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

The authors thank the reviewer for their thoughtful comments, which helped us improve the quality of our manuscript.

The article "The importance of round-robin validation when assessing machine- learning-based vertical extrapolation of wind speeds" by Bodini & Optis details a round robin approach to vertical extrapolation from the 4 ARM SGP Doppler lidars using a random forest algorithm. The paper is well written and the reviewer agrees the necessity of a round-robin type approach to assess the accuracy of machine learning algorithms and make it more universal. Below are some comments/questions which probe into some of the details of the paper and would improve the paper if addressed in the next version.

1. Line 61: Extrapolation is not only generally done up to Hub-height but through the rotor swept area. So, I am not sure I follow the author's argument here, that if hub- height winds are available extrapolation is unnecessary. The approach to go above hub-height can also be treated as a "Gap filling" approach for met-masts (when Lidars are moved around from one location to the other for a short period). Please clarify.
   We agree with the reviewer that the approach should not be limited to hub-height wind speed extrapolation, but can rather be used to obtain wind resource at any height of interest for wind energy production. To make this clear, we have changed the wording "hub-height wind speed" to expressions such as "heights relevant for wind energy production" or "heights of the rotor swept area" throughout the manuscript.

2. For power law type extrapolations, measurements not only at the surface but at multiple heights is needed to estimate the dynamic power law exponent. So please define what you mean by near-surface in the paper? Is it within surface layer or also above surface layer?
   We have rephrased the sentence in the introduction as "By contrast, conventional extrapolation approaches do not have nor require knowledge of hub-height wind speeds and therefore can generalize to any location where measurements are available at a single level near the surface (for the logarithmic law) or at two levels in the lower part of the boundary layer (for the power law).".

3. The authors mention LLJs, frequently observed in the ARM site, how does this effect the ML output at higher heights?
   We have extensively studied ML extrapolation for LLJ events in a companion conference paper which is currently in review. We have added the following sentence to the Results section, after the analysis of the ML extrapolation performance with height: "As an application of the performance of the random forest in predicting wind speed at higher heights, we present the case study of a LLJ in a companion paper (Bodini and Optis, in review)." We will update the reference if the conference paper is reviewed before this manuscript is accepted for publication.

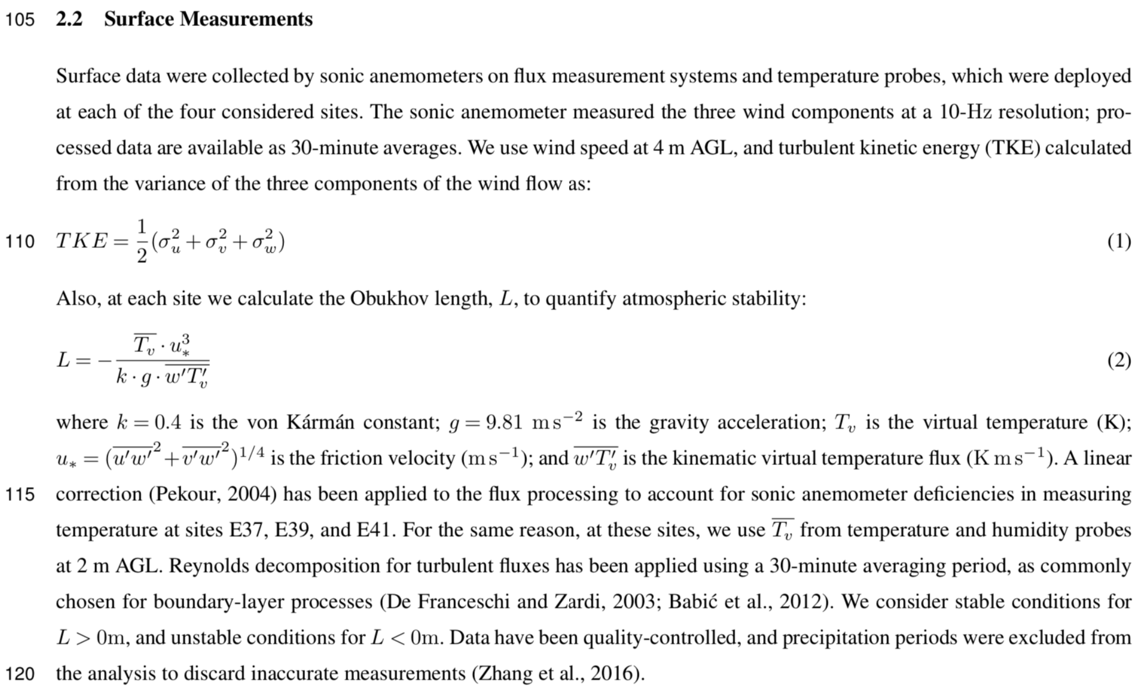
4. The idea of round-robin is fair for machine-learning based extrapolation, but only if the training has been done accounting for all atmospheric conditions that would be representative of other sites. As you know, ML models can only learn what is in their

training dataset. Therefore, the round-robin type approaches come with a caveat that the search space of the variables expands to many of the common conditions (including external forcings specific to each site) observed in the atmosphere and at all the evaluated sites. This comment needs to be addressed in the paper with supporting evidence.
See answer to comment 10.

5. For the SNR filtering, not only precipitation, but fog is also prevalent at SGP and it diminishes the range considerably at lower heights. Therefore, an upper limit on SNR could be important to filter out any abnormalities in radial velocity data.
We have re-done our analysis by setting an upper limit on SNR, chosen after inspecting the data. We have rephrased the sentence in Section 2.1 as: "We discard from the analysis measurements measurements with a signal-to-noise ratio lower than −21 dB or higher than +5 dB (to filter out fog events), along with periods of precipitation, as recorded by a disdrometer at the C1 site.".

6. Line 94: Maybe I am picky, but the poor data quality is because those measurements fall within the lidar blind zone? The Blind zone is generally 2 times the range-gate size, which fits the heights. If yes, please mention that for clarity.
We have rephrased the sentence as follows: "Data recorded at two lowest heights (13 and 39 m AGL) could not be used because of their poor quality, as they lie in the lidar blind zone."

7. Equation 2: The temperature used was from the sonic or from the cup anemometer for the fluxes? Sonic anemometer temperature measurements have significant biases and are not considered very accurate (Berg et al., 2017). This would cause errors in classifying stability or L.
We have discussed with the instrument mentor at ANL, who confirmed that the flux data provided on the ARM website have been linearly corrected to account for the instrument issues the reviewer is mentioning. On the other hand, for the average temperature data, which were not corrected, we have now switched to use data from the 2-m temperature and humidity probe as done in Berg et al. 2017. Section 2.2 now reads:



2.2 Surface Measurements

Surface data were collected by sonic anemometers on flux measurement systems and temperature probes, which were deployed at each of the four considered sites. The sonic anemometer measured the three wind components at a 10-Hz resolution; processed data are available as 30-minute averages. We use wind speed at 4 m AGL, and turbulent kinetic energy (TKE) calculated from the variance of the three components of the wind flow as:

$$TKE = \frac{1}{2}(\sigma_u^2 + \sigma_v^2 + \sigma_w^2) \tag{1}$$

Also, at each site we calculate the Obukhov length, L, to quantify atmospheric stability:

$$L = -\frac{\overline{T_v} \cdot u_*^3}{k \cdot g \cdot \overline{w'T_v'}} \tag{2}$$

where $k = 0.4$ is the von Kármán constant; $g = 9.81\ \mathrm{m\,s^{-2}}$ is the gravity acceleration; T_v is the virtual temperature (K); $u_* = (\overline{u'w'}^2 + \overline{v'w'}^2)^{1/4}$ is the friction velocity ($\mathrm{m\,s^{-1}}$); and $\overline{w'T_v'}$ is the kinematic virtual temperature flux ($\mathrm{K\,m\,s^{-1}}$). A linear correction (Pekour, 2004) has been applied to the flux processing to account for sonic anemometer deficiencies in measuring temperature at sites E37, E39, and E41. For the same reason, at these sites, we use $\overline{T_v}$ from temperature and humidity probes at 2 m AGL. Reynolds decomposition for turbulent fluxes has been applied using a 30-minute averaging period, as commonly chosen for boundary-layer processes (De Franceschi and Zardi, 2003; Babić et al., 2012). We consider stable conditions for $L > 0\mathrm{m}$, and unstable conditions for $L < 0\mathrm{m}$. Data have been quality-controlled, and precipitation periods were excluded from the analysis to discard inaccurate measurements (Zhang et al., 2016).

8. How is atmospheric stability defined? Based on Richardson number of MO length? Please provide the thresholds or a reference from which you picked the thresholds for classifying stability for the MO type extrapolation.
   We have added the following sentence in Section 2.2: “We consider stable conditions for L > 0 m, and unstable conditions for L < 0 m.”.

9. MO length is not known to be valid for complex terrain (Fernando et al., 2015), therefore these parameters would not fit well for all types of terrain/sites. Therefore, a note about applicability of the chosen parameters to different conditions/terrains would be needed to address the universality of these parameters for such an approach.
   We have added the following sentence to Section 3.3: “We note that when similar techniques are applied to more complex sites, the Obukhov length might not be well-suited to capture atmospheric stability in complex terrain (Fernando et al., 2015), and therefore an accurate choice of the input variables as a function of the specific topography is recommended.”.

10. The effect of external forcings at different ARM sites are not considered, which is important in this context of machine learning (comment #4 above). The wakes from wind turbines have major impact on the hub-height winds at some of these sites. Sites E37 and E39 are far away from turbines or wind farms, while C1 and E41 are relatively closer and have considerable impact on the winds at hub-height in certain predominant wind directions. Please see attached the wind directions and distance from wind turbines at each of these sites and something similar must be included in your analysis. Therefore, I would recommend you can either discard the below sectors from your analysis or test the accuracy in waked conditions.

We agree with the reviewer that different forcings experienced at different sites have an importance when assessing the round-robin validation of the proposed machine learning method, and that explicit emphasis on this caveat should be included in the analysis. For the specific comment about the impact of wind farms, we have extensively studied this topic in the aforementioned companion conference paper.

We have added the following discussion paragraph to the Results section to make all these thoughts explicit to the reader:

“Moreover, we can expect the performance comparison to be influenced not only by the pure separation between training and testing sites, but also by the different forcings that each specific site experiences. Notably, Bodini and Optis (in review) compared the extrapolation performance of the proposed random forest approach before and after a wind farm was built in the vicinity of site C1, and found an increase in MAE up to 10% if waked data are not included in the training set. Therefore, to fully exploit the performance of the proposed machine learning approach in extrapolating the wind resource at sites different from the training one it is essential to build a training set of observations which can encompass the specific atmospheric conditions representative of the desired testing site.”

11. How much of these chosen parameters (TKE, L, WS4, WS65) explain the variance in the RF model? What is the unbiased predictor importance estimates of the chosen variables?

We have added the predictor importance analysis as suggested by the reviewer. The following paragraph has been added:

“The results of the analysis of the predictor performance are listed in Table 5. As already suggested by the partial dependence analysis, wind speed at 65 m AGL is the predictor with the largest importance in extrapolating wind speed at 143 m AGL. However, all the considered surface observations account for over 30% of the overall performance of the random forest. In particular, the addition of the Obukhov length to include direct atmospheric stability information in the algorithm has a not-negligible 8% importance.”

The following table has also been included in the manuscript:

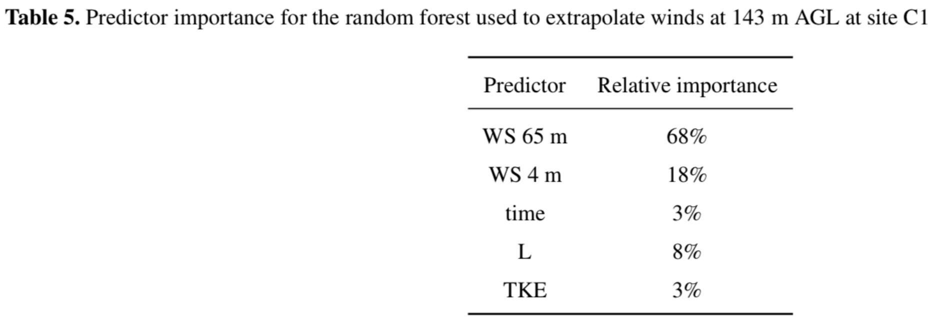


Table 5. Predictor importance for the random forest used to extrapolate winds at 143 m AGL at site C1

Predictor	Relative importance
WS 65 m	68%
WS 4 m	18%
time	3%
L	8%
TKE	3%

12. Figure 7: Maybe some additional explanation is required on how the dependence is calculated. It’s not very clear if it’s just a correlation type analysis or something else. Please provide more details here. Also, the extrapolated wind speeds (Y-axis) for all plots are not same and it’s not very clear why.

We have improved our introduction to Figure 7 in the results section, and added a reference where more information on partial dependence analysis can be found. The paragraph now reads: “Figure 7 shows the partial dependence plots, which show the marginal effect of

> each input feature on the predicted extrapolated wind speed (Friedman, 2001). We note that the values on the y-axes have not been normalized, so that large ranges indicate strong dependence of extrapolated wind speed on the feature, whereas small ranges show weaker dependence.”.

Very Minor comment: The language is a bit colloquial for a journal and would urge the authors to take that into consideration for their revised manuscript. For example: Line 225: Maybe you can but I am not sure if it’s formal to end a sentence with “are”: please rephrase. Similar sentence structuring needs to be considered throughout the document.

We have rephrased the sentence as “Distributions of the input features are also shown, which help distinguish densely populated regions, with strong statistical relationships, and sparsely populated regions, with weaker statistical relationships.” The whole manuscript has undergone editorial review by a professional native English-speaking editor.