Ref: wes-2019-58
Title: Utilizing Physics-Based Input Features within a Machine Learning Model to Predict Wind Speed Forecasting Error
Journal: Wind Energy Science

**Referee:** Anonymous Reviewer

We would like to thank the reviewer for his/her time in reading and commenting on the manuscript that led to considerable improvement of the paper. We have tried to address all comments and hope that this next version is acceptable for publication.

Major Comments:

1. I find the choice of the authors of (extensively) describing the methods used before the data a bit confusing. I personally had to go back to the methods section after reading the data section to make sure I got everything right. I would recommend switching the order of the two sections.
We agree with the reviewer that the order of the sections may lead to confusion/frustration. We have moved the section "Site, Data, & Instrumentation" ahead of the methodology section in order to relieve this issue.

2. More clarification is needed on what averaging time is used in the calculations of the variables considered in this work. For example, what averaging time is used to calculate the Reynolds decomposition for turbulent averaging, for example for TKE, TI, friction velocity? Why did you choose it? How does that conciliate with the different lead times of the ML models?
We have copied a majority of our response from the previous reviewer, as he had a similar question:
The utilized data had already been preprocessed and 5-minute averaged by NCAR; we had forgotten to include this information in the original and have added a reference on P3 L28.
5-minute averaging is a common averaging period used in most meteorological studies (for example CASES-99, RASEX, Perdigão, etc.) as it helps minimize flux sampling errors (systematic, random, and mesoscale variability error) and provides necessary flags to categorically distinguish between instrumental problems and plausible physical behavior (Mahrt et al. 1996, Sun et al. 1996, Vickers and Mahrt 1997). A local average of 5 minutes seems to adequately capture most of the turbulent fluxes in stationary time periods compared to one-hour local averaging (Mahrt et al., 1996, Sun et al., 1996). A 20 to 30-minute time-averaging protocol has become standard eddy-covariance practice for idealized conditions (i.e., quasi-stationary and horizontally homogeneous), but one can combine these 5-minute averages to obtain more statistically significant averages over longer time periods without much loss of information (Aubinet et al., 2012). Therefore, the authors did not venture out into testing other input averaging intervals.

References:
Aubinet, Marc, Timo Vesala, and Dario Papale, eds. *Eddy covariance: a practical guide to measurement and data analysis*. Springer Science & Business Media, 2012.
Mahrt L., D. Vickers, J. Howell, J. Højstrup, J. M. Wilczak, J. Edson, and J. Hare, 1996: Sea surface drag coefficients in the Risø Air Sea Experiment. J. Geophys. Res., 101, 14 327–14 335.

Sun, J., J. Howell, S. K. Esbensen, L. Mahrt, C. M. Greb, R. Grossman, and M. A. LeMone, 1996: Scale dependence of air–sea fluxes over the western equatorial Pacific. J. Atmos. Sci., 53, 2997–3012.
Vickers, D., and L. Mahrt, 1997: Quality Control and Flux Sampling Problems for Tower and Aircraft Data. J. Atmos. Oceanic Technol., 14, 512–526.

3. In addition, the authors should better clarify how the random split of train-test set mentioned in the paper is implemented. Do you mean that you are randomly picking 25% of data for testing, and then using those time stamps, once the algorithm has been trained, to predict wind speed 10-min, 1-hour, 3-hour ahead of each of the randomly selected time stamps in the testing?
The reviewer is correct, we have randomly selected 75% of the samples (input-output pairs) to train the model and then test the algorithm's efficacy on the final 25% of the samples. We have changed our wording on P5 L5-6 to reflect this altered explanation and have added two sentences on P4 L3 to better explain the data processing.

4. With such a huge data set as the one used in this analysis, I feel like the results shown could be greatly expanded, as a lot of additional analysis relevant to the topic could be made. After all, you are using 2 sonic anemometers out of an array of almost 200. For example, how does the performance of the used ML algorithms vary with atmospheric stability? Or with height? Do you find that different input features are more relevant close to the surface compared that at let's say hub height? Or how do the results vary in different complex terrain locations, for example comparing results from the valley and from the ridge tops? Please consider adding more analysis to this piece of work.
We agree with the reviewer that a plethora of studies could have been done covering a wide variety of topographical and climatological conditions, but we had to confine ourselves to a practicable endeavor requiring a reasonable effort that may lay the foundation for a variety of future studies with the dataset. In response to the referee's comments, we have added more analysis as to how the model performs with respect to changing wind speeds, wind direction, time of day, and turbulence. This additional analysis, which includes Figs. 7 and B1 as well as Table B3, can be found beginning on P12 L22. Other analysis mentioned by the reviewer is kept as a part of future work.

Minor Comments:

5. Abstract: introducing the symbols of each feature are not necessary in the abstract
Input feature symbols have been removed from the abstract

6. Figure 2-a: the map is not super clear.
Figure 2a has been replaced with a contour plot (Fig. 1) with a marker for the tower position

7. Figure 2-b: not really needed.
Figure 2b has been removed from the manuscript

8. P. 4: was wind speed at Perdigão really stationary? Over which time scales? Please clarify.
We have copied our response from the previous reviewer, as he had a similar question:

We have used the Augmented Dickey Fuller Test to test for long-term statistical stationarity within a given times series. This test has a null hypothesis that a given time series has a unit root, i.e. that it has a stochastic trend/drift that pervades throughout the entire time series. The testing procedure is applied to the model:

$$\Delta y_t = \alpha + \beta t + \sum_{i=1}^{n} (\delta_i \Delta y_{t-i}) + \varepsilon_t$$

where $\Delta y_t$, in our case, is the change in wind speed from one period to the next, $\alpha \neq 0$ represents a constant drift term, $\beta \neq 0$ represents a trend in the data, $\delta_i$ represents the dependency on the past $\Delta y_{t-i}$ term, and $\varepsilon_t$ is the residual. The number of lags, $n$, is chosen based on the Akaike information criterion (a standard process). The test results in a test statistic (the Dickey-Fuller test statistic) which can be transformed into a p-value that informs the user as to whether or not the null hypothesis (that the time series has a trend/drift) is likely to be true. The goal of this test is determining if the time series has any trend or drift that must be accounted for when running the ARIMA model. Generally speaking, we would like a p-value of $\leq 0.01$ (1% likelihood) to prove that the null hypothesis is false.

We tested for the likelihood that the data (the 10-minute, hourly, and 3-hour time series) could be represented by two basic regression models (these are the models most commonly tested in this type of analysis): a time series with a constant and a trend ($\alpha$ and $\beta \neq 0$) and a time series with a constant and no trend ($\alpha \neq 0$ and $\beta = 0$). Tests of all three time series on both regression models showed a p-value $\ll 0.01$ (the computer-generated p-values were all at least four orders of magnitude smaller than the 0.01 cut-off, meaning there was at most a 0.0001% chance of the null hypothesis being true), providing strong evidence that there is no underlying trend (i.e. change in the mean or variance of the wind speed over the course of the 3+ month campaign) in any of the time series.

To clear up what we believe may be the source of confusion, this test does not take into account any type of diurnal wind speed variations, instead testing to ensure there are no long-term trends/drift in the data. These diurnal variations are expected to be one constituent piece of the ARIMA forecasting error. We have changed the wording to "long-term statistical stationarity" and "wind speed data contains no embedded trends or drift (e.g. changes in the mean or variance of the wind speed due to long-term variability)" (beginning P5 L18) in order to relieve any confusion. We have also added the Python library utilized to perform the tests. However, we would prefer not to include the more detailed analysis above as this test was only one small ancillary piece of the analysis performed.

9. P. 6 L. 10: Rephrase as "a feature set that utilizes all input features is tested"

This phrasing has been added to the manuscript (P8 L9).

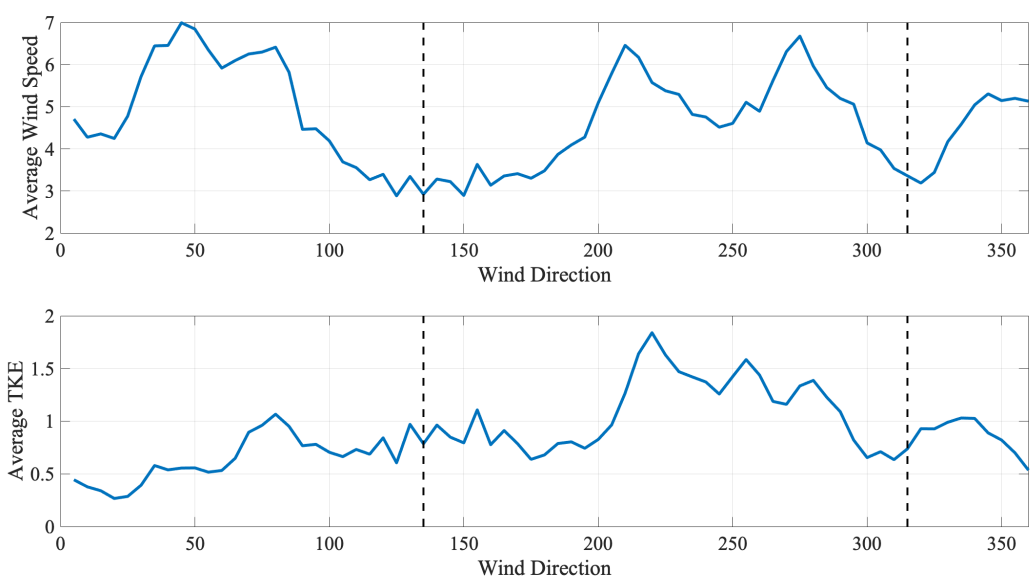
10. Did you apply any cross-validation for your ML models? If not, why?

By cross-validation we assume the reviewer is referring to the use of a validation set during the training process. We did not use a validation set. Unlike an artificial neural network, the random forest model does not require a validation set as it is inherently robust against the problem of overfitting (Breiman, 2001). The bagging process, when combined with a large number of trees and effective pruning (all described in Sec. 3.2), effectively obviates the necessity of a validation set – no changes.

11. P. 8 L. 1: Please specify what you mean by "sensors at 20 and 100 m AGL were chosen based on data availability."
These sensors were chosen because they had a relatively high percentage of clean data. We have clarified this statement: "the high percentage (>99% for all variables except temperature at 100 m AGL, which was available for ~95% of the periods) of clean data at these elevations" (P3 L30).

12. Please state the native time resolution of the sonic data you are using.
A statement has been added to P3 L25: "(20 Hz native measurement resolution)".

13. Have sonic anemometer data been filtered for tower wake effects? These effects would artificially increase turbulence (and reduce wind speed) for some wind direction bins, thus invalidating the quality of quite some data.
A line has been added to P3 L27 stating that no clear tower wake effects could be discerned. During quality control, the data has been checked for tower wake effects, but the two primary effects cited in the literature (reduced wind speeds alongside increased TKE; Barthlott and Fiedler 2003; McCaffrey et al. 2017) were not discerned in the dataset. The boom was angled at ~135º from northerly, meaning that the center of the tower wake would be expected at ~315º, approximately parallel to the ridge. Both the average wind speed and TKE in the expected wake region were similar to that seen in the opposite direction (135º), as can be seen in the figure below (dashed vertical lines indicate directions of along-ridge flow; 135º is opposite to expected wake, 315º is expected wake region; data is separated into 5º bins). Because we did not perceive wake effects, no corrections were made – no changes.




References:
Barthlott, Christian, and Franz Fiedler. "Turbulence structure in the wake region of a meteorological tower." *Boundary-layer meteorology* 108.1 (2003): 175-190.

McCaffrey, Katherine, et al. "Identification of tower-wake distortions using sonic anemometer and lidar measurements." *Atmospheric Measurement Techniques (Online)* 10.NREL/JA-5000-68031 (2017).