The authors present a simulation study of the Perdigao field campign, using WRF with 5 nested domains, with the three finest being in LES mode. Comparisons with the field measurements are discussed for two selected cases, i.e. one stable and one convective. Overall, this simulation study can be of interest to the community. However, as it is presented now, I do not believe it should be accepted for publication. Results are poorly discussed, with a lot of hand waving and qualitative comparisons of instantaneous fields, the paper is overly long, and lacks detail in many instances (see further comments below). The author's see the merit of their study mainly in the fact that they are the first to use realistic inflow conditions for LES of the Perdigao field campign, but that is not strictly true, since Wagner et al (Atmos. Chem. Phys, 2019) also presented a nested simulation, albeit with only three levels, with only the finest LES. Given all this, I recommend to thoroughly rewrite this paper, and better analyze results.

We thank the associate editor for the time spent reviewing our work and for the detailed suggestions and comments. Our study is innovative in that it is the first to nest large-eddy simulations down to 10-m resolution, forced with mesoscale simulations. Further, we use these new simulations to examine turbine wake behavior in highly complex terrain. Current wake models do not account for the flow effects of complex terrain when designing wind farms. At 10-m resolution, we can parameterize the wind turbine rotor and analyze the wake's interaction with turbulent terrain-induced flow features under different stability conditions. To the authors' knowledge, there is no previous study that combines highly complex terrain, various atmospheric stabilities, and a wind turbine rotor parameterization at such high resolution. The study by Wagner et al. 2019 analyzes flow characteristics and low-level jets at a grid resolution of 200 m, over an order of magnitude coarser resolution than used in our work here. The wind turbine's rotor diameter is 80 m and requires a minimum of 10 m grid spacing. We view the merit of our study as an examination of the WRF-LES-GAD configuration's ability to capture both relevant and complicated meteorological and terrain impacts, as well as the impacts of those flow regimes on operating turbines, using the Perdigão site as the test location. These points have been added to the manuscript through responses to specific comments, as detailed below. Note that manuscript has decreased from ~10,000 words to under 8,000 words after these major revisions.

Here is a list of the major revisions made to the manuscript:

- The abstract has been reworded to better frame the motivation and to include justification for the two chosen cases.
- Parts of the introduction have similarly been rewritten to better frame the motivation
- Section 2.2: the Froude number has been defined as the internal Froude number. An average internal Froude number has been calculated and the justification for the two cases has been related to the Froude number.
- The semi-idealized section and any associated discussion have been removed
- Sections 4.1 and 4.2 have been renamed as Validation of the Stable Case Study and Validation of the Convective Case Study, respectively.
    - In these sections, we have interpolated the DLR lidar data onto the same grid as the model, which allowed us to add a third panel to this figure to show the difference between the model and the observations, and additional discussion has been added.
    - We also removed many non-objective statements and tried to be as quantitative as possible.
    - Discussion, in general, has been reduced or removed and this includes comparisons with soundings. Figure 12 (previous submitted manuscript numbering) has also been removed.
- Section 4.3

  - Figure 16 (previous submitted manuscript numbering) has been time-averaged and the associated discussion has been modified.
  - Figure 19 (previous submitted manuscript numbering) now includes experimental results using DLR and DTU lidars. We have also included error bars for the model results and measurements in this figure.
- We have added a section in the appendix called: Sensitivity of Model Results to Grid Resolution. In this appendix, we discuss how errors and biases for mean flow fields are not necessarily reduced by adding layers but that more fine-scale turbulence is able to be resolved (spectra are shown). We also show time-averaged transects highlighting how recirculation depends on the terrain that is able to be resolved on the grid.

Comments

1. I do not see the point of the semi-idealized runs. It is obvious that this will not work with periodic BCs. There is also not much connection to the rest of the paper. This has probably been a useful internal verification exercise, but is not worthwhile reporting on. I suggest to remove this part.

Thank you for the suggestion. The semi-idealized section and any associated discussion have been removed.

2. Discussion is qualitative at best and seems often overly optimistic when reporting own results. For instance: "As seen in Fig. 12, the flow inside the valley and near the two ridges is highly dynamic but very accurately predicted by …" Such a statement can not be simply based on instantaneous snapshots such as shown in Figure 12. Overall, the paper is very lengthy in discussing a lot of these instantaneous figures (e.g. 6, 7, 11, 12, 16, 17). This part could be cut significantly

We thank the reviewer for bringing this to our attention. The manuscript's discussion of instantaneous figures has been trimmed significantly. Qualitative discussion in general has been reduced. Specifically (using figure numbers from the previous submission):

- For Figure 6, subpanels (c) and (d) have been removed and a paragraph of discussion has been removed
- The discussion for Figure 7 has been significantly reduced. Additionally, we have added a brief examination on the sensitivity of model results to the uncertainty in the TLS positioning
- Figure 12 and the associated discussion have been removed
- Figure 16 has been time-averaged and the discussion has been modified.

3. Although there are various mean-field results are shown, I would expect a more careful discussion, and in particular figures showing comparisons between 1h averages simulation and experiment along selected vertical and/or horizontal transects (= a line), including error bars (e.g. using bootstrapping). A side by side comparison of color plots in vertical planes is not really very informative (e.g. Fig. 10, 15, 18). Why are experimental results not added to Figure 19?

Figures 10 and 15 (now Figures 5 and 9) have been modified. The DLR lidar data have been interpolated onto the same grid as the WRF model output. This allows us to take a difference between the simulations and the measurements. In both Figure 10 and 15 (now Figures 5 and 9), a third sub panel has been added which shows the difference in the along-transect velocities. Brief discussion on the differences between the simulations and measurements has been added to the manuscript as well.

Figure 19 (now Figure 14) now includes experimental results by taking the difference between the DLR and DTU lidar scans. In this figure, we have also included error bars as +/- 1 standard deviation over the hour used for averaging to provide insight into uncertainty.

4. Figure 7 (lower panel), Figure 8, Figure 13 – this seems to be the long-time evolution of metmast measurements. How much of this is already contained in the d01 and d02 levels. In general, the authors could do a much better job discussing the added benefit of adding LES layers to the prediction quality of their results. E.g. can you quantify a significant reduction of bias per added layer, etc.

Thank you for the suggestions. We have added a subsection to the appendix titled "Sensitivity of Model Results to Grid Resolution" which examines the wind speed outputs from the model on d03 (dx = 150 m), d04 (dx = 50 m), and d05 (dx = 10 m) for the convective and stable case studies. We have added figures (Figure A1 and A3) that show time-series and spectra for all three LES domains and the metmast measurements. Tables A1 and A2 look at the biases and RMSE for the time-series for all three domains. Biases and errors are not significantly reduced as the grid resolution increases; however, the resolved wind speed spectra are different with only d05 able to capture similar smaller-scale energy-containing eddies as in the metmast measurements.

The long-time evolution of the met mast measurements are not captured on d01 or d02 because of the coarse grid resolution. With a grid spacing of 6.75 km or 2.25 km for d01 and d02, respectively, the double ridge in the terrain is not resolved. Additionally, the ridge-to-ridge distance is ~1400 m so all three met towers would be within a single cell. However, after we nest to finer grid resolutions, the terrain is more accurately resolved.

Another figure is added, Figure A2, that looks at time-averaged vertical transects of the wind speed across the rotor plane during the convective case. Reverse flow or recirculation is only present on the finest domain, d05, because it is able to resolve the steeper slopes in the terrain. Figure A4 is the same figure but for the stable case and we see that the large scale dynamics do not require a 10 m grid, but the wake and turbulence dynamics do.

5. Discussion on Brunt-Vaisala frequency and gravity waves throughout the paper is very vague and imprecise. In particular: p6 – there are inversion waves and internal waves. Both are determined by a different Froude number (the first based on the inversion step, the second based on the lapse rate). Both relate to phenomena above the ABL. Please properly discuss and define, incl. height and way in which you measure the Fr numbers, velocity scale, length scale (also e.g. in Figure 3). P 12-13. Where/how do you measure the wavelength of the mountain wave event. This seems to be the horizontal wavelength. P 17. The horizontal wavelength of the wave is not influenced by the Froude number, but forced by the terrain feature, only the vertical wavelength is influenced by Fr number. P. 19 Discrepancy between vertical wavelength in model and measurements seems to be related to the fact that your simulation domain and Rayleigh damping layer is not sufficiently high. Properly discuss and justify your set-up or discuss limitations

Thank you for the suggestions. To clarify, the Rayleigh damping layer in the multi-scale configuration is for the top 5 km of the domain and the domain height is approximately 20 km. This has been added into the manuscript at lines 196 and 216. The Rayleigh damping layer may not have been sufficiently high in the semi-idealized configuration; the idealized portion of the manuscript has now been removed at the reviewer's suggestions.

The discussion has been modified significantly. Section 2.2 now reads as follows (lines 124-132):

> "Mountain waves can occur when stably stratified flow approaches a topographic disturbance, such as a mountain ridge. These waves can be described using a ratio of inertial to buoyant forces represented by the internal Froude number (Stull, 1988):
>
> $Fr = \pi U/(WN)$
>
> Where W is the mountain ridge width (defined for the southwest ridge of Perdigão to be 586.5 m on average according to Palma et al. (2020), U is the free stream wind speed and N is the Brunt-Väisälä frequency, defined as:
>
> $N = \mathrm{sqrt}(g/\theta\ d\theta/dz)$
>
> Here g is the gravitational constant, θ is the potential temperature of the environment, and dθ/dz is the lapse rate or vertical gradient in potential temperature. The internal Froude number can also be defined as the ratio of the natural wavelength of the air ($\lambda = 2\pi U/N$) to the effective wavelength of the mountain ridge (2W)."

Figure 3 (previous submission numbering), which had shown the time-series of the Brunt-Vaisala frequency and Froude number has been removed. Section 2.2.1 reads as follows (lines 164-167):

> "...using the average wind speed (6.3 m/s), temperature gradient or lapse-rate (0.031 K/m), potential temperature of the environment (296.7 K), and width of the southwestern ridge of Perdigão (586.5 m), the average internal Froude number during the period of interest is calculated to be 1.05. With a Froude number close to one, we expect a resonant mountain wave to occur."

Relevant discussion in Section 4.1 reads (lines 239-243):

> "The wavelength of the mountain wave is defined as the distance from the first ridge (where the low-level jet begins to deform) to the first crest of the mountain wave. The model predicts a wavelength of 1220 m (Fig. 4(a)), 13% less than the 1410 m wavelength from the DTU lidars, which is almost exactly the ridge-to-ridge distance of 1400 m (Palma et al., 2020)."

The soundings and any associated discussion has been removed. This includes discussion about the Brunt-Vaisala frequency and Froude number, which has been replaced by the new text above.

6. page 8: why these particular cases. Better justify

Previous to Page 8, in Section 2.2, we mention (lines 118-122):

> "Two terrain-induced flow features characteristic of the field site are of interest for the present work: (a) recirculation zones and (b) mountain waves. These flow features may occur at the Perdigão site depending on the time of day (Menke et al., 2019; Fernando et al., 2019; Wagner et al., 2019) and are characteristic of the site, in general. Fernando et al. (2019) found that mountain waves occurred for almost 50% of the nights during the IOP while Menke

et al. (2019) found that recirculation occurred over 50% of the time when the wind direction was perpendicular to the ridges."

These particular cases are chosen because either a recirculation zone or mountain wave is expected to occur and those phenomena were commonly observed during the field campaign. We relate the justification of the cases to the expected behavior of the flow based on the internal Froude number. In Section 2.2.1 (Case 1: Convective Conditions), we have added the following sentence on lines 172-173:

> "The negative lapse rate results in an infinite internal Froude number and with wind perpendicular to the ridge a turbulent mountain wake with recirculation should form."

In Section 2.2.2 (Case 2: Stable Conditions), we have added the following sentences on lines 164-167:

> "Using the average wind speed (6.3 m s-1), temperature gradient or lapse-rate (0.031 K m-1), potential temperature of the environment (296.7 K), and width of the southwestern ridge of Perdigão (586.5 m), the average internal Froude number during the period of interest is calculated to be 1.05. With a Froude number close to one, we expect a resonant mountain wave to occur."

Additionally, the abstract has been modified to include the following sentence on lines 4-7:

> "Two case studies are selected to be representative of typical flow conditions at the site, including the effects of atmospheric stability: a stable case where a mountain wave 5 occurs (as in 50% of the nights observed), and a convective case where a recirculation zone forms in the lee of the ridge with the turbine (as occurred over 50% of the time with upstream winds normal to the ridgeline)"

7. page 12: why set the surface roughness to 0.5m. This seems rather arbitrary. Did you fit this value to obtain the best results?

This is explained in lines 225-229 which now read:

> "The CORINE dataset seemingly misclassifies the land type in the valley as mixed shrub- land/grassland when the vegetation is mostly tall eucalyptus and fir trees. Likewise, Wagner et al. (2019) concluded that the surface roughness lengths at the Perdigão site based on the CORINE Land Cover data were too small. To account for this, we set the surface roughness length for the mixed shrubland/grassland land use category in the valley to 0.5 m, the same value used in the LES studies of Berg et al. (2017) and Dar et al. (2019)."

8. page 13, line 290. This is a very qualitative statement. Also in earlier work (e.g. Berg et al), agreement was quite remarkable. Please try to be objective, and focus on verifiable facts

Thank you for bringing this up. We have removed this sentence.

9. Eq. 3, 4: better explain that these eqs are only valid for the surface layer. Where, at which height to do you measure them?

The manuscript states that Eq. 3 and 4 are calculated using data from the tower on the southwest ridge at a height of 10 m (the lowest sensors available). We have added to the manuscript that the equations are only valid for the surface layer. The sentence now reads on lines 151-152:

> "The Obukhov length and friction velocity, valid for the surface layer, are calculated at SW_TSE04 using 5 minute statistics from the lowest sensors available at 10 m. "

I thank the authors for carefully revising the manuscript. All comments raised during the interactive discussion have been taken into consideration and addressed in the author's response. In some places, however, I found that the clarifications stated in the author's response could have been transferred more to the primary manuscript. The information will remain available as the author's response will be published together with the primary manuscript upon final acceptance, so I don't think this calls for another round of reviews. However, I strongly suggest to go over the author's response once more to make sure that all additional information regarding numerical parameters, relevant references, more detailed explanations, reasoning behind certain choices etc. stated in the author's response can also be found in the primary manuscript if it is in any way relevant for the clarity of the manuscript or the reproducibility of the simulations or the postprocessing.

We thank the reviewer for the time spent reviewing our work and for the detailed suggestions and comments in their first report. From the previous review, we have made major revisions to the manuscript. All of the major revisions are listed below; however, the most relevant changes related to the reviewer's previous response are summarized here:

1. The Froude number, Equation 1, is now defined as the internal Froude number, which is the ratio of the natural wavelength of the air to the effective wavelength of the mountain ridge (From Stull 1988 Chapter 14). This equation is appropriate for the horizontal wavelength and we apologize for misunderstanding the reviewer's previous comments.
2. The semi-idealized section and any associated discussion has been removed, which should alleviate any confusion that there may have been between the two model set ups and configurations.
3. Vertical profiles of 1 h time-averaged along-transect velocity profiles are now included, comparing the model output with lidar data. Error bars +/- 1 standard deviation aim to address uncertainties in the comparison.

Below is a comprehensive list of the major revisions made to the manuscript:

- The abstract has been reworded to better frame the motivation and to include justification for the two chosen cases.
- Parts of the introduction have similarly been rewritten to better frame the motivation
- Section 2.2: the Froude number has been defined as the internal Froude number. An average internal Froude number has been calculated and the justification for the two cases has been related to the Froude number.
- The semi-idealized section and any associated discussion have been removed
- Sections 4.1 and 4.2 have been renamed as Validation of the Stable Case Study and Validation of the Convective Case Study, respectively.
    - In these sections, we have interpolated the DLR lidar data onto the same grid as the model, which allowed us to add a third panel to this figure to show the difference between the model and the observations, and additional discussion has been added.
    - We also removed many non-objective statements and tried to be as quantitative as possible.
    - Discussion, in general, has been reduced or removed and this includes comparisons with soundings. Figure 12 (previous submitted manuscript numbering) has also been removed.
- Section 4.3
    - Figure 16 (previous submitted manuscript numbering) has been time-averaged and the associated discussion has been modified.

- Figure 19 (previous submitted manuscript numbering) now includes experimental results using DLR and DTU lidars. We have also included error bars for the model results and measurements in this figure.

- We have added a section in the appendix called: Sensitivity of Model Results to Grid Resolution. In this appendix, we discuss how errors and biases for mean flow fields are not necessarily reduced by adding layers but that more fine-scale turbulence is able to be resolved (spectra are shown). We also show time-averaged transects highlighting how recirculation depends on the terrain that is able to be resolved on the grid.

Reviewer general comment: The current manuscript presents results of a suite of WRF-LES configurations around the complex topography of the Perdig~ao field site. Results represent the first time that WRF-LES is used to study the flow interaction in complex terrain and wind turbines. Results of the numerical simulations are qualitatively compared to the experimental measurements. According to the authors, the features of interest for the present work are: (a) mountain waves, (b) recirculation zones, and how these interact with a wind turbine and the corresponding wake as a function of atmospheric thermal stratification.
The manuscript is very nicely written, clear, and with superb "candy-to-the-eye" type figures. Also the a priori aimed research question is of interest, and providing a thorough analysis of the problem posed would be of high interest to the community. Unfortunately the scientific analysis and content remains largely the same as the original version, mostly based on qualitative metrics and comparisons of numbers. The truth is that review of this manuscript has become quite a frustrating experience (hence in part the long delay in the review; my apologies for that), given that the authors instead of carefully considering the constructive suggestions and comments made, have instead taken a defensive approach. Review of scientific manuscripts is a volunteering act in support of the community. I don´t do it with any aim to go against anyone, or their work. I mostly do it to keep learning, and try to provide my opinion for others to consider and stir some additional thinking. As a critical scientist with myself, and with the community, I would reject publication of the manuscript as is. However, I also realize that this might be a controversial decision, hence I rather leave the community to judge it by themselves, and time will tell whether the work herein presented is worthy of publication. Beyond this rather important general comment, below I provide a few additional comments, that if the authors are willing to consider, could probably strengthen their manuscript.

We thank the reviewer for the time spent reviewing our work and for the detailed suggestions and comments on how to strengthen our manuscript.

First, we would like to provide a list of the major revisions made to the manuscript:

- The abstract has been reworded to better frame the motivation and to include justification for the two chosen cases.
- Parts of the introduction have similarly been rewritten to better frame the motivation
- Section 2.2: the Froude number has been defined as the internal Froude number. An average internal Froude number has been calculated and the justification for the two cases has been related to the Froude number.
- The semi-idealized section and any associated discussion have been removed
- Sections 4.1 and 4.2 have been renamed as Validation of the Stable Case Study and Validation of the Convective Case Study, respectively.
    - In these sections, we have interpolated the DLR lidar data onto the same grid as the model, which allowed us to add a third panel to this figure to show the difference between the model and the observations, and additional discussion has been added.
    - We also removed many non-objective statements and tried to be as quantitative as possible.
    - Discussion, in general, has been reduced or removed and this includes comparisons with soundings. Figure 12 (previous submitted manuscript numbering) has also been removed.
- Section 4.3
    - Figure 16 (previous submitted manuscript numbering) has been time-averaged and the associated discussion has been modified.
    - Figure 19 (previous submitted manuscript numbering) now includes experimental results using DLR and DTU lidars. We have also included error bars for the model results and measurements in this figure.
- We have added a section in the appendix called: Sensitivity of Model Results to Grid Resolution. In this appendix, we discuss how errors and biases for mean flow fields are not necessarily reduced by adding layers but that more fine-scale turbulence is able to be resolved

(spectra are shown). We also show time-averaged transects highlighting how recirculation depends on the terrain that is able to be resolved on the grid.

More Specific Technical Comments

1. Reviewer comment: In line 9-10; it might be appropriate to try to justify why the authors have selected to two study the mentioned periods. Are they relevant to something, or representative of something? Or is just a random selection. Providing this additional info my help motivate. This also relates to the text in Line 14-15. Are the authors expecting the dependence on the wake behavior to be continuous and smooth, or illustrating a sharp change in behavior at a certain thermal stratification?

We have added in the abstract at lines 4-7 why we have chosen the two selected studies:

> "Two case studies are selected to be representative of typical flow conditions at the site, including the effects of atmospheric stability: a stable case where a mountain wave 5 occurs (as in 50% of the nights observed), and a convective case where a recirculation zone forms in the lee of the ridge with the turbine (as occurred over 50% of the time with upstream winds normal to the ridgeline)"

We have also related the justification of the cases to the expected behavior of the flow based on the internal Froude number (Equation 1). In Section 2.2.1 (Case 1: Stable Conditions), we have added the following sentences on lines 164-167:

> "Using the average wind speed (6.3 m s-1), temperature gradient or lapse-rate (0.031 K m-1), potential temperature of the environment (296.7 K), and width of the southwestern ridge of Perdigão (586.5 m), the average internal Froude number during the period of interest is calculated to be 1.05. With a Froude number close to one, we expect a resonant mountain wave to occur."

In Section 2.2.2 (Case 2: Convective Conditions), we have added the following sentence on lines 172-173:

> "The negative lapse rate results in an infinite internal Froude number and with wind perpendicular to the ridge a turbulent mountain wake with recirculation should form"

We expect the wake behavior to be largely governed by the background flow and ambient turbulence, which in turn depends on the thermal stratification and wind speed. Determining continuous vs abrupt change in wake behavior would require more simulations closer to the transitions. This initial work demonstrates two distinctly different cases that are highly representative of this field site and future work can explore the transition between them. We would expect WRF-LES-GAD to be able to capture the full range of flow conditions in different cases and associated effects on the wind turbine wake.

2. Reviewer comment: In line 18-20; 'This study demonstrates the ability of the ...'; I wonder has this been questioned in the literature? If the WRF model has been used and validated in a continuous manner through the years, and the GAD model has also been developed and tested (as mentioned by the authors in line 61 to 63 in the text), where is the need for another comparison? Have the authors introduced any new element in the WRF platform that requires testing and verification? – This doesn't seem to be the case, but maybe I missed it Again.

The WRF model at the mesoscale has been used and validated in a continuous manner; however, the LES capability is becoming more widely used and is in need of validation, especially in cases where terrain effects are important.. Additionally, WRF-LES-GAD has been previously validated for flat and simple terrain but has yet to be tested for complex terrain.

We motivate our study based based on the following:

- An examination of the WRF-LES-GAD configuration to capture both relevant and complicated meteorological and terrain impacts, as well as the impacts of those flow regimes on operating turbines,
- The literature disagrees on vertical wake deflection in different stability conditions
- Terrain-induced flow phenomena and their effects on turbine wake behavior are not typically accounted for in turbine wake models used for designing wind farms. LES can be used to provide guidance to these lower-fidelity tools.

We feel that this demonstration and validation of the WRF-LES-GAD model for two distinct stability classes shows its utility for simulating wakes over a range of stability conditions. However, because the computational expense of running additional simulations in other stability conditions (e.g. neutral, slightly stable, etc.) is prohibitive, we have chosen to focus on two representative cases that are similar to those commonly observed during the field experiment.

3. Reviewer comment: In line 55; ‘and more realistic turbulence compared to’; this is not necessarily true. Why would the turbulence initiation method based on the cell perturbation method provide more accurate turbulence than that on an idealized LES simulation? Maybe, and only maybe, under certain atmospheric conditions where very large scale forcing is of relevance, affecting the near surface turbulence, this could be true. However, this is not necessarily Generalizable.

The quoted sentence does not directly relate to CPM or the turbulence initiation method. Rather, it refers generally to our multi-scale nested setup. To be more clear, the sentence has been rewritten in lines 41-42 as follows:

> “Such setups can provide LES with realistic time-varying inflow conditions directly from the mesoscale simulations.”

4. Reviewer comment: In line 74-75, the objective is nicely stated, which is great. It would be nice if the conclusions came around and provided an answer to this fundamental questions. Unfortunately I have the impression this is not the case, since the study only provides information for two independent stability cases. For example, the authors could also follow up on the results/hypotheses from previous works cited by the authors in line 80-81, and explore self similarity of the turbine wakes, in this case as a function of varying thermal stratification. This would add some interesting generalizable science, beyond the case specific observations.

As mentioned in the response to the second comment, this demonstration and validation of the model at two distinct stabilities shows that the WRF-LES-GAD model can be useful for simulating wakes over a range of stability conditions. However, the computational expense of these simulations prohibits running a large number of cases. Thus, we chose two cases that aim to be representative of commonly observed conditions. On lines 193-194, we have added the following note:

> “the current setup takes roughly eight hours of wall-clock time for five minutes of simulation time on 288 cores”

Additionally, the goal of the work has been reframed and added to the manuscript on lines 78-80:

> "The goal of this work is to examine the ability of the WRF-LES-GAD framework to capture terrain-induced flows and their interaction with an operating wind turbine at the Perdigão site (described further in Sect. 2), to thereby demonstrate the efficacy of mesoscale-to-microscale coupled simulations to improve wind plant simulations in complex terrain."

5. Reviewer comment: In lines 87-88, why do the authors want to use idealized LES? – This hardly makes any sense for two reasons: 1. the authors mention from the beginning that they are interested in real conditions; 2. they run the simulations, show two instantaneous snapshots, and don't use that data any further. What is the point? I think the manuscript would be a lot simpler just using directly the realistic WRF-LES platform.

Thank you for the suggestion. The semi-idealized section and any associated discussion has been removed.

6. Reviewer comment: In line 86, 'based on atmospheric stability'; this statement is a bit generous since the authors only consider 2 different stratifications. This is equivalent to that experimentalist that goes to develop a field experiment and only takes two data values to study a complex problem. One would expect at least three points, to discard the obvious linear fit, isn't it? Maybe saying something about the neutral stratification case to complement? If at the neutral stratification regime results are similar to those observed in convective conditions, what is the intensity on stable stratification needed for the wake to start changing behavior? – These are the kind of research question I would have expected to get resolved or studied in this manuscript as I read the introduction.

While we appreciate the reviewer's desire for a description of how wakes change with atmospheric stability, our work shows two distinct, disparate, complex regimes of behavior in the atmospheric stability regimes experienced at this site. A smooth trend between these two regimes is not necessarily expected. As an analogy, smoke dispersion in a convective boundary layer is fundamentally different from that in a stable boundary layer, and no smooth transition between those regimes is observed. Neither would such a trend be expected here, as we assess the dispersion of a wind turbine wake (which is not as passive as a smoke plume) in an atmosphere influenced by complex terrain. Furthermore, neutral stratification rarely occured at this site during the experimental period (Menke et al. 2020). Specifically, here we analyze how the vertical deflection and dissipation of the wake varies in two distinct but representative atmospheric stability regimes, as discussed in the response to #1 above.

7. Reviewer comment: In line 161, note that the additional figure sent in the review seems to include periods of counter gradient fluxes (maybe my eyes are betraying me,...), but if this was the case a brief comment could be added, mentioning that those time periods have been discarded for example.

We have added the following sentence to the manuscript at line 157-158:

> "The met tower measured periods of counter-gradient heat fluxes during both stable and convective conditions but not during the time periods selected for this tower location."

8. Reviewer comment: In line 206-207, when talking about the cooling rate, I interpret that being at the surface, is that correct? Mind you it could also be at the top boundary condition. Maybe it is worth clarifying this, is you decide to keep the text about the idealized LES.

That is correct; however, we have removed the semi-idealized section.

9. Reviewer comment: In line 209-210, how is the turbulence initialized in the ideal LES cases? To that end, what is the turn over time of the flow around the periodic domain? Does the turbulent flow at least have time to properly develop before reaching the mountain? Does the turbulence reach some sort of convergence prior to reaching the region of interest, meaning the mountains?

These are all great questions but we have removed the semi-idealized section to improve the readability of the manuscript, as suggested above.

10. Reviewer comment: In line 219, when discussing the surface roughness. If the authors have access to wind lidar data and three tall met towers, why don't they use the experimental data to extract an approximate z0 value? Regardless of how challenging the experimental data might be, at least the value will be of higher significance than that chosen by the authors (which doesn't seem to have much justification). Unless there is a stronger reason behind that decision that is not explained in the manuscript.

We appreciate the reviewer's suggestion. We now more thoroughly explained the surface roughness on lines 225-229:

> "The CORINE dataset seemingly misclassifies the land type in the valley as mixed shrub- land/grassland when the vegetation is mostly tall eucalyptus and fir trees. Likewise, Wagner et al. (2019) concluded that the surface roughness lengths at the Perdigão site based on the CORINE Land Cover data were too small. To account for this, we set the surface roughness length for the mixed shrubland/grassland land use category in the valley to 0.5 m, the same value used in the LES studies of Berg et al. (2017) and Dar et al. (2019)."

11. Reviewer comment: In line 224-225; 'Having demonstrated ..'; once again, I wouldn't feel comfortable myself making that statement. The authors have only showed two beautiful colored figures showing reasonable results, but there is no strict 'demonstration'.

This sentence has been removed along with the rest of the semi-idealized section.

12. Reviewer comment: In line 245; instead of going around a rather lengthy justification of why the method to generate turbulence might or might not work, why don't the authors compare the turbulence spectra with that of the met towers? Wouldn't that be a simple, and really robust way to demonstrate there is no doubt on the approach? If the goal of the paper is to study the effect of thermal stratification in realistic conditions, then what better way to proof that the resultant flow is realistic than comparing turbulence spectra? – This seems to me a fundamental requirement.

We have added an Appendix titled "Grid Sensitivity Resolution" which examines the wind speed outputs from the model on d03 (dx = 150 m), d04 (dx = 50 m), and d05 (dx = 10 m). We have added Figures A1 and A3 that show time-series and spectra for all three LES domains and the metmast measurements for both the convective and stable cases.

13. Reviewer comment: In line 269-270, where exactly is the surface roughness being manually

modified? It might be interesting to have a figure illustrating the surface roughness used in the different regions of the domain, so if someone else wants to redo these simulations for new purposes can actually do so.

We appreciate this suggestion to improve the reproducibility of our simulations. Although we have decided not to add a figure, we have added the following text description on line 228 (see response to #10 above):

> "...we set the surface roughness length of the mixed shrubland/grassland land use index in the valley to 0.5 m..."

14. Reviewer comment: In line 275, why does one need to compare the numerical data with the doppler data 'to understand'? One can compare numerical data and experimental data to verify the accuracy of the numerical simulations, as it seems to be done in this section, or alternatively one can study in detail one of the datasets, or both at the same time to unravel the physics of the problem, which is not what is done in this section. It might be worth rephrasing this opening sentence to clarify the expectations for the reader.

We have rephrased the sentence to focus on model validation. The sentence now reads (lines 234-235):

> "The stable case is influenced by a mountain wave event. To validate the accuracy of this event in WRF-LES-GAD, we compare the model with multi-Doppler lidar scans obtained by the DTU lidars."

15. Reviewer comment: In the paragraph contained between lines 300 to 305, it is discussed that comparison of averaged data in a non-stationary problem is complicated. This is indeed very much true. In fact, as the authors well indicate, given the random nature of the turbulent flow, comparison of instantaneous snapshots is not very significant. As an alternative, the authors provide some sort of 'cone averaging strategy (data within an easting and northing position)', as well as an additional spanwise averaging within 30m, if understood correctly. While one could discuss whether this approach is significant or not (which is not what I intend here), what at least the authors could do is better justify their decision. Why is this method robust? How sensitive are the results on this averaging strategy? – Given that this is more or less a hand-waved approach, I think it would be nice for the reader to have a better sense of the sensitivity of the outcomes on this analysis.

We have added the following sentence to the manuscript on lines 260-262:

> "To partially account for this and for any uncertainty of the TLS positioning (estimated to be ±30 m), the wind fields in Fig. 6(a-d) have also been spatially averaged by ±30 m in the span-wise direction."

Additionally, figure 6e (current submission numbering) now includes time-series for an uncertainty of +/- 60 m in addition to +/ 30 m to briefly examine the sensitivity of the model results. We have added the following sentences (lines 265-267):

> "In Fig. 6(e), the predictions for ±30 m and ±60 m are similar because the flow is relatively homogeneous in the spanwise direction (a result of the limited terrain variability in this direction, as seen in Fig. 11 and discussed more in Section 4.3)."

Note that Figure 11 was previously instantaneous and has now been time-averaged.

16. Reviewer comment: In lines 368-369; 'the height of the mountain wave does not extend as high in the model compared to the results'; any thoughts on why this could be? – This kind of physical analysis and interpretation is what could make the manuscript a lot more interesting.

This is likely due to the GFS forcing used as the lateral boundary conditions for the coarsest domain d01. Previous work conducted a sensitivity study on how to force the nested simulations and found that GFS forcing produced more accurate results when compared quantitatively with surface observations. We have thus added "near the surface" to line 220. Additionally, we have added the following sentence to line 247-248:

> "The height of the mountain wave does not extend as high in the model compared to measurements, likely a result of errors in the GFS forcing."

17. Reviewer comment: I would strongly encourage the authors to consider shortening the manuscript. While the main goal of the project, as stated in the introduction, is to understand the effect of thermal stratification on wind turbine wakes in complex terrain, the analysis related to this point only expands between pages 25 to 30, among which only 1.5 pages would be of text and analysis, the rest are figures. So, said otherwise, the main focus and goal of the paper is discussed in 1.5 pages of text out of a manuscript of about 30 pages without counting references. The complementary information provided in the Results section (pages 13 to 25, more than 10 pages) relates to the comparison of the WRF obtained flow with data obtained from the met towers and wind lidars. I have the impression that this is quite out of balance, and maybe the paper should then be framed as a comparison between WRF data and experimental data, which once again I am not so sure on its value given that WRF is used daily around the world, and the manuscript does not introduce any new element or variation of elements in terms of WRF configuration. I think this final page-count also helps illustrates the point I was making in my opening remarks.

Thank you for the suggestions. Again, we would like to highlight that while mesoscale WRF is used daily around the world, its LES capability is not used widely, especially in a multi-scale setup with parameterized wind turbines. Fine-scale terrain-induced dynamic flow features, such as those seen here, require resolving steep terrain and are therefore not represented in mesoscale WRF simulations. Overall, the manuscript has been shortened from close to 10,000 words to under 8,000 words. The complementary information in the Results section has been reframed to focus on validation and has been made more concise. These sections (Sections 4.1 and 4.2) now take up only 6 pages compared to 12 previously.