In this document, the reviewer's comments are in black, the authors' responses are in red.

The authors thank the reviewer for their thoughtful and productive comments.

General comments

The paper focuses on the turbulence dissipation rate and whether three machine learning techniques can outperform parametrizations of dissipation rate commonly used in numerical models. For this purpose, the authors use the Perdigão dataset with an unprecedented number of 184 sonics on towers ranging up to 100m in height. This paper is both timely and relevant as the turbulence dissipation rate is one of the most important turbulence diagnostics and its incorrect representation in models and related biases have wide ranging consequences. The machine learning approach is also the appropriate methodology to tease out the information about possible many influences from such a large dataset and the results are encouraging. Despite its merits, however, the paper still needs to address a number of points listed below, some of which might change the results, before I can recommend it for publication. Given my expertise I focus more on the physical aspect of the paper than details of machine learning. I therefore recommend major revision.

Specific comments

1. **Information on data analysis**
   I find the information on the data post-processing and analysis insufficient.
   - Particularly missing is the information on the averaging times which is confusing. It is stated that the dissipation rates were calculated from the second order structure functions at 30 s intervals, but that TKE was calculated at 2 min intervals (ln 91). Are the other averaging times 30 min (ln 96)? Why is there a difference between the averaging times of different variables and how are they then reconciled for the purposes of machine learning where predictor and response variables need the same length?
     We have added details on the variables used in our analysis. Moreover, we have now calculated all variables using the same 30-minute average period. This has been clarified in many places throughout the manuscript.

   - What is the motivation of calculating TKE at 2 min intervals and not 30 min like the other variables? Are the authors trying to say that the relevant TKE for the dissipation is not the one of the energy containing eddies but the one at smaller scales? Is then $TKE^{2/3}$ calculated at 30s, 30 min intervals or 2 min? And is there other motivation for having TKE and $TKE^{2/3}$ a part for testing for its nonlinear influence?
     As stated above, we have now calculated TKE using a 30 minute averaging period. We have also removed $TKE^{2/3}$ from the set of input features used in the analysis to reduce the correlation between the variables used.

   - Turbulence data (dissipation rates included) calculated at 30s intervals have a large random error due to under-sampling. Are the authors then averaging the 30s

dissipation rates and 2 min TKE values to the 30 min period (Ln 96) to reduce this random error?
We have now addressed this issue by calculating dissipation rates every 30s, and then averaging data at a 30-minute resolution. This has been clarified in the manuscript: "We calculate ε every 30 s, and then average values at a 30-minute resolution." And again: "For each variable, we calculate and use in the machine learning algorithms 30-minute average data, to reduce the high autocorrelation in the data and limit the impact of the high-frequency large variability of turbulent quantities."

- Apart from tilt correction, are data rotated into the mean wind?
  As described at the DOI of the data (included in the data availability section), data have been rotated into a geographic coordinate system. We have now also included this specification in the manuscript.

- Given the forested nature of Perdigão, has the displacement height been taken into account? Is it assured that the measurements are above the canopy layer and roughness sublayer or are the authors testing the parametrization irrespective of the PBL layer that is probed?
  To include the effect of canopy in the machine learning models, we have now added a vegetation-related feature as input to the ML algorithms:
  - the mean vegetation height $\overline{h_{veg}}$ in the upwind 1-km radius sector centered on the measurement point. Given the forested nature of the Perdigão region, we expect canopy to have an effect in triggering turbulence, especially at lower heights. To compute this variable, we use data from a lidar survey during the season of the field campaign, at a 20 m horizontal resolution.

- What is the number of data points used when all the quality criteria are satisfied?
  We have added the following sentence in Section 2.2 "After all the quality controls have been applied, a total (from all sonic anemometers) of over 284,000 30-minute average ε data remains for the analysis."

2. Dissipation calculation

Given the very large array of different sonic anemometers, can the author discuss if there were any noticeable differences in the estimated dissipation rates? Is aliasing observed at tau = 0.1s for any of the datasets especially in stable conditions? Have the authors performed a quality control of the dissipation rate based on if the slope of the structure function is really 2/3 (plus/minus some uncertainty interval)?
We agree with the reviewer that it is important to add some quality control on the dissipation rate values used in the analysis. To this regard, we have implemented the following QC based on the propagation of errors:

To account for the uncertainty in the calculation of ϵ, we apply the law of combination of errors, which tracks how random errors propagate through a series of calculations (Barlow, 1989). We apply this method to equation 2 and quantify the fractional standard deviation in the ϵ estimates (Piper, 2001; Wildmann et al., 2019) as

$$\sigma_\epsilon = \frac{3}{2}\frac{\sigma_I}{I}\epsilon \tag{3}$$

where I is the sample mean of $\tau^{-2/3}D_U(\tau)$, and σ_I^2 is its sample variance. To perform our analysis only on lowly-uncertain ϵ values, we discard dissipation rates characterized by $\sigma_\epsilon > 0.05$. About 3% of the data are discarded based on this criterion.

3. Multivariate linear regression

- The multivariate linear regression shows the worst results of the machine learning methods used. At a first glance this comes as no surprise given that the dissipation rate is not necessarily related to other variables in a linear way, despite the fact that it is commonly accepted that dissipation and TKE are strongly coupled. The authors also mention that it is due to dissipation rate spanning multiple order of magnitude more than the TKE. However, the method might be underperforming because of a different reason. Since the response variable is the logarithm of the dissipation rate, there is no reason to expect that the predictor variables should be variables themselves rather than logarithms. For example, equation (5) shows that the parametrization of dissipation is related to TKE through:

$$\epsilon = \frac{TKE^{\frac{2}{3}}}{BL_M}$$

If we now want to see how logarithm of dissipation rate is related to TKE we see that it is related to the logarithm of TKE and not to TKE itself:

$$\log_{10}\epsilon \approx \frac{2}{3}\log_{10} TKE - \log_{10} L_M$$

Indeed, plotting $\log_{10}\epsilon$ vs TKE produces a similar shape to the one observed in Fig. 8, while $\log_{10}\epsilon$ vs $\log_{10} TKE$ are linearly related.

I expect that the multi-linear regression will produce a much more significant results with better R^2 and less bias if the predictor variables (TKE, u*, z/L) are switched with their logarithms. In the logarithmic representation there will also be only one TKE representation necessary. I suspect the same approach will produce an even better result for random forests.

We agree with the reviewer, and thank her for pointing this out. We have now modified the set of input features used in our study, and re-done the analysis accordingly. Section 4.4 describes in detail the new set of input features used:

4.4 Input features for machine-learning algorithms

Given the large variability of ϵ, which can span several orders of magnitude (Bodini et al., 2019b), we apply the machine-learning algorithms to predict the *logarithm* of ϵ. To select the set of input features used by the learning models, we take advantage of the main findings of the observational studies on the variability of ϵ to select as inputs both atmospheric- and terrain-related variables to capture the impact of topography on atmospheric turbulence. For each variable, we calculate and use in the machine learning algorithms 30-minute average data, to reduce the high autocorrelation in the data and limit the

impact of the high-frequency large variability of turbulent quantities. We use the following input features (calculated at the same location and height as ϵ) for all of the considered learning algorithms:

- wind speed (WS), which has been shown to have a moderate correlation with ϵ (Bodini et al., 2018);
- the logarithm of TKE, which is expected to have a strong connection with ϵ according to Eq. (4), calculated as

$$\log(\text{TKE}) = \log\left[\frac{1}{2}\left(\sigma_u^2 + \sigma_v^2 + \sigma_w^2\right)\right] \tag{13}$$

  where the variances of the wind components are calculated over 30-minute intervals. The choice of using the *logarithm* of TKE is justified by the fact Eq. 4 suggests this quantity is linearly related to the logarithm of ϵ;
- the logarithm of friction velocity u_*, which is calculated as

$$u_* = (\overline{u'w'}^2 + \overline{v'w'}^2)^{1/4}. \tag{14}$$

  An averaging period of 30 minutes (De Franceschi and Zardi, 2003; Babić et al., 2012) has been used to apply the Reynolds decomposition and calculate average quantities and fluctuations.
- the log-modulus transformation (John and Draper, 1980) of the ratio $\zeta = z_{\text{son}}/L$, where z_{son} is the height above the ground of each sonic anemometer, and L is the 30-minute average Obukhov length:

$$\text{sign}(\zeta)\log(|\zeta| + 1) \tag{15}$$

  The use of ζ is justified within the context of the Monin Obukhov similarity theory (Monin and Obukhov, 1954). The use of the logarithm of ζ is consistent with the use of the logarithm of ϵ as target variable. Finally, the log-modulus transformation allows for the logarithm to be calculated on negative values of ζ and be continuous in zero.
- the standard deviation std(z_{terr}) of the terrain elevation in a 1-km radius sector centered on the measurement point (i.e., the location of the sonic anemometer). The angular extension of the sector is set equal to $\pm 30°$ from the recorded 30-minute average wind direction (an example is shown in Figure 7). While we acknowledge that some degree of arbitrariness lies in the choice of this variable to quantify the terrain influence, it represents a quantity that can easily be derived from numerical models, should our approach be implemented for practical applications, to capture the influence of upwind topography to trigger turbulence. To compute this variable, we use Shuttle Radar Topography Mission (SRTM) 1 Arc-Second Global data, at 30 m horizontal resolution.
- the mean vegetation height $\overline{h_{\text{veg}}}$ in the upwind 1-km radius sector centered on the measurement point. Given the forested nature of the Perdigão region, we expect canopy to have an effect in triggering turbulence, especially at lower heights. To compute this variable, we use data from a lidar survey during the season of the field campaign, at a 20 m horizontal resolution.

The distribution of the input features and of $\log(\epsilon)$ are shown in the Supplement.

The distribution of the input features in the Supplement have been modified accordingly.

- I miss the information on what variables were chosen by the multivariate model? The results are only presented for the random forest. With so many related variables the full model should be penalized.
  We are not sure we exactly understand this comment. If the reviewer is asking about the input features used in the model, these are the same used for all three the models used in our analysis. We have specified this in Section 4.4: "We use the following input features for the three learning algorithms considered in our study:".
  If the reviewer is instead asking about the model weights (i.e. the coefficients of the multivariate regression), these are not shown as they cannot be directly

determined from the nested cross validation approach followed in our analysis. Such an approach is aimed at getting the most accurate estimate of the generalization error of the learning algorithm, but will not provide a single estimate of the model weights, as more one "optimal" model is found for each nested run. Nevertheless, we are reporting a detailed analysis of the physical interpretation of the machine learning results in Section 5.2.

- Can the authors discuss more in depth their motivation for choosing the parameters they did maybe within the Monin-Obukhov framework or HOST framework for stable conditions?
  The description of the input variables now includes more comments in this sense.

4. Influence of measurement height

Given that there are only a couple of towers that are 100 m high I am wondering about the representativity of these very high measurements as they will occupy only a very small fraction of the training. If one uses z/L then this influence will be normalized and will no longer be an issue, however, the authors use height of the sonic zson which is not normalized and therefore subject to representativity issues.

We agree with the reviewer. We have now removed z and L from the set of input features, and used instead a variable derived from log(z/L):

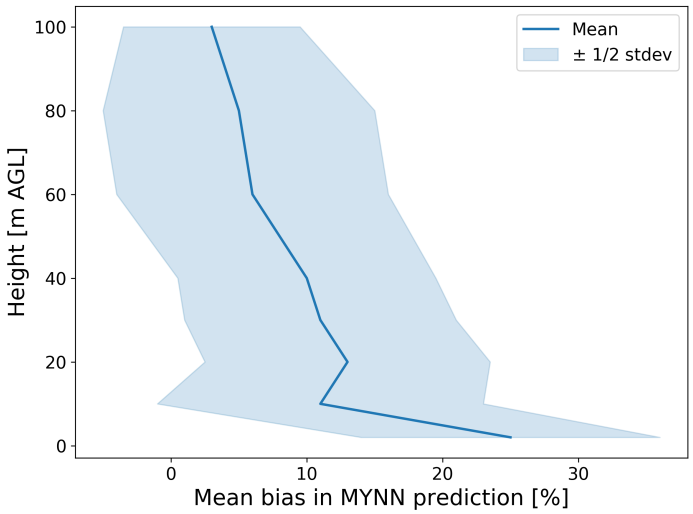
– the log-modulus transformation (John and Draper, 1980) of the ratio $\zeta = z_{\text{son}}/L$, where z_{son} is the height above the ground of each sonic anemometer, and L is the 30-minute average Obukhov length:

$$\text{sign}(\zeta)\log(|\zeta|+1) \tag{15}$$

The use of ζ is justified within the context of the Monin Obukhov similarity theory (Monin and Obukhov, 1954). The use of the logarithm of ζ is consistent with the use of the logarithm of ϵ as target variable. Finally, the log-modulus transformation allows for the logarithm to be calculated on negative values of ζ and be continuous in zero.

In the same way I wonder about the results of Figure 6 in which the mean bias according to height of MYNN is shown. The results for lower heights will include a more varied set of conditions than for higher heights. I would find it justified to compare the bias for different heights only on the towers with similar heights (for example the two 100m towers).

We have added some error quantification to Figure 6 to quantify the spread of the results shown at each height:



We have also performed the same analysis only using data from the three 100-m towers, and added a comment in the main paper and a figure in the Supplementary Information: "We obtain comparable results when computing the bias in the MYNN parameterization only for the sonic anemometers mounted on the three 100-m meteorological towers (Figure shown in the Supplement), thus confirming that the observed trend is not due to the larger variability of the conditions sampled by the more numerous sonics at lower heights. Therefore, our results show how the MYNN formulation fails in accurately representing atmospheric turbulence especially in the lowest part of the boundary layer."

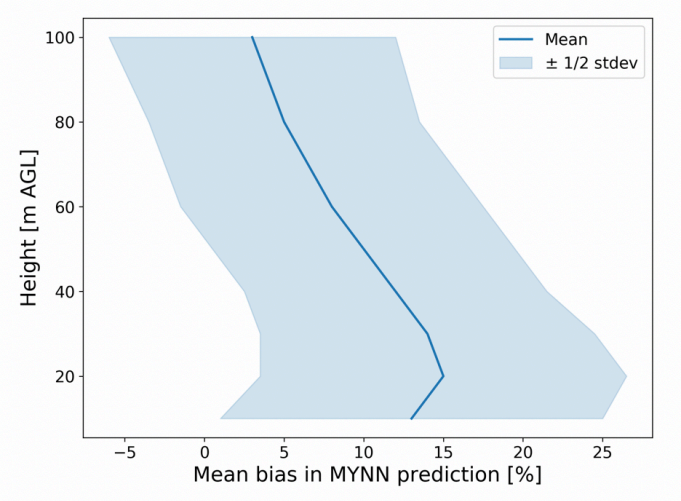


Figure S1: Mean bias in the MYNN-parameterized log(ϵ) at different heights, as calculated from the sonic anemometers on the three 100-m towers at Perdigão.

5. Terrain influence

The terrain influence in the paper is quantified through a standard deviation of the terrain within 1 km upstream of the measurement. I presume that this is because such a variable is readily available in numerical models, but this motivation is missing in the paper. On the other hand, from a physical point of view I am wondering how this variable can be justified. Given the variety of measurement heights in the dataset the flux footprint and therefore also the terrain that influences the measurement is going to vary substantially. It would

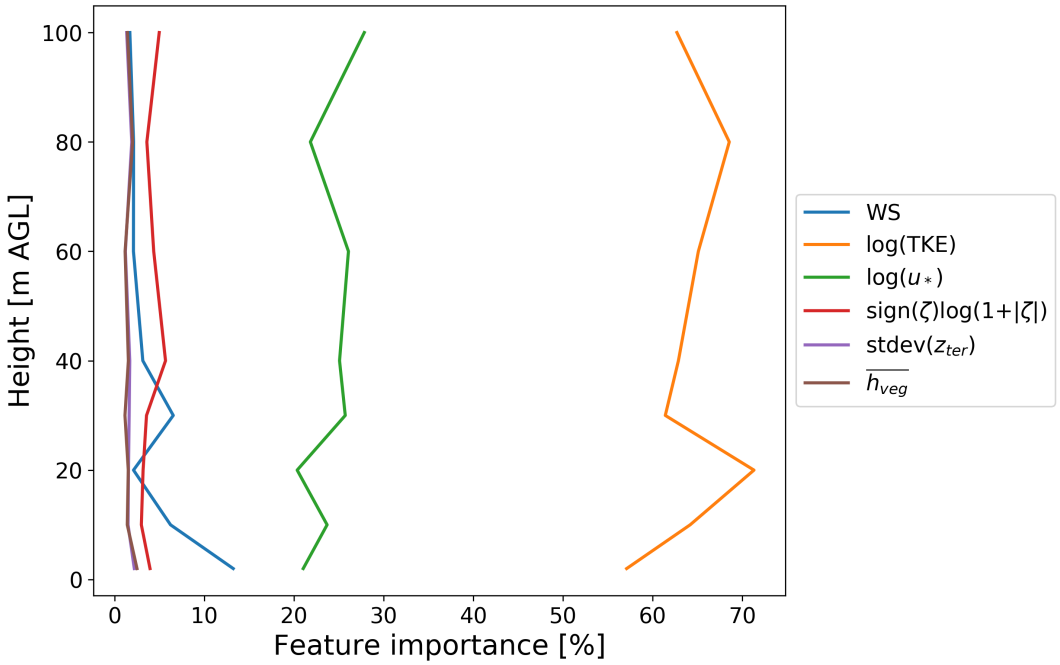
therefore be good to either motivate the choice in detail, or to present some footprint analysis which convinces us that the choice of 1 km is meaningful. I also miss the information on how this standard deviation is computed. What is the resolution of the digital elevation model used for this computation? And what is the reasoning behind using standard deviation as opposed to for example slope angle? Given the change of the footprint with height, wouldn't it be more appropriate to estimate the effect of terrain only for measurements with similar height?

In the description of the 'new' set of input features used (see answer to specific comment #2) we have added a comment on how the standard deviation of upwind terrain has been chosen as it can be easily computed from numerical models. We have also added details on the DEM dataset used to compute this variable in our analysis.

In Section 5.2, we state that "Though not negligible, the importance of topography and canopy might increase by considering different parameters that could better encapsulate their effect."

Finally, we have performed an additional analysis on the importance of the input features for the random forest prediction when single heights are considered:

"We have tested how the feature importance varies when considering several random forests, each trained and tested with data from all the sonic anemometers at a single height only, and did not find any significant variation of the importance of the considered variables in predicting ε (plot shown in the Supplement)."



6. Separation according to stability

Results of Fig 5 show very large difference in the success of the parametrization for stable and unstable stratification. Looking at the results I would say that there is visually almost no need for improvements on the unstable side. With this in mind, I wonder why the approach is then followed which lumps all the data together.

We have now added a more detailed analysis of the random forest results based on stability:

Given the large gap in the performance of the MYNN parameterization of ϵ between stable and unstable conditions, it is worth exploring how the machine learning algorithms perform in different stability conditions. To do so, we train and test two separate random forests: one using data observed in stable conditions, the other one for unstable cases. We find that both algorithms eliminate the bias observed in the MYNN scheme (Figure 9). The random forest for unstable conditions provides, on average, more accurate predictions (RMSE = 0.37, MAE = 0.28) compared to the algorithm used for stable cases (RMSE 0.44, MAE = 0.33), thus confirming the complexity in modeling atmospheric turbulence in quiescent conditions. However, when the error metrics are compared to those of the MYNN parameterization, the random forest for stable conditions provides the largest relative improvement, with a 50% reduction in MAE, while for unstable conditions the reduction is of 20%.

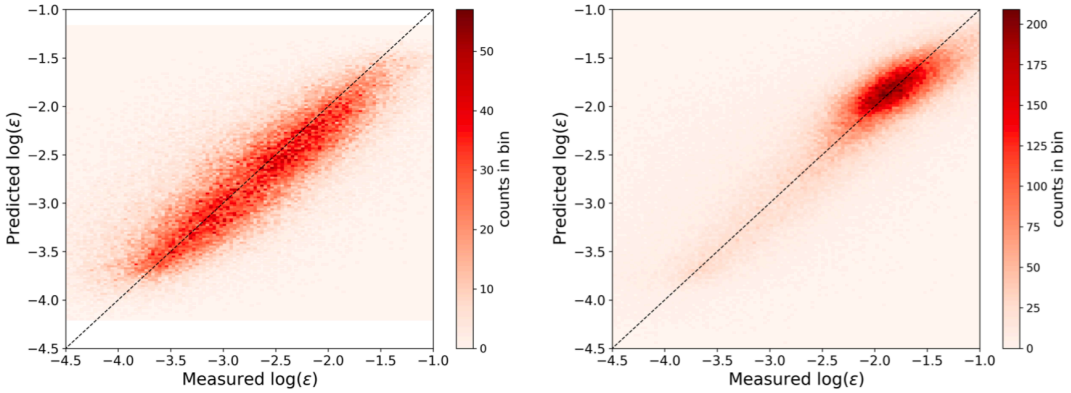


Figure 9. Density histogram showing the comparison, performed on the testing set, between observed and machine-learning-predicted ϵ from a random forest for stable conditions (left) and unstable conditions (right).

7. Paper structure

The paper structure could be improved if machine learning algorithms were introduced before the predictor variables that are used to feed these algorithms.

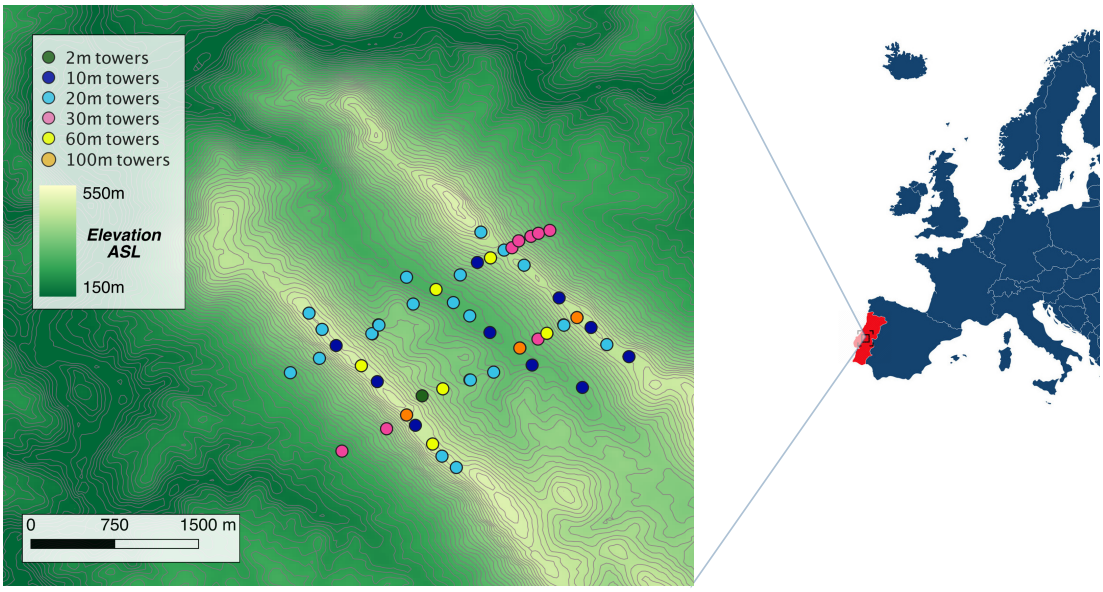
We have changed the structure of the paper following your feedback, and the machine learning algorithms are now presented before the input features.

Minor points

1. Ln 35: "; for example" should be ": for example"
   Changed.

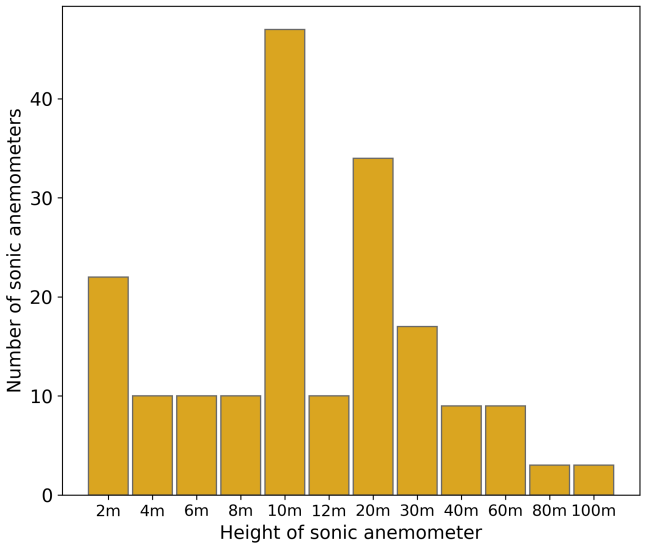
2. Figure 1: It would be good to color the points according to the height of the tower their represent

Done:



3. Ln 69: “are recorded” should be “were recorded”
   Changed.

4. Figure 2: given that this figure is only for presentation purposes I would suggest replacing a histogram for a bar plot which correctly represents the measurement heights. This could still be done in some meaningful increments but would not bundle 2m heights under 0 and would not have gaps for say 90 m height which does not exist
   We have replaced the figure with the following:



We have also added the following table to make the information provided more detailed:

Table 1. Details on heights where sonic anemometers were mounted on the meteorological towers at the Perdigão field campaign.

Nominal tower height	Sonic anemometer heights (m AGL)	Number of towers
2 m	2	1
10 m	10	5
	2, 10	5
20 m	10, 20	10
	2, 10, 20	6
	2, 4, 6, 8, 10, 12, 20	4
30 m	10, 30	3
	2, 4, 6, 8, 10, 12, 20, 30	5
60 m	10, 20, 30, 40, 60	5
	2, 4, 6, 8, 10, 12, 20, 30, 40, 60	1
100 m	10, 20, 30, 40, 60, 80, 100	3
	Total number of towers	48
	Total number of sonic anemometers	184

5. Ln 75-77: it is not necessary to mention that it is a structure function of horizontal velocity twice in this sentence
   We have rephrased as follows:
   TKE dissipation rate from the sonic anemometers on the meteorological towers is calculated from the second-order structure function $D_U(\tau)$ of the horizontal velocity U (Muñoz-Esparza et al., 2018):

$$\epsilon = \frac{1}{U\tau}\left[aD_U(\tau)\right]^{3/2} \quad (1)$$

   where τ indicates the temporal increments over which the structure function is calculated, and $a = 0.52$ is the one-dimensional

6. Ln 79: the part of the sentence "is done using the temporal separation between" is not very clear. Do you mean that you calculate the dissipation rate for lags between 0.1 and 2s by assuming that this is the inertial subrange?
   We have rephrased the sentence as: "We calculate ε every 30 s, and then average values at a 30-minute resolution. At each calculation of ε, we fit experimental data to the Kolmogorov model (Kolmogorov, 1941; Frisch, 1995) using time lags separation between τ_1 = 0.1 s and τ_2 = 2 s, which represent a conservative choice to approximate the inertial subrange (Bodini et al., 2018).".

7. Ln 87: algorithms haven't been introduced yet
   See answer to specific comment #7.

8. Ln 102: this is not sensible heat flux but buoyancy flux, given that the authors mention no Schotanus correction. Also, is θ_V really virtual temperature or rather sonic temperature?
   We have corrected this sentence and stated we are using buoyancy flux. We have also specified that "ϑ_v is the virtual potential temperature (K, here approximated as the sonic temperature)".

9. Ln 103: Why are the authors using values of L to define stability ranges, when it is more common to define them through z/L, where neutral stratification has a clear meaning, whereas L is not as clearly specified?
   We have now classified atmospheric stability based on z/L instead of L: "For atmospheric stability, we classify unstable conditions as $\zeta = z/L < -0.02$; and stable conditions as $\zeta > 0.02$; nearly-neutral conditions as $|\zeta| \leq 0.02$."

10. Ln 129: Given the many profiles that exist in the data, I wonder why it is impossible to estimate the L_T and L_B scales. The TKE is not expected to vary so erratically to not be possible to estimate its vertical variability with an analytical function.
    While we agree with the reviewer that some assumptions could be made to approximate the other two length scales, we think this is not strictly necessary in the context of our paper. To better explain this point, we have added the following comment: "The observed bias would be even larger if LM was calculated including all the contributions according to Eq. (5), and not Ls only as in our approximation. Therefore, while the approximation in Eq. (9) is major and could be eased by making assumptions on the vertical profile of TKE at Perdigão, it does not affect the conclusion of a high inaccuracy in the MYNN parameterization of ε."
    We have also added to the Supplementary Information the analytical proof that our approximation determines an overestimation of LM.

11. Ln 140: so is TKE then calculated at 30s?
    See answer to your specific comment #1.

12. Ln 163: What do you mean by "time stamps with missing data"? Do you mean that only those periods when all the instruments had all the values were used?
    We have clarified as: "No data imputation was performed, and missing data were removed from the analysis."

13. Ln 165 – 166: What do you mean by hyperparameters? Are you referring to the ones defined in Table 1? This should be referenced here.
    We have rephrased as "hyperparameters (model parameters whose values are set before the training phase and that control the learning process)".
    Table 1 only shows the hyperparameters of the random forest, while the linear and polynomial regression only have one hyperparameter (i.e. the alpha parameter for Ridge regression). To make this clear, we have added the following sentence: "Before testing the models, however, it is important to avoid overfitting by setting the values of hyperparameters. Each learning algorithm has specific model-specific hyperparameters that need to be considered, as will be specified in the description of each algorithm."

14. Ln 194: Mention that Scikit-learn is a python library.
    We have rephrased as "python's library Scikit-learn".

15. Ln 195: what are the variables chosen by the ridge regression?
    See answer to specific comment #3.

16. Ln 223: I find this sentence not very clear. Values of what were sampled in the cross-validation search? And what do you mean by five sets of parameters?
    We have clarified the sentence as: "Table 2 describes which hyperparameters we considered for the random forest algorithm. For each hyperparameters listed, we include the range of values that are randomly sampled in the cross-validation search to form the ten sets of hyperparameters used in the training phase."

17. Table 1: How were these values chosen?
    For some hyperparameters, the choice of their values is constrained by the problem: for example, the maximum number of features has to be picked based on the number of features of the specific problem. For other parameters, the minimum value is often 1, while the maximum sampled values are chosen (after some empiric tests and/or past experience) to avoid allowing for a model that is complicated enough to overfit the problem.

18. Ln 232: How do you explain this "optimistic result" that using a reduced parametrization is actually beneficial to using the full one?
    We have clarified what we mean by "optimistic result": "We note that, because the length scale approximation we made in calculating MYNN-predicted ε led to a better agreement with the observed values compared to what would be obtained using the full MYNN parameterization, the RMSE and MAE for the MYNN case would in reality be higher than what we report here, and so the error reductions achieved with the machine-learning algorithms would even be greater than the numbers shown in the Table."

19. Ln 252: Is R^2 the adjusted one that takes into account the penalization for overfitting? Are all the variables statistically significant and at which p value?
    To remove ambiguity and be consistent with the error metrics used throughout the paper, we have removed R^2 from the table.

20. Ln 265: Within Monin-Obukhov similarity theory L is not the relevant variable but z/L. The use of logarithm of (z/L) might improve the importance of this variable.
    As already mentioned, we have now used a variable derived from log(z/L) as input feature for the machine learning algorithms.