General feedback

This is a very interesting study that I find to be complete, focused and clearly written. It brings many clear and novel results. The methods are explained well, the chosen figures are clear and well-chosen and there is a very good discussion about implications and limitations of the work. Still, I have taken my time to formulate several general and specific comments that I believe can improve the manuscript.

General comments

1. As you have the wind speed timeseries for a 10-year period available, I believe it would be a valuable addition to, at least briefly, discuss how the models compare in terms of the full wind speed distribution, not just the extremes. It would be a very interesting finding if the overall distributions (means, variability, histogram overlap) match well, as you find strong inter-model differences for the extremes. In other words, good performance on the overall distribution might not imply that you can just use any single model for the extremes.

2. There is still quite a lot of variability in the U50 estimates within individual spatial categories. Of course, the sites within a category might still be different in several ways, which can explain this. However, could it also be that over the 10-year period, some of these sites encountered a much more severe convective storm than other sites? Or more generally, that the intra-category, point-specific sets of ordinary events differ due to different sampling of meteorological events in these 10 years? I can imagine that this would lead to other U50 estimates. This would have several implications for wind energy: Using the site-specific U50 estimate, or even the ensemble mean U50, might not be a “safe” choice but one might be better off considering the grid cell of that category that was most “unlucky” during this 10-year period due to stochasticity. If it is correct that meteorological stochasticity over the 10 years plays an important role in the spread, please explain this to reader, e.g. by integrating into section 5.4 and 5.5. If, on the other hand, you believe that the spread is mainly due to varying site surface conditions, I would recommend to add a small section in the discussion that it is crucial to perform a site-specific assessment for extreme winds when a new project is developed somewhere. Perhaps both of these perspectives deserve a space?

3. The SMEV analysis you present is very interesting. However, I think it would be valuable to the work to add some sensitivity testing to illustrate the robustness of the method. For example, how sensitive is the method to the chosen parent distribution (I believe there are other distributions used to model extreme wind speeds)? Or how sensitive is it to the threshold chosen for the left-censoring? Perhaps you can also elaborate somewhere how your bootstrapping supports the robustness of this method.

Specific comments

Line 8-9: after reading a bit further, I realized that this sentence is quite confusing for several reasons. You summarize your complete methodological package into one short sentence, which I think needs elaboration. I would rework this to make things more clear. Three reasons why it is unclear:

- The term 'surface-based spatial categorisation' is vague when no additional information is added. "Surface" on itself can refer to many things (e.g. also surface winds) and only later you explain that you classify into terrain, roughness and climate zones, I think you can be a bit more elaborate and specific already in the abstract.
- The way the sentence is written now, the reader might think you used PCA for the spatial categorisation.
- You mention PCA, a very widely used and general method, but do not explain here what you use it for in your work, so also here I advise to be a bit more specific and elaborate.

Line 26-27: This sentence is unclear to me. It could be because regional variability is mentioned twice, but in different ways, so I think the sentence can be more concise.

Line 31: A few comments

- Please distinguish between GCM and RCM resolutions, as GCMs have a nominal horizontal grid increment of >=100 km, not 10 km.
- Using a term like "horizontal grid spacing" might also be a better idea than resolution, because the effective resolution is many times lower than the grid spacing. You could also clarify with a sentence that you are referring to grid spacing, not effective resolution.
- Please be more specific with what you mean with "general" wind patterns. GCMs only capture synoptic and larger scales. RCMs (dx > 10 km)also capture some mesoscales. Perhaps it's better to describe which scales both of them do not capture, that the CORDEX FPS runs do capture.
- Also, I recommend changing "cannot capture convective motions" to "cannot explicitly resolve…"

Line 55-64: This entire section is about methods you will not use eventually. Therefore, I would suggest to shorten it a bit and also to merge it with the section where you introduce SMEV, the method that you will be using. More generally, I would check where else in the introduction you could make the sequence more logical, because I feel like you switch back and forth between surface characteristics, inter-model comparison and EV estimation quite a bit, which is a bit difficult for the reader.

Line 68-69: This sentence is a bit confusing for me, as I don't see how the parts are connected. With addressing the limitations of observations with an inter-model comparison, do you mean that models provide spatial fields that observations cannot and therefore capture the spatial variability of signals? That would be valid, but I don't understand why the part of performing a PCA-based inter-model comparison is merged into that sentence. Perhaps elaborate and/or split it up into more simple chunks.

Line 91-96: After reading the manuscript I went back to these objectives. I think that it would help to be a bit more explicit that objective 1 is a full-domain analysis based on spatial maps. And that objective 3 and 4 build on objective 2 to perform analyses for the different categories, because the link is not very clear the way it is written now.

Line 113-116: Model physics (e.g. convection scheme) and stochastic variability are indeed an important part of why the CPMs might differ, but I think that the following elements can also be important:

- Model numerics: e.g. the number of vertical levels used and/or the amount of numerical diffusion applied, ... Could you add something about it to the text?
- Simulation protocol: continuous vs. chunked simulations, temporal spin-up. Perhaps the models had the same settings here? Could you add some notes on this?
- Also, next to the convection scheme, also the parametrization of the planetary boundary layer (PBL) is important in shaping low-level wind variability – the CPMs might be using different ones. Please mention this too. It would be helpful to add into the table the convection & PBL schemes used by the different models. Or at least to give a general idea of how they differ between models?

Fig 2: If the 25th percentile of the relative frequency of the categories is used, How can 35 out of 51 categories have relative frequency below this value? That does not agree with the definition of percentiles.. Please clarify in the text so that the reader can better understand this.

Line 208: I think "temporal variability" should be changed to "temporal co-variability", because you do not perform the PCA on the individual time series.

Line 210: What do you mean with "fluctuate"? If you mean "why models differ", then I would not use the word fluctuate. If you mean some form of fluctuation, please be more specific about what you mean.

Line 224: "under identical atmospheric conditions" – I don't think this is correct because the model itself creates the atmospheric conditions in the domain. What I think you can say is "under identical atmospheric boundary conditions".

Section 3.2.2: I agree that this seasonal correlation analysis is a nice addition to the PCA. Yet, you perform this on the absolute monthly maximum – could it not be informative to also perform this analysis on, for example, the 99th percentile? I expect that the P99 wind speed value is also very relevant for wind energy planning and will be less "noisy" than the monthly maximum, sort of a "less anomalous, but still very extreme" wind speed. You could add this, or you could add a good argumentation about why the analysis of both is equivalent.

Line 268: "Ordinary events" you say this is the entire set of independent events, but it is not clear what these independent events means at this point. In section 3.3.1 it becomes clear that these ordinary events are extracted local maxima separated with your correlation window. Therefore, I would be more careful with introducing the term already here because it can be confusing to the reader. If you do introduce it here, please explain what it means.

Line 271: "instead of assuming this is infinitely high" – why would an algorithm assume that the frequency of occurrence of events is infinitely high? Can you please explain this part better please?

Line 272: When you say that it accounts for "intensity" do you mean that it takes into account the intensity distribution? Because "intensity" seems to refer to one specific intensity value. Please clarify this a bit more in the text.

Section 3.3.2: This comment links back to my last general comment. Normally, Weibull distributions are used to fit an entire wind speed distribution, but you already filter for local maxima and then fit a Weibull distribution. Even for full wind speed timeseries, Weibull distributions don't always work well for complex sites and I am also surprised that here it is applied to a maxima-filtered timeseries, with the "left-censoring" added to it. I think it is necessary to convince the reader (visually) that this fitting works. Could you either add a few supplementary plots or refer to some other work where this is clearly demonstrated?

Line 315: Can you be more explicit what you bootstrapped? Is it the ordinary event dataset for each model? Is it the different points in each category? This is currently not very clear from this small section. Please clarify in the text.

Line 316: "for each disaggregated layer" – I think this is the first time you use the term "layer" and it is unclear what you are referring to. Could you please clarify in the text?

Figure 6: Can you add a physical interpretation in the text on why the lowest roughness class, i.e. open water areas, is characterized by the lowest U50 wind speed estimates? Because low roughness corresponds to less frictional deceleration so in this sense you would expect higher extreme wind speeds. Are there perhaps less summer convective storms over water? (I don't know immediately myself).

Conclusion: I think it would be good to – very briefly – repeat some of the suggestions/limitations you mentioned here again. For example, the need for CPM evaluation for observations, measurement campaigns and the need to involve more CPMs. Also, I would recommend to use the past tense for summarizing the stuff that you did. For example "We first investigate…"➔ "First, we investigated…". General truths or findings can remain in present tense of course (e.g. "This highlights the need for ensemble approaches..."). I personally think this makes more sense, but you can also leave it like this if you don't agree.