Review of “Can machine learning improve the model representation of TKE dissipation rate in the boundary layer for complex terrain?” by Nicola Bodini, Julie K. Lundquist, and Mike Optis

## 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 Perdigao 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?
   - 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$^{2/3}$ apart for testing for its nonlinear influence
   - 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?
   - Apart from tilt correction, are data rotated into the mean wind?
   - Given the forested nature of Perdigao, 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?
   - What is the number of data points used when all the quality criteria are satisfied?

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)?

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.
   - 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.
   - 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.
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 $z_{son}$ which is not normalized and therefore subject to representativity issues. 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).
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?

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.
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.

Minor points:

**Ln 35:** "; for example" should be ": for example"

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

**Ln 69:** "are recorded" should be "were recorded"

**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

**Ln 75-77:** it is not necessary to mention that it is a structure function of horizontal velocity twice in this sentence

**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?

**Ln 87:** algorithms haven't been introduced yet

**Ln 102:** this is not sensible heat flux but buoyancy flux, given that the authors mention no Schotanus correction. Also, is $\theta_v$ really virtual temperature or rather sonic temperature?

**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?

**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.

**Ln 140:** so is TKE then calculated at 30s?

**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?

**Ln 165 – 166:** What do you mean by hyperparameters? Are you referring to the ones defined in Table 1? This should be referenced here.

**Ln 194:** Mention that Scikit-learn is a python library.

**Ln 195:** what are the variables chosen by the ridge regression?

**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?

**Table 1**: How were these values chosen?

**Ln 232:** How do you explain this "optimistic result" that using a reduced parametrization is actually beneficial to using the full one?

**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?

**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.