Review of 'Methods for high-accuracy wind resource assessment to support distributed wind-turbine siting' by Menear et al.

This is a very thorough piece of work that seeks to unify a set of publicly available wind resource products into a single product with improved performance relative to any model by itself. I think this is an important study for end users who may need to deal with conflicting results based on different data products. I do think there are a few aspects of the wind climate that could be addressed in more detail in this study. Moreover, the paper seems unnecessarily long, with similar information reproduced in multiple figures (see specific comments).

Major comments:

1. I am very interested in the statement in line 228 that "a quantile summary over all available years is essentially as informative as a summary restricted to any subset of years". This is problematic for a number of reasons. Firstly, the pdf of interannual wind speed range (inset in figure 4) shows a peak in the distribution at 0.5 m/s, and many sites are to the right of this with larger differences. Claiming that this is not important is in direct contradiction to lines 22-24, which state that "biases of order 0.5-1.0 m/s … can translate into substantial changes in annual energy production". Moreover, it is not clear how many years are actually used at each site: Line 159 states that "only years with at least 95% completeness are retained". Is there any threshold for how many years should be retained? Would a station with only a single valid year be included?

2. Following on from point 1, there have been a small number of studies looking at the interannual variability in wind speed in the US. For example, Hamlington et al. 2014 found up to 20% variabilty in the wind resource with ENSO, particularly in the western part of the United States. This suggests that a representative number of years that includes all three phases of ENSO should be included in the analysis. I think this is an important point because it can provide concrete guidance on how many years of wind data the industry should use for wind resource assessment. I'm not sure it is correct to be claiming that the number years does not really matter.

(Hamlington, B. D., P. E. Hamlington,S. G. Collins, S. R. Alexander, and K.-Y. Kim (2015), Effects of climate oscillations on wind resource variability in the United States, Geophys. Res. Lett., 42, 145–152,doi:10.1002/2014GL062370.)

3. To address these concerns, as a minimum it would be useful to see:
(a) A map of the interannual (max - min) variability plotted on a map, to see whether the spatial variability matches with that of Hamlington et al. (2014).
(b) A sensitivity experiment where only stations that have above a threshold number of years
(c) A sensitivity experiment where all stations are restricted to a small number of years, versus including all years

4. It looks like table 2 and figure 6 are essentially showing the same information. Moreover, many of the numbers in the table are also repeated in the text (lines 408-430). I suggest this material can be condensed - for example, show the median numbers on the box plots, and get rid of table 2, and then just pull out the key points briefly in the text. This same data is then repeated in a spatial sense in figure 8, and as scatter plots in figure 9. There are many different ways that this data could be combined, but one idea would be to retain figure 9, but instead of colour coding the dots by bias (which is shown in the scatter plot anyway, colour code the points by height. This would remove the need for figure 7, and also be quite interesting. Another idea would be a heat map of bias, with wind speed on the horizontal axis and height on the vertical axis. This might actually bring out some additional structure in the data.

5. The authors mention several times that the mean bias across all stations in the new dataset is reduced to zero. Is it not the case that a ML model will always reduce the bias to virtually zero, since this is what it is optimising for? Therefore, it is less important that the mean bias is reduced to zero, and more important that the spread about this mean is reduced. I believe that if you just trained the ML model on one input dataset (eg ERA5 only), then it would also reduce the mean bias to zero. Did you do this test? For example, in figure 10, could you add the model error if you trained it only on WTK or only on WTK-LED CONUS. This would add weight to your argument that it is the combination of the multiple datasets that leads to the model improvement, not just that you're essentially bias correcting a dataset.

6. Can the authors comment on the representivity of the datasets, versus actual model errors? For example, is ERA5 really providing a poor result, or is it providing a result that is reasonable but at a larger scale? Is it reasonable to compare ERA5 to an observations that have been influenced by small-scale obstacles and changes in surface roughness?

7. Can the authors include uncertainty measurements for each station in the model output? This seems like it should be within reach, given that multiple inputs go into each station, and also that the ML algorithm presumably gives some metric of uncertainty.

8. Please clarify the relationship between figures 10 and 11, and associated text. Is figure 10 the result of the "combinatorial sweeps", and then after that a separate experiment was run where all datasets were included, but then withheld one by one? Are both these figures really necessary, and why are the results presented in different ways (model error in figure 10, and box plots in figure 11)?

Minor commments

1. Abstract: "National gridded set of wind speed quantiles". It is nowhere stated in the title or abstract that this study is about the US, so it doesn't really work to say 'national' here.

2. Line 33: "Intermittency" -> I don't think this study looks at intermittency either, given that it just uses quantiles.

3. Line 43: "at a single dataset" -> I think the main innovation here is learning the quantiles, rather than the mean. This approach could actually be used on a single dataset.

4. Line 104: and surface layer parametrisation

5. Line 135: energy-producing or hazard-prone

6. Line 155: Does this mean that within the (0,2kt] range, every wind speed is treated as equally likely? Does that lead to discontinuities in the distribution?

7. Line 161: "large set of consistent long-term 10m wind records". What is the actual number? It was mentioned earlier that there were 1842 stations. Was that before or after the selection for having at least one year with 95% completeness? If there only needs to be at least one year with 95% completeness, then some of these are not actually 'long-term'.

8. Line 166: "306 distinct sites and 368 unique height-location combinations". Is the difference in these numbers because some sites have more than one height?

9. Line 170: "whereas fewer sites sample only near-surface levels" -> can this be more precise?

10. Line 240: How is the interpolation handled if the site s on the coast, with some neighbouring grid points over the sea?

11. Line 345: After this subsampling, how many stations are included? Is this just one once, so that the 1842 number of stations quoted earlier is actually much less? Or is it done multiple times with a different random sample?

12. Line 395. Should Optuna have a citation?

13. General comment about figures and tables: Could you keep the ordering of the datasets consistent, for ease of reading?

14. Line 415: “Perform slightly worse” - this would be much stronger if there was a significance test to see whether the difference in the distributions of errors between the datasets is significant.

15. Line 456-462. The weak winds being too windy, and the strong winds not being windy enough in the models hints at a stability effect. Did the authors consider using any more physically determined predictors from the models, for example wind shear?

16. Figure 10: What do the colours mean?

17: Equation 23: What does this mean? The left and right hand sides of this equation are the same.

18. Line 572: Wouldn’t the ensemble have corrected both positive and negative bias across sites even if there was only one input dataset?

19. Section 6. I got very confused here. How does what is being described here relate to figures 18 and 19. How were figures 18 and 19 made, if not by applying the algorithm to every grid point?

20. Figures 4-5: Please consider using the same vertical axis scale for the three plots in figures 4 and 5, and the same vertical axis scale on the three insets. The insets lack an axis label on their vertical axes.

21. The height-error plots in figure 7 are fascinating. It looks like there is a lot of data hidden in the 10m, 40m and 80m heights, due to many points being on top of each other. Consider contouring this with density of points. I suspect there might be slight upward trend in these plots. Would it be possible to show a linear fit (or perhaps log-linear) and note the correlation and significance?