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

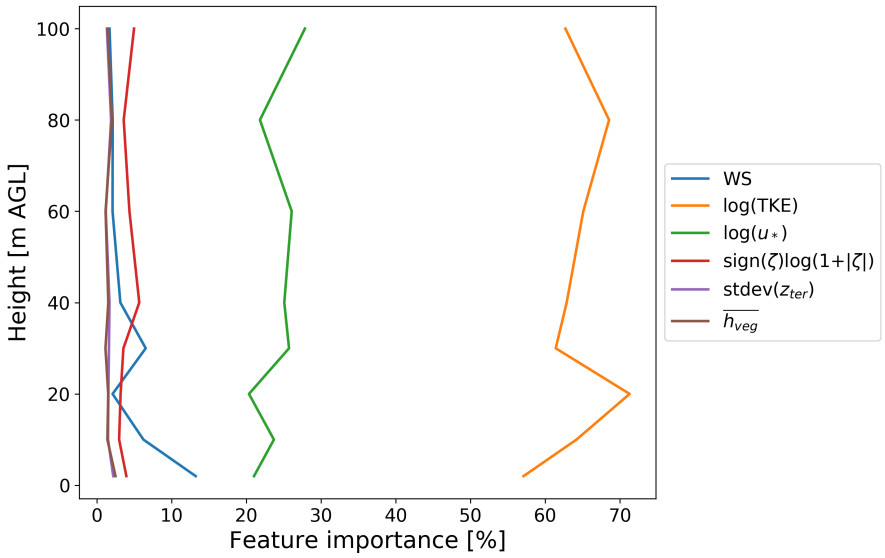
The authors thank the reviewer for their thoughtful and productive comments.

The manuscript "Can machine learning improve the model representation of TKE dissipation rate in the boundary layer for complex terrain?" by Bodini et al. provides an interesting look at using machine learning techniques to generate estimates of TKE dissipation rate and comparing those results to the approach used in the MYNN parameterization. This work should be of interest to the community and provides a useful road map for scientists wanting to apply a similar approach to other data sets. Overall, I think the manuscript will be acceptable for publication in Geoscientific Model Development after relatively minor revisions. The text is generally clearly written and straight forewarned to follow. I wonder, given that focus on data analysis rather than atmospheric model development, if the manuscript is a better fit for Atmosphere Chemistry and Physics or Atmospheric Measurement Techniques. I leave that, however, up to the editor.

Thank you for finding our work interesting and well-structured. Regarding the choice of the journal, we would like to emphasize that GMD has already published at least another paper (reference below) with a focus similar to ours, and therefore we think that adding another publication on the topic in the same journal would strengthen both papers. In addition, the focus of our work is on explaining weaknesses in MYNN parameterization and working towards a possible replacement, hence we think this fits into GMD's scope of "new methods for assessment of models, including work on developing new metrics for assessing model performance and novel ways of comparing model results with observational data".
Leufen, L. H. and Schädler, G.: Calculating the turbulent fluxes in the atmospheric surface layer with neural networks, Geosci. Model Dev., 12, 2033–2047, https://doi.org/10.5194/gmd-12-2033-2019, 2019.

General comments

- Machine learning techniques generally do not increase our physical understanding. The authors try to address this in Section 5.1 and 5.2 where additional analysis is provided. Section 5.2, however, is very brief and should be developed more to provide additional insight into the results.
  To give more importance to the physical interpretation of the machine learning results, we have now unified Sections 5.1 and 5.2 and used "Physical interpretation of machine learning results" as header.
  We have also added a new analysis on the performance of the random forest for different stability conditions – see answer to the next general comment.
  In addition, we have added more comments on the description of the partial dependence analysis, and added plots for all the input features used.
  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)."



- In section 3, the authors show that the MYNN approach does a reasonable job in unstable conditions, but much worse when the boundary layer is statically stable. I was surprised that the authors didn't carry this analysis into the subsequent sections. It would seem natural to examine the model behavior with stability in Section 5.

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%.

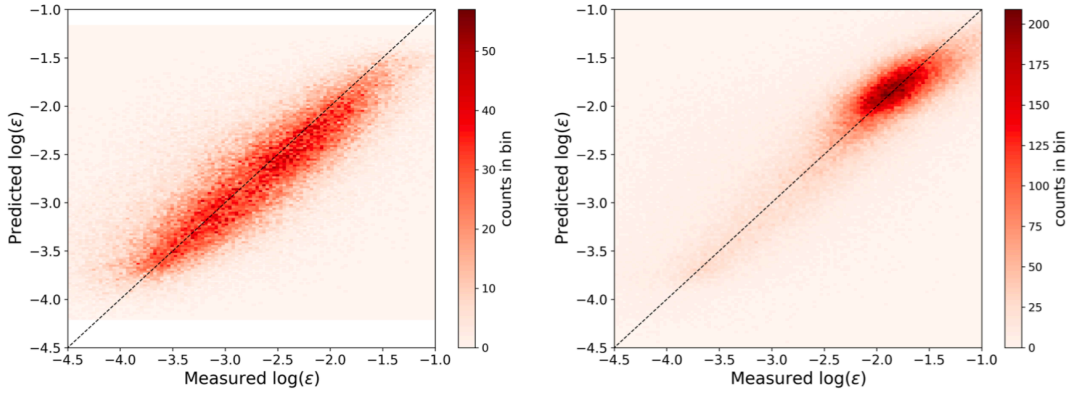
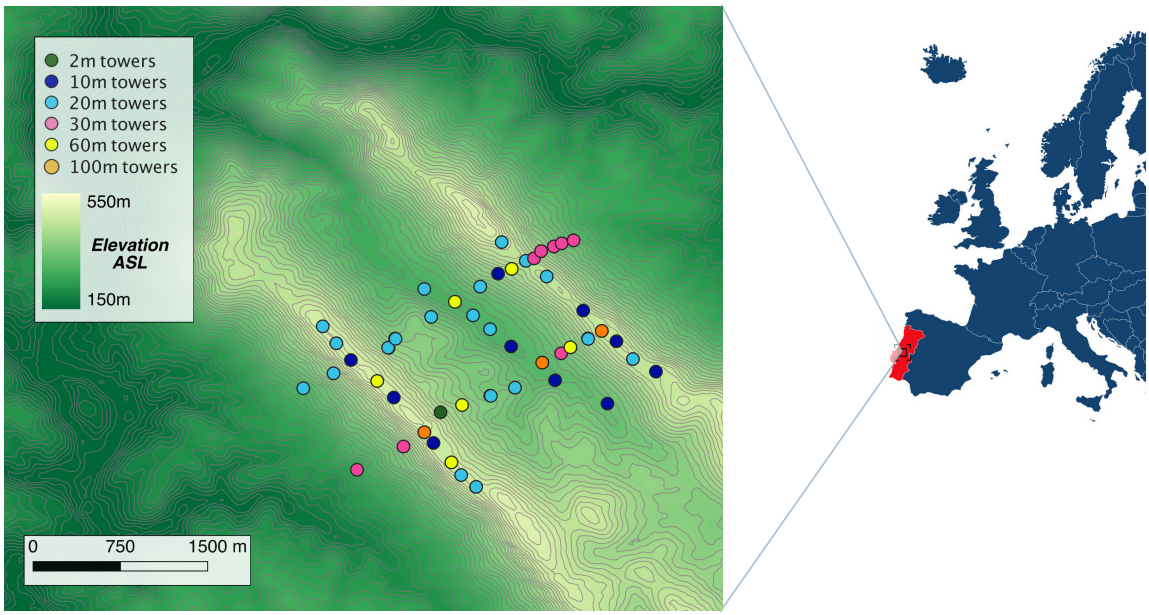


Figure 9. Density histogram showing the comparison, performed on the testing set, between observed and machine-learning-predicted ϵ from a random forest for stable conditions (left) and unstable conditions (right).

Specific comments

1. Figure 1. I appreciate the histogram shown in Figure 2, but could you also differentiate the points in Figure 1 to indicate measurement heights? Maybe that doesn't work well if the measurements made at a single location are at several heights?
   Yes, multiple sonics at several heights were installed on each tower. However, to give the reader a better idea of the distribution of the tower heights, we have changed the map to reflect this information:



2. Section 2.1: Can you say anything more about how the sonics are distributed on the towers? For example, how many were deployed on the 100 m tower?
   We have added the following table to include more details on the measurement heights of the sonic anemometers:

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

3. Lines 78-80: Double check this sentence, the wording seems odd.

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 $\tau_1 = 0.1$ s and $\tau_2 = 2$ s, which represent a conservative choice to approximate the inertial subrange (Bodini et al., 2018).".

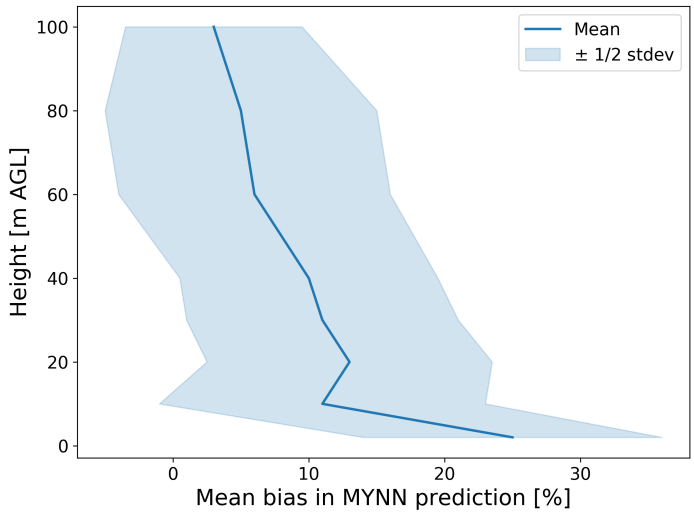
4. Line 101: Is the mean potential temperature computed from the sonic data or does it come from a different source?
   Yes, and we have now specified it: "ϑ_v is the virtual potential temperature (K, here approximated as the sonic temperature)".

5. Lines 104-109: Can you point the reader to the terrain data set that was used? What was the resolution of that data set? Does that have any impact on the results?
   We have added additional details on this:
   - 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 ±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.

   We have also included the relevant information in the Data availability section.

6. Line 138: I agree that the length scale assumption is the best you can do given the data set that you have, but I think some additional discussion is warranted to help defend that selection. Can you argue that Ls is likely dominate near the surface?
   To better explain our approximation, and why we don't think that additional assumptions are strictly needed for our analysis, 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.

7. Figure 6: You show the mean bias in Figure 6, could bars be added to indicate the standard deviation of the bias? This would help show how significant the biases are. In addition, the figure shows a decrease with height. Is this significant, or could it (at least partially) be related to the horizontal distribution of the measurements taken at different heights?
   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."

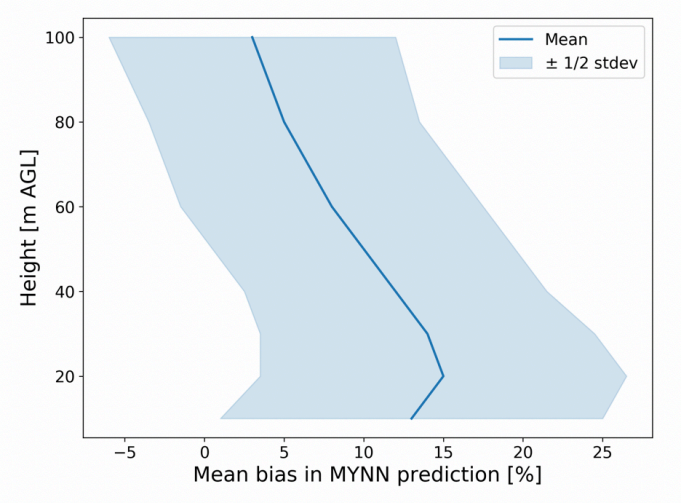


Figure S1: Mean bias in the MYNN-parameterized log(ϵ) at different heights, as calculated from the sonic anemometers on the three 100-m towers at Perdigão.

8. Section 4: It would be helpful if you could include a brief discussion of why you selected these particular algorithms for this application.
We have added the following comment: "Given the proof-of-concept nature of this analysis in proving the capabilities of machine learning to improve numerical model parameterizations, we defer an exhaustive comparison of different machine-learning models to a future study, and only consider relatively simple algorithms in the present work."

9. Section 5.2: Is there a better header for this section to help the reader understand the importance of the analysis that is presented?
   We have unified Sections 5.1 and 5.2 and used "Physical interpretation of machine learning results" as header.

10. Section 5.2: This section seems to end abruptly. Can you guide the reader to anything important? What additional insight is gained from the analysis? What does it tell us about what is controlling the dissipation rate at large values of wind speed and/or TKE?
    See answer to general comments #1 and #2.