Authors' response

Response to all reviewers

The authors appreciate the feedback from all reviewers. The authors' response to each of the reviewers' comments (in black) can be found below (in red), as well as the rephrased sentences or added text (in blue).

Reviewer 1

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.
Thank you for this comment. We have now added a new subsection to the discussion.

"4.4: Implications for applications

The results provide several practical implications for the use of mesoscale models and their wind farm parameterizations in wind resource assessments. Firstly, a sufficiently fine horizontal resolution that limits the number of turbines within the same grid cell is highly recommended. Secondly, the choice of WFP affects the power estimates substantially, with Fitch-pAIM shown to provide the most consistent representation of undisturbed turbine power and average wake losses. Lastly, unavoidable biases in the background wind speed, as well as the impact of model choices on the results, should be carefully considered when absolute energy yield is of interest.

Mesoscale models are less suitable than engineering wake models for applications that depend on detailed turbine interactions, such as layout optimization or analyses of narrow wind direction sectors. However, their performance improves when considering broader wind direction ranges and therefore appear well suited for wind farm scale AEP assessments, provided that the AEP is not dominated by only a few narrow wind direction sectors. Their strength lies in the representation of interactions between wind farms and the atmosphere, something that is lacking in engineering wake models, making them particularly useful for assessing future wind farm buildout scenarios.

In general, the presented results support applications based on longer-term averaged quantities, while phenomena on the short timescale, and therefore turbine scale, remain insufficiently resolved. This makes mesoscale models with WFPs complementary to, rather than a replacement for, engineering and higher-fidelity wake models."

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.
   Agreed; replaced with "wakes that are too weak"
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°?).

We agree that "narrow wind direction sectors" was not sufficiently clear. However, we feel like it would not be appropriate to define a fixed angular threshold here. We here therefore rephrased the sentence as follows:

"Despite the limitations in accurately reproducing distinct wake features in aligned and staggered configurations, ..."

3. Line 14: WPF => WFP

   Fixed

4. Line 21: atlases => atlas

   Fixed

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

   Fixed

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.

   We have not carried out an analysis on how accurate this WRF setup is w.r.t. stability profiles. The measurements used in this study also don't allow for such an analysis. However, because the variable serves only to broadly distinguish between atmospheric conditions, high precision is not essential. We have added the following sentence to Section 3.4.3 to indicate that results in this section should not be overinterpreted:

   "Uncertainties associated with WRF's h_{BL} estimates should be considered, hence the classification in this section is intended for qualitative interpretation only."

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?

   Thank you for this comment. Both the SCADA power and the power coefficient used in WRF refer to the electrical power.

   We made the following changes:

   Section 2.2.1: "..., and the generated electrical power were used for analysis"

   Section 3.1: "Both P_{SCADA} and P_{WRF} refer to electrical power, hence including electrical losses. "

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

   Thanks, great idea!

"Details on six considered wind farms in the German Bight, indicated in Fig.1b, are presented in Table 3. Since the Hohe See and Albatros wind farms are adjacent, they were considered as one wind farm in this study."

Table 3. Overview of six wind farms in the German Bight considered in this study

Wind farm	Operator	Turbine type	Number of turbines	Hub Height [m]	Diameter [m]
Sandbank	Vattenfall	Siemens 4MW	72	95	130
DanTysk	Vattenfall	Siemens 3.6MW	80	88	120
Amrumbank West	RWE	Siemens 3.6MW	80	90	120
Nordsee Ost	RWE	Senvion 6.15MW	48	96.2	126
Hohe See	EnBW Hohe See GmbH & Co. KG	Siemens 7MW	71	105	154
Albatros	EnBW Hohe See GmbH & Co. KG	Siemens 7MW	16	105	154

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?

   We have answered this together with the next comment.

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.

    Thank you for the feedback. We reframed the paragraph to hopefully make the procedure on the wind direction bias correction easier to follow:

    "Each turbine's wind-direction signal was compared with the wind-direction average of the whole wind farm using weekly running averages. Sudden changes in the difference between an individual turbine and the farm average of more than 2° per day were treated as discontinuities, for example due to sensor changes or resets. The offsets between these periods were then removed so that each turbine had a consistent wind-direction signal over time. Even after removing discontinuities, the nacelle wind vanes may still have a constant angular offset relative to true north. To estimate this offset, wake losses from the SCADA data were compared with wake losses simulated using FOXES. An error in the wind-direction measurement causes the observed wake-loss pattern to appear rotated relative to the true compass directions. Therefore, the wind-direction offset for each turbine was adjusted until the measured direction-dependent wake-loss distribution matched the simulated one as closely as possible. Overall, the corrected mean absolute bias per turbine was approximately 3 ° with 4 % of the turbines exceeding a 10° true north bias."

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.

    Revised that actually not nominal rated wind speed was used as threshold but a wind-farm individual buffer was applied. We would like to highlight that our focus of this paper was not on the SCADA data processing workflow but on the comparison of the data to mesoscale simulations. As elaborated at the end of the SCADA-processing section, the uncertainty of the processed data is estimated to be rather low compared to the sensitivity of model choices. SCADA-data processing for model validation deserves separate publication and there is some published work which has highlighted certain challenges. One of them is that for the five wind farms in the manuscript, there is no completely unified automatized processing possible, and it always requires expert knowledge to understand the implications of processing choices.

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.

    Done.

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.

    Agreed, thanks

    "All sub-figures: Only undisturbed turbines are included, and the corresponding undisturbed wind speeds are obtained from a separate simulation without turbines."

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.

Thanks, we have adjusted the figures accordingly. We also updated figures 9, 15 and 16 for consistency. Only in figure 11 now the dashed/solid lines don't refer to staggered/aligned conditions, but we believe that this is clear from the caption.

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.

    Fixed

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"?

    Thanks for catching this. Instead of "thrust coefficient", here it should read "thrust force".

    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.

    In Fitch-pAIM the induction correction is multiplied by the number of turbines in the cell, see Eq. A16. This means when there are more turbines in a cell, the correction is larger and therefore the corrected wind speed increase is larger. As can be seen in Fig. B1b, T1 on shares its cell with two other turbines, whereas T6 shares its cell with five other turbines. The correction is therefore larger for T6 than for T1, which increases the power ratio discussed in Fig. 12a. As a brief clarification, we added the following:

    "At 2 km, however, this effect is outweighed by the issue of multiple turbines in one cell, which amplifies the induction correction more for T6 (in a cell containing six turbines, see Fig. B1b) than for T2 (in a cell containing three turbines), resulting in a higher power ratio than Fitch-O."

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).

Filters for WD and WS have been added to the caption.

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

    Done

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.

    Correct, no filter for wind speed was applied here (mentioned in line 413-414 or the original manuscript).

    And agreed with a more nuanced statement. The induction correction effectively increases the thrust force below rated wind speed and decreases the thrust force above rated wind speed. The mean thrust is then a function of the wind speed distribution. The following has been added:

    "..., due to increased thrust from the induction correction. This represents an average between a thrust force increase below and a decrease above rated wind speeds, weighted by the wind speed distribution."

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

    Fixed: "wind speed in the wake."

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.

    Fixed

22. Line 432: And => An

    Fixed

23. Line 472: courser => coarser

    Fixed

24. Section 4.3: I think this section could be made more complete by adding some other elements.

    a. 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.

       Agreed

       "Lastly, when other mesoscale models or other Planetary Boundary Layer schemes (currently not available in WRF) are used, the wake recovery rate could be affected, thereby constraining the transferability of the presented results."

b. 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.

   Thanks, we added the following two sentences now:

   "Many other regions employ lower turbine densities, so the criterion of one turbine per cell will be achieved at a coarser resolution."

   "However, the absolute spacing between turbines will increase, so the criterion of one turbine per cell will be achieved at a coarser resolution."

c. 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.

   Agreed

   "Planned and future wind farms will consist of many more turbines and contain larger turbines than the wind farms considered in this study. The accuracy of existing wind farm parameterizations on these scales is unknown."

   Additionally, in the conclusions:

   "Additionally, the applicability of the results for larger wind farms and future turbines is unclear."

d. 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.

   The aim of this manuscript was to evaluate the performance of WRF's wind farm parameterizations. These WFPs are not responsible for how WRF handles the mentioned specific conditions; rather this is the job of the PBL scheme and the dynamical core. We therefore consider this out of scope for the current manuscript.

Following your first comments, we have added the following sentence to the discussion:

"In general, the presented results support applications based on longer-term averaged quantities, while phenomena on the short timescale, and therefore turbine scale, remain insufficiently resolved. This makes mesoscale models with WFPs complementary to, rather than a replacement for, engineering and higher-fidelity wake models."

Reviewer 2

Overall Assessment

Validating wind farm parameterizations with offshore SCADA data by Sengers et al. demonstrates the value of the recently developed Fitch-pAIM parameterization through a multi stage validation. The authors first compare the energy yield of Fitch-pAIM against six existing wind farm parameterizations using SCADA observations from six offshore wind farms in a well studied German Bight region. Based on this initial evaluation, the standard Fitch and EWP schemes are retained for more detailed comparison with Fitch-pAIM at a finer grid cell resolution. Results ultimately suggest Fitch-pAIM's improved turbine power curve representations translate into more accurate energy-yield predictions.

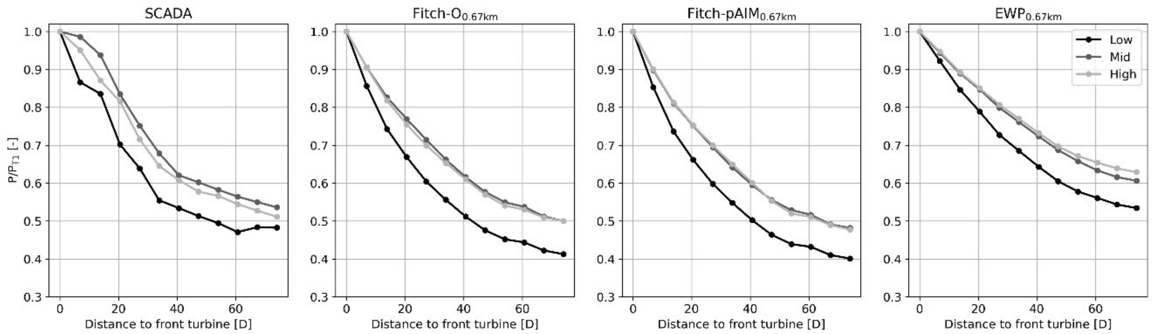
The manuscript is well written and addresses a novel topic of clear interest to the offshore-wind meteorological modeling community. I believe the manuscript would be further strengthened by addressing the following comments prior to publication.

Specific comments

1. Defining uncertainty:
   The manuscript frequently refers to "uncertainty," and only some of these use cases are accompanied by formal uncertainty estimates. I would encourage the authors to distinguish more carefully between "uncertainty", "variability", and "sensitivity" and more clearly distinguish what quantity is being evaluated in each case.
   Thank you for this comment; we fully agree. With made the following changes (line numbers correspond to those in the original manuscript):
   "The remainder focused on addressing the main sources of error affecting the total energy yield." (l 6-7)
   "(2) assess the major sources of error independently," (l 84-85)
   "Therefore, the analysis in the following subsections attempts to look at the main sources of error individually. Those considered to have the greatest impact on the model result are ..." (l 222-224)
   "3.2 Wind speed ~~uncertainty~~" (l 226)
   "3.3 Gross yield ~~uncertainty~~" (l 247)
   "The second source of error concerns the gross yield, ..." (l 248)
   "3.4 ~~Uncertainty of~~ internal wake and blockage losses" (l 279)
   "In the evaluation of the wind farm parameterizations, it can be considered another source of error." (l 401-402)
   "3.5 External wake losses ~~uncertainty~~" (l 403)
   "Considering the ~~uncertainty related to~~ wind speed bias" (l 459)

"The main sources of error were addressed in the remainder of the work" (l 524-525) "The presented results should be interpreted with care due to the associated sources of error, including deviations in mean wind speed, model configuration, and SCADA data uncertainty." (l 538-539)

2. Atmospheric stability vs. boundary-layer structure:
   Section 3.4.3 uses model PBLH as "an alternative to any stability metric" to evaluate model performance across different boundary-layer regimes. Can the authors clarify how the binning thresholds were determined and to what degree the modeled NWF PBLH is representative of the observed boundary layer structure? Because boundary-layer depth and atmospheric stability are related but distinct properties, it would also be helpful to discuss this distinction explicitly or, if feasible, corroborate the analysis using an independent stability diagnostic. Such clarification would strengthen the manuscript's conclusion that the parameterizations perform robustly across a range of atmospheric conditions.
   Thanks for this comment. "any stability metric" is clearly too strong and has now been replaced with the intended "a stability metric".
   The reason for using PBLH instead of RMOL is that the latter is an instantaneous, surface layer variable that does not necessarily reflect the conditions in the rotor area. The PBLH has been shown to affect the wake recovery rate, and has therefore been chosen in this section to assess its impact on the wake magnitude. The following short clarification has been added to the text:
   "It was chosen instead of the Obukhov length as it better reflects the conditions affecting the wake recovery."
   Lastly, the point about the binning thresholds is absolutely valid, as these were not based on anything other than convenience. We have now reworked this by following this paper (https://doi.org/10.5194/wes-8-787-2023), where the 33.3rd and 66.7th percentile was used for stability classes. This did not affect the results much, and did not affect the interpretation of these results.
   "Three discrete classes were considered, separated by the 33.3rd and 66.7th percentiles (following Baas et al., 2023) of the h_{BL} in the results of the WRF simulation without turbines after filtering for the wind speed and direction considered here: Low (h_{BL} < 616 m), Mid (616 < h_{BL} < 943 m) and High (h_{BL} > 943 m) boundary layers."



3. Generalizability

   Throughout the Discussion, the authors identify several factors (e.g., the number and distribution of vertical levels, tuning parameters within the schemes, site characteristics, etc.) that are framed as unlikely to alter the principal conclusions while simultaneously serving as promising avenues for future work. It would be helpful to clarify which aspects of the results are believed to be fundamentally insensitive to these factors, and which may change quantitatively or qualitatively as they are explored in future studies.

   While we agree this would be valuable, without carrying out dedicated sensitivity tests it is incredibly difficult to speculate how these factors affect the results, e.g., whether momentum extraction or wake recovery are affected.

   To clarify what findings we don't expect to change, we have moved (and slightly reformulated) the last paragraph of Section 4.3 now to the top of Section 4.

   "The aspects mentioned in this section are expected to affect the results. However, the general findings of this work are likely to remain valid: All WFPs except for Fitch-pAIM will systematically underestimate power at undisturbed turbines, EWP continues to generate substantially weaker wakes compared Fitch-O and Fitch-pAIM, while external wakes are similar for all WFPs."

   Specifically for the impact of vertical levels on the model results, we have reformulated this with more focus on what should be studied in future work:

   "Secondly, Mangara et al. (2019) illustrated that the number of vertical model levels inside the rotor area affects the wind speed deficit in the wake. Future work should study in more detail how aspects related to the wind farm parameterizations (momentum extraction and turbulence generation) and aspects related to the boundary layer schemes (wake recovery, turbulence dissipation) are affected."

   Lastly, we took the opportunity to include our recently published work on the TKE coefficient, which addresses one of the concerns mentioned in Section 4.2.

   "Recent work by Sengers et al. (2026) proposed a new analytical equation derived from the 1D momentum theory to compute the TKE coefficient, making this correction factor obsolete. Using this new equation, slightly more TKE is added to the flow than when using α_{TKE} = 0.25, resulting in a small power increase due to faster wake recovery."

Organizational suggestions

1. Methods versus Results. Several methodological descriptions are embedded within the Results section. While this organization can sometimes improve readability, expanding the Methods section to introduce the evaluation metrics and their scientific motivation may provide a clearer roadmap for readers here.
   We agree that this is a trade-off between readability and structure. We believe that the preferred structure is subjective, and that it depends on the complexity of the paper. The current manuscript touches on many different aspects of wake modeling with a mesoscale model and uses a range of analyses to discuss these aspects. We fear that if all methodological descriptions were taken out of the results section and introduced in the methods sections, we would have to reintroduce the analysis again in the results section in the form of summaries and references. We believe this is detrimental to the readability of the manuscript and have therefore chosen to keep the structure as is. Should the reviewer have specific preferences regarding this point, we would appreciate a clear indication of which paragraphs they recommend moving to the Methods section.
2. Model characteristics versus model performance. Separating the description of how the parameterizations differ from the subsequent evaluation of their performance may improve the flow.
   Unfortunately, it is not entirely clear to us what exactly the reviewer meant with this comment. Much of the parameterization description has been separated from the main text by moving it to the appendix. In the results sections we only discuss the performance of these parameterization, and aim to explain differences between the parameterizations by referring to their theory and assumptions. We believe that this is a common and clear way of evaluation the results.
   If deemed necessary, we kindly ask the reviewer to provide a more detailed explanation of their comment in a next revision round.

Editorial Comments

1. L14: "WPFs"
   Fixed
2. L36: Consider "financial" rather than "economical."
   Since the reference paper itself uses economical instead of financial, we decided to not make the suggested change.

3. Figures 1, 8, and 10 could benefit from larger legends, labels, and axis text. The EWP color is also difficult to distinguish in several plots.
   Changed EWP color to orange in all figures. Figures font sizes revised.
4. L60: Perhaps “incompatible” is more appropriate here?
   Disagree. The older results would be incompatible for direct comparison with newer results, but it itself the results are questionable.
5. L115–129: Consider summarizing this information in a table or incorporating into Figure 1.
   Done, see answer to Reviewer 1, specific comment 8
6. A summary table of the evaluated model characteristics and performance metrics may help orient readers.
   Thanks, great idea. We have now added the following to the conclusions:
   “The main results of the following are summarized in Table 4.

   Caption Table 4: Qualitative summary of the main results. Downward and upward arrows indicate underestimation and overestimation of generated power as a result of these sources of error. Double arrows indicate a stronger bias, while circles indicate no substantial bias.”

	Fitch-O	Fitch-pAIM	EWP
Wind Speed	↓	↓	↓
Power Curve	↓	○	↓
Internal Wakes	○	○	↑↑
External Wakes	○	○	○

7. L195: "wind speed deficit"
   Fixed
8. L276: Should Section 3.4.3 also be referenced?
   Thanks for catching this; this has been added now.
9. L472: Consider "generated" rather than "produced."
   Agreed
10. L472: "coarser."
    Fixed
11. L479: “This section”
    We are not sure what this comment means, but we have added the following: “More details are discussed below.”
12. L467: “, as shown in Chang et al. (2026).”

Fixed

13. L 479: “Its”

    Fixed

Community Comment by Paul vd Laan

The authors validate the output of several wind farm parametrization models with SCADA measurements of five offshore wind farms. The article is well written and contains important information that can have a high impact. This is motivated from the fact that the wind industry needs accurate models for wind farm losses in order to design new wind farm layouts in areas with neighboring wind farms. Furthermore, I find the discussion and application of different validation metrics very interesting. However, I have some concerns with the validation procedure because the authors employ gap filling of the SCADA measurements, which essentially creates new data points that have not been measured; more details are added below.

1. When wake models are validated with SCADA, it is common to filter the SCADA for time intervals where turbines are curtailed, boosted, or not operating normally. When a limited (raw) SCADA set is available, as is often the case in scientific validation studies, the remaining filtered SCADA set may become too small in order to perform a statistically valid validation. It is likely, that the authors have recognized this limitation and then chose to perform an alternative approach by replacing these data points with expected power values obtained from the manufacturer's power curve and ERA5 wind speed values. As a result, the proposed SCADA now includes data points obtained from a model and one could argue that the comparison of gap-filled SCADA with wind farm parametrization is a model-to-model comparison and not a validation. Therefore, I strongly recommend the authors to avoid the gap-filling procedure and perform filters instead. The authors write that the gap-filling procedure leads to a more fair comparison with SCADA, however, the filters for curtailment, downtime and non-normal operation could also be employed as a post filter to the meso-scale simulation results. If the authors prefer to keep using the gap-filling method then they should at least write how much of the data (in percentage) is filled, which could be reported as an average over all five wind farms to avoid showing potential confidential information. Another analysis could be made by showing the impact of the gap-filling procedure on a SCADA set where the gap-filling is not required (i.e. most turbines are operating normally). One could then perform the gap-filling on a number of random turbines (so gaps are artificially made). The results with and without gap filling can then be compared.

   Thanks for critically checking this, Paul. We agree that many of the manuscript's outcomes hinge on the validity of the reference data set, and with that on an appropriate processing of the raw SCADA data.

   You're right that after filtering for full operational availability of all turbines, the remaining data sets became too small to provide a meaningful reference for this validation exercise. For most of the wind farms only less than 10% of periods remain in which all turbines are fully operating.

We agree that a filter of certain turbines or periods in the post-processing workflow could have been an alternative. In the end, we selected the gap-filling method because if we filter for periods in the SCADA-data, the wind conditions might be not equivalent during these periods in the WRF simulations because they are not fully correlated in time and may have phase-shifts. Thus, removing periods from the data may induce unintended biases between data and simulation results.

To your last part: This is actually what we have done and written in the paper. Apparently, it has not been formulated clearly enough so we revised the last paragraph:

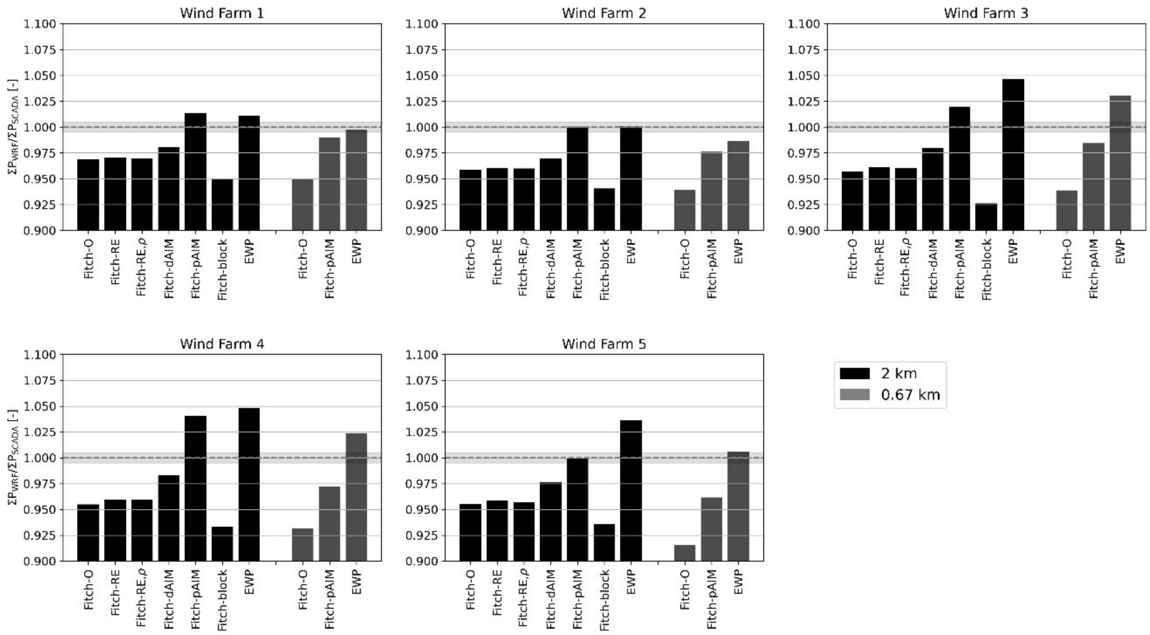
"On average about 15 % of the power values were replaced with the estimated power. In below-rated conditions this decreases to 10 %. Because the gap-filling method represents the main source of uncertainty of the validation data, the uncertainty of the potential power estimate was calculated using a Monte-Carlo approach. Reference periods with more than 95 % of turbines in normal operation were selected. At randomly selected time stamps, the measured wind speed was replaced by the KNN-based wind-speed estimate, and the resulting estimated power was compared with the measured power. The differences between both values were used to quantify the uncertainty associated with the potential-power estimate.

The uncertainty of the energy-yield estimate at 100 % turbine availability was calculated by multiplying the fraction of flagged periods by the uncertainty of the potential-power estimate during those periods. For instance, an uncertainty of 10 % in the estimated power during 15% of the time results in an overall uncertainty of 1.5 %. Across the wind farms considered in this study, the resulting wind-farm-level uncertainty ranged from 0.1 % to 0.5 %. Individual wind-farm uncertainties are not reported for confidentiality reasons."

2. The validation figures lack error bars showing the uncertainty of the mean of the processed SCADA. Adding this information would help the reader to quantify the difference between the models with respect to the SCADA uncertainty. The authors may find that some wind farm parametrization models perform similar when taking into account this uncertainty.

   In Section 2.2.2, Monte-Carlo simulations were performed to estimate the error introduced by the processing methodology. Considered is the AEP on wind farm level. The upper limit of this uncertainty (0.5%) was added as a shaded area to Figure 5, and the following was added to the caption.

“The shaded area around unity indicates the uncertainty of the SCADA (maximum of 0.5%, as mentioned in Sect. 2.2.2).”



Note that this 0.5% error is at wind farm level. Errors of individual turbines could be larger, depending on their data availability and consequently how much gap-filling was done. However, these are excluded to protect the confidentiality of the data.

3. The authors do not provide a script used for processing the SCADA and this makes it difficult to replicate the validation procedure. I am aware that the employed SCADA is confidential, but the authors could use an open SCADA set for the purpose of providing a test to the script. For example, the authors could use the GLoBE dataset, which is 1.5 years data containing the Amrumbank West and Nordsee Ost wind farms.

   To our knowledge, the GLoBE data set is not publicly available. For confidentiality reasons we cannot directly share our scripts, but we hope that our rewrite of Section 2.2.2 following your first comment and the specific comments from Reviewer 1 has clarified any questions related to the SCADA processing methodology.

Additional changes

1. The paragraph below was deleted from Sect. 4.2. Between submission of the manuscript and now, the authors realized that in the WFP code, WRF actually uses u- and v- components aligned with the grid, not with north-south and east-west direction. The issue mentioned in this paragraph therefore does not exist, and therefore the paragraph was deleted.
   "Fitch-pAIM uses a correction factor to account for how much of the mesh cross-section is occupied by the turbine area. For simplicity, this correction assumes that the mesh orientation is aligned in north-south and west-east direction, which due to the map projection is only true at the domain center. The figures in Appendix B, however, illustrate that the misalignment of the cells containing the wind farms of interest was quite small."