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):

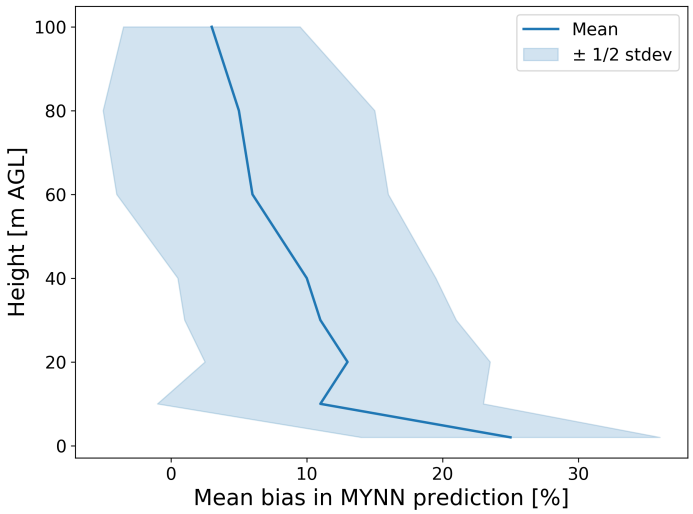
– the log-modulus transformation (John and Draper, 1980) of the ratio $\zeta = z_{\text{son}}/L$, where z_{son} is the height above the ground of each sonic anemometer, and L is the 30-minute average Obukhov length:

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

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."

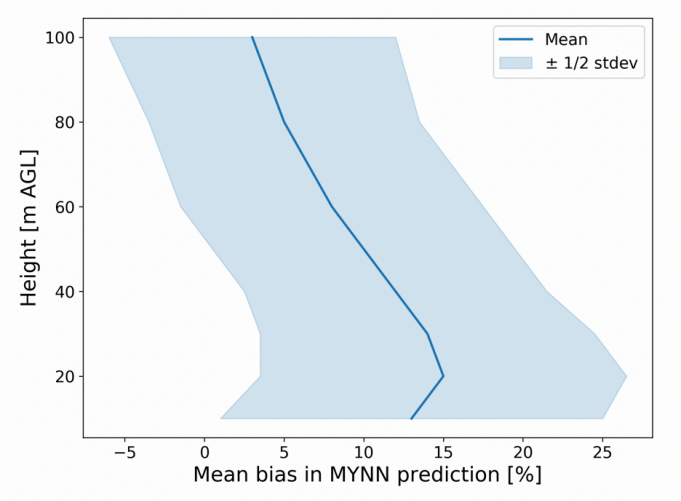


Figure S1: Mean bias in the MYNN-parameterized log(ϵ) at different heights, as calculated from the sonic anemometers on the three 100-m towers at Perdigão.

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

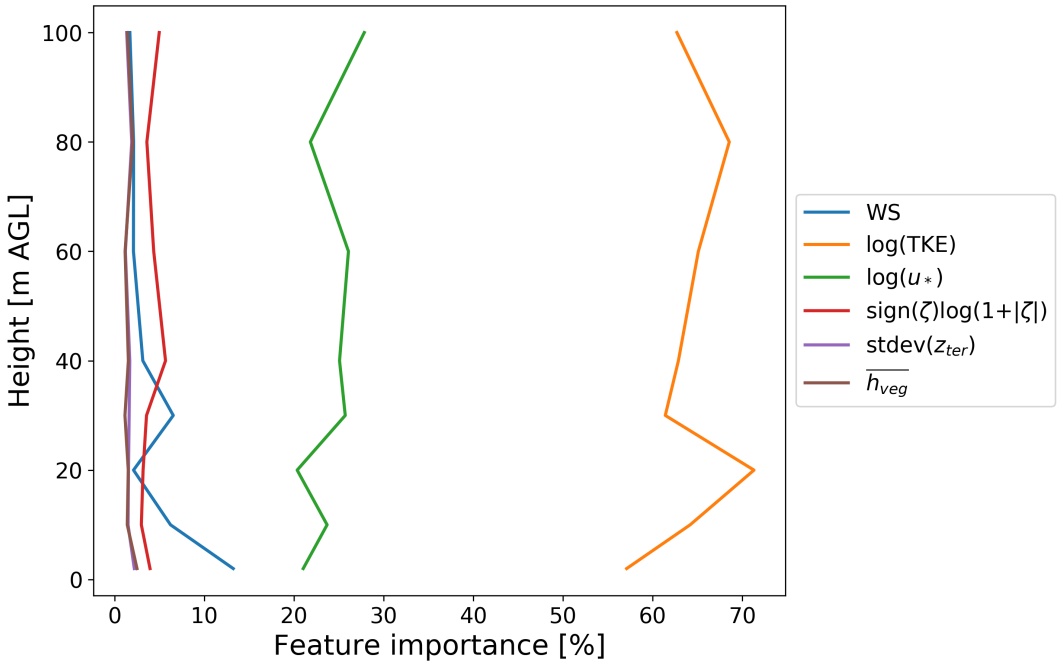
therefore be good to either motivate the choice in detail, or to present some footprint analysis which convinces us that the choice of 1 km is meaningful. I also miss the information on how this standard deviation is computed. What is the resolution of the digital elevation model used for this computation? 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%.

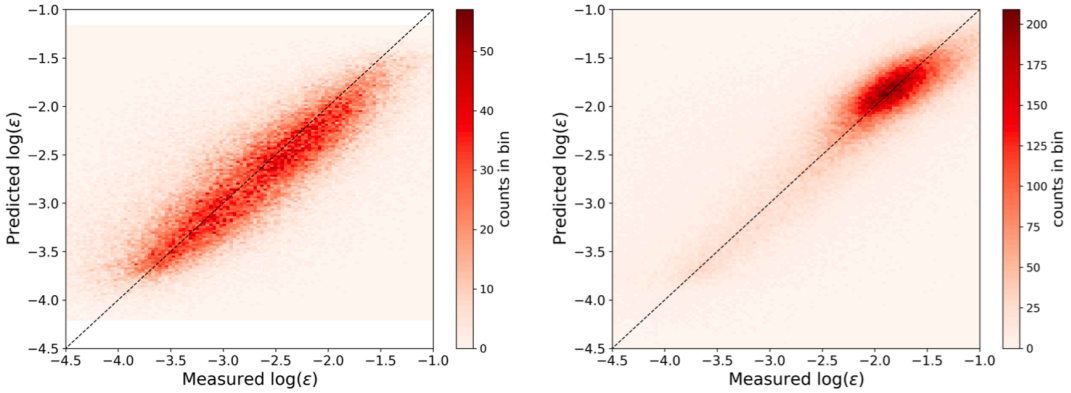


Figure 9. Density histogram showing the comparison, performed on the testing set, between observed and machine-learning-predicted ϵ from a random forest for stable conditions (left) and unstable conditions (right).

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.

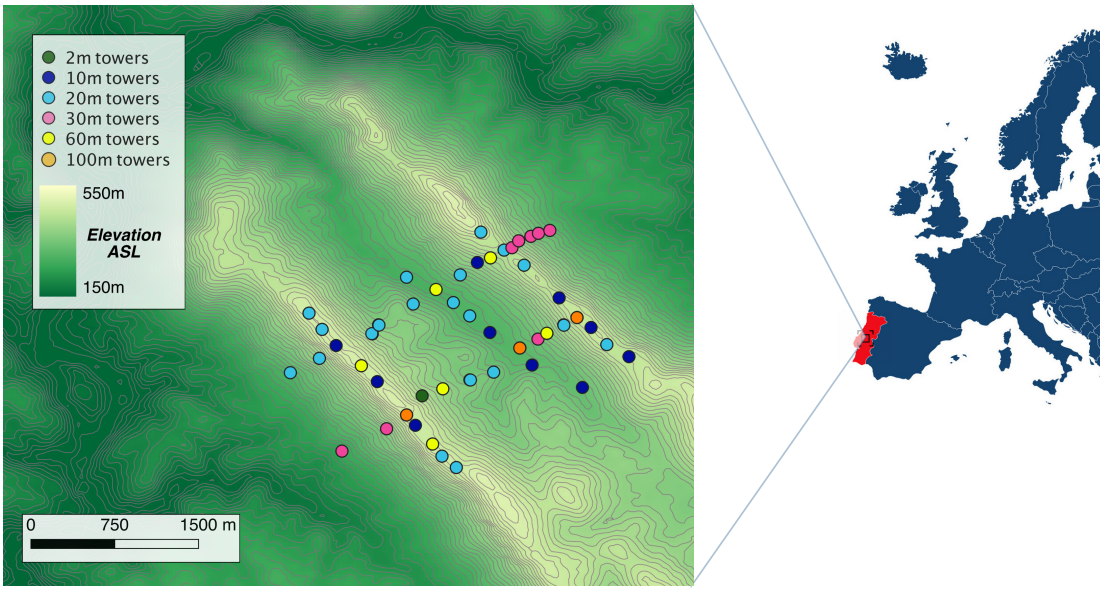
We have changed the structure of the paper following your feedback, and the machine learning algorithms are now presented before the input features.

Minor points

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

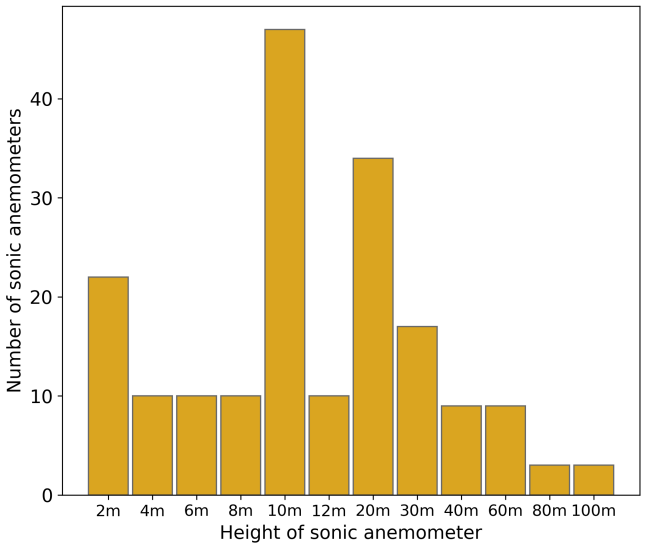
2. Figure 1: It would be good to color the points according to the height of the tower their represent

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.

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

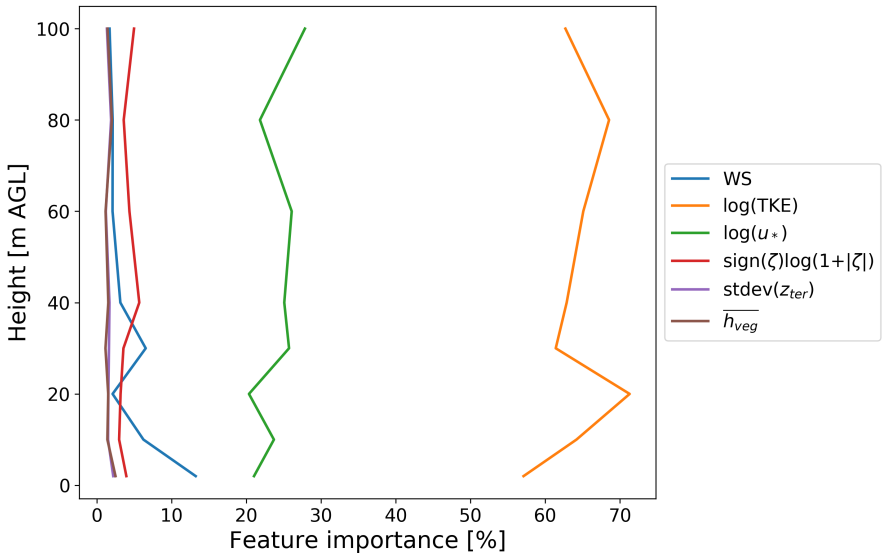
The authors thank the reviewer for their thoughtful and productive comments.

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%.

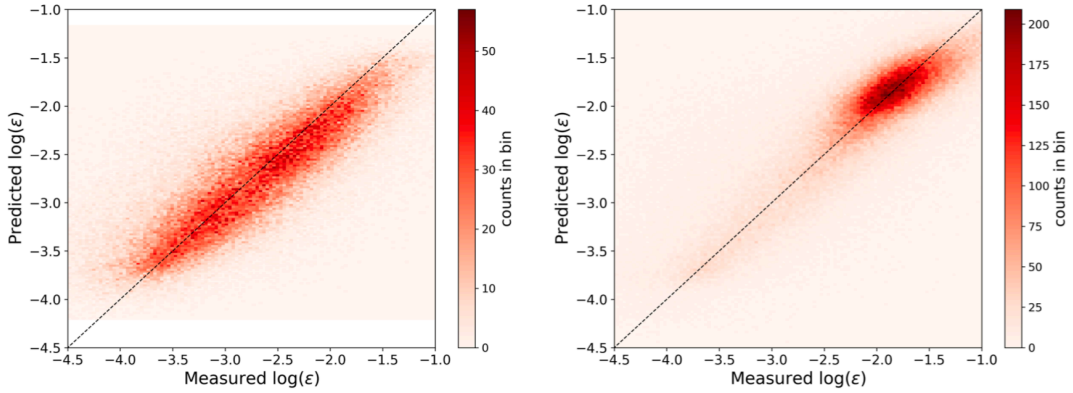
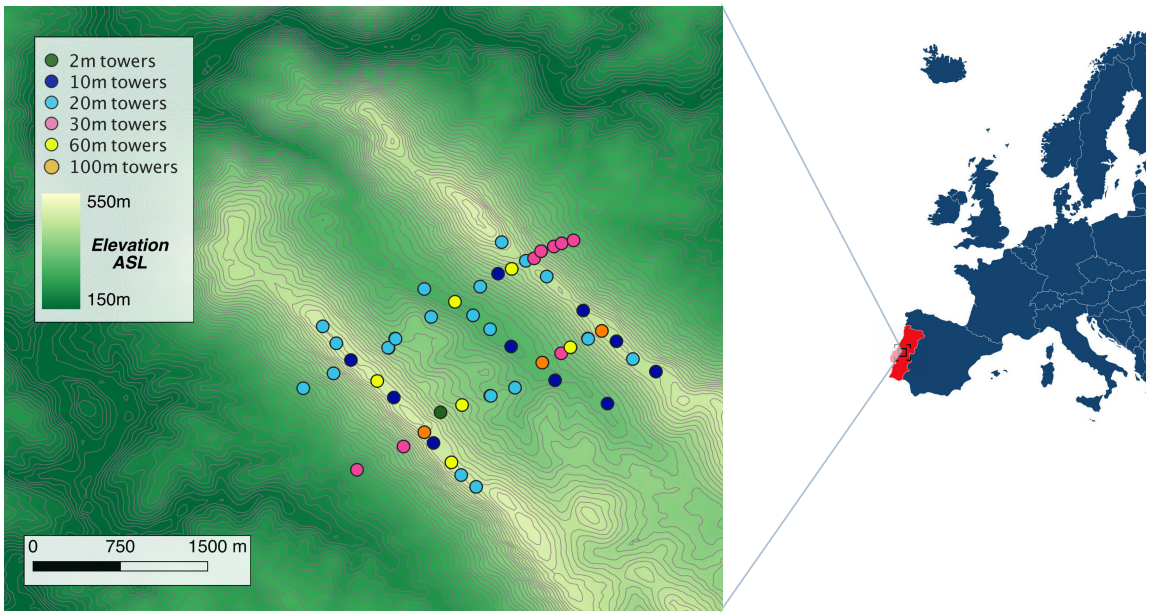


Figure 9. Density histogram showing the comparison, performed on the testing set, between observed and machine-learning-predicted ϵ from a random forest for stable conditions (left) and unstable conditions (right).

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).".

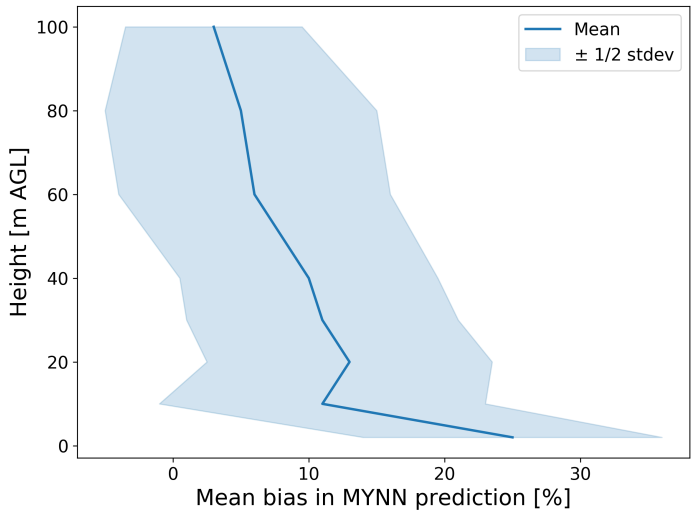
4. Line 101: Is the mean potential temperature computed from the sonic data or does it come from a different source?
   Yes, and we have now specified it: "ϑ_v is the virtual potential temperature (K, here approximated as the sonic temperature)".

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."

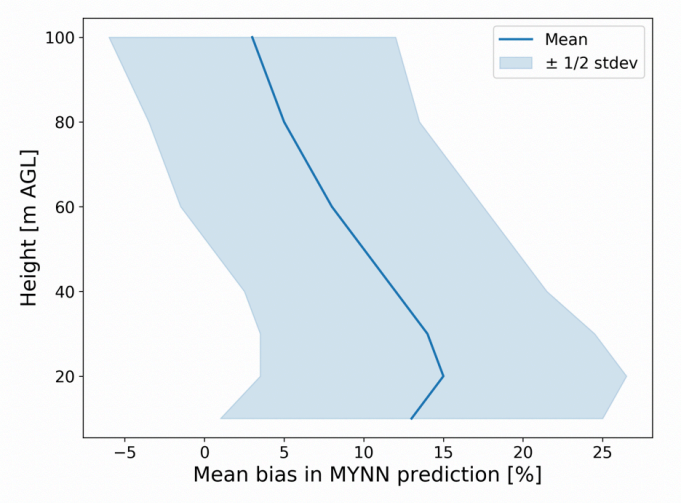


Figure S1: Mean bias in the MYNN-parameterized log(ϵ) at different heights, as calculated from the sonic anemometers on the three 100-m towers at Perdigão.

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.

Can machine learning improve the model representation of TKE dissipation rate in the boundary layer for complex terrain?

Nicola Bodini[1,2], Julie K. Lundquist[1,2], and Mike Optis[2]

[1]Department of Atmospheric and Oceanic Sciences, University of Colorado Boulder, Boulder, Colorado, USA
[2]National Renewable Energy Laboratory, Golden, Colorado, USA

Correspondence: Nicola Bodini (nicola.bodini@nrel.gov)

Abstract. Current turbulence parameterizations in numerical weather prediction models at the mesoscale assume a local equilibrium between production and dissipation of turbulence. As this assumption does not hold at fine horizontal resolutions, improved ways to represent turbulent kinetic energy (TKE) dissipation rate (ϵ) are needed. Here, we use a 6-week data set of turbulence measurements from 184 sonic anemometers in complex terrain at the Perdigão field campaign to suggest improved representations of dissipation rate. First, we demonstrate that ~~a~~ the widely used Mellor, Yamada, Nakanishi, and Niino (MYNN) parameterization of TKE dissipation rate leads to a large inaccuracy and bias in the representation of ϵ. Next, we assess the potential of machine-learning techniques to predict TKE dissipation rate from a set of atmospheric and terrain-related features. We train and test several machine-learning algorithms using the data at Perdigão, and we find that ~~multivariate polynomial regressions and random forests can~~ the models eliminate the bias MYNN currently shows in representing ϵ, while also reducing the average error by up to ~~30~~almost 40%. Of all the variables included in the algorithms, TKE is the variable responsible for most of the variability of ϵ, and a strong positive correlation exists between the two. These results suggest further consideration of machine-learning techniques to enhance parameterizations of turbulence in numerical weather prediction models.

1 Introduction

While turbulence is an essential quantity that regulates many phenomena in the atmospheric boundary layer (Garratt, 1994), numerical weather prediction models are not capable of fully resolving it. Instead, they rely on parameterizations to represent some of the turbulent processes. Investigations into model sensitivity have shown that out of the various parameterizations

currently used in mesoscale models, that of turbulent kinetic energy (TKE) dissipation rate (ϵ) has the greatest impact on the accuracy of model predictions of wind speed at wind turbine hub height (Yang et al., 2017; Berg et al., 2018).

Current boundary layer parameterizations of ϵ in mesoscale models assume a local equilibrium between production and dissipation of TKE. While this assumption is generally valid for homogeneous and stationary flow (Albertson et al., 1997), as the horizontal grid resolution of mesoscale models is constantly pushed toward finer scales thanks to the increase of the computing resource capabilities, the theoretical bases of this assumption are violated. In fact, turbulence produced within a model grid cell can be advected farther downstream in a different grid cell before being dissipated (Nakanishi and Niino, 2006; Krishnamurthy et al., 2011; Hong and Dudhia, 2012).

The inaccuracy of the mesoscale model representation of ϵ impacts a wide variety of processes that are controlled by the TKE dissipation rate. In fact, the dissipation of turbulence affects the development and propagation of forest fires (Coen et al., 2013), it has consequences on aviation meteorology and potential aviation accidents (Gerz et al., 2005; Thobois et al., 2015), it regulates the dispersion of pollutants in the boundary layer (Huang et al., 2013), and it affects wind energy applications (Kelley et al., 2006)~~;~~ : for example, in terms of the development and erosion of wind turbine wakes (Bodini et al., 2017).

Several studies have documented the variability of ϵ using observations from both in-situ (Champagne et al., 1977; Oncley et al., 1996; Frehlich et al., 2006) and remote-sensing instruments (Frehlich, 1994; Smalikho, 1995; Shaw and LeMone, 2003). Bodini et al. (2018, 2019b) showed how ϵ has strong diurnal and annual cycles onshore, with topography playing a key role in triggering its variability. On the other hand, offshore turbulence regimes (Bodini et al., 2019a) are characterized by smaller values of ϵ, with cycles mostly impacted by wind regimes rather than convective effects. Also, ϵ greatly increases in the wakes of obstacles, for example wind turbines (Lundquist and Bariteau, 2015; Wildmann et al., 2019) or whole wind farms (Bodini et al., 2019b).

This knowledge on the variability of TKE dissipation rate provided by observations lays the foundation to explore innovative ways to improve the model representation of ϵ in the atmospheric boundary layer. In this study, we leverage the potential of machine-learning ~~techiniques~~ techniques to explore their potential application to improve the parameterizations of ϵ. Machine-learning techniques can successfully capture the complex and nonlinear relationship between multiple variables without the need of representing the physical process that governs this relationship. They have been successfully used to advance the understanding of several atmospheric processes, such as convection (Gentine et al., 2018), turbulent fluxes (Leufen and Schädler, 2018), and precipitation nowcasting (Xingjian et al., 2015). The renewable energy sector has also experienced various applications of machine-learning techniques, in both solar (Sharma et al., 2011; Cervone et al., 2017) and wind (Giebel et al., 2011; Optis and Perr-Sauer, 2019) power forecasting. Applications have also been explored at the wind turbine level, for turbine power curve modeling (Clifton et al., 2013), turbine faults and controls (Leahy et al., 2016), and turbine blade management (Arcos Jiménez et al., 2018).

Here, we train and test different machine-learning algorithms to predict ϵ from a set of atmospheric and topographic variables. Section 2 describes the Perdigão field campaign and how we retrieved ϵ from the sonic anemometers on the meteorological towers. In Section 3, we then evaluate the accuracy of one of the most common planetary boundary layer parameterization schemes used in numerical weather prediction: the Mellor, Yamada, Nakanishi, and Niino (MYNN) parameterization scheme

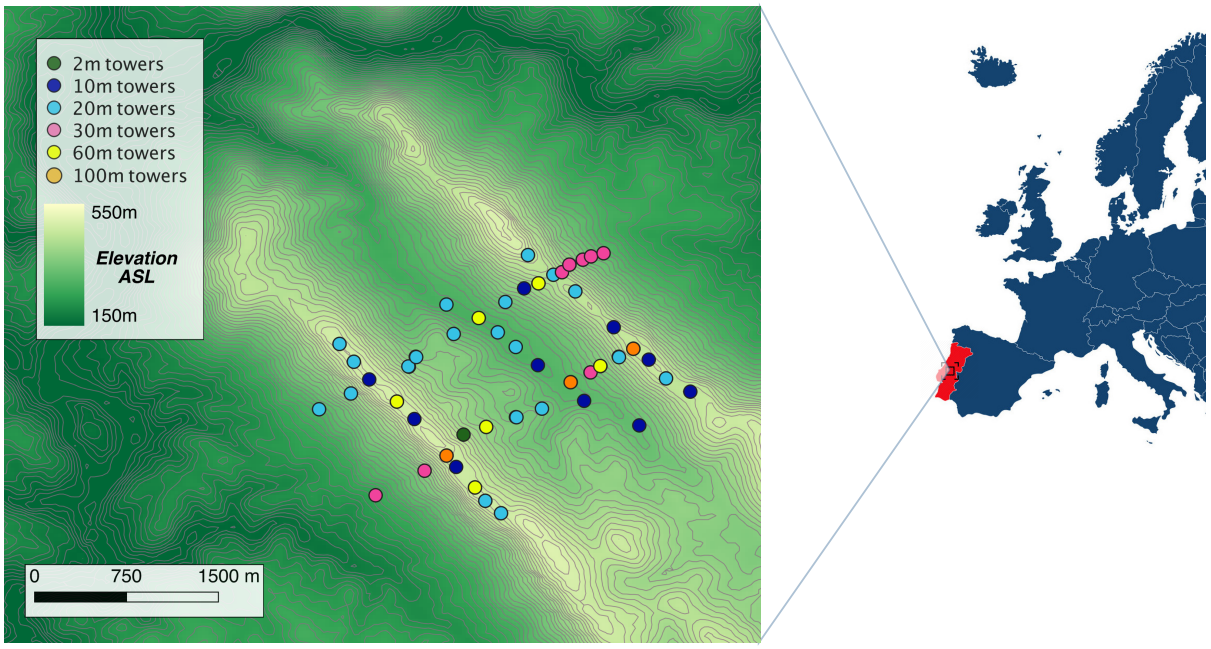


Figure 1. Map of the Perdigão valley showing the location and height of the 48 meteorological towers whose data are used in this study. Digital elevation model data courtesy of the U.S. Geological Survey.

(Nakanish, 2001). Section 4 presents the machine-learning algorithms that we used in our analysis. The results of our study are shown in Section 5, and discussed in Section 6, where future work is also suggested.

2 Data

2.1 Meteorological towers at the Perdigão field campaign

The Perdigão field campaign (Fernando et al., 2018), an international cooperation between several universities and research institutes, brought an impressive number of instruments to a valley in central Portugal to survey the atmospheric boundary layer in complex terrain. The Perdigão valley is limited by two mountain ridges running from northwest to southeast (Figure 1), separated by ∼1.5 km. The intensive operation period (IOP) of the campaign, used for this study, was from 1 May to 15 June 2017.

At Perdigão, 184 sonic anemometers were mounted on 48 meteorological towers, which provided an unprecedented density of instruments in such a limited domain (Figure 1). Observations from the sonic anemometers (a mix of Campbell Scientific CSAT3, METEK uSonic, Gill WindMaster, and YOUNG Model 81000 instruments) ~~are~~ were recorded at a 20-Hz frequency. The height of the towers ranged from 2 m to 100 m, with the sonic anemometers mounted at various ~~heights (~~levels on each tower, as detailed in Table 1 and summarized in the histogram in Figure 2~~)~~, allowing for an extensive survey of the variability

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

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

of the wind flow in the boundary layer. ~~Histogram of the heights of the 184 sonic anemometers considered in this analysis. The bin for the lowest height includes sonic anemometers at 2 m above ground level (AGL).~~ Data from the sonic anemometers have been tilt-corrected following the planar fit method (Wilczak et al., 2001), and rotated into a geographic coordinate system.

To classify atmospheric stability, we calculate the Obukhov length L from each sonic anemometer as

$$L = -\frac{\overline{\theta_v} \cdot u_*^3}{k \cdot g \cdot \overline{w'\theta'_v}}. \quad (1)$$

θ_v is the virtual potential temperature (K, here approximated as the sonic temperature); u_* is the friction velocity ($\mathrm{m\,s^{-1}}$); $k = 0.4$ is the von Kármán constant; $g = 9.81\,\mathrm{m\,s^{-2}}$ is the gravity acceleration; and $\overline{w'\theta'_v}$ is the kinematic buoyancy flux ($\mathrm{m\,K\,s^{-1}}$). 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.

2.2 TKE dissipation rate from sonic anemometers

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):

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

where ~~$D_U(\tau)$ is the second-order structure function of the horizontal velocity, U, calculatedover temporal increments,~~ τ indicates the time lags over which the structure function is calculated, and $a = 0.52$ is the one-dimensional Kolmogorov

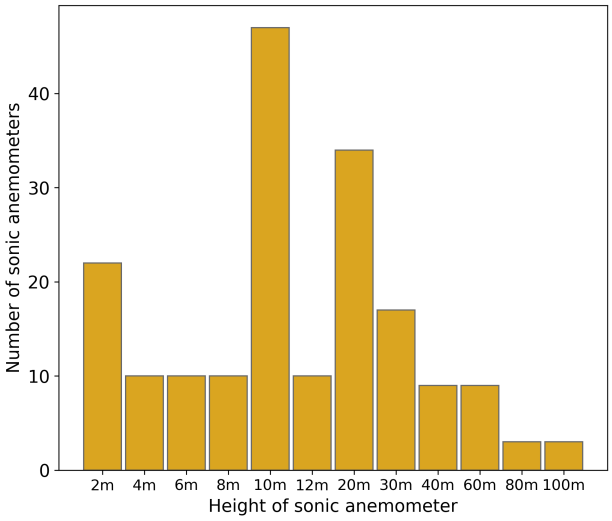


Figure 2. Histogram of the heights AGL of the 184 sonic anemometers considered in this analysis.

constant (Paquin and Pond, 1971; Sreenivasan, 1995). We calculate ϵ ~~is calculated~~ every 30 s, and ~~the fit~~ 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) ~~is done using a temporal separation~~ using time lags between $\tau_1 = 0.1$ s and $\tau_2 = 2$ s~~(Bodini et al., 2018). To~~, which represent a conservative choice to approximate the inertial subrange (Bodini et al., 2018).

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. As additional quality controls, to exclude tower wake effects, data have been discarded when the recorded wind direction was within $\pm$ 30° of the direction of the tower boom. Data during precipitation periods (as recorded by a precipitation sensor on the tower 'riSW06' on the southwest ridge) have also been discarded from further analysis.

2.3 ~~Input features for machine-learning algorithms~~

~~In order to train machine-learning algorithms to predict (the logarithm of)~~ After all the quality controls have been applied, a total (from all sonic anemometers) of over 284,000 30-minute average ϵ ~~, a set of input features needs to be provided. By taking advantage of the main findings of the observational studies on the variability of ϵ, we decided to use both atmospheric- and~~

terrain-related variables to capture the impact of topography on atmospheric turbulence. Therefore, we use the following input features (calculated at the same location and height as ϵ) for all of the considered algorithms:

- wind speed;
- TKE, calculated as

  $$TKE = \frac{1}{2}\left(\sigma_u^2 + \sigma_v^2 + \sigma_w^2\right)$$

  where the variances of the wind components are calculated over 2-minute intervals;
- $TKE^{3/2}$, which is the main variable currently used to represent ϵ in parameterizations in numerical weather prediction models (Nakanish, 2001);
- friction velocity, calculated as

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

  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.
- height above the ground z_{son};
- Obukhov length, L, calculated from each sonic anemometer to classify atmospheric stability. We calculate L as

  $$L = -\frac{\overline{\theta_v} \cdot u_*^3}{k \cdot g \cdot \overline{w'\theta_v'}}.$$

  θ_v is the virtual potential temperature (K); $u_* = (\overline{u'w'}^2 + \overline{v'w'}^2)^{1/4}$ is the friction velocity ($\mathrm{m\,s^{-1}}$); $k = 0.4$ is the von Kármán constant; $g = 9.81\,\mathrm{m\,s^{-2}}$ is the gravity acceleration; and $\overline{w'\theta_v'}$ is the kinematic sensible heat flux ($\mathrm{m\,K\,s^{-1}}$). For atmospheric stability, we classify neutral conditions as $L \leq$ -500 m and $L >$ 500 m; unstable conditions as -500 m $< L \leq$ 0 m; and stable conditions as 0 m $< L \leq$ 500 m (Muñoz-Esparza et al., 2012);
- the ratio z_{son}/L;
- 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 wind direction (an example is shown in Figure 7). In this way,we can capture the influence of upwind topography to trigger turbulence.

Example of an upwind terrain elevation sector with a 1-km radius centered on the location of one of the meteorological towers at Perdigão. The distribution of the input features and of $\log(\epsilon)$ are shown in the Supplement.

~~While we acknowledge that the input features are not fully uncorrelated, we found that including all these features provides a better predictive power for the learning algorithms, despite negatively affecting the computational requirements of the training phase. The application of principal component analysiscan help reduce the number of dimensions in the input features while preserving the predictive power of each, but it is beyond the scope of the current work~~data remains for the analysis.

3 Accuracy of current parameterization of TKE dissipation rate in mesoscale models

Before testing the performance of machine-learning algorithms in predicting TKE dissipation rates, we first assess the current accuracy of the parameterization of ϵ in numerical models. In the Weather Research and Forecasting model (WRF~~)~~ ~~(Skamarock et al., 2005)~~, Skamarock et al. (2005)), the most widely used numerical weather prediction model, turbulence in the boundary layer can be represented with several planetary-boundary-layer (PBL) schemes, most of which implicitly assume a local balance between turbulence production and dissipation. Among the different PBL schemes, the MYNN scheme is one of the most commonly chosen. Turbulence dissipation rate in MYNN is given (Nakanish, 2001) as a function of TKE as ~~(Nakanish, 2001)~~

$$\epsilon = \frac{(2\ \mathrm{TKE})^{3/2}}{B_1\ L_M} \tag{4}$$

where $B_1 = 24$, and the master length scale, L_M, is defined with a diagnostic equation, based on large-eddy simulations, as a function of three other length scales

$$\frac{1}{L_M} = \frac{1}{L_S} + \frac{1}{L_T} + \frac{1}{L_B}. \tag{5}$$

L_S is the length scale in the surface layer, given by

$$L_S = \begin{cases} \kappa\ z/3.7 & \zeta \geq 1 \\ \kappa\ z\ (1 + 2.7\ \zeta)^{-1} & 0 \leq \zeta < 1 \\ \kappa\ z\ (1 - \alpha_4\ \zeta)^{0.2} & \zeta < 0 \end{cases} \tag{6}$$

where $\kappa = 0.4$ is the von Kármán constant, $\zeta = z/L$ (with L the Obukhov length), $\alpha_4 = 100.0$.

L_T is the length scale depending upon the turbulent structure of the PBL (Mellor and Yamada, 1974), defined as

$$L_T = \alpha_1 \frac{\int_0^\infty q\ z\ dz}{\int_0^\infty q\ dz} \tag{7}$$

where $q = \sqrt{2\ TKE}$, and $\alpha_1 = 0.23$.

L_B is a length scale limited by the buoyancy effect, given by

$$L_B = \begin{cases} \alpha_2\ q/N & \partial\Theta/\partial z > 0 \text{ and } \zeta \geq 0 \\ \frac{\alpha_2\ q + \alpha_3\ q\ (q_c/L_T\ N)^{1/2}}{N} & \partial\Theta/\partial z > 0 \text{ and } \zeta < 0 \\ \infty & \partial\Theta/\partial z \leq 0 \end{cases} \tag{8}$$

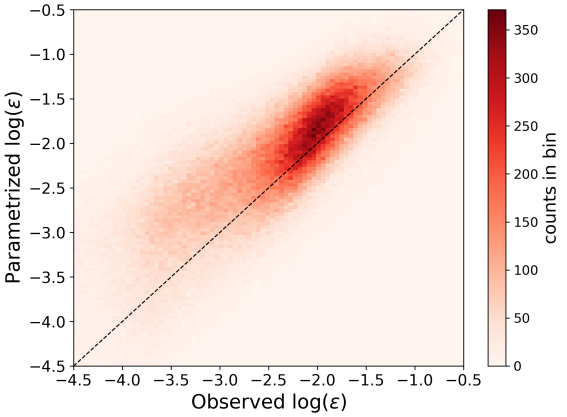


Figure 3. Density histogram showing the comparison between observed and MYNN-parameterized ϵ from the 184 sonic anemometers at Perdigão.

with N the Brunt-Väisälä frequency, Θ the mean potential temperature, $\alpha_2 = 1.0$, ~~and~~ $\alpha_3 = 5.0$, and $q_c = \left[(g/\Theta_0 \overline{w'\theta'} L_T)\right]^{1/3}$.

From the available observations from the meteorological towers at Perdigão, only L_S can be determined, while the calculation of L_T and L_B would only be possible with critical assumptions about the vertical profile of TKE. Therefore, we decide to approximate L_M as

$$\frac{1}{L_M} \approx \frac{1}{L_S}. \tag{9}$$

By doing so, L_M is overestimated (proof shown in the Supplement), which in turn implies that ϵ ~~as calculated from~~ calculated using Eq. (4) will be underestimated.

To evaluate the accuracy of the MYNN parameterization of ϵ, we calculated, ~~every 30 s~~using 30-minute average data, the parameterized ϵ using Eq. (4) (with the approximation in Eq. (9)) from all of the 184 sonic anemometers considered in the study, and compared with the observed values of TKE dissipation rate (Figure 3) derived from the sonic anemometers with Eq. (2). Given the extremely large range of variability of ϵ, we calculate all the error metrics using the logarithm of predicted and observed ϵ. The TKE dissipation rate predicted by the MYNN parameterization shows, on average, a large positive bias compared to the observed values, with a mean bias ~~(in logarithmic space) of~~ of ~~+28% . This~~ 12% in terms of the logarithm of ϵ, +47% in terms of ϵ. The root-mean-square error (RMSE) is 0.61, and the mean absolute error (MAE) is 0.46. The observed bias would be even larger if L_M was calculated including all the contributions according to Eq. (5), and not L_s only as in our approximation. ~~The root-mean-square error (RMSE) is 0.76, and the mean absolute error (MAE) is 0.57~~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.

Different atmospheric stability conditions give different biases. Figure 4 compares observed and parameterized ϵ values for stable and unstable conditions, classified based on ~~the Obukhov length, L,~~ $\zeta = z/L$, measured at each sonic anemometer,

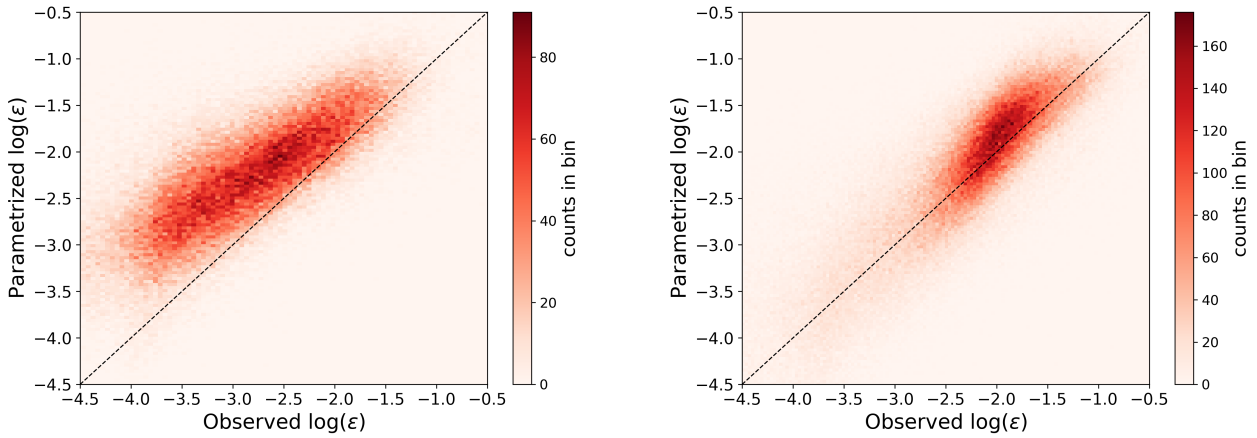


Figure 4. Density histogram showing the comparison between observed and MYNN-parameterized ϵ from the 184 sonic anemometers at Perdigão for stable conditions (left) and unstable conditions (right), as quantified by ~~the Obukhov length~~ $\zeta = z/L$ calculated at each sonic anemometer.

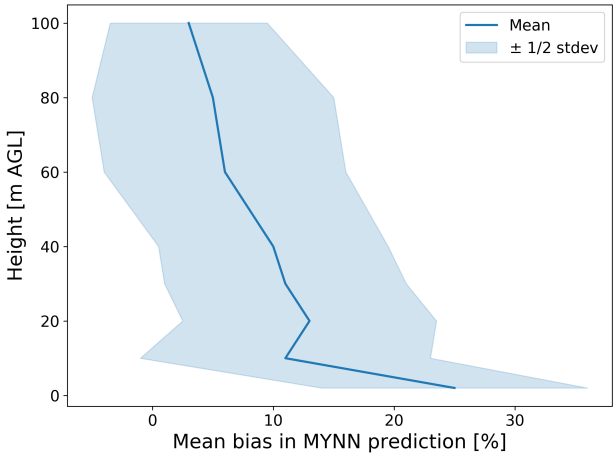


Figure 5. ~~Mean bias~~ Bias in the MYNN-parameterized $\log(\epsilon)$ at different heights, as calculated from the 184 sonic anemometers at Perdigão.

according to the thresholds described in Section 2.1. Stable cases show the largest bias (mean of +~~63~~24% in terms of the ~~logarithms~~ logarithm of ϵ, +101% in terms of ϵ), whereas for unstable conditions the bias is smaller (mean of +~~7~~6% in terms of the ~~logarithms~~ logarithm of ϵ, +19% in terms of ϵ). The MYNN parameterization of ϵ is therefore especially inadequate to represent small values of ϵ, which mainly occur in stable conditions.

Different heights also impact the accuracy of the parameterization of ϵ. As shown in Figure 5, the mean bias in parameterized $\log(\epsilon)$ decreases with height, while its spread (quantified in terms of the standard deviation of the bias at each height) does not

show a large variability at different levels. Close to the surface (data from the sonic anemometers at 2 m AGL), a mean bias (in logarithmic space) of ~~almost~~ about +~~30~~25% is found, whereas for the sonic anemometers at 100 m AGL, we find a mean bias of just $\sim$ +~~20~~3%. This difference in bias with height becomes much larger if the bias is calculated on the actual ϵ values (and not on their logarithm). 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.

4 Machine-learning algorithms

To test the power of machine learning to improve the numerical representation of the TKE dissipation rate, we consider three learning algorithms in this study: multivariate linear regression, multivariate third-order polynomial regression, and random forest. ~~These models~~ 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. The learning algorithms are trained and tested to predict the logarithm of ϵ using 30-minute average data. For all but the random forest algorithm, the data were scaled and normalized. ~~Time stamps with~~ No data imputation was performed, and missing data were removed from the analysis.

For the purpose of machine-learning algorithms, the data set has to be divided into three subsets: training, validation, and testing sets (Friedman et al., 2001). The algorithms are first trained multiple times with different hyperparameters (model parameters whose values are set before the training phase and that control the learning process) on the training set, then the validation set is used to choose the best set of hyperparameters, and finally the predicting performance of the trained algorithm is assessed on the testing set. Usually, the data set is split randomly into training, validation, and testing sets. However, as the data used in this study consist of observations ~~taken~~ averaged every 30 ~~s~~minutes, data in contiguous time stamps are likely characterized by ~~a large~~ some auto-correlation. Therefore, the traditional random split between training and testing data would lead to an artificially enhanced performance of the machine-learning algorithms, which would be tested on data with a large auto-correlation with the ones used for the training. Therefore, here we use one concurrent week of the data for testing ($\sim$17% of the data), whereas the other 5 weeks are split between training (4 weeks, 66% of the data) and validation (1 week, 17% of the data). The 1-week testing period is shifted continuously throughout the considered 6 weeks of observations at Perdigão, so that each model is trained and its prediction performance tested six times. For each algorithm, we evaluate the overall performance based on the RMSE between the actual and predicted (logarithm of) ϵ, averaged over the different week-long testing periods.

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. To test different combinations of hyperparameters and determine the best set, we use cross validation with randomized search, with ~~five parameter settings~~ 20 parameter sets sampled for each learning algorithm. For each set of

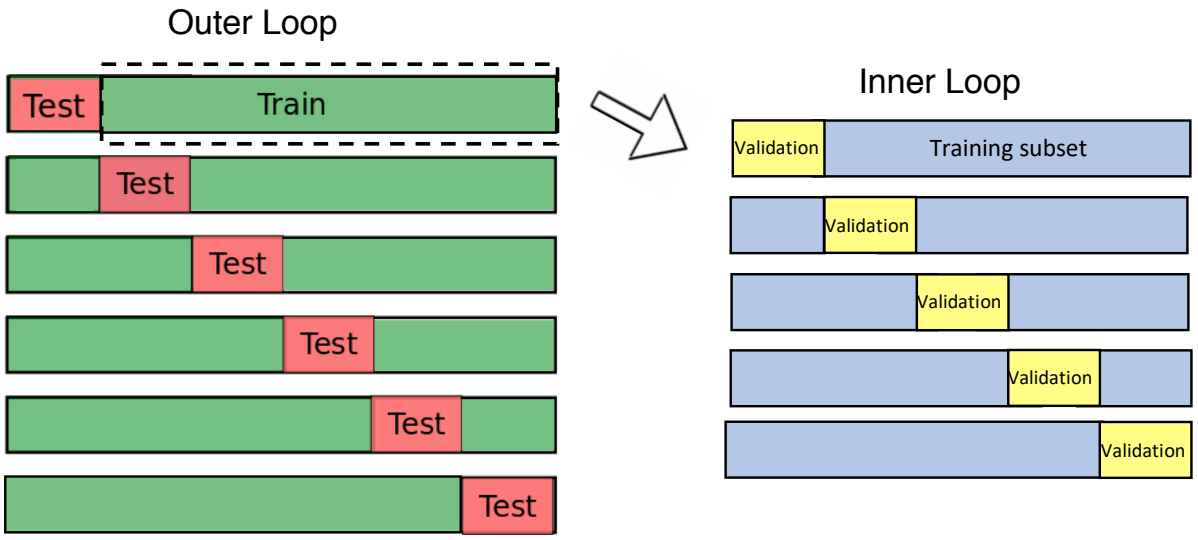


Figure 6. Cross-validation approach used to evaluate the performance of the machine-learning models considered in this study.

hyperparameters, the RMSE between the actual and predicted $\log(\epsilon)$ in the validation test is calculated. For each model, we select the hyperparameter combination (among the ones surveyed in the cross validation) that leads to the lowest mean (across the five validation sets) RMSE. We then use this set as the final combination for assessing the performance of the models on the testing set. Overall, the procedure is repeated six times, by shifting the 1-week testing set (Figure 6).

In the following paragraphs, we describe the main characteristics of the three machine-learning algorithms used in our study. A more detailed description can be found in machine-learning textbooks ~~(Géron, 2017)~~(Hastie et al., 2009; Géron, 2017).

4.1 Multivariate linear regression

To check whether simple learning algorithms can improve the current numerical parameterization of ϵ, we test the accuracy of multivariate linear regression

$$\log(\hat{\epsilon}) = \theta_0 + \theta_1\, x_1 + \theta_2\, x_2 + ... + \theta_n\, x_n \tag{10}$$

where $\hat{\epsilon}$ is the machine-learning predicted value of ϵ, n is the number of features used to predict ϵ (here ~~8~~6 - see Section 4.4), x_i is the i^{th} feature value, and θ_j is the j^{th} model ~~parameter~~weight.

To avoid training a model that overfits the data, regularization techniques need to be implemented, so that the learning model is constrained: the fewer degrees of freedom the model has, the harder it will be for it to overfit the data. We use Ridge regression (Hoerl and Kennard, 1970) (`Ridge` in python's library Scikit-learn) to constrain the multivariate regression. Ridge regression constrains the weights of the model θ_j to have them stay as small as possible. The Ridge regression is achieved by adding a

regularization term to the cost function (MSE)

$$J(\theta) = MSE(\theta) + \alpha \sum_{i=1}^{n} \theta_i^2 \tag{11}$$

where the hyperparameter ~~,~~α ~~,~~controls how much the model will be regularized. The optimal value of the hyperparameter ~~,~~α ~~,~~is determined by cross validation, as described earlier, with values sampled in the range from 0.1–10.

4.2 Multivariate third-order polynomial regression

Multivariate polynomial regression can easily be achieved by adding powers of each input feature as new features. The regression algorithm is then trained as a linear model on this extended set of features. For a third-order polynomial regression, the model becomes

$$\begin{aligned}\log(\hat{\epsilon}) = \theta_0 &+ \sum_{i=1}^{n} \theta_i \, x_i + \sum_{i=1}^{n} \theta_{ii} \, x_i^2 + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \theta_{ij} \, x_i \, x_j \\ &+ \sum_{i=1}^{n} \theta_{iii} \, x_i^3 + \sum_{i=1}^{n} \sum_{j \neq i} \theta_{iij} \, x_i^2 \, x_j \\ &+ \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^{n-1} \theta_{ijk} \, x_i \, x_j \, x_k \end{aligned} \tag{12}$$

Ridge regression (`Ridge` in Scikit-learn) is used again to constrain the multivariate polynomial regression, with the hyperparameter α in Eq. (11) determined via cross validation, with values sampled in the range from 1–2000.

4.3 Random forest

Random forests (`RandomForestRegressor` in Scikit-learn) combine multiple decision trees to provide an ensemble prediction.

A decision tree can learn patterns and then predict values by recursively splitting the training data based on thresholds of the different input features.

As a result, the data are divided into groups, each associated with a single predicted value of ϵ, calculated as the average target value (of the observed ϵ) of the instances in that group.

As an ensemble of decision trees, a random forest trains them on different random subsets of the training set. Once all the predictors are trained, the ensemble (i.e., the random forest) can make a prediction for a new instance by taking the average of all the predictions from the single trees. In addition, random forests introduce some extra randomness when growing trees: instead of looking for the feature that, when split, reduces the overall error the most when splitting a node, a random forest searches for the best feature among a random subset of features.

Decision trees make very few assumptions about the training data. As such, if unconstrained, they will adapt their structure to the training data, fitting them closely, and most likely overfitting them, without then being able to provide accurate predictions on new data. To avoid overfitting, regularization can be achieved by setting various hyperparameters that insert limits to the

Table 2. Hyperparameters considered for the random forest algorithm.

Hyperparameter	Meaning	Sampled values
Number of estimators	Number of trees in the forest	10 - 250
Maximum depth	Maximum depth of the tree	1 - 50
Maximum number of leaf nodes	Maximum number of leaf nodes in the decision tree	2000 - 500,000
Maximum number of features	Number of features to consider when looking for the best split	1 - ~~8~~ 6
Minimum number of samples to split	Minimum number of samples required to split an internal node	1 - 200
Minimum number of samples for a leaf	Minimum number of samples required to be at a leaf node	1 - 50

structure of the trees used to create the random forests. Table 2 ~~shows the hyperparameters considered here and the values we~~ 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 ~~(with five sets of parameters sampled)~~to form the twenty sets of hyperparameters considered in the training phase.

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 the three learning algorithms considered in our study:

- 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

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

  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

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

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:

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

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.

While we acknowledge that the input features are not fully uncorrelated, we found that including all these features provides a better predictive power for the learning algorithms, despite negatively affecting the computational requirements of the training phase. The application of principal component analysis can help reduce the number of dimensions in the input features while preserving the predictive power of each, but it is beyond the scope of the current work.

5 Results

5.1 Performance of machine-learning algorithms

To evaluate the prediction performance of the three machine-learning algorithms we considered, we use, for each method, a density histogram showing the comparison between observed and machine-learning-predicted ϵ ~~. As~~ (Figure 8). The prediction

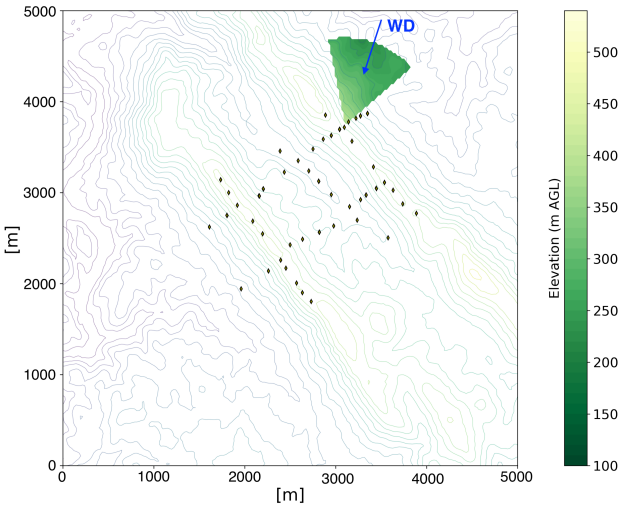


Figure 7. Example of an upwind terrain elevation sector with a 1-km radius centered on the location of one of the meteorological towers at Perdigão.

from all the considered learning algorithms do not show a significant mean bias, as found in the MYNN representation of ϵ. As specific error metrics, we compare RMSE ~~, MAE, and coefficient of determination (R^2) of~~ and MAE of the machine-learning predictions with what we obtained from the MYNN parameterization, with the caveat that while the MYNN scheme is thought to provide a universal representation of ϵ, the machine-learning models have been specifically trained on data from a single field campaign. Each machine-learning algorithm was tested on six 1-week-long testing periods, as described in Section 4. For each method we present the RMSE ~~, MAE, and R^2~~ and MAE averaged across the different testing periods. ~~Because the length scale approximation we made in calculating MYNN-predicted ϵ lead to an optimistic result, the RMSE and MAE for the MYNN case would in reality be higher, and the R^2 lower, and so the error reductions achieved with the machine-learning algorithms would even be greater than the numbers we report here.~~

~~Predictions from the~~ Even the simple multivariate linear regression (Figure **??**) ~~only show a slight improvement constrained to values of ϵ greater than 10^{-3} m^2 s^{-3}, whereas a severe overestimation of ϵ is found for low turbulence regimes, when the linear regression predicts almost constant values of ϵ. We speculate that this limited performance is mainly because of the fact that the linear regression can only accurately model ϵ over the same range of orders of magnitude of TKE (which is the dominant input feature, as will be systematically proven with the random forest algorithm), while the dissipation rate varies over more orders of magnitude than TKE (see histograms in the Supplement). Despite the poor agreement with the observations, the multivariate linear regression~~ 8-a) improves, on average, on MYNN. Overall, the average RMSE (~~0.66) is 13~~0.47) is 23% smaller than the MYNN parameterization, and the average MAE (~~0.51) is 12~~0.36) is 22% lower than the MYNN prediction. ~~Density histogram showing the comparison between observed and machine-learning-predicted ϵ from a multivariate linear regression on the test set.~~

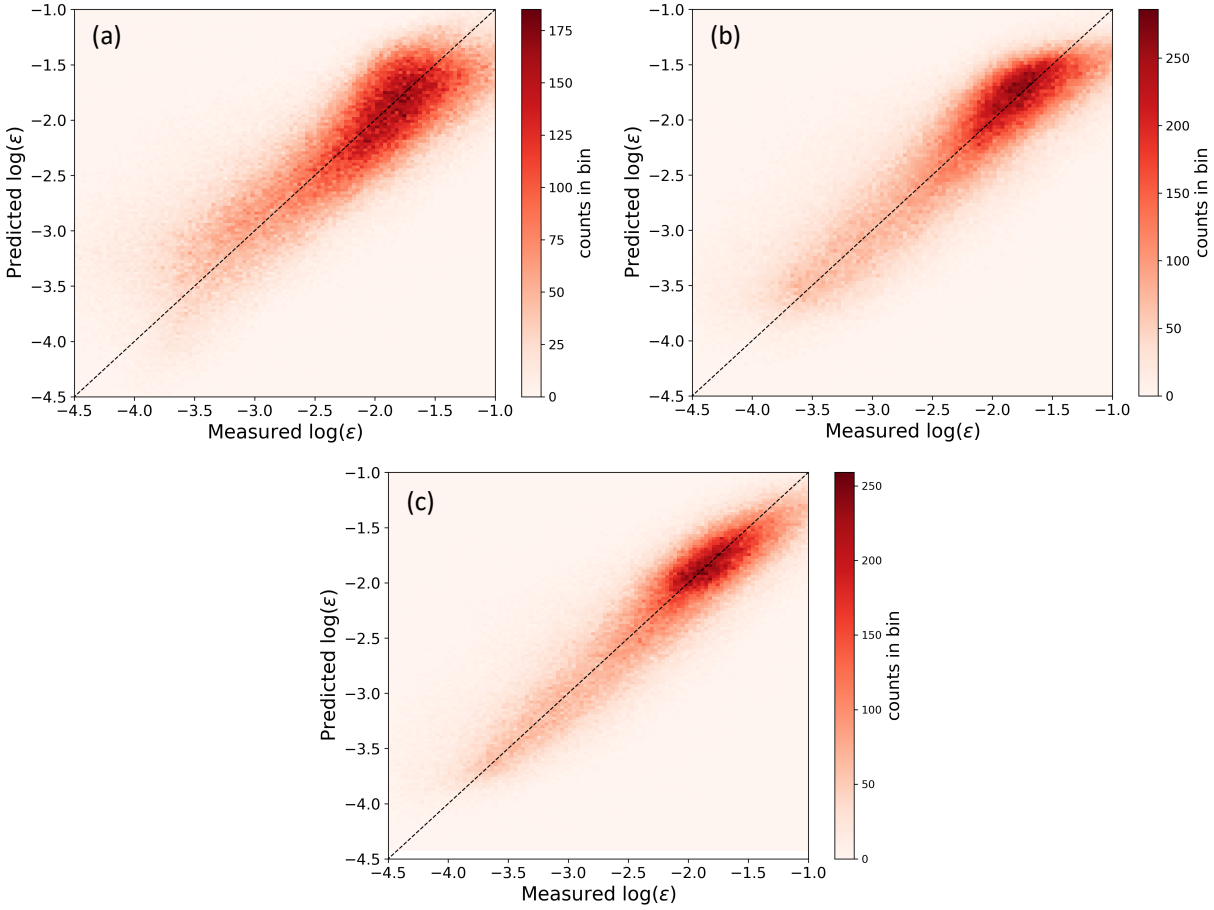


Figure 8. Density histogram showing the comparison, performed on the testing set, between observed and machine-learning-predicted ϵ from a multivariate linear regression (a), a multivariate third-order polynomial regression (b), and a random forest (c).

~~On the other hand, the~~ The multivariate third-order polynomial regression provides ~~a significant~~ an additional improvement (Figure ~~??~~8-b) for the representation of ϵ, ~~by eliminating the average bias found in the MNYNN parameterization. Density histogram showing the comparison between observed and machine-learning-predicted ϵ from a multivariate third-order polynomial regression on the test set.~~ with the average RMSE (0.44) over 28% smaller than the MYNN parameterization, and the average MAE (0.33) 28% lower than the MYNN representation. The additional input features created by the polynomial model allow for an accurate prediction of ϵ even at the low turbulence regime~~, as they allow the model to cover the full range of variability of~~ . Finally, the random forest further reduces the spread in machine-learning predicted ϵ~~. We find the average RMSE (0.59) to be over 23% smaller than the MYNN parameterization, and the average MAE (0.45) is 22% lower than the MYNN representation.~~

~~Results from the random forest (Figure ??) provides accurate predictions at both high and low turbulence regimes. The RMSE (0.53) is~~ , with the RMSE (0.40) reduced by about ~~30~~35% from the MYNN case, and the MAE (~~0.41) by 29~~0.29) by 37%, with no average bias between observed and predicted values of ϵ. ~~Density histogram showing the comparison between observed and machine-learning-predicted ϵ from a random forest on the test set.~~

Table 3 summarizes the performance of all ~~of~~ the considered algorithms~~, and includes results also in terms of the coefficient of determination, R^2.~~

.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.

Table 3. Performance of the machine-learning algorithms trained and tested at Perdigão, measured in terms of RMSE ~~, MAE,~~ and ~~R^2~~ MAE between the logarithm of observed and MYYN-parameterized ϵ.

	MYNN parameterization	Linear regression	Third-order polynomial regression	Random forest
RMSE	~~0.76~~ 0.61	~~0.66~~ 0.47	~~0.59~~ 0.44	~~0.53~~0.40
% change in RMSE		~~-13~~ -23%	~~-23~~ -28%	~~-30~~ -35%
MAE	~~0.58~~ 0.46	~~0.51~~ 0.36	~~0.45~~ 0.33	~~0.41~~0.29
% change in MAE		~~-12%~~ -22%	~~-29% R^2~~ -28%	~~0.40 0.53 0.62 0.68% change in R^2 +33% +55% +70%~~ -37%

5.2 ~~Feature importance~~

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%.

~~Random~~

5.2 Physical interpretation of machine-learning results

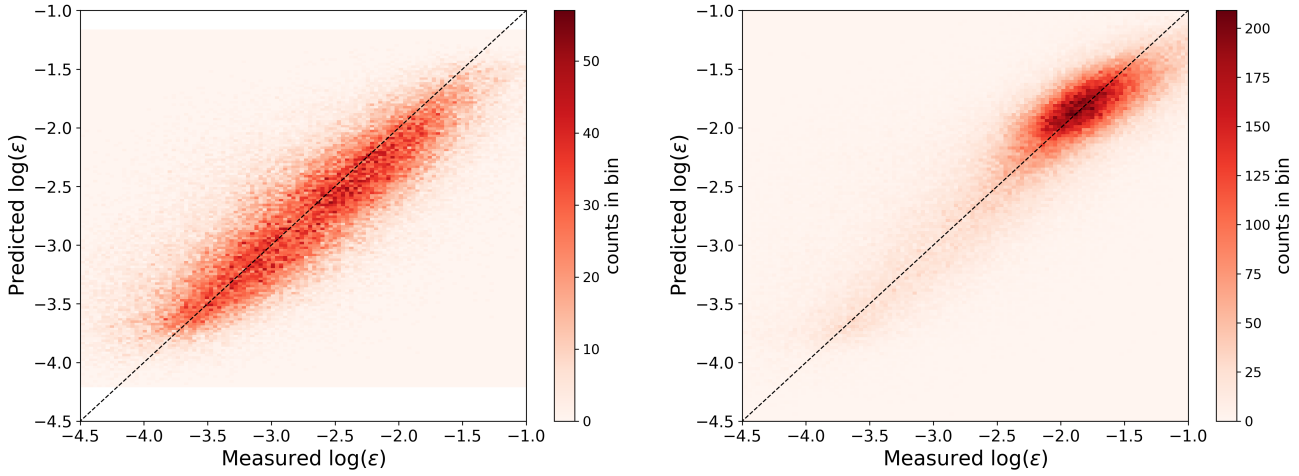


Figure 9. Density histogram showing the comparison, performed on the testing set, between observed and machine-learning-predicted ϵ from a random forest for stable conditions (left) and unstable conditions (right).

Table 4. Feature importance classification as derived from the random forest.

Input feature	Feature importance
~~TKE~~ log(TKE)	~~33~~47%
~~TKE$^{3/2}$~~ log(u_*)	~~29~~24%
~~z_{son}~~ sign(ζ) log($\lvert\zeta\rvert + 1$)	~~11~~13%
~~Wind Speed~~ WS	~~8~~11%
~~u_{star} 8%~~ std(z_{terr})	~~6% z_{son} / L~~ 3%
~~L~~ $\overline{h_{veg}}$	2%

Not only machine learning techniques provide accuracy improvements to represent atmospheric turbulence, but additional insights on the physical interpretation of the results can - and should - be achieved. In particular, random forests allow for an assessment of the relative importance of the input features used to predict (the logarithm of) ϵ. The importance of a feature is calculated by looking at how much the tree nodes that use that feature reduce the MSE on average (across all trees in the forest), weighted by the number of times the feature is selected. Table 4 shows the feature importance for the ~~eight~~ six input features we used in this study.

The feature importance results are affected by the correlation between some of the input features used in the models. We find how the logarithm of turbulence kinetic energy is the preferred feature for tree splitting, with the largest importance (~~62% in total, considering TKE and TKE$^{3/2}$~~47%) in reducing the prediction error for ϵ in the random forest. This result, which can be expected as both TKE and ϵ are variables connected to turbulence in the boundary layer, agrees well with the current

formulation of the MYNN parameterization of ϵ, which includes TKE as main term. As TKE is correlated to ~~u_{star}~~ u_* and L, we find that the decision trees more often split the data based on TKE, so that the feature importance of its correlated variables is found to be ~~much~~ lower. The limitations of the Monin-Obukhov similarity theory (Monin and Obukhov, 1954) in complex terrain might also be an additional cause for the relatively low feature importance of the ~~features~~ feature associated with L. The ~~height of the measurement is the next most important feature (11%), followed by wind speed (at 8%). The~~ standard deviation of the upwind elevation ~~has an importanceof 6%~~ and the mean vegetation height have the lowest importance, of respectively 3% and 2%. Though not negligible, the importance of topography and canopy might increase by considering different parameters that could better ~~capture the topographic effect.~~ encapsulate their effect. 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).

5.3 ~~Partial dependence plots~~

~~To~~ Finally, to assess the dependence of TKE dissipation rate on the individual features considered in this study, Figure 10 shows partial dependence plots for the ~~features with the largest importance, namely wind speed, friction velocity, TKE, and z (plots for the rest of the features are included in the Supplement)~~input features considered in the analysis. These are obtained, for each input feature, by applying the machine-learning algorithm (here, random forests~~using about 120,000 instances to speed up the computation~~) multiple times with the other feature variables constant (at their means) while varying the target input feature and measuring the effect on the response variable (here, $\log(\epsilon)$).

In each plot, the values on the y-axes have not been normalized, so that large ranges show a strong dependence of $\log(\epsilon)$ on the feature, whereas small ranges indicate weaker dependence. ~~A strong relationship appears between TKE dissipation and low wind speed, with a steep increase in~~ The strong relationship between ϵ and TKE is confirmed, as the range shown on its y-axis is the largest among all features. As TKE increases, so does ϵ~~for wind speeds up to $\sim$2 m s^{-1}. Stronger wind speeds do not appear to have a systematic impact on~~ . A similar trend, tough with a weaker influence, emerges when considering the dependence of ϵ on friction velocity. The relationship between ~~TKE and its dissipation shows a similar trend, with~~ ϵ and wind speed shows a less clear trend, and with a weaker dependence: ϵ ~~systematically increasing for TKE values~~ increases for 30-minute average wind speeds up to $\sim$2 m ~~^{2} s^{-2}, and staying constant for stronger TKE values. TKE dissipation also increases with friction velocity, u_*, with a less steep trend. Finally, height above the ground shows a systematic impact on~~ s^{-1}, and then decreases for stronger wind speed values. A more distinct trend could emerge when considering data averaged at shorter time periods. The dependence between TKE dissipation and atmospheric stability shows a moderate impact, with stable conditions (positive values of the considered metric) showing smaller ϵ values compared to unstable cases (negative values of the considered metric). Interestingly, the largest ϵ values seems to be connected to neutral cases. Finally, both terrain elevation and vegetation height show weak impact on determining the values of ϵ, ~~with a strong downward trend~~as quantified by the narrow range of values sampled on the y-axis for these two variables.

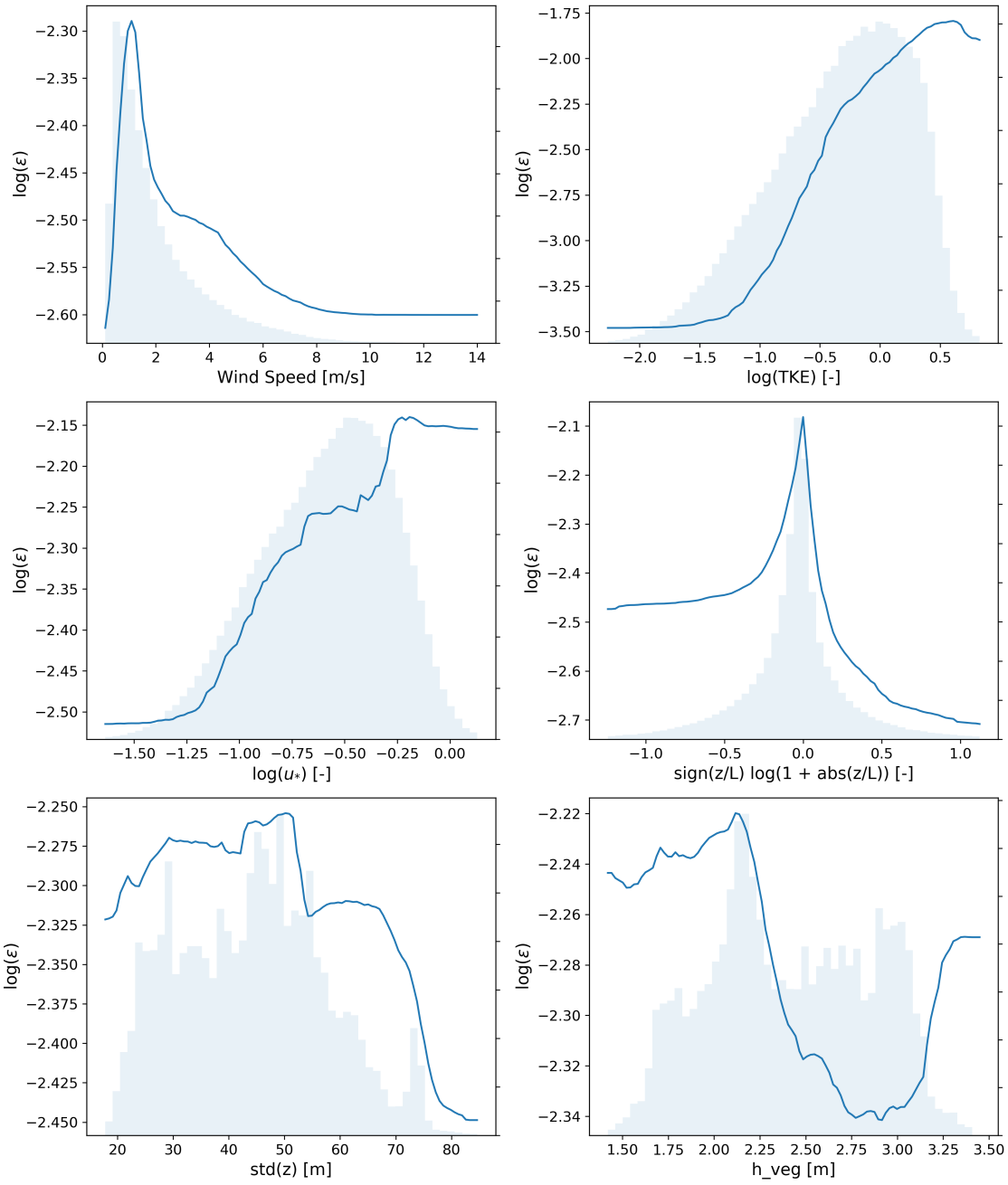


Figure 10. Partial dependence plots for ~~wind speed; friction velocity, u_*; TKE; and measurement height, z~~the input features used in the analysis. Distributions of the considered features are shown in the background.

6 Conclusions

Despite turbulence being a fundamental quantity for the development of multiple phenomena in the atmospheric boundary layer, the current representations of TKE dissipation rate (ϵ) in numerical weather prediction models suffer from large inaccuracies. In this study, we quantified the error introduced in the MYNN parameterization of ϵ by comparing predicted and observed values of ϵ from 184 sonic anemometers from 6 weeks of observations at the Perdigão field campaign. A large positive bias (average +~~28~~12% in logarithmic space, +47% in natural space) emerges, with larger errors found in atmospheric stable conditions. The need for a more accurate representation of ϵ is therefore clearly demonstrated.

The results of this study show how machine learning can provide new ways to successfully represent TKE dissipation rate from a set of atmospheric and topographic parameters. ~~While~~ Even ultrasimple models such as ~~linear regression cannot provide a satisfactory~~ a multivariate linear regression can provide an improved representation of ϵ ~~, more~~ compared to the current MYNN parameterization. More sophisticated algorithms, such as a random forest approach, lead to ~~over a 30~~the largest benefits, with over a 35% reduction in the average error introduced in the parameterization of ϵ, and eliminate the large bias found in it, for the Perdigão field campaign. When considering stable conditions only, the reduction in average error reaches 50%. Although the generalization gap between the universal nature of the MYNN parameterization of ϵ and the campaign-specific training and testing of the machine-learning models ~~has~~ needs to be acknowledged, the results of this study can be considered as a proof of concept of the potentialities of machine-learning-based representations of complex atmospheric processes.

Multiple opportunities exist to extend the work presented here. In the future, additional learning algorithms, such as support vector machines and extremely randomized trees, should be considered. Deep learning methods, such as recurrent neural networks, and specifically long-short term memory, which are well-suited for time-series-based problems, could also be considered to obtain a more complete overview of the capabilities of machine-learning techniques for improving numerical representations of ϵ. Moreover, additional input features could be added to the learning algorithms to possibly identify additional variables with a large impact on atmospheric turbulence. ~~To this extent, we expect vegetation height to be an important parameter to consider. The results could also be analyzed with a focus on the temporal variability of the performance of the proposed algorithms; for example, at different times of the day, or for different atmospheric stability conditions.~~ Finally, the learning algorithms developed here would need to be tested using data from different field experiments, to understand whether the results obtained in this study can be generalized everywhere. Once the performance of a machine-learning representation of ϵ has been accurately tested, its implementation in numerical weather prediction models, such as the Weather Research and Forecasting model, should be achieved.

Code and data availability. High-resolution data from sonic anemometers on the meteorological towers (UCAR/NCAR, 2019) are available through the EOL project at https://doi.org/10.26023/8X1N-TCT4-P50X. Digital Elevation Model data are taken from the SRTM 1 Arc-Second Global at https://doi.org/10.5066/F7PR7TFT. The vegetation height data are available at XXXXXXXXXX. The machine learning code used for the analysis is stored at https://doi.org/10.5281/zenodo.3754710.

Author contributions. NB and JKL designed the analysis. NB analyzed the data from the sonic anemometers and applied the machine-learning figures, in close consultation with JKL and MO. NB wrote the paper, with significant contributions from JKL and MO.

Competing interests. The authors declare that they have no conflicts of interest.

Acknowledgements. We thank the residents of Alvaiade and Vale do Cobrão for their essential hospitality and support throughout the field campaign. In particular, we are grateful for the human and logistic support Felicity Townsend provided to our research group in the field. We thank Prof. Jose Laginha Palma for providing the vegetation data used in the analysis. We thank Dr. Ivana Stiperski for her exceptionally thoughtful review of our discussion paper, which greatly improved the quality of this work. Support to NB and JKL is provided by the National Science Foundation, under the CAREER program AGS-1554055 and the award AGS-1565498. This work utilized the RMACC Summit supercomputer, which is supported by the National Science Foundation (awards ACI-1532235 and ACI-1532236), the University of Colorado Boulder, and Colorado State University. The Summit supercomputer is a joint effort of the University of Colorado Boulder and Colorado State University.

References

Albertson, J. D., Parlange, M. B., Kiely, G., and Eichinger, W. E.: The average dissipation rate of turbulent kinetic energy in the neutral and unstable atmospheric surface layer, Journal of Geophysical Research: Atmospheres, 102, 13 423–13 432, 1997.

Arcos Jiménez, A., Gómez Muñoz, C., and García Márquez, F.: Machine learning for wind turbine blades maintenance management, Energies, 11, 13, 2018.

Babić, K., Bencetić Klaić, Z., and Večenaj, Ž.: Determining a turbulence averaging time scale by Fourier analysis for the nocturnal boundary layer, Geofizika, 29, 35–51, 2012.

Barlow, R. J.: Statistics: a guide to the use of statistical methods in the physical sciences, vol. 29, John Wiley & Sons, 1989.

Berg, L. K., Liu, Y., Yang, B., Qian, Y., Olson, J., Pekour, M., Ma, P.-L., and Hou, Z.: Sensitivity of Turbine-Height Wind Speeds to Parameters in the Planetary Boundary-Layer Parametrization Used in the Weather Research and Forecasting Model: Extension to Wintertime Conditions, Boundary-Layer Meteorology, pp. 1–12, 2018.

Bodini, N., Zardi, D., and Lundquist, J. K.: Three-dimensional structure of wind turbine wakes as measured by scanning lidar, Atmospheric Measurement Techniques, 10, 2017.

Bodini, N., Lundquist, J. K., and Newsom, R. K.: Estimation of turbulence dissipation rate and its variability from sonic anemometer and wind Doppler lidar during the XPIA field campaign, Atmospheric Measurement Techniques, 11, 4291–4308, 2018.

Bodini, N., Lundquist, J. K., and Kirincich, A.: US East Coast Lidar Measurements Show Offshore Wind Turbines Will Encounter Very Low Atmospheric Turbulence, Geophysical Research Letters, 2019a.

Bodini, N., Lundquist, J. K., Krishnamurthy, R., Pekour, M., Berg, L. K., and Choukulkar, A.: Spatial and temporal variability of turbulence dissipation rate in complex terrain, Atmospheric Chemistry and Physics, 19, 4367–4382, 2019b.

Cervone, G., Clemente-Harding, L., Alessandrini, S., and Delle Monache, L.: Short-term photovoltaic power forecasting using Artificial Neural Networks and an Analog Ensemble, Renewable energy, 108, 274–286, 2017.

Champagne, F. H., Friehe, C. A., LaRue, J. C., and Wynagaard, J. C.: Flux measurements, flux estimation techniques, and fine-scale turbulence measurements in the unstable surface layer over land, Journal of the Atmospheric Sciences, 34, 515–530, 1977.

Clifton, A., Kilcher, L., Lundquist, J., and Fleming, P.: Using machine learning to predict wind turbine power output, Environmental research letters, 8, 024 009, 2013.

Coen, J. L., Cameron, M., Michalakes, J., Patton, E. G., Riggan, P. J., and Yedinak, K. M.: WRF-Fire: coupled weather–wildland fire modeling with the weather research and forecasting model, Journal of Applied Meteorology and Climatology, 52, 16–38, 2013.

De Franceschi, M. and Zardi, D.: Evaluation of cut-off frequency and correction of filter-induced phase lag and attenuation in eddy covariance analysis of turbulence data, Boundary-Layer Meteorology, 108, 289–303, 2003.

Fernando, H. J., Mann, J., Palma, J. M., Lundquist, J. K., Barthelmie, R. J., Belo Pereira, M., Brown, W. O., Chow, F. K., Gerz, T., Hocut, C. M., et al.: The Perdigão: Peering into Microscale Details of Mountain Winds, Bulletin of the American Meteorological Society, 2018.

Frehlich, R.: Coherent Doppler lidar signal covariance including wind shear and wind turbulence, Applied Optics, 33, 6472–6481, 1994.

Frehlich, R., Meillier, Y., Jensen, M. L., Balsley, B., and Sharman, R.: Measurements of boundary layer profiles in an urban environment, Journal of Applied Meteorology and Climatology, 45, 821–837, 2006.

Friedman, J., Hastie, T., and Tibshirani, R.: The elements of statistical learning, vol. 1, Springer series in statistics New York, 2001.

Frisch, U.: Turbulence: the legacy of A.N. Kolmogorov, Cambridge University Press, 1995.

Garratt, J. R.: The atmospheric boundary layer, Earth-Science Reviews, 37, 89–134, 1994.

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., and Yacalis, G.: Could machine learning break the convection parameterization deadlock?, Geophysical Research Letters, 45, 5742–5751, 2018.

Géron, A.: Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems, " O'Reilly Media, Inc.", 2017.

Gerz, T., Holzäpfel, F., Bryant, W., Köpp, F., Frech, M., Tafferner, A., and Winckelmans, G.: Research towards a wake-vortex advisory system for optimal aircraft spacing, Comptes Rendus Physique, 6, 501–523, 2005.

Giebel, G., Brownsword, R., Kariniotakis, G., Denhard, M., and Draxl, C.: The state-of-the-art in short-term prediction of wind power: A literature overview, ANEMOS. plus, 2011.

Hastie, T., Tibshirani, R., and Friedman, J.: The elements of statistical learning: data mining, inference, and prediction, Springer Science & Business Media, 2009.

Hoerl, A. E. and Kennard, R. W.: Ridge regression: Biased estimation for nonorthogonal problems, Technometrics, 12, 55–67, 1970.

Hong, S.-Y. and Dudhia, J.: Next-generation numerical weather prediction: Bridging parameterization, explicit clouds, and large eddies, Bulletin of the American Meteorological Society, 93, ES6–ES9, 2012.

Huang, K., Fu, J. S., Hsu, N. C., Gao, Y., Dong, X., Tsay, S.-C., and Lam, Y. F.: Impact assessment of biomass burning on air quality in Southeast and East Asia during BASE-ASIA, Atmospheric Environment, 78, 291–302, 2013.

John, J. and Draper, N. R.: An alternative family of transformations, Journal of the Royal Statistical Society: Series C (Applied Statistics), 29, 190–197, 1980.

Kelley, N. D., Jonkman, B., and Scott, G.: Great Plains Turbulence Environment: Its Origins, Impact, and Simulation, Tech. rep., National Renewable Energy Laboratory (NREL), Golden, CO., https://www.nrel.gov/docs/fy07osti/40176.pdf, 2006.

Kolmogorov, A. N.: Dissipation of energy in locally isotropic turbulence, in: Dokl. Akad. Nauk SSSR, vol. 32, pp. 16–18, 1941.

Krishnamurthy, R., Calhoun, R., Billings, B., and Doyle, J.: Wind turbulence estimates in a valley by coherent Doppler lidar, Meteorological Applications, 18, 361–371, 2011.

Leahy, K., Hu, R. L., Konstantakopoulos, I. C., Spanos, C. J., and Agogino, A. M.: Diagnosing wind turbine faults using machine learning techniques applied to operational data, in: 2016 IEEE International Conference on Prognostics and Health Management (ICPHM), pp. 1–8, IEEE, 2016.

Leufen, L. H. and Schädler, G.: Calculating the turbulent fluxes in the atmospheric surface layer with neural networks, Geoscientific Model Development Discussions, 2018, 1–22, https://doi.org/10.5194/gmd-2018-263, https://www.geosci-model-dev-discuss.net/gmd-2018-263/, 2018.

Lundquist, J. K. and Bariteau, L.: Dissipation of Turbulence in the Wake of a Wind Turbine, Boundary-Layer Meteorology, 154, 229–241, https://doi.org/10.1007/s10546-014-9978-3, http://link.springer.com/10.1007/s10546-014-9978-3, 2015.

Mellor, G. L. and Yamada, T.: A hierarchy of turbulence closure models for planetary boundary layers, Journal of the Atmospheric Sciences, 31, 1791–1806, 1974.

Monin, A. S. and Obukhov, A. M.: Basic laws of turbulent mixing in the surface layer of the atmosphere, Contrib. Geophys. Inst. Acad. Sci. USSR, 151, 1954.

Muñoz-Esparza, D., Cañadillas, B., Neumann, T., and van Beeck, J.: Turbulent fluxes, stability and shear in the offshore environment: Mesoscale modelling and field observations at FINO1, Journal of Renewable and Sustainable Energy, 4, 063 136, https://doi.org/10.1063/1.4769201, 2012.

Muñoz-Esparza, D., Sharman, R. D., and Lundquist, J. K.: Turbulence dissipation rate in the atmospheric boundary layer: Observations and WRF mesoscale modeling during the XPIA field campaign, Monthly Weather Review, 146, 351–371, 2018.

Nakanish, M.: Improvement of the Mellor–Yamada turbulence closure model based on large-eddy simulation data, Boundary-layer meteorology, 99, 349–378, 2001.

Nakanishi, M. and Niino, H.: An improved Mellor–Yamada level-3 model: Its numerical stability and application to a regional prediction of advection fog, Boundary-Layer Meteorology, 119, 397–407, 2006.

Oncley, S. P., Friehe, C. A., Larue, J. C., Businger, J. A., Itsweire, E. C., and Chang, S. S.: Surface-layer fluxes, profiles, and turbulence measurements over uniform terrain under near-neutral conditions, Journal of the Atmospheric Sciences, 53, 1029–1044, 1996.

Optis, M. and Perr-Sauer, J.: The importance of atmospheric turbulence and stability in machine-learning models of wind farm power production, Renewable and Sustainable Energy Reviews, 112, 27–41, 2019.

Paquin, J. E. and Pond, S.: The determination of the Kolmogoroff constants for velocity, temperature and humidity fluctuations from second- and third-order structure functions, Journal of Fluid Mechanics, 50, 257–269, 1971.

Piper, M. D.: The effects of a frontal passage on fine-scale nocturnal boundary layer turbulence, 2001.

Sharma, N., Sharma, P., Irwin, D., and Shenoy, P.: Predicting solar generation from weather forecasts using machine learning, in: 2011 IEEE International Conference on Smart Grid Communications, pp. 528–533, IEEE, 2011.

Shaw, W. J. and LeMone, M. A.: Turbulence dissipation rate measured by 915 MHz wind profiling radars compared with in-situ tower and aircraft data, in: 12th Symposium on Meteorological Observations and Instrumentation, https://ams.confex.com/ams/pdfpapers/58647.pdf, 2003.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Wang, W., and Powers, J. G.: A description of the advanced research WRF version 2, Tech. rep., National Center For Atmospheric Research, Boulder, CO, Mesoscale and Microscale Meteorology Div, 2005.

Smalikho, I. N.: On measurement of the dissipation rate of the turbulent energy with a cw Doppler lidar, ATMOSPHERIC AND OCEANIC OPTICS C/C OF OPTIKA ATMOSFERY I OKEANA, 8, 788–793, 1995.

Sreenivasan, K. R.: On the universality of the Kolmogorov constant, Physics of Fluids, 7, 2778–2784, 1995.

Thobois, L. P., Krishnamurthy, R., Loaec, S., Cariou, J. P., Dolfi-Bouteyre, A., and Valla, M.: Wind and EDR measurements with scanning Doppler LIDARs for preparing future weather dependent separation concepts, in: 7th AIAA Atmospheric and Space Environments Conference, p. 3317, 2015.

UCAR/NCAR, E. O. L.: NCAR/EOL Quality Controlled High-rate ISFS surface flux data, geographic coordinate, tilt corrected. Version 1.1, https://doi.org/10.26023/8x1n-tct4-p50x, 2019.

Wilczak, J. M., Oncley, S. P., and Stage, S. A.: Sonic anemometer tilt correction algorithms, Boundary-Layer Meteorology, 99, 127–150, 2001.

Wildmann, N., Bodini, N., Lundquist, J. K., Bariteau, L., and Wagner, J.: Estimation of turbulence dissipation rate from Doppler wind lidars and in-situ instrumentation in the Perdigão 2017 campaign, Atmospheric Measurement Techniques, 2019, 6401–6423, 2019.

Xingjian, S., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-C.: Convolutional LSTM network: A machine learning approach for precipitation nowcasting, in: Advances in Neural Information Processing Systems, pp. 802–810, 2015.

Yang, B., Qian, Y., Berg, L. K., Ma, P.-L., Wharton, S., Bulaevskaya, V., Yan, H., Hou, Z., and Shaw, W. J.: Sensitivity of turbine-height wind speeds to parameters in planetary boundary-layer and surface-layer schemes in the weather research and forecasting model, Boundary-Layer Meteorology, 162, 117–142, 2017.