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

The authors thank the reviewer for their thoughtful comments, which helped us improve the quality of our manuscript.

The authors present a machine learning approach to vertical extrapolation of wind speeds compared to standard approaches. The research is quite robust, described in detail and well written. The conclusion that the machine learning approach, a random forest, can be extrapolated to other sites as shown by this round robin evaluation is an important scientific discovery.

There are a few areas that can be further described or clarified to make this an excellent paper.

1.) Page 5 line 108 - it is stated that precipitation periods were excluded from the analysis. Please explain why and what impact this has on the analysis.
We have rephrased the sentence and added a reference to a study on the impact of precipitation on the accuracy of sonic anemometer data. It now reads: "precipitation periods were excluded from the analysis to discard inaccurate measurements (Zhang et al. 2016)."

2.) Page 5 Line 11 - it is stated that a 30-min average is used. Is there a reason why 30-min was chosen and would that averaging period affect the results? No further analysis is needed - just an explanation or including in the results discussion how the averaging period may impact the analysis.
Because of the wrong line number listed, we could not determine whether the reviewer is referring to the 30-minute average period used for the lidar and sonic anemometer data, or for the 30-minute average period used for the Reynolds decomposition to calculate Obukhov length. In any case:
1) For the 30-minute average period used as resolution for the main data used in the analysis, the choice was due to the fact the sonic anemometer data were only publicly available at that time resolution. We have added the following sentence to Section 2.2 "processed data are available as 30-minute averages". We have also added the following comment to the beginning of Section 3: "We acknowledge that the resolution of the data used will have an impact on the magnitude of the error values shown in the analysis (as observations at a higher time resolution would likely cause larger extrapolation errors). However, we do not expect the relative comparison between the different extrapolation techniques and the analysis of the predictor importance to be strongly affected by the resolution of the input features used."
2) For the 30-minute average period used for the Reynolds decomposition, as stated in the paragraph (with appropriate references listed), 30-minute is the most common averaging period used to calculate fluxes for boundary layer processes, as it is considered to be shorter than the period of large-scale fluctuations, but longer than the period of short-term turbulence fluctuations, following considerations related to the spectral gap (Van Der Hoven, 1957).

3.) Page 7 Line 146 - please cite Pedregosa et al 2011 for Scikit-learn. http://www.jmlr.org/papers/v12/pedregosa11a.html

We have added the suggested reference.

4.) Page 7 Lines 155-157 - the explanation of training, testing and cross-validation is not clear. Is the 5-fold cross validation performed on the 80% training data and the 20% testing data is held out for independent validation after the hyperparameters are chosen? Please describe so that it is clear the testing data was not used in the choosing of the hyperparameters.
We have rephrased the paragraph as: "We use a five-fold cross validation to evaluate different combinations of the hyperparameters, with 30 sets randomly sampled at each site. We use 80% of the data in the cross-validation, while the remaining 20% (selected without shuffling the original data set to avoid unfair predicting performance improvement because of auto-correlation in the data) is held out for independent testing."

5.) Figure 7 are the partial dependence plots for the random forest, which are an important aspect of the interpretability of the machine learning models. However, it is better to show both predictor importance and partial dependence plots so that the relative importance of each variable and its associated partial dependence is known. Recommend adding in predictor importance plots or list to add value to the interpretability.
We have added the predictor performance analysis as suggested by the reviewer. The following paragraph has been added:
"The results of the analysis of the predictor performance are listed in Table 5. As already suggested by the partial dependence analysis, wind speed at 65 m AGL is the predictor with the largest importance in extrapolating wind speed at 143 m AGL. However, all the considered surface observations account for over 30% of the overall performance of the random forest. In particular, the addition of the Obukhov length to include direct atmospheric stability information in the algorithm has a not-negligible 8% importance."
The following table has also been included in the manuscript:

Table 5. Predictor importance for the random forest used to extrapolate winds at 143 m AGL at site C1

Predictor	Relative importance
WS 65 m	68%
WS 4 m	18%
time	3%
L	8%
TKE	3%