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 paper presents a robust measurement campaign and analysis results. The thoroughness of the various analyses is commendable. The sensitivity analysis at the end of the paper is particularly insightful as it helps shed light on the reasons for the performance of the machine-learning (ML) approach used by the authors. Indeed, the risk when using ML is to blindly depend on a black box which may, or may not, provide reliable output. especially for new situations for which no data was included in the training set.

The following questions and comments are provided in the hope of enhancing the readability and overall reach of the paper:

0. One could argue that power law and log law are also machine learning approaches - even though they are simple regressions!
   Fair point! As the power law and log law are so well-known and broadly used in the wind energy community, we think referring to them as "conventional techniques" will be of easier understanding for the general reader of the paper.

1. In the intro, low-level jets (LLJ) are mentioned. Provide some more background as they are not ubiquitously present, nor relevant. Or specify that " in some regions ...".
   We have added "in some regions" in the introduction sentence.
   Also, 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)."

2. Provide a better presentation of the measurement campaign - notably, do not forget to add the missing paragraph which was posted: - Site description - Typical wind regime description - Lidar precision/accuracy/validation/testing discussion as the wind industry is still considering scanning lidars with a lot of caution. Or, provide discussion that high lidar accuracy is irrelevant in this context because ... - Provide an idea of the total number of data samples used. - Any data quality applied?
   We have included the paragraph that was missing in the first draft, and added details across Section 2 to include the suggestions of the reviewer. Section 2 now reads as follows:

2 Data: The Southern Great Plains (SGP) Atmospheric Observatory

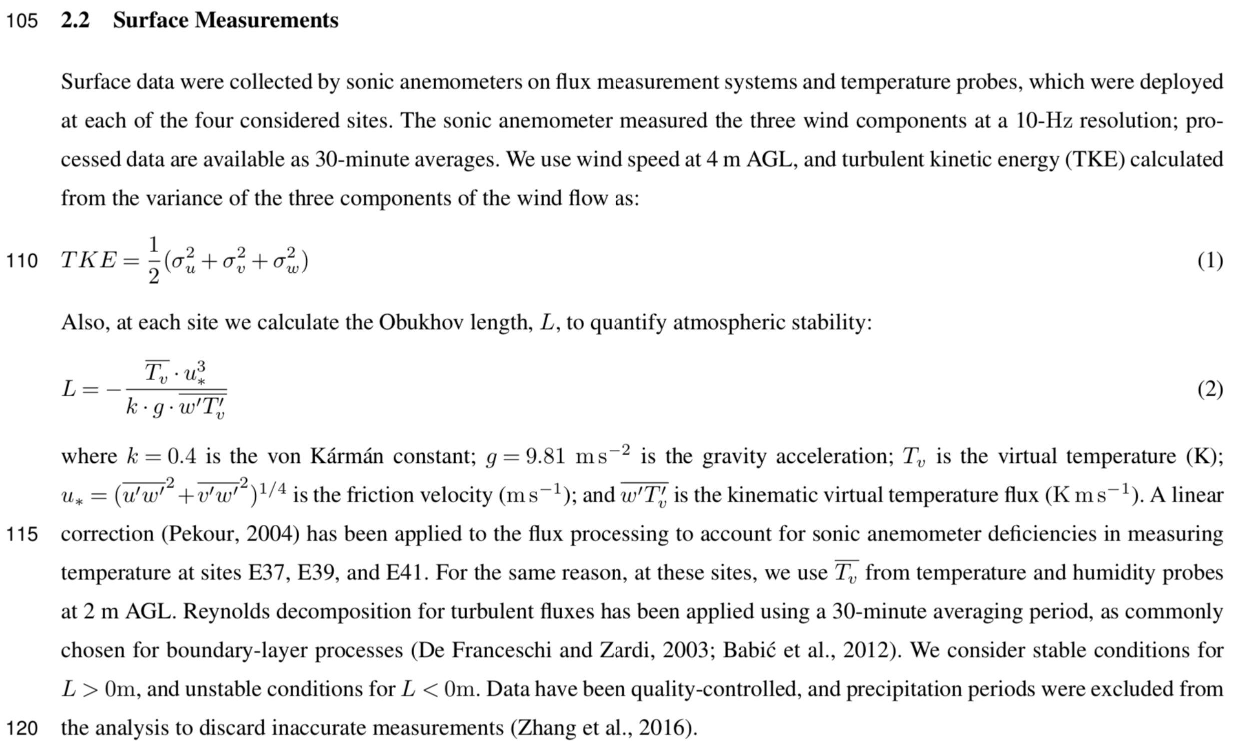
We use observations collected at the Southern Great Plains (SGP) atmospheric observatory, a field measurement site in north-central Oklahoma, managed by the Atmospheric Radiation Measurement (ARM) Research Facility. To assess the variability in space of the performance of machine-learning-based wind speed vertical extrapolation, we focus on four different locations at the site (Figure 1), over a region about 100 km wide. The site is primarily flat, and its land use is characterized by cattle pasture and wheat fields. Winds mostly flow from the South, with more variability observed in the winter. For our analysis, we use 30-minute average data from 13 November 2017 to 23 July 2019 (for a total of over 29,000 timestamps).

2.1 Lidars

At each of the four locations considered in our study, a Halo Streamline lidar (main technical specifications in Table 1) was deployed. A preliminary intercomparison study of the lidars performed by Atmospheric Radiation Measurement (ARM) research confirmed that all the lidars produce consistent measurements, with correlation coefficients greater than 0.9, and precision less than 0.1 m/s (Newsom, 2012). The lidars performed a variety of scan strategies. For this analysis, we retrieved horizontal wind speed from the full 360° conical scans, which were performed every $\sim$10-15 minutes and took about 1 minute to complete. We use the velocity-azimuth-display approach in Frehlich et al. (2006) to retrieve the horizontal wind speed from the line-of-sight velocity recorded in the scans. To do so, we assume that the horizontal wind field is homogeneous over the scan volume, and that the average vertical velocity is zero (Browning and Wexler, 1968). We discard from the analysis measurements with a signal-to-noise ratio lower than $-21\,\mathrm{dB}$ or higher than $+5\,\mathrm{dB}$ (to filter out fog events), along with periods of precipitation, as recorded by a disdrometer at the C1 site. Finally, processed data were averaged over 30-minute periods. For this study, data from five range gates are used, corresponding to heights of 65, 91, 117, 143, and 169 m AGL. 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.

Table 1. Main Technical Specifications of the ARM Halo Lidars

Wavelength	$1.5\,\mu\mathrm{m}$
Laser pulse width	$150\,\mathrm{ns}$
Pulse rate	$15\,\mathrm{kHz}$
Pulses averaged	$20,000$
Points per range gate	10
Range-gate resolution	$30\,\mathrm{m}$
Minimum range gate	$15\,\mathrm{m}$
Number of range gates	200



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).

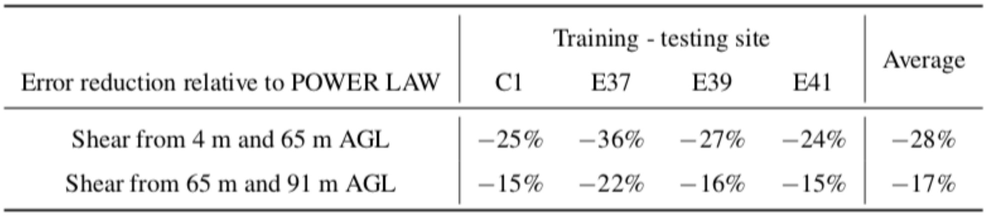
3. The wind industry also uses by-sector and/or by-hour-of-day vertical extrapolation. These are targeting a couple of shortcomings the authors note, namely: stability and terrain complexity. It would be useful to add this in the discussion - or even better, in the analysis.
   We have added the following analysis to the results section:
   "In addition, it is important to check whether the results of the performance comparison are affected by the time resolution at which the shear exponent α is calculated. Wind energy consultants apply a variety of methods to calculate shear (Brower, 2012): one could calculate shear values at each timestamp (as done in our analysis), or use a single average shear exponent, or consider various shear values based on bins of wind direction and/or time of day. To compare the time series-based shear calculation with its most different approach, we test the performance of the power law in extrapolating the average wind resource from 65 m AGL to 143 m AGL using only a single mean value for the shear exponent, calculated as the average of the α values at each considered timestamp. We find that the average extrapolated wind speed from the random forest approach still has a smaller error compared to the average extrapolated wind speed using the mean shear value, at all the considered sites (across-site MAE for random forest is 0.01 m s−1, for power law is 0.13 m s−1). Given the overall small MAE values found for both methods, we can also conclude that machine-learning-based extrapolation approaches are most beneficial for time series-based extrapolations, as deficiencies in conventional approaches tend to average out more when considering the long-term average results."

4. My understanding is that the authors optimized the hyper-parameters by making use of available target-height measurements. So what is the authors' suggestion to fine-tune these parameters in the absence of target-height measurements? Could we contemplate a

database of parameters for specific site conditions? Other? More generally, how their round-robin results could be leveraged, used on site?
We have added the following sentence to the Conclusions of the paper: “In real world applications, a machine learning algorithm could be trained on observations collected by a single lidar, and then used to extrapolate wind speed at nearby locations, where only much cheaper short meteorological masts would need to be installed”.

5. Lines 192 and following: any particular reason comparison results for the specific use case under discussion were not more thoroughly reported?
   We have added the following table to the Supplement (and added reference to it in this paragraph in the main paper) to support our description of the comparison between ML and power law performance when data at 91 m AGL are included in both methods:

   Table 1: Percentage reduction in wind-speed extrapolation MAE from the random forest approach over the power law when wind shear is calculated using data at 4 m and 65 m AGL versus at 65 m and 91 m AGL. In the latter case, wind speed at 91 m AGL is included as input feature for the random forest model.



	Training - testing site				Average
Error reduction relative to POWER LAW	C1	E37	E39	E41	
Shear from 4 m and 65 m AGL	−25%	−36%	−27%	−24%	−28%
Shear from 65 m and 91 m AGL	−15%	−22%	−16%	−15%	−17%

6. Personally, I find the last sections of the paper to be the most valuable ones! Without suggesting to re-write the whole paper, I submit the following ideas for author’s consideration: - Put the emphasis on the fact that more physical parameters where included in a data-driven model, and their impact on model performance was investigated and fully understood (cf. sensitivities). - The model seems to out-perform standard models, even under round-robin conditions (which is indeed a better way of assessing the model). - The model could be used for a given site as follows (might need more thought to be put here ...)
   In the Results section, we have added an extensive discussion of feature importance to further emphasize the importance of being able to understand and quantify the different input features used in the machine learning model:

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%

"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."

We have also added the following sentences to the Conclusions:

"The benefit of including more physical parameters in a data-driven model clearly demonstrates its importance."

"In real world applications, a machine learning algorithm could be trained on observations collected by a single lidar, and then used to extrapolate wind speed at nearby locations, where only much cheaper short meteorological masts would need to be installed."

We have also rephrased the following sentence in the Conclusions to further emphasize that the round-robin validation still outperforms conventional techniques:

"Therefore, we have confirmed that the random-forest approach outperforms conventional techniques for wind-speed vertical extrapolation, even under a more robust round-robin validation, which we recommend to avoid overestimating the potential performance of machine-learning techniques, which could lead to underestimation of the uncertainty in wind speed estimates."

Thank you for having submitted a paper which makes a balanced and useful use of ML!

Thank you for taking the time to review our manuscript!