Response to Referee #2

Daniel Hatfield, Charlotte Bay Hasager, Ioanna Karagali

February 14, 2023

First of all, thank you very much for taking the time to review this article and for your positive comments and suggestions. Below are responses (in black) to all of the referee's comments (in blue).

The authors test a new model, based on machine learning (ML), to extrapolate ocean surface winds to hub heights of offshore wind masts, which is of relevance for the operation of offshore wind farms. Scatterometer ocean surface winds together with air-sea temperature differences appear the most important parameters for training the ML model. The model was trained for different time periods and the verification with independent data (not used for training) shows that the ML model outperforms an NWP model based on WRF.

General Comments

Although the authors have demonstrated that ML techniques can be used to extrapolate ocean surface winds to 100 meter altitude, they have not demonstrated that the methodology outperforms the use of already available NWP models.

We would like to remind the referee of the purpose of the paper in line 66: "The aim of this study is to assess the potential of using machine learning models with two-dimensional wind field observations at lower atmospheric levels in order to predict the wind at higher heights". The goal was obviously to try to out-perform model data but this is a *proof-of-concept* paper exploring the potential of satellite extrapolation through ML techniques vertically, horizontally, spatially and temporally. This method may not have out-performed the state-of-the-art NORA3 model, but we do believe that this article is still relevant for the area of wind energy and satellite wind retrievals.

It would be very helpful to shortly outline current operational practice of operators of wind farms. ASCAT winds and SST are freely available, so a ML model using these two input parameters could be an interesting option for operators in case they have no access to actual mesoscale model data. But is that the case? Although WRF is freely available as well, the NEWA dataset is of limited use for operators as it needs ERA5 as hosting model. However, ERA5 availability is at best a couple of days behind real time, and as such the NEWA approach of limited value for daily operations.

This method (as well as satellite wind observations) are more applicable in terms of wind resource assessment as opposed to forecasting or daily operational use in wind farms. Yes, we agree that the NEWA (WRF) approach is limited for daily operations, that is why in every comparison with NEWA (and for that matter NORA3) we have compared long-term statistics.

In the context of the above, can the authors please explain the relevance of the use of NEWA in their study? Also given that NORA3 outperforms NEWA WRF (line 326)

NEWA is well studied and provides a larger domain than that of NORA3. NORA3 does outperform NEWA and falls within the domain of interest in this work and therefore the inclusion of NEWA was to strengthen the argument for the comparison with NORA3.

The comparison against NEWA (WRF based) is not fair in the sense that NEWA does not make use of scatterometer winds explicitly (NEWA has no data assimilation), but only implicitly through the boundaries of the hosting model. A more fair comparison would be to use NORA3 in Table 5, because it outperforms NEWA as stated by the authors and probably makes explicit use of scatterometer winds in the reanalysis, although this was not mentioned by the authors.

NEWA was used in previous study in conjunction with ASCAT (in Hatfield et al. 2022) and was initially used in the ML comparisons until NORA3 was realized to have a better comparison. NEWA has been well study and used in many papers cited throughout this work, whereas NORA3 has few in comparison. We felt as though we needed to justify the use of NORA3 as a comparison, which is the initial idea behind the use of NEWA. The first comparison with the ML models and the FINO masts in Table 5 with the addition of a NORA3 comparison could help in further justify the use of NORA3 and the reason why NEWA is no longer used throughout the study.

"[NORA3] ... makes explicit use of scatterometer winds in the reanalysis, although this was not mentioned by the authors"

NORA3 "runs explicitly resolved deep convection and yields hindcast fields that realistically downscale the ERA5 reanalysis." (Haakenstad et al. 2021) whereas NEWA uses ERA5 for dynamical forcing (Dorenkamper et al. 2020) so both use ERA5 in the model chain. But even so, the use of satellite wind observations is more to do with the wind resource than forecasting. The two simulated datasets may also have different use cases, however they are still a large wind dataset that is used over a 12-year period which provides similar domains and resolutions.

Major Comments

1. Although it was not the ultimate goal of the study to test if a model based on ML does outperform an NWP model, it is of importance to operators of offshore wind farms to know if ML outperforms a mesoscale NWP model. NORA3 represented the latter, so please add NORA3 to Table 5.

The ultimate goal of the paper was to (line 66) "The aim of this study is to assess the potential of using machine learning models with two-dimensional wind field observations at lower atmospheric levels in order to predict the wind at higher heights". The use of NWP was a means of evaluation/validation due to the lack of offshore wind observations other than model data.

As to the addition of NORA3 to Table 5, we completely agree, it is a little pointless to have just the NEWA comparison. It also justifies the use of only NORA3 in the spatial extension of the ML model. This is now added.

2. Table 5. The column denoted 'N' shows for ML the number for “Concurrent data with ASCAT“ (although the numbers are not exactly the same with those in Table 3). This is misleading as the numbers in the other column (RMSE, bias, ..) are based on “Data used for validation”. Please use the correct numbers in Table 5.

Thank you for pointing this out. At the FINO1 site there are 6180 data points concurrent with ASCAT (as stated in Table. 5 & 6) this was double-checked and changed in Table 3. The $N = 5739$ for FINO3 is consistent across all three tables. As to the N-values in Table 5. We will leave them as is. The point of Table 3 was to demonstrate that 4/5 of N was used to train the model and the other 1/5 was used as evaluation.

3. Section 2.5. As a non-expert in ML techniques, this section was too abstract and hard to read and understand. It would help to relate the parameters in Table 2 to, X,Y and T in the text. A formula would help (see below). How does y_overbar relate to the parameters in Table 2? Is it wind speed at e.g. 100m?

This is a good point. We have related the RFM variable with the inputs and outputs used in the paper in line 154: "In the case of this study, $\overline{y}$ will be the predicted wind speed at higher heights (107 m at FINO3 for example), where Y will be the concurrent wind speed measurements at the mast at the desired height." and in line 154 "... input data (X in the equation above)". The only variable that does not related to data in this study is T, the trees, which is inherent to the RFM itself.

4. The paragraph on hyper-parameters and K-fold cannot be understood without any background knowledge of these techniques. For me it was totally unclear. We would remove it from the text. The last paragraph in section 3.1 concludes that wind at height is mainly modelled through wind at the surface (WS) and the air-sea temperature difference (AT-SST). In formula: ML(j,FINOi) = a(i,j)WS + b(i,j)(AT-SST), with j denoting altitude and i the FINOi (I=1,2,3) station. The training set then aims to estimate a(i,j) and b(i,j). Is that right?

We agree, it is worth mentioning the hyperparameters and the ranges evaluate for the RFM optimization, but the explanation of the K-fold method may be too much. We have removed section on the k-fold method leaving the necessary information of the Hyperparameters:

"While model parameters are "learned" during the training phase, *hyper-parameters* are set before the training to create a more accurate algorithm. Hyper-parameter tuning relies on experimental results of combinations of model parameters to evaluate the performance of each model. ~~To avoid over-fitting the model, the K-Fold cross-validation method is applied. The data is split into testing and training sets which is split further into five subsets, K. The model is trained iteratively K times, evaluating on the K-th fold, changing on each iteration.~~ The hyper-parameters are varied and their associated ranges are outlined in Table 4."

5. (See also remark to section 2.5 above). Line 426: "Results from this study show the prospect of applying machine-learning methods for the purpose of extrapolating surface winds to higher atmospheric levels".I think this statement is too strong as the study does not show that RFM outperforms mesoscale models which assimilate ASCAT. Please correct.

We agree the language throughout the text needs to be softened and has been. In this particular example we have change the word "prospect" to "potential" as this work lays the groundwork for the potential of extrapolation through the use of machine-learning algorithms.

6. I can imagine a seasonal dependence of the ML model parameters for the different station locations. Was this tested? Please comment.

Interesting point. We have looked at monthly averaged spatial extensions but decided that the number of samples was too small and that yearly spatial comparisons were more realistic and inline with other satellite extrapolation methods (i.e. see Badger et al. 2015 and Karagali et al. 2018 using the long-term stability correction). Optis et al. 2021 show the dependency of stability conditions on the ML extrapolation of low level lidar winds measurements (lower RMSE in unstable conditions compared to stable conditions) which is a seasonal phenomena. Thus we would expect, for example, the model to perform better in the Autumn months (predominantly unstable conditions) than in the Spring months (predominantly stable conditions).

Minor Comments

Line 91. "This 12.5 km product has a standard deviation of 1.7 m s-1 and a bias of 0.02 m s-1 (Verhoef and Stoffelen, 2019)." I guess this is for wind speed, not the wind components? Please make clear in the text.

It is now clarified

Line 144; why 3x3 grid. Given the NORA3 2 km grid size and ASCAT 12.5 km product, I would expect 6x6, since 6*2=12, which is close to the 12.5 km ASCAT footprint. Please explain the choice for the number 3.

NORA3 is a 3 km grid size hence the choice of a 3x3 grid. This is outlined in section 2.4 in line 134.

In Table 3, how were the "Data used for validation" selected? Randomly from the "Concurrent data with ASCAT"?

This is a good question and is now clarified in the text. The "Data used for validation" is 1/5 of the "Concurrent data with ASCAT" whereas the "Data used for model training" is 4/5. These are randomly chosen 4/5 and 1/5 splits as we tried to explain through the k-fold method, but our explanation of the k-fold method was unclear in your Major Comments, hopefully it is now clearer.

Figure 2, right panel. Why does the number 1148 differ from 1147 in Table 3? Please correct.

Corrected

Figure 5.These numbers are based on validation data only, so N=1148 (or 1147), right?

For both Figure 4b and Figure 5b, the model that was trained/validated (4591/1148) was then applied to the data for 2018; this includes the ASCAT 10m wind speeds (over N=750, Figure 7b), the FINO low-level atmospheric data and the FINO/satellite SST values. Basically, the model that was trained earlier was then applied to the 2018 data alone. We have now tried to explain this in the text that introduces these figures.

Table 6. The number -0.003 in the 7-th column should be -0.004, in agreement with Table 5? Please check.
Yes, good spot. We have re-run these to verify and it is now corrected.

The numbers below Table 6 are not clear at all. 63% comes from 1.196 − > 1.949. Why is this relevant? 8% comes from 1.803 − > 1.949. This change is indeed important to report. The same issue applies to the numbers in ". . . . 65% or by 4% . . . . . .".
The first (i.e. 63%) shows the spatial increase in the RMSE of the model trained at FINO3 whereas the second one shows the increase in RMSE from a model that was trained at FINO1. Although we see a large increase in the RMSE from extending the model from FINO3 (63%) we are only slightly increasing the error from a model optimized for FINO1 (8%) suggesting either the FINO1 91m measurements could be flow distorted (which can also be seen in the higher RMSE with NEWA or in validation such as in Witha et al 2018) or that the ASCAT gird cell could be influence from the high density of wind turbines (Figure 1) giving higher scatter - both of which are interesting consequences. We do agree with the referee that the comparison with the model trained at FINO1, for example, is more relevant. The comparison with the FINO3 model is now removed and the percentage changes mentioned above are now included in Table 6 for clarity.

Line 218: how do you arrive at 1% and 2% (Table 6). There is lots of information in Table 6, but not clear enough explained in the text.
Table 6 now includes the percentage changes and these 1% and 2% (FINO1 traning model extended to the FINO3 location compared to the original model trained at FINO1) are irrelevant and removed from the text.

Line 134: "while the structure and features of spatial variability in the wind field are not maintained". Was that expected? Why?
No this was not expected. Similar wind fields to the 10m observations are seen but at a much smaller scale as seen in Figure 4b & 5b (changes in wind speeds of around 0.3 m/s across the entire 125km^2 area). Even with the spatial SST addition to the training (i.e. Figure 5a) there is very small changes across the entire 125km^2 area. This suggests the constants or unchanging data used the in the model training (i.e. air temperature, pressure, humidity) from the FINO3 mast may have a larger influence on the extrapolation than expected.

Figure 4d. The title mentions: 'ASCAT-NORA difference 100m', which is confusing given that ASCAT is representative of 10m. Clearly, RFM based on ASCAT is meant. Please correct.
Corrected

Line 340: the text: "The noticeable improvement of the model compared to NORA3 is evident" is misleading as it suggests that the RFM model outperforms NORA3, but this has not been assessed in the study (see comments on Table 5). What is meant is ". . . taking NORA3 as reference (truth) . . ."
This is true, and has been changed in line 345 as "An improvement of the RFM model taking NORA3 as reference is seen in Figure ...". It should be noted that we have tried to soften the language throughout the text to not over-exaggerate the results of the RFM.

Line 372: "vertical wind measurements". "vertical wind" is confusing. What is meant is: wind profile or profile of (horizontal) wind. Please correct.
Corrected, thank you

Typos

All typos were fixed that were noted by the referee, thank you for taking the time to read the manuscript so carefully.