In this document, the reviewer's comments are in black, the authors' responses are in red.

General comments:

Overall, the manuscript is well written and structured in an understandable way. The introduction provides enough background and references to connect the article to the current state-of-the-art and the methodology section provides a high degree of details to understand and follow the workflow. Some elements in the methodology and data descriptions however require, in my opinion, additional description or additional references. While the analysis in general is presented concise and in a convincing way, I see the need for further investigation & provision of statistics with respect to the claimed generalization of the machine learning approach (see specific comment #2). Some additional words on limitations of the methodologys range of application would be appreciated as well. While I would suggest also some minor changes in the text and data presentation, my recommendation is publication after the questions/comments below have been addressed adequately.

We thank the reviewer for their thoughtful comments, which we addressed in our response to the specific comments below.

Specific comments:

general remarks:

#1 Usage of the term "boundary condition and parametric uncertainty"
The authors introduce above mentioned term to describe the share of uncertainty that can described by NWP ensemble runs appearing several times in the manuscript. This term, however, can be highly misleading and misinterpreted especially in the context of regional NWP where the term "boundary conditions" and boundary condition "uncertainties" are used in a different context. I would suggests a term like "Ensemble-derived uncertainty" or something along those lines to avoid misinterpretation.
We now use "ensemble-derived uncertainty" as suggested throughout the paper. We kept only one instance of the old terminology to specify that this was the terminology used in the Bodini et al. 2021 article.

#2 Generalization of the machine learning approach
Section 5 describes the details of the validation of the machine learning approach. While the round-robin cross- validation approach shows promising behavior, the authors acknowledge the impact of spatial correlation due to the close vicinity of locations of validation. Here, I think it very crucial to quantify this impact by e.g. calculating mutual correlation coefficients of the Lidar time series to fully understand how independent the training and validation data actually are (maybe presented in a similar manner to Fig.4). This would also provide more details to the generalization skill and to what extent the high generalization skill is achieved purely due to high correlation of training and validation data. In this context, it would be also valuable if the mutual distance between the Lidars could be stated.
We added the following paragraph to Section 5:

When interpreting these results, it is important to also consider the correlation existing between the considered data sets. In fact, the R^2 coefficient between observed hourly average 140 m wind speed from the closest pairs of lidars is quite high. For example, $R^2 = 0.84$ when considering the NYSERDA E05 and E06 lidars (about 75 km from each other), and $R^2 = 0.88$ between the NYSERDA E06 and Atlantic Shores 06 lidars (about 55 km from each other). These values drop significantly when considering larger distances, so that we have $R^2 = 0.55$ between the observations of the NYSERDA E05 and Atlantic Shores 04 lidars, which are over 145 km apart. Finally, it is important to remember that the Atlantic Shores 04 and 06 lidars do not have any overlapping time in their periods of record, so that their time series can be considered independent from each other. Given these considerations, it is reasonable to expect that the existing correlations between the data sets have an impact on the good generalization skills found here, but only up to a certain level. For example, the remarkably strong generalization skill found between the two NYSERDA lidars is likely connected to a combination of their long period of records and strong autocorrelation. On the other hand, the random forest still performs well when trained and tested at lidars over 140 km apart (the NYSERDA E05 and Atlantic Shores 04 lidars), and even when trained and tested at two lidars (Atlantic Shores 04 and 06) with no overlapping period of records. Therefore, while numbers in Figure 3 are not immune from existing correlations, the overall good generalization performance of the extrapolation algorithm in the relatively limited geographical region considered in our analysis is confirmed.

Finally, we added information on the distance between the lidars in the caption of Figure 1: "The distance between the two NYSERDA lidars is about 75 km, the two Atlantic Shores lidars are about 20 km from each other, and finally the distance between the NYSERDA E05 and Atlantic Shores 04 lidars is about 145 km."

#3 More details in description of numerical WRF setup
While the setup is described fairly detailed, the following information is missing:
- Nudging is mentioned in l. 72, but it remains unspecified if grid or spectral nudging has been used. Please specify and provide details on parameter settings if they differ from the default settings in WRF.
- Land surface model, Microphysics, Longwave/Shortwave radiation, topographic data base and land use data in Table 1 lack references. Please add them for completeness in line with the other specifications in Table 1.

We added a specification that we used spectral nudging, and references to all missing entries in the table.

#4 Limitations of the proposed methodology
In the current version of the manuscript, very little is talked about the limitations of this methodology. While there is a statement made about the atmospheric conditions at the buoy north of Cape Cod (l. 244), a more in-dept critical elaboration is required in my opinion (maybe as additional paragraph in Sect. 6). This concerns especially the reliability / trustworthiness of the method when the random forest is applied to locations that are very different from the training data (geographic location, distance to training data, atmospheric conditions).

We have added a whole new paragraph to critically address the choice of the random forest model, and whether using fewer input variables could be enough to provide an accurate uncertainty quantification, to more critically explore some of the potential limitations of the proposed methodology.

We also added some text to the conclusions to further stress that the results are location-dependent, and how having access to more and more hub-height observations would be beneficial in lowering the uncertainty associated to the machine learning extrapolation model.

Remarks addressing specific lines or sections:

l. 8: Since it is the abstract, please be specific about your used method (random forest) instead of the generic term "machine learning technique".

Changed as suggested.

l. 22: The stated reference for the currently installed offshore wind farms in the US is around 7 years old, please update with newer reference (maybe GEWCs Global Wind Report) to confirm and to use more up-to-date data.
We changed the sentence to "While the United States currently only have 42 MW of installed offshore wind capacity (Global Wind Energy Council, 2023), …".

l. 31: The reference (Skamarock et al. 2008) points towards Version 3 of the WRF model, but your WRF version seems to be 4.2.1 (Table 1). Is there a particular reason why the Version 3 reference is used here and not Version 4? Otherwise, please update.
Thanks for catching this – we updated the reference.

l. 125/126: I would recommend to mention here again that the variables used for training are coming from the Lidar to avoid any ambiguity about the input for the training process.
We added the following specification: "We use the following observed variables as inputs to the model" (we did not specify "lidar" as some variables are technically coming from other instruments mounted on the lidar buoys).

l. 135 (footnote): The part " [...] which are both needed because each value of sine only (or cosine only) is linked to two different values of the cyclical variable" is unclear to me. What does it mean? Please consider elaboration or reformulation.
We added details and rephrased this as "To preserve the cyclical nature of this variable, we calculate and include as inputs its sine and cosine. We note that both sine and cosine are needed to identify a specific value of the cyclical variable, because each value of sine only (or cosine only) is linked to two different values of the cyclical variable. For example, the sine of wind direction is 0 for both 90◦and 270◦, but once their (different) cosines are introduced as well, the two can be identified in a univocal way".

l. 149: What exactly do you mean by "typical single-site uncertainty"? Is this the averaged standard deviation of the residual time series for a particular location or something else? Please elaborate.
We added the following specification "(i.e., the average of each site's standard deviation of the residuals)".

l. 270: Replace "wind energy" with "wind turbine power production"
Changed.

Technical corrections:
Language corrections:
l. 3 [...] heavier relative weight [...] → [...] heavier is the relative weight [...] Rephrased as "and the resulting heavier relative weight".
l. 144 Then, to assess the uncertainty → To assess the uncertainty Changed.
l. 161 Introduction → introduction Changed.
l. 256 [...] to numerically model [...] → [...] to model [...] Changed.

Figures:
Fig. 1: For completeness, please state in the figure caption which markers indicate Lidar locations and which markers indicate buoy locations. This is currently unclear by looking at the figure only.
We added the following sentence to the caption "Lidar locations are shown as diamonds, NDBC buoys are shown as dots."

Fig. 6: I would suggest to transform this figure to a table since the bars for the parameters in the middle do not convey much information. For completeness, I would also suggest adding the explanation of "SST" to the caption similar to "WS" and "WD".
Changed.