Dear authors,

Congratulations with this excellent paper on a very thorough study. To my opinion there are far too little of these validation studies in the wind energy society!

I think the results are presented in a very clear way, although I had the feeling that the text could be a bit more concise. On the other hand I understand you want to be complete. I anyhow like the tabular summaries of these long pieces of text e.g. tables 1 and 2.

I wish I had more time to do a more thorough review but still I have a few comments:

- On some places you quantify differences between calculations and measurements but on other places you use very subjective assessments with terms like a good, poor or fair agreement. I realize very well that this is difficult to prevent but be aware that another person may come up with a completely different assessment. For example: You write on page 4, line 1 "It can be seen that the drag coefficient C_D is **slightly** different". I would write that the differences are huge….
- It is good to compare CP, CT and CMyaw but please be aware that a comparison on basis of these global integrated properties has little meaning. The only conclusion you may draw from a comparison of integrated loads is that a bad agreement means that there is something wrong. A good agreement doesn't say much because we very often see 'compensating errors'. An overprediction at the root may be compensated by an underprediction at the tip or vice versa. Some examples can be found in my PhD thesis. For yawed conditions with a delicate balance between root and tip vortex effects *) I expect this to be even more the case. In that sense I think that a comparison of e.g. local yawing moments from the different calculational methods would be extremely useful to better assess the aerodynamic modelling of the different partners (I realize this is a lot of work and it is not mandatory for me, but I strongly recommend this for a follow-up study)
- Does any of the lifting line methods, which use airfoil data, consider dynamic stall effects?
- On page 6, line 27 you write that the thrust is measured at the tower foot. I do note that the tower is included in the simulation (which is good, so you make a fair comparison) but to my point of view it still obscures the comparison a bit. Some of the differences in CT might come from the tower which are not so relevant for the wake properties since the wake is measured several ROTOR diameters (and very many TOWER diameters) behind the first turbine. In this respect: 1) Do you have "rotor off measurements", i.e. measurements of the tower alone, 2) how large is the tower drag compared to the overall thrust 3) has there been any treatment of the tower to prevent vortex shedding
- My most essential comments are about quality:
  - I am extremely happy that you include measurement uncertainties in the results Still I am a little bit surprised to see that the uncertainties are independent of the conditions. I would expect a dependency?
  - Are there any calculations which include the tunnel geometry. How large are tunnel effects? They might be more important than the measurement uncertainties in particular for yawed conditions.
  - What about the turbine quality: Have the blade geometries been scanned (small differences in blade geometry may lead to huge differences in airfoil polars at these low Reynolds numbers) and how accurate are the pitch angles? Are the pitch angles of all blades similar, and are the blade geometries similar? If not the aerodynamic

unbalance may obscure many of the results.
How accurate is the rotational speed?

I hope these comments are useful and I wish you success with the finalization of this article.

Kind regards

Gerard Schepers

*)_ For more information on this delicate balance between root and tip vortices, please have a look at: H. Rahimi, A. Martinez Garcia, B. Stoevesandt, J. Peinke, G. Schepers (2018). *An engineering model for wind turbines under yawed conditions derived from high fidelity models*:. Wind Energy. 10.1002/we.2182.