Wind Energ. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/wes-2021-43-RC2, 2021

Comment on wes-2021-43

Anonymous Referee #2

Referee comment on "A physically interpretable data-driven surrogate model for wake steering" by Balthazar Arnoldus Maria Sengers et al., Wind Energ. Sci. Discuss., https://doi.org/10.5194/wes-2021-43-RC2, 2021

Purely data-driven flow/wake models are indeed 'under-published' over literature - but perhaps for good reasons. Although the approach itself is quite interesting, I would call for caution when calling it 'a wake model' as it reads more like a clever surrogate to map inputs to outputs for the given problem. Accordingly, I would suggest a more methodology-focused tone in the article rather than reading too much into the results as they are most likely not generalizable.

- Introduction: The novelty is clear but the motivation for purely data – driven wake modelling is lacking. Since physics – based, data – informed (through parameter fitting) wake modelling is a well – developed area, with several open access toolboxes/implementations as indicated in the article; why would we go for statistical models with even more parameters to fit? Especially if we have limited data?
- Section 2.3 and 3: Every step in Figure 3 should be explained anyway (again, more methodology oriented focus is recommended than too much weight on the results). It reads as if there exist another surface response modelling between the 'LES input variables' and 'input parameters'. Why is it needed? 15deg steps in yaw control settings are potentially too broad to have a smooth surface for any type of interpolations – its sensitivity and potential effects should be discussed.
- Section 3.2: "Although these input parameters might all have their own isolated effect on the wake propagation, they are heavily correlated in LES as shown in Fig. 5… Because of the highly correlated clusters, it is hypothesized that one is able to achieve reasonable accuracy in estimating key wake steering parameters with the regression model as long as each cluster is represented." ☐ High correlation between input and output features are indeed desirable for any regression problem but why a high-correlation among the input variables would help with better accuracy? How these *heavily correlated* effects are to be isolated so that the results are interpretable? How do you see the distribution of the parameters and their sensitivity in line with this high correlation? How these input features are correlated to the outputs in Table 2 should be the actual motivation to define the input parameters (together with their availability) but is not discussed at all…

- Section 3.4: Figure is frequently discussed here so it might help the reader to move it further down in the article, closer to Section 3.4.
- Section 3.5: Firstly, the subsection heading should be re-worded to something like "optimum model architecture" as optimization has many applications in the model fit and flow control discussed in the article. Secondly, the subsection generates more questions than answers – does it mean different set of input parameters might be used for different clusters? So there might be several, case-specific (or BL-specific) oSWSM and/or cSWSM?
- Section 4.1: "...The high correlation between variables is not an issue due to the use of the regression method described in Sect. 3.3" □ not true as stated earlier. It is an *issue* at least to isolate the effects of the change in input parameters, hence an *issue* for the interpretability.
- Section 4.1: Given the results in Figure 7, I suggest to omit oSWSM and cSWSM approaches to sharpen the focus of the article and avoid potential use of case-specific models (within already limited parameter space defined in LES). If agreed, all results should be updated with allSWSM.
- Section 4.2: "… Arguably, this is currently not a major disadvantage of the benchmark models as turbines tend to operate without derating (β = 0 â□¦)." □ should be omitted as turbines do operate under intentional derating with β > 0 â□¦ more often than intentional misalignment for wake steering. Therefore, it is an important achievement to be able capture this.
- Section 4.2 and 4.3: The main advantage of using data-driven approaches is the reduction of uncertainties. The reduction in the variability of the results should be mentioned/discussed at least once throughout Figures 7, 8 and 9.
- Discussion & Conclusion: A stronger discussion on interpretrability is required here. Great that the computational cost is presented but the number of parameters that is in SWSM as well as the other two benchmark models should also be mentioned. It would also be nice to discuss how explainable the whole procedure is, given all the collective data manipulations, normalizations etc. compared to a black box model. Additionally, do you see a potential trade-off between interpretablity and accuracy, i.e. do you think you might have got a better performance out of a black-box model?