Referee review of Bourdin et al: “Intercomparison of four tropical cyclone detection algorithms on ERA5”

Overall comments

In this article, the authors describe a comparison of tropical cyclones in ERA5, derived using four different tracking methods, against observations. The strengths and weaknesses of the different tracking methods are evaluated which tracks are found or not found compared to observations, and what properties these different tracks have. The authors also devise some post-processing to remove some of the false alarms.

This is a really valuable contribution to the literature on tropical cyclone analysis, since many papers on tropical cyclones use different methods and a thorough comparison of some of these methods is overdue. This is also important given the uncertainty in future projections of tropical cyclones, and having better understanding of strengths and weaknesses of trackers is useful for understanding some of these uncertainties. The paper is very well written, logically organised, and contains inciteful analysis techniques to better understand the tracking methods.

I have a several broader questions that I’d like to see clarified in the text.

For the benefit of potential users of the tracking algorithms I think it would be useful to have some mention of their respective complexities, for example ease of obtaining and use. Having used several of these trackers myself, there is a very wide range of cost and complexity involved (some can simply be downloaded and run, others not), and I think that this information together with the scientific quality of the tracker outputs would be valuable to communicate.

You choose to use six hourly data from ERA5 for the tracking. I’d like to see this justified more robustly (that is, I am not expecting you to retrack using hourly data). I could have imagined that this would have been an excellent opportunity to use the available hourly data from ERA5, hence removing uncertainties that come from using six hourly data (particularly from stitching points of the tracks together, and from missing points within a track), and helping to clarify whether it remains good enough for climate models only to produce six hourly data for tracking, or whether key aspects of the tropical cyclones are missed. Hourly tracks could be just as easily compared to six hourly observations after all.

The North Atlantic seems to be a difficult basin for all the trackers as they all have a low bias. Is there a reason for this that you can suggest? Since many climate models also struggle with TCs over the Atlantic I feel this may be an interesting observation.

Detailed comments

L20: bottleneck impediment – maybe choose one of these words, rather than both

L69: “in a forthcoming paper…” – perhaps you could just say “in future work”, if the manuscript is not available yet.

L128: As in the overall comments, could you add a sentence to justify using six hourly data? Since there is hourly reanalysis data, is this primarily because IBTrACS generally does not have hourly data (although you could match the tracks even if the times are not the same)? Do you think that the tracks would be more accurate using hourly data, certainly the uncertainty from gaps and distances

would be smaller? We found in recent work on Mediterranean cyclones that using hourly data is important, and it would be interesting to hear something about this, given that (by historic default?) most models produce six hourly data.

L202: Dulac et al – again, should you cite if not available?

Figure 1: Can you clarify what the y-axis is? Is it normalised frequency, or percentage?

Figure 3: could you clarify the caption, it is not quite clear.

L306: I think you should have a citation for the statement "…potential poleward shift with climate change".

Figure 8: Maybe clarify – does offset delay mean when the storm finishes (lysis)?

Figure 9: What does OWZ look like in terms of precursors, since you don't show it but I think it is one of its selling points? Given it is mslp-based, UZ cannot really detect over land, but OWZ might be able to, or is TRACK the only one? Perhaps you could add a figure to any supplementary data, I think this would help to distinguish the utility of the trackers or combinations of them.

L479: But Patricola et al. filtered out AEWs and found no change in TCs – how does this compare to your statement?

Figure 12: I really like the Venn diagram, but some of the numbers are difficult to read, could you make them a little clearer?

L511: Are there any characteristics of observed storms that make all four trackers miss (92% POD means 8% miss)?

L544: "The third Venn diagram" – I think you are referring to the fourth one, bottom right.

L550: This paragraph is true for the reanalysis, does it necessarily hold for climate models? Could models' representation of TCs be distorted (since there is no data assimilation) such that different trackers would not necessarily find the stronger TCs, or not?

L581: "Contrasted results …varying resolution of observations" – it is not clear to me what you are trying to say in these several sentences. Perhaps you could rewrite to make clearer? I really don't know what varying resolution of observations means, do you mean how frequent they have been over time, or how densely observed pre-/post- satellite era?

L585: binarity – my dictionary does not have this word. Perhaps binary choice?

L587: "To some extent, trackers' thresholds are arbitrary" – this may be true, but could one also argue that observational classification of TCs is also somewhat arbitrary, satellite images with an interpreted core of the storm or other methods? Perhaps you disagree, but maybe you could qualify the statement a little.

L887: You could mention here that understanding precursors is becoming a key question for future projections (in particular to understand diverging projections of increase or decrease in frequency in future), so having trackers able to identify such features is really important.

Table B1: For TRACK, I think here you are using the version averaging vorticity across 850,700,600 hPa, hence rather than eta_850 I think it should be eta_bar_T63, as you have written on L192.

Figure B2: What is the difference between Fig. B2 and the earlier, similar one, can you clarify?

Looking at the algorithm parameters within StitchNodes, the different trackers have different minimum durations set. But duration (lifetime) of the tracks are compared in the results section. Would it not have made more sense to use the same minimum duration across all trackers?

Technical corrections

Pacific Ocean should be capitalised – in various places in the manuscript.

L272: "extrem" → "extreme"

Figure 6 caption: missing a (for "top panel)".

L409: "..for each tracker.." or "..for each of the trackers.."

L447: "maximum 10 meter wind speed"

L559: "..missing tracks correspond to"

Figure B1: Second to bottom box: "Remove tracks lasting less than one day"

L638: "single circulation at that point"

L657: Ullrich et Zarzycki

L669: several times in this section you have "r_thresholds" when I think you mean r_threshold.

L677: Several references in this section around here have lost their formatting, e.g. Bell et al. 2018 rather than Bell et al. (2018).

L738: Figure D1 shows that…