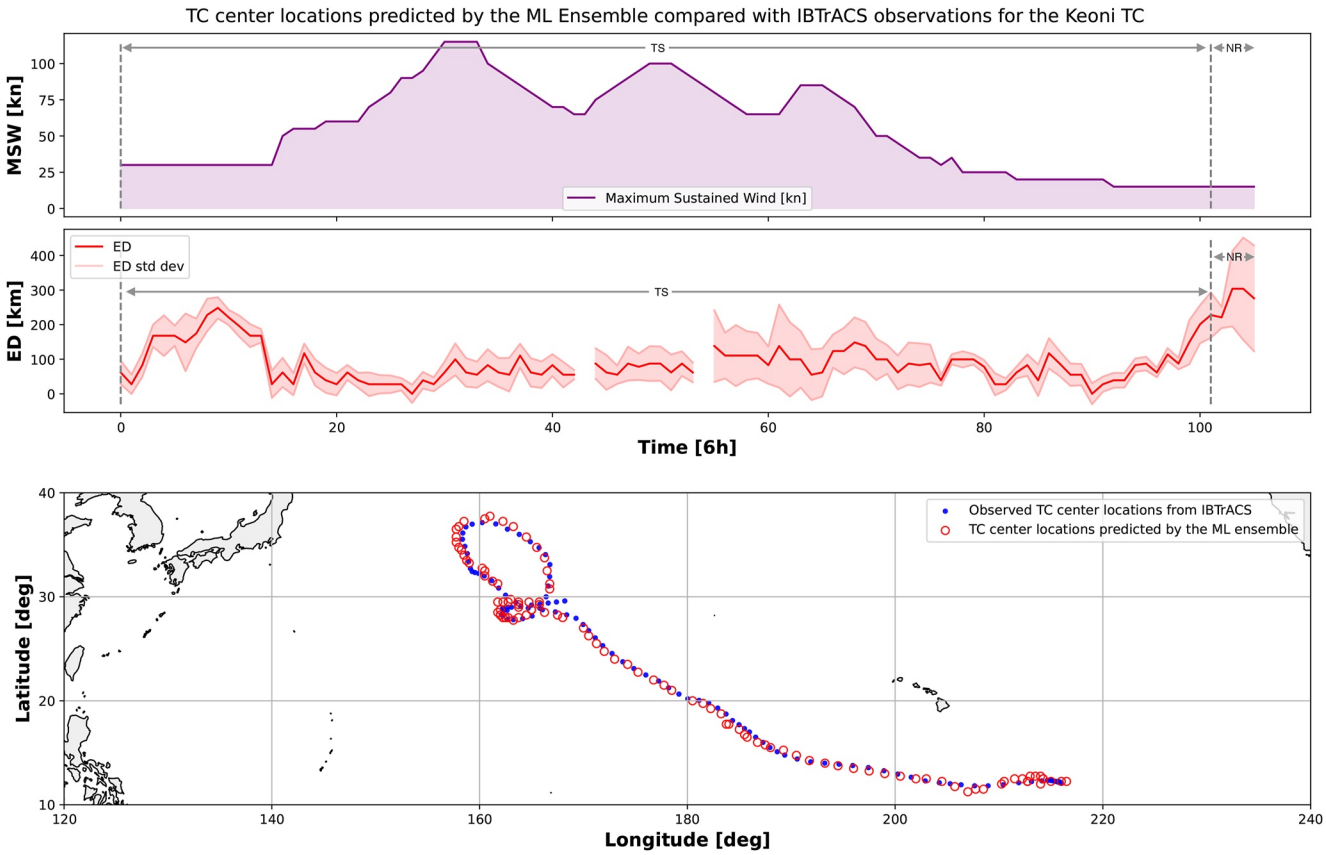

Figure D1. Comparison between the observed Tropical Cyclone (TC) center locations from International Best Track Archive for Climate Stewardship (IBTrACS) with those estimated by the Machine Learning (ML) ensemble. The upper panel represents the Maximum Sustained Wind (MSW) of the TC (in purple), expressed in knots. The middle panel shows the Euclidean Distance (ED) (in red) between the observed TC center coordinates from IBTrACS and the estimated ones produced by the ML ensemble, along with the standard deviation among models in agreement, as resulting from the IQR method (light red area). The discontinuities of the ED in the middle panel correspond to time steps in which the ensemble did not detect the TC center locations. Furthermore, the indication of the TC stages during its lifetime (i.e., *Tropical Storm* [*TS*] and *Not Reported* [*NR*]) is also reported for the upper and middle panels as vertical dashed lines. In the bottom panel, the observed TC center coordinates are depicted as blue points, whereas the ML ensemble estimates are reported as red circles.

D2. Julio TC

Similarly to the Keoni test case, the early and final stages of the Julio TC (see Figure D2) are characterized by low MSW and high levels of MSLP, which correspond to higher localization errors for the ML ensemble. Nevertheless, even though in the aforementioned stages the cyclone is classified as a *Disturbance Storm*, the ML ensemble is still able to capture the phenomena at these stages of the TC lifecycle. This is a remarkable result since both *Disturbance Storm* and *NR* TCs were filtered out from the training set, and therefore their characteristics were not shown during training. Over the TC evolution, the ED remains stable on average and slightly increases as the TC dissipates its energy. A discontinuity in the ED corresponds to time steps in which the ML ensemble did not reach the minimum level of consensus about the presence of a TC center in input patch (see Section 2.3.4). Nevertheless, the tracking algorithm fixes this issues and is able to reconstruct the whole trajectory from the set of points located from the ML ensemble, as discussed in Section 3.3.

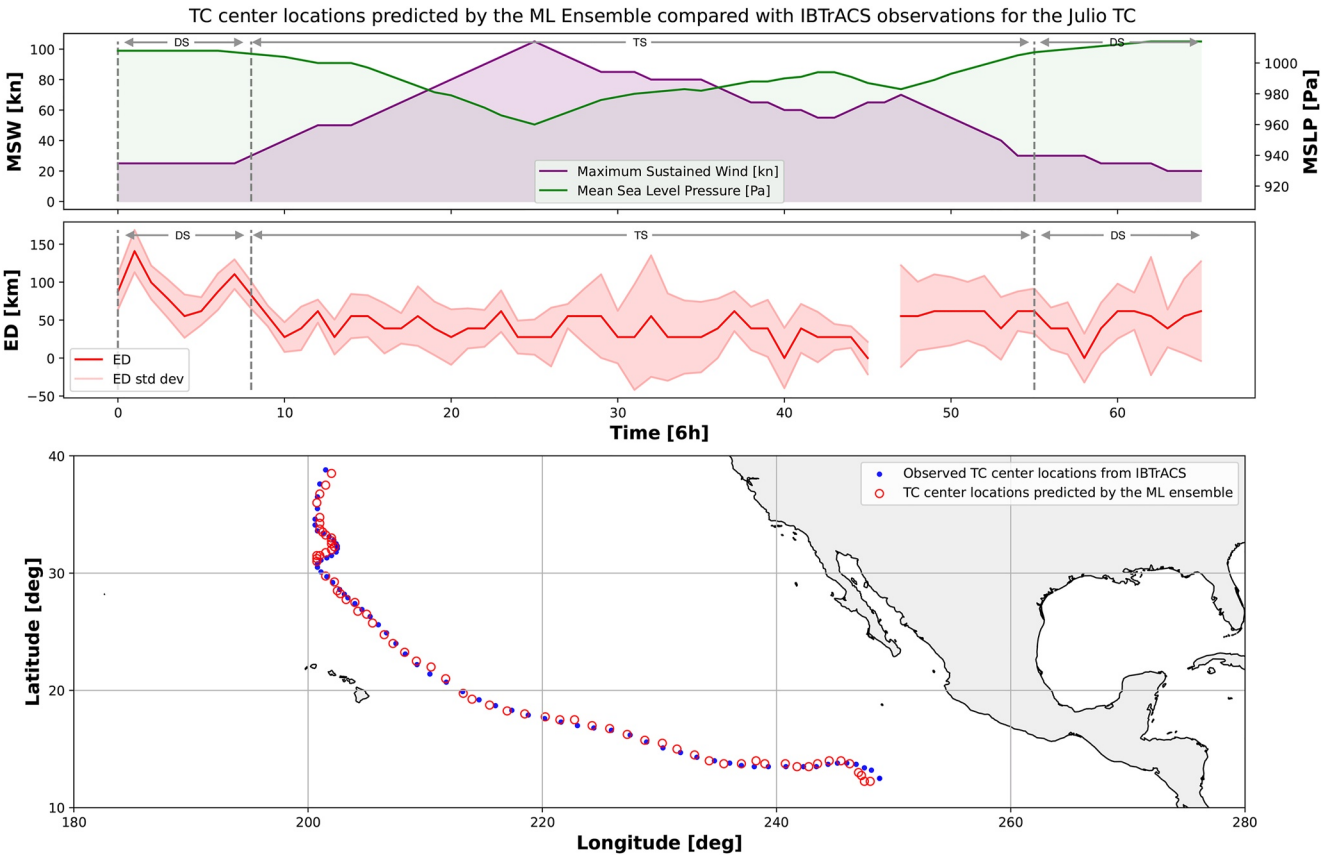

Figure D2. Comparison between the observed Tropical Cyclone (TC) center locations from International Best Track Archive for Climate Stewardship (IBTrACS) with those estimated by the Machine Learning (ML) ensemble. The upper panel represents the Maximum Sustained Wind (MSW) of the TC (in purple), expressed in knots. In addition, the Mean Sea Level Pressure (MSLP) is also reported (green line). The middle panel shows the Euclidean Distance (ED) (in red) between the observed TC center coordinates from IBTrACS and the estimated ones produced by the ML ensemble, along with the standard deviation among models in agreement, as resulting from the IQR method (light red area). The discontinuities of the ED in the middle panel correspond to time steps in which the ensemble did not detect the TC center locations. Furthermore, the indication of the TC stages during its lifetime (i.e., *Tropical Storm* [*TS*] and *Disturbance Storms* [*DS*]) is also reported for the upper and middle panels as vertical dashed lines. In the bottom panel, the observed TC center coordinates are depicted as blue points, whereas the ML ensemble estimates are reported as red circles.

Data Availability Statement

The data sets used in this study are freely accessible from public repositories:

- Copernicus ERA5 reanalysis data sets:
  - Single levels [Dataset] (i.e., mean sea level pressure, 10 m wind gust since previous post-processing and instantaneous 10 m wind gust): https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview (Hersbach et al., 2023b).
  - Pressure levels [Dataset] (i.e., relative vorticity at 850 mb, temperature at 300 mb and temperature at 500 mb): https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels?tab=overview (Hersbach et al., 2023a).
- International Best Track Archive for Climate Stewardship (IBTrACS) from National Centers for Environmental Information (NCEI) [Dataset]: https://www.ncei.noaa.gov/data/international-best-track-archive-for-climate-stewardship-ibtracs/v04r00/access/csv/ (Knapp et al., 2010, 2018).
- Source code for the Machine Learning Tropical Cyclones Detection and Tracking approach presented in this work [Software]: https://dx.doi.org/10.5281/zenodo.8321138 (Donno et al., 2023).

Acknowledgments
This work was supported in part by the eFlows4HPC and InterTwin projects. eFlows4HPC has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 955558. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Spain, Germany, France, Italy, Poland, Switzerland and Norway. In Italy, it has been preliminarily approved for complimentary funding by Ministero dello Sviluppo Economico (MiSE) (ref. project prop. 2659). InterTwin has received funding from Horizon Europe under grant agreement No 101058386. Moreover, the authors would like to thank Dr. Enrico Scoccimarro from the CSP (Climate Simulations and Prediction) Division of the CMCC for his scientific support. Moreover, we thank Antonio Aloisio from the ASC (Advanced Scientific Computing) Division of the CMCC for his editing and proofreading work on this article.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). *TensorFlow: Large-scale machine learning on heterogeneous sytems* [Software]. Retrieved from https://www.tensorflow.org/

Barrera, F. Y. (2022). What are "bottlenecks" in neural networks? Retrieved from https://www.baeldung.com/cs/neural-network-bottleneck

Befort, D. J., Kruschke, T., & Leckebusch, G. C. (2020). Objective identification of potentially damaging tropical cyclones over the western north pacific. *Environmental Research Communications*, *2*(3), 031005. https://doi.org/10.1088/2515-7620/ab7b35

Bloemendaal, N., de Moel, H., Mol, J. M., Bosma, P. R. M., Polen, A. N., & Collins, J. M. (2021). Adequately reflecting the severity of tropical cyclones using the new tropical cyclone severity scale. *Environmental Research Letters*, *16*(1), 014048. https://doi.org/10.1088/1748-9326/abd131

Bourdin, S., Fromang, S., Dulac, W., Cattiaux, J., & Chauvin, F. (2022). Intercomparison of four algorithms for detecting tropical cyclones using ERA5. *Geoscientific Model Development*, *15*(17), 6759–6786. https://doi.org/10.5194/gmd-15-6759-2022

Camargo, S. J., & Zebiak, S. E. (2002). Improving the detection and tracking of tropical cyclones in atmospheric general circulation models. *Weather and Forecasting*, *17*(6), 1152–1162. https://doi.org/10.1175/1520-0434(2002)017〈1152:ITDATO〉2.0.CO;2

Carmo, A. R., Longépé, N., Mouche, A., Amorosi, D., & Cremer, N. (2021). Deep learning approach for tropical cyclones classification based on C-band sentinel-1 SAR images. In *2021 IEEE international geoscience and remote sensing symposium IGARSS* (pp. 3010–3013). https://doi.org/10.1109/IGARSS47720.2021.9554111

Chauvin, F., Royer, J.-F., & Déqué, M. (2006). Response of hurricane-type vortices to global warming as simulated by ARPEGE-climat at high resolution. *Climate Dynamics*, *27*(4), 377–399. https://doi.org/10.1007/s00382-006-0135-7

Chollet, F. (2015). Keras. Retrieved from https://keras.io

Dabhade, A., Roy, S., Moustafa, M. S., Mohamed, S. A., El Gendy, R., & Barma, S. (2021). Extreme weather event (cyclone) detection in India using advanced deep learning techniques. In *2021 9th International conference on orange technology (ICOT)* (pp. 1–4). https://doi.org/10.1109/ICOT54518.2021.9680663

Donno, D., Accarino, G., Immorlano, F., Elia, D., & Aloisio, G. (2023). Machine learning TC detection and tracking [Software]. Zenodo. https://doi.org/10.5281/zenodo.8321138

ECMWF. (2020a). *ECMWF fact sheet* (Tech. Rep.). European Centre for Medium-Range Weather Forecasts.

ECMWF. (2020b). Reanalysis. Retrieved from https://www.ecmwf.int/sites/default/files/medialibrary/2020-06/ecmwf-fact-sheet-reanalysis.pdf

Ejarque, J., Badia, R. M., Albertin, L., Aloisio, G., Baglione, E., Becerra, Y., et al. (2022). Enabling dynamic and intelligent workflows for HPC, data analytics, and AI convergence. *Future Generation Computer Systems*, *134*, 414–429. https://doi.org/10.1016/j.future.2022.04.014

Elia, D., Fiore, S., & Aloisio, G. (2021). Towards HPC and big data analytics convergence: Design and experimental evaluation of a HPDA framework for escience at scale. *IEEE Access*, *9*, 73307–73326. https://doi.org/10.1109/ACCESS.2021.3079139

Elsner, J. B., Kossin, J. P., & Jagger, T. H. (2008). The increasing intensity of the strongest tropical cyclones. *Nature*, *455*(7209), 92–95. https://doi.org/10.1038/nature07234

Emanuel, K. A. (2003). Tropical cyclones. *Annual Review of Earth and Planetary Sciences*, *31*(1), 75–104. https://doi.org/10.1146/annurev.earth.31.100901.141259

Emanuel, K. A., & Nolan, D. S. (2004). Tropical cyclone activity and the global climate system. In *Preprints, 26th conf. on hurricanes and tropical meteorology, Miami, FL* (Vol. 10). Amer. Meteor. Soc. A.

Enz, B. M., Engelmann, J. P., & Lohmann, U. (2022). Parallel use of threshold parameter variation for tropical cyclone tracking. *Geoscientific Model Development Discussions*, *2022*, 1–29. https://doi.org/10.5194/gmd-2022-279

Ganaie, M., Hu, M., Malik, A., Tanveer, M., & Suganthan, P. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, *115*, 105151. https://doi.org/10.1016/j.engappai.2022.105151

Gray, J., Liu, D. T., Nieto-Santisteban, M., Szalay, A., DeWitt, D. J., & Heber, G. (2005). Scientific data management in the coming decade. *ACM SIGMOD Record*, *34*(4), 34–41. https://doi.org/10.1145/1107499.1107503

Gray, W. M. (1975). Tropical cyclone genesis (Unpublished doctoral dissertation). Colorado State University. Libraries.

Haque, M. N., Ashfaqul Adel, A. A. M., & Alam, K. S. (2022). Deep learning techniques in cyclone detection with cyclone eye localization based on satellite images. In M. S. Arefin, M. S. Kaiser, A. Bandyopadhyay, M. A. R. Ahad, & K. Ray (Eds.), *Proceedings of the international conference on big data, IoT, and machine learning* (pp. 461–472). Springer Singapore.

Haynes, K., Lagerquist, R., McGraw, M., Musgrave, K., & Ebert-Uphoff, I. (2023). Creating and evaluating uncertainty estimates with neural networks for environmental-science applications. *Artificial Intelligence for the Earth Systems*, *2*(2), 220061. https://doi.org/10.1175/AIES-D-22-0061.1

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., et al. (2023a). *ERA5 hourly data on pressure levels from 1940 to present*. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). https://doi.org/10.24381/cds.bd0915c6

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., et al. (2023b). *ERA5 hourly data on single levels from 1940 to present*. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). https://doi.org/10.24381/cds.adbb2d47

Hey, T., Butler, K., Jackson, S., & Thiyagalingam, J. (2020). Machine learning and big scientific data. *Philosophical Transactions of the Royal Society A*, *378*(2166), 20190054. https://doi.org/10.1098/rsta.2019.0054

Hodges, K., Cobb, A., & Vidale, P. L. (2017). How well are tropical cyclones represented in reanalysis datasets? *Journal of Climate*, *30*(14), 5243–5264. https://doi.org/10.1175/JCLI-D-16-0557.1

Horn, M., Walsh, K., Zhao, M., Camargo, S. J., Scoccimarro, E., Murakami, H., et al. (2014). Tracking scheme dependence of simulated tropical cyclone response to idealized climate simulations. *Journal of Climate*, *27*(24), 9197–9213. https://doi.org/10.1175/JCLI-D-14-00200.1

Hoyer, S., & Hamman, J. (2017). xarray: N-D labeled Arrays and Datasets in Python. *Journal of Open Research Software*, *5*(1), 10. https://doi.org/10.5334/jors.148

IBTrACS Science Team. (2019). *International best track archive for climate stewardship (IBTRACS) technical documentation* (Tech. Rep.). National Oceanic and Atmospheric Administration.

Kepert, J. D. (2010). Tropical cyclone structure and dynamics. In *Global perspectives on tropical cyclones* (pp. 3–53). https://doi.org/10.1142/9789814293488_0001

Kim, M., Park, M.-S., Im, J., Park, S., & Lee, M.-I. (2019). Machine learning approaches for detecting tropical cyclone formation using satellite data. *Remote Sensing*, *11*(10), 1195. https://doi.org/10.3390/rs11101195

Kim, S., Kim, H., Lee, J., Yoon, S., Kahou, S. E., Kashinath, K., & Prabhat, M. (2019). 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

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Knapp, K. R., Diamond, H. J., Kossin, J. P., Kruk, M. C., & Schreck, C. J. (2018). *International best track archive for climate stewardship (IBTRACS) project* version 4. NOAA National Centers for Environmental Information. https://doi.org/10.25921/82TY-9E16

Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J., & Neumann, C. J. (2010). The international best track archive for climate stewardship (IBTRACS): Unifying tropical cyclone data. *Bulletin of the American Meteorological Society*, *91*(3), 363–376. https://doi.org/10.1175/2009BAMS2755.1

Kumler-Bonfanti, C., Stewart, J., Hall, D., & Govett, M. (2020). Tropical and extratropical cyclone detection using deep learning. *Journal of Applied Meteorology and Climatology*, *59*(12), 1971–1985. https://doi.org/10.1175/JAMC-D-20-0117.1

Lam, L., George, M., Gardoll, S., Safieddine, S., Whitburn, S., & Clerbaux, C. (2023). Tropical cyclone detection from the thermal infrared sensor IASI data using the deep learning model YOLOv3. *Atmosphere*, *14*(2), 215. https://doi.org/10.3390/atmos14020215

Mendelsohn, R., Emanuel, K., Chonabayashi, S., & Bakkensen, L. (2012). The impact of climate change on global tropical cyclone damage. *Nature Climate Change*, *2*(3), 205–209. https://doi.org/10.1038/nclimate1357

MetOffice. (2023). Development of tropical cyclones. Retrieved from https://www.metoffice.gov.uk/weather/learn-about/weather/types-of-weather/hurricanes/development#:~:text=Several\%20conditions\%20are\%20needed\%20for,known\%20as\%20low\%20wind\%20shear

Murakami, H. (2014). Tropical cyclones in reanalysis data sets. *Geophysical Research Letters*, *41*(6), 2133–2141. https://doi.org/10.1002/2014GL059519

Nair, A., Srujan, K. S. S. S., Kulkarni, S. R., Alwadhi, K., Jain, N., Kodamana, H., et al. (2022). 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

National Oceanic and Atmospheric Administration. (2023). International best track archive for climate stewardship (IBTRACS). Retrieved from https://www.ncei.noaa.gov/products/international-best-track-archive

Pang, S., Xie, P., Xu, D., Meng, F., Tao, X., Li, B., et al. (2021). NDFTC: A new detection framework of tropical cyclones from meteorological satellite images with deep transfer learning. *Remote Sensing*, *13*(9), 1860. https://doi.org/10.3390/rs13091860

Roberts, M. J., Camp, J., Seddon, J., Vidale, P. L., Hodges, K., Vanniere, B., et al. (2020). Impact of model resolution on tropical cyclone simulation using the HighResMIP–PRIMAVERA multimodel ensemble. *Journal of Climate*, *33*(7), 2557–2583. https://doi.org/10.1175/JCLI-D-19-0639.1

Roy, C., & Kovordányi, R. (2012). Tropical cyclone track forecasting techniques — A review. *Atmospheric Research*, *104–105*, 40–69. https://doi.org/10.1016/j.atmosres.2011.09.012

Rüttgers, M., Lee, S., Jeon, S., & You, D. (2019). Prediction of a typhoon track using a generative adversarial network and satellite images. *Scientific Reports*, *9*(1), 6057. https://doi.org/10.1038/s41598-019-42339-y

Scoccimarro, E., Fogli, P. G., Reed, K. A., Gualdi, S., Masina, S., & Navarra, A. (2017). Tropical cyclone interaction with the ocean: The role of high-frequency (subdaily) coupled processes. *Journal of Climate*, *30*(1), 145–162. https://doi.org/10.1175/jcli-d-16-0292.1

Scoccimarro, E., Gualdi, S., Villarini, G., Vecchi, G. A., Zhao, M., Walsh, K., & Navarra, A. (2014). Intense precipitation events associated with landfalling tropical cyclones in response to a warmer climate and increased CO_2. *Journal of Climate*, *27*(12), 4642–4654. https://doi.org/10.1175/JCLI-D-14-00065.1

Sebestyén, V., Czvetkó, T., & Abonyi, J. (2021). The applicability of big data in climate change research: The importance of system of systems thinking. *Frontiers in Environmental Science*, *9*, 619092. https://doi.org/10.3389/fenvs.2021.619092

Shakya, S., Kumar, S., & Goswami, M. (2020). Deep learning algorithm for satellite imaging based cyclone detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 827–839. https://doi.org/10.1109/JSTARS.2020.2970253

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Stewart, S. R., & Jacobson, C. (2016). *Hurricane Julio NOOA report*. National Hurricane Center & Central Pacific Hurricane Center. Retrieved from https://www.nhc.noaa.gov/data/tcr/EP102014_Julio.pdf

Strachan, J., Vidale, P. L., Hodges, K., Roberts, M., & Demory, M.-E. (2013). Investigating global tropical cyclone activity with a hierarchy of AGCMS: The role of model resolution. *Journal of Climate*, *26*(1), 133–152. https://doi.org/10.1175/JCLI-D-12-00012.1

Sun, Y., Zhong, Z., Li, T., Yi, L., Hu, Y., Wan, H., et al. (2017). Impact of ocean warming on tropical cyclone size and its destructiveness. *Scientific Reports*, *7*(1), 8154. https://doi.org/10.1038/s41598-017-08533-6

The Pandas Development Team. (2023). pandas-dev/pandas: Pandas (Version v1.5.3) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.7549438

Tory, K. J., Chand, S. S., Dare, R. A., & McBride, J. L. (2013a). An assessment of a model-grid-and basin-independent tropical cyclone detection scheme in selected CMIP3 global climate models. *Journal of Climate*, *26*(15), 5508–5522. https://doi.org/10.1175/JCLI-D-12-00511.1

Tory, K. J., Chand, S. S., Dare, R. A., & McBride, J. L. (2013b). The development and assessment of a model-grid-and basin-independent tropical cyclone detection scheme. *Journal of Climate*, *26*(15), 5493–5507. https://doi.org/10.1175/jcli-d-12-00510.1

Tropical Cyclones 1993 NOOA Report. (1993). NOAA Central Pacific Hurricane Center. Retrieved from https://www.nhc.noaa.gov/data/tcr/CP1993_Seasonal_TCR.pdf

Wang, P., Wang, P., Wang, C., Yuan, Y., & Wang, D. (2020). 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://doi.org/10.1109/JSTARS.2020.2995158

Weaver, M. M., & Garner, A. J. (2023). Varying genesis and landfall locations for North Atlantic tropical cyclones in a warmer climate. *Scientific Reports*, *13*(1), 5482. https://doi.org/10.1038/s41598-023-31545-4

World Meteorological Organization. (2022). Essential climate variables. Retrieved from https://public.wmo.int/en/programmes/global-climate-observing-system/essential-climate-variables

World Meteorological Organization. (2023). Tropical cyclones. Retrieved from https://public.wmo.int/en/our-mandate/focus-areas/natural-hazards-and-disaster-risk-reduction/tropical-cyclones

Xie, M., Li, Y., & Cao, K. (2020). Global cyclone and anticyclone detection model based on remotely sensed wind field and deep learning. *Remote Sensing*, *12*(19), 3111. https://doi.org/10.3390/rs12193111

Xie, M., Li, Y., & Dong, S. (2022). A deep-learning-based fusion approach for global cyclone detection using multiple remote sensing data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *15*, 9613–9622. https://doi.org/10.1109/JSTARS.2022.3219809

Zarzycki, C. M., & Ullrich, P. A. (2017). Assessing sensitivities in algorithmic detection of tropical cyclones in climate data. *Geophysical Research Letters*, *44*(2), 1141–1149. https://doi.org/10.1002/2016gl071606

Zarzycki, C. M., Ullrich, P. A., & Reed, K. A. (2021). Metrics for evaluating tropical cyclones in climate data. *Journal of Applied Meteorology and Climatology*, *60*(5), 643–660. https://doi.org/10.1175/JAMC-D-20-0149.1

Zhao, M., Held, I. M., Lin, S.-J., & Vecchi, G. A. (2009). Simulations of global hurricane climatology, interannual variability, and response to global warming using a 50-km resolution GCM. *Journal of Climate*, *22*(24), 6653–6678. https://doi.org/10.1175/2009jcli3049.1