Review of "Objective Identification of high wind features within extratropical cyclones using a probabilistic random forest (RAMEFI). Part 1: Method and illustrative case studies" by L. Eisenstein, B. Schulz, G. A. Qadir, J. G. Pinto and P. Knippertz.

This study aims to objectively identify different mesoscale wind features associated with extratropical cyclones by applying a probabilistic random forest approach. Overall, the topic is very interesting, the method novel, and very promising results are presented. My three main issues are (1) the structure of the paper needs to be improved and streamlined, (2) there are many terms concerning the random forest and its verification that a typical reader of WCD will not be able to understand and (3) there are notable limitations of this method that should be highlighted more clearly. These issues, and more minor points, are detailed below.

Major Comments

1. **The structure of the manuscript could be improved.** I appreciate that this is not a simple manuscript to organise, but I think with some careful re-structuring the manuscript would become shorter and much easier to read.

   (a) Currently there is quite a bit of repetition of the same or very similar topics. One main cause of this is the large amount of overlap between section 2 (theoretical description of the wind features) and section 4.1 (how they were labelled). I think if these could be combined, it would be much easier for a reader. Some specific examples of repetition include: lines 204-210, also the text in lines 196-199 is repeating information that has already been stated, Lines 259 -260.

   (b) It is not clear exactly what threshold / criteria were used to manually label the features. Was a certain set of "rules" created for each variable and each feature or was it done purely subjectively? In lines 315 – 318 it is mentioned some of the characteristics a WJ requires. It is also said that "signs for a CJ as missing" but it is not clear what these signs are. I think my confusing comes from the structure of the manuscript (see point above) with the "theoretical" characteristics described in section 2 but then the characteristics / values of the actual labelled features are loosely described in section 4 and 5. I strongly encourage the authors to reconsider the structure and combine the theoretical basis and what was actually used in the same section.

   (c) Methods are distributed throughout the manuscript breaking up the flow for a reader and making it hard to find the method details if a reader wants to go back and check. Furthermore, some details from the appendices should be included in the main text. Section 4 is headed "Method" but it needs to be better structured and expanded to include extra details / extra subsections of all of the methods used to develop "*the method*".

      1. First, a clear subsection of how the 12 cases were selected needs to be included somewhere as currently it is easy for a reader to miss. For this, some basic details of the SSI should be included in the main text, not in the appendix.

      2. All of section 6.1 is describing a method. It interrupts the flow of the results. Could this be moved to the "methods" section and revised to make it more accessible for a typical WCD reader (see below)?

3. Details of why the threshold of 80% of the 98th percentile is used should be more clearly stated. Currently, this is first stated on line 201 but it should be earlier. Furthermore, this should not be under a subheading "subjective labelling using an interactive tool" as it is really part of the method.

(d) In a few places in the text, future sections of the manuscript are referred to which is somewhat distracting for a reader but may also indicate that topics are not in the most optimal order (e.g. line 123, line 212, line 262, line 290, line 325, line 350, line 382, line 399).

2. **Limitations of the method should be clearly highlighted and considered more.**

(a) The requirement of having either observations or model data at 1 hour temporal resolution is a limitation of this study and potentially will limit how this method / software can be applied in the future e.g. data from CMIP models is generally not available at 1 hour time resolution and nor are observations from other parts of the world. This limitation should be stressed more clearly – it is briefly noted in section 3, where it is stated that few stations report precipitation with 1 hour resolution. Furthermore, the justification for using changes over 1 hour should be motivated and justified in the methods / data section.

(b) In Line 416 it is stated "The RF should only be applied to regions affected by extra-tropical cyclone". Given the text that comes before this, this seems a little inaccurate. Is it not more accurate to state that the RF should only be applied in the region where the training was performed (over Europe and over land)? This also raises concerns about how valid this method would be in other areas of world. A specific example is that extra-tropical cyclones certainly affect mountainous regions (e.g. Norway, Iceland) so is it likely that the RF method will not work in those area which certainly are commonly affected by extra-tropical cyclones? At a minimum some text should be added to considered how accurate the method will work over ocean and in other parts of the world e.g. at the start of the storm track, in mountainous areas.

(c) Subjective choices. Although the method is objective, the training and some threshold are subjective. It should be considered how sensitive the results are to the subjective training. It should also be considered how the results depend on the threshold used for "strong wind features" e.g. 80% of the 98th percentile? For examples, is it a limitation of this method that the same threshold is used for all wind features even though some are well known to be stronger than others? For example, a warm jet is likely to have weaker winds than a cold jet so does this mean fewer warm jet features are identified, or that more extreme WJs are identified that CJs.

(d) Are computation costs another limitation that should be more clearly noted. On line 231: "For computational reasons..." This needs to be more clearly explained. Does this mean that the method cannot work on a very large data set? It would be nice if some indication of the computing cost of this probabilistic method could be included in the manuscript.

3. **Technical terms are hard for a meteorologist / typical WCD reader to understand.** I think overall it is great that new methods are being used in the field of synoptic-dynamic

meteorology but not all readers of WCD will be familiar with these methods and the terminology that goes along with them (myself included). For these reason, special care is needs to ensure terms are clearly explained and also put into context of the data / synoptic terms used. It would also be clearer to explain some of these terms (particularly those about the evaluation of the forecasts) in the methods section rather than in the middle of the results (see # 1 above).

(a) Section 4.2 / machine learning terms. As a meteorologist I found this section very hard to understand and I am probably quite a typical reader of WCD. There were quite a few terms in this section that I do not understand (order criteria, predictor space, target variable). I feel that for this journal these terms need to be better described. Potentially some schematic diagram describing the probabilistic random forest could be included to make it easier for a reader to understand.

(b) Section 6.1 / terminology. Again, in this section there were quite a few terms that I did not fully understand. For example, "multi-class probability forecast", "binary predictions", "COPR-approach" (the abbreviation should be explained). For a reader it would help if these terms could be out into content with the case here. e.g. I think for "multi-class" class = the different wind features.

(c) Section 6.2 / terminology. More examples of terms that are not easy for a meteorologist to understand: "For each fold of the cross-validation scheme...", "binary probabilistic forecasts"

Minor Comments

1. The title is very long. I don't think it is necessary to have the abbreviation "RAMEFI" in the title. "illustrative" could also be removed.

2. The synoptic meteorology terminology that is used is not clearly explained. Specifically, it is not clear whether the "short warm jet" is exactly the same as a warm conveyor belt (WCB), and whether the "short cold jet" is the same as a cold conveyor belt (CCB). How these terms relate to the commonly used terms (WCB, CCB) should be clearly defined in the introduction. Potentially also considered revising the subheadings of 2.1 and 2.2 as it is unclear why "conveyor belt" is in parenthesis.

3. Introduction – first 2 paragraphs. Here the Norwegian cyclone model and the Shapiro Keyser cyclone model are introduced. It would be beneficial to include a more in-depth discussion of how all of the wind features typically different between the two types of cyclone models – currently the only reason to introduce these different cyclone models appears to be to state that sting jets only occur in Shapiro-Keyser cyclones. This could be included in section 2 (or a revised combined section 2, section 4.1) rather than in the introduction.

4. Related to minor comment #3 - Lines 33 – 37 and Figure 1. This could be incorporated in section 2. Or at least Figure 1 should be referred to much more from the text in section 2 as currently Figure 1 is not referred to / utilised to its full extent.

5. Line 42. Can some additional details about the type of post-processing Pantillon et al (2021) used be added here?

6. After reading the introduction, I was wondering if any other studies had applied the probabilistic random forest method to meteorological data sets as its not clear from the introduction. If previous studies have used this method, it should be stated.

7. The discussion of the warm conveyor belt in section 2.1 does not consider the cyclonic and anticyclonic branches which are shown in Figure 1. These features should be briefly noted since they are shown in Figure 1.

8. Line 92. I disagree with the statement that there is little precipitation in the warm sector. The warm conveyor belt is often the cause of a significant proportion of precipitation in extra-tropical cyclones.

9. Line 108. Something is missing in the text here "...stays close to the ground 850 hPa..".

10. Section 2.3 / Sting jets. Please give some indication of the spatial and temporal scale of a sting jet.

11. Line 154. The computation of the potential temperature is unclear as is the reason why potential temperature is used over temperature. If mean sea level pressure (p) is used, and a reference pressure of 1000-hPa is used to compute the potential temperature, then the effect of station altitude is not removed – which was one reason I assumed potential temperature was used over temperature. Can both the method to compute potential temperature and the reason for doing so be more clearly explained in the manuscript as I think I have misunderstood something here.

12. Line 165. The justification for normalising the wind speed by its 98th percentile is not clear here – is it to remove seasonal / diurnal cycle or to focus on high wind events? I think it is the former and that the 0.8 threshold on v/v98 mentioned later is to ensure only high wind events are considered.

13. Line 170. Can the Euro Cordex domain be shown on a map or at least the longitudes / latitudes it covers added to the text?

14. Section 3.2. Can the vertical levels of the variables taken from COSMO-REA6 be added to the text?

15. Lines 290. Is this text really needed?

16. Line 293. (int.) - this is not clear. Maybe it would be better to write also known elsewhere as Eleanor?

17. Line 322. Suggest to add "training labels" and "predictions" in parenthesis after the reference to Figs 3d and 4b.

18. Line 338. I don't understand what is meant by "the edges of the boundary or cyclone". Please revise and clarify this sentence.

19. Lines 342, 344. The terminology is a little inconsistent here and thus potentially confusing. In line 342, the comparison of station observations to gridded forecasts are discussed whereas in line 344 it is the comparison between station observations and reanalysis-based results. If this is the same thing, please be consistent with the terminology. (forecasts made me think more of operational weather forecasts than the forecasts from RAMEFI)

20. Section 5.3 / comparison of Figure 5 to Figure 4b. Here the biggest difference is that in figure 5 the different wind features / different colours join up with each other (not many white grid points) whereas in Figure 4b there are gaps. Why is this?

21. Line 373. What is the "climatological benchmark"? If this is some average of the 12 cases, then I think using the term "climatological" is misleading.

22. Line 430. Again there is a term used that I do not understand – please explain what the "BS permutation importance" is.

Figures and Tables

1. Figure 1: The long cold front shown in this schematic is not typical of what a Shapiro-Keyser cyclone looks like. Could this front be shortened? Would figure 1 also have a panel b showing a Norwegain cyclone model type of cyclone as well?

2. Table 1. Can units be added to this table?

3. Figure 2. How are the cyclone tracks shown in this figure produced? e.g. from an objective tracking algorithm or subjectively? Also Eberherd does not seem to have any markers indicating MSLP or are they just too small to see?

4. Figure 3d. Are the labels only for stations where $v > 0.8$? This does not seem to be consistent with the contours e.g. there are labelled features outside of the dotted contour.

5. Figure 3 caption "set labels" in the caption is not a very clear term. Is "training labels" clearer?

6. Figure 4. Could the mean sea level pressure field be added to these panels to give a reader some synoptic context for where the different wind features are predicted and to get a rough idea of what the different cyclones looked like.

7. Figure 6: What does CEP stand for? Please clarify what "via consistency bands (under the assumption of calibration" means. These details should be in the text rather than just in the caption. Also for figure 6, are all 12 storms included? This information should be added to the caption.

8. Figure 7. Please define CEP.

9. Figures 8, 9 and 10. the delta symbol does not seem to be displaced correctly.

10. Figure 10. Please add the units for each of the predictor variables.