We thank both the editors and reviewers for considering this manuscript. Below, reviewer comments are in black and **our responses are in blue bold.**

Reviewer 2

Review of Characterizing atmospheric stability in complex terrain by Nathan J. Agarwal and Julie K. Lundquist

Recommendation: Major revision

This is a well written, careful study that leverages the unusually rich Perdigão tower dataset to assess how three stability metrics (Obukhov length L, turbulence kinetic energy TKE, and dissipation rate ϵ) behave in complex ridge–valley terrain and how representative surface (10 m) measurements are of hub-height (100 m) conditions. The manuscript provides useful guidance on Reynolds-decomposition windows (ogive analysis) and introduces sensible metrics for quantifying representativeness (HHPI and horizontal homogeneity). Overall the work is scientifically sound and valuable to the wind-energy and boundary-layer communities, but several important methodological clarifications and additional discussion (especially on canopy effects and comparisons to flat-terrain benchmarks) are required before publication.

We thank the reviewer for their thoughtful and supportive appraisal.

Major comments

1.   Expand and tighten the Introduction with regard to prior Perdigão work.

o   The Perdigão campaign generated many multi-instrument studies on ridge–valley and canopy effects. Readers would benefit from a clearer map of what previous Perdigão studies (and other complex-terrain works) have shown and how this manuscript's contribution differs.

We thank the reviewer for this comment. We agree that the Perdigão field campaign has allowed for several laudable scientific contributions. We include a more complete peer review that begins on (updated) line 109 and extends fully to line 144, in section 2.1:

"Perdigão addressed fundamental understandings of physical mechanisms in complex terrain. Menke et al. (2019) documented recirculation zones throughout the valley with doppler radar data. Letson et al. (2019) used data from both sonic anemometers and Doppler lidars to assess wind gust characteristics. Wagner et al.

(2019) as well as Wildmann et al. (2019) and Venkatraman et al. (2023) analyzed detection of low-level jets (LLJs). Wise et al. (2021) and Robey and Lundquist (2024) have documented mountain wave patterns. This field campaign has also supported research into wind-energy-specific mechanisms like wakes. Menke et al. (2018) used data from six scanning lidars to analyze the dependence of wake deflection patterns on the atmospheric stability, and these wake deflection patterns were also scrutinized by Barthelmie and Pryor (2019) and Wise et al. (2021). Similarly, Dar et al. (2019) simulated the wake behind a turbine and showed that the self-similarity of the wake implied that the wake pattern depended on neither the complexity nor the shape of the terrain.

"Improving the understanding of these physical mechanisms has also enabled model improvement. Wagner et al. (2019) established a foundation with a long-term Weather Research and Forecasting (WRF) large eddy simulation (LES) simulation set for the full observational period. Palma et al. (2020) developed a digital terrain model (DTM) for the site and also defined a necessary horizontal grid cell spacing criteria of 40 m to appropriately characterize the site. Several modeling works have also expanded upon this need for improved representation of horizontal variability for the region. Quimbayo-Duarte et al. (2022) established a forest canopy parameterization for the WRF-LES model that improved wind modeling within the lowest 500 m of the atmosphere and attributed the imposed drag parameterization to a need for even higher-resolution land surface inputs. Venkatraman et al. (2023) altered the forest canopy parameterization to consider more realistic (i.e. shorter) heights in a set of OpenFOAM simulations. Finally, Al Oqaily et al. (2025)—building upon the ensemble models in Giani and Crippa (2024)—again underscored the importance of land cover in an additional set of LES simulations. Connolly et al. (2021) demonstrated the utility of the cell perturbation method (CPM) (Muñoz-Esparza et al., 2015) in capturing flow features in LES simulations of the region, especially in weakly convective conditions. The Perdigão field campaign has also enabled improvements in wind turbine specific modeling in complex terrain. Wise et al. (2021) introduced the WRF-LES-GAD to the Perdigão dataset, and leveraging the output from this model, Robey and Lundquist (2024) demonstrated the utility of a virtual lidar model (Robey and Lundquist, 2022) in simulating range height indicator (RHI) scans. The modeling improvements have also extended beyond numerical weather prediction (NWP) models. Barthelmie and Pryor (2019) developed a novel detection algorithm that could both identify and characterize wake patterns, Vassallo et al. (2020) used artificial neural networks (ANNs) to improve wind speed error by up to 52%, Bodini et al. (2020) used machine-learning models to reduce the average error in model representation of turbulence dissipation rate by up to 40%, and Mosso et al. (2025) predicted turbulence anisotropy for the region using random forest models.

"The Perdigão field campaign has also enabled improvements in measurement techniques. Vasiljevic et al. (2017) outlined a methodology for multi-lidar Doppler

lidar experiments and Wildmann et al. (2019) presented a new wake measurement strategy that used three synchronized lidars to adapt to the prevailing wind direction automatically. Bell et al. (2020) combined multiple Doppler lidar scans to create virtual towers that could extend beyond the range of traditional in situ meteorological towers. Coimbra et al. (2025) also performed a virtual mast analysis of the Perdigão region with three coordinated dual lidar. These measurements allowed for not only the mean quantity comparisons available in Bell et al. (2020), but also turbulent quantities as well."

We also welcome an opportunity to more clearly motivate the novelty of the present work. We include an additional paragraph in the Introduction starting on line 91:

"(E)xisting Perdigão analyses share two important limitations. First, most focus on vertical extrapolation, virtual towers, and rotor-layer prediction while largely neglecting how horizontal heterogeneity imposed by steep terrain and vegetation modulates near-surface flow. Second, although technological solutions have enabled reliance on measurements above 40–80 m, the representativeness of surface or 10 m observations remains poorly understood. This gap directly affects wind-energy stakeholders, who often depend on near-surface measurements for resource assessment and may be unable to deploy tall towers in complex terrain.

"These limitations motivate the present study…"

o   Menke et al. (2019; Characterization of flow recirculation zones at Perdigão), a study co-authored by one of the present authors, also investigates stability effects and should be discussed — particularly in the context of how recirculation zones may provide alternative explanations for some of the vertical decoupling observed here and why the stability characterization approach used in Menke et al. is not used in this work. If the authors consider Menke et al. (2019) not directly comparable, they should explicitly state and justify this.

We thank the reviewer for this comment. We agree that the Menke et al. (2019) work is an important contribution to the Perdigão literature and include it in the literature review. We avoided using a Richardson number because, as our analysis demonstrates, the stability at one level may not necessarily reflect the stability at another level. To reflect this concern in the manuscript, we have added a line, now line 206:

"Richardson Numbers were also intentionally not employed for this analysis. While other analyses such as Menke et al. (2019) have relied on a gradient Richardson number to assess stability, here we avoid any layer-based stability characterizations. These metrics were excluded because these metrics are sensitive to how the layer is defined."

We also understand the reviewer's interest in a discussion as to the potential role of recirculation zones in contributing to the vertical decoupling.

We also address TKE decoupling directly on line 521:

"Canopy effects may help explain this 10 m TKE decoupling during unstable conditions in the valley and NE ridge. These 10 m measurements, located with the canopy layer but not necessarily the vegetation layer, could experience heightened turbulence that would necessarily be reflected at taller measurements. Recirculation zones, also present during unstable conditions in the valley and NE ridge, may also influence this TKE decoupling. However, as noted in Menke et al. (2019), the depth of these recirculation zones exceeds 100 m. Thus, because these recirculation zones would not necessarily impact 10 m measurements uniquely, their effect may be secondary to the more direct influence of the canopy."

We also discuss the potential role of recirculation zones in contributing to the observed friction velocity behavior on the NE ridge during unstable conditions. On line 431 in a new Vertical Profile section, where we discuss how–during unstable conditions, NE ridge friction velocities align more with valley friction velocities than SW ridge friction velocities–we note:

"Both flow and terrain features may be contributing to this NE ridge subversion. Menke et al. (2019) documented recirculation zones during over half of the unstable periods. These flows could then explain why this unique behavior is reflected only during unstable cases. Menke et al. (2019) also documented that these recirculation flows occurred more frequently near the valley and NE ridge than the SW ridge. Thus, recirculation zones provide a mechanism that is aligned in terms of geography and stability condition to explain this behavior..."
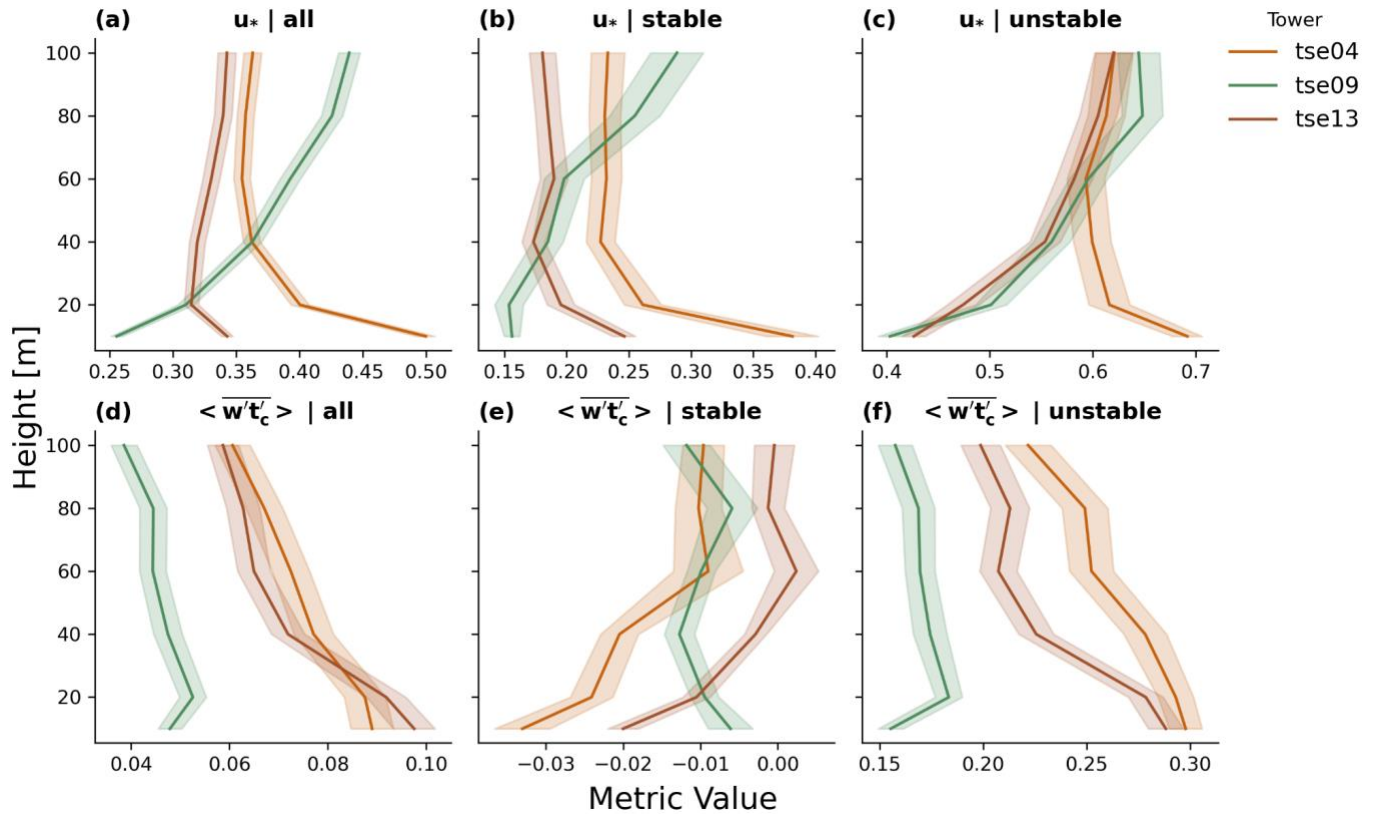
Figure 6. Perdigão field campaign vertical profiles with 30 min averaging at the three 100 m towers. The solid line indicates the mean value and the band represents the standard error. (a) friction velocity all (00-23 UTC); (b) friction velocity stable (00-02 UTC); (c) friction velocity unstable (12-14 UTC); (d) heat flux all (00-23 UTC); (e) heat flux stable (00-02 UTC); (f) heat flux unstable (12-14 UTC).

2.   Significantly expand discussion and quantification of canopy/vegetation effects.

o   Why it matters: Several of the manuscript's central findings — in particular the apparent 10 m vs 100 m decoupling in the valley and the NE ridge — could be strongly influenced by canopy height and density. The manuscript repeatedly notes "canopy" or "canopied NE ridge" but does not quantify canopy height/cover or show photos/maps of tower surroundings. The text suggests remote generation/advection as the explanation (e.g., upside-down boundary layer / advection), but canopy sheltering is an equally plausible (and in places more parsimonious) explanation for decoupling at 10 m. See the discussion and TKE stratification text and figures.

o   Suggested fixes:

§ Add a table or figure that lists canopy/land-cover at each tower (or at least for the three focal towers), **including approximate canopy height and whether the 10 m sonic is inside the roughness sublayer/canopy.**

We thank the reviewer for this comment. We include a new table and associated text (starting on line 156) with the requested information. We base our estimates on existing literature for this case study.

"The three 100 m towers were strategically placed to sample flow conditions within differing terrains. Specifically, tse04 was located on an exposed ridge on the southwest, tse09 in the valley with eucalyptus and fir trees, and tse13 on a ridge with a heterogeneous canopy to the northeast (Table 1). Although translating CORINE Land Cover data (Bossard et al., 2000) into U.S. Geological Survey land use types to obtain surface roughness lengths (Pineda et al., 2004) suggests surface roughness lengths on the order of 0.01–0.05 m, Wise et al. (2021), Wagner et al. (2019), and Palma et al. (2020) (among many others) have suggested that larger surface roughness lengths are more appropriate for the site. The mean canopy heights–as documented in Letson et al. (2019)–are below 10 m (Table 1 and Fig. 2), although individual trees may extend to 15 m (Mosso et al., 2025)."

Table 1. Shortname Reference for 100 m towers

Tower	Shortname	Topography	Location [WGS84]	Elevation above sea level [m]	Mean canopy height [m]
tse04	Southwest Ridge	Turbine Ridge	7°44′33.37″W 39°42′21.47″N	473	1.5
tse09	Valley	Valley	7°44′5.40″W 39°42′40.36.″N	305	5.4
tse13	Northeast Ridge	Canopy Ridge	7°43′49.38″W 39°42′48.97.″N	453	3.6

§ Include **site photographs** of the three focal towers (or reference an existing photo repository) to make interpretation of 10 m measurements transparent to readers.

"We thank the reviewer for this suggestion. We include the below figure, Fig. 2, which includes available site photographs of each of the three focal towers.

Figure 2. Canopy cover for (a) tse04, (b) tse09, and (c) tse13. The image of tse04 was identified as a still photo in the Danish Technical University (DTU) Perdigão video (DTU Wind and Energy Systems, 2017) and the images of tse09 and tse13 can be found in the Perdigão data archive (EOL).

§   Re-analyze (or at least discuss) whether the valley / SE ridge 10 m decoupling is better explained by canopy effects than by remote advection.

We thank the reviewer for their proposal of an alternate decoupling mechanism. We include a more complete discussion of TKE decoupling mechanisms, emphasizing

3.   Justify the choice to focus horizontal-homogeneity (HoH) on 10 m measurements only, or extend analysis to higher levels.

o   The HoH analysis determines the minimum number of towers needed to represent site variability but is performed at 10 m because only 18 towers have 10 m temperature. For wind-energy stakeholders, representativeness at hub-height (or intermediate heights) is more relevant than representativeness at 10 m.

to 20 m would only allow 38 towers to be analyzed. Shifting to 30 m would only leave 17 towers. Further, only 3 towers had measurements at hub height.

o   Suggested fixes:

§   Either: (a) repeat the HoH analysis at other heights where sufficient data exist (e.g., 20/40/60 m subsets) and show differences, or **(b) clearly justify why a 10 m-based HoH is still useful for turbine siting.**

We thank the reviewer for this suggestion. While including measurements from taller heights may seem ideal, in this particular case, shifting the analysis to a height above 10 m reduces available data and thus the impact of the HoH analysis. Shifting to 20 m would only allow 38 towers to be analyzed. Shifting to 30 m would only leave 17 towers. Analyzing 100 m would only leave 3 data points. Further, 10-m towers are much more common in wind resource assessment than any of the other heights. Maintaining the analysis at 10 m also reveals that stakeholders may still be able to leverage the commonly-used 10-m towers to inform site characterization. To acknowledge this methodological concern in the manuscript, we have added additional text in a paragraph, now starting on line 319:

"Each metric was also evaluated for HoH. This assessment was designed to quantify the smallest number and location of meteorological towers necessary to capture the site's variability in surface (10 m) stability. For each metric, all 47 towers with sonic anemometers were considered. Because only 18 of these towers had available 10 m temperature measurements, the virtual potential temperature was assumed to be a uniform 300 K for the Obukhov Length calculation just for the HoH analysis. Thus, this analysis could not account for differences in virtual potential temperature that might occur between towers. This approach to addressing the lack of available temperature measurements was determined optimal because other potential solutions–such as performing the analysis at either a shorter or taller height–further limited the number of available towers or relied on data fully within the vegetation layer. Shifting to 20 m would only allow 38 of the 47 towers to be analyzed. Shifting to 30 m would only leave 17 towers. Analyzing 100 m would only leave 3 data points. These other solutions also devalued the assessment as a practical demonstration of how stakeholders could leverage surface-level measurements to inform site characterization."

§   Add a short paragraph discussing pros/cons of using Louvain community detection (e.g., sensitivity to input correlation metric and partition resolution) and whether other clustering algorithms were tested.

We thank the reviewer for this careful observation. Our decision to use the Louvain community detection algorithm is largely because this algorithm is most appropriate for the use case. Other clustering algorithms, such as a kNN, require the analyst to pre-determine the number of clusters. As such, the kNN approach ambiguously defines the marginal value of an additional cluster. To reflect this decision, along with further details on some of the benefits and limitations of this algorithm, we revise this paragraph starting on line 329:

"These 47 towers were then represented as a graph network, with a node for each tower location and the edge weights representing the strength of similarity in surface stability between any two given towers. This similarity metric S mapped the correlation in surface stability Cij with only positive edge weights such that S = (Cij +1)/2 . The smallest number of meteorological towers were then determined based on the Louvain community detection algorithm (Blondel et al., 2008), as implemented by the NetworkX Python package (Hagberg et al., 2008). Note that the Louvain algorithm, unlike kNN methods, treats the number of clusters as an output instead of an input. The Louvain algorithm determines which towers are redundant by identifying tower subgroups and assigning each tower to one of these subgroups. The Louvain community detection algorithm also enforces (non)treatment of overlapping communities, further ensuring that each community is treated independently with a unique tower. While the Louvain community detection algorithm is also sensitive to several tuning parameters, we make no adjustments to the default modularity gain threshold (1e-7) or resolution (1) and introduce no artificial seeds or establish restrictions on the number of optimization cycles."

§   Consider adding the widely used TRIX (or equivalent) representativeness test from the wind-energy community to complement the HoH analysis — readers from industry will find that comparison useful.

We appreciate the reviewer's suggestion to introduce a model-based HoH analysis to complement the observational analysis. We agree that a model-based HoH is a more complete demonstration of our methodology. While we did consider the TRIX representativeness test, we instead offer an LES-based analysis, consistent with the recommendation from Reviewer 1. We believe that an analysis of LES data provides a more natural comparison with our current observational analysis. This analysis is described in Appendix C:

"Here we demonstrate a proof-of-concept implementation of the Louvain methodology applied to large-eddy simulation (LES) data. This Louvain implementation considers near-surface static stability from the LES of Robey and Lundquist (2024) (Table A1), based on Wise et al. (2021). These data represent a roughly 3 hr period early in the morning. Although simulation output is available at 1

second resolution, here we sample every 30 min, consistent with other analyses in this work. We also spatially coarsened the 100 m output to the effective model resolution of 500 m (Skamarock, 2004) and then subset to the identified Perdigão tower locations. Static stability was then calculated from these spatially- and temporally-coarsened data as:

$$\frac{\partial \theta}{\partial z} = \frac{\theta_2 - \theta_1}{z_2 - z_1}$$

such that θ represents a potential temperature (K), z represents a model height above ground (m) and the two model levels taken are those closest to the surface (10 m and 2 m). The Louvain algorithm was then applied to these static stability calculations, resulting in seven partitions (Fig. C1). Overall, these tower groupings reflect broad terrain boundaries, with opportunities for mixing. Orange is predominately restricted to the upper valley. Dark blue exists mostly on ridges and the outer transects. The relatively small gray partition exists in the slope between the NE ridge and the valley. Light blue reflects most of the SW ridge, with occasional representation in the valley. The salmon, brown, and green transects are less geographically informed. Of course, because this analysis is based on 3 hours of LES during stably stratified conditions rather than the full measurement campaign period, the partitions are different from those presented in the main paper.

Table A1. Model parameters used in LES Louvain analysis, based on Robey and Lundquist (2024).

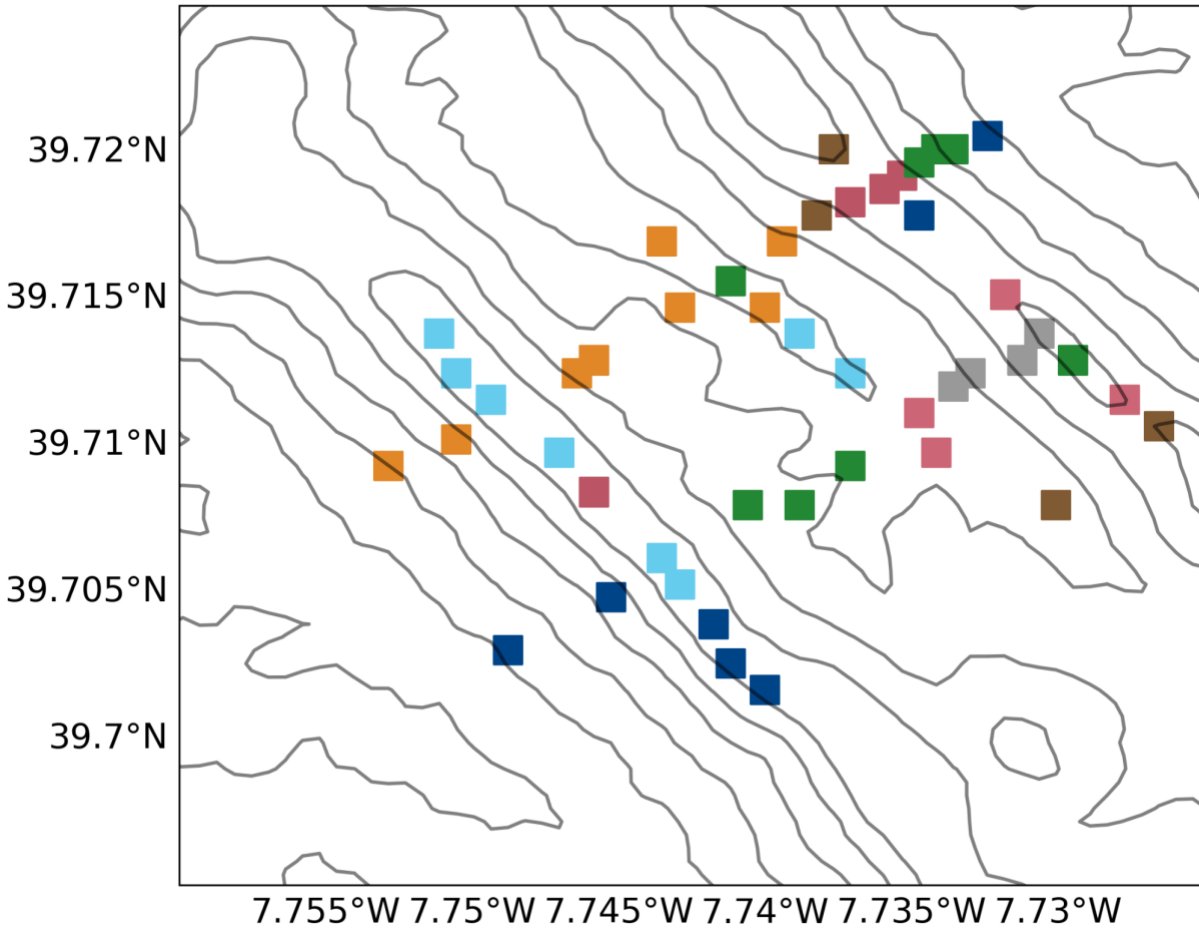| Model Parameter	Value
Start time	2017-06-14 03:24:00
End time	2017-06-14 05:16:25
Effective horizontal resolution	500 m
Horizontal grid resolution	100 m
Vertical grid resolution	160 vertical levels
Vertical levels	2 m and 10 m
Land cover	CORINE Land Cover 2006 with mixed shrubland–grassland roughness length updated to 0.5 m
Terrain	1 arcsec terrain from the Shuttle Radar Topography Mission (SRTM)
Turbulence subgrid-scale model	non-linear backscatter and anistropy (NBA2)

Figure A1. Static stability partitions determined by the Louvain community detection algorithm with LES output from Robey and Lundquist (2024) at the locations of the station towers. Model parameters are defined in Table A1.

"This implementation is designed to serve several stakeholders and future research directions. Because this analysis builds on publicly available LES output from Robey and Lundquist (2024), it provides a reproducible benchmark for future Louvain or LES-based studies. The short duration and dense observations create opportunities to explore similarity metrics beyond the temporal correlation weight used here. We also introduce static stability as an alternate metric that depends only on temperature measurements, recognizing that many stakeholders, including wind farm developers, rely on sparse instrumentation. To further mimic these practical constraints, we adopt a 30-minute temporal resolution.

"This case study therefore provides both a benchmark and a flexible framework that can be extended to other datasets, parameterizations, and applications."

4. Provide a direct comparison or benchmark of HHPI with flat-terrain results (or literature).

o   The conclusion that "surface measurements are unable to make hub-height predictions" is important, but the reader needs context: are the HHPI values unusually low compared to flat terrain or other published sites? That context would quantify how "bad" the Perdigão complex terrain results are and support the recommendation for tall towers. The manuscript suggests future work comparing to other sites but does not provide **literature benchmarks**.

We thank the reviewer for this suggestion. After assessing the available literature to confirm that our HHPI methodology is novel, we established our own HHPI benchmark with available eXperimental Planetary boundary layer Instrumentation Assessment (XPIA) data as described in Appendix B:

"Here we provide a flat terrain comparison to the Perdigão HHPI analysis. This flat terrain benchmark serves to contextualize the role of complex terrain in affecting the relative agreement between surface and hub-height stability. Sonic anemometer data from the Boulder Atmospheric Observatory (BAO) tower during the three-month eXperimental Planetary boundary layer Instrumentation Assessment (XPIA) (Lundquist et al., 2017; Bodini et al., 2018) field campaign provide the data source for this comparison. Available measurements include data from 5m, 50m, and 100m. As such, we compare both the 5 m - 100 m and 50 m - 100 m. Each stability metric calculation employs a screening process similar to that employed for the Perdigão data, including removal of the approximately 20% of data affected by precipitation, 99.5th percentile extrema removal, and tower wake distortion screening.

"HHPI is higher in flat terrain than in complex terrain at shorter heights. HHPI in flat terrain is similar to HHPI in complex terrain in several ways. Both improve with taller towers. In both cases, log10ϵ has the lowest HHPI of all metrics and this metric distinction is especially-pronounced for shorter heights (Fig. B1c). Both terrains also exhibit similar curvatures for each metric. L HHPI consistently is flatter at lower heights, and increases more sharply at higher heights (Fig. B1a). TKE HHPI increases more sharply for shorter heights before increasing more gradually above (Fig. B1b). log10 ϵ HHPI is unique in its consistent linear increase (Fig. B1). HHPI in the two terrains also differ. Notably, HHPI is higher in flat terrain than in complex terrain (Fig. B1). Surface L HHPI is between 0.1 and 0.2 higher in flat terrain than in complex terrain (Fig. B1a). Surface TKE HHPI in flat terrain leads that in complex terrain by sometimes as much as 0.3 (Fig. B1b). Surface log10 ϵ demonstrates the smallest flat terrain lead (Fig. B1c). These flat terrain leads, pronounced at the surface, are not guaranteed at higher heights. While L HHPI continues to reflect a flat terrain premium (Fig. B1a), both TKE HHPI (Fig. B1b) and log10 ϵ HHPI (Fig. B1c) do not. Thus, while shorter towers may be more useful in flat terrain than in complex terrain, flat terrain towers may also benefit less from height increases."
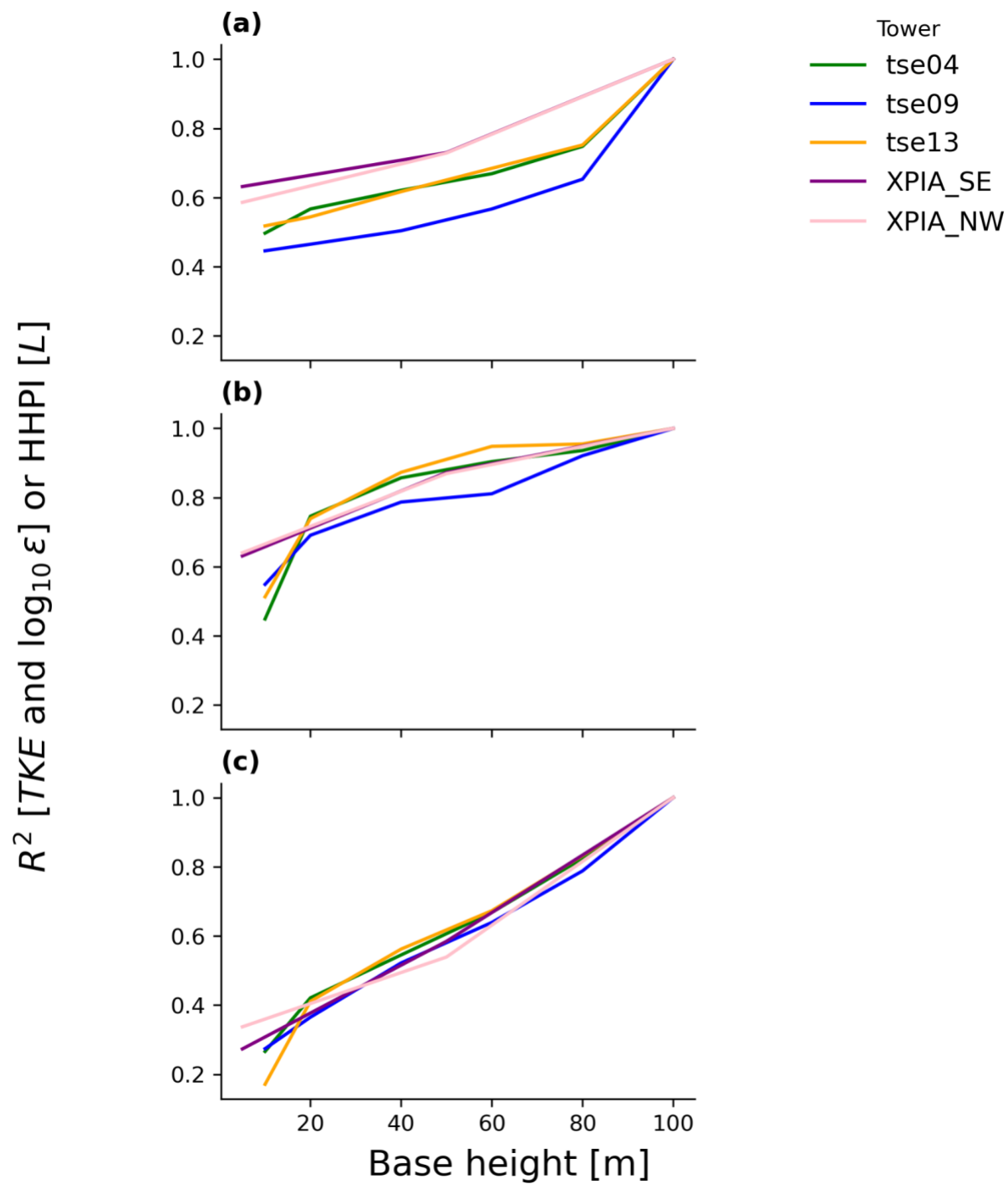
Figure B1. Flat terrain (XPIA) benchmark to the Perdigão HHPI analysis. (a) L; (b) TKE; (c) log10ϵ.

5. Re-evaluate some interpretations of Figures (wind roses and 10 m vs 100 m diffuseness).

o The manuscript states "100 m winds are more diffuse than 10 m"; however, the reviewer's read of Fig. 4 suggests 100 m sometimes shows a clearer prevailing direction than 10 m. Please verify the figure interpretation and ensure the text statement is correct.

We revise this sentence, now on line 448, to read "100 m winds are less bidirectional than 10 m winds" as suggested by another reviewer.

Minor comments and editorial suggestions

1. Introduction: Many studies cited that are not explicitly identified as flat or complex-terrain studies. For each "stability affects wind generation" citation, briefly state terrain type (flat, coastal, complex) so readers immediately see which conclusions generalize to Perdigão-like terrain.

We thank the reviewer for offering an opportunity to make our literature review more useful to other researchers. We have introduced more site-specific characteristics throughout our literature review.

2. Be consistent with units formatting: use parentheses "[m s⁻¹]" consistently instead of mixing "[]" and "()" in figure captions. (Figure 1 and others show mixed usage.)

We thank the reviewer for this observation. We have adjusted Figure 1 as well as other figures to reflect a consistent "[]" usage.

3. Move subplot descriptions from the long caption into a one-line title for each subplot (improves readability for multi-panel figures such as Figs. 3 -13). This was hard to follow when flipping between caption and panels.

We thank the reviewer for helping improve the readability of our figures. We have adapted figures 2-15 with titles.

4. Line 140: double words "to to"

Thank you. We have removed the additional "to".

5. Two sentences in 3.2 (lines ~394ff) are nearly identical and can be combined for concision: the point about TKE HHPI improving with height for all sites can be consolidated.

Thank you. We adapt this line 547 to read:

"The TKE HHPI profile consistently shows higher HHPI than the other two metrics overall, regardless of the location (Fig. 16a,b,c)."