Investigating differences between Tropical cyclone detection systems

Article

Published Version

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

Investigating Differences between Tropical Cyclone Detection Systems

DANIEL GALEA,[a] KEVIN HODGES,[b] AND BRYAN N. LAWRENCE[a,b]

[a] *Department of Computer Science, University of Reading, Reading, United Kingdom*
[b] *National Centre of Atmospheric Science, Department of Meteorology, University of Reading, Reading, United Kingdom*

ABSTRACT: Tropical cyclones (TCs) are important phenomena, and understanding their behavior requires being able to detect their presence in simulations. Detection algorithms vary; here we compare a novel deep learning–based detection algorithm (TCDetect) with a state-of-the-art tracking system (TRACK) and an observational dataset (IBTrACS) to provide context for potential use in climate simulations. Previous work has shown that TCDetect has good recall, particularly for hurricane-strength events. The primary question addressed here is to what extent the structure of the systems plays a part in detection. To compare with observations of TCs, it is necessary to apply detection techniques to reanalysis. For this purpose, we use ERA-Interim, and a key part of the comparison is the recognition that ERA-Interim itself does not fully reflect the observations. Despite that limitation, both TCDetect and TRACK applied to ERA-Interim mostly agree with each other. Also, when considering only hurricane-strength TCs, TCDetect and TRACK correspond well to the TC observations from IBTrACS. Like TRACK, TCDetect has good recall for strong systems; however, it finds a significant number of false positives associated with weaker TCs (i.e., events detected as having hurricane strength but are weaker in reality) and extratropical storms. Because TCDetect was not trained to locate TCs, a post hoc method to perform comparisons was used. Although this method was not always successful, some success in matching tracks and events in physical space was also achieved. The analysis of matches suggested that the best results were found in the Northern Hemisphere and that in most regions the detections followed the same patterns in time no matter which detection method was used.

KEYWORDS: Tropical cyclones; Classification; Deep learning; Machine learning; Neural networks; Pattern recognition

1. Introduction

Tropical cyclones (TCs) are extreme weather events that can have a large effect on any environment. These can be and are detected and tracked in satellite data, numerical weather prediction (NWP) simulations, and longer simulations with global circulation models (GCMs) via automatic means.

Previous studies (section 2) have shown that the performance of various detection algorithms is comparable when addressing strong TCs, that is, those that have obtained hurricane status according to the Saffir–Simpson scale. Some show that the detection algorithms did not perform well when used on datasets other than that on which the algorithm was first tuned.

Galea et al. (2023) introduced a deep learning technique, TCDetect, for detecting the presence or absence of a hurricane-strength TC in a field of simulation data. In this study, we compare the performance of this model with a state-of-the-art algorithm (which does not use machine learning) and an observational dataset. We begin with a description of previous literature comparing various detection algorithms (section 2) and then describe the data and detection algorithms used in this study (section 3). The main body of the paper (section 4) describes the results obtained when comparing TCDetect with a version of a state-of-the-art non–machine learning algorithm (TRACK; see below) applied to reanalysis data and when comparing TCDetect with reality as recorded by the International Best Track Archive for Climate Stewardship (IBTrACS; Knapp et al. 2010, 2018) archive. The final summary (section 5) presents our understanding of the limitations of the application of TCDetect for detecting tropical cyclone features in simulation data.

2. Previous studies

There is extensive previous literature comparing different automatic TC detection algorithms; we here provide some context for our work by highlighting some previous intercomparison between other techniques and some previous investigations into how the type of input data used impacts results.

Horn et al. (2014) compare five different detection algorithms. The first two were a modified version of the Commonwealth Scientific and Industrial Research Organization (CSIRO) tracking scheme (Horn et al. 2013) and the Zhao tracking scheme (Zhao et al. 2009). The last methods were those developed by the modeling groups whose data were involved, that is, the groups from the Meteorological Research Institute (MRI), the National Aeronautics and Space Administration (NASA) Goddard Institute for Space Studies (GISS), and the Centro Euro-Mediterraneo per i Cambiamenti Climatici-Istituto Nazionale di Geofisica e Vulcanologia (CMCC-INGV). The simulations examined used the CMCC-INGV ECHAM5 model, which has ~90-km grid spacing at the equator (Roeckner et al. 2003); the NASA-GISS model, which has ~110-km grid spacing at the equator (Schmidt et al. 2014); the National Centers

Corresponding author: Daniel Galea, galea.daniel18@gmail.com; d.galea@pgr.reading.ac.uk

for Environmental Prediction (NCEP) Global Forecast System (GFS), which has ~110-km grid spacing at the equator (Saha et al. 2014); and version 3.2 of the Meteorological Research Institute Atmospheric General Circulation Model (MRI AGCM3.2), which has ~60-km grid spacing at the equator (Mizuta et al. 2012).

The authors showed that the method tuned to the underlying data achieved the best performance when comparing hurricane-strength TC counts with observations and usually outperformed the other methods applied to the same data, without being tuned. They also show that detection methods that were not optimized on the data being tested do not work as well as if they had been optimized. Similarly, Onogi et al. (2007) also found that a detection algorithm developed for the Japanese Meteorological Agency (JMA) obtained 80% of hurricane-strength TCs in their JRA-25 reanalysis but less than 60% of hurricane-strength TCs in the ERA-40 reanalysis (Uppala et al. 2005).

Given that the requirement for these automatic tracking algorithms is to detect hurricane-strength TCs in a particular set of data that corresponds to those that occurred in real life, it is only natural that the threshold values are tuned to obtain the same number of systems. This could lead to resolution-dependent thresholds as in Walsh et al. (2007).

Zarzycki and Ullrich (2017) conducted sensitivity analysis on the thresholds used for one tracking algorithm, Tempest-Extremes (Ullrich and Zarzycki 2017), applied to four different reanalysis datasets. They found that the most sensitive thresholds were those defining the TC vortex strength, two examples of which are the depth of the minimum of sea level pressure (MSLP) and the maximum sustained surface wind speed. They reported a larger difference when comparing storm count rather than integrated or weighted metrics such as the number of days with a hurricane-strength TC present or accumulated cyclone energy (ACE). Zhao et al. (2009) also found that there is a large sensitivity to the choice made for the threshold for minimum duration of a TC. Strachan et al. (2013) also note that any wind speed threshold should vary linearly with the resolution of the data used, and any deviations from this relationship are likely due to model biases and errors.

It was also noted in previous literature that the intensity of TCs, whether defined by MSLP or maximum sustained surface wind speed, is underestimated in all of the reanalysis datasets. Strachan et al. (2013) noted that resolution alone does not explain this observation, due to feedback processes present in the model. Schenkel and Hart (2012) also noted that the choice of data assimilation method is important to get realistic surface wind speeds. For this reason, the JRA-25 and JRA-55 reanalyses are most realistic, likely due to a vortex re-location step performed during their creation.

Despite all these considerations when it comes to the resolution of different reanalyses, Strachan et al. (2013) show that while those datasets with a resolution finer than 100 km are capable of showing the correct interannual variability, even datasets with a resolution of 20 km are not capable of producing the right intensities.

Hodges et al. (2017) investigated how TRACK performed using six different reanalysis datasets. It was found to work well (probability of detection: 97% in the Northern Hemisphere; 92% in the Southern Hemisphere) at detecting TC tracks across all basins, but it had a high false detection rate, especially in the Southern Hemisphere, when considering only TCs that fulfilled criteria that considered intensity and presence of a warm core. Most of these false positives had their genesis poleward of the latitude of 20°S, leading to the conclusion that these may have been hybrid TCs. An additional conclusion was that the observations may have missed recording some storms, as there were around 20% more advisories issued than storms present in the data. Hodges et al. (2017) opined that such storms may have been omitted due to the lack of human impact and/or accurate measurements.

Although this previous work shows that there are differences in the results of the multiple existing TC detection and tracking algorithms, previous literature also agrees that there is little disagreement in the detection and tracking of strong TCs, that is, those that are at least of category 3 on the Saffir–Simpson scale.

3. Data and methods

The goal of this paper is to understand the characteristics and applicability of our deep learning cyclone detection method, TCDetect, when applied to simulations of the real world. Doing this requires going beyond the normal deep learning metrics, as there are additional complications for real-world applications, arising from the fact that both the observations (the ground truth) and the input simulation data (to the deep learning) introduce detection biases.

In the real world, IBTrACS provides the best source of ground truth. Initially developed by the National Oceanic and Atmospheric Administration (NOAA), it combines all of the best-track data for hurricane-strength TCs from all the official Tropical Cyclone Warning Centers, the WMO Regional Specialized Meteorological Centers (RSMCs), and other sources.

Reanalysis data provide the best possible simulation data by using synthesized observations of meteorological variables; we choose to use the ERA-Interim product (Dee et al. 2011), as the newer ERA5 dataset had not been published when this work was started. ERA-Interim utilizes version CY31r2 of the European Centre for Medium-Range Weather Forecasts (ECMWF) numerical weather prediction system, the Integrated Forecasting System (IFS), together with assimilation of observations from 1979 through 2019.

Unfortunately for our purposes, reanalysis data are not the same as the actual observations, and so, to try to understand the influence of the difference between the observations (IBTrACS) and the input data (ERA-Interim), we use another detection method—TRACK—as a gold standard comparator. TRACK (Hodges 1994, 1995, 1999) is a state-of-the-art automatic detection and tracking system for different types of atmospheric disturbances with considerable use since inception. We then use TCDetect and TRACK applied to reanalysis data and compare them with each other and with the IBTrACS dataset.

ERA-Interim data are produced at a spatial resolution of 79 km at a temporal resolution of 6 h and has 60 vertical levels up to 0.1 hPa. Of the many parameters produced, only the MSLP, 10-m wind speed, and relative vorticity at 850, 700, and 600 hPa are used in this study, and thus, are the input variables used by both TRACK and TCDetect.

The comparison presented here is limited to the 25 months from 1 August 2017 to the end of August 2019 because the earlier data have been used in training the deep learning algorithm.

a. IBTrACS

The IBTrACS dataset has information about reported storms, such as the storm center in latitude and longitude, maximum surface wind speed, minimum sea level pressure, and category. It records tracks of hurricane-strength TCs. In what follows, we consider a TC event to be a snapshot in time of a particular TC recorded by IBTrACS.

While IBTrACS is the best available observational dataset, the dataset is split into observations from seven global basins, or regions. As a consequence, there is some inhomogeneity in coverage and methodology as the different contributing centers have differing observing systems. Such observing systems can be limited in time and space, leading to an incomplete record of their evolution (i.e., events not detected). Such omissions are most likely if they had limited or no human impact, or they were out of range of detection systems such as airborne missions. Also, different centers have different approaches to making their observations and reporting them. An example of this is that the exact minimum surface pressure is often not directly observed but interpolated from nearby observations. Different centers have different methods for making this calculation.

b. TRACK

While TRACK is well described in the literature [see Hodges (1994), (1995), (1999) and references therein], some facets of the way it works have an impact on the comparison to come, so we provide a brief summary of some key aspects of the technique.

TRACK has four different stages: data preparation, segmentation, feature point detection, and tracking.

In the first step, TRACK treats relative vorticity data so that features of interest are easier to detect. This is done with the help of spectral filtering to keep only spatial scales related to the features of interest. With regard to tropical cyclones, these are wavenumbers 5–63, corresponding to spatial extents of around 1300 and 100 km, respectively. This filtering is applied to the vertical average of vorticity between the heights of 850 and 600 hPa.

During the segmentation stage, each point is classified as a background or an object point, depending on whether the value for the vertical average of vorticity is above or below the threshold of $5 \times 10^{-6} \text{ s}^{-1}$. The object points are then collected into objects. Feature point detection then allocates a feature point to each object, representing its center. This feature point is selected as the local extremum in the vertically

averaged vorticity field. Last, the tracking stage uses the feature points generated to minimize a constrained cost function to get the smoothest possible tracks.

The complete TRACK algorithm finds a range of TCs, only some of which may be of hurricane strength. The tracks produced can be processed to identify those particular tracks. The original tracks produced by TRACK are then processed to remove any non-hurricane-strength TC cases from only the start of each track. This is done by using criteria similar to those given by Bengtsson et al. (2007):

- The lifetime must be at least 2 days.
- The initial point in the track must be in between latitudes 30°S and 30°N.
- The maximum in T63 spectrally filtered data (to keep features greater than 180 km) of vertically averaged relative vorticity intensity at 850 hPa must be over $5 \times 10^{-6} \text{ s}^{-1}$.
- A warm-core check must be passed: there must be a T63 vorticity maximum at each atmosphere level at 850, 600, 500, 400, 300, and 250 hPa and within 5° (geodesic) of the maximum at the lower level. Also, the difference between the maxima at 850 and 250 hPa must be greater than $5 \times 10^{-6} \text{ s}^{-1}$.
- The previous two conditions must hold for the last n time steps, where n is a user-defined value.

The tracks are then reformatted so that they only last from the first point that satisfies these criteria to the last point of the original track. We refer to the resulting set of tracks as truncated-TRACK (T-TRACK) dataset.

c. TCDetect deep learning model

The TCDetect deep learning scheme was introduced and described in Galea et al. (2023), where full details of the scheme, including structure and training, can be found. Here we only summarize the inputs, outputs, and results. It is trained on ERA-Interim data, utilizing MSLP, 10-m wind speed, and vorticity at 850, 700, and 600 hPa; all coarsened to an input resolution of approximately 320 km. This data-coarsening step was arrived to as a result of hyperparameter tuning and helped to filter out small-scale noise. Each time step of ERA-Interim data was then split into eight equally sized regions (Fig. 1), which are loosely based on those used in the IBTrACS dataset. While IBTrACS has seven regions, we choose to split the southern Pacific into two regions to maintain equality in region size.

These inputs were then used to infer a classifier value (effectively a probability) ranging between 0 and 1; a hurricane-strength TC is inferred to be present if the value is greater than 0.5, and absent if less than or equal to 0.5.

In testing, the TCDetect algorithm obtained a recall rate, or a probability of detection, of 92% with a precision rate, or a success ratio, of 36%. In practice, this means that while most of the actual hurricane-strength TCs were detected, many of the cases identified were technically false negatives (i.e., non-hurricane-strength TCs). However, as discussed in Galea et al. (2023) and further discussed below, most of these were actually meteorologically significant.
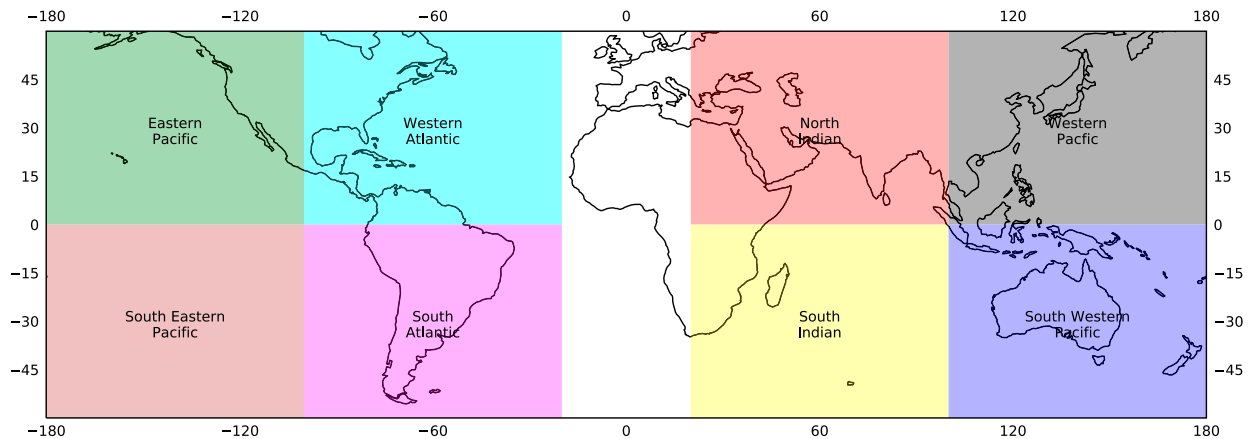
FIG. 1. The eight regions used by TCDetect, which are loosely based on those used in the IBTrACS dataset.

The recall rate and precision were calculated in terms of the application of the technique to ERA-Interim data, but the labels came from IBTrACS. It is reasonable then to ask, "to what extent does the ability of ERA-Interim to reproduce the original storm strength and timing impact these results?" We address this question by applying both T-TRACK and TCDetect to ERA-Interim, and comparing the results with the IB-TrACS, which is considered as the ground truth.

The TC center is not given by TCDetect, so a way to extract it was needed. For this, the gradient-weighted class activation map technique (Grad-CAM; Selvaraju et al. 2017) was used. This works by passing the output of the deep learning model back through the model; then, via gradient maximization, a heat map of the areas in the original input used in a selected layer en route to the output is produced. We selected the first convolutional layer to obtain a heat map and assumed that the TC central position in latitude and longitude is collocated with the maximum activation. This layer was chosen to obtain the heat map with the finest possible resolution so as to get the most precise location of the TC.

Because the heatmaps used for Grad-CAM were generated from the coarsened (320-km resolution) data, the resulting TC centers were coarsely quantized and only poor-quality comparisons were possible. To mitigate this effect, the Grad-CAM centers ("preliminary centers") were then passed through an additional refinement step to generate more accurate locations. A box with sides of 10° in latitude and longitude was centered on the preliminary centers, and the original full resolution ERA-Interim vorticity values at 850, 700, and 600 hPa were obtained and vertically averaged. These three heights in the atmosphere were selected such that the process mirrored that of T-TRACK as closely as possible. The TC center was assumed to be located at the position of the maximum in the absolute value of the averaged vorticity.

These TC centers were then used to make up TC tracks. Given that the TCDetect methodology simply records the presence or absence of a hurricane-strength TC in each of the eight regions defined in Fig. 1, a track was first defined as having TC centers that were present in consecutive time steps in

the same region. However, this produced many short (<2 days) tracks. To attempt to fix this, these short tracks are stitched together to make longer tracks. For a single region, if one track ended at most 2 days (eight time steps) before the next track started and the last TC center of the first track was at most 20° (geodesic) away from the initial TC center of the second track, the two tracks were joined to make up one track. This process was carried out until no more tracks could be joined. The separation distance criterion might intuitively seem to be too wide, but as will be shown below, TCDetect has trouble with locating TC centers, so some buffer was built into this criterion.

4. Results

The first question to consider is "To what extent do the two detection algorithms recover the hurricane-strength TC events, where events are snapshots in time of the system being considered, seen in the observations?" We can then ask, "How well do the two algorithms (combined with ERA-Interim data) position the TCs in space?" Last, we ask, "To what extent does the detection success depend on the TC structure and strength?"

a. Detection

Figure 2 shows the relationship between detection and observations for all the events during the period of interest. For these purposes, for each region, an event was counted when at least one hurricane-strength TC was observed and/or detected in any time step, and the event counts from each region are summed into the global and hemispheric counts as appropriate. While we acknowledge that having the matching criterion set as two tracks matching at just one point could be viewed as too lenient, this was done to reduce the chances of an incorrect nonmatch due to TCDetect missing an event in the track. More stringent matching is performed later in the section.

It is also important to note that for IBTrACS and T-TRACK, this relationship between detection and observations is an underestimate of the total number of events, because only one
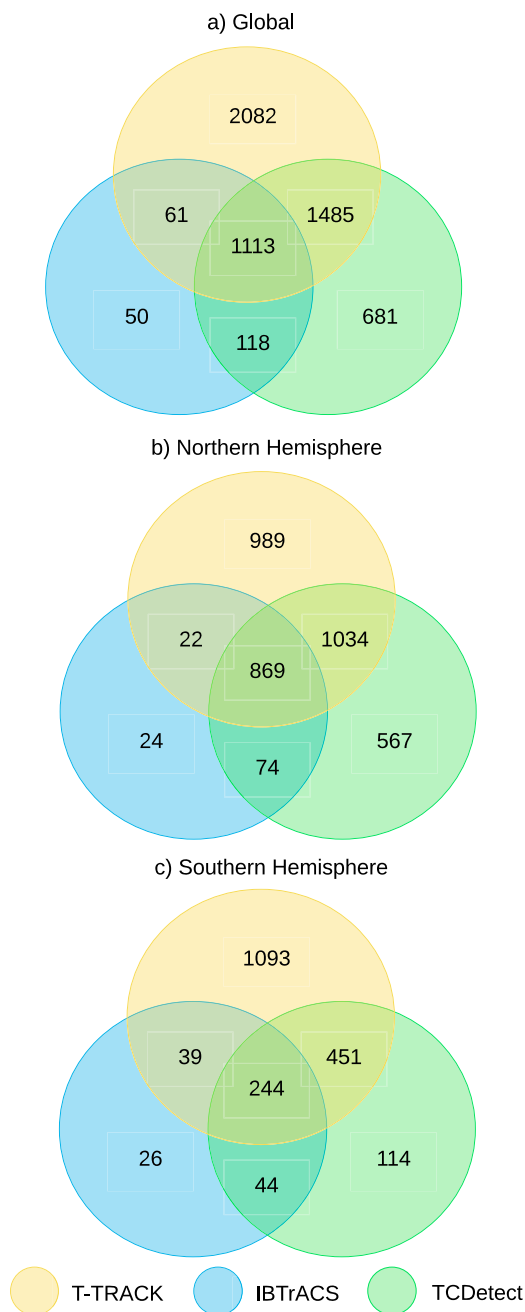
FIG. 2. Events reported by observations (IBTrACS), events detected by T-TRACK, and events detected by TCDetect applied to ERA-Interim data for (a) the whole globe, (b) the Northern Hemisphere, and (c) the Southern Hemisphere.

TABLE 1. Percentage of IBTrACS TC events detected (recall) by T-TRACK and deep learning in ERA-Interim data for all regions (global), the Northern Hemisphere (NH), and the Southern Hemisphere (SH).

	Global	NH	SH
T-TRACK	87%	90%	80%
TCDetect	92%	95%	82%

found by both detection algorithms, with TCDetect finding slightly more than T-TRACK. Relatively few (50) IBTrACS events were not found by either detection method, consistent with the expected high recall rates. However, more events were detected by one or both of T-TRACK and TCDetect than were present in the observations, which suggests many meteorological events were being incorrectly classified as TCs. This finding is discussed further below.

With an a priori expectation that IBTrACS may be undersampling hurricane-strength TC events in the Southern Hemisphere, the data were also split into hemispheres to investigate Figs. 2b and 2c. In terms of recall—that is, the ability for IBTrACS TCs to be detected in ERA-Interim, it can be seen (Table 1) that TCDetect is doing slightly better than T-TRACK in both hemispheres, and slightly more so in the north.

Note that the criteria used to supposedly screen TRACK to identify hurricane-strength TCs are responsible for some of the "missing" detections. If TRACK alone is used, then the recall rate is much higher, reaching 96% globally, with 97% and 92% in the Northern and Southern Hemispheres, respectively, albeit with many more false positives.

Of the 3397 cases in which TCDetect detects a hurricane-strength TC, 681 cases, or around 20%, are not observed or detected by T-TRACK; similarly, of the 4741 cases in which T-TRACK detected the presence of a hurricane-strength TC, 2082 cases, or around 44%, are not observed or detected by TCDetect. These "extra" events found by the detection algorithms require more investigation. Formally, they represent poor precision in the detection (a high proportion of false positives), but the significant overlap using two different techniques is interesting, and suggests the techniques are identifying things that are nearly hurricane-strength TCs—nearly because they are just outside the tropics, or nearly TC-like in structure and strength, consistent with the results reported previously. It could also be that the underlying ERA-Interim data has deficiencies in its representation of these systems, which is causing the methods to produce false positives.

Thus far the analysis has considered time step "events" since the algorithms (TRACK and TCDetect) are applied to one time step after another—but, in reality, these steps form part of the life cycle of a meteorological phenomenon, and it is that thinking that informs the criteria that distinguish T-TRACK from TRACK. These phenomena move along tracks, and so we can consider track detection independently of event detection.

In terms of tracks, Fig. 3a shows how many hurricane-strength TC tracks match, whereby two tracks are matched across datasets if they share one or more detection events in

event was counted even if multiple hurricane-strength TCs were present in the region. This undercounting was used for consistency with TCDetect, which can only report the presence or absence of at least one hurricane-strength TC.

In total there were 1342 such events in the IBTrACS data, and 4741 and 3397 detected by T-TRACK and TCDetect, respectively (Fig. 2a). The majority of the observed events were
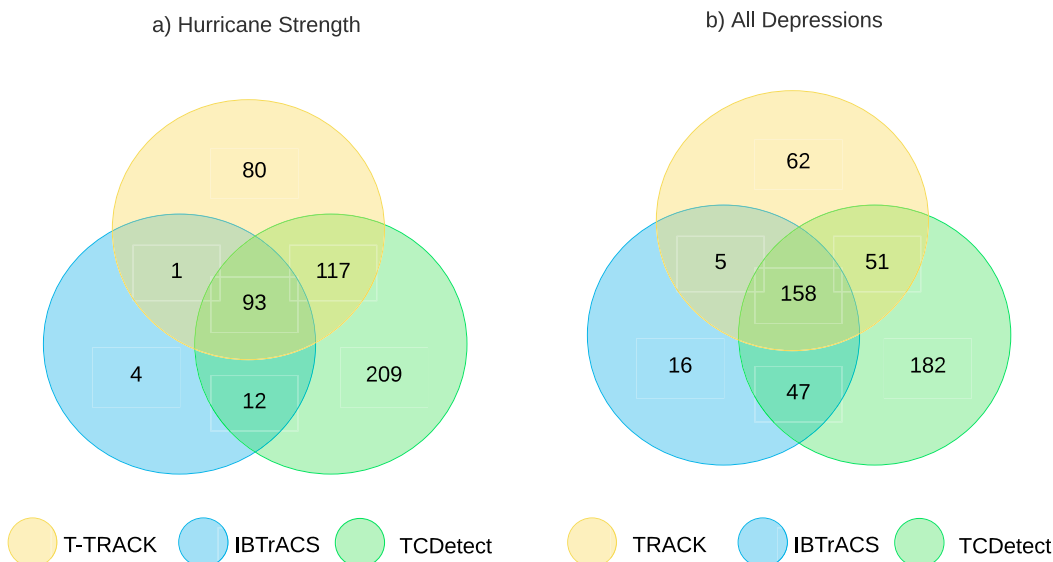
FIG. 3. Tracks reported by observations (IBTrACS), tracks detected by T-TRACK, and tracks detected by TCDetect applied to ERA-Interim data. Overlaps occur when they share a detection event at some point along the track in the same region at the same time step. Tracks are matched for (a) only hurricane-strength TCs and (b) all depressions [i.e., a superset of (a)].

the same region at the same time step. (Note that this means that a single track from one dataset can be matched to multiple tracks from another dataset if multiple TCs are detected in the second dataset.) Similarly, Fig. 3b shows matching tracks where non-hurricane-strength events were also considered.

The majority (96%) of IBTrACS tracks, whether depressions or full hurricane-strength TCs, are matched by at least one of the two detection algorithms. Similar to the events, TCDetect matched to more IBTrACS tracks than T-TRACK, but a majority of the hurricane-strength tracks in IBTrACS were matched with both detection algorithms. Also, there were many tracks that matched between T-TRACK and TCDetect but not with IBTrACS. These could be evidence of TC-like structures being picked up by the detection algorithms that either had not strengthened to hurricane strength or were nontropical systems, an argument supported by the increased number of three-way matches seen when including all non-hurricane-strength tracks (Fig. 3b) and our earlier analysis for TCDetect and IBTrACS alone. The most unmatched tracks come from TCDetect and were due to many nonmeteorological false positives, that is, regions where no tropical system of any kind was defined by IBTrACS but TCDetect erroneously classified the region as having a TC present. However, it is encouraging that most of the tracks either produced by TRACK or found in IBTrACS are being matched by tracks produced by TCDetect.

Thus far, we have been considering track matches that occurred with very loose criteria. We now make some additional track-matching constraints to further understand the differences between the two methods and the observations. The new criteria are very similar to those used by Hodges et al. (2017):

- the mean separation distance between all matching TC centers (i.e., those present in the same time step) between tracks is less than 5° (geodesic),
- the tracks need to overlap for at least 10% of the base track's lifetime (the base track originates from T-TRACK if present; otherwise, it is taken to be from IBTrACS), and
- the track with the least mean separation distance is chosen if multiple matching tracks exist.

These constraints remove any of the unmatched TC tracks. It should also be noted that some events can still be detected by only one method if the tracks considered are not of the same length and/or have a different starting time step.

After these criteria were applied, the events from the remaining tracks were again split by method. The matches between detection and observations now includes fewer events (cf. Figs. 4 and 2a). The number of cases with the presence of a hurricane-strength TC found in IBTrACS decreases from 1342 to 1327. The same occurs for T-TRACK (from 4741 to 3357) and TCDetect (from 3397 to 1067).

The small change in total events for IBTrACS is expected, given the vast majority of observed tracks are detected by one or other method at some point during their evolution. The biggest change is seen in the TCDetect results, where many events are rejected, either because they did not lie on a track, or because the tracks were incorrectly positioned and lay outside the 5° (geodesic) criterion.

We now turn our attention to when events occur. The events from Fig. 2 are recast as events seen per month and shown in Fig. 5. While T-TRACK and TCDetect detect more hurricane-strength TC tracks, they share the same intra-annual variability seen in the IBTrACS observations. Regions in the Northern
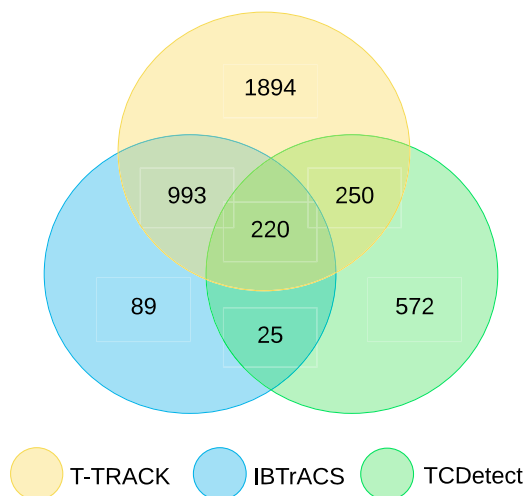
FIG. 4. Events detected by TRACK, events detected by TCDetect, and/or events reported by IBTrACS that fall on matching tracks, defined by applying constraints similar to those of Hodges et al. (2017).

Hemisphere show an uptick in the number of such events in the months between July and October, with a similar increase in the months between December and June for regions in the Southern Hemisphere.

Table 2 shows the Pearson correlations between the time series given by IBTrACS and T-TRACK and IBTrACS and TCDetect. Confidence intervals for these correlations were also calculated using Fisher's transform. When considering all regions, there is a correlation of 0.69 between IBTrACS and T-TRACK, showing that these two time series have a reasonably strong correlation. When examining this correlation by region, correlations higher than 0.7 are obtained in four regions, while another two have correlations between 0.5 and 0.7. The north Indian Ocean has a correlation less than 0.5, whereas no value is obtained for the South Atlantic Ocean because there were no TCs recorded in IBTrACS.

The correlations are not as good when considering IBTrACS and TCDetect. A correlation of −0.08 is obtained when considering TCs over the whole globe. Six regions have a weak correlation, with values between 0.4 and 0.55, while the north Indian Ocean has a correlation of 0.27.

These results show that the time series of IBTrACS is better correlated to that of T-TRACK than TCDetect, both when considering the whole globe and on the level of an individual region. The problems in the Indian Ocean and the South Atlantic are further discussed below.

b. Location

Using the constrained matched track data, we now address in more detail the question as to how well TCDetect locates hurricane-strength TC centers. Figure 6 shows the location of the events in these data.

The IBTrACS dataset is again considered to be the ground truth. It shows that most TCs are found in a few well-defined regions:

- close to the eastern shores of the North American continent and farther out to the middle of the Atlantic Ocean,
- to the west of the North American continent and in the middle of the Pacific Ocean,
- to the east of Asia, over the western Pacific Ocean, and
- over the middle of the Indian Ocean and to the north of Australia.

By comparison, T-TRACK shows a larger number of events and longer tracks, some extending well into the subtropics. There are also more TC centers present in the Southern Hemisphere than IBTrACS, especially the central southern Pacific Ocean. As Hodges et al. (2017) discuss, weak TCs in the Southern Hemisphere are often not reported to the best-track database, because of uncertainties associated with the lack of direct observations in this region. Hence, the use of an objective method to obtain the locations off the eastern coast of the South American continent, which are nonexistent in IBTrACS, could point to the use of reanalysis data and tracking algorithms providing better ground truth in observation-poor regions of the globe and/or certain operational procedures that mean that the center responsible excludes these storms from their reports.

The locations reported by TCDetect are positioned mostly in the right regions, but some centers are located well inland or well into the subtropics, where TCs are not expected. Also, the centers over the Indian Ocean are more spread out than those found in IBTrACS or T-TRACK. It is clear that the geolocation part of the algorithm is not working as well as the detection algorithm—consistent with the way the deep learning model was developed (it was trained for detection, not location).

Location accuracy can also be seen in the spatial correlation of all the TC events that have centers within 10° (Fig. 7). The data for both two-way and three-way matches show a tight grouping and a good correlation, but more scatter is seen in the two-way matches involving TCDetect. Analysis of the temporal matches where the centers were farther than 10° apart suggest that in addition to the TCDetect location issues, further complications could arise from the way tracks for TCDetect were created, where some tracks from two separate events were erroneously joined together.

Figure 8 shows the distribution of all hurricane-strength TC cases by latitude. While the peak of the distributions for both detection algorithms in both hemispheres is biased equatorward (with respect to IBTrACS observations), the two detection algorithms broadly agree in the Northern Hemisphere. However, the distribution for the deep learning–based algorithm shows two peaks in the Southern Hemisphere: one at around 10°S and a peak at around 40°S. The first peak matches up well with that from T-TRACK. The second is consistent with the southern bias in positions seen in the Indian Ocean and the excess of detections in and around the Tasman Sea.

c. Structure

It is feasible that the physical structure of cyclones in terms of their representation in ERA-Interim might affect the results presented here. To investigate this, we created composites of the events, presented in Fig. 9 for Northern Hemisphere TCs and Fig. 10 for Southern Hemisphere TCs, using the ERA-Interim
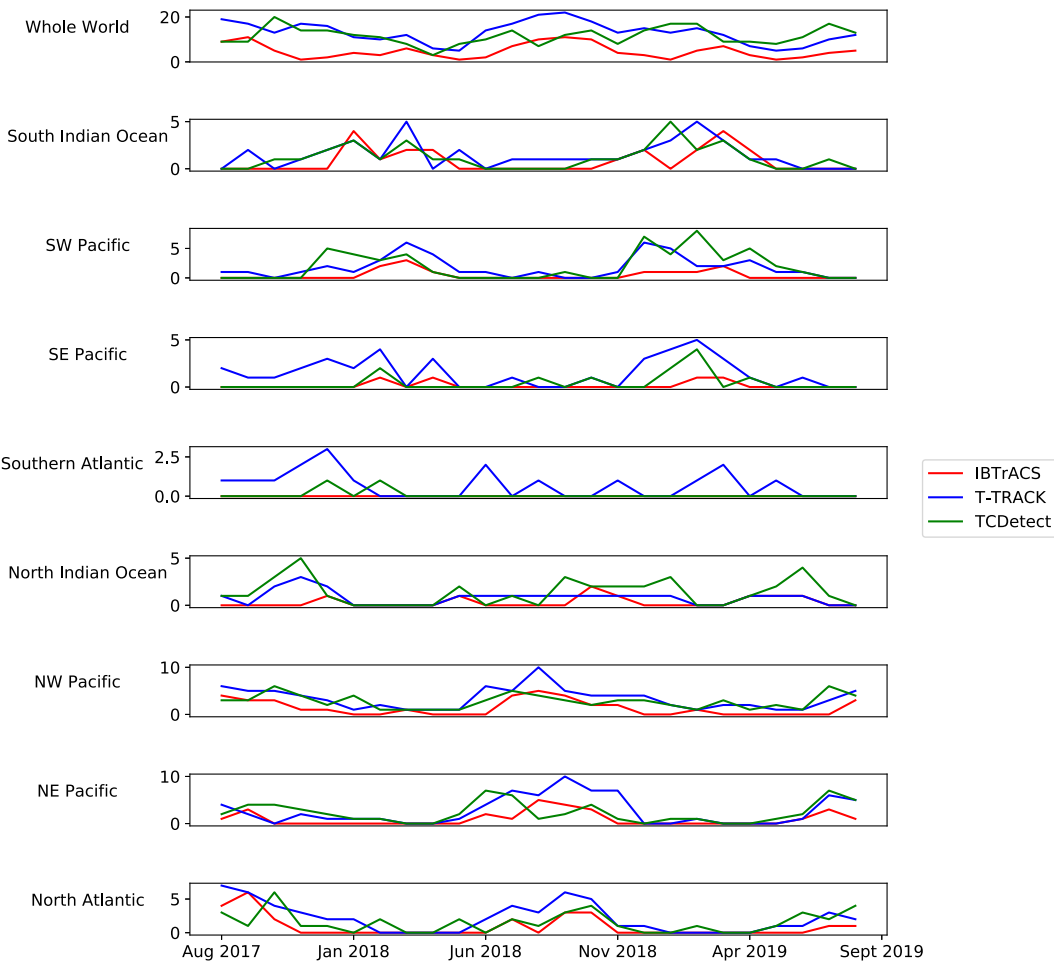
FIG. 5. Hurricane-strength TC frequency: the number of hurricane-strength TC tracks present in a month, as reported by IBTrACS, T-TRACK, and TCDetect, stratified by the regions used by TCDetect.

data. For each method, the composites were created by averaging boxes with sides 30°, centered on the reported TC center. For cases in which the TC was detected by T-TRACK, the TC center used was that as given by T-TRACK. Of the remaining cases, if the TC was present in IBTrACS, the center used was that as given by IBTrACS, and for those TCs that were only

detected by the deep learning model, the TC center used was that as derived from the deep learning model, with the help of the Grad-CAM technique.

The data fields examined were those used as input to the deep learning algorithm: MSLP, 10-m wind speed, and the magnitude of vorticity at 850, 700, and 650 hPa.

TABLE 2. Frequency correlations between IBTrACS and T-TRACK and between IBTrACS and TCDetect. The large correlation intervals (calculated using Fisher's z transform) are indicative of the small sample size and the noisy data. See the main text for an explanation of the South Atlantic (non)result.

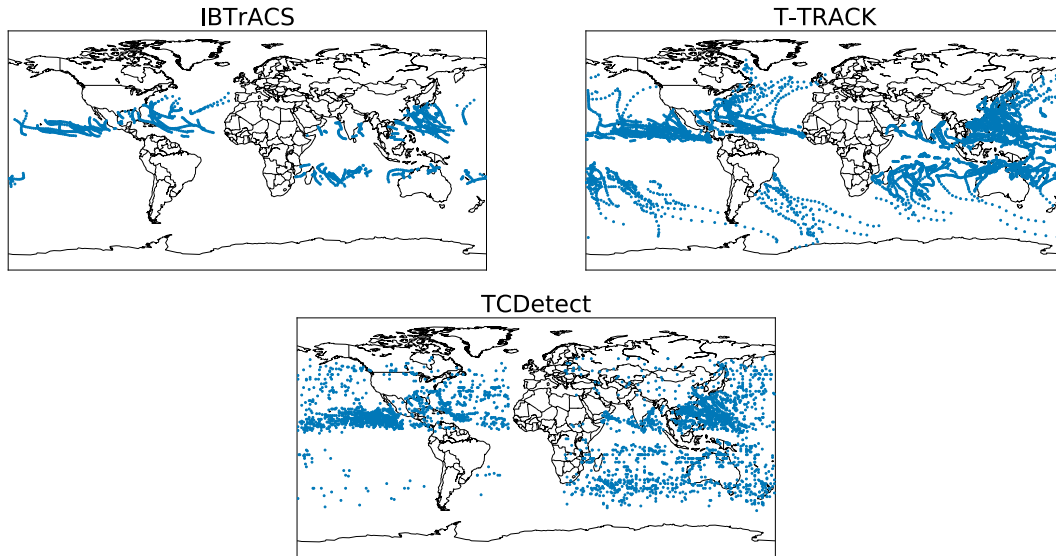|	IBTrACS–T-TRACK		IBTrACS–TCDetect	
	Correlation	95% Confidence interval	Correlation	95% Confidence interval
South Indian Ocean	0.53	0.17–0.77	0.54	0.18–0.77
SW Pacific Ocean	0.73	0.47–0.87	0.47	0.10–0.73
SE Pacific Ocean	0.66	0.36–0.84	0.49	0.12–0.74
South Atlantic Ocean	—	—	—	—
North Indian Ocean	0.25	−0.16–0.59	0.27	−0.14–0.60
NW Pacific Ocean	0.79	0.45–0.87	0.45	0.06–0.71
NE Pacific Ocean	0.72	0.45–0.87	0.42	0.03–0.70
North Atlantic Ocean	0.86	0.71–0.94	0.46	0.07–0.72

FIG. 6. Position of each hurricane-strength TC event center in the constrained matched tracks for (top left) IBTrACS; (top right) T-TRACK; and (bottom) TCDetect.

The composite case for TCs detected by all three methods shows a fairly symmetric low pressure area with a minimum of around 998 hPa. It also shows a wind field with a maximum wind speed of around 13.5 m s^{-1} in the top-right quadrant of the TC and a clear eye. Vorticity is very concentric with very little noise with highs of 0.00024, 0.00021, and 0.000175 s^{-1} at the 850, 700, and 600 hPa levels, respectively. All of the features and magnitudes are similar for composites in both hemispheres.
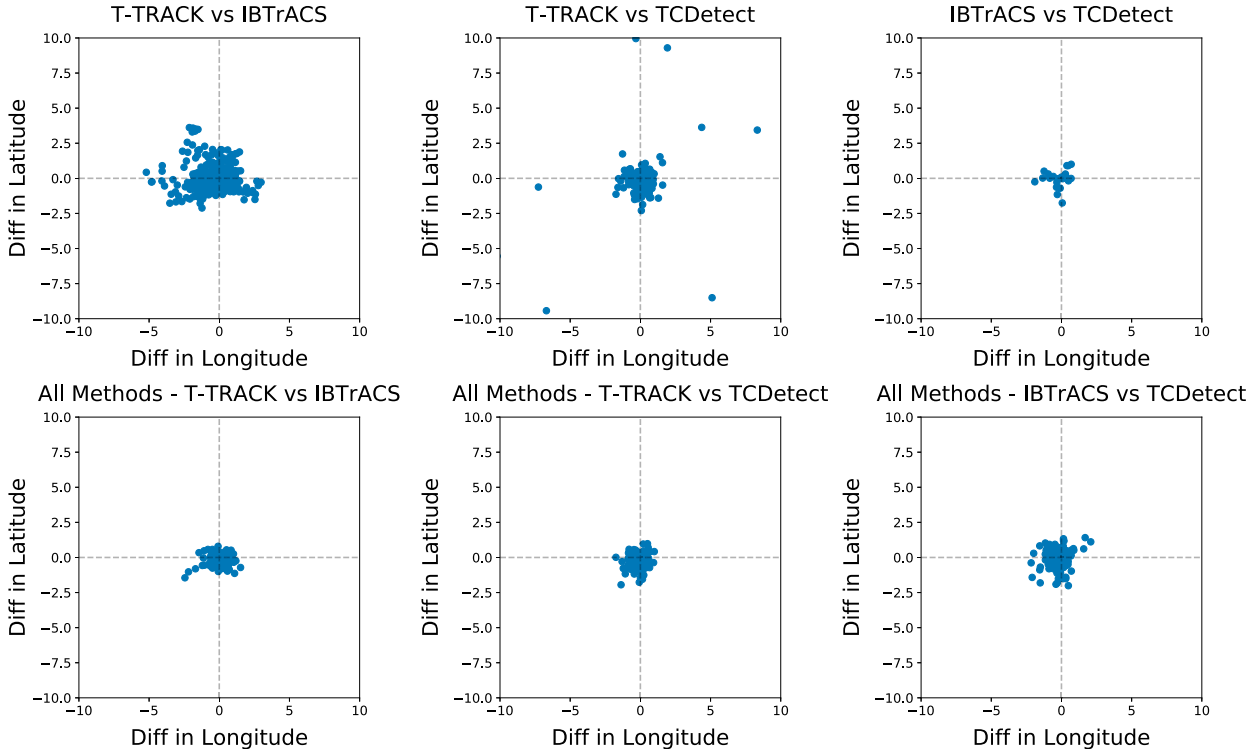

FIG. 7. Spatial correlation of the overlapping regions shown in Fig. 4, i.e., for matches with constraints applied: (top) two-way matches showing pairwise correlation, and (bottom) pairwise comparisons for three-way matches. Differences are calculated using the TC centers obtained from the first method mentioned in the title of each plot minus TC centers obtained from the second-mentioned method.
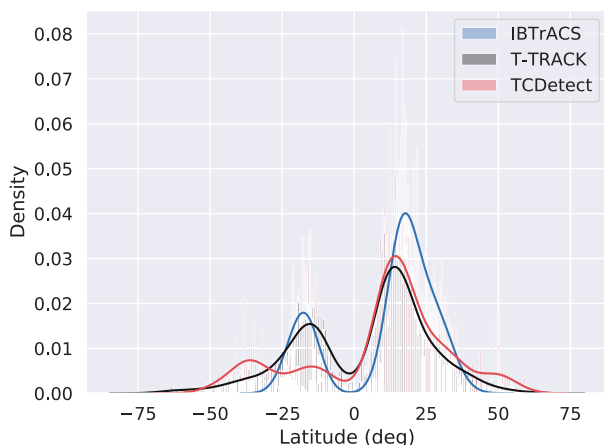
FIG. 8. Density plots of TC center latitude as given by IBTrACS (blue), T-TRACK (black), and the deep learning–based algorithm (red).

The picture is similar, but with some subtle differences, for the composite cases of TCs detected by two of the three detection methods. MSLP fields for these cases have wider low pressure centers, and all have a weaker low, with a central pressure no lower than 1000 hPa. The wind speed field is similar. All cases show more noise in the composite, especially in the composite case derived from TCs detected by T-TRACK and IBTrACS but not TCDetect, but this is somewhat expected, as relatively few TCs are present only in IBTrACS when compared with the other composites. Also, maximum wind speeds are weaker and do not exceed 10.4 m s^{-1}. The vorticity fields show a similar situation, where all vorticity centers are wider and have their maximum magnitude between a third and a half that of the composite case of TCs detected by all detection methods.

When examining these composites when split up by hemisphere (Figs. 9 and 10), one thing of note emerges. It is seen that both MSLP and wind speed fields have a tighter center of circulation for the composite case coming from cases from the Northern Hemisphere than from those originating in the Southern Hemisphere.

Last, the composites for TCs detected by only one of the detection methods show some differences from the composite for the TCs detected by all three methods.

As a general note, it is noticeable that wind speed values in the Northern Hemisphere in cases detected by only one of the two methods, or present only in the observational data, are weaker than those in the Southern Hemisphere.

The composite for TCs present only in IBTrACS shows a low pressure with a considerably higher minimum pressure of 1008 hPa. The maximum wind speed is also down to 8.4 m s^{-1} and does not show a clear eye at the center of the composite. The vorticity fields show wider but much shallower centers, with the maximum vorticity much lower than that of the composite for TCs detected by all three methods. This weaker structure could be due to TC positions given by IBTrACS not lining up well with the position of the TC in ERA-Interim. Considerable noise is also present outside the vorticity

centers, but this is somewhat expected, as relatively few TCs are detected by IBTrACS only when compared with the other composites.

When split up by hemisphere, the composite for cases detected only by IBTrACS shows some differences. First, the MSLP field in the composite for the Northern Hemisphere (Fig. 9) cases shows a wave structure rather than a well-defined low. The wind speed field also shows a lack of a center. The vorticity fields do show clear centers but have considerable noise present.

The composite for cases detected only by IBTrACS originating in the Southern Hemisphere (Fig. 10) shows a much more organized situation. A clear but wide low pressure center is noted, as well as a center in the wind speed field. The vorticity fields also have well-defined but not concentric centers, but there is also a considerable amount of noise present on the outskirts of the centers.

When examining the composite case for TCs detected by the deep learning model only, an MSLP low area is observed with a minimum of around 1009 hPa. A clear center is also seen in the wind speed and vorticity fields. The maximum wind speed is around 7.2 m s^{-1} and the magnitude of the vorticity fields is around half that of the composite case for TCs detected by all three methods.

The same patterns can be observed when the composite of cases detected only by TCDetect is split by hemisphere. However, one thing to note is that the composite for the Southern Hemisphere shows a relatively shallow area of low pressure in the MSLP field when compared with the composite for TCs detected by all three methods.

Last, the composite for TCs detected only by T-TRACK is very similar to that of TCs detected by all three detection methods. The only differences are that the magnitudes for vorticity in the former are about one-half that of the latter. This does not change when the TCs are split by hemisphere.

From the above analysis, it could be concluded that the TCs detected by all three detection methods are the strongest and most well-defined in the data. Hodges et al. (2017) also show this when comparing non-ML TC-tracking algorithms. Furthermore, those detected by two of the methods are weaker, usually with a lack of a clear area of maximum wind speed and somewhat less organized. Those TCs detected by only one detection method are even weaker, with the most noticeable decrease in strength in the vorticity fields. There is also the possibility of wrongly identified TCs, that is, systems that are not a TC, being included in this category.

d. Strength

With the results from the analysis of life cycles and composites in mind, the obvious question that arises is how good a job is TCDetect doing at finding hurricane-strength tropical cyclone events as opposed to all tropical cyclone events?

The match between any class of depression is shown in Fig. 11, which extends Fig. 2a by allowing matches between any class of depression. After doing so, it can be seen that only 17 hurricane-strength events were left unmatched. Also, 80 of the events that were detected by TCDetect and present
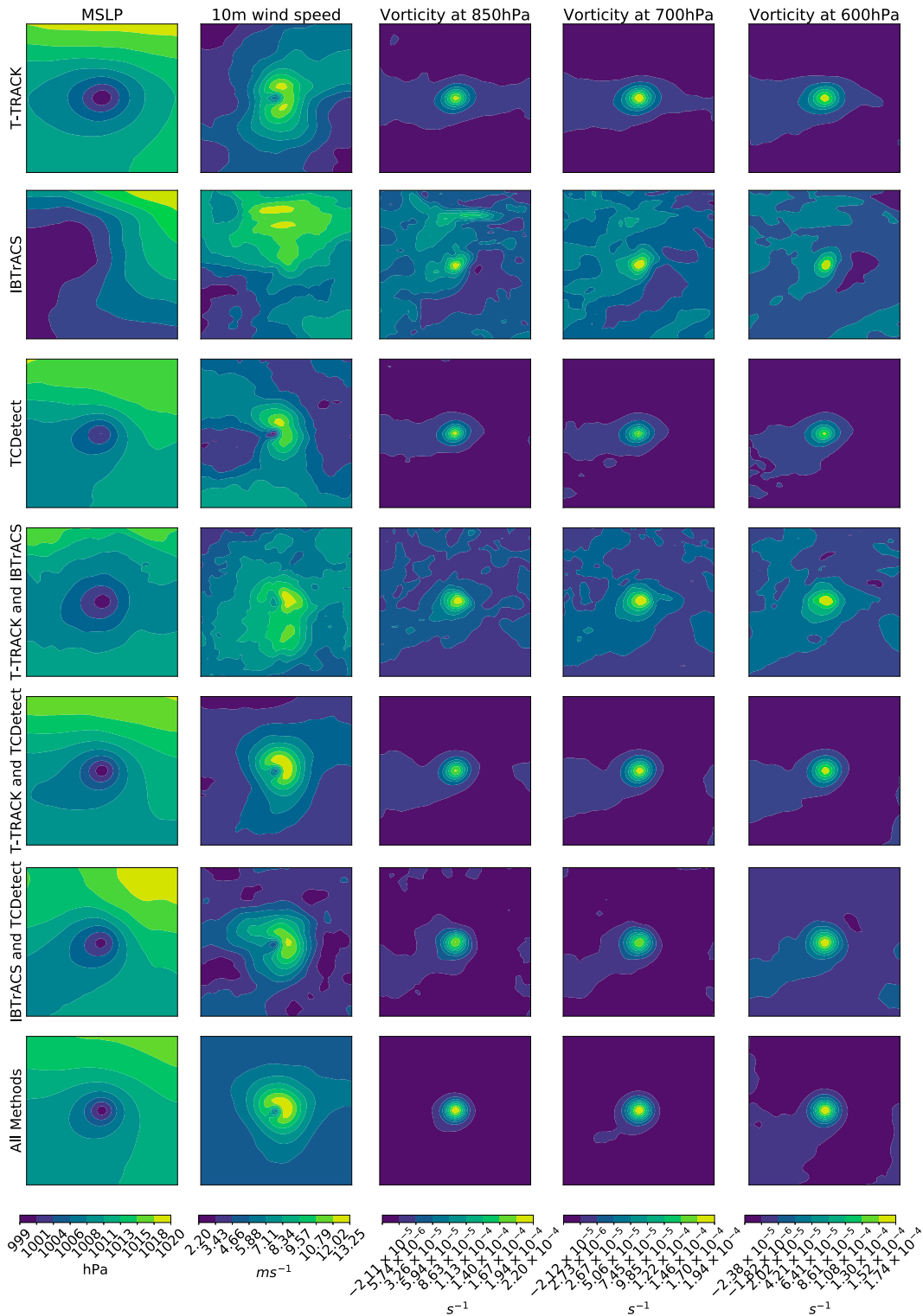
FIG. 9. Composite view of Northern Hemisphere events by detection algorithm or observations that pick up the TC. The total number of cases used to produce each composite can be obtained from Fig. 2 for the variables (left) MSLP, (left center) 10-m wind speed, (center) vorticity at 850 hPa, (right center) vorticity at 700 hPa, and (right) vorticity at 600 hPa.
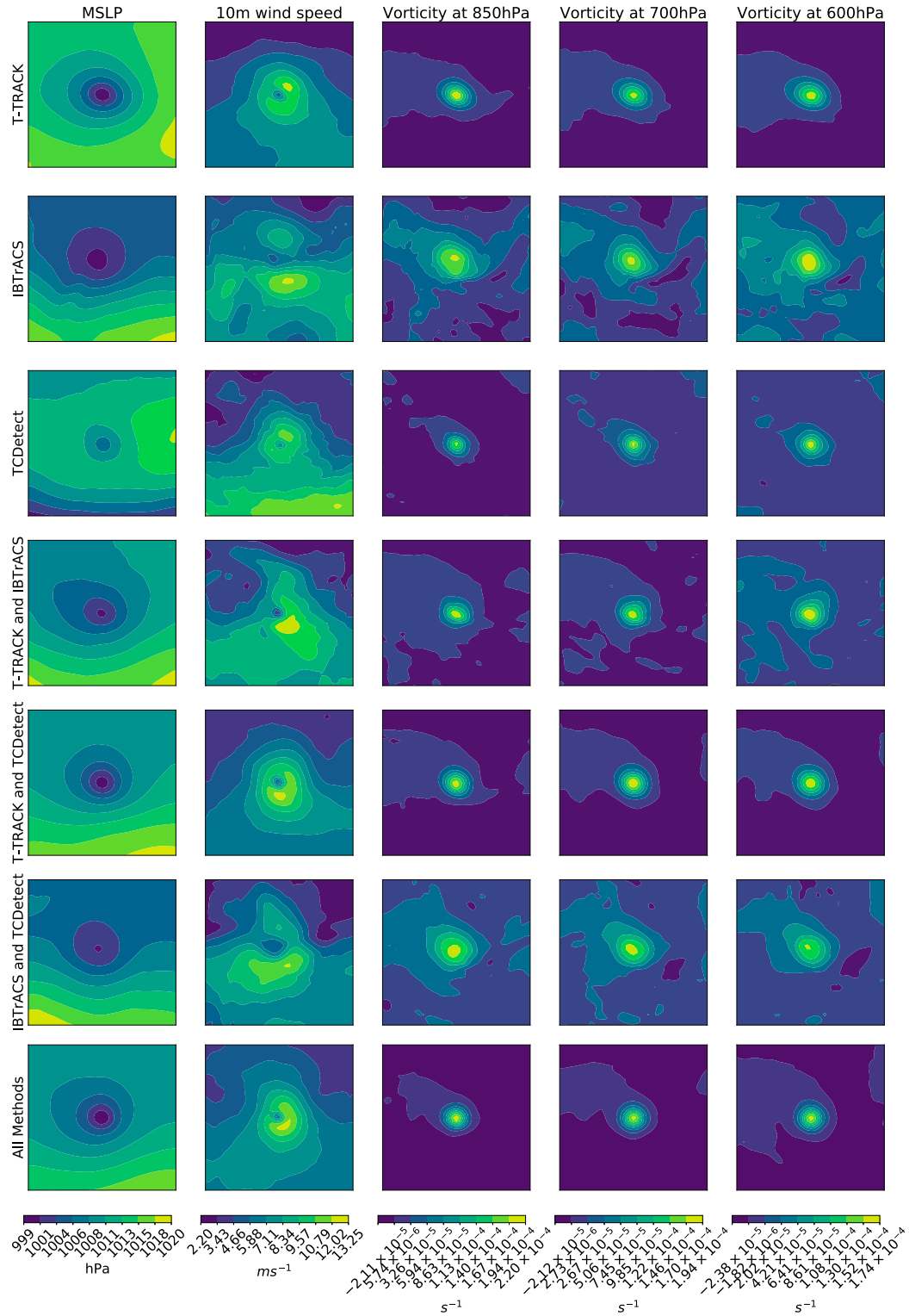
FIG. 10. As in Fig. 9, but for the Southern Hemisphere. Note that the sign of the vorticity has been reversed for ease of comparison.
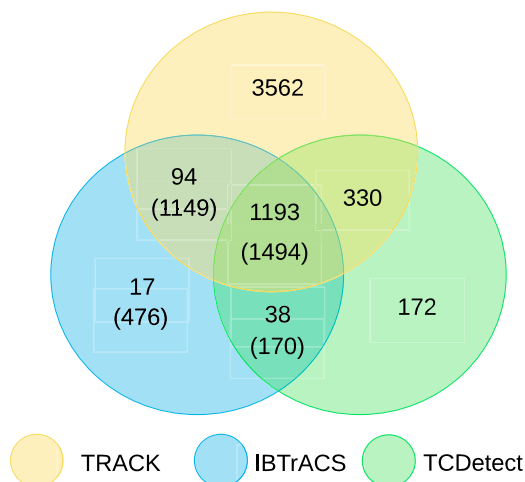
FIG. 11. Events detected by TRACK, events detected by TCDetect, and events reported by IBTrACS. All meteorological systems are included from IBTrACS and TRACK, not just category-1-and-higher systems. Events present in IBTrACS (blue area) were split into TCs of hurricane status (numbers not in parentheses, defined as true positives for TCDetect) and other depressions (values in parentheses, defined as false positives for TCDetect and T-TRACK).

in IBTrACS are now detected by TRACK as well. Similarly, many fewer events are now seen by both TRACK and TCDetect, but not reported in IBTrACS (330 vs 1485 in Fig. 2a). The additional 1494 events found in observations and by both detection methods are showing that TCDetect is doing a good job of finding a range of depressions and storms; although the recall is not as high as for all storms as it is for hurricane-strength TCs (cf. Tables 1 and 3), the precision with respect to recovering storms is good (Table 3).

The details of the matches between TCDetect and IBTrACS are shown in Table 4. The overall precision results from the fact that only 506 out of 3397 TCDetect detections were cases that had no meteorological system present and the vast majority of cases with no TCs are being classified as such, that is, true negatives. Some cases with hurricane-strength TCs present are being misclassified (false negatives), with a greater portion of lower-category hurricane-strength TCs being misclassified than higher-category TCs, consistent with the results in Galea et al. (2023). Also, around two-thirds of non-hurricane-strength TCs are being classified as hurricane-strength TCs (false positives). This is all consistent with the deep learning model recognizing the pattern but struggling to distinguish between strong (deep) and weak (shallow) systems.

TABLE 3. Recall and precision of meteorological disturbances seen in ERA-Interim and labeled by IBTrACS, as recorded by T-TRACK and TCDetect (cf. with Fig. 11, which shows TCDetect recall for TCs).

	Recall	Precision
T-TRACK	85%	50%
TCDetect	63%	85%

TABLE 4. Inferences generated by TCDetect, split by storm type reported by IBTrACS. Positive inferences are where TCDetect detected the presence of a TC; negative inferences are where TCDetect detected no TC. For example, of the 19 759 cases that had no meteorological system, TCDetect classified 506 as having a TC present (i.e., false positives). Similarly, of the 484 cases in which a category-1 TC was the strongest system present, 426 were classified as having a TC (i.e., true positives).

Storm type	Positive inference	Negative inference
No meteorological system	506	19 253
Unknown	2	30
Post-tropical systems	18	47
Disturbances	165	337
Subtropical systems	32	51
Tropical depressions	348	625
Tropical storms	1095	501
Category-1 TCs	426	58
Category-2 TCs	281	26
Category-3 TCs	243	15
Category-4 TCs	212	12
Category-5 TCs	69	0

5. Summary

We have investigated the influence of the structure and location of the underlying events on the detection of hurricane-strength tropical cyclones in eight planetary regions, loosely based on the regions used in the IBTrACS database.

While our primary aim has been to investigate the performance of our new deep learning–based algorithm (TCDetect), of necessity we have had to use an additional technique (T-TRACK) to put our results in context, both in terms of the state of the art, and in terms of the impact of the input (ERA-Interim) not being a perfect copy of the real world that yielded the IBTrACS observations.

T-TRACK is designed to find and track cyclones and storms, and while TCDetect was not trained to find locations and tracks, it is possible to estimate the position of the systems it detects, so the comparison extends not only to counting detections but also to the location and structure of the detected events.

A priori, we might have expected that the events recorded by IBTrACS would be stronger in the observations than in the reanalysis (Strachan et al. 2013; Hodges et al. 2017), and that some events in the Southern Hemisphere would be omitted by the observations (Hodges et al. 2017). These expectations were confirmed in this analysis, but there are also interesting differences in the characteristics of what was detected and where.

Both T-TRACK and TCDetect found more events than appeared in the IBTrACS observations, with both finding more events in the Indian Ocean and TCDetect more over land. The latitudinal distribution of where the events were found differs as well: both T-TRACK and TCDetect find distributions skewed to higher latitudes than those seen in observations, albeit with the peak in numbers at a lower latitude, with the bias to high latitudes more pronounced in the TCDetect data. TCDetect also detected more TCs over the Indian Ocean, giving an extra peak at about 40°S.

Although both TCDetect and T-TRACK found more cyclones in the Southern Hemisphere, for TCDetect the matching of detected and observed cyclones was not as good as for T-TRACK. As already noted, a good number of the extra Southern Hemisphere TCs were found in the Indian Ocean by both techniques although poorly positioned by TCDetect. T-TRACK additionally found many TCs in the South Pacific and east of South America, which might have been omitted in IBTrACS because of a paucity of observing systems in those sectors—but they were not found by TCDetect. The relatively poor geolocation of the TCDetect storms is not unexpected, given that TCDetect was not trained to locate storms, and the method used to find their positions is very post hoc. A future extension to this work could look at training for both detection and location.

Those TCs found by both detection methods and observed in IBTrACS were the strongest and most well defined. Those detected by any two of T-TRACK, IBTrACS, and TCDetect were weaker and had more disorganized fields, and those detected by only one of the methods were the weakest storms present and had considerable noise in their fields.

It was found that most of the false positives (hurricane-strength TC reported but not present in IBTrACS) generated by TCDetect were associated with some sort of TC, albeit without hurricane status. In fact, the overall precision of TCDetect in terms of recovering singular events, that is, TC snapshots, recorded by IBTrACS was higher than T-TRACK—an unexpected result. However, T-TRACK has superior recall (it detects a higher percentage of such storms). This is consistent with the results shown in Galea et al. (2023), who showed that recall in TCDetect is related to storm strength; however, both techniques have similar results in terms of recall of hurricane-strength TCs.

In a companion paper in preparation, TCDetect is used in a GCM, and results are compared at a range of resolutions, for current and future climate. Future work could also look at the sensitivity of TCDetect with different reanalysis data, and, as already noted, the technique could be redeveloped to improve the locations obtained.

REFERENCES

Bengtsson, L., K. I. Hodges, and M. Esch, 2007: Tropical cyclones in a T159 resolution global climate model: Comparison with observations and re-analyses. *Tellus*, **59A**, 396–416, https://doi.org/10.1111/j.1600-0870.2007.00236.x.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, https://doi.org/10.1002/qj.828.

Galea, D., J. Kunkel, and B. N. Lawrence, 2023: TCDetect: A new method of detecting the presence of tropical cyclones using deep learning. *Artif. Intell. Earth Syst.*, **2**, e220045, https://doi.org/10.1175/AIES-D-22-0045.1.

Hodges, K. I., 1994: A general method for tracking analysis and its application to meteorological data. *Mon. Wea. Rev.*, **122**, 2573–2586, https://doi.org/10.1175/1520-0493(1994)122<2573:AGMFTA>2.0.CO;2.

——, 1995: Feature tracking on the unit sphere. *Mon. Wea. Rev.*, **123**, 3458–3465, https://doi.org/10.1175/1520-0493(1995)123<3458:FTOTUS>2.0.CO;2.

——, 1999: Adaptive constraints for feature tracking. *Mon. Wea. Rev.*, **127**, 1362–1373, https://doi.org/10.1175/1520-0493(1999)127<1362:ACFFT>2.0.CO;2.

——, A. Cobb, and P. L. Vidale, 2017: How well are tropical cyclones represented in reanalysis datasets? *J. Climate*, **30**, 5243–5264, https://doi.org/10.1175/JCLI-D-16-0557.1.

Horn, M., K. Walsh, and A. Ballinger, 2013: Detection of tropical cyclones using a phenomenon-based cyclone tracking scheme. US CLIVAR Hurricane Workshop Rep., 30 pp., https://usclivar.org/sites/default/files/documents/2014/2013HurricaneReportFinal.v3_0_1.pdf.

——, and Coauthors, 2014: Tracking scheme dependence of simulated tropical cyclone response to idealized climate simulations. *J. Climate*, **27**, 9197–9213, https://doi.org/10.1175/JCLI-D-14-00200.1.

Knapp, K. R., M. C. Kruk, D. H. Levinson, H. J. Diamond, and C. J. Neumann, 2010: The International Best Track Archive for Climate Stewardship (IBTrACS). *Bull. Amer. Meteor. Soc.*, **91**, 363–376, https://doi.org/10.1175/2009BAMS2755.1.

——, H. J. Diamond, J. P. Kossin, M. C. Kruk, and C. J. Schreck, 2018: International Best Track Archive for Climate Stewardship (IBTrACS) Project, version 4. NOAA/National Centers for Environmental Information, accessed 2 January 2019, https://doi.org/10.25921/82ty-9e16.

Mizuta, R., and Coauthors, 2012: Climate simulations using MRI-AGCM3.2 with 20-km grid. *J. Meteor. Soc. Japan*, **90A**, 233–258, https://doi.org/10.2151/jmsj.2012-A12.

Onogi, K., and Coauthors, 2007: The JRA-25 reanalysis. *J. Meteor. Soc. Japan*, **85**, 369–432, https://doi.org/10.2151/jmsj.85.369.

Roeckner, E., and Coauthors, 2003: The atmospheric general circulation model ECHAM 5. Part I: Model description. Max Planck Institute for Meteorology Rep. 349, 140 pp., https://pure.mpg.de/rest/items/item_995269_4/component/file_995268/content.

Saha, S., and Coauthors, 2014: The NCEP Climate Forecast System version 2. *J. Climate*, **27**, 2185–2208, https://doi.org/10.1175/JCLI-D-12-00823.1.

Schenkel, B. A., and R. E. Hart, 2012: An examination of tropical cyclone position, intensity, and intensity life cycle within atmospheric reanalysis datasets. *J. Climate*, **25**, 3453–3475, https://doi.org/10.1175/2011JCLI4208.1.

Schmidt, G. A., and Coauthors, 2014: Configuration and assessment of the GISS modelE2 contributions to the CMIP5 archive. *J. Adv. Model. Earth Syst.*, **6**, 141–184, https://doi.org/10.1002/2013MS000265.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, 2017: Grad-CAM: Visual explanations from deep networks via gradient-based localization. *2017 IEEE*

Int. Conf. on Computer Vision (ICCV), Venice, Italy, Institute of Electrical and Electronics Engineers, 618–626, https://doi.org/10.1109/ICCV.2017.74.

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

Ullrich, P. A., and C. M. Zarzycki, 2017: TempestExtremes: A framework for scale-insensitive pointwise feature tracking on unstructured grids. *Geosci. Model Dev.*, **10**, 1069–1090, https://doi.org/10.5194/gmd-10-1069-2017.

Uppala, S. M., and Coauthors, 2005: The ERA-40 Re-Analysis. *Quart. J. Roy. Meteor. Soc.*, **131**, 2961–3012, https://doi.org/10.1256/qj.04.176.

Walsh, K. J. E., M. Fiorino, C. W. Landsea, and K. L. McInnes, 2007: Objectively determined resolution-dependent threshold criteria for the detection of tropical cyclones in climate models and reanalyses. *J. Climate*, **20**, 2307–2314, https://doi.org/10.1175/JCLI4074.1.

Zarzycki, C. M., and P. A. Ullrich, 2017: Assessing sensitivities in algorithmic detection of tropical cyclones in climate data. *Geophys. Res. Lett.*, **44**, 1141–1149, https://doi.org/10.1002/2016GL071606.

Zhao, M., I. M. Held, S.-J. Lin, and G. A. Vecchi, 2009: Simulations of global hurricane climatology, interannual variability, and response to global warming using a 50-km resolution GCM. *J. Climate*, **22**, 6653–6678, https://doi.org/10.1175/2009JCLI3049.1.