Ref: wes-2019-58
Title: Decreasing Wind Speed Extrapolation Error via Domain-Specific Feature Extraction and Selection
Journal: Wind Energy Science

Referee #1: Mr. Leonardo Alcayaga

We would like to thank Mr. Alcayaga 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 next version is acceptable for publication.

P. 1 Top "How does the ANN perform against other physics based models? Logarithmic profiles (which includes friction velocity, a measure of TI) and surface layer similarity theory (M-O) includes this and stability effects that perform much better than the power law. A comparison with these more physics-based approaches would make this assessment fairer."
We have added the log law as a base case for the study, with an explanation in Section 2.2 (P. 5 L. 3). We have also stated that M-O theory is not expected to hold in two of the three campaigns due to the complex terrain (P. 5 L. 8).

P. 2 L. 1 "Why beneficial?"
This explanation has been expanded to describe why physical models may struggle in certain regions, where machine learning methods may be a better option for modeling (P. 2 L. 1).

P. 3 L. 10 "What is the reference for this?"
We were not able to find a reference for this in the literature, so we have presented a quick explanation in Appendix A. The phrasing was also changed to "insensitivity to non-dimensionalization" rather than "universality" in order to better define the goal (P. 4 L. 5).

P. 5 L. 10 "What about wind speeds higher than the cut-off limit?"
A line was added stating that there were no wind speeds above 23 m s^{-1}, so no data needed to be removed due to a typical cut-off limit of 25 m s^{-1} (P. 6 L. 9).

P. 5 L. 18 Language
Fixed spelling

P. 5 L. 20 "What is reasonably complex?"
This has been clarified to mean that the terrain is complex, but less so than the Perdigão Campaign (P. 6 L. 20).

P. 5 L. 21 "What value was used here? SNR or CNR is in general in dB"
The CNR thresholds were -23 dB for both Casper and Perdigão, and -22 dB for WFIP2. – no changes

P. 6 L. 11 "There is also a dependency on the data available apparently"

Included a brief phrase mentioning the dependence on data availability (P. 7 L. 15). Also made a comment on the next line (P. 8 L. 1) stating that the WFIP2 project, which has the lowest extrapolation error, is also the site with the most robust dataset.

P. 7 Fig. 3 “Table”
Fixed the labeling of both Fig. 3 and 4 as Tables 1 and 2, as well as all references in the paper.

P. 7 L. 7 “This phrase is not clear. Because it is a feature like the others, it should have been neglected also for other cases”
The reason that we selected TI as a second input for the cases with three inputs is that it was the most beneficial input feature that included information about the flow’s turbulence levels, something that is expected to be highly influential particularly in the complex terrain sites. The number of tests would also become unwieldy if we were to test all features as a secondary feature. We have further clarified our reasoning in the manuscript (P. 8 L. 11).

P. 7 L. 10 “Why is this feature (which is a model, not data) is included as an input? How do you determine the difference of alpha and dudz in this case?”
One of the nice things about machine learning methods is that they can be used as the foundation for ensemble or hybrid modeling. U_α was used as an input to determine whether knowledge of another predictive model would be helpful as an aid for the ANN. In most cases it did not prove to be beneficial because it provided similar information as other inputs. The network is provided with the power law prediction (i.e. the extrapolated wind speed from the power law prediction) as opposed to the α value. Therefore U_α is distinct from dudz, discretizing the two input features. – no changes made

P 10 Fig. 6 “I suppose that this means percentage over the whole dataset”
This is the percent of data at each wind speed over the testing set, which is very similar to the distribution over the entire dataset. We have added a line beneath the figure to this effect.

P. 11 L. 1 “Since there is less data available for higher wind speeds, the sigma_error should increase for both cases (or at least become highly random). How is the normalization done in the non-dimensional case? Is normalized by the dimensional or non-dimensional mean WS in Equation (4)?”
Two additional sentences have been added below Equation 6 (P. 12 L. 9) further detailing why the non-dimensional network’s uncertainty remains so low at high wind speeds.

A sentence has been added just after Equation 6 (P. 12 L. 1) stating that, to calculate Equation 6 for the non-dimensional cases, U_n is first transformed back into U, so that we may find the true wind speed extrapolation uncertainty for all cases.

Referee #2: Ms. Tuhfe Gocmen

We would like to thank Ms. Gocmen for taking the time to reviewing this manuscript and helping us to improve the manuscript substantially. We have tried to address all comments and have specifically cleaned up much of the terminology in order to more accurately reflect what we see as the novelty and major findings in this study. We hope that this revised manuscript is acceptable for publication.

Please address to the attached comments in the pdf regarding the normalization/scaling being a standard procedure for network training, widely known for increasing the performance anyway.
We have changed the language used in the manuscript for clarification of our intent. Much of the time when we have said "normalization" what we have meant is non-dimensionalization, which is a feature extraction technique in fluid dynamics. We have also added a paragraph in Section 2.2 (beginning P. 5 L. 29) pertaining to the typical normalization process and why it was not used for this study. We have referred to normalization when a variable is divided by a quantity of same dimensions to scale it to a convenient value without considering dynamical implications.

Two sentences have been altered to further clarify the novelty of this study. These sentences may be found at the end of the Introduction (P. 2 L. 26).

P. 2 L. 4 "These references are highly redundant – suggest to keep the focus of deep learning applications to (at least) the wind energy field."
The references have been removed, as they did not add anything to the manuscript. More relevant meteorological citations have been added (P. 2 L. 15).

P. 2 L. 22 "Use either Sect. or Section – just a note on consistency"
All uses of "Sect." have been changed to "Section"

P. 4 L 7 "What is the ratio between the # of parameters to be fitted in your network to the number of data points?"
There were 986 model parameters, so the sample-parameter ratio ranged from approximately 2:1 to 23:1. – no changes

P. 4 L. 21 "Not clear if it is only the mean WS or more statistics extracted? (for the Second Base)"
The second base case only utilizes the mean streamwise wind speed at lower heights, no further statistics are extracted. – no changes

P. 4 L 22 "So the third base does not have WS as an input?"
The third base case does use WS as an input alongside wind direction and hour of the day – no changes

P. 4 L. 29 "More commonly referred to as the standard deviation of wind speed"
Language is changed to standard deviation of wind speed.

P. 4 L. 29 "What is meant is horizontal wind speed is normalized by the streamwise wind speed @ 20m below the height in question? (i.e. it is not clear if WS refers to the stream-wise or axial wind speed only)

Clarification has been added to this subsection that U refers to streamwise wind speed (P. 5 L. 11). WS has also been changed to U for improved interpretability.

P. 4 L. 31 "What does that mean exactly? How is WS_p different than WS_n (except of the time step of normalization) $WS_p = WS_{t-1}$, $WS_n = WS_t$, isn't it?"

That is correct, except we took U_p all the way up to extrapolation height, so it is assumed that the previous period's wind speed at extrapolation height is known. Clarification has been added in Section 2.2 (P. 5 L. 20).

Bottom of P. 4 "It would help with clarification if the authors include brief formulation of the non-dimensionalized inputs listed in this paragraph"

Clarifying notes have been added to Appendix B to describe how non-dimensional inputs were extracted.

P. 5 L. 1 "If the authors use the conventional cartesian nomenclature for wind speed, then it would be more clear to use u, v, w for all the wind speeds referred here (instead of WS, that does not have a clear direction defined for it so far)

We have changed the nomenclature for streamwise wind speed from WS to U in order to ease interpretability.

P. 5 L. 2 "In addition to the physically non-dimensionalized features (e.g. TI), it is not clear if additional scaling is applied as a part of data pre-processing for NN. If not, why? It is a pretty standard processing technique for deep learning problems, and shown to improve the performance of the networks majority of the times…"

We did not normalize the data in the typical manner because it did not have any discernable impact on the network's performance. However, we did rescale the wind direction inputs from $-180 \rightarrow 180$ to $-1 \rightarrow 1$ in order to relieve scaling issues. This resulted in changes to the results seen for the third base case, and updates to the manuscript have been made accordingly. We have added a paragraph in Section 2.2 addressing the question of input normalization (P. 5 L. 29).

P. 6 L. 7 "It is unclear if this uncertainty is taken into account for the input feature sampling…"

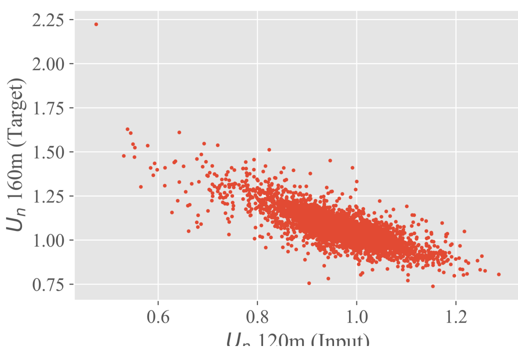
A sentence has been added at the end of Section 3 (P. 7 L. 9) stating that the lidar measurements are treated as true due to a lack of secondary measurements at the lidar locations.

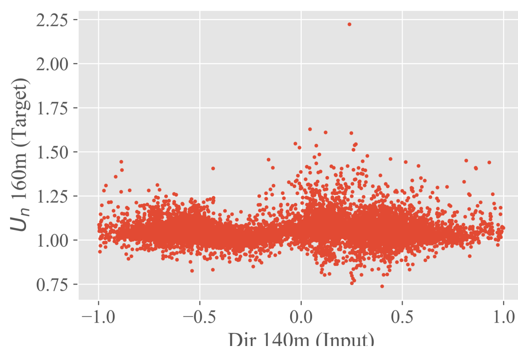
P. 6 L. 10 "Is it training + validation + test data? Or validation + test (I would suggest to use that)? Or only the test dataset?

The data shown in Table 1 was the training+validation+test, but have been changed to reflect the number of samples for validation+test. This is clarified at the beginning of Section 4 (P. 7 L. 12) and in the Table 1 description.

P. 7 Figure "Would be nice to see a pair plot or similar to see the correlation between these input features and the output – at least for the training dataset."
We agree that visualizing the correlation between inputs and the target variable is useful, but such a pair plot visualization would require dozens if not hundreds of plots (each input feature is found at multiple heights, and many correlations are different for each site/extrapolation height) and would be more overwhelming than beneficial for readers.

For the benefit of the reviewers, we have attached two scatter plots showing the relationship between exemplary input (x axis) and target (y axis) variables. The first plot, showing the relationship between U_n at 120m and 160m AGL, is a "best-case scenario," with an input variable that is highly correlated with the target variable (U_n at 160m AGL at Perdigão). These variables have a negative correlation that is scattered and contains some outliers. The second plot is a "worst-case scenario," or an input variable that is almost completely uncorrelated with the target variable. This plot shows the relationship between wind direction at 140m AGL and U_n at 160m, which are almost completely uncorrelated. – no changes





P. 7 Figure "Since WS_p as an input feature seems to include a time dependency in the problem, randomized split for train/validation/test dataset should be considered or clarified further."
A brief statement, in parentheses, is added at the beginning of Section 4 (P. 7 L. 12) stating that the data is randomly split. A more detailed description of data splitting is given at the end of Section 2.1 (P. 4 L. 18).

P. 8 Figure "It is not realistic to expect the selected NN architecture is the most optimum for all these essentially different problems… The potential effects of having more/less features fed into the training on the overfit/underfit should be addressed. That might be one of the reasons why there seems to be no clear trend in the performance of the networks when more information is added, or in terms of different terrain complexity…"
It is true that this exact network architecture is likely not the optimal architecture for each of these input scenarios and sites. However, tests were run where we both increased the number of nodes in each layer and added more layers to the network. None of these tests showed that a larger network provided more accurate results when more input variables were added (as mentioned in Section 2.1). This is likely due to two reasons:

The first is that the existing network architecture contains two dropout layers that reduce overfitting for testing cases with fewer inputs (Section 2.1). This allows us to use a larger network for cases with fewer input variables, essentially allowing us to have tailored the network architecture for cases with more inputs.

Second, many of the inputs contain very similar or interdependent information. Including more inputs, all of which contain the same information, should not improve ANN accuracy because the ANN is not receiving anything new that would improve its accuracy (e.g. inputting dudz or U_{α} repeats information already given by the U values). Instead, the additional redundant or noisy inputs actually decrease the ANN's prediction accuracy because they reduce its ability to pick up on meaningful patterns contained in the pre-existing input features (first paragraph of Section 5). – no changes

P. 8 L. 10 "That is yet another indication that scaling was not applied and should seriously be considered in order to assess the full potential of the developed networks."
This same result occurs both with and without standardization of dimensional inputs. Additional scaling of dimensional variables had a negligible effect. This is addressed in the paragraph added to Section 2.2 (P. 5 L. 29).

P. 9 L. 6 "Both are quite high actually… although the 'dimensional' features don't clearly work for higher wind speeds."
Correct, both models perform well when the wind speed is between 5-14 ms^{-1}. However, the non-dimensional network performs much better for high wind speeds. – no changes

P. 9 L. 13 "which error? Percentage error? MAPE? RMSE?..."
Clarification added: it is the standard deviation of RMSE.

P. 11 L. 11 "Rewording is suggested; 15-16 m/s is not necessarily extreme wind speed"
"Extreme wind speeds" changed to "high wind speeds" (P. 10 L. 5; P. 12 L. 18).

P. 11 L. 24 "More 'objective' wording is suggested"
"Unexpected" is removed as it does not add anything to the explanation.

P. 11 L. 27 "This is exactly why we normalize/scale all the input/output features, so the improvement in performance compared to non-normalized is not so surprising…"
"Normalizing" has been replaced with "extracting non-dimensional" for clarification.

P. 12 L. 28 "Domain knowledge definitely adds value to the resulting ML applications. However, a thorough investigation of how to best apply the ML technique as well as possible optimum architecture for different problems are also needed."
We agree that model optimization is something that must be investigated. However, a thorough investigation as to the optimal ANN architecture (as well as the optimal ML technique) is beyond the scope of this investigation. - no changes

P. 12 L. 29 "The references after the highlights sentence do not belong to the Conclusion."
These sentences have been moved to the introduction (P. 2 L. 7).

P. 13 L. 4 “I am afraid I have a hard time seeing the novelty in this conclusion, as further detailed in my previous comments attached to the normalization discussions.”
We appreciate the comment, these sentences have been altered to clarify that the benefit is seen from extracting non-dimensional variables rather than simple normalization (P. 14 L. 7).

P. 13 L. 19 “How so? It is not clear from the manuscript…”
The variables used are the only things altered in the code. The phrasing has been changed to add clarity (P. 14 L. 23).