Referee report:

"A physically interpretable statistical wake steering model"

by Balthazar Sengers, Matthias Zech, Pim Jacobs, Gerald Steinfeld, and Martin Kühn

1 General comments

This study presents a data-driven model to model the physics of a wind turbine wake under yawed conditions. The model employs a regression model, with inputs from large eddy simulation (LES) data, resulting in a linear model of the wake behavior. They then evaluate the performance of this model, as well as that of two other wake models, the Gaussian wake model and the Gaussian-Curl Hybrid model, against LES data. All of these models require tuning and/or training, so the same LES data was used for to prepare the models. Then a single case LES case that was left out of the training data was used for evaluation. The main focus of the study is to present a data-driven model that includes more physics and provides a linear model.

This paper provides an interesting new data-driven model, which results in a linear model for the wake behavior of yawed turbines. In particular, the scaling of the input parameters to enable the application of the model to a wide array of atmospheric boundary layer conditions is interesting, as changing conditions have been challenging for data-driven models. The paper presents an interesting approach and I would recommend for publication after minor revisions. Below is a detailed list of comments that should be addressed in a revision of this manuscript:

Specific comments

1. In line 66: the authors reference 'default numerical schemes' when describing the LES code. A more detailed description of these should be provided.

2. The caption for Figure 1: the authors give results for 'over the 15 main simulations'. It was unclear which simulations these were, specifically pertaining to their number. In Table 1, eight simulations are listed, so this is a confusing statement.

3. Section 3.4 seems very similar to the description in Section 3.1 and comes off a bit repetitive. Could the differences between them be clarified more?

4. In line 298: the authors mention '...which is due to the 'top head' shape of the wake deficit as a result of temporal averaging.' I'm familiar with the 'top-hat' wake shape but not the 'top-head'. If not a typo, the author should provide a more detailed description of this shape.

5. In line 235: 'To test whether a higher accuracy is achieved when more variables are included, allSWSM uses all (non-transformed) available variables of Table 3 as input.' The authors mention using all the variables rather than three. What (if any) is the time savings in training using only three variables versus all of the variable in Table 3?

6. In line 238: the authors mention that 'since the near wake is more dynamic and therefore needs more parameters to explain its variability'. The authors should provide citations

for this. The authors also mention that the near wake requires more parameters for this description. Are these factors already included in the parameters used in the input parameters or are they additional parameters outside this study? Do the authors have thoughts on what these parameters are? In Figure 7, we only see the results from x/D=4 onwards. Does the allSWSM method improve the near-wake performance at all?

7. In line 247: 'The models are trained or tuned with seven out of the eight BLs (Fig. 1) and tested on the remaining one representing a new inflow condition.' The authors mention the training and tuning of the three models compared in this paper. What is the order of magnitude for how long the training takes for this model? Is there a significant difference between the training of this model and the tuning of the other two wake models it is compared with here?

8. In line 302: 'In this study, large eddy simulation data were used to train the model, the generation of which is computationally expensive.' The authors mention that the model needs to be trained using LES data, which is a limitation. Have they considered whether the model could be trained using some combination of data obtained from an operating wind farm? This would save computational resources and customize the model to that specific wind site. The wake cross-sectional area data could possibly be obtained from strain measurements from the wind turbine blades, such as in Bottasso, Cacciola, and Schreiber, Renewable Energy, Vol 116, Part A, 2018.

9. In line 315-16: 'If desired, further development of the model is needed to include the near wake, which can for instance be done by including the super-Gaussian description introduced by Blondel and Cathelain (2020).' To my knowledge, the super-Gaussian description was also used in Shapiro et al. Energies, 2019; 12, no. 15: 2956.

Technical comments

1. In the description of Table 1: "...except for the domain size which is extended in streamwise direction". Typo, missing 'the'

2. In lines 60-61: "A precursor without and a subsequent main simulation with one turbine make up the simulation chain." Sentence is a little confusing. Hard to figure out what the simulation looks like.

3. In line 80: "(Sec. 2.2 and averaged over a line of size 2 d in crosswise direction and a period..." Typo, missing a parentheses) at the end of Sec. 2.2? Also a missing 'the'.

4. In line 125: "An expression for the wake width as function", typo

5. In line 153: The variable Y is mentioned twice, but X is not mentioned at all

6. In line 168: " IN this section..." typo

7. Figure 6: the second plot in the upper right corner of Figure 6 has no labels and is not mentioned in the caption.