Utilizing Physics-Based Input Features within a Machine Learning Model to Predict Wind Speed Forecasting Error

Daniel Vassallo, Raghavendra Krishnamurthy and Harindra J.S. Fernando

REVIEW – round 2

GENERAL COMMENT:

The authors have addressed most of my comments from the previous review, and the manuscript has improved. There are, however, a couple of major points that I think still need to be tested/addressed before the manuscript can be recommended for publication.

MAJOR COMMENTS:

- Page 3 line 28: In response to my previous comment on the topic, you now state that "No clear tower wake effects could be discerned." However, I do not think you have performed a necessarily correct check. The fact that wind speed and TKE are similar in magnitude for opposite wind directions does not necessarily mean that tower wake effects are not present: why should wind speed and TKE be equal in the first place for opposite wind directions? The correct way to assess potential impacts of tower wake effects would be to compare concurrent wind speed and TKE values as a function of wind direction as measured by two sonic anemometers mounted on opposite booms on the same met tower at the same height. From my knowledge, this configuration can be found in one of the meteorological towers at Perdigão. I strongly advise the authors to re-assess this aspect.

- Page 4: Given the randomized splitting between training and testing datasets, together with the absence of cross-validation, I am still concerned about potential overfitting, also considering the large autocorrelation your data have (due to the sum of that introduced by the overlapping averages and that naturally present in the data). While you state that random forests do not overfit the data at all, this is a debatable statement: I am sure you can find plenty of papers that can support both opinions. To make this reviewer happy, I would love to see whether the performance of the proposed model varies if the splitting between training and testing set is not performed randomly, but rather with a hybrid approach. For example, what happens if you keep all observations from one week for testing? Or from one full day from each week? Such tests would definitely give an answer to potential autocorrelation impacts on your results.

MINOR COMMENTS:

- Figure 1: this map still looks somewhat incomplete to me: at the very least, please add some reference to understand the horizontal distances.

- Page 4 l. 3: what do you mean by 'augmented' data here?

- Page 4 l.3: "data were averaged into 10-minute, hourly, and three-hour segments at a 5-minute moving average in order to create a robust dataset" is still not clear to me. Do you mean that you are creating three datasets, both with one data point every 5-minute, but in dataset A all data are 10-minute average, in dataset B hourly averages, and in dataset C 3-hourly averages?

- Figure 5: these scatterplots could be improved. Can you please change the color in the scatterplot based on density (e.g. https://matplotlib.org/api/_as_gen/matplotlib.pyplot.hist2d.html)?