Investigating differences between Tropical Cyclone detection systems

Daniel Galea,[a] Kevin Hodges,[b] Bryan N. Lawrence,[a,b]

[a] *Department of Computer Science, University of Reading, UK*

[b] *National Centre of Atmospheric Science, Department of Meteorology, University of Reading, UK*

Corresponding author: Daniel Galea, galea.daniel18@gmail.com

ABSTRACT: Tropical cyclones (TCs) are meteorological events that are identified in simulations by detection algorithms. Where the simulations are re-analyses, detection algorithms can be compared with observations recorded using the International Best Track Archive for Climate Stewardship (IBTrACS). In this work, a novel deep-learning-based detection algorithm, TCDetect, is compared with a state of the art tracking system and IBTrACS to provide context for use in analysing climate simulations. TCDetect is applied to ERA-Interim data, and focuses on whether the tropical cyclone events are detected in a given region. Previous work has shown the scheme has good recall, here the question addressed is whether the structure of the TCs detected or observed play a part in the performance? A key part of the comparison is the recognition that ERA-Interim itself does not fully reflect the observations, and so no detection algorithm operating on ERA-Interim will fully recover the IBTrACS observations. However, for strong well-defined cyclone events, the two detection algorithms operating on reanalysis agree well with each other and the IBTrACS observations, with comparable performance across all areas of the globe. Where events are detected by only one algorithm (or only in observations) they are the weakest events with around half the maximum vorticity seen in the events detected by both algorithms and the observations. Furthermore, the events detected by both algorithms and the observations have the least amount of noise in their fields and have a clear centre of circulation.

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, i.e. 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. (2022b) introduced a deep learning technique for detecting the presence or absence of a TC in a field of simulation data, named TCDetect. In this study, we compare the performance of this model to a state-of-the-art non-machine-learning algorithm and an observational dataset. Section 2 describes previous literature comparing various detection algorithms, Section 3 describes the data and detection algorithms used in this study, and Section 4 describes the results obtained when comparing TCDetect with a version of TRACK (Hodges et al. 2017) applied to re-analysis data and compared with reality as recorded by the International Best Track Archive for Climate Stewardship (IBTrACS, Knapp et al. 2010, Knapp et al. 2018) archive to understand some the limitations of the application of our technique for feature identification in simulation data.

2. Previous Studies

There is extensive previous literature comparing different automatic TC detection algorithms.

Horn et al. (2014) compare four different detection algorithms, namely a modified version of the Commonwealth Scientific and Industrial Research Organisation (CSIRO) tracking scheme (Walsh et al. 2007, Horn et al. 2013), the Zhao tracking scheme (Zhao et al. 2009), and those developed by the modelling groups whose data was involved, i.e. the groups from the Meteorological Research Institute (MRI), the National Aeronautic 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 models used were the CMCC-INGV ECHAM5 model which has ~90-km grid spacing at equator (Roeckner et al. 2003); the NASA-GISS model which has ~110-km grid spacing at the equator (Schmidt 2014); the National Center for Environmental Prediction (NCEP) Global Forecast System (GFS) which has ~110-km grid spacing at the equator (Saha 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).

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

Given that the requirement for these automatic tracking algorithms is to detect TCs in a particular set of data which correspond to those that occurred in real-life, it is only natural that the threshold values are tuned to obtain the same number of TCs. 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, TempestExtremes (Ullrich and Zarzycki 2017) applied to four different reanalysis datasets. They found that the most sensitive thresholds were those defining the TC vortex strength, for example for the depth of the minimum of sea level pressure (SLP) or warm core strength. They reported a larger difference when comparing storm count rather than integrated or weighted metrics such as the number of days with a TC present or accumulated cyclone energy (ACE). Zhao et al. (2009) also found that the threshold for minimum duration of a TC was sensitive to the choice made while previous literature also seems to agree that even though some differences might be observed between detection methods, there are little disagreements on strong TCs, i.e. those that are at least of category 3 on the Saffir-Simpson scale.

It was also noted in some of this work that the intensity of TCs, whether surface winds or the depth of the minimum mean sea level pressure (MSLP), is underestimated in all of the reanalyses datasets. Strachan et al. (2013) noted that resolution alone does not explain this observation due to feedback processes present in the model. Despite this, they still noted that any wind speed threshold should vary linearly with resolution and any deviations from this relationship are likely due to model biases and errors.

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 JRA25 and JRA55 reanalyses are most realistic, due to a vortex relocation 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 higher than 60km are capable of showing the correct inter-annual variability even a resolution of 20km is 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 (97% in NH; 92% in SH) at tracking TCs across all basins, but that it had a high false detection rate, especially in the Southern Hemisphere, when considering only TCs which fulfilled criteria which considered intensity and presence of a warm core. Most of these false positives had their genesis at a latitude greater than 20^{o}S, leading to the conclusion that these may have been hybrid TCs of some sort. 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.

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: both the observations (the ground truth) and the simulation data used as input 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 the best-track data for TCs from all the official Tropical Cyclone Warning Centers, the WMO Regional Specialized Meteorological Centers (RSMCs), and other sources.

TRACK is a state-of-the-art automatic detection and tracking system for different types of atmospheric disturbances with considerable use since inception. Here the TC tracking component is used as a gold-standard comparator against which to compare the results from the deep learning model.

We use TCDetect and TRACK applied to re-analysis data and compare them to each other and the IBTrACS dataset. Re-analysis data provides the best possible synthesized observations of meteorological variables; we choose to use the ERA-Interim product (Dee et al. 2011). ERA-Interim utilises 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 to 2019.

ERA-Interim data is produced at a spatial resolution of 79km, a temporal resolution of 6 hours and has 60 vertical levels up to 0.1hPa. Of the many parameters produced, only the mean-sea level pressure (MSLP), 10-metre wind speed and relative vorticity at 850hPa, 700hPa and 600hPa are used in this study.

The comparison presented here is limited to the 25 months between the 1st of August 2017 until the end of August 2019 as the earlier data has been used in training the deep learning algorithm.

a. IBTrACS

The IBTrACS dataset has information about reported storms, such as the storm centre in latitude and longitude, maximum surface wind speed, minimum sea level pressure and category. It records both events and tracks (sequences of events from the same system).

While IBTrACS is the best available observational dataset, some inhomogeneity exists between each contributing source as the different centres have differing observing systems and parametric approaches. Such observing systems can be limited in time and space, leading to the omission of events not detected or an incomplete record of their evolution, particularly if they had limited or no human impact, or they were out of range of detection systems such as airborne missions.

b. The TCDetect Deep Learning Model

The TCDetect deep learning TC detection scheme was described in Galea et al. (2022b). It uses a deep learning scheme trained on ERA-Interim data which utilises mean sea-level pressure (MSLP), 10-metre wind speed, and vorticity at 850hPa, 700hPa and 600hPa; all coarsened to a sixteenth of ERA-Interim's native spatial resolution, resulting in an input resolution of approximately 320km. This data coarsening step was arrived to as a result of hyperparameter tuning, and helped filter out small-scale noise.

These data were passed through a convolutional base connected to a fully-connected dense classifier trained to detect TCs labelled using IBTrACS. The system inferred a classifier value ranging between 0 and 1; a tropical cyclone is inferred to be present if the value is greater than 0.5, and absent if less than or equal to 0.5.

For the identification of tropical cyclones and using 0.5 as the boundary, when trained on ERA-Interim, the TCDetect algorithm obtained a recall rate of 92% with a precision rate of 36%. In practice, this means that the while most of the actual TCs were detected, many of the TCs identified were technically false negatives (i.e. not storms of strength 1 or greater on the Saffir-Simpson scale). However, as discussed in Galea et al. (2022b) and further discussed below, most of these were actually meteorologically significant.

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 on these results"? We address this question by applying both T-TRACK and TCDetect to ERA-Interim, and comparing the results with the IBTrACS "ground truth-labels".

The TC centre is not given by TCDetect, so a way to extract it was needed. For this, the Gradient Class Activation Map technique (Grad-CAM, Selvaraju et al. (2017)) was used: for a given input, the output of the deep learning model is passed back through the model and together with gradient maximisation, produces a heatmap of the input areas used in a selected layer en route to the output. For TC location, we selected the first convolutional layer, and assumed that the TC central position in latitude and longitude is co-located with the maximum activation.

Because the heatmaps used for Grad-CAM were generated from the coarsened (320km resolution) data, the resulting TC centres were coarsely quantized and only poor quality comparisons were possible. To mitigate this effect, the Grad-CAM centres ("interim centres") were then passed through an additional refinement step to generate more accurate locations. A box with sides of 10 degrees in latitude and longitude was centred on the interim centres, and the original full resolution

ERA-interim vorticity values at 850hPa, 700hPa and 600hPa were obtained and vertically averaged. The TC centre was assumed to be located at the position of the maximum in the absolute value of the averaged vorticity.

These TC centres were then used to make up TC tracks. Given that only one TC centre could be produced per region at any one timestep, a track was first defined as having TC centres which were present in consecutive timesteps in the same region. However, this produced many short (< 2 days) tracks. To try and fix this, tracks for a single region which had at most 2 days (8 timesteps) of no TC being detected and a separation distance of 20 degrees (geodesic) between the final TC centre from one track and the initial TC centre of the next track 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 had some trouble with locating TC centres, so some buffer was built into this criterion.

c. TRACK

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

In the first step, TRACK treats the data so that features of interest are easier to detect. This is done with the help of spectral filtering to only keep features which have spatial scales in the range of the features of interest. With regards to tropical cyclones, the features present in wavenumbers 5 to 63 are kept in the vertical average of vorticity between the heights of 850hPa and 600hPa.

During the segmentation stage, each point in each timestep of any data used 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 5×10^{-6} s^{-1}. The object points are then collected into objects.

Feature point detection then allocates a feature point to each object, representing its centre. This feature point is selected as the local extremum in the vertically-averaged vorticity field.

Finally, the tracking stage uses the feature points generated to minimise a constrained cost function to get the smoothest possible tracks.

The complete TRACK algorithm finds a range of cyclones, some of which may be TCs. The tracks produced can be processed to identify only TC tracks.

The original tracks produced by TRACK are then processed to remove any non-TC cases from the start of each track. This is done by using criteria similar to those given by Bengtsson et al. (2007):

- a lifetime of at least 2 days
- the initial point in the track must be in between the latitudes of 30°S and 30°N
- a maximum in T63 vertically-averaged relative vorticity intensity at 850hPa over 5×10^{-6} s^{-1}
- a warm core check: a T63 vorticity maxima for each atmosphere level up to 250hPa and that the difference between the maxima at 850hPa and 250hPa is above a 5×10^{-6} s^{-1}
- the last two conditions holding for the last *n* timesteps, 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 this new set of tracks which conform to these criteria as the "truncated-TRACK" dataset or T-TRACK.

4. Results

The first question to consider is “To what extent do the two detection algorithms recover the TC events seen in the observations?”. We can then ask “How well do the two algorithms (combined with ERA-Interim data) position the TCs in space?”. Finally, we ask “To what extent does the detection success depend on the TC structure and strength?"

a. Detection

Figure 1 shows the relationship between detection and observations for all the events during the period of interest. For these purposes, an event was counted when a TC (Cat-1 or greater on the Saffir-Simpson scale) was observed and/or detected in any timestep. TCs in different regions in the same timestep would give an event for each region in which a TC is seen. However, if multiple TCs are in the same region in the same timestep, this is considered to be one event.

In total there were 1342 such events in the IBTrACS data, and 4741 and 3397 detected by T-TRACK and TCDetect respectively (Figure 1a). The majority of the observed events were found by both detection algorithms, with TCDetect finding slightly more than T-TRACK. Relatively few

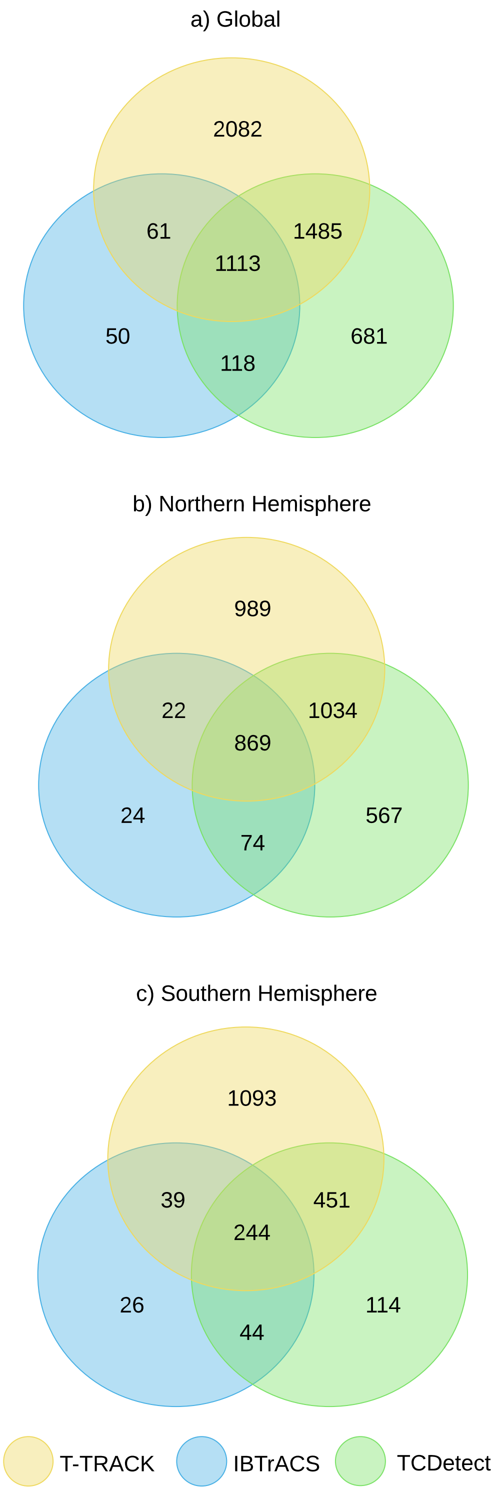

FIG. 1. Events reported by observations (IBTrACS) and detected by T-TRACK and TCDetect applied to ERA-Interim data for (a) the whole globe, (b) the Northern Hemisphere and (c) the Southern Hemisphere.

(50) IBTrACS events were not found by one or other 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 non-TC meteorological events were being incorrectly classified as TCs. This finding is discussed further below.

With an a priori expectation that IBTrACS may be undersampling TC events in the Southern hemisphere, the data was also split into hemispheres to investigate (Figure 1b/c). 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.

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

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

It is worth noting that the criteria used to supposedly screen TRACK to identify 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 97/92% in the northern/southern hemispheres respectively, albeit with many more false positives.

Of the 3397 cases in which TCDetect detects a TC, 681 cases, or around 20% are not observed or detected by T-TRACK, and similarly, of the 4741 cases in which T-TRACK detected the presence of a 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 TCs (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 timestep "events" since the algorithms (TRACK and TCDetect) are applied to one timestep 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 which distinguish T-TRACK from TRACK. These phenomena move along tracks and so we can consider track detection independently of event detection.

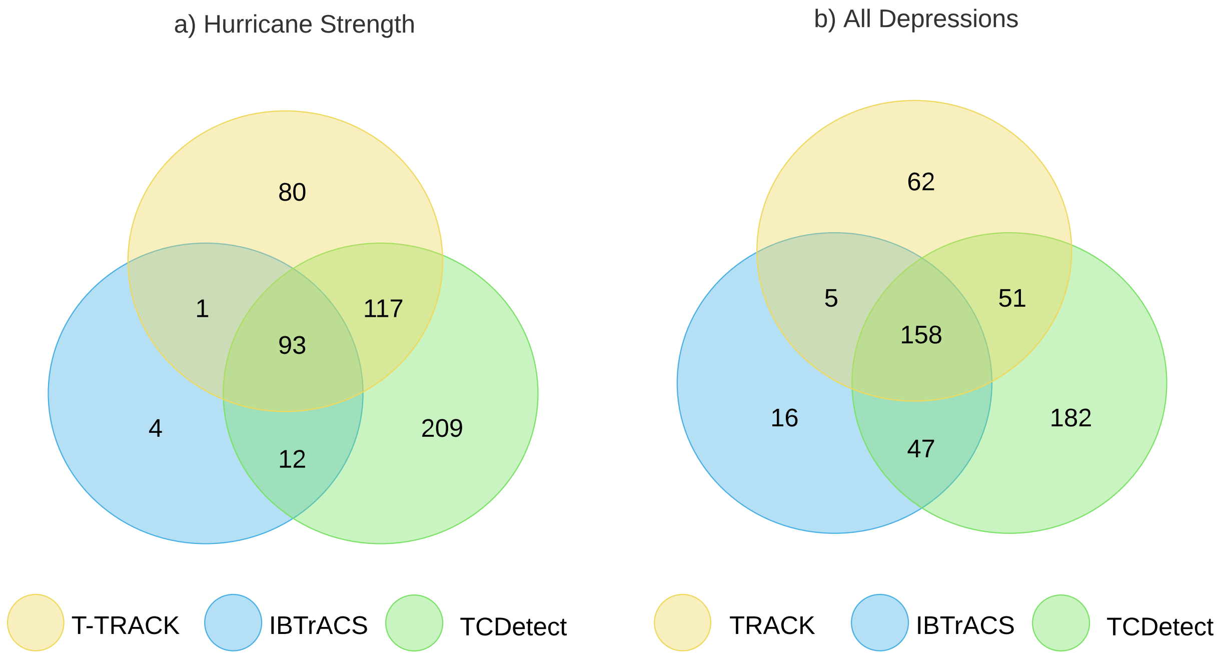

FIG. 2. Tracks reported by observations (IBTrACS) and detected by T-TRACK and TCDetect applied to ERA-Interim data. Overlaps occur when they share a detection event at some point along the track in the same region at the same timestep. Tracks are matched for (a) only TCs (hurricane-strength) and (b) all depressions (i.e. a superset of a).

In terms of tracks, Figure 2a shows how many TC tracks match, whereby two tracks are matched across datasets if they share one or more detection events — in the same region at the same timestep. (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, Figure 2b shows matching tracks where depression events were also considered.

The majority (96%) of IBTrACS tracks, whether depressions or hurricanes, 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 (88% of hurricanes) of the matched IBTrACS tracks were 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

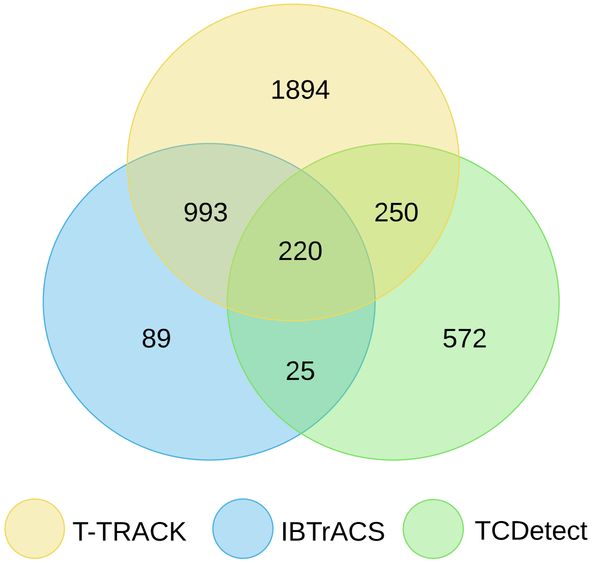

FIG. 3. Events detected by TRACK, TCDetect and/or reported by IBTrACS which fall on matching tracks, defined by applying constraints similar to those of Hodges et al. (2017).

picked up by the detection algorithms which either had not strengthened to hurricane strength or were non-tropical systems, an argument supported by the increased number of three-way matches seen when including all depression tracks (2b), and our earlier analysis for TCDetect and IBTrACS alone. The most unmatched tracks come from TCDetect and were due to many non-meteorological false positives. However, it is encouraging that most of the tracks either produced by TRACK or given in IBTrACS are being matched by tracks produced by TCDetect.

To further understand the differences between the two methods and the observations, the tracks from both detection algorithms and the observations for hurricane-strength TCs were matched using the following criteria. These were similar to those used by Hodges et al. (2017):

- the mean separation distance between all overlapping points between tracks is less than 5^o (geodesic)
- the tracks need to overlap for at least 10% of the base track's lifetime
- the track with the least mean separation distance is chosen if multiple matching tracks exist

These constraints remove any of the unmatched TC tracks, but events can still not match, since they may fall on part of a track where those events were not detected/observed by another method.

After these criteria were applied, the TC events from the remaining tracks were again split by method. The matches between detection and observations now includes fewer TC events (compare Figure 3 and Figure 1a). The number of cases with the presence of a TC reported by IBTrACS decreases from 1342 to 1327. The same occurs for those given by T-TRACK (4741 to 3357) and TCDetect (3397 to 1067).

The small change in total TCs for IBTrACS is expected, given the vast majority of tracks are detected at some point during their evolution. The biggest change is seen in the TCDetect results where many events are rejected because they either didn't lie on a track, or the tracks were incorrectly positioned and lay outside the 5^o (geodesic) criterion.

The overall frequency of TC events shown in Figure 1a is seen in the observations and the detected events is shown in Figure 4. While T-TRACK and TCDetect detect more TC tracks, they share the same intra-annual variability seen in the IBTrACS observations. Regions in the Northern Hemisphere show an uptick in TC frequencies in the months between July and October, while TC frequencies increase in the months between December and June for regions in the Southern Hemisphere.

b. Location

The question as to how well TCDetect locates TC centres given the matching technique is now addressed in more detail. Figure 5 shows the location of the events reported using each technique following the matching technique discussed above.

The IBTrACS data is here 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 further out to the middle of the Atlantic
- 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
- over the middle of the Indian ocean and to the north of Australia

In comparison, T-TRACK shows a larger number of events and longer tracks, some extending well into the sub-tropics. There are also more TC centres present in the Southern Hemisphere

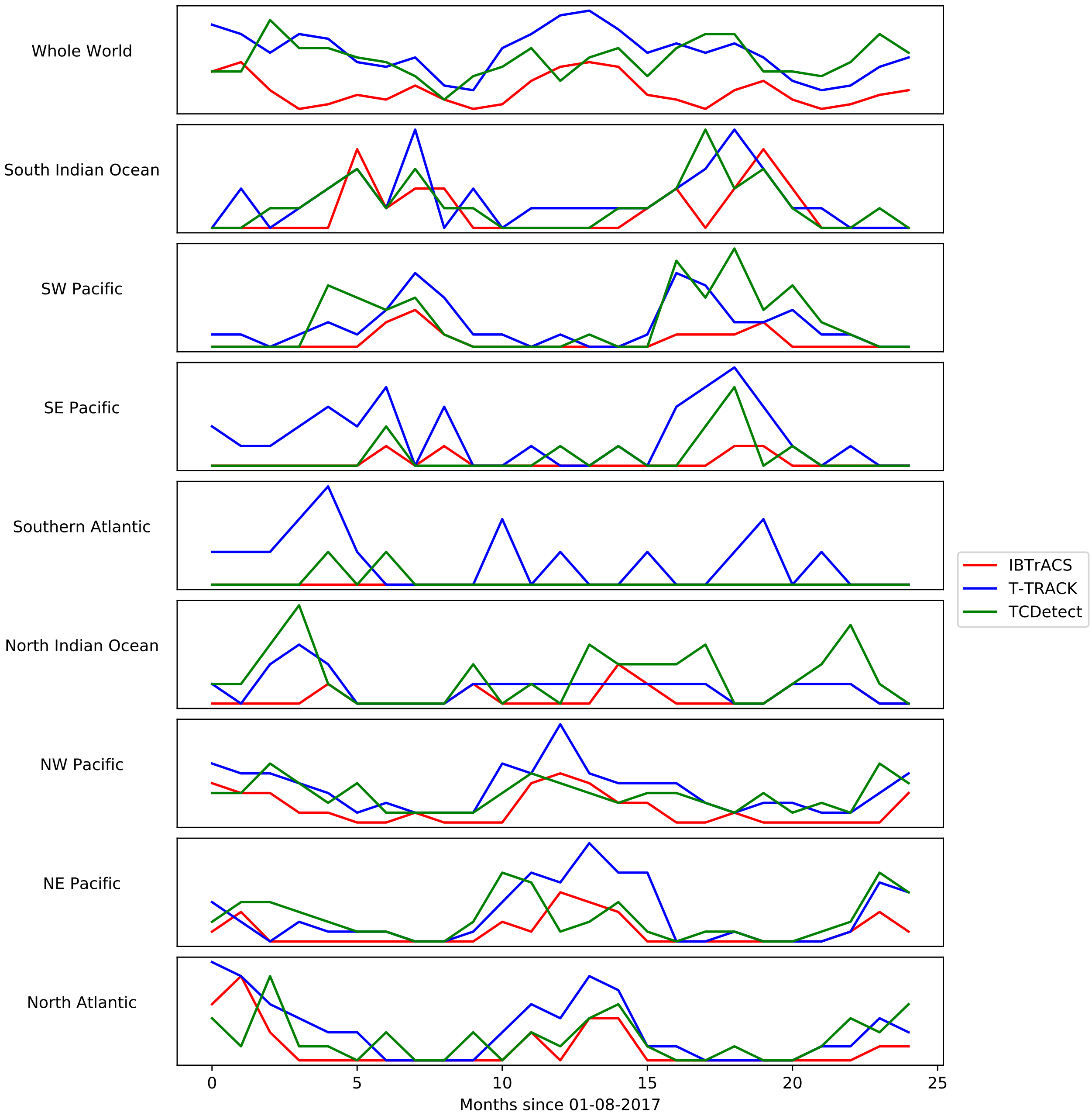

Fig. 4. TC frequency: the number of TC tracks present in a month, as reported by IBTrACS, T-TRACK and TCDetect, stratified by the regions used by TCDetect.

than IBTrACS, especially the Central Southern Pacific ocean. The locations off the eastern coast of the South American continent, which are non-existent in IBTrACS, could point to the use of re-analysis data and tracking algorithms providing better ground truth in observation poor regions

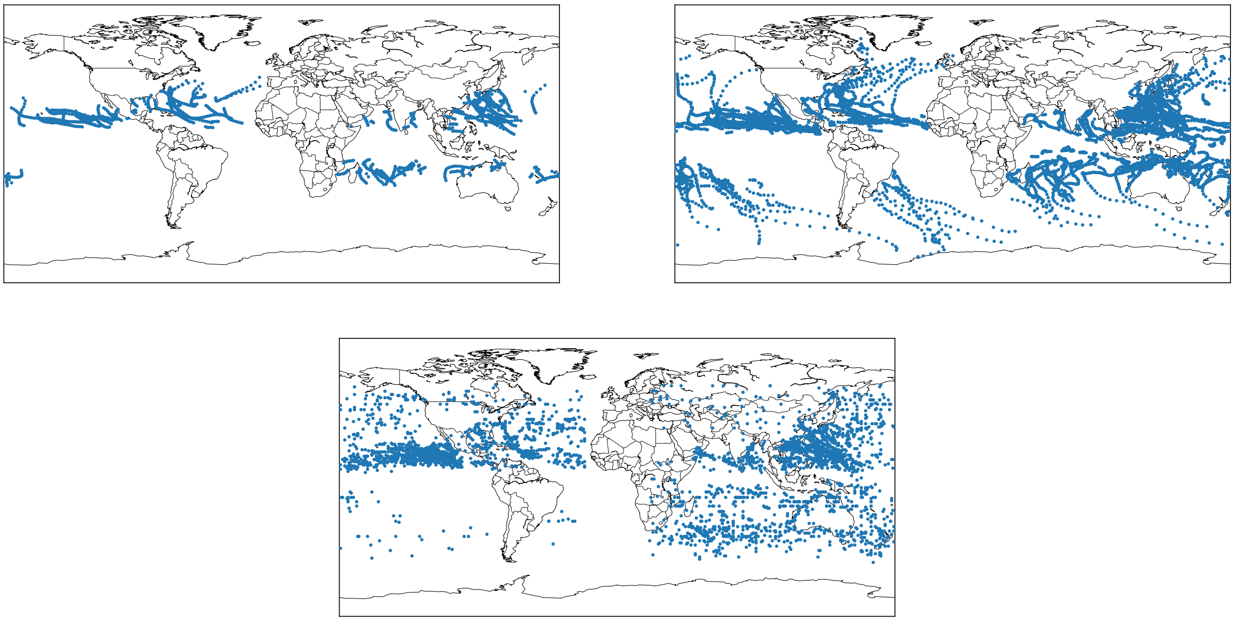

Fig. 5. Position of each Tropical Cyclone event center as given by IBTrACS (top-left); T-TRACK (top-right) and the deep learning model (bottom).

of the globe and/or certain operational procedures which mean that the centre responsible excludes these storms from their reports.

The locations reported by TCDetect are positioned mostly in the right regions, but some centres are located well inland or well into the subtropics, where TCs are not expected. Also, the centres 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 which have centres within 10^{o} (Figure 6). 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 centres were further than 10^{o} 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.

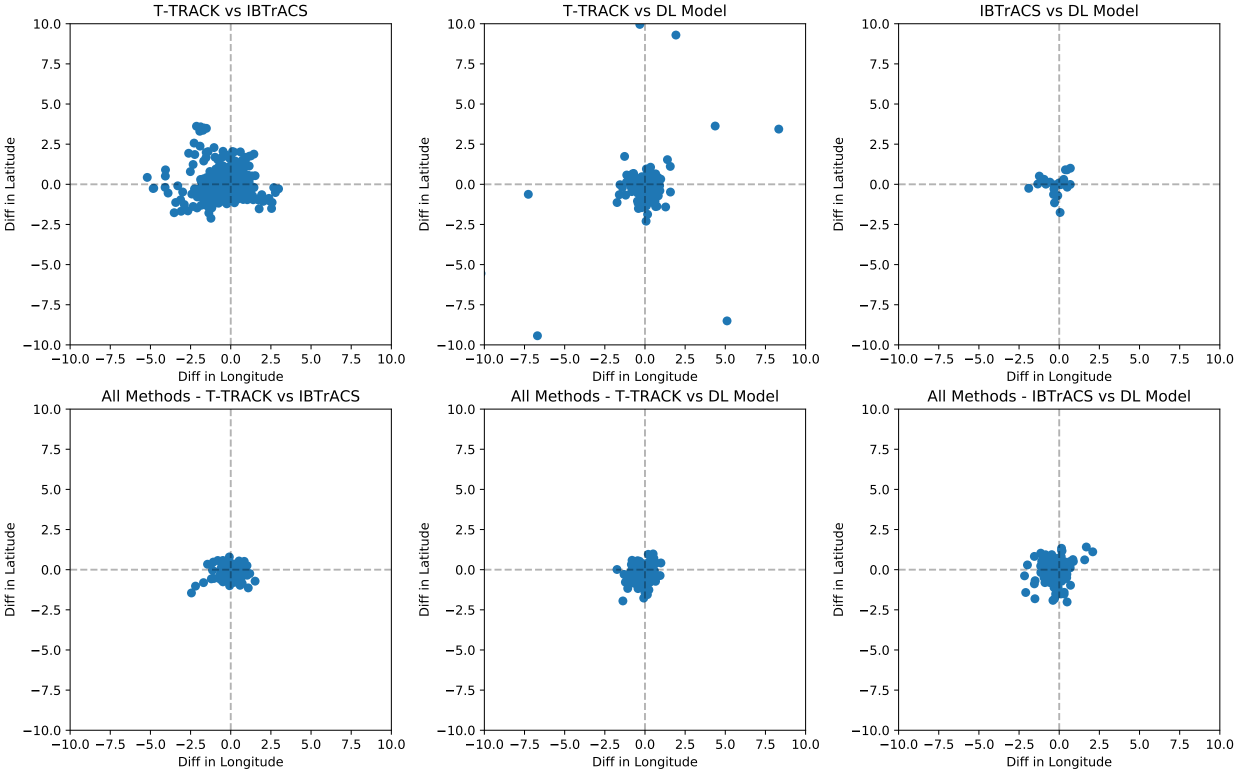

Fig. 6. Spatial correlation of the overlapping regions shown in Figure 3, i.e. for matches with constraints applied. Top row pairwise matches showing pairwise correlation. Bottom row, matches in all three methods, but still pairwise correlations.

Figure 7 shows the distribution of all TC cases by latitude. While the peak of the distributions for both detection algorithms in both hemispheres is biased equatorwards (with respect to IBTrACS observations) the two detection algorithms broadly agree. 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 Figure 1 using the ERA-interim data. For each method the composites were created

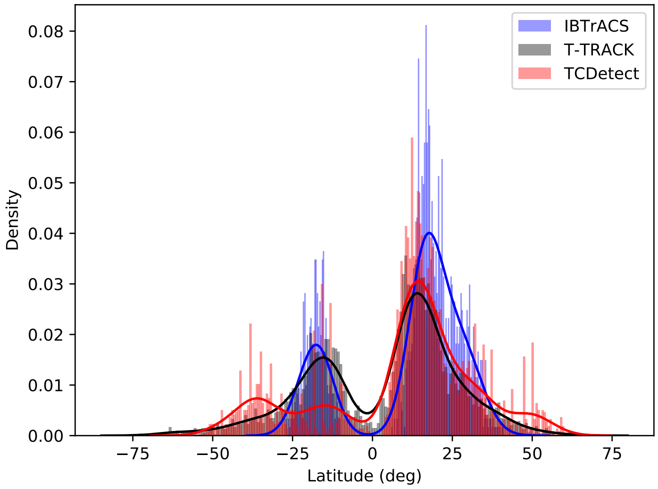

Fig. 7. Density plots of TC centre latitude as given by IBTrACS (blue), T-TRACK (black) and the deep learning based algorithm (red).

by averaging boxes with sides 30°, centered on the reported TC centre. For cases in which the TC was detected by T-TRACK, the TC centre used was that as given by T-TRACK. Of the remaining cases, if the TC was present in IBTrACS, the centre used was that as given by IBTrACS, and for those TCs that were only detected by the deep learning model, the TC centre 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: mean sea level pressure (MSLP), 10-m wind speed, and the magnitude of vorticity at 850, 700 and 650 hPa.

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 the maximum wind speed of around 13.5 m s^{-1} in the top-right quadrant of the TC and a clear eye. Finally, vorticity is very concentric with very little noise with highs of 0.00024 s^{-1}, 0.00021 s^{-1} and 0.000175 s^{-1} at the 850hPa, 700hPa and 600hPa levels respectively. All the features and magnitudes are similar for composites in both hemispheres.

The picture is similar with some subtle differences for the composite cases of TCs detected by two of the three detection methods. MSLP fields for these cases have slightly wider low centres and all have a weaker low with a central pressure no lower than 1000hPa. The wind speed field is similar. All cases show more noise in the composite, especially in the composite case derived from TCs

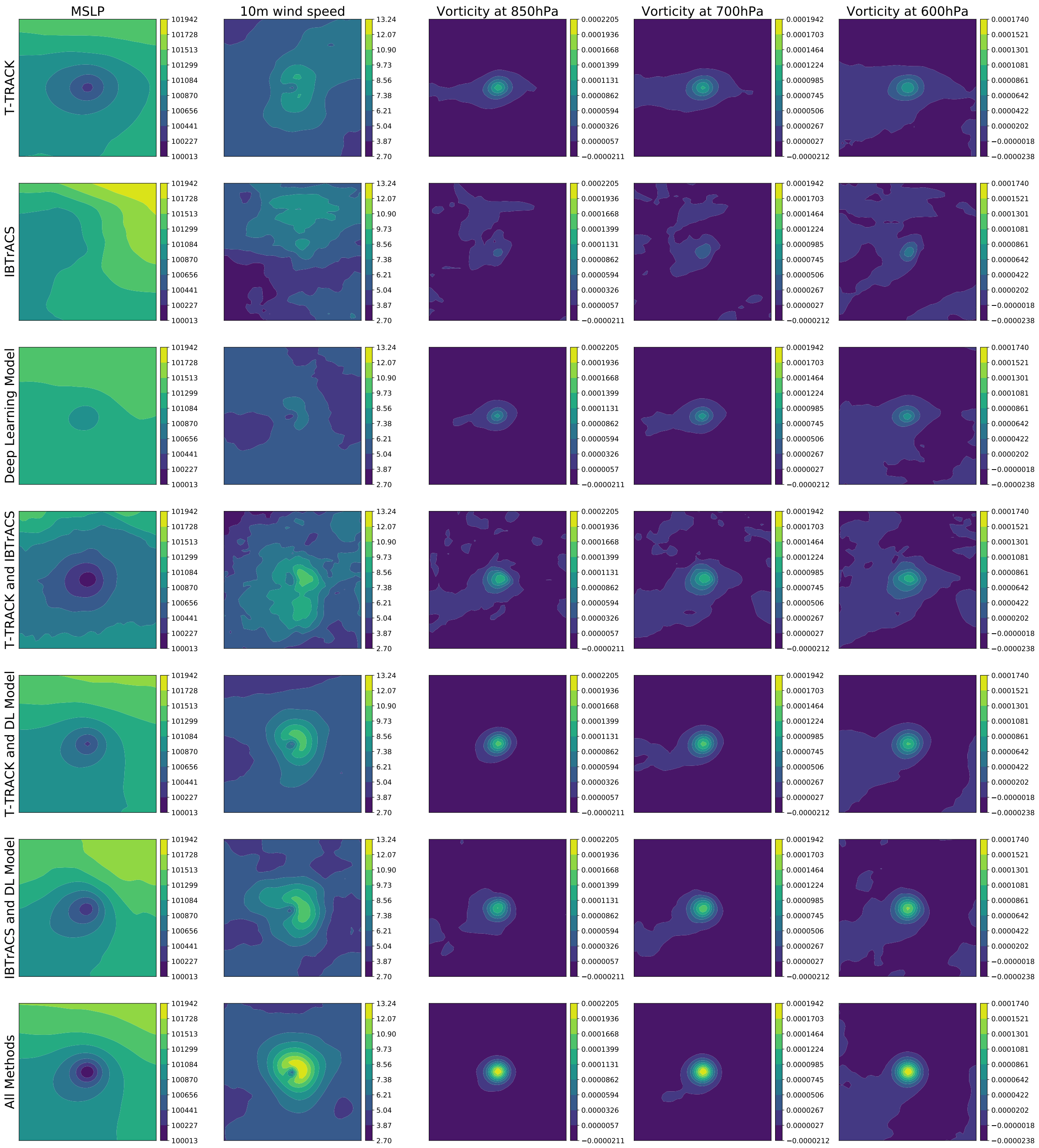

Fig. 8. Composite view of Northern Hemisphere events by detection algorithm or observations which pick up the TC. Total number of cases used to produce each composite can be obtained from 1. Columns correspond to the variables used: MSLP (first column), 10-metre wind speed (second column), vorticity at 850hPa (third column), vorticity at 700hPa (fourth column) and vorticity at 600hPa (fifth column).

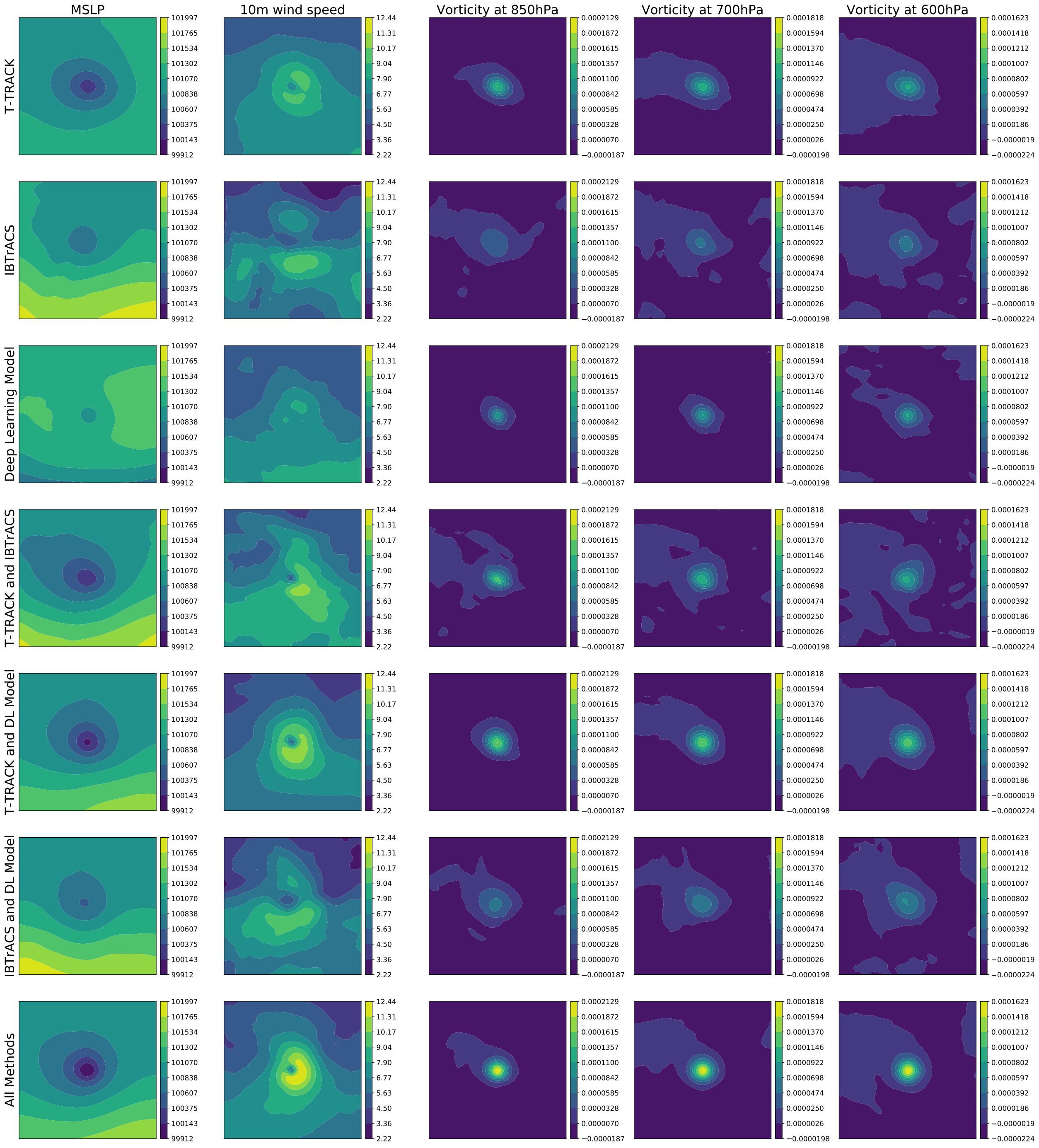

FIG. 9. Composite view of the Southern Hemisphere cases (rows and columns as described in Figure 8) - but the sign of vorticity has been reversed for ease of comparison).

detected by T-TRACK and IBTrACS but not the deep learning model but this is somewhat expected as relatively few TCs are present only in IBTrACS when compared to 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 centres are wider and those at 850hPa and 700hPa 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, 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.

Finally, 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 1008hPa. The maximum wind speed is also down to 8.4 m s^{-1}, and does not show a clear eye at the centre of the composite. The vorticity fields show wider but much shallower centres, with the maximum vorticity around half an order of magnitude 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 centres, but this is somewhat expected as relatively few TCs are detected by IBTrACS only when compared to the other composites.

When split up by hemisphere, these composite cases show some differences. First, the MSLP field in the composite for the Northern Hemisphere cases shows a wave structure rather than a well-defined low. The wind sped field also shows a lack of a centre. The vorticity fields do show clear centers but have considerable noise present.

The composite for cases originating in the Southern Hemisphere shows a much more organised situation. A clear, but wide, low pressure centre is noted, as well as a centre in the wind speed field. The vorticity fields also have well-defined but not concentric centres but there is also a considerable amount of noise present on the outskirts of the centres.

When examining the composite case for TCs detected by the deep learning model only, a concentric centre is observed in the MSLP field with a minimum MSLP of around 1009hPa. A clear centre is also seen in the wind speed and vorticity fields as well. 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.

This situation does not change much when the composite is split by hemisphere. The one difference is that the composite for the Southern Hemisphere shows a relatively shallow area of low pressure in the MSLP field when compared to the composite for TCs detected by all three methods.

Finally, 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 voriticty in the former are about 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 organised. Finally, those TCs detected by only one detection method are even weaker, with the most noticeable decrease in strength in the vorticity fields.

d. Strength

With the results from the analysis of life-cycles and composites in mind, the obvious question that arises it what extent TCDetect is doing a good job of finding "cyclone events" as opposed to "tropical cyclone events"?

The match between any class of depression is shown in Figure 10 which extends Figure 1a 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 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 of 1485). 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

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

TABLE 2. Recall and precision of meteorological disturbances seen in ERA-Interim labelled by IBTrACS as recorded by T-TRACK and TCDetect (compare with figure 10 which shows TCDetect recall for TCs).

not as high as for all storms as it is for TCs (compare tables 1 and 2) the precision with respect to recovering storms is good (table 2).

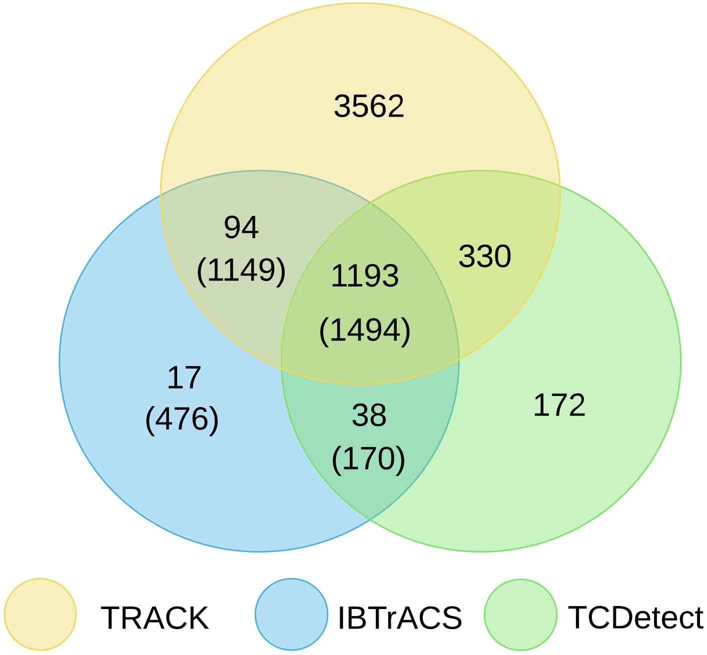

FIG. 10. Events detected by TRACK and TCDetect and reported by IBTrACS. All meteorological systems are included from IBTrACS and TRACK, not just category 1 and higher systems. Events present in IBTrACS (blue area) were split into TCs of hurricane status (non-bracketed; defined as true positives for TCDetect) and other depressions (bracketed values; defined as false positives for TCDetect and T-TRACK).

The details of the matches between TCDetect and IBTrACS is shown in Table 3. The overall precision results from the fact that only 506 out of 3397 cases were cases that had no meteorological system present and the vast majority of cases with no TCs are being classified as such, i.e. true negatives. Some cases with TCs present are being misclassified (false negatives) with a greater portion of lower category TCs are being misclassified than higher category TCs, consistent with

	+ve	-ve
No meteorological system	506	19253
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

TABLE 3. Inferences generated by TCDetect, split by storm type (rows) reported by IBTrACS. Positive inferences are where TCDetect detected the presence of a TC, negative inferences where TCDetect detected no TC. For example, of the 19759 cases which 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).

the results in Galea et al. (2022b). This is all consistent with the deep learning model recognising the pattern but struggling to distinguish between strong (deep) and weak (shallow) systems.

5. Summary

In this study, two automatic detection methods for TCs, namely T-TRACK and TCDetect (a deep learning based algorithm) were applied to ERA-Interim data, and compared to an observational dataset for TCs, the International Best Track Archive for Climate Stewardship (IBTrACS) database. 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 of the detected events, and a comparison of the structure of those 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 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. These differences can be explained by noting that when matching detected cyclones with observed cyclones little difference is observed, that T-TRACK is finding more cyclones than were observed, and the TCDetect locations are generally distributed further from the actual locations of events.

While both TCDetect and T-TRACK found more (presumably real) 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 disorganised fields, and those detected by only one of the methods were the weakest storms present and had considerable noise in their fields.

Finally, it was found that most of the false positives (TC reported, but not present in IBTrACS) generated by TCDetect were associated with some sort of storm, albeit without hurricane status. In fact, the overall precision of TCDetect in terms of recovering events (as opposed to TCs) 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. (2022b), who showed that recall in TCDetect is related to storm strength, however both techniques have similar results in terms of recall of actual TCs.

In a companion paper in preparation, TCDetect is used in a General Circulation Model (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.

Acknowledgments. DG was funded by Natural Environment Research Council (NERC) as part of the UK Government Department for Business, Energy and Industrial Strategy (BEIS) National Productivity Investment Fund (NPIF), grant number NE/R008868/1. KH and BL were funded by NERC via the national capability funding of the National Center for Atmospheric Science, NCAS.

Data availability statement. The data and code to produce the dataset, deep learning model/s and subsequent results is available to access at Galea et al. (2022a). Also included is the IBTrACS version 4 dataset, obtained from Knapp et al. 2018.

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 A: Dynamic Meteorology and Oceanography*, **59 (4)**, 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. *Q. J. R. Meteorol. Soc.*, **137 (656)**, 553–597, https://doi.org/10.1002/qj.828.

Galea, D., K. Hodges, and B. N. Lawrence, 2022a: Investigating differences between Tropical Cyclone detection systems. Zenodo, URL https://doi.org/10.5281/zenodo.6595117.

Galea, D., J. Kunkel, and B. N. Lawrence, 2022b: Tcdetect: A new method of detecting the presence of tropical cyclones using deep learning. *Artificial Intelligence for the Earth Systems*, submitted.

Hodges, K., A. Cobb, and P. L. Vidale, 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., K. Walsh, and A. Ballinger, 2013: Detection of tropical cyclones using a phenomenon-based cyclone tracking scheme. Tech. rep., US CLIVAR. URL https://usclivar.org/sites/default/files/documents/2014/2013HurricaneReportFinal.v3_0_1.pdf.

Horn, M., and Coauthors, 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.

Knapp, K. R., 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, URL https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.ncdc:C01552.

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. Am. Meteorol. Soc.*, **91 (3)**, 363–376, https://doi.org/10.1175/2009BAMS2755.1.

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. *Journal of the Meteorological Society of Japan. Ser. II*, **85 (3)**, 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 Report*, **349**.

Saha, S., 2014: The ncep climate forecast system version 2. *J. Climate*, **27**, 2185–2208.

Schenkel, B. A., and R. E. Hart, 2012: An examination of tropical cyclone position, intensity, and intensity life cycle within atmospheric reanalysis datasets. *Journal of Climate*, **25 (10)**, 3453 – 3475, https://doi.org/10.1175/2011JCLI4208.1.

Schmidt, G. A., 2014: Configuration and assessment of the giss modele2 contributions to the cmip5 archive. *J. Adv. Model. Earth Syst.*, **6**, 141–184.

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 International Conference on Computer Vision (ICCV)*, 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. *Journal of Climate*, **26 (1)**, 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. *Geoscientific Model Development*, **10 (3)**, 1069–1090, https://doi.org/10.5194/gmd-10-1069-2017.

Uppala, S. M., and Coauthors, 2005: The era-40 re-analysis. *Quarterly Journal of the Royal Meteorological Society*, **131 (612)**, 2961–3012, https://doi.org/https://doi.org/10.1256/qj.04.176.

Walsh, K., 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.

Zarzycki, C. M., and P. A. Ullrich, 2017: Assessing sensitivities in algorithmic detection of tropical cyclones in climate data. *Geophysical Research Letters*, **44 (2)**, 1141–1149, https://doi.org/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. *Journal of Climate*, **22 (24)**, 6653 – 6678, https://doi.org/10.1175/2009JCLI3049.1.