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: Javier Sanz Rodrigo

We would like to thank Dr. Rodrigo for his 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 revision is acceptable for publication.

Have the authors tested different input intervals to see the impact in the error reduction?
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.

P4 L7: Can you elaborate further on how wind speed data is stationary? Is this tested at the prediction timescales (10 min – 3 hr)? I would also expect wind speed to be subject to seasonal and diurnal variability. Please clarify.
We have used the Augmented Dickey Fuller Test to check 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 Δy_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, δ_i represents the dependency on the past Δy_{t-i} term, and ε_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 ≤ 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 (α 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.

P5, L28: I'm curious why are you not using the Obukhov length (or z/L)? Isn't it a more commonly used parameter to characterize stability? You may want to motivate this selection even though from the results of Figure 6 it seems that stability parameters are not that important in the improvement of forecasts.

We have decided against using the Obukhov length for this study because a few of the theory's critical assumptions (specifically spatial homogeneity) are broken in this location because of the complex terrain. Studies have shown that using the Obukhov length in complex terrain can lead to poor results (e.g., Fernando et al., 2015), and thus we have removed it from the list of potential input features. This has been noted on P7 L10.

References:

Fernando, H. J. S., et al. "The MATERHORN: Unraveling the intricacies of mountain weather." *Bulletin of the American Meteorological Society* 96.11 (2015): 1945-1967.

P7, L28: How are sonic measurements corrected for tilt? What is the interval used when deriving the fluxes? Is it equal, shorter or longer than the 5-min interval used in the moving average? This is just to know if 5-min is the actual filter in the data or if the data already came with a longer averaging time. This could also be relevant to understand the potential impact of this filter in the performance at 10 min prediction horizon (Figure 3).

The sonics were corrected for tilt via the technique described in Wilczak, Oncley, and Stage, 2001, "Sonic Anemometer Tilt Correction Algorithms," *Boundary Layer Meteor.,* 99, pp.127-150. This has been noted on P3 L26. The mean values for turbulent flux calculations were taken at 5-minute intervals, so the filter should not have had any effect on the 10-minute prediction forecasts.

P8, Figure 2: The map is difficult to read. It would be better to show an elevation contour plot where we can read the relative heights. I don't think it is necessary to provide an illustration of the mast levels if they are described in the text.

Fig. 2 (now Fig. 1) has been replaced with a contour plot with a marker for the tower position

P8, L8: You end up using 5-min averaged data to build predictive models with prediction horizons at 10 min, 1 hour and 3 hours. You previously mentioned that these are single-step forecasts. Wouldn't you have to use input data that is averaged at the same interval than the forecast step (e.g. use 3-hour moving averages to predict 3 hr ahead)? Or do you forecast {10min, 1hr, 3h} ahead based always based on 5-min data? If the latter is true, please clarify why not using a consistent interval between input and prediction data or, alternatively, how dependent are the results to the chosen interval in the time series.

The reviewer is correct, the data had already been 5-minute averaged (default from NCAR). We then had to average multiple 5-min periods in order to get the 10-min, hourly, and 3-hour averages for all variables. We have clarified this point on P4 L3.

P10. Figure 4: One may wonder how a Persistence-RF model would work. This might be a good result to include in the paper so that you can just isolate the impact of RF from that of the forecasting model to make the results more generally applicable. Maybe you get to the same conclusions with a simpler model.

We agree that this would be an interesting aspect to investigate, but we would rather exclude such tests for a few reasons. First, we worry that adding a Persistence-RF model would cause additional complications around justifying the feature set for such a model and the perceived need of designing another "optimal RF architecture". We would also rather keep the focus of the manuscript on how the random forest benefits a non-naïve model that is a function of previous atmospheric conditions (this idea is further described in the introduction). Additionally, we are currently working on a paper investigating the efficacy of a Persistence-RF model on decadal datasets (FINO1 and ARM SGP), and if readers are curious about such a setup for wind speed forecasting, they may refer to the upcoming manuscript. Therefore, we prefer to refrain from adding results from a Persistence-RF model – no changes.

P8. L11: How is the flux Richardson calculated between 100-20 m? Isn't it a local quantity derived from a sonic level? Is it the mean value between the two levels? Please clarify.

In order to calculate the flux Richardson number, we used the $\overline{w'T'}$, $\overline{u'w'}$, $\overline{v'w'}$, and T_{pv} measurements from the 100 m sonic anemometer (this has been noted on P16 L12). However, we used the u and v measurements from both the 100 and 20 m locations in order to calculate $\partial u/\partial z$

and $\partial v / \partial z$. While we recognize that the flux Richardson number is typically more localized, this technique was deemed as the best suited for this particular use case.