Response to Referee #1

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).

General Comments

"The paper describes the application of machine learning algorithms to vertically extrapolate near surface winds derived from satellites data. The focus is offshore, in Northern Europe. The paper is generally well written (although some grammar improvements here and there would be recommended), and it represents a nice academic application of machine learning to the wind energy sector. However, I question the practical utility of this approach in real world applications. The main reason for this is that satellite-derived observations of wind speed at a given location are only available twice a day, and always at around the same times. I am really struggling in finding a situation where someone would be interested in knowing hub-height wind speed only at these two hours of the day. While practical applicability is not a strict requirement for having a paper published, I still think this limit should be at the very minimum highly stressed in the paper, and language softened to reflect the limited applicability of the results described in the analysis."

Thank you very much for you comments and attention to detail while reviewing the manuscript. We agree that the practical application of the methods presented in the study are limited, but would like to remind the referee that this was a *proof of concept* style of paper, exploring the potential of using machine-learning method as a way of extrapolating satellite data as opposed to the alternatives that have not had much success. This was done in hopes to further utilize the invaluable daily satellite surface wind coverage for longer-period wind resource assessment.

We do agree that the language used in the paper needs to be "softened", in which sections and word-choices have been changed to stress the limitations/potential of applying machine learning methods.

Specific Comments

1. L.19: add "above the sea surface" or something similar. Added

2. L. 26: 275 m seems like a very specific threshold – can you provide a reference? We have included references for both ferry and buoy based lidars with heights up to 275m in line 26 as (Rubio et al. 2022 and Hatfield et al. 2022)

3. L.45: "predicting" instead of "predict" (or "to predict") Added

4. L. 66: "higher" instead of "greater" Added

5. L. 106: in the text you mention water temperature, while in Table 1 you mention sea surface temperature; please be consistent. This has been changed in Table 1 to Water Temperature

6. L.111: do you mean that all variables have exactly an availability of 85%, or greater than 85%? We have clarified this in line 113 "All measured quantities show a data availability above 85% with the exception of WT (76%)"

7. L.123: what's the temporal resolution? The temporal resolution is *daily* coverage (which is now clarified in the text in line 121). With the high heat capacity of water small diurnal changes are observed and uncertainties are provide by Høyer & Karagali 2016 (also mentioned in the text). It should be noted that the SST product has a mean difference of -0.06°C compared to moored buoys and a 0.46°C standard deviation of the differences.

8. Consider moving all the details regarding data access to the 'Data availability' section towards the end. Thank you for the suggestion, we will however include the data in the text to remain consistent with other papers using the satellite/FINO data.

9. Figure 1 caption: "shapes" instead of "rectangles". Added

10. It is hard to fully understand Table 3 (and some of the discussion in this section) without a clear explanation of the temporal frequencies considered here for the various data sources. Does "total data" refer to 30-min average time periods? And "Concurrent data with ASCAT" to 30-min average time periods in which an ASCAT data point was recorded? Please clarify in the paper. This is correct, "total data" refer to 30-min average time periods at the masts, "concurrent data with ASCAT" refers to 30-min average time periods in which an ASCAT data point was recorded. This is now updated in the Table 3 label.

11. Table 2: "FINO" is repeated twice in the left column. Removed

12. Why using the cosine of wind direction only (and not the sine, too)? We have used both at one point in the early stages of this work where using the cosine of the wind speed performed better in the ML training (resulting in lower RMSE).

13. Once again, being clear about the temporal resolution of the data used is key to understand whether the random split between train and test sets is a right choice, or auto-correlation effects might play a role in artificially enhancing the ML results. Added in Table 2 caption as: "All of the data measured from the FINO masts are 30 minute averaged."

14. Table 5: I would argue that the most relevant comparison is ML vs NEWA at each site, so I would suggest adding a horizontal line after each site, and highlight in bold the "winner" metrics at each site. This is a very good suggestion and is now incorporated in Table 5 along with the addition of the same comparison with NORA3 at FINO for the same periods as NEWA.

15. L.170: clarify that the error values refer to the test set. This is clarified in line 171 as "...with the test dataset..."

16. L.173: "Note" instead of "not". Added

17. Section 3.2: please clarify how the mean wind profile from the RF was computed. Do you simply apply the RF to the whole period, and average results? Or to the test set only? This is now clarified in the text in line 172: "The model trained in Section 3.1 is applied to the entire 12-year collocated dataset at all heights at FINO3 from 31 m to 107 m."

18. Section 3.3: why not including FINO2 as well? FINO2 is much too far away and the Baltic sea has a very different marine atmosphere. As this is still an article exploring the *proof-of-concept*, we have limited the spatial extension / round-robin approach to the North Sea. It should also be noted in the ML papers extrapolating wind speeds (i.e. Optis et al. (2021) and Bodini et al. (2020)) that use the round-robin approach, their distance are less than 100km apart, whereas the distance between FINO1 and FINO3 already exceeds that of previous work (136km). Thus, the comparison with FINO2 is beyond the scope of this article.

19. Section 3.4: can you somewhat verify your hypothesis of horizontal homogeneity by looking at spatial variability of the meteorological variables from NEWA and the reanalysis product? In Figures

4a and 5a we can see the spatial variation of both the wind speed (from NORA3) and SST (from DMI L4 SST) both with very low variation across the 125km^2 area. The FINO3 area is also in open ocean, far from the coast and without islands within the study area. We think this is a very fair assumption to make. We have also verified using ERA5 data for the North Sea that there is small variations in air temperature (± 1°C similar to that of the SST variation) across the chosen area which is not shown.

20. Figure 4: once again, more clarity is needed when explaining what is being plotted and described. What's the temporal extent of what is shown? Are NORA data taken only at time stamps at which ASCAT data are available? How about the ML-extrapolated winds? Are we really comparing apples to apples? Once that is clarified, please adjust text accordingly (it is misleading to state you are showing 2018-averaged winds, if you are only cherry picking time stamps). ASCAT is the limiting feature, all time-stamps need to be concurrent with ASCAT. This is however is not the case in Figures 4 & 5, these included yearly averaged 30 minute NORA3 grids and have now been updated to be concurrent with ASCAT so that we are comparing "apples to apples". The Figures 4 & 5 are now updated to include concurrent NORA3 data instead of yearly averaged for 2018.

Thank you very much for the comment!

21. Figure 4: why wasn't NEWA included? Cheynet et al. (2022) has shown that NORA3 out-performs the NEWA at FINO1 but this is not mentioned until the Discussion. A clarification is now added in lines 228 "It should be noted that only NORA3 will be included in the spatial comparison with the RFM as it has out-performed NEWA at the FINO3 mast in Table 5 and in Cheynet et al. (2022) at FINO1.". NORA3 comparison was also added in Table. 5 to further this argument.

22. Conversely, why wasn't NORA3 included in the earlier analysis (Table 5)? It is essential to know how well it compares to observed winds in order to use it as proxy for the truth here. Added

23. Figure 6: please change labels and instead list ALL the satellites available in each time period. Also, specify that this is a cumulative plot. What do you mean by "concurrent" in the caption? Concurrent is confusing, it is now changed to "Cumulative number of samples of ASCAT observations at the FINO3 location from 2010-2022. The vertical lines represent the launch of each MetOp satellite as well as the decomission date of MetOp-A." in the Figure 6 label. The Figure is now updated to show which satellites were also in operation during this time period.

24. Discussion of Figure 10 is key (see my main major comment above), and in my opinion should be moved way earlier in the paper. I agree that this is a major point of discussion, but we do think it is in the right place within the results. This paper explores the implications of the use of ML on satellite extrapolation where we explore vertically (wind profile), horizontally (round-robin) such as previous work but also the spatial (in comparison with NORA3) and temporal (in sampling) domains due to the nature of satellites. We completely agree that the discrete nature of the satellites is one, if not, the largest limiting factor within this paper, however we think that is explored in the discussion. Whereas, the layout of the paper slowly builds on expanding the simple vertical extrapolation of the satellite at each FINO mast.

25. Figure 10 caption: specify when referring to bars vs lines. Also, specify you are referring to local time and not UTC time (I believe). All times throughout the paper have been recorded in UTC (ASCAT, FINO, SST, NORA3) which we have added in the Data section in line 146 as "all data used from all sources is recorded in Coordinated Universal Time (UTC)". But we have added in the Figure for clarity and clarified the bars/lines.

26. L.332: decreases instead of increases? Yes, changed.

27. Data availability: why not sharing the model algorithm scripts as well? Unfortunately, there is no intention from the authors to publish the scripts as well.

28. Please double check references and make sure each has a DOI. Done

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.