Reviewer 2:

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.

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 some references here and also mention the wind atlases. In our paper we perform some analyses which add to the understanding of factors that are relevant for mesoscale wind speed variability. We hope that rewriting part of the introduction results in a better framing for the paper.

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.

We have set up our own simulation of this area since this gives us more freedom regarding the generation of output data. This is mainly important for the temporal variability analysis, which uses 10-minute interval output data. We also believe that the spatial and temporal analysis methods used in our paper aid the understanding of factors driving mesoscale variability. In addition to this we also plan on using this simulation setup for further studies, for which control of the model setup is required. As this last point is not relevant for this study it is (in contrast to the other points) not added to the manuscript.

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.

The simulation timestep is 10 seconds. The output frequency differs from variable to variable, but for the wind speed on specific height levels (10m, 80m and 100m) it is set to 10 minutes. For practical reasons the calculation of the 10-year simulation has been divided into smaller periods, but using the restart files generated by COSMO these periods were initialised with a warm start. Subgrid-scale turbulence is parametrised by a one-dimensional diagnostic level 2.5 closure scheme based on a prognostic TKE equation (Raschendorfer, 2001; Schulz, 2008). This is the standard PBL scheme used in COSMO-CLM. A file containing the configuration of the model will be added to the Zenodo, and made available to the public.

This information is added to the model setup section

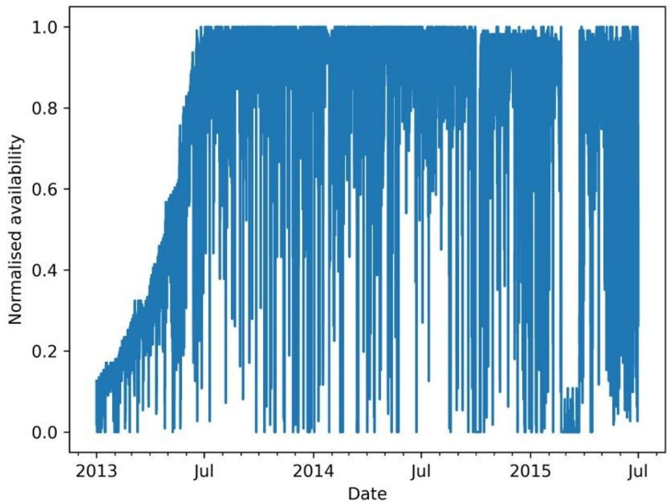
Validation efforts: Researchers often validate modeled winds against measured winds in order to built trust, but the validation study here instead erodes my confidence. I also don't see how validation furthers the authors stated goals in the intro. I believe that all the lidar analysis should be struck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

The validation with lidar data is included in the paper in order to validate the wind speeds at 100 metre, as we mainly use these wind speeds in the paper. The lidar data is also a nice addition to the ASCAT data, as 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.

The lidar is positioned 2 kilometres west from the Anholt wind farm. The wind farm was under construction during the measurement campaign, and the availability of the wind farm is shown in the figure hereunder (this figure is added to the appendix). The first 100 days of this measurement campaign see a rise in wind farm availability, but this availability stays under 50%. As there is not a large difference in model performance compared to lidar data for the first 100 days, relative to the full dataset, while there is a large difference in wind farm availability means that the impact of the wind farm on the lidar data is not that large. This does not mean that the Anholt wind farm does not have an impact on the surrounding atmospheric conditions, but that due to the wind mainly blowing from the west in this area (Karagali et al. 2013), resulting in a lidar that is largely unwaked.

This discussion has been added to the manuscript.



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.

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. A statement about this has been added to the manuscript.

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.

A direct comparison between ASCAT and model output 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 (yet) (G. -J. Marseille and A. 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 compared. This statement has been added to the 2.2 section explaining the methods used in the validation of the model with scatterometer data.

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.

This corresponds to approximately 20,800,000 grid cells of our COSMO model, and this statement has been added to the manuscript.

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).

We have added to the paper that the ASCAT device only measures the near surface wind speeds at 9h UTC and 21h UTC and that the lidar validation has been added to the paper to validate the model further at 100m.

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.

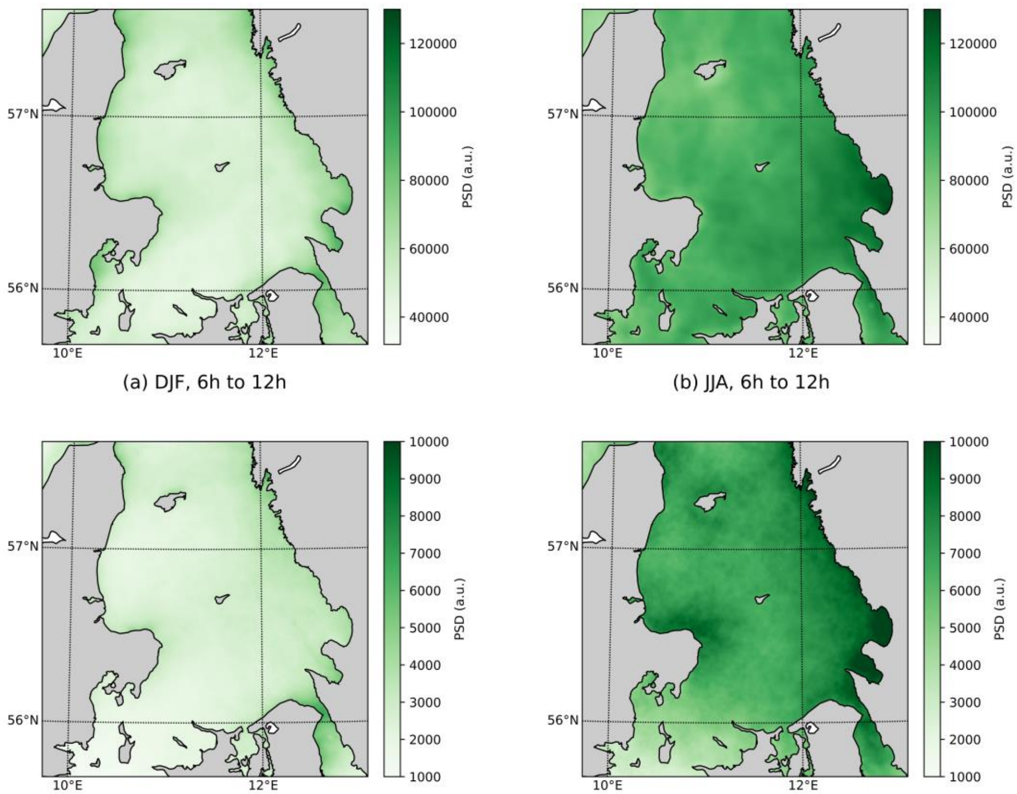
Relatively warm SSTs result in turbulence and unstable conditions. Unstable conditions are a driver for convective cells. The systems are not bound to a specific place and are advected with the wind. Therefor they can move relatively fast and can result in short timescale variability. Sea breezes on the other hand 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.

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.

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.

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.



(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.

A statement about this analysis has been added to the introduction of the paper.

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.*

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.*

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*

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*

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*

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*

   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*

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 do think that in the climate community the terms spectrum and periodogram can be used interchangeably, and I'm certain that I have called a periodogram a spectrum and vice versa. As spectra is a better known term, I guess it's better to use that term, but since I've given already too many presentations calling them periodograms I think I'll keep calling them that way.

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.*

   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.

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)*

A statement addressing this has been added to the manuscript.

10. *L151: power curve*

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.*

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

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

A 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.

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.*

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.

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

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

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

This is comparable to the RMSD that for instance Wang et al. 2019 found. References to other papers have been added to the manuscript.

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.