Wind Energ. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/wes-2022-55-RC2, 2022

Comment on wes-2022-55

Anonymous Referee #2

Referee comment on "Introducing a data-driven approach to predict site-specific leading-edge erosion from mesoscale weather simulations" by Jens Visbech et al., Wind Energ. Sci. Discuss., https://doi.org/10.5194/wes-2022-55-RC2, 2022

The paper proposes a data-driven approach for predicting leading edge erosion damages. The main contribution of the paper is building a prediction model using realistic data. Data-driven methods for predicting edge erosion damages already exist, however according to the authors, they rely on data that are hard and/or expensive to acquire. The prediction model is an ensemble of feedforward neural networks that is trained using leave-p-out cross-validation to better utilize the limited amount of training data.

1) A comparison between the minimal feature approach followed in this paper and more expensive data-driven approaches mentioned in related work is needed to understand how the “data minimization” affects the quality of predictions.

2) The choice of the ensemble model should be better justified and a comparison to other types of models from interpretable ones like decision trees and linear regression models to black-box models should be added. Interpretability was mentioned as a requirement when discussing dimensionality (around line 255) so an interpretable model might reveal further insights regarding the important features for different damage types

3) Several points need further clarifications:

- Sparsity (Section 2.3) refers to missing data or missing labels? Please expand the discussion on ”sparsity and limitations in data availability” and provide e.g. % of missing data.

- it is not clear what the cardinalities of the training, validation and testing sets are

- the input features should become clear from the very beginning. This information comes too late now and becomes crystal clear only in section 4. A better idea would be to summarize the features in a table, including their value domain.

4) The organization of the paper needs improvement. The introduction section is too long and also covers related work, part of which is also discussed in section 4. I suggest a separate related work section. Section 2.3 is also too long and could be better organized into e.g., data, model and parameter tuning. The discussion section is very interesting but too long to follow. I suggest you split it into different subsections regarding e.g., feature/data choices, model choices, experimental findings etc. Also the discussion includes suggestions for future extensions, the title therefore should be changed accordingly.

5) I believe the novelty of this work is not the ML model but rather the minimal data approach that is followed to train such a model. The title of the paper should therefore be updated.

6) Figure 1 needs improvement, for example, the input features could be clearly indicated. Also, I find the current “ensemble splitting”, “model selection”, “training and validation”, “ML trained model” components not very informative, I believe the input data should be clearly depicted. The training and testing parts are confusing, it seems that only the weather data are used during testing.

7) Statements like “simple neural networks are also well suited as weak learners for ensemble modeling, whereas simple linear regression models are not” (line 315) “they are able to interpolate and, to some extent, extrapolate, which is not the case for other machine learning classes such as support vector machines or those based on decision trees” (line 315) should be supported by appropriate references.