Please note that references to line numbers are for the revised manuscript.

Reviewer general comment: The current manuscript presents results of a suite of WRF-LES configurations around the complex topography of the Perdigão 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. The manuscript is very nicely written, clear, and with superb "candy-to-the-eye" type figures. Therefore one could say that the article is excellent (despite a few typos or unclear elements).

However, to this reviewer's opinion, while the general idea of the work is interesting and filled with interesting challenges and scientific unknowns, the work presented here remains poor in scientific content, and falls short of addressing any of the initial science-based goals. Let me explain, intercomparisons between numerical models and experimental measurements are critical to be able to objectively determine the quality of the simulation results. However, if the effort stops at that, then it only becomes a mere technical work that anyone outside of the academic world could do. There are indeed publications where industry models are intercompared with experimental data, which have been used in the past to provide a sense of confidence or trust to industry. Alternatively, there are also publications from the academic world, where intercomparisons between models and experimental data are done, but in those there is traditionally a critical analysis of the influence of using different numerical schemes, subgrid models, filtering approaches, etc. So at the end there is an evident scientific gain. Unfortunately in the present work the potential scientific gain remains hard to find. What have we learned out of this manuscript? – That we are now entitled to do WRF-LES simulations in complex terrain with turbines? WRF simulations are done on a daily basis around the world, so how will this change what is done on a daily basis? One could have taken advantage of these great simulations to make a more robust objective/quantitative comparison between the simulations and the experimental data. For example, reading that there are wind differences observed at certain times of 1-2 m/s doesn't really mean much. That could be a 50% difference in a weak mean wind, or a tiny % difference in a strong wind scenario,... Or that the flow looks similar or dissimilar here and there, has little scientific rigor. Once the rigourous/strict comparison of results done, then one could have an additional section with more scientific insight. Research questions that come to mind could be:

As stated in the introduction of the paper, "The focus of this work is to model realistic atmospheric conditions and the associated turbulent flow phenomena to better understand wind turbine wake propagation in complex terrain, using the Perdigão site as the test location. Specifically, we analyze how the vertical deflection and dissipation of the wake varies based on atmospheric stability." This is an open question in the literature based on field observations, and WRF-LES-GAD is now applied for the first time to a real-case simulation with steep slopes to answer the question about wake deflection. The errors in the simulation are quantified via standard metrics (bias and RMSE) compared to a range of field observations. We are cautious to not generalize their results too broadly, but we have attempted to add language that addresses how the results could be interpreted for similar atmospheric conditions and for other sites that may be similar to the Perdigão site (lines 514-518).

1. What is the effect of the surface conditions? – I want to understand that your WRF-LES is using IBM for the topography in regards to the momentum, but is it using the same approach for the ther- mal/moisture field? Based on what I was able to gather from the manuscript it seems that there are no surface conditions for temperature,... does this mean that your flow is insensitive to the surface con- ditions? This is strange since one of the driving mechanisms to thermal stratification is the ground,... but not much is discussed about the impacts or not of the surface conditions on the flow?

The simulations do not use an immersed boundary method - they use the standard terrain-following vertical coordinate system in WRF. We have added language making this more clear in Section 3.3 in line 230. Additionally, the terrain-following coordinate system is mentioned in Table 1. The flow is sensitive to the surface conditions. We use the MYNN surface layer scheme (mentioned in Section 3.3), where the lower boundary conditions are determined from Monin-Obukhov similarity theory, see lines 253-255. The surface temperature is determined based on the Noah land-surface model and a radiative transfer model (the Rapid Radiative Transfer Model, RRTM, stated in Section 3.3). The land-use type is determined using the high-resolution CORINE dataset.

2. What is the effect of your turbulence initialization, and is it really worth it to run a suite of mesoscale WRF simulations just to provide time varying boundary conditions? Mesoscale WRF provides an ensemble flow solution that evolves with time, said otherwise, provides a more or less accurate

mean flow and thermodynamic conditions of the region. However, the LES simulations provide an instantaneous realization of the turbulent ABL, the question then could be posed as how does that compare to instead using profiles of experimental data to force the LES? Also, one can only wonder, what is the effect of the cell perturbation method to generate turbulence? Sure enough the mountains will spur some turbulence, but are the incoming turbulent flow conditions representative of all turbulent scales, including large scale perturbations?

The primary idea behind using multiple grid nests for the simulations is that large scale forcing and perturbations can be passed down to the finer nests where turbulence is resolved. The ultimate goal of such a setup is to use WRF in forecast mode over complex terrain. Although we do not attempt to run LES forecasts in the current work, using observations from the field campaign to force the LES would not achieve progress towards this goal. Furthermore, idealized simulations are unable to provide fully realistic turbulence conditions because the larger forcing scales are not captured (see further discussion in response to specific comment #7 below).

The role of the cell perturbation method has been addressed by Connolly et al. (2021), who studied the effect of using the cell perturbation method versus mountain-generated turbulence in weakly convective, strongly convective, and weakly stable atmospheric conditions for Perdigão. As we mention in line 242-244, they found that the cell perturbation method improves the representation of turbulence relative to the use of high-resolution complex terrain alone. Additionally, Muñoz-Esparza et al. (2014) found that using the cell perturbation method accelerates the generation of turbulence on the inner nest, with no adverse impact on the flow field and negligible computational cost. Also note that the CPM provides perturbations that lead to development of a full range of turbulent scales, as previously presented in several CPM studies, and this detail has been added to the manuscript at line 241.

3. An alternative, potentially the most interesting research question is how can one use the outcome of the rich simulations to understand flow configurations at other locations? Can results be scaled such that the results become generalizable in terms of stratification, mountain slope, terrain complexity? It would be a lot more interesting if the authors used the rich dataset to come up with generalizable relations that enabled one to extract conclusions at other locations without having to run expensive simulations,...

The major research question this article addresses is the vertical deflection of the wake in two distinct atmospheric conditions. Given conflicting literature (Barthelmie et al. 2019 and Menke et al. 2018) regarding the vertical deflection of the wake in different atmospheric stability conditions based on measurements, we use large-eddy simulations to address vertical wake deflection in complex terrain. The wake deflection has important ramifications for any turbines that could be located downstream.

As previously mentioned, we are cautious to not generalize their results too broadly, but we have attempted to add language that addresses how the results could be interpreted for similar atmospheric conditions and for other sites that may be similar to the Perdigão site (lines 514-518):

"We expect that the conclusions in terms of the wind turbine wake behavior would hold for all convective and stable atmospheric conditions at the Perdigão site as long as the phenomena of interest (recirculation zones and mountain waves) are present. Other phenomena could be modeled using WRF-LES-GAD to examine wind turbine wake behavior in other cases. With regards to other sites, the wind turbine wake behavior would depend largely on the vegetation (surface roughness) and steepness of the terrain."

More Specific Technical Comments

1. Line 42; when talking about 'length-scales' one should clarify that these are "turbulent length-scales"

This has been added.

2. Line 53; it is important to clarify the difference in time scales between the LES and the WRF mesoscale simulations. LES provides an instantaneous realization of the flow at high-temporal frequency, while WRF mesoscale provides results representing the outcome of an ensemble of flow marching in time, but that per the ergodicity definition can not be interpreted at the same frequency than otherwise the LES inout/output.

WRF is designed to operate as both a RANS and LES model, and in practice, the output of the model is a time-evolving 3D field, regardless of whether it is in RANS or LES mode. The timesteps in WRF are chosen according to the spatial resolution and stability limits. Multiple grid nests used to transition from RANS to LES are standard practice, e.g. Wiersema et al. (2020), Rai et al. (2017), and Arthur et al. (2020).

3. In line 57; can the authors provide a brief description of the GAD model? Meaning, I don't need to read the reference to find out whether the models is an actuator disk with/out rotation, etc.

We have added the following text to the manuscript at line 59-61: "In the GAD parameterization, thrust and rotational forces computed at the turbine's blades are averaged over a discretized two-dimensional disk formed by their rotation. These forces are then applied to the flow surrounding the turbine."

4. In line 82; the authors mention that "the goal of this work is to model realistic atmospheric conditions and the associated turbulent flow phenomena to better understand wind turbine wake propagation in complex terrain". At the end of the manuscript, what have we learned about this that maybe of use as a function of thermal stratification, or in other locations, mountains, terrain, etc?

As mentioned in the introduction (paragraph from lines 64-75), there is conflicting literature regarding the vertical deflection of the wind turbine wake in different atmospheric stability conditions. In this manuscript, we have learned that during stable conditions where a mountain wave occurs, the wake deflects downwards. During convective conditions where a recirculation zone forms, the wake does not mix with the recirculation zone and deflects above it.

Mountain waves occurred for almost 50% of the nights during the intensive observation period (Fernando et al. 2018). For recirculation, Menke et al. 2019 found that reverse flow with wind speeds greater than 0.5 m/s occurred over 50% of the time when the wind direction was perpendicular to the ridges. We have added these details into the manuscript at lines 130-132.

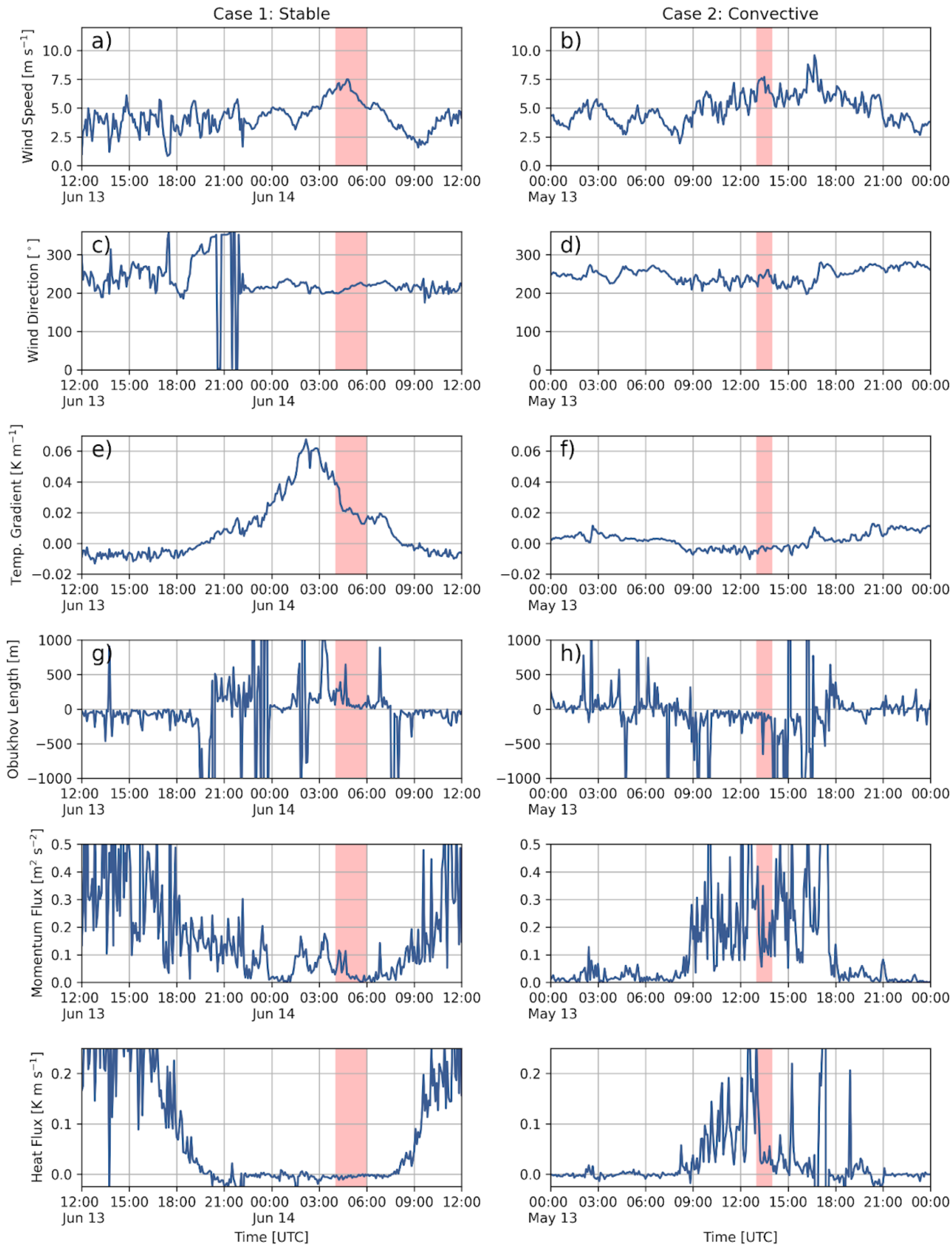
As previously mentioned in this response, these phenomena are site specific and depend on the characteristics of the terrain such as the slope and vegetation (surface roughness). But generally, we would expect that in other places with hilly terrain the wake would deflect up or down depending on stability.

5. By the end of section 2.1; what about the surface conditions? If they were not measured (besides roughness) how are they taken care of in the simulations? Is z0,t taken as a fraction of z0?

As mentioned in a response to a previous comment, we use the MYNN surface layer scheme (mentioned in Section 3.3), where the lower boundary conditions are determined from Monin-Obukhov similarity theory. The surface roughness lengths are based on the 100 m CORINE Land Cover dataset and we also use 30 m terrain from SRTM (mentioned in Section 3.3). Aside from the CORINE land cover which is specific to Europe, we have not changed any of the land-surface options which are standard in WRF. For the reviewer's reference, in WRF, z0,t is parameterized as a function of the Zilitinkevich parameter (Zilitinkevich 1995), z0, and Reynolds roughness number.

6. In Figure 2; I could only wonder but what happens to the near surface heat and momentum fluxes? – It is well known in the community the existence of counter gradient heat fluxes, specially in complex terrain. Why not include subpanels with the near surface heat fluxes?

In response to the other reviewer, we have added the Obukhov length to Fig. 2. Shown here is the new Fig. 2 with additional sub panels of the near surface heat and momentum fluxes (included here only for the reviewer - we do not believe the additional panels are needed in the paper). The near surface heat flux during the stable case is around -0.0025 K m/s and 0.025 K m/s during the convective case. During both case studies, the near surface heat fluxes are relatively constant. Comparing the near surface fluxes with large-eddy simulations is complicated above the surface because there is a combination of resolved and subgrid LES quantities. This analysis is the subject of our ongoing work, which focuses on turbulence quantities, but is not included in the present paper.


Case 1: Stable
Case 2: Convective
a)
b)
c)
d)
e)
f)
g)
h)
Wind Speed [m s^{-1}]
Wind Direction [$^\circ$]
Temp. Gradient [K m^{-1}]
Obukhov Length [m]
Momentum Flux [m^2 s^{-2}]
Heat Flux [K m s^{-1}]
12:00 15:00 18:00 21:00 00:00 03:00 06:00 09:00 12:00
Jun 13
Jun 14
00:00 03:00 06:00 09:00 12:00 15:00 18:00 21:00 00:00
May 13
Time [UTC]

7. Around line 185; It seems like simulations are initialized with a uniform heat flux. However, it is known also that the heat flux will be rarely uniform in complex terrain. How is that potential effect assessed? Are we left to interpret that the results are indifferent to that surface forcing?

The idealized simulations are indeed initialized with a uniform heat flux because they are formulated as relatively simple studies to inform the setup of the more realistic multi-scale simulation. A recent study showing that the multiscale setup gives better turbulence properties than the idealized setup is presented in Wiersema et al. (2020). This is because idealized turbulence is missing the large scale forcing that results from dynamic downscaling. A sentence in the introduction of the manuscript at lines 54-56 has been edited to mention this:

"Such setups can provide LES with more realistic time-varying inflow conditions directly from the mesoscale simulations, as these setups include the large scale forcing that results from dynamic downscaling capturing a wider range of atmospheric phenomena and more realistic turbulence compared to conventional idealized LES setups (Wiersema et al., 2020)."

In our multi-scale simulation setup, we use topographic shading and a land-surface model to account for non-uniform heat fluxes in complex terrain. This has been added in Section 3.3 from lines 255-256:

"Additionally, topographic shading is enabled to account for shading effects on the surface heat flux in the complex Perdigão terrain."

8. First line in Section 3.3; "Having demonstrated the ability..."; I dare say that to this point not much has been demonstrated besides showing two beautiful figures. Maybe the authors can tone this a bit down.

Our goal with the semi-idealized simulations was to see if the idealized model can capture the deflection of the wake with different phenomena of interest. We have edited the first line in Section 3.3 as follows: "Having demonstrated the ability of WRF-LES-GAD to capture the different types of wake behavior in a semi-idealized setup, we now proceed to the full multi-scale simulation."

9. Line 210; the authors mention that "adequate turbulence is developed", where is this shown? What is the premised/argument used

to judge that? A certain amount of energy at certain scales? A tke scaling of k−5/3? A certain spectral comparison of tke with the field experiment at different heights? – Some of this might be more "objective".

The text in the manuscript has been rephrased to say "spun up for 9 hours". This amount of spin up time is consistent with common practice for WRF. For example, Connolly et al. (2020) used an 8 hour spin-up time, used a 15 hour spin-up time, and Arthur et al. (2020) also used a 9 hour spin-up time.

10. It is unclear how the Cell Perturbation Method generates different type of turbulence for the stable and unstable stratification.

The cell perturbation method provides perturbations based over the boundary layer depth. The perturbation temperature is drawn from a uniform distribution within a range of potential temperatures that is based on an optimal Eckert number of 0.2 (see Muñoz-Esparza et al. 2015). This is the subject of the studies by Muñoz-Esparza et al. (2018) and Connolly et al. (2020) and more detail regarding the methodology for CPM can be found there and in Muñoz-Esparza et al. (2015).

11. From line 265 onwards, the text is riddled with subjective comparisons. I would suggest using percentages instead of absolute values when for example comparing wind speeds. Or avoid using comments like:"matches the measurements well", without an actual metric of it. See line 311; "errors are small" in comparison to what?

At lines 311, 317, and 388, subjective comparisons have been removed.

At lines 338, 346-349, 354, and 397-398, when discussing and comparing the errors, we have now quantified the errors with metrics of bias and RMSE within the text.

12. Why is the wind speed and direction outputted at different frequencies?

The wind speed and wind direction are output at the same frequency, 10 seconds. Perhaps the reviewer is commenting about the temperature gradient output frequency, which was 2.5 minutes, simply because of computational storage limitations.

13. Around line 367; the authors comment that the errors/discrepancies observed during the convective periods are larger than during the stable periods, but they don't provide any comment, hypothesis, argument, detailed inspection, trying to investigate that in details,...

The sentence containing this observation has been removed from the manuscript.

14. Figure17, great visualization for oral presentations, proposals or others where "candy-to-the-eye" is well accepted; but what is the use of such an image here? What is the intended message, outcome extracted? – I am not suggesting the authors remove the figure but instead include scientific argumentation around it.

Figure 17 shows the meandering and full development of the wake over complex terrain. The entirety of the wake cannot be visualized in two dimensions (vertical transects or plan slices) because of wind veer and horizontal/vertical wake meandering. The discussion for Figure 17 is qualitative while Figure 19, later on in Section 4.3, aims to be more quantitative regarding vertical wake behavior. We have added the following sentences in lines 429-431:

"The entirety of the wake cannot be visualized in two-dimensions because of wind veer and horizontal/vertical meandering. Three-dimensional visualizations provide insight into the wind turbine wake advection, meandering, and direction downstream as the flow evolves and develops over the first ridge and through the valley."

In conclusion, I am convinced the data presented in this manuscript is of high quality, and has the potential to become of high value for the community if made publicly available. My only concern is that there is not enough scientific value at presenting the data itself, when additional analysis (some of which does not require extensive work) could be added that would increase the scientific value of the work.

We thank the reviewer for the comments. We think that the revised manuscript and the responses above help to address most of these concerns.