Ref: wes-2019-58
Title: Utilizing Physics-Based Input Features within a Machine Learning Model to Predict Wind Speed Forecasting Error
Journal: Wind Energy Science

Referee: Anonymous Reviewer

We want to sincerely thank the reviewer for once again taking the time to provide an in-depth review of the paper. The comments have helped us make considerable improvements to the manuscript and we hope the updated version is acceptable for publication.

Major Comments:

1. Page 3 Line 28: In response to my previous comment on the topic, you now state that "No clear tower wake effects could be discerned." However, I do not think you have performed a necessarily correct check. The fact that wind speed and TKE are similar in magnitude for opposite wind directions does not necessarily mean that tower wake effects are not present: why should wind speed and TKE be equal in the first place for opposite wind directions? The correct way to assess potential impacts of tower wake effects would be to compare concurrent wind speed and TKE values as a function of wind direction as measured by two sonic anemometers mounted on opposite booms on the same met tower at the same height. From my knowledge, this configuration can be found in one of the meteorological towers at Perdigão. I strongly advise the authors to re-assess this aspect.

We were unaware that there were two sonic anemometers at the same height on a single tower, and we have since performed the suggested assessment. The test showed small shadow effects (maximum of 7% decrease in wind speed, shown below) in a directional sector spanning about 30º (~310-340º from northerly), a much smaller shading effect than we have seen quoted in much of the literature, which are generally more than 30% (e.g. Moses & Daubek, 1961; Cermak & Horn, 1968; Orlando et al., 2011; McCaffrey et al., 2017, Lubitz et al., 2018). After examining this effect, we have added a statement in the manuscript stating that small tower shading effects were present for this sector (beginning P3 L26). However, we would still prefer to keep the unaltered data in the study for two reasons.

First, the ARIMA model is particularly useful for continuous datasets and has a built-in assumption that the dataset is continuous. As mentioned in Section 2, we have filled in missing data periods because we wanted to ensure that the model would have a continuous dataset. The 310-340º directional sector constitutes approximately 5% of the dataset and removing this data would lead to considerable data partitioning, thereby negating much of the efficacy of the ARIMA model. Further, all periods which utilize the removed data as inputs would themselves be negated. The dataset is already relatively small for a machine learning application, and we would therefore prefer to avoid removing these periods, especially due to the finding that the shading effect is on the order of 7%.

We would also prefer not to replace the data via a correction such as the measure-correlate-predict (MCP) method utilized to fill the missing periods (Section 2). While we were required to use the method to fill missing periods, it can produce errors (~17% mean absolute error) that are greater than that seen from the tower shadow effect. We do not expect that this would be a problem

for the filled periods, as they constitute only a very small portion of the dataset (less than 1%). However, filling 7-8% of the dataset via MCP could lead to misleading results.

We would therefore prefer to keep the dataset as-is, as any adjustment would likely lead to further complications and potentially deleterious results. As stated previously, we have added a statement in the manuscript that slight tower shading was observed, but we have left the data unaltered as it was far less prominent than that quoted in the literature. We would like to thank the reviewer for referring to the information that led to this finding.

References:

Cermak, J. E., and J. D. Horn. "Tower shadow effect." *Journal of Geophysical Research* 73.6 (1968): 1869-1876.

Lubitz, William David, and Andrew Michalak. "Experimental and theoretical investigation of tower shadow impacts on anemometer measurements." *Journal of Wind Engineering and Industrial Aerodynamics* 176 (2018): 112-119.

McCaffrey, Katherine, et al. "Identification of tower-wake distortions using sonic anemometer and lidar measurements." *Atmospheric Measurement Techniques (Online)* 10.NREL/JA-5000-68031 (2017).

Moses, Harry, and Hugh G. Daubek. "Errors in wind measurements associated with tower-mounted anemometers." *Bulletin of the American Meteorological Society* 42.3 (1961): 190-194.

Orlando, Stephen, Adam Bale, and David A. Johnson. "Experimental study of the effect of tower shadow on anemometer readings." *Journal of Wind Engineering and Industrial Aerodynamics* 99.1 (2011): 1-6.

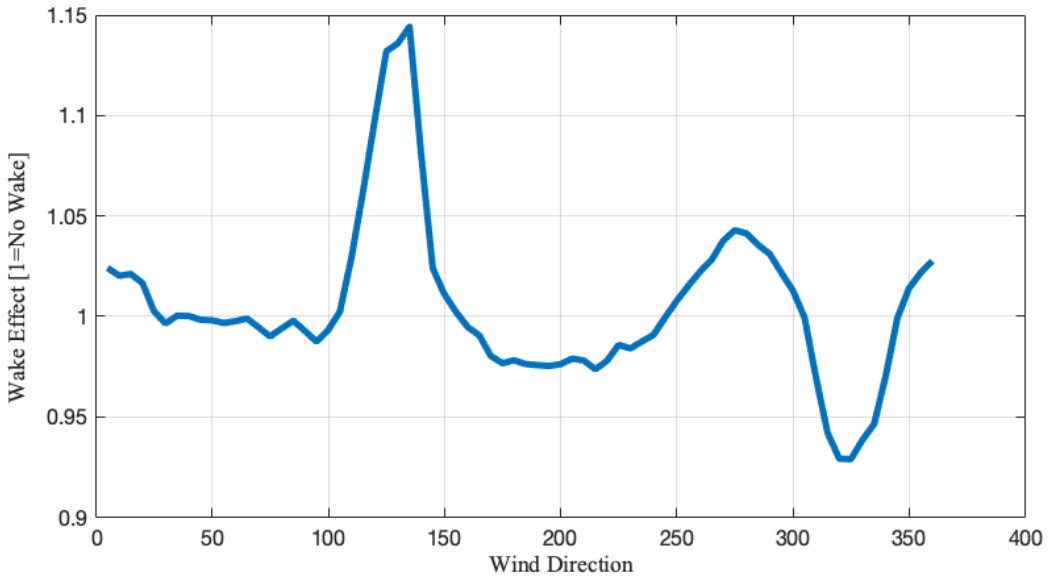

Figure: Ratio of 5-minute average wind speed at tse04 between two sonics in opposite boom directions at ~80 m AGL from April – June 2017. The sonic of interest (SE) observed tower wake effects of approximately 7% from the NW (~310-340º), and hence a decrease in velocity is observed. The spike seen in the SE direction is the wake observed by the NW sonic and is irrelevant for the current case.

2. Page 4: Given the randomized splitting between training and testing datasets, together with the absence of cross-validation, I am still concerned about potential overfitting, also considering the large autocorrelation your data have (due to the sum of that introduced by the overlapping averages and that naturally present in the data). While you state that random forests do not overfit the data at all, this is a debatable statement: I am sure you can find plenty of papers that can support both opinions. To make this reviewer happy, I would love to see whether the performance of the proposed model varies if the splitting between training and testing set is not performed randomly, but rather with a hybrid approach. For example, what happens if you keep all observations from one week for testing? Or from one full day from each week? Such tests would definitively give an answer to potential autocorrelation impacts on your results.

We appreciate this comment. In reviewing the cross-validation, we indeed discovered that the model was overfitting the data. We have made appropriate changes to the manuscript to reflect the new findings. Much of the results and discussion (which are now a single section to ease the explanation process) have been changed to reflect the changes in findings. The new findings reflect the results obtained over 10 testing sets partitioned via stratified k-fold cross validation. Similar to the reviewer's suggestion, this cross-validation technique splits the data nearly chronologically while ensuring the target variable distribution is consistent among the training and testing sets. We would like to thank the reviewer for his/her suggestion of performing cross validation, as it has led to many positive updates to the manuscript.

Minor Comments:

Figure 1: this map still looks somewhat incomplete to me: at the very least, please add some reference to understand the horizontal distances.
We have added both a reference for horizontal distances as well as a north arrow to make it easier for readers to orient themselves.

Page 4 L. 3: what do you mean by 'augmented' data here?
Our intention was to mention that the data has been quality controlled, and we have changed "augmented" to "quality controlled" to better reflect this (P4 L7).

Page 4 L. 3: "data were averaged into 10-minute, hourly, and three-hour segments at a 5-minute moving average in order to create a robust dataset" is still not clear to me. Do you mean that you are creating three datasets, both with one data point every 5-minute, but in dataset A all data are 10-minute average, in dataset B hourly averages, and in dataset C 3-hourly averages?
Yes, the reviewer is correct. We have changed the language of the sentence to "data were averaged over 10-minute, hourly, and three-hour segments at a 5-minute moving average in order to create three robust datasets, each consisting of over 28,000 samples" in order to relieve any potential confusion (P4 L7).

Figure 5: these scatterplots could be improved. Can you please change the color in the scatterplot based on density (e.g. https://matplotlib.org/api/_as_gem/matplotlib.pyplot.hist2d.html)?
We have removed this figure as it is no longer a useful indicator of model performance. We have replaced the R^2 metric with mean absolute error (MAE) as we believe it is a more telling metric.

Utilizing Physics-Based Input Features within a Machine Learning Model to Predict Wind Speed Forecasting Error

Daniel Vassallo[1], Raghavendra Krishnamurthy[2, 1], and Harindra J.S. Fernando[1]

[1]University of Notre Dame, Indiana, USA
[2]Pacific Northwest National Laboratory, Washington, USA

Correspondence: Daniel Vassallo (dvassall@nd.edu)

Abstract. Machine learning is quickly becoming a commonly used technique for wind speed and power forecasting. Many machine learning methods utilize exogenous variables as input features, but there remains the question of which atmospheric variables are most beneficial for forecasting, especially in handling non-linearities that lead to forecasting error. This question is addressed via creation of a hybrid model that utilizes an autoregressive integrated moving average (ARIMA) model to make an initial wind speed forecast followed by a random forest model that attempts to predict the ARIMA forecasting error using knowledge of exogenous atmospheric variables. Variables conveying information about atmospheric stability and turbulence as well as inertial forcing are found to be useful in dealing with non-linear error prediction. Streamwise wind speed, time of day, turbulence intensity, turbulent heat flux, vertical velocity, and wind direction are found to be particularly useful when used in unison for hourly and three-hour timescales. The prediction accuracy of the developed ARIMA-random forest hybrid model is compared to that of the persistence and bias-corrected ARIMA models. The ARIMA-random forest model is shown to improve upon the latter commonly employed modeling methods, reducing hourly forecasting error by up to 5% below that of the bias-corrected ARIMA model and achieving an R^2 value of 0.84 with true wind speed.

1 Introduction

Global wind power capacity reached almost 600 GW at the end of 2018 (GWEC, 2019), making wind energy a vital component of international electricity markets. Unfortunately, integrating wind power into an existing electrical grid is difficult because of wind resource intermittency and forecasting complexity. For utility companies employing wind power, it is important to estimate the aggregated load over a period of time to better balance grid resources, including long-term (1+ days ahead), short-term (1-3 hours ahead) and very-short term (15 minutes ahead) forecasts (Soman et al., 2010; Wu et al., 2012). Forecasting accuracy depends on site conditions, surrounding terrain, and local meteorology. Many wind farms are built in locations which are known to amplify winds due to surrounding terrain (such as Lake Turkana in Kenya, Tehachapi Pass in California etc.), requiring bespoke forecasts for accurate predictions. Numerical weather prediction models (NWPs) fail at such complex sites due to a lack of appropriate parameterization schemes suitable for local conditions (Akish et al., 2019; Bianco et al., 2019; Olson et al., 2019; Stiperski et al., 2019; Bodini et al., 2020). Therefore, statistical models and computational learning systems (such as an artificial neural network or random forest) are likely better suited to provide accurate power forecasts. Since wind

power production is heavily reliant upon environmental conditions, improvements in wind speed forecasting would allow for more reliable wind power forecasts.

If we simplify our wind speed prediction process down to its core (which has no true relation to atmospheric motions), we can imagine a system of atmospheric flow without external forcing. This would result in a constant streamwise wind speed U (i.e. $U_\tau = U_{\tau-1}$; U is streamwise wind speed, τ a timestep; this assumes discrete timesteps for simplicity). In this case, a persistence or autoregressive forecast would have zero forecasting error and uncertainty. However, uncertainty increases once we add an external force that we may represent by some variable x_1. Now future wind speed may be seen to be $U_\tau = f(U_{\tau-1}, x_{1,\tau-1})$. Assuming the external force is notable in strength and coupled with the inertia associated with winds, the previous autoregressive model will now struggle to predict U_τ because it does not take into account our external forcing $x_{1,\tau-1}$, resulting in an error ε (ε_τ is abbreviated to ε for simplicity). We can then break down our future wind speed into two parts: $U_\tau = \hat{U}_\tau + \varepsilon$ where $\hat{U}_\tau$ is our autoregressive forecast that is only dependent on $U_{\tau-1}$ (i.e. $\hat{U}_\tau = f(U_{\tau-1})$). The prediction error is thus skewed to represent the effects of the external force $x_{1,\tau-1}$ upon $U_{\tau-1}$.

If we continue to add external forces (x_1, x_2, ... x_n; n is the number of external forcing variables), our atmospheric system becomes much more complex and non-linear due to interactions between forcing mechanisms. We can again obtain our forecasting error as $\varepsilon = f(U_{\tau-1}, x_{1,\tau-1}, x_{2,\tau-1}, ... x_{n,\tau-1})$, which we can discretize as $\varepsilon = \mu_\varepsilon + \varepsilon'$ (μ_ε is the error bias, ε' the error fluctuations about μ_ε) given that we have a statistically significant sample size and the process is stationary. Squaring this equation and taking the average gives us the discretized equation for the mean squared error $\overline{\varepsilon^2} = \overline{\mu_\varepsilon^2} + \overline{\varepsilon'^2}$, with $\overline{\varepsilon'^2}$ representing the error variance and overlines denoting the average over all samples (Lange, 2005). $\overline{\mu_\varepsilon^2}$ represents the bias and may be removed via a simple bias-correction. The true concern is the error fluctuation term (ε') which constitutes the error variance. Assuming the external forcing variables (x's) are normally distributed, we can break down $\overline{\varepsilon'^2}$ into two constituents (Ku et al., 1966):

$$\overline{\varepsilon'^2} = \sigma_{x_j}^2 \left(\frac{\partial \varepsilon}{\partial x_j} \right)^2 + 2 \left[\sigma_{x_j, x_k} \frac{\partial \varepsilon}{\partial x_j} \frac{\partial \varepsilon}{\partial x_k} \right], \qquad j \neq k \tag{1}$$

where $\sigma_{x_j}^2$ is the variance of x_j and σ_{x_j, x_k} is the co-variance between x_j and x_k (subscript τ removed for simplicity). Unless external forcing (or its coupling with $U_{\tau-1}$) is minimal, the error is likely highly non-linear and chaotic (i.e. large $\overline{\varepsilon'^2}$). Therefore, it behooves us to discover which forcing mechanisms and atmospheric variables are the best predictors of individual fluctuations ε', which we will call "exogenous error".

Many studies that use machine learning (ML) techniques for wind speed or power forecasting utilize a handful of unadulterated atmospheric variables such as wind speed, pressure, and temperature as input features (Mohandes et al., 2004; Ramasamy et al., 2015; Lazarevska, 2018; Chen et al., 2019). Recently, a handful of investigations have begun to determine which variables may be most useful for these models. Vassallo et al. (2020) showed that invoking turbulence intensity (TI) can vastly improve vertical wind speed extrapolation accuracy. Similarly, Li et al. (2019) showed that TI improves wind speed forecasting on multiple timescales, while Optis and Perr-Sauer (2019) showed that both atmospheric stability and turbulence levels are important indicators for wind power forecasting. Markedly, it has been shown by Cadenas et al. (2016) that multivariate sta-

tistical models consistently outperform univariate models for wind speed forecasting. However, to the authors' knowledge, the question of which atmospheric variables are most useful in predicting exogenous error has not been addressed in the literature.

This investigation aims to determine if exogenous error may be, at least in part, predicted via a list of common meteorological measurements by following a methodology similar to that performed by Cadenas and Rivera (2010). The autoregressive integrated moving average (ARIMA) model first obtains an autoregressive forecast, and the forecasting error is extracted and bias-corrected. A random forest model is then utilized to discover patterns in the exogenous variables (and their relations to the endogenous variable U) that are predictive of exogenous error. The ARIMA-random forest hybrid model so constructed is referred to as the ARIMA-RF model.

This study is not intended to provide a catch-all list of input features that should or should not be used for every future study. Rather, it aims to inform future researchers and industry professionals as to what types of meteorological information must be used as ML inputs to predict the non-linear interactions between various atmospheric forces. Section 2 describes the Perdigão field campaign (the data source for the work), site characteristics, and instrumentation used for data collection. Section 3 provides an overview of the models utilized, testing process, and feature extraction/selection methodology. Section 4 provides testing results and includes a discussion of the obtained results. Finally, conclusions can be found in Section 5.

2 Site, Data, & Instrumentation

Data for this study were taken from the Perdigão campaign, a multinational project located in central Portugal that took place in the spring of 2017 (Fernando et al., 2019). The project site is characterized by two parallel ridges, both about 5 km in length with a 1.5 km wide valley between them. These ridges, which are represented by the elevated contours in Fig. 1, run northwest to southeast and rise about 250 m above the surrounding topography, making the site highly complex and increasing forecasting difficulty. The ridges will be referred to as the northern and southern ridge.

A variety of remote and *in situ* sensors were positioned in and around the valley to provide an accurate and thorough description of the surrounding flow field. Foremost among these sensors was a grid of meteorological towers which ran both parallel and normal to the ridges. One 100 m tower located on top of the northern ridge (white star in Fig. 1) is utilized in this study. This tower had sonic anemometers (20 Hz native measurement resolution) at 10, 20, 30, 40, 60, 80, and 100 m above ground level (AGL) as well as temperature sensors at 2, 10, 20, 40, 60, 80, and 100 m AGL. Information about tower data quality control, including corrections for boom orientation and tilt, may be found in NCAR/UCAR (2019). One of the towers in Perdigão was instrumented with sonic anemometers on both ends of the boom, allowing for an investigation into the effects of tower shadow. Minimal tower shadow effects were observed from the northwest ($\sim 310-340°$), with a maximum of only 7% flow deceleration. Wake effects were much smaller than those reported in previous studies (Moses and Daubek, 1961; Cermak and Horn, 1968; Orlando et al., 2011; McCaffrey et al., 2017; Lubitz and Michalak, 2018), which generally exceed 30%. We have therefore left the data unaltered. The tower data in the Perdigão database has been averaged into 5-minute increments by data managers at NCAR.

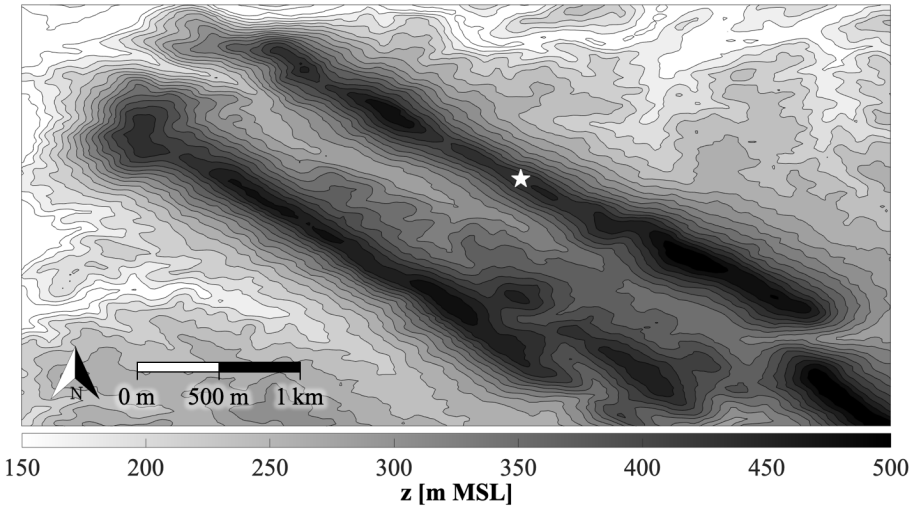

Figure 1. Contour plot of the campaign topography in meters above mean sea level (MSL). The white star represents the 100 m tower location on the northern ridge.

Sensors at 20 and 100 m AGL were chosen because of the high percentage ($> 99\%$ for all variables except temperature at 100 m AGL, which was available for $\sim 95\%$ of the periods) of clean data at these elevations. The utilized data spans three months, running from 10 March – 16 June 2017. Data at 100 m were correlated with that at 20 m, and missing data were filled using the variance ratio measure-correlate-predict method (Rogers et al., 2005). Any periods unavailable at both heights were filled using linear interpolation with Gaussian noise. All periods are required for proper functionality and assessment of the ARIMA model, and manually filled periods are not expected to make a noticeable difference in the findings.

The quality controlled data were averaged over 10-minute, hourly, and three-hour segments at a 5-minute moving average in order to create three robust datasets, each consisting of over 28,000 samples. These three datasets were split via stratified 10-fold cross validation (Diamantidis et al., 2000). The target streamwise wind speed, or that to be forecasted, is located at 100 m AGL. Squared buoyancy frequency (N^2), Richardson numbers (flux Ri_f and gradient Ri_g), and temperature gradient ($\partial T/\partial z$) were calculated between 20 – 100 m AGL. Friction velocity (u^*) was found at 20 m, just above surface roughness height (Fernando et al., 2019). All other input variables utilized were from 100 m AGL.

3 Methodology

This investigation utilizes two modeling methods, ARIMA and random forest regression, to create a hybrid model (ARIMA-RF) wherein the ARIMA model is first used to get a linear, univariate wind speed forecast. The ARIMA forecast is bias-corrected and the exogenous error ε' is then extracted and used as the target variable for the random forest. The random forest's goal (and the goal of the study) is to predict ε' (predictions denoted as $\hat{\varepsilon}'$) and determine which atmospheric variables and

forcing categories are useful for the predictive process. After the most important variables have been established, combinations of these input features are tested in an effort to determine whether specific groupings of input features may provide similar (or improved) forecasts compared to tests which incorporate all inputs. Finally, the ARIMA-RF results are compared with those of the persistence and bias-corrected ARIMA (hereafter referred to as the BCA) models. Section 3.1 details the ARIMA model, while Section 3.2 describes random forest regression. Sections 3.3 and 3.4 provide more detail on the feature extraction and selection methodology as well as the testing procedure.

3.1 ARIMA

ARIMA (Box et al., 2015) is a univariate statistical model used for time series forecasting. It is predicated on the combination of three functions: an autoregressive function that uses lagged values as inputs, a moving average function that uses past forecasting errors as inputs, and a differencing function used to make a time series stationary. In its simplest form, the next term in a time series sequence, y_τ, is given by

$$y_\tau = \sum_{i=1}^{p} \phi_i y_{\tau-i} + \sum_{j=1}^{q} \Theta_j \varepsilon_{\tau-j} + \varepsilon_\tau \tag{2}$$

where p and q are the orders of the autoregressive and moving average functions, respectively, ϕ_i and Θ_j the i^{th} autoregressive and j^{th} moving average parameters, respectively, $y_{\tau-i}$ the i^{th} lagged value, $\varepsilon_{\tau-j}$ the j^{th} past prediction error, and ε_τ the error term at time τ. The order of differencing is given by the parameter d and does not show up directly in Eqn. 2.

The dataset was tested for long-term statistical stationarity via the Augmented Dickey Fuller Test (Dickey and Fuller, 1979) using the statsmodels Python module (Seabold and Perktold, 2010). The test, to a statistically significant degree, proved that the wind speed data contains no embedded trends or drift (e.g. changes in the mean or variance of the wind speed due to long-term variability). Therefore, the differencing parameter d was set to 0 (This turns the ARIMA model into an ARMA model, but we stick with the term ARIMA for uniformity). The autoregressive and moving average parameters used, $p = 2$ and $q = 1$, were determined via minimization of the Akaike information criterion (Shibata, 1976) and empirical testing. Increasing parameters beyond this point did not lead to improved ARIMA accuracy.

3.2 Random Forest Regression

Random forest regression (Breiman, 2001) is an ensemble method that is made up of a population of decision trees. Bootstrap aggregation (bagging) is used so that each tree can randomly sample from the dataset with replacement, while only a random subset of the total feature set is given to each individual tree. The trees can be pruned (truncated) to add further diversification. After construction, the population's individual predictions are averaged to give a final prediction of the target variable. Ideally, this process results in a diversified and decorrelated set of trees whose predictive errors cancel out, producing a more robust final prediction.

An advantage of random forests is their ability to determine the importance of all input features for the predictive process. This is done by calculating the mean decrease impurity, or the decrease in variance that is achieved during a given split in each

decision tree. The decrease in impurity for each input feature can be averaged over the entire forest, providing an approximation of the feature's importance for the prediction (feature importance estimates sum to 100% to ease interpretability). To assist the random forest in representing the dynamic nature of atmospheric processes, input variables are taken from the previous two timesteps (i.e. input feature U comprises $U_{\tau-1}$ and $U_{\tau-2}$).

The constructed random forest model contains 1,000 trees for tests of individual variables and 1,500 trees for tests of variable combinations. To ease concerns of overfitting, each internal node was required have at least 100 samples in order to split (this truncation is a form of regularization). The random forest model was built using the scikit-learn Python library (Pedregosa et al., 2011).

3.3 Feature Extraction and Selection

In an effort to ensure that the findings are applicable to real-world campaigns, we limit our sources of information to those which may be measured by a typical meteorological mast containing sonic anemometers alongside temperature sensors. Using this information, we can write our future wind speed U_τ as a function of the following variables, which were broken down into their mean and fluctuating values:

$$U_\tau = f\left(U_i, \theta_i, W_i, T_i, t_i, u_i', \theta_i', w_i', T_i'\right) \tag{3}$$

where U_i and θ_i are the mean streamwise wind speed and direction, respectively, W_i the mean vertical wind speed, T_i the mean temperature, t_i the time of day, u_i' the fluctuating horizontal velocity, θ_i' the fluctuating wind direction, w_i' the fluctuating vertical velocity, and T_i' the fluctuating temperature at each previous timestep i. Unfortunately, θ' was not available within the dataset utilized (which had already been 5-minute averaged) and is therefore ignored for this study. Previous analysis, however, has shown that θ' varies inversely with U in complex terrain (Papadopoulos et al., 1992), and we may therefore assume its influence is largely captured by U.

Although these unadulterated features give us an idea as to how the system is working at the moment, they may not explicitly represent the relevant atmospheric forcing mechanisms. Our list of measurements allows us to break down our system into two principal forcing components: buoyancy and inertial forcing (which indirectly includes pressure gradient forces). Each of these forces can be further discretized into large and small scales (also called mean and fluctuating values; typically separated by at least one order of magnitude).

Fig. 2 shows an illustrative breakdown of the two main forcing mechanisms alongside a list of extracted descriptor variables. The definitions and formulations of all non-obvious extracted variables used in this study can be found in Appendix A. From this figure, it is clear that the variables in Eqn. 3, when manipulated, are able to describe both the inertial and buoyant forces at multiple scales. Large-scale inertial forcing can be described by the local mean wind speed (U) and direction (θ, broken down into North-South and East-West components in an attempt to eliminate any discontinuities) or vertical velocity W, while small-scale inertial forcing can be described by variables such as the fluctuating (standard deviation of) velocity σ_u, friction velocity u^*, and the turbulence kinetic energy TKE. Likewise, large-scale buoyancy forcing can be described by the squared buoyancy

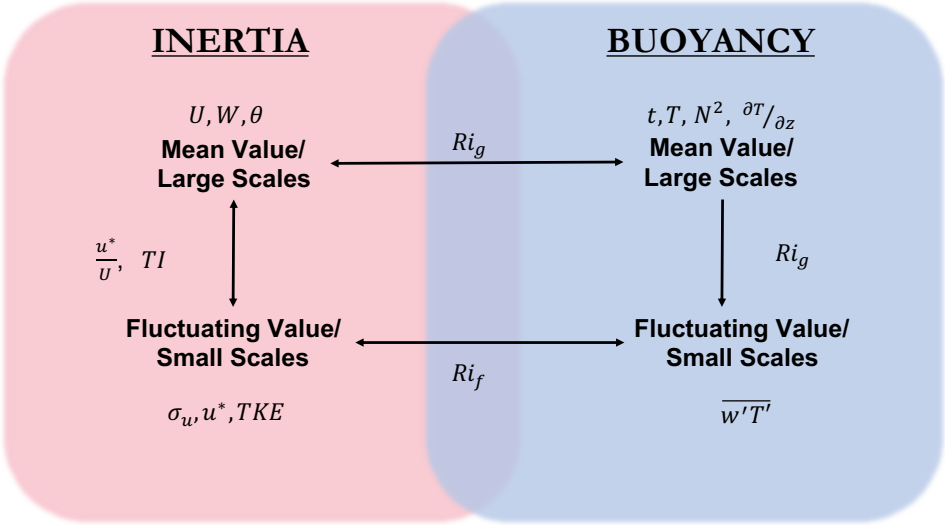

Figure 2. Illustrative breakdown of the scales and variables related to inertial and buoyant forcing. θ' is not shown as it is not utilized in the analysis.

frequency N^2, the temperature gradient ${}^{\partial T}/_{\partial z}$, or by proxy values such as the time of day t (broken down into sine and cosine components, one of which relates to 0000–1200 local time and the other to 0600–1800 local time) or temperature T (which, on average, is higher during the day and lower at night; stability parameters based on Monin-Obukhov similarity theory have been considered ill-suited for complex terrain flows because of the breakdown of underlying assumptions (Fernando et al., 2015), and hence were not used in this study). Small-scale buoyancy effects can be described by the turbulent heat flux $\overline{w'T'}$. The correspondence between forces and internal parameters can also be described by non-dimensional variables such as the gradient Richardson number Ri_g, flux Richardson number Ri_f, turbulence intensity TI, and normalized friction velocity ${}^{u^*}/_U$. These derived non-dimensional variables, or extracted features, are typically ignored by current ML models in lieu of raw features such as those listed in Eqn. 3.

Extracted variables like those in Fig. 2 may not provide any more information than the raw variables in Eqn. 3. However, they may ease the burden on the model by discretizing (or directly relating) informational categories, therefore reducing informational overlap and noise, providing more periodic patterns, and more accurately describing the underlying system. Further, such well-conceived meteorological variables have been shown to be useful for atmospheric prediction (Kronebach, 1964; Li et al., 2019; Bodini et al., 2020). In theory, given enough data, the model should be able to decipher and interpret these extracted features on its own. Unfortunately there often isn't enough collected data for this to happen organically. Instead, by providing better information we can create a simpler, cheaper, more robust model that requires less training data and construction time. Selected features will ideally represent the underlying system as accurately as possible without providing noisy or redundant information.

3.4 Testing

Initial tests utilize a full feature set (i.e. all input variables are included). Feature importance estimates are then extracted from the random forest model and various user-selected combinations of the most important input features are tested. It must be noted that only select input feature sets were tested in this investigation due to the sheer multitude of potential feature sets.

In order to relieve any timescale bias, forecasts are made across multiple timescales. Typically, wind power utility operators require single-step short range power forecasts run hour-by-hour for a few days to reduce unit commitment costs. The forecast skill of observation-based methods generally reduces with forecast lead time within an hour, and numerical models have higher skill in forecasting larger lead times (> 3 hours; Haupt et al. (2014)). Statistical learning methods have proved to be particularly effective from about 30 minutes to approximately three hours ahead (Mellit, 2008; Wang et al., 2012; Yang et al., 2012; Morf, 2014), and roughly this time frame is thus the focus for this study. The shortest forecast predicts wind speeds 10 minutes ahead, roughly within the turbulent spectral band (Van der Hoven, 1957). Forecasts are also made one and three hours ahead, which are within the spectral gap between the turbulent and synoptic spectra and approach the six-hour period wherein NWP models become particularly useful (Dupré et al., 2019). These are all single-step forecasts, which is to say that the averaging timescale increases with the forecasting timescale (e.g. a 10-minute forecast predicts 10-minute averaged wind speed, whereas a three-hour forecast predicts three-hour averaged wind speed). Each dataset is split via stratified k-fold cross validation (Diamantidis et al., 2000), a technique that splits the dataset into k sections (in this case, we set $k = 10$) and uses each section as a separate test set (i.e. tests consist of 10 runs, each of which utilizes a unique test set). This technique splits data nearly chronologically, ensuring that the model does not overfit the dataset. Forecasting metrics for each test are obtained by averaging the ensemble of all 10 runs.

The root mean squared error (RMSE) and mean absolute error (MAE) of the BCA are found for each timescale, giving two metrics of the true exogenous error ε'. The random forest model is then trained to predict ε', combined with the ARIMA model, and the newly constructed ARIMA-RF is used to forecast wind speeds. The reduction in RMSE and MAE (which come exclusively from the random forest's prediction of exogenous error $\hat{\varepsilon}'$) is then found for the test set. Eqn. 4 and Eqn. 5 describe both metrics, wherein U_m is the target wind speed, $\hat{U}_m$ the predicted wind speed, m each individual sample, and M the sample size.

$$RMSE = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(U_m - \hat{U}_m)^2} \tag{4}$$

$$MAE = \frac{1}{M}\sum_{m=1}^{M}\left|U_m - \hat{U}_m\right| \tag{5}$$

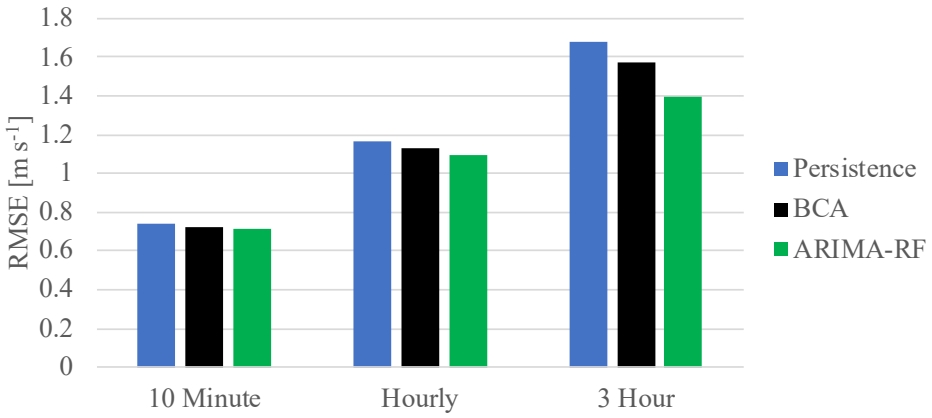

Figure 3. Comparison of RMSE obtained by the persistence, BCA, and ARIMA-RF with the full input feature set for all forecasting timescales.

4 Results and Discussion

The ARIMA-RF model utilizing the full feature set obtained lower RMSE than the BCA for all timescales. Fig. 3 shows a comparison between the RMSE obtained by the ARIMA-RF model and that obtained by the persistence and BCA models. The BCA's RMSE amounted to 0.726, 1.132, and 1.575 m s^{-1} for the 10-minute, hourly, and three-hour forecasts, respectively. The ARIMA-RF, utilizing all input features, reduced these RMSE values (as well as the MAE values) by ∼2%, 4%, and 11%, respectively (RMSE and MAE values given in Table 1). The random forest is clearly able to ascertain more prudent physical patterns at larger timescales (up to three hours), as large-scale atmospheric dynamics provide a more predictable signal for the prediction of exogenous error.

All feature importance values were extracted from the random forest and are shown in Fig. 4. The variables are broken down into three distinct categories: inertial (large-scale dimensional variables signifying inertial forces in Fig. 2), stability (blue and purple regions in Fig. 2 which are akin to atmospheric stability), and turbulence variables (small scale and non-dimensional inertial variables in Fig. 2). θ is the most important variable for the prediction of ε' at all timescales, achieving up to 20% importance at the three-hour timescale. Fig. 5a shows the partial dependence on the East-West component of θ (i.e. the random forest model's average predictions, $\hat{\varepsilon}'$, across the range of a given variable, in this case the East-West component of θ) alongside the variable's distribution. The model is clearly able to discern an East-West directional pattern in the training data. The climatology above the Perdigão ridges displays a proclivity for northeasterly and southwesterly flows, the former of which exhibit comparatively high velocities (Fernando et al., 2019). The BCA tends to under-predict flow from the northeast (i.e. the random forest's average target variable $\overline{\varepsilon'}$ is positive; Fig. B1, Appendix B). Accordingly, the random forest predicts positive ε' values, correcting for the BCA's under-prediction. Fig. 6c, which displays both the ARIMA-RF (solid lines) and BCA (dashed lines) RMSE values by directional sector, shows that the ARIMA-RF successfully improves hourly and three-hour forecasting accuracy when winds are northeasterly.

Table 1. RMSE and MAE (m s^{-1}) obtained by the persistence, BCA, and ARIMA-RF models when utilizing select input feature combinations. The final two rows show the testing results when utilizing the full input feature set and the full feature set except T. Underlined values show the best performance from each column.

	10 Minute		Hourly		3 Hour	
Model	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
Persistence	0.737	0.531	1.165	0.884	1.676	1.315
BCA	0.726	0.528	1.132	0.863	1.575	1.240
Input Features	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
U, θ, t	0.733	0.533	1.109	0.834	1.533	1.194
U, θ, t, T	0.752	0.549	1.164	0.876	1.597	1.231
U, θ, t, W	0.729	0.531	1.095	0.825	1.518	1.185
U, θ, t, TI	0.730	0.531	1.072	0.812	1.533	1.191
$U, \theta, t, \overline{w'T'}$	0.731	0.532	1.095	0.825	1.521	1.189
$U, \theta, t, W, TI, \overline{w'T'}$	0.728	0.530	1.073	0.812	1.521	1.180
$U, \theta, t, W, TI, \overline{w'T'}, T$	0.738	0.539	1.115	0.842	1.565	1.215
Full input feature set	0.714	0.520	1.092	0.830	1.395	1.100
All features except T	0.712	0.518	1.071	0.813	1.379	1.083

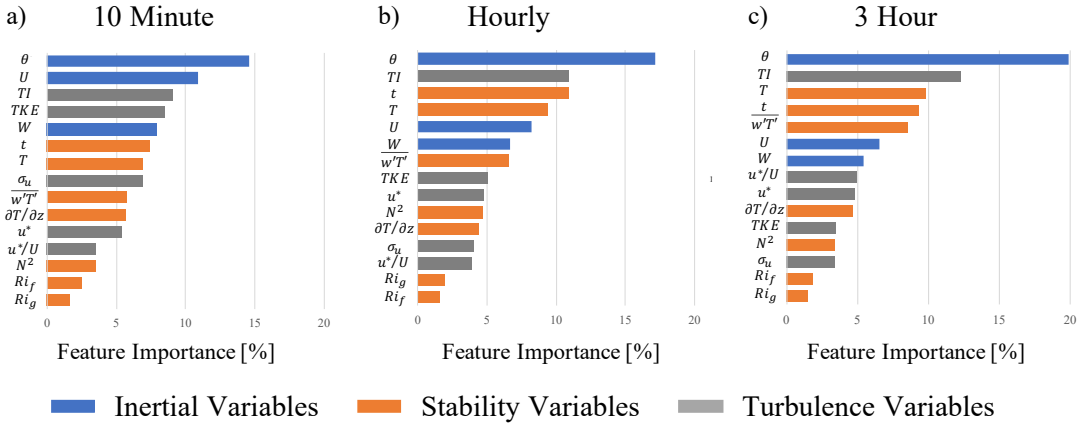

Figure 4. Feature importance for the prediction of exogenous error when all input features are given to the random forest model. a) shows importance for the 10-minute prediction, b) for the hourly prediction, and c) for the three-hour prediction. Blue bars denote inertial variables, orange denote stability variables, and grey bars denote turbulence variables. Importance values for each test sum to 100%.

Even larger improvement is seen for westerly winds which pass over the southern ridge prior to reaching the measurement location (Fig. 6c). Westerly winds are common between 1300 – 2100 local time (Fernando et al., 2019), and Fig. 6b shows

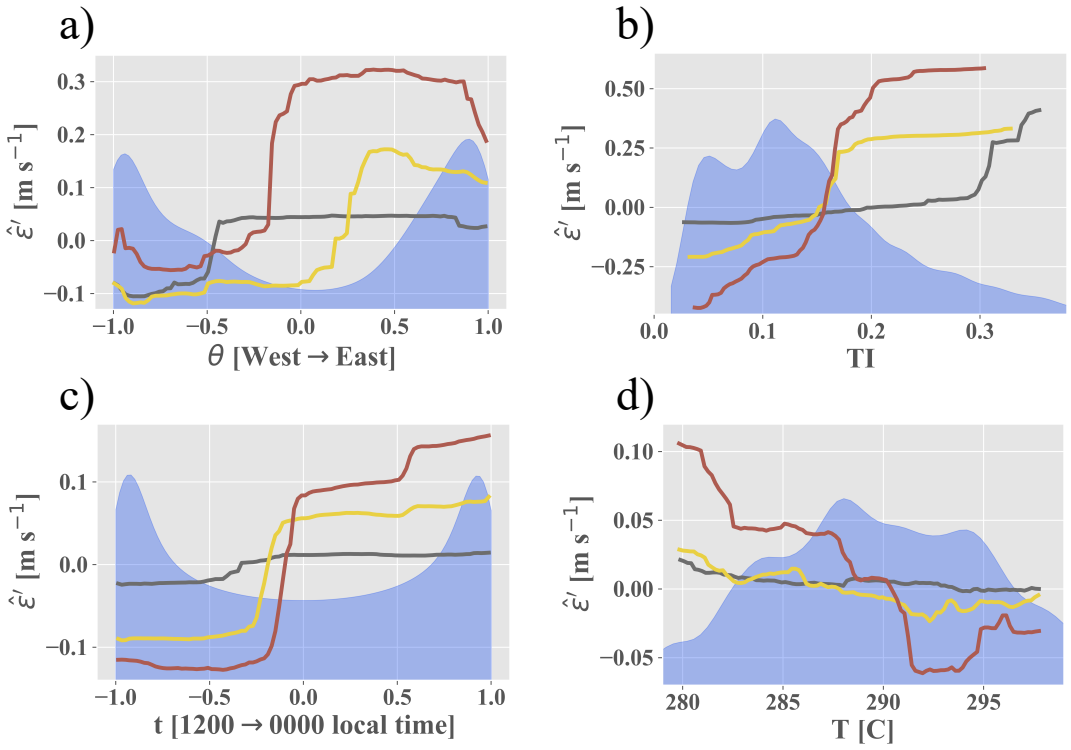

Figure 5. Lines show dependence between the random forest prediction $\hat{\varepsilon}'$ and (a) the East-West component of θ, (b) TI, (c) the noon-midnight component of t, and (d) T. Blue shading shows variable distribution. Arrows in (a) and (c) correspond to direction (on the x-axis) of the normalized oscillatory component.

that the ARIMA-RF is able to improve upon the BCA forecast during these hours. The BCA tends to over-predict wind speeds around 1200 local time (i.e. negative $\overline{\varepsilon'}$ in Fig. B2, Appendix B), as wind speeds reach a relative nadir (Fernando et al., 2019). The model's over-prediction is captured and partially corrected by the random forest, which predicts negative ε' values around noon ($t \approx -1$ in Fig. 5c). Wind speeds then pick up throughout the afternoon as the atmosphere becomes more convective. Increased convection leads to high TKE and TI values, which peak in the mid-afternoon (not shown). As wind speeds rise and the atmosphere becomes more convective, the BCA begins to under-predict wind speeds. The under-prediction is once again captured by the random forest and the artifacts can be seen in Fig. 5b and c. The random forest identifies periods more than five hours from noon ($t \geq -0.25$ in Fig. 5c) and those with high TI as periods wherein the BCA will likely under-predict wind speeds and corrects the BCA forecast accordingly by predicting positive ε' values.

A comparison of both the BCA and ARIMA-RF models' RMSE values in stable and unstable conditions (Table 2) shows that the BCA performs better under unstable conditions for both the hourly and three-hour forecasts, but the opposite is true for the 10-minute timescale. Increased turbulence during the daytime clearly hampers the BCA when forecasting 10 minutes and,

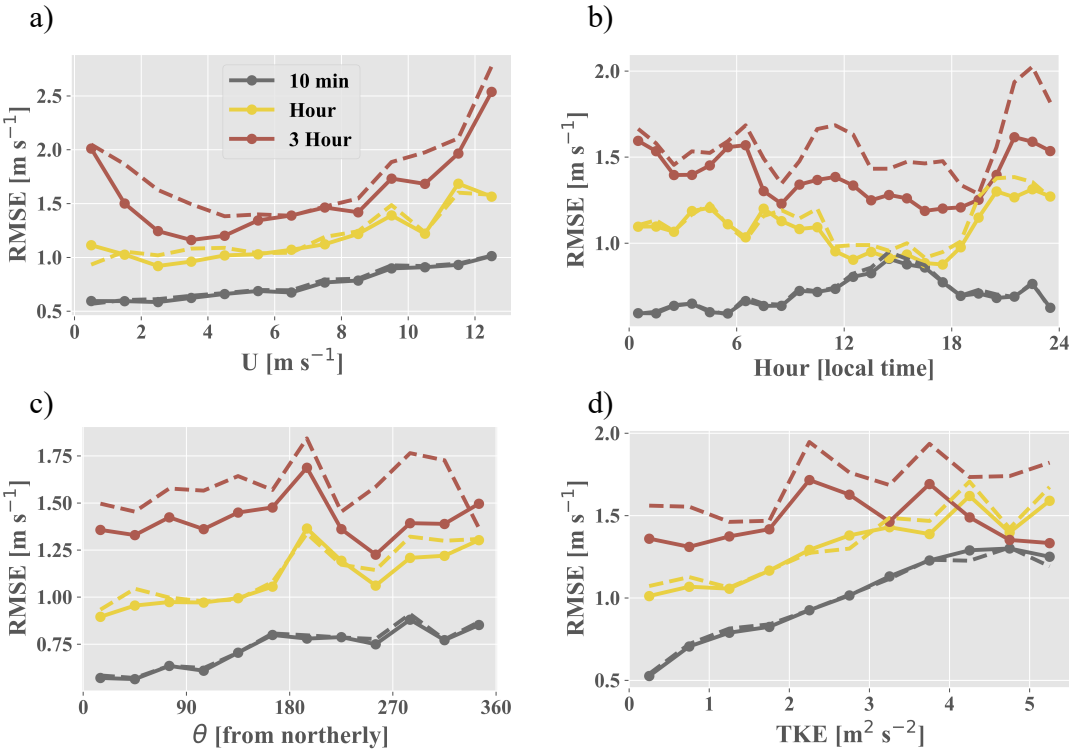

Figure 6. RMSE obtained by the ARIMA-RF (with the full feature set; solid lines with points) and BCA (dashed lines) partitioned by (a) U, (b) hour of the day (local time), (c) θ, and (d) TKE.

to a lesser extent, one hour ahead (dashed grey and yellow lines, respectively, in Fig. 6d). Notably, the random forest is only able to make minimal forecasting improvements ($\sim 1.5\%$) on the 10-minute and hourly timescales during unstable conditions, but is able to improve the three-hour forecast by almost 15% during such conditions. Wind speeds can be highly chaotic during convective conditions, leading to large fluctuations as high-energy eddies pass through the measurement location. Typically the large-eddy turnover timescale for the lower atmosphere is 10-20 minutes (specifically during daytime), and averaging timescales approaching or less than this timescale exclude information on more stable and deterministic large eddies, thus making predictions more prone to random errors. The lack of large eddy influence results in a wind speed signal that is replete with random fluctuations originating in the inertial subrange, adding substantial noise to the prediction. These fluctuations may overwhelm the ML model's pattern recognition capabilities, even up to the hourly timescale, reducing the random forest prediction to a noisy guess. Such ML models will always make predictions based on patterns in the training data, even when those patterns are erroneous and do not hold for the testing dataset. This results in error predictions that are only minimally correlated with the true exogenous error.

The highest RMSE values produced by the hourly and three-hour BCA occur during the evening transition period (Fig. 6b; sunset typically between 2000 – 2100 local time). There is a drastic reduction in both wind speed and atmospheric TKE during this period (Fernando et al., 2019) as the atmosphere transitions from a convective to a stable regime. Wind ramps (defined as wind speed changes of 20% and 50% for hourly and three-hour forecasts, respectively) are particularly prevalent between 1900 – 2300 local time (not shown). Such ramp events are difficult for a simple statistical model such as ARIMA to predict as they are not only highly situational, they are also statistical outliers. The random forest is able to partially discern such transitional events occurring between 1900 – 2300, reducing RMSE by an average of 1%, 6%, and 16% for the 10-minute, hourly, and three-hour timescales, respectively.

Table 2. RMSE (m s^{-1}) obtained by the BCA and ARIMA-RF (with the full input feature set) based on stability (as defined by N^2) of the forecasted time period.

	Stable		Unstable	
Timescale	**BCA**	**ARIMA-RF**	**BCA**	**ARIMA-RF**
10 minutes	0.711	0.700	0.770	0.758
1 Hour	1.166	1.118	1.017	1.003
3 Hours	1.590	1.418	1.526	1.299

The seven most important features for the hourly and three-hour predictions, namely θ, U, TI, t, T, W, and $\overline{w'T'}$, are identical (although scrambled; Fig. 4), and were therefore used to test discriminate feature set combinations. All tests with multiple input features contained U, θ, and t. There are two reasons for prioritizing these three variables: they prove to be some of the most important input features for all timescales and they can all be captured by a simple cup anemometer and wind vane rather than a more expensive sonic anemometer. These three features, when used in conjunction, were able to achieve about 58% of the error reduction obtained by the test incorporating all features at the hourly timescale (Table 1). Although T appears to be one of the most important input features (Fig. 4), it clearly hinders the model's predictive capabilities and decreases prediction accuracy across all timescales. The case with an input set of U, θ, t, and T consistently performs the worst of all cases shown in Table 1. Simply adding T to the base input feature set (U, θ, and t) decreases forecasting accuracy by up to 5%, whereas removing T from the full input feature case improves prediction accuracy at all timescales. T is highly seasonal and, because stratified k-fold cross validation splits the training and testing sets nearly chronologically (the distribution of the target variable ε' is kept constant between training and testing sets), the discrepancy between mean T values can be as high as 10° C between the training and testing set. The disparity between the training and testing distributions clearly hampers the random forest's predictive capabilities by providing training information that is nugatory or deleterious for prediction on the testing set. As can be seen in Fig. 5d, the random forest appears to be somewhat dependent upon T, particularly on the three-hour timescale, as low T leads to positive $\hat{\varepsilon}'$ and high T leads to negative $\hat{\varepsilon}'$. T is a clear example of the inherent risks associated with utilizing dimensional or seasonal inputs within an ML forecasting model, although such issues may be negated for a dataset spanning several years.

The discriminate input set incorporating only U, θ, t, W, TI, and $\overline{w'T'}$ produces an hourly forecast that is nearly equivalent to that incorporating all features except T (Table 1). W, TI, and $\overline{w'T'}$ all improve forecast accuracy, particularly at the hourly timescale. The 10-minute and three-hour models, however, appear to derive a majority of their forecasting skill from the entire array of input features rather than the discriminate list tested. Notably, many of the most important input features (U, θ, t, and W) are directly measurable and need not be extracted (although W cannot be captured by a cup anemometer). The most important variables that require extraction (i.e. values that are not direct measurements), TI and $\overline{w'T'}$, both contain small-scale (fluctuating) forcing components, indicating that small-scale processes may be more easily captured by ML models after domain-specific interpretation. These small-scale variables provide significant forecasting improvements, even at a multi-hour timescale. The testing results from the study show that, in order to achieve an optimal forecast of exogenous error, information about these small scales must be included as inputs for the predictive model.

The results of the discriminate tests, which may be found in Table 1, show a stark distinction between the 10-minute and the hourly/three-hour forecasts. None of the 10-minute tests except that with the full input feature set (and all features except T) were able to improve upon the BCA forecast. In contrast, all hourly and three-hour tests except those utilizing U, θ, t, and T were able to improve upon the BCA forecast. The disparity in the findings likely reflects the inherent challenges associated with forecasting wind speeds within the turbulent peak of the wind speed spectrum (Van der Hoven, 1957). 10-minute forecasts are more prone to turbulent fluctuations induced by eddies in the inertial subrange. Hourly and three-hour forecasts, however, incorporate information from more stable large-scale eddies, allowing for a more predictable meteorological pattern.

A majority (if not all) of the random forest's predictive capability derives from the utilization of multiple atmospheric variables within the input feature set. Table B1 in Appendix B shows that t is the only input feature that, when used individually, leads to a decrease in RMSE below that of the BCA. Individual atmospheric variables effectively represent the magnitude of the first term on the right side of Eqn. 1, $\sigma_{x_j}^2(\frac{\partial \varepsilon}{\partial x_j})^2$. The random forest model is more powerful when utilizing multiple atmospheric variables within the input feature set because the model can incorporate the second term on the right side of Eqn. 1 $\left(2[\sigma_{x_j,x_k}\frac{\partial \varepsilon}{\partial x_j}\frac{\partial \varepsilon}{\partial x_k}]\right)$, an indication of how the exogenous error changes depending on the input features' co-variance. This is especially true for the testing case incorporating all input features except T, which typically provides the most accurate predictions of exogenous error. The ARIMA-RF's improvement over the BCA forecast increases with increasing timescales, providing more than 11% improvement at the three-hour timescale. The ARIMA-RF hourly forecast (with the full input feature set) obtains an R^2 value of 0.84 with true wind speed, akin to that of numerical models in complex terrain (Yang et al., 2013). This study shows that the forecasting improvement, which comes from prediction of non-linear exogenous error ε', can be directly attributed to prudent feature engineering.

5 Conclusions

Exogenous error (ε') arises from atmospheric forcing that is ignored or misrepresented in the modeling process. It has been shown that this error, or a portion thereof, can be predicted by an ML model given relevant atmospheric information. U, θ, t, W, TI, and $\overline{w'T'}$ were the most important input features, whereas T provided information that was particularly deleterious.

Domain-specific feature extraction was found to be particularly useful for input features relating small-scale forcing, and these turbulence variables were found to reduce forecasting error even for multi-hour forecasts. Predictions of ε' were shown to be particularly dependent upon TI, but feature dependence patterns tend to be relatively uniform across timescales. Atmospheric stability and turbulence appear to play a large role in the model's ability to predict ε', as the site's specific climatology is shown to produce many of the patterns captured by the random forest. Finally, it is shown that utilizing multiple atmospheric variables which relate various forcing mechanisms and scales is necessary in order to forecast ε'.

While the exact results of this investigation are site-specific, the findings are expected to be generally applicable to numerous wind projects, especially those located in complex terrain. Prudent implementation of atmospheric forcing information, particularly that which is non-linear or derived via coupling of multiple forces, is crucial for the prediction of exogenous error and must be addressed to obtain optimal forecasting results. This study supports the supposition that a hybrid model using ML techniques to correct a simpler statistical predictor (such as an ARIMA model) can be effective for wind speed forecasting.

Further improvements are still required to more accurately represent atmospheric forcing. Gridded meso or synoptic-scale information would allow the model to predict transitional periods including weather fronts and drastic wind ramp events. Multiple scales of forcing should also be incorporated to improve the pattern recognition capabilities of ML techniques. Additional information about microscale, mesoscale, and synoptic events would better depict atmospheric forcing and momentum, and the effects of seasonality must be accounted for when possible. Hopefully, this study will be a forerunner for the improved incorporation of atmospheric physics within ML modeling.

Code and data availability. Data from the Perdigão campaign may be found at https://perdigao.fe.up.pt/. Due to the multiplicity of cases analyzed in this study, example processing and modeling codes can be found at https://github.com/dvassall/.

Appendix A: Input Features

Atmospheric variables were measured using sonic anemometers and temperature sensors along a single 100 m tower. When possible, missing data from the 100 m sensors were filled via correlation with the 20 m sensors using the variance ratio measure-correlate-predict method (Rogers et al., 2005). There were no periods with functional 100 m sensors and nonfunctional 20 m sensors. All periods without any measurements from both sets of sensors (15 5-minute periods) were filled using linear regression with Gaussian white noise. Many of the input features used in the study required derivation. A description of necessary derivations are given below.

Friction velocity is defined as $u^* = \left(\overline{u'w'}^2 + \overline{v'w'}^2\right)^{1/4}$ and was measured at 20 m AGL, just above canopy height (Fernando et al., 2019). Turbulence kinetic energy is defined as $TKE = \frac{\overline{u'^2}+\overline{v'^2}+\overline{w'^2}}{2}$ and was measured at 100 m AGL. Buoyancy frequency squared is typically defined as (see Kaimal and Finnigan (1994) for details of all parameters that appear below)

$$N^2 = \frac{g}{\rho_0}\frac{\partial \rho}{\partial z} = \frac{g}{T_{pv0}}\frac{\partial T_{pv}}{\partial z} \tag{A1}$$

where g is the gravitational force, ρ the air density, z the height AGL, T_{pv} the virtual potential temperature, and subscript 0 indicates reference variables in using the Boussinesq approximation. The gradient Richardson number is defined as

$$Ri_g = \frac{N^2}{\left(\frac{\partial u}{\partial z}\right)^2 + \left(\frac{\partial v}{\partial z}\right)^2} \tag{A2}$$

where u and v are the two horizontal wind speed components. The flux Richardson number is defined as

$$Ri_f = \frac{\frac{g}{T_v}\overline{w'T'}}{\overline{u'w'}\left(\frac{\partial u}{\partial z}\right) + \overline{v'w'}\left(\frac{\partial v}{\partial z}\right)}, \tag{A3}$$

where T_v is the virtual temperature while $\overline{u'w'}$ and $\overline{v'w'}$ (both measured at 100 m AGL alongside $\overline{w'T'}$ and T_v) are the Reynolds stresses that indicate the flow's vertical momentum flux. Ri_f is typically used in conjunction with a stably stratified atmosphere (Lozovatsky and Fernando, 2013). It is used here in the general sense as it is a measure of the ratio between buoyant energy production and mechanical energy production (associated with inertial forces) related to Fig. 2. Negative N^2 values, corresponding to convective atmospheric conditions, are set to 0. Ri_g and Ri_f are limited to a maximum of 5 and minimum values of 0 and -5, respectively, to remove extremes in both variables. Turbulence intensity is the ratio of fluctuating to mean wind speed, or $TI = {}^{\sigma_u}/_{U}$. Both hour of the day and wind direction were broken into two oscillating components in order to eliminate any temporal or directional discontinuity.

Appendix B: Testing Results & Analysis

Figs. B1 and B2 show the average exogenous error $\overline{\varepsilon'}$ produced by the bias-corrected ARIMA (BCA) as partitioned by direction and hour, respectively. Table B1 presents the RMSE and MAE obtained by the BCA (total exogenous error) and the ARIMA-RF using individual input features.

Author contributions. Daniel Vassallo prepared the manuscript with the help of all co-authors. Data processing was performed by Daniel Vassallo, with technical assistance from Raghavendra Krishnamurthy. All authors worked equally in the manuscript review process.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work was funded by the National Science Grant numbers AGS-1565535 and AGS-1921554, Wayne and Diana Murdy Endowment at University of Notre Dame and Dean's Graduate Fellowship for Daniel Vassallo. The Pacific Northwest National Laboratory is operated for the DOE by Battelle Memorial Institute under Contract DE-AC05-76RLO1830. Special thanks to the teams at both EOL/NCAR and DTU who collected and managed the tower data utilized in this study.

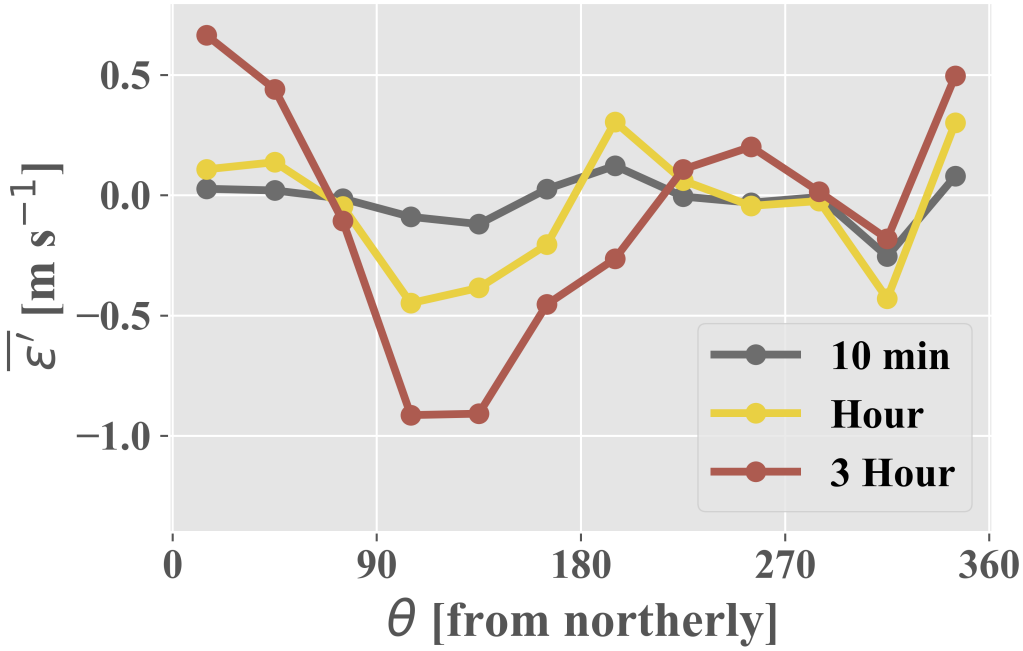

Figure B1. Average exogenous error ($\overline{\varepsilon'}$) produced by the BCA partitioned by direction.

References

Akish, E., Bianco, L., Djalalova, I. V., Wilczak, J. M., Olson, J. B., Freedman, J., Finley, C., and Cline, J.: Measuring the impact of additional instrumentation on the skill of numerical weather prediction models at forecasting wind ramp events during the first Wind Forecast Improvement Project (WFIP), Wind Energy, 2019.

Bianco, L., Djalalova, I. V., Wilczak, J. M., Olson, J. B., Kenyon, J. S., Choukulkar, A., Berg, L. K., Fernando, H. J., Grimit, E. P., Krishnamurthy, R., et al.: Impact of model improvements on 80 m wind speeds during the second Wind Forecast Improvement Project (WFIP2), Geoscientific Model Development (Online), 12, 2019.

Bodini, N., Lundquist, J. K., and Optis, M.: Can machine learning improve the model representation of turbulent kinetic energy dissipation rate in the boundary layer for complex terrain?, Geoscientific Model Development, 13, 4271–4285, 2020.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M.: Time series analysis: forecasting and control, John Wiley & Sons, 2015.

Breiman, L.: Random forests, Machine learning, 45, 5–32, 2001.

Cadenas, E. and Rivera, W.: Wind speed forecasting in three different regions of Mexico, using a hybrid ARIMA–ANN model, Renewable Energy, 35, 2732–2738, 2010.

Cadenas, E., Rivera, W., Campos-Amezcua, R., and Heard, C.: Wind speed prediction using a univariate ARIMA model and a multivariate NARX model, Energies, 9, 109, 2016.

Cermak, J. and Horn, J.: Tower shadow effect, Journal of Geophysical Research, 73, 1869–1876, 1968.

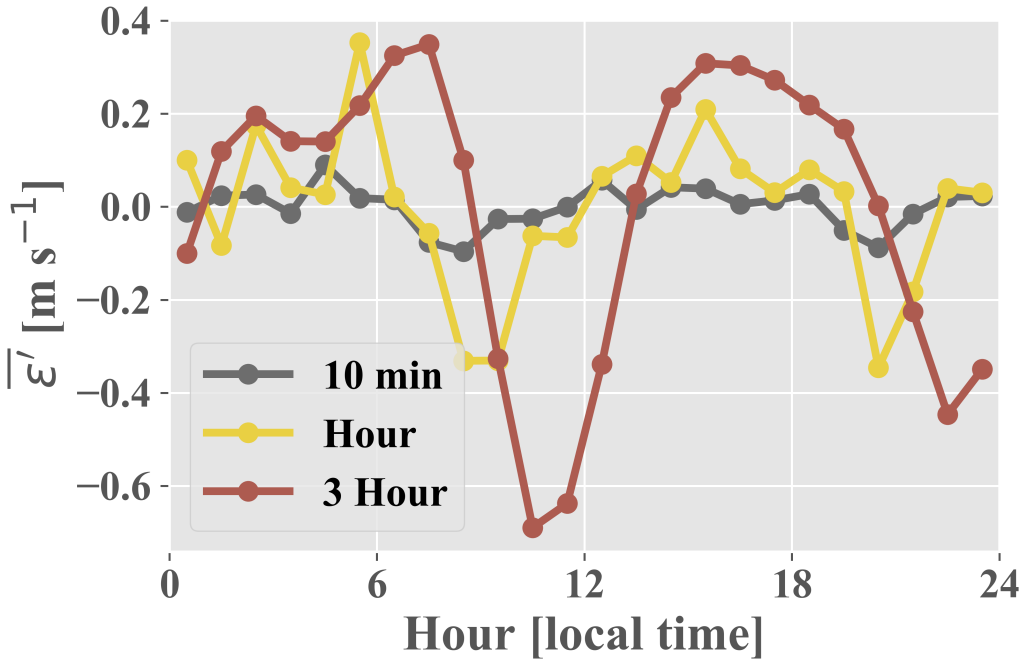

Figure B2. Average exogenous error ($\overline{\varepsilon'}$) produced by the BCA partitioned by hour of the day.

Chen, Y., Zhang, S., Zhang, W., Peng, J., and Cai, Y.: Multifactor spatio-temporal correlation model based on a combination of convolutional neural network and long short-term memory neural network for wind speed forecasting, Energy Conversion and Management, 185, 783–799, 2019.

Diamantidis, N., Karlis, D., and Giakoumakis, E. A.: Unsupervised stratification of cross-validation for accuracy estimation, Artificial Intelligence, 116, 1–16, 2000.

Dickey, D. A. and Fuller, W. A.: Distribution of the estimators for autoregressive time series with a unit root, Journal of the American statistical association, 74, 427–431, 1979.

Dupré, A., Drobinski, P., Alonzo, B., Badosa, J., Briard, C., and Plougonven, R.: Sub-hourly forecasting of wind speed and wind energy, Renewable Energy, 2019.

Fernando, H., Pardyjak, E., Di Sabatino, S., Chow, F., De Wekker, S., Hoch, S., Hacker, J., Pace, J., Pratt, T., Pu, Z., et al.: The MATERHORN: Unraveling the intricacies of mountain weather, Bulletin of the American Meteorological Society, 96, 1945–1967, 2015.

Fernando, H., Mann, J., Palma, J., Lundquist, J., Barthelmie, R. J., Belo-Pereira, M., Brown, W., Chow, F., Gerz, T., Hocut, C., et al.: The Perdigao: Peering into microscale details of mountain winds, Bulletin of the American Meteorological Society, 100, 799–819, 2019.

GWEC: Global Wind Report 2018, https://gwec.net/wp-content/uploads/2019/04/GWEC-Global-Wind-Report-2018.pdf, 2019.

Haupt, S. E., Mahoney, W. P., and Parks, K.: Wind power forecasting, in: Weather Matters for Energy, pp. 295–318, Springer, 2014.

Kaimal, J. C. and Finnigan, J. J.: Atmospheric boundary layer flows: their structure and measurement, Oxford university press, 1994.

Kronebach, G. W.: An automated procedure for forecasting clear-air turbulence, Journal of Applied Meteorology, 3, 119–125, 1964.

Table B1. The top row shows RMSE and MAE (both in m s^{-1}) obtained by the BCA. Below are the resulting RMSE and MAE values from ARIMA-RF predictions utilizing individual inputs for all forecasting timescales. Input features are separated into inertial (top), stability (middle), and turbulence (bottom) variables, as described in Section 4.

	10 Minute		Hourly		3 Hour	
Model/Input	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
BCA	0.726	0.528	1.132	0.863	1.575	1.240
U	0.731	0.534	1.148	0.877	1.605	1.270
θ	0.731	0.535	1.143	0.871	1.641	1.295
W	0.730	0.534	1.145	0.876	1.627	1.278
t	0.729	0.533	1.127	0.864	1.559	1.233
N^2	0.731	0.535	1.147	0.875	1.627	1.282
$\partial T/\partial z$	0.732	0.536	1.155	0.880	1.622	1.277
T	0.742	0.546	1.170	0.897	1.824	1.412
$\overline{w'T'}$	0.730	0.533	1.149	0.877	1.631	1.298
Ri_f	0.729	0.532	1.145	0.874	1.620	1.278
Ri_g	0.728	0.531	1.135	0.868	1.590	1.251
σ_u	0.731	0.533	1.140	0.869	1.628	1.297
u^*	0.730	0.534	1.149	0.881	1.611	1.277
TKE	0.730	0.532	1.144	0.871	1.626	1.296
TI	0.730	0.534	1.131	0.868	1.603	1.263
u^*/U	0.730	0.534	1.151	0.871	1.598	1.258

Ku, H. H. et al.: Notes on the use of propagation of error formulas, Journal of Research of the National Bureau of Standards, 70, 1966.

Lange, M.: On the uncertainty of wind power predictions—Analysis of the forecast accuracy and statistical distribution of errors, Journal of solar energy engineering, 127, 177–184, 2005.

Lazarevska, E.: Wind Speed Prediction based on Incremental Extreme Learning Machine, in: Proceedings of The 9th EUROSIM Congress on Modelling and Simulation, EUROSIM 2016, The 57th SIMS Conference on Simulation and Modelling SIMS 2016, 142, pp. 544–550, Linköping University Electronic Press, 2018.

Li, F., Ren, G., and Lee, J.: Multi-step wind speed prediction based on turbulence intensity and hybrid deep neural networks, Energy Conversion and Management, 186, 306–322, 2019.

Lozovatsky, I. and Fernando, H.: Mixing efficiency in natural flows, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 371, 20120 213, 2013.

Lubitz, W. D. and Michalak, A.: Experimental and theoretical investigation of tower shadow impacts on anemometer measurements, Journal of Wind Engineering and Industrial Aerodynamics, 176, 112–119, 2018.

McCaffrey, K., Quelet, P. T., Choukulkar, A., Wilczak, J. M., Wolfe, D. E., Oncley, S. P., Brewer, W. A., Debnath, M., Ashton, R., Iungo, G. V., et al.: Identification of tower-wake distortions using sonic anemometer and lidar measurements, Atmospheric Measurement Techniques (Online), 10, 2017.

Mellit, A.: Artificial Intelligence technique for modelling and forecasting of solar radiation data: a review, International Journal of Artificial intelligence and soft computing, 1, 52–76, 2008.

Mohandes, M. A., Halawani, T. O., Rehman, S., and Hussain, A. A.: Support vector machines for wind speed prediction, Renewable Energy, 29, 939–947, 2004.

Morf, H.: Sunshine and cloud cover prediction based on Markov processes, Solar Energy, 110, 615–626, 2014.

Moses, H. and Daubek, H. G.: Errors in wind measurements associated with tower-mounted anemometers, Bulletin of the American Meteorological Society, 42, 190–194, 1961.

NCAR/UCAR: NCAR/EOL Quality Controled 5-minute ISFS surface flux data, geographic coordinate, tilt corrected, version 1.1. UCAR/NCAR - Earth Observing Laboratory, https://doi.org/10.26023/ZDMJ-D1TY-FG14, 2019.

Olson, J. B., Kenyon, J. S., Djalalova, I., Bianco, L., Turner, D. D., Pichugina, Y., Choukulkar, A., Toy, M. D., Brown, J. M., Angevine, W. M., et al.: Improving wind energy forecasting through numerical weather prediction model development, Bulletin of the American Meteorological Society, 100, 2201–2220, 2019.

Optis, M. and Perr-Sauer, J.: The importance of atmospheric turbulence and stability in machine-learning models of wind farm power production, Renewable and Sustainable Energy Reviews, 112, 27–41, 2019.

Orlando, S., Bale, A., and Johnson, D. A.: Experimental study of the effect of tower shadow on anemometer readings, Journal of Wind Engineering and Industrial Aerodynamics, 99, 1–6, 2011.

Papadopoulos, K., Helmis, C., and Amanatidis, G.: An analysis of wind direction and horizontal wind component fluctuations over complex terrain, Journal of Applied Meteorology, 31, 1033–1040, 1992.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in Python, Journal of machine learning research, 12, 2825–2830, 2011.

Ramasamy, P., Chandel, S., and Yadav, A. K.: Wind speed prediction in the mountainous region of India using an artificial neural network model, Renewable Energy, 80, 338–347, 2015.

Rogers, A. L., Rogers, J. W., and Manwell, J. F.: Comparison of the performance of four measure–correlate–predict algorithms, Journal of wind engineering and industrial aerodynamics, 93, 243–264, 2005.

Seabold, S. and Perktold, J.: Statsmodels: Econometric and statistical modeling with python, in: Proceedings of the 9th Python in Science Conference, vol. 57, p. 61, Austin, TX, 2010.

Shibata, R.: Selection of the order of an autoregressive model by Akaike's information criterion, Biometrika, 63, 117–126, 1976.

Soman, S. S., Zareipour, H., Malik, O., and Mandal, P.: A review of wind power and wind speed forecasting methods with different time horizons, in: North American Power Symposium 2010, pp. 1–8, IEEE, 2010.

Stiperski, I., Calaf, M., and Rotach, M. W.: Scaling, Anisotropy, and Complexity in Near-Surface Atmospheric Turbulence, Journal of Geophysical Research: Atmospheres, 124, 1428–1448, 2019.

Van der Hoven, I.: Power spectrum of horizontal wind speed in the frequency range from 0.0007 to 900 cycles per hour, Journal of meteorology, 14, 160–164, 1957.

Vassallo, D., Krishnamurthy, R., and Fernando, H. J.: Decreasing Wind Speed Extrapolation Error via Domain-Specific Feature Extraction and Selection, Wind Energy Science, 5, 959–975, 2020.

Wang, F., Mi, Z., Su, S., and Zhao, H.: Short-term solar irradiance forecasting model based on artificial neural network using statistical feature parameters, Energies, 5, 1355–1370, 2012.

Wu, W., Zhang, B., Chen, J., and Zhen, T.: Multiple time-scale coordinated power control system to accommodate significant wind power penetration and its real application, in: 2012 IEEE Power and Energy Society General Meeting, pp. 1–6, IEEE, 2012.

Yang, D., Jirutitijaroen, P., and Walsh, W. M.: Hourly solar irradiance time series forecasting using cloud cover index, Solar Energy, 86, 3531–3543, 2012.

Yang, Q., Berg, L. K., Pekour, M., Fast, J. D., Newsom, R. K., Stoelinga, M., and Finley, C.: Evaluation of WRF-predicted near-hub-height winds and ramp events over a Pacific Northwest site with complex terrain, Journal of applied meteorology and climatology, 52, 1753–1763, 2013.