Response to reviewer 2

We would like to thank the reviewer for the interesting insights and constructive comments made during the critical assessment of our work. Their thorough review has greatly enriched the quality and depth of our manuscript. In what follows we address the reviewer's concerns point by point. We give a motivation to the changes, and we mention the additions to the manuscript (*in italic*).

This paper conducts a 10-year mesoscale atmospheric simulation in Kattegat using the COSMO model. The goal of this paper, as stated by the authors is "to investigate what factors influence mesoscale wind speed variability, on what timescales this variability occurs, and how it affects wind power output in offshore wind farms". While I have come across many papers that conduct mesoscale wind resource assessments, I believe that a paper like this that digs deeper into the atmospheric mechanics is valuable. I really enjoyed the spectral analysis as well as the spatial analysis. I also appreciate that the authors have stored code and data on Zenodo. This is also the first mesoscale wind energy paper that I have read that doesn't use WRF, which is refreshing. That being said, I have major concerns regarding connections to the broader literature, novelty, methodology, and analysis. To the editor, I recommend a status of Major Revisions.

Reply: We appreciate the comments by the reviewer and are happy to hear that our efforts to dig deeper into the atmospheric processes is appreciated. A multi-model approach is very common for many research questions (impact of urbanisation, deforestation, increasing greenhouse gasses, ...) and we also see much value in doing this also for the wind energy sector. We appreciate the comments given and have done an effort to mitigate the concerns that the reviewer poses here.

Major Concerns:

Framing of the paper

To put it succinctly, I'm uncertain if the authors are claiming that either (a) few mesoscale wind simulations studies have been published or (b) if they are very specifically claiming that the mesoscale wind simulation studies that have been published have not sufficiently analyzed "mesoscale wind speed variability".

If they are claiming (a), then I strongly disagree. The authors categorize the following as mesoscale phenomena: "thunderstorms, but also sea breeze systems, low-level jets and gravity waves". Many many papers have been published on how wind energy is affected by these, e.g. (Tomaszewski and Lundquist 2020 for thunderstorms, sea breezes in the North Sea by Steele et al. (2014), many papers for LLJs, many papers by Allaerts on gravity waves). Additionally, there has been extensive work on the mesoscale through large-scale wind resource assessments (like the New European Wind Atlas and the WIND Toolkit) as well as smaller but important studies (e.g., Hahmann et al. 2015). Relatedly, I believe that NEWA, the Dutch Offshore Wind Atlas (Kalverla et al. 2020), and possibly NORA3 (Cheynet et al 2022) all include Kattegat in their simulations, and these products should be mentioned somewhere if so.

If (b), the citations in L62-63 are a small set of mesoscale wind resource papers, and I don't believe these papers do anything unique regarding wind speed variability.

Reply: Indeed, in retrospect we agree that the claim that "*Less is known about mesoscale weather systems..."* is not appropriate. We have restructured and rewritten large parts of the paragraph that used to start with "Less is known about mesoscale weather systems, for example in organised convection." We have added references and also mention the wind atlases. We hope that rewriting part of the introduction results in a better framing for the paper. The novelty of the paper is addressed in the reply to the next point.

Variations in wind speed can also arise from mesoscale weather systems. With their length scales ranging up to a hundred kilometres and timescales spanning from ten minutes to a few hours, mesoscale weather systems occupy an intermediary position between turbulence and synoptic weather systems. (...) Wind atlases utilising these models at a kilometre grid have been made available to the public, such as the Dutch Offshore Wind Atlas (DOWA) (Wijnant et al., 2019) and the New European Wind Atlas (NEWA) (Petersen et al., 2014; Hahmann et al., 2020; Dörenkämper et al., 2020), and provide thoroughly validated information at the mesoscale level (Kalverla et al., 2020). Both Hahmann et al. (2015) and Wang et al. (2019) have evaluated kilometre scale climate models against in situ data and have shown the capacity of these models to reproduce mesoscale variability. (...)

Novelty

I normally don't comment on novelty, but many mesoscale simulations have been conducted specifically in the North Sea, with domains that include Kattegat. A non-exhaustive list includes Hahmann et al. (2015) and the New European Wind Atlas (see the two papers in GMD). I believe that the Dutch Offshore Wind Atlas (see papers by Kalverla) and possibly NORA3 (Cheynet et al. 2022) may also cover this region. I don't see any references to these papers. The authors should clearly state why their work is novel to help the readers out. I feel like I've seen many papers also do similar-but-different seasonal analysis before (e.g., Wang et al 2019 in California), and these should also at least be referenced.

Reply: We included several lines on the novelty of our study in the introduction, and we included the references mentioned above.

In our paper we use a 10-year integration of the convection permitting climate model COSMO-CLM with a horizontal resolution of 1.5 km to study the atmospheric mechanics behind mesoscale variability. Augmenting existing research, we added here the implications for power production for a full 10-year period. Moreover, this paper introduces a new index, which identifies spatially coherent mesoscale systems, allowing for detection of situations with a strong mesoscale system. We add on this spatial index with temporal analysis methods, since these methods are inherently complementarity. Additionally, this paper uses COSMO-CLM, which complements the frequently used model WRF, enabling a multi-model approach to applications in the wind energy sector in the future. Our analysis has been performed using 10-minute wind speed data, which is not available from wind atlases. (...)

Methodology

Simulation setup: What is the timestep of the simulation? How frequently is data saved out? Did you run a single, 10-year long simulation or did you chunk up the job into smaller periods? Did you use spectral nudging to encourage the long simulation to not drift too far away from the expected ERA5 values? I'm unfamiliar with COSMO, but in WRF, it is known that the PBL scheme choice is very important. If multiple options are allowed in COSMO, which PBL scheme is used here? While not required by this journal, I highly encourage the authors to upload an example configuration file somewhere so that others may more easily reproduce this study in the future. There is possibly one on Zenodo, but the link is private, so I cannot view its contents.

Reply: We added this information to the manuscript, and we added a configuration file to the Zenodo link. Additional information about the nesting strategy is mentioned in appendix A. Except for the 10, 80 and 100 m wind speed which are saved out every ten minutes, all variables are saved out hourly.

The dynamical core of this model solves the primitive thermo-hydrodynamical equations describing a compressible flow in a moist atmosphere (Doms and Baldauf, 2018) with a timestep of 10 seconds. (...) Subgrid-scale turbulence is parametrised by a one-dimensional diagnostic level 2.5 closure scheme based on a prognostic TKE equation (Raschendorfer, 2001; Schulz, 2008a, b). (...) For practical reasons the calculation of the 10-year simulation has been divided into smaller periods, and by using the restart files generated by COSMO these periods were initialised with a warm start. (...)

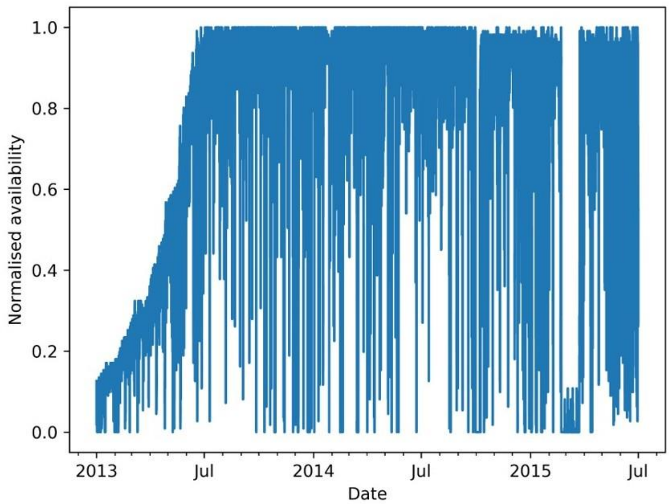
Different simulation setups have been tested for 3-month integrations. These tests included a larger domain (340 × 360 grid points compared to 180 × 184 grid points), adding an intermediate nesting at ≈ 12 km resolution and applying spectral nudging. We found no added value of using a larger domain, of adding an in between nesting step and of applying spectral nudging compared to the wind data from scatterometer. This result is in line with the findings of Ban et al. (2021) where different nesting strategies of COSMO-CLM in ERA-Interim do not show any substantial differences.

Validation efforts: Researchers often validate modeled winds against measured winds in order to built trust, but the validation study here instead erodes my confidence. I also don't see how validation furthers the authors stated goals in the intro. I believe that all the lidar analysis should be struck entirely. Please write a sentence or two that explicitly ties the motivation behind the validation back to the goal of the paper.

Reply: The validation with lidar data is included in the paper in order to validate the wind speeds at 100 metres, as we mainly use these wind speeds in the paper. We added these lines to the introduction to motivate the validation efforts. We do not agree that the lidar comparison should be removed from the manuscript (see point below).*The goals mentioned above can only be achieved if the model represents the real atmospheric winds. Therefore, we have incorporated an evaluation of the near-surface winds using ASCAT data (ref) and the 100-m winds using lidar data. These nicely complement each other as the ASCAT measures multiple locations, but only at 9h UTC and 21h UTC, while the lidar data provides measurements during the whole day, albeit at one single location.*

Lidar: The simulated winds do not include wakes, but the lidar is probably being waked. That waking can be significant (>1 m/s modifications). We also don't know how close the lidar is to the nearest turbine, and also the number of operational turbines changes throughout the comparison period, changing the waking strength. The authors attempt to minimize the effects of waking by looking at periods when wind farm availability stayed below 50%, but that is insufficient in my opinion.

Reply: In response to the editor’s comments in an earlier stage about the fact that only near surface winds were initially evaluated but not the hub height wind speeds, the comparison with lidar was added. We agree that it is not an ideal location, but it is the only offshore data that are available. In addition, the fact that there is not a large difference in performance between the first 100 days and the entire period gives us confidence of the added value of including the lidar data (see the wind farm availability figure added hereunder). We agree that the reader should be aware of the caveats and therefore explicitly state the ones that you mentioned in the manuscript now.



Using scatterometer data only the near surface wind speeds at 9h and 21h UTC of our model can be evaluated. As a complementary validation of the scatterometer a light detection and ranging (lidar) device located 2 kilometres west of the Anholt wind farm is used. This is not an ideal location, since it might be affected by the wind farm which is not implemented in our model, but it is the only offshore measurement at hub height in the Kattegat area. Therefore, a sub period of the measurement campaign was used, one in which the Anholt wind farm was still under construction and the availability of operation wind farms increased from xx to xx%. Note that eventually xx windfarms were operational in the Anholt wind farm. Moreover, the wind mainly blows from the west in this area (Karagali et al., 2013), resulting in a lidar signal that is largely unwaked. (...)

ASCAT: I am less familiar with this instrument, but I suggest that the authors point out that others in wind energy have used this data source before (e.g., Hasager et al. 2020). I mention this because I know there is some controversy regarding validating WRF against SAR (a different instrument), but if others in the wind energy sector have looked at ASCAT previously, then there is a stronger case for its use here.

Reply: A reference to the validation of the DOWA with ASCAT has been added to the paper, as well as a reference to Hasager et al. 2020.

Validation with ASCAT data has already been used for a variety of offshore wind datasets (Hasager et al., 2020; Duncan et al., 2019).

ASCAT: I haven't seen others make maps of winds at different percentiles. Why did you do this instead of simply comparing mean wind speed maps? You give a justification on L110, but I don't quite follow. Maybe if you provide a summary of what the "double penalty" is, that would help clarify things.

Reply: We have added an explanation of how a double penalty could impact the comparison between scatterometer data and simulation output.

A direct point-to-point comparison between ASCAT and model output, like RMSD, might underestimate the quality of the model to represent the mesoscale variability. In the absence of strong forcing over sea, it can easily happen for a model to reproduce a mesoscale system, albeit slightly shifted in time and/or space. The reproduced mesoscale system then results in two errors. First it induces an error over the place where it should have been but is not reproduced now, and secondly it induces an error over the place where it is now but should not have been, which is referred to as the double penalty (Marseille and Stoffelen, 2017). That is why, apart from the RMSD, statistical methods to assess the distribution of wind speeds, like the 25th, 50th and 75th percentile wind speeds, are also used. The comparison of these distribution parameters has an added value compared to an evaluation of the mean, as an erroneous distribution can still have a good representation of the mean due to error compensation. Evaluating distribution parameters is therefore more rigorous than evaluating only the mean.

ASCAT: L106: It's hard to judge if 229,503 WVC is a lot of data or a small amount of data. Could you give more context? Maybe an easy number to calculate is the number of aggregated WRF grid cells over this period.

Reply: We have added a calculation of the number of aggregated simulation grid cells to the manuscript.

The scatterometer does not cover the Kattegat on every overpass it makes, yet a total of $299,503$ Wind Vector Cells (WVC), which is equivalent to ≈ 20,800,000$ aggregated (i.e. averaged) model grid cells, is available for comparison.

ASCAT: If a model validates well at 10 m, that doesn't necessarily mean that the model is accurate at hub-height (see the Bodini extrapolation papers if you're interested). Please add that caveat in somewhere, as many today would consider validating against near-surface winds to be of minimal utility (for what it's worth, I am not one of those people).

Reply: We have added this caveat to the manuscript.

Using scatterometer data only the near surface wind speeds at 9h and 21h UTC of our model can be evaluated, which is not necessarily representative for the hub-height wind speed. As a complementary validation of the scatterometer a light detection and ranging (lidar) device located 2 kilometres west of the Anholt wind farm is used.

Analysis

Fig 5: I really like the spectral analysis. One of the goals of this paper is to "to investigate what factors influence mesoscale wind speed variability", and as such, I would like to see a stronger physical justification as to why the wintertime shows stronger short-timescale TKE and the summertime shows stronger long-timescale TKE. Why would relatively warm SSTs impact the 20 min - 1 hour range as opposed to a different range? Why would sea breezes and nocturnal jets contribute to TKE specifically on the 6-12 hour range? Consider citing papers in this section to justify your analysis, I don't think you necessarily need to write any code to address this point. Also, consider mentioning here that you also further investigate this line of questioning later in the paper.

Reply: We added a more physical explanation of how winter conditions favour short timescale variability, and summer conditions result mainly in long timescale variability to the manuscript.

Relatively warm SSTs in winter result in turbulence and unstable conditions. Unstable conditions are a driver for convective cells. The systems are not bound to a specific place and are advected with the wind. Therefore they can move relatively fast and result in short timescale variability (20min-1h) (Ahrens, 1994). (...) The difference between DJF and JJA on the longer timescales may be due to the sun being higher in summer, and it heating the land more effectively. With the sun over land, the air above it expands and generates a breeze over the sea during the morning or afternoon. During the night, due to the land cooling nocturnal jets may be formed. Sea breezes are fed by the contrast in land surface temperature and sea surface temperature. This keeps them confined to the vicinity of the intersection between land and sea, and this results in longer timescale variability (Ahrens, 1994).

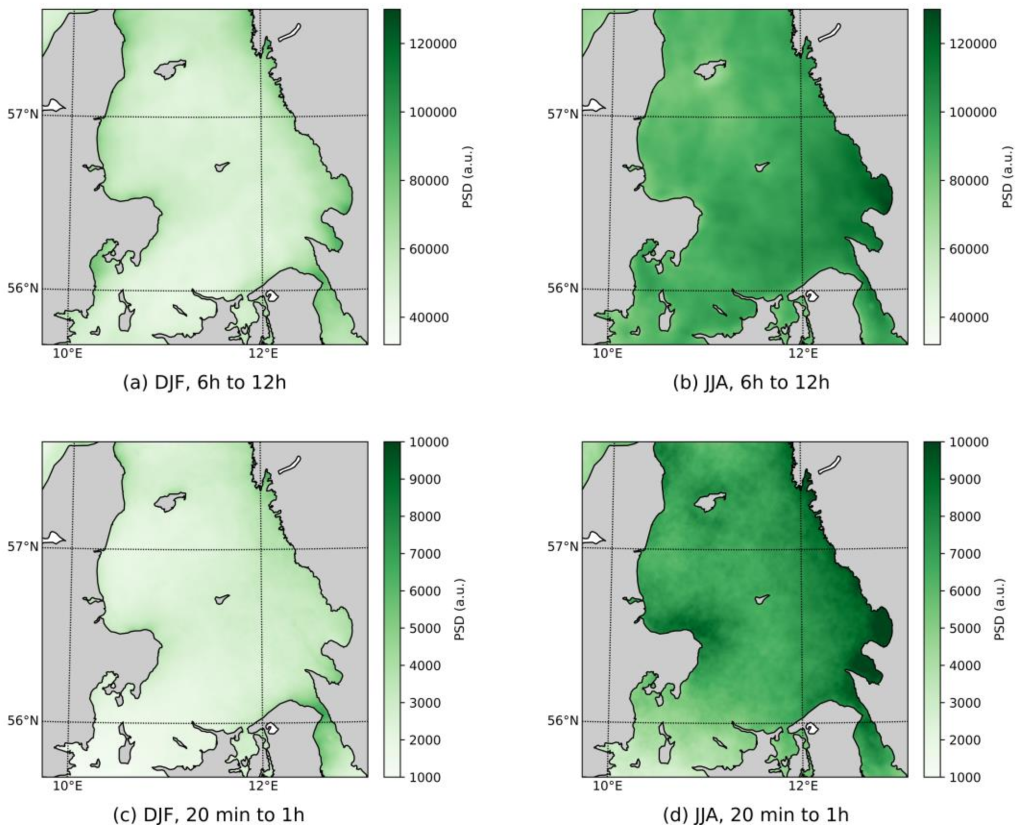
Fig 6: Nice figure! I think your analysis in L234-235 makes sense. I like that you have the 11 day case study to justify your hypothesis, but please include the figure somewhere, perhaps in an Appendix. Consider talking about Vincent et al (2011) citation in the preceding paragraph.

Reply: The integrated periodogram for the 11-day case has been added to the appendix, and there have been made more references to the study of Vincent et al. (2011).

Fig 7: You use a power curve from a 120 m turbine, but I assume you are calculating power using winds measured at 100 m, correct? You should use a turbine for which you have reliable simulated winds. If you didn't save out winds at 120 m, I believe the NREL 5 MW turbine has a 90 m hub height, and you could interpolate between your modeled winds at 80 m and 100 m. In theory you could extrapolate your modeled winds up to 120 m, but I have a feeling that would introduce a lot of noise and also uncertainty, so I recommend against that approach.

Reply: The power curve of the Siemens SWT-3.6-120 3.6 MW turbine has been used in this case. This turbine does have a hub height of 90m (the 120 in the name is a reference to the diameter of 120 metres). We have updated the data in figure 7 and figure 8 with the interpolated 90m wind speeds and included these figures in the paper. The difference in figure 7 is hardly noticeable (but there is a difference), for the wind speeds in figure 8 there is a small but noticeable difference. The conclusions remain the same, applying the 90m wind speeds or the 100m wind speeds.

Periodograms can also be used to examine the fluctuations in potential wind power. The power curve of a wind turbine converts a wind speed time series to a wind power time series, and of the latter time series a periodogram can be calculated. The power curve of the Siemens SWT-3.6-120 wind turbine is available in table form (Bauer and Matysik, 2022) and a cubic spline interpolation is used to obtain the power for every possible wind speed. As the hub height for this type of turbine is 90 metres the 80 metre and 100 metre model wind speeds are interpolated to 90 metres via the power law given in equation 1.



(a) DJF, 6h to 12h (b) JJA, 6h to 12h

(c) DJF, 20 min to 1h (d) JJA, 20 min to 1h

L270-279: I don't think that the reader was warned that this type of analysis was going to be conducted. This paragraph felt like it came out of nowhere.

L281-286: I don't think that the reader was warned that this type of analysis was going to be conducted. This paragraph felt like it came out of nowhere.

Reply: These analyses are now introduced in the introduction and explained in the methods section of the manuscript.

The added value of a convection permitting simulation over ERA5 for wind and power resources is assessed in this section.

A comparison between ERA5 and our simulation output was performed using the MSVI, assessing the added value of convection permitting simulations for mesoscale variability. This is done by regridding ERA5 to the grid of our COSMO simulation output, and then applying the MSVI metric to the ERA5 data.

Given the offshore wind speed variability, wind power fluctuations are expected. Wind speed and wind power are related to each other via the power curve of a wind turbine, and given this relation an MSVI analysis on the wind power could in principle be made. Yet the shape of this power curve imposes restrictions on our methodology: below the cut-in wind speed of a turbine (3 m/s) no power is produced, prohibiting a MSVI calculation for power fluctuations as the denominator would become zero. Instead we opt for a RMSD comparison between power time series for a stationary 10 x 10 window and a 30 x 30 window. The stationary windows are positioned in the area of the Anholt wind farm. The small window is *in area nine times smaller than the large window. In order to cover the whole large window, nine small window power time series are calculated and compared to the large window power time series. The average over these nine RMSD values is then taken to assess the differences in wind power.*

Minor concerns:

1. Figs 2 and 3: In accordance with WES "colour vision deficiency" publication guidelines, please use a different colormap than the rainbow ones.

Reply: The colourmap for figs 2, 3 has been adjusted to matplotlibs 'viridis'. Figs 10 and 11 have also been changed to this colourmap, as they also used the rainbow ones.

2. L26-34: This is a suggestion and not a requirement. I found the discussion on farm density a bit hard to follow, and I wasn't certain why the authors were talking about density. Consider reorganizing the paragraph to move the thunderstorm example higher up.

Reply: This paragraph has been reorganised.

3. L37-38: Is there a latitude dependency for this peak? I would imagine that the timescale isn't also 4 days near the equator, but I may be wrong

Reply: The location of this peak can change from place to place and depends on the local climate I presume. The spectrum of Kang et al. (2016) measured over Boulder for instance does not feature a on timescales longer than the diurnal cycle.

4. L41/42: Consider citing the review papers of Stevens and Meneveau (2017) as well as the Porté-Agel et al. (2020) review paper

Reply: These references are indeed quite monumental papers in this research area, and are added to the manuscript.

5. L96-97: You should cut this statement. If you wish to retain it, please consult the ERA5 wind energy validation that was done as part of NEWA and the Olauson (2018) paper

Reply: This statement has been left out of the manuscript.

6. L109: Why use RSMD instead of bias? I feel like every wind validation paper I have seen has used bias, not RSMD

Reply: We used the RMSD here since this metric somewhat takes the shape of both distributions into account. Two distributions with quite different means, but similar means will have a low bias.

7. L128: You don't need to change this in the paper, and this is more for my education: do COSMO researchers talk about "periodograms"? In WRF, we call them spectra, though I suppose periodogram is more correct

Reply: I don't have the impression that periodogram is specifically used by the COSMO community. As far as I understand a periodogram is an estimate of the underlying spectrum of a signal. For real signals it’s impossible to know the underlying spectrum of a signal, due to finite time series and sample frequencies and so on. It is similar to a sample mean, which is an estimate for the full population mean. I therefore think the term periodogram is better. I do think that in the climate community the terms spectrum and periodogram can be used interchangeably, so we added the following line to the manuscript.

The spectral density of a signal is estimated using a periodogram, also referenced to as a spectrum, calculated with the Welch method.

8. L131: Why use a window of 7 days? Could you put that into context of the mesoscale timescales you're interested in? As an aside, thank you for giving all these details on your FFTs, because people often neglect to mention these important details.

Reply: The 7-day window is used primarily for including (an indication of) the synoptic weather peak in the spectrum. The 1024 output intervals are not exactly equal to seven days, but in general FFT algorithms work most efficient when the length of the input vector is equal to a power of two.

In the Welch method the 10-year time series of 10-minute interval wind speeds is cut using a Hann window in overlapping sections of approximately seven days (1024 output intervals) in our case, with an overlap of 50%. This window length allows for part of the synoptic peak in periodogram to be seen.

9. L147-149: If you integrate the periodogram over all bins, that's just the TKE, right? If so, maybe mention here that you take a spectral approach because you can then focus on specific scales (which would be harder to do in the time-domain)

Reply: A statement addressing this has been added to the manuscript.

These averaged periodograms can be integrated over a time interval of interest resulting in one single value per grid point that quantifies the temporal variability over that time interval. This makes it possible to visually compare the different grid points for each season and allows us to focus on specific scales, which would be challenging in the time-domain.

10. L151: power curve

Reply: This has been changed in the manuscript.

11. L159 and 165: I recommend the authors state that "We define the MSVI..." and "We define the size of the small window...". When I read these sentences, I got the impression that some other paper specified these definitions, but I believe the MSVI is invented here.

Reply: It has been made clearer that we constructed the MSVI metric.

12. L191-192: This statement about the double-penalty seems very hand-wavey

Reply: An explanation of the double penalty and why this metric might make a climate model appear worse than it actually is, is added to the methods section. It’s hard to estimate the effects of this double penalty, and it is therefore not possible to attribute the RMSD of 1.35 m/s to this double penalty alone. This has been added to the manuscript.

Moreover, the 1.35 m/s RMSD may be explained by the double penalty mentioned earlier, but there is no straightforward way of testing this. However, tests with spectral nudging did not substantially improve the performance indicating that the lateral boundaries to a large extent control the timing and location of weather systems for this domain and model configuration.

13. L212-214: I strongly disagree with this statement. COSMO may underpredict winds in simulations without turbines, and the wakes on the lidar would conveniently also lower the observed wind speed.

Reply: A figure plotting the availability of the Anholt wind farm has been added to the appendix. Here it can be seen that the wind farm availability over the first 100 days is much lower than over the rest of the measurement campaign. The low difference in bias and PSS between these two periods despite the large difference in availability indicates that the impact of the Anholt wind farm on the lidar is not that large. This does not mean that the Anholt wind farm has no impact on the atmosphere, but thanks to the wind blowing mainly from the west the lidar is often unwaked and this dataset is still quite useful for validation. A few lines on this have been added to the results of our manuscript.

Even though there is some uncertainty about the effect of the wind farm on the lidar data, the small difference in performance between the two periods, together with the close correspondence between lidar data and COSMO gives us some confidence that the model performance is adequate. It might however be possible that COSMO under predicts winds in the simulation without turbines, which is masked by the wind farm wakes experienced by the lidar. However, the effect is likely small, due to the wind mainly blowing from the west in this area (Karagali et al., 2013).

14. L224-225: I appreciate that you conduct statistical testing. Is this test done to 95% confidence?

Reply: Yes it is, and this has been added to the manuscript.

Using a Student's t-test we find that the differences over these time slots are significant at the 0.05 level.

15. L289: Is an RMSD of 1.35 m/s "good agreement"? Relative to what? Either compare to other papers or reword

Reply: This is comparable to the RMSD that for instance Wang et al. 2019 found.

The simulation showed good agreement with scatterometer observations away from coasts and small islands with a spatiotemporal root mean square difference of 1.35~m/s, which is comparable to for instance Wang et al. (2019) found.

16. L298: I thought the short-timescale variability came from SST/air temperature differences, not convective systems?

This statement is indeed a bit confusing. The gradient in the short-timescale variability implies that there is a certain spin-up period for the variability to reach its full potential, during which it is advected over the Kattegat. The variability at this timescale is therefore probably related to the unstable conditions over sea offshore, and unstable conditions result in more convective systems. This has been clarified in the manuscript.

Our results show that more variability in wind speed is expected in winter due to unstable conditions over sea. These unstable offshore conditions result in increased turbulence and induce convective systems, which generate wind speed variability on short timescales (20 minutes to 1 hour).