Review of "Composite Sharpening by Vortex Symmetrization and Normalization of Tropical Cyclones"

Submitted by Caratsch et al. to GMD

In this paper, the authors introduce a new methodology meant to improve the compositing approaches often used in the study of cyclones' structure. Their suggestion is to normalize the cyclones according to their size before averaging their structure. Their method appears as a relevant improvement of the usual centered-only approach, especially in the era of km-scale modelling when models are expected to better resolve cyclones' structures (which, surprisingly, is not mentionned as a motivation?). Their approach is rigorous and the improvement they bring is clearly demonstrated. The subject of the paper is also very well suited for the journal, so I think the study deserves to be published in GMD. However, the manuscript needs important revisions, as detailed below.

My comments are separated in general ones that apply to the whole manuscript, and specific ones that correspond to specific parts of the manuscript, but the former are not necessarily more important than the latter.

General Comments

1. As detailed in the specific comments below, I think some of the structure of the manuscript should be revised to better serve the main purpose of the paper, i.e. to demonstrate the improvement brought by your new method. The part on statistical comparison between the observed climatology and the simulations (3.2) should, in my opinion, be substantially reduced. Moreover, an outline and more intermediary intros/transition paragraph would help better follow through the manuscript by always knowing what is the purpose of the section the reader is going through within the manuscript.
2. The question of model resolution is not explicitly treated in the manuscript, and I think it should:
   - In my opinion, the advent of km-scale models is the main interest of this method: traditionnal 100-km-or-so models could not resolve cyclone structure well enough for the details of the compositing method to matter. But we expect 10-km-or-less models to better resolve the structure of the storms, and this is why having more rigorous methods like yours is important. This opportunity could even be extended to the apparition of more systematic high-resolution observations such as EarthCare. They might provide enough observationnal references that we can apply the same methods to.

- Along with this lacking motivation, there is a confusion/lack of clarity in the paper between the stakes of ~25km models and those of km-scale. You could better highlight the different expectations for the two classes of models.
    - L. 121, Bourdin et al. 2024 used 200-25km resolution simulations, whereas Judt et al. 2021 used km-scale simulation. Which of those does the statement L. 120-121 applies to?
    - L. 264-265: Again, Bourdin et al. 2024 highlighted the challenge of resolving storm intensity for models with resolution 25km or coarser, which is much more coarse than your own data.
- This is also an important distinction because for models with ~25km resolution or coarser, it can be challenging to define the R_17 (your R_out), since wind intensities are largely underestimated. Can you comment on the applicability of your method to coarser data?

3. In the introduction, you mention extra-tropical cyclones (ETCs), but you only apply your method to TCs. Is your method relevant for ETCs as well?
4. Might be out-of-scope, but I would be curious to see what you method yields for precipitations?

Specific comments

1. L. 42: Which "size" are you talking about?
2. L. 47-65: I assume the role of this paragraph is to introduce the "small-scale features of interest". I find it lacks clarity as it is very dense in information and it is not immediately clear where you are going with it. I suggest reframing it, especially since having a clear idea of what features you are interested in is essential.
3. L.75: I suggest providing an outline for your paper at the end of the introduction to let the reader know what to expect.
4. If your simulations start July 1st and you remove the first two weeks, it means you are counting TCs from July 15th only. In 2005, 4 observed tropical storms, including Cindy (Cat. 1) & Dennis (Cat. 4) occured before July 15th. In general, many seasons see at least one storm before July 15th. This will impact in particular in figure 2, the number of TCs, the intensity distribution (early-season storms are usually weaker), the month of genesis, as well as potentially the genesis location. Can you justify your choice of timing? I expect this might have been due to other constraints and/or cannot easily be changed. I therefore recommend you crop the HURDAT data to exclude the storms before July 15th for a fair comparison, and acknowledge this limitation.
5. L.83: I assume you are using the ECMWF IFS HRES **analysis**? If yes, please precise so, as "ECMWF IFS HRES" could mean the forecast. If not, please explain what you are using.
6. Section 2.2: The TC tracking section is not clear enough. I understand you do not want to include an extensive description of the scheme, but even without lengthening it, you could

improve clarity. You also mention a sensitivity test, but I do not understand if it was done in Enz et al. or in your own study. And which thresholds did you keep from this sensitivity study? Also, given some of your analysis is sensitive to when the tracker starts/ends the tracks, can you give some details about how you method compares to HURDAT to this respect?

7. Section 2.3: I think it should be re-structured, as it has three somewhat disconnected parts: The filtering of detected tracks, the introduction of HURDAT as a reference, and the time normalization. I would merge the first with the previous section, isolate the second, and maybe keep the third for the paragraph where the corresponding results are discussed.
8. Section 2.4: Why not use a polar grid for the initial projection, which would make the subsequent computations easier, and why re-project the data onto a cartesian grid at the very end, since the composites are azimuthal averages?
9. L. 159: I do not understand what this sentence means.
10. L. 160: Why is it important to get rid of the eyewall tilt? Is it not a structural feature we may want to keep?
11. L. 167: If R_out is capped to 700km and composites are cropped beyond R_out, why output the data on a 1000×1000km grid in the first place?
12. L. 168: What happends when R_out is found within 30km of RMW?
13. Table 1 could be an opportunity to schematically detail better the different steps of thection 2.4, which is fairly intricate.
14. Figure 1: If you have 10 detection sectors, why does the figure only show 6 RMW and 7 R_out.
15. L. 223-224: You say it is more meaningful to compare wind and pressure at the same level, yet you test for winds at 8 km (upper troposphere) vs surface pressure. Your approach is confusing.
16. Figure 2: I am surprised that the wind speed values do not decrease more with height, when we know that the strongest TC winds are found near the surface. Can you comment on this?
17. Section 3.1 : Does it make sense to compare absolute values of wind at 10m (in HURDAT) and at different model levels? Computing the MSE score this way means you will favor most the level that have average wind values close to the observed 10m winds, rather than the one that has the wind-pressure relationship shape closer to your observation target, isn't it?
18. I don't understand the role of your wind-pressure analysis within the manuscript.
19. Section 3.2 constitutes a very tricky statistical exercise: First, you are working with a small amount of data, and second, more importantly, it is not obvious that it makes sense to compare the 1980-2024 climatology to an ensemble representative of the very extreme 2005 season. As shown by my numerous comments on this section, it raises a lot of questions. I would refrain from making a precise statistical comparison of two ensembles that are not equivalent, especially since I don't think you need it within the manuscript (which is meant to demonstrate the improvement brought by your compositing method, not the ability of the ICON model to

simulate TCs). Assuming this analysis plays the role of a validation of the TCs simulated by the model, I would recommend reducing a lot the complexity of this section to give only essential keypoints with respect to validating the model's TCs: For example, the low number of simulated TCs is obvious as soon as you mention that it is within the 1-7 TCs/season range, compared to an average 7 in the climatology, and 16 in 2005.

20. L. 249-250: I see as many grey as black arrows in fig. 1 for the comparison of the climatology to the simulated ensemble. Why do you say "predominantly insignificant differences"? also, shoudl all metric be considered equals to that respect?
21. L. 251: You cannot compare the seasonnality of the two ensembles since one starts earlier than the other. I will also impact you counts and intensity statistics (see comment above).
22. L. 255-256: The lower termination latitudes could also be due to the tracker having a propensity to stop detecting the tracks earlier than HURDAT would.
23. Figure 3: The counts between panels (b) and (c) do not match. In panel (c) I count 3×2=6 major TCs in observed 2005, but in panel (b) you indicate 5. Also, I was confused for a while before I noticed the mention burried in the caption that squares in panel (c) account for more than 1. Maybe use a different shape to warn the reader these are not the same? Plus, you should precise which categories does "major" correspond to (I assumed 3-5).
24. Do you have an idea why the model produces so few TCs?
25. Figure 4: Showing the x-axis grid lines would help with comparing the distributions across the two panels.
26. Figure 5: Why are the dots not equally spaced in azimuth? I thought you divided the space in 10 sectors of equals size.
27. Figures 5 & 6: The colormap/palette could be better chosen to better identify the small scale features and how they change with the transformation. You don't really need the negative side of the colormap, so you could use a non-divergent colormap with more hues so that the details of the circulation are clearer.
28. L.307: I expected the AI would be minimum at the lifetime maximum intensity, as I thought the structure would be less symmetric in the early stages when the cyclone is still forming, and the late stages when it is likely being disrupted by vertical wind shear or land interaction. Can you discuss this more? Was this result expected based on the literature? If my naive expectation is wrong, why? And if it's reasonnable, why do your result differ?
29. Figure 7: It is difficult to see the faint grey enveloppes. Could you add thinner lines to indicate the quartile values, which would contains the IQR, to make it easier to read?
30. Figure 7: Why do you use less interpolation points in the decay than in the intensification stage, whereas you showed previsouly that both stages have in general the same length?
31. L. 332: Can you specify how many TCs there are in each group? Could you also comment on how sensitive your composites are on the number of cyclones being aggregated?

32. Section 6: It is not clear to me at this stage if you aggregate only snapshots of TCs at their lifetime maximum intensity (most usually done, but smaller sample), or all snapshots at all stages (in which case longer-lived cyclone would be over-represented)?
33. L. 360: If the improvement is greater for weak cyclone, why did you chose to show the intense ones in the main body?
34. The conclusion is very good. The abstract should follows the same lines, and probably not mention the not-robust results with respect to the ability of the ICON model to simulate TCs.