General feedback

I find this work on the validation of wind farm parametrizations with SCADA an excellent contribution to the mesoscale wind farm modelling literature. The chosen analyses are designed well, provide many novel insights and are presented clearly. Particularly the analyses for individual turbines and filtered conditions reveal the strengths and weaknesses of the parametrizations well and go beyond typical farm-integrated analyses. For this, the complexities of comparing gridded simulations and real wind farm data were carefully taken into account and explained well. Although the analysis of blockage and external wakes is rather limited, I recognize that it is not the main focus of this paper and I don't believe that additional analyses are necessary as these topics require more targeted studies (which already exist to some extent). Furthermore, the paper reads smoothly, has a logical flow and the figures are clear overall. In general, the interpretation of the results is also complete and to-the-point. Therefore, I recommend minor revisions with one bigger comment and a bunch of smaller comments of which I do believe they can still improve the paper's clarity and completeness.

Big comments

1. It would be valuable to synthesize your findings into slightly more explicit recommendations/guidelines for the future use of mesoscale models equipped with wind farm parametrizations. The initial motivation of the paper presented in the introduction is to employ them for wind resource assessment. What do your findings imply with respect to this application? For example, in what use cases or in what aspects is this approach a contender for the use of fast wake models based on the accuracies reported here? And what are the most important choices/uncertainties to consider when using this approach? Furthermore, are there other applications next to wind resource assessment for which this WRF-WFP setup would be an excellent tool? For example, I could see this being a good validation basis to assess power production changes under future buildout scenarios. It would be valuable to add some discussion on this. Potential limitations of this mesoscale WFP approach also connect to my last comment's last bullet point about short-timescale phenomena.

Specific comments

1. Line 10: What do you mean with "too shallow" - for me this is ambiguous: is it too weak? or is the vertical wake extent too small? Perhaps choose another adjective.

2. Line 13-14: I think it would be good to make more explicit already here in the abstract that with "narrow" wind direction sectors you really mean that the mesoscale model-WFP setup cannot capture either aligned and staggered configurations. Because without clarifying, the reader might not really know what level of "narrow" you are talking about (5°? 30°?).

3. Line 14: WPF => WFP

4. Line 21: atlases => atlas

5. Line 63: "several adaptation of this models" => correct this

6. Line 94: could you provide a reference or a small analysis that confirms that this WRF setup can properly capture PBL stability profiles? This aspect could also influence wake recovery patterns and power estimates analyzed later.

7. Line 119: What power are we looking at? Is it the electrical power (which can be subject to losses)? Please clarify in the manuscript. Also, could you please clarify in the text if this matches with the type of Cp values used in the WFP?

8. Line 121-127: I believe this section would work better in a table as it is quite tedious to read through.

9. Line 149: "Discontinuities were detected using a threshold condition of a peak difference of at least 2◦ per day." - This statement is unclear to me, could you elaborate/rephrase this part a bit please? What do you mean with discontinuities? What do you mean with peak difference of 2° per day? Does this mean only turbines with a deviation larger than 2° from the farm-averaged wind direction are corrected for that week?

10. Line 150: It is not fully clear to me how this is done. I suppose the FOXES runs are based on a grid-based wind direction. So do you correct the measurements to grid north first to derive the wind direction bias based on this procedure? And could you please specify explicitly whether this matching is done individually per turbine? It would also be helpful to get an idea of the typical correction magnitudes encountered.

11. Line 154-155: Does this 98% of rated power filtering not risk of flagging "normal" operation cases with positive TI anomalies or negative air density anomalies when the wind speed is only slightly above rated? Both of these could lead to lower power than expected based on the "standard" power curve. Similarly, power production can be higher than "normal" when there is a positive air density anomaly. If these effects are also not in the WFP, I can understand that you filter these out, but some more clarification about this would be good.

12. Figure 6a: could you work with linestyles so all the lines become visible? You could also simplify by showing only one of the TNW lidars and mention they coincide.

13. Figure 7: I think it would be good to repeat/emphasize in the caption that the points are plotted for undisturbed turbines ánd undisturbed wind speeds coming from a separate simulation. I believe this will speed up interpretation by the reader.

14. Figure 7, 11, 12: The slightly thinner lines for 0.67 km are hard to distinguish from the 2 km runs and might complicate interpretation. I would recommend to use different linestyles depending on the resolution. In figure 11, this conflicts with the use of different linestyles for the aligned and staggered plots which would then need to be solved.

15. Line 346: I would not say "as the wind direction veers" as that suggests you look at a time period along which actual veering occurs, but you are actually just looking at different wind direction bins.

16. Line 352-353: Two comments:
    - At 0.67 km, you mention that the thrust coefficient increases in Fitch pAIM due to the induction. But, in the wind speed range you are looking (6 to 10 ms-1), I would think that the thrust coefficient either stays constant or decreases slightly as you estimate C_T based on a corrected, higher wind speed. The thrust force does increase and therefore the momentum tendency and the wake strength, but a reformulation is necessary.
    - Could you elaborate on this "issue"? Do you mean Fitch-pAIM gets a higher T6 power ratio at 2 km resolution because the cell-integrated induction factor increases the inflow speed and therefore the power production? Would this also not be the case for T1 so that it at least partly cancels out in the ratio P_T1/P_T6? It would be good to elaborate slightly in the text.

17. Line 366: Are you using the same wind speed range (6 to 10 ms-1) here for filtering figure 13? If so, please add this in the caption. If not filtered for wind speeds, please comment briefly on the underlying wind speed distributions of each PBL height class to make sure this is not explaining the power ratio differences per class (for example if the mid/high classes have significantly more above-rated inflow wind speeds than the low class this could also lead to this pattern).

18. Figure 15: please provide some extra info on the directional filtering here in the caption.

19. Line 418: I don't think you filtered for wind speed here, right? Because at above-rated wind speeds I would expect the thrust force to actually be highest in Fitch-O because the wind speeds are not corrected for induction. So I think the explanation should be more nuanced.

20. Line 426: "Of the wind speed" - do you mean of the wind speed deficit? Best to make this explicit.

21. Line 431: Was the wind speed filter applied throughout 3.4? I thought it was only in 3.4.2? If so, it should be specified correctly.

22. Line 432: And => An

23. Line 472: courser => coarser

24. Section 4.3: I think this section could be made more complete by adding some other elements.
    - Several other mesoscale models are being used for wind farm modelling and they may employ different model physics, notably PBL schemes, which could impact wake recovery. So, these results only pertain to WRF and mesoscale model intercomparisons are an important additional step.
    - I think it needs to be made explicit that in many geographic areas that employ lower turbine densities, resolutions of 2 km may be sufficient to meet the one-turbine-per-cell criterion so a sub-km resolution is not needed per sé. The same can be said for future wind farms which will have lower turbine densities.
    - Next, you mentioned that we cannot generalize to next-generation turbines but I would say the same is true for planned, multi-GW wind farms. More so because you observed that the streamwise internal wake evolution did not flatten out in the model runs like in the SCADA data.
    - Finally, it should be noted that this is mostly a "bulk" or "climatological" validation which does not reveal how well the WFP performs for specific weather systems (convective systems, fronts, other ramp events, ...) so this would require targeted investigation. In relation to this, no information was provided on the WRF power spectrum at short timescales (e.g. 10 min to 1 hour) so we don’t have a good view on the short-timescale variability experienced by the wind farms in these simulations. Please provide some discussion or additional analysis on this and its implications for specific use cases of mesoscale WFPs.