RESEARCH ARTICLE

WILEY

Quantifying sensitivity in numerical weather prediction-modeled offshore wind speeds through an ensemble modeling approach

Mike Optis[1] | Andrew Kumler[1] | Joseph Brodie[2] | Travis Miles[2]

[1]National Renewable Energy Laboratory, Golden, Colorado, USA

[2]Center for Ocean Observing Leadership, Rutgers, The State University of New Jersey, New Brunswick, New Jersey, USA

Correspondence
Mike Optis, National Renewable Energy Laboratory, 15013 Denver W Pkwy, Golden, CO, USA.
Email:mike.optis@nrel.gov

Summary

A decade of research has shown that numerical weather prediction (NWP)-modeled wind speeds can be highly sensitive to the inputs and setups within the NWP model. For wind resource characterization applications, this sensitivity is often addressed by constructing a range of setups and selecting the one that best validates against observations. However, this approach is not possible in areas that lack high-quality hub height observations, especially offshore wind areas. In such cases, techniques to quantify and disseminate confidence in NWP-modeled wind speeds in the absence of observations are needed. We address this need in the present study and propose best practices for quantifying the spread in NWP-modeled wind speeds. We implement an ensemble approach in which we consider 24 different setups to the Weather Research and Forecasting (WRF) model. We construct the ensemble by considering variations in WRF version, WRF namelist, atmospheric forcing, and sea surface temperature (SST) forcing. Our analysis finds that the standard deviation produces more consistent estimates compared to the interquartile range and tends to be the more conservative estimator for ensemble variability. We further find that model spread increases closer to the surface and on shorter time scales. Finally, we explore methods to attribute total ensemble variability to the different ensemble components (e.g., atmospheric forcing and SST product) and find that contributions by components also vary depending on time scale. We anticipate that the methods and results presented in this paper will provide a reasonable foundation for further research into ensemble-based wind resource data sets.

KEYWORDS
model sensitivity, numerical weather prediction, offshore, wind resource

1 | INTRODUCTION

1.1 | U.S. offshore wind and lack of observations

Offshore wind energy is poised for considerable development in the U.S. North and Mid-Atlantic waters. The first U.S. offshore wind farm—Block Island Wind Farm—has been providing 30-megawatt (MW) capacity (5 × 6-MW turbines) off the coast of Martha's Vineyard since 2016. There
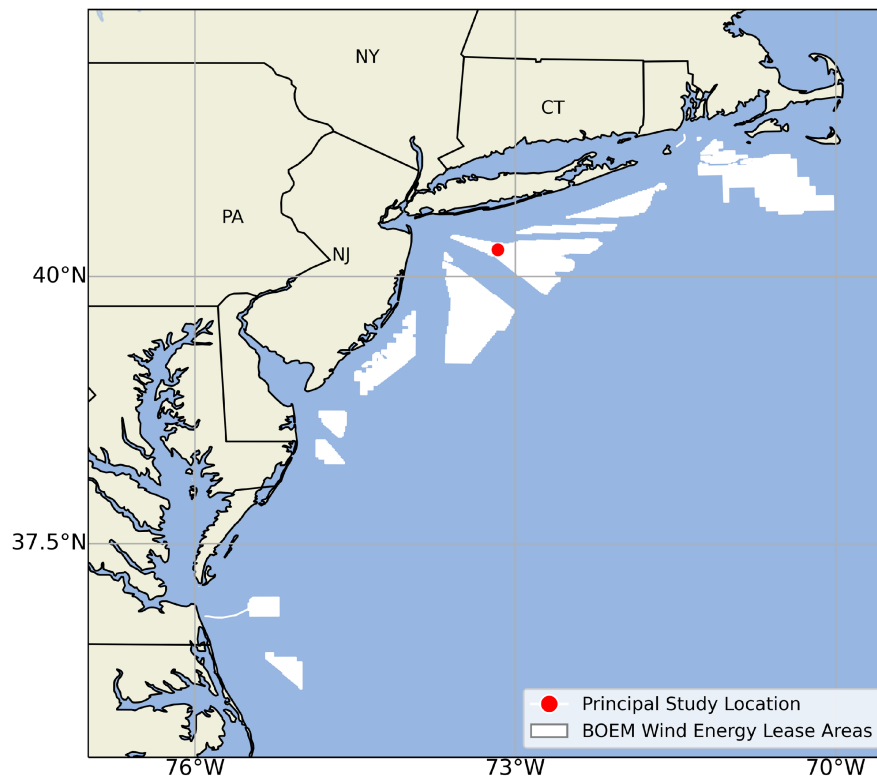
FIGURE 1 Offshore wind energy lease and call areas in the U.S. Atlantic, as of July 2020. The red point denotes the location where much of the analysis in this study was performed. BOEM = Bureau of Ocean Energy Management

are currently 15 active lease areas with over 21 gigawatts (GW) of capacity spanning from Massachusetts to North Carolina (Figure 1) and a planned 86-GW capacity in all U.S. waters by 2050.[1] Altogether, tens of billions of dollars are poised to be invested in the U.S. offshore wind industry in the coming decade.

This investment will come with considerable financial risks, not the least of which is the uncertainty in the U.S. offshore wind resource. A robust wind resource assessment requires, at a minimum, wind measurements at hub height for at least 1 year. Ideally, multiple years of data with wind speeds at multiple heights and at multiple locations across a proposed wind farm would be available.[2] However, there are sparse wind observations available in U.S. North and Mid-Atlantic waters, which are largely limited to buoy observations below 10 m from the National Data Buoy Center (NDBC). Observational data coverage is beginning to improve through the emergence of floating lidar[3,4] and a planned offshore meteorological mast.[5] Despite these recent advancements, the high installation and operation costs of these emerging observational technologies will continue to be a limitation; as a result, spatial and temporal gaps in the data will persist.

1.2 | Mesoscale models and their sensitivity

Wind energy analyses that require more complete data coverage than that provided by offshore observations generally rely on simulated data from a numerical weather prediction (NWP) model. Analyses that require these data include the estimation of wind plant annual energy production (AEP),[6] grid integration[7] and life-cycle cost analyses,[8] and capacity expansion studies.[9] Accuracy in these analyses is vital for the financial health of the offshore wind industry. For example, Hale[10] found that a 3% change in AEP for a 200-MW onshore wind farm translated to a difference in $17 M in net present value. In terms of new project financing, a small uncertainty in the annual wind resource of 1%–2% can translate to 3%–4% in AEP uncertainty and can considerably impact the interest rate on loans for new wind farm construction.[2] Accurate quantification of this uncertainty is especially important offshore given the sparsity of offshore observations, the associated difficulty in validation, and therefore increased wind resource uncertainty.

A key source of uncertainty in NWP-modeled wind speeds is their sensitivity to model inputs and setup. Over a decade of research into this topic has highlighted sensitivity in NWP-modeled wind speeds to the planetary boundary layer (PBL) scheme,[11-16] large-scale atmospheric forcing,[14,16-18] data assimilation techniques,[19] nesting techniques,[13] grid size,[16] vertical resolution,[14] horizontal resolution,[15] spin-up time,[14] sea surface temperature (SST),[14] convection schemes,[11] and soil models.[11]

The most comprehensive analysis of NWP model sensitivity for wind energy applications was recently published as part of development of the New European Wind Atlas (NEWA).[20] The NEWA project assessed the sensitivity of mean annual wind speeds to variations across a broad range of model inputs and parameterizations, including reanalysis products, PBL scheme, SST product, radiation scheme, land surface model, domain size, atmospheric nudging, model version, number of vertical levels, surface roughness, simulation length, one-way versus two-way nesting, and the high-performance computing compiler used to run the simulations. These sensitivity analyses were used to find the best model setup to run the 30-year NEWA.[21]

1.3 | Need to quantify and disseminate NWP sensitivity

Despite the extensive literature demonstrating NWP model sensitivity, no work to our knowledge has focused on methods to synthesize that sensitivity into usable confidence metrics for wind energy analyses. The identification, quantification, and dissemination of NWP-modeled wind speed sensitivity and its impact on wind energy applications that use such data are a critical knowledge gap that should be addressed to support the U.S. offshore wind industry.

The goal of this study is to develop and recommend methods for quantifying and disseminating the sensitivity arising from NWP-modeled wind speeds. To conduct this research, the National Renewable Energy Laboratory (NREL) collaborated with the Rutgers University Center for Ocean Observing Leadership (RU-COOL) to perform a sensitivity analysis of the Weather Research and Forecasting (WRF) mesoscale model[†,22] over a spatial domain encompassing the New Jersey, New York, and Delaware offshore wind energy lease areas. NREL and RU-COOL have each independently developed WRF-based time series wind resource data sets for this region: NREL has produced 7 years of WRF-based wind resource data in its Wind Integration National Data Set (WIND) Toolkit,[23] while RU-COOL has developed and runs a daily WRF-based forecast model called RU-WRF for the New Jersey and neighboring offshore wind energy areas.[24] By considering the main differences in model inputs and setups between the WIND Toolkit and RU-WRF, we construct an ensemble of 24 WRF model setups based on differences in the WRF version, WRF namelist, atmospheric forcing, and SST forcing. We first explore and compare different metrics for quantifying wind speed sensitivity from the ensembles. Then we investigate how sensitivity changes with height from the surface and over different time scales. Finally, we explore methods to attribute total ensemble variability to the different ensemble components (e.g., WRF namelist and WRF version) and investigate how those attributions change with time scale.

2 | EXPERIMENTAL SETUP

In this section, we describe the WRF simulations that were conducted for this analysis. A total of 24 different WRF setups were created and run over a full year; these setups are informed by the differences between NREL's WIND Toolkit and RU-COOL's RU-WRF model.

From here on, we use the term sensitivity to represent the spread in modeled wind speeds from different ensembles.

2.1 | Study domain

We implement a one-way nested 9- to 3-km domain for the WRF simulations centered on the New Jersey offshore wind energy region (Figure 2). The 3-km domain spans from New York to Maryland, and the 9-km domain spans from Maine to North Carolina. We model a 1-year time period from June 1, 2015, through May 30, 2016. This period of record is selected because of good data coverage for RU-COOL's custom SST product used in this analysis (described in Section 2.2).

We run the WRF model in separate 3-day simulation periods each offset by 2 days with no nudging. The first 24 h in the simulation are used to initialize or "spin-up" the WRF model. The full 1-year time series is then constructed by concatenating the 2-day simulation periods. This approach is used to limit the influence of model "drift" over long simulation periods (i.e., the tendency for the WRF simulation to deviate from the large-scale forcing in the absence of atmospheric nudging).

2.2 | Ensemble members for WRF simulations

We construct the ensemble of WRF simulations by considering the differences between NREL's WIND Toolkit setup and RU-COOL's RU-WRF setup. A summary of the main differences between the two models is given in Table 2. Based on these differences, we identify four principal

---

[†]The WRF model is a state-of-the-art, open-source NWP model produced by the National Center for Atmospheric Research. The model is used by tens of thousands of users globally for operational NWPs, atmospheric research, and wind energy analyses.
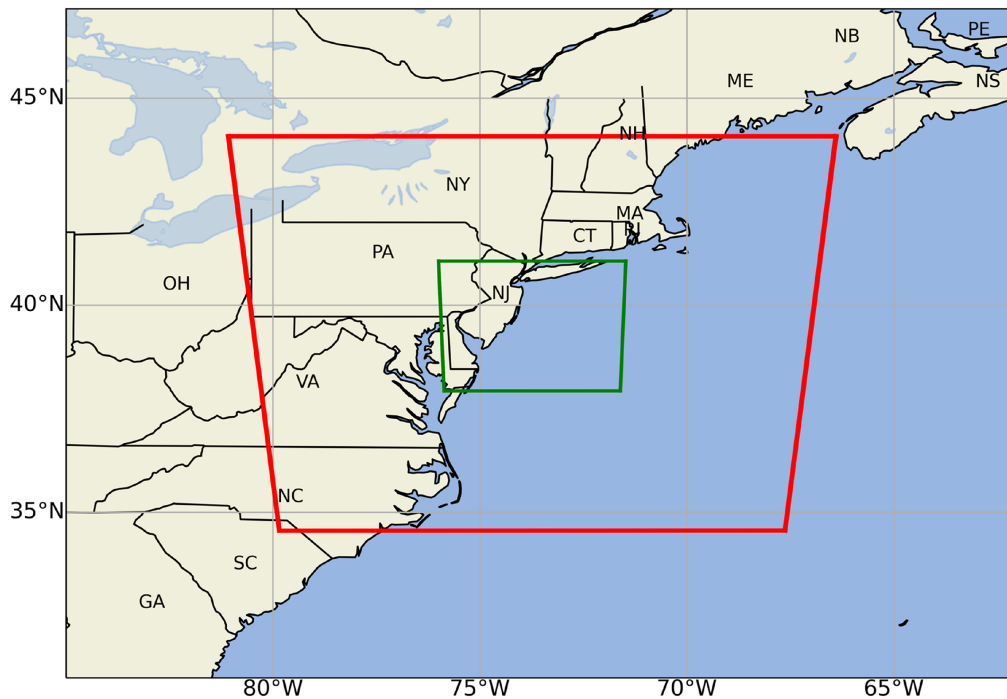
FIGURE 2 The 9-km (red) and 3-km (green) WRF domains used in this analysis

categories that are expected to produce sensitivity in the modeled wind speeds. These categories are described as follows and summarized in Table 1:

1. *Atmospheric forcing product*: The WIND Toolkit uses the ERA-Interim (ERAI) reanalysis product,[25] which provides boundary conditions to WRF at 6-h intervals. The RU-WRF model uses the Global Forecasting System (GFS) product from the National Centers for Environmental Prediction (NCEP).[26] Given that GFS is a forecast product and less accurate than reanalysis (or hindcast) products such as ERAI, we replace the GFS in this ensemble with NCEP's Final Operational Global Analysis (FNL). The FNL is based on the same model as the GFS but is prepared about an hour after the GFS is initialized. This later preparation allows for more observational data to be incorporated and results in a more accurate product. To construct a hindcast product from FNL, we consider output at model initialization every 6 h and concatenate these outputs together to produce a continuous 6-h time series over the validation period.

2. *SST input*: RU-COOL developed a customized SST product designed to better capture the frequent occurrence of surface cold water caused by upwelling of the subsurface cold pool in the Mid-Atlantic Bight.[27,28] This SST product is used as input to RU-WRF. Similar to other SST products, the RU-COOL product is based on satellite data; however, the product differs in both the method of composting and the declouding algorithm in order to capture more cold pool events, which are often identified as clouds by standard algorithms and therefore discarded. By contrast, NREL's WIND Toolkit uses the NCEP real-time global (RTG) SST product. Both data products are produced at high spatial resolution

TABLE 1 Summary of the WRF model components used to construct the 24-member ensemble

Category	Values considered for hindcast ensemble
Atmospheric forcing	NCEP FNL
	ERAI
SST input	RU-COOL coldest-pixel
	NCEP RTG
	Default from atmospheric forcing
WRF version	3.9
	4.0
WRF namelist	RU-WRF
	WIND Toolkit

TABLE 2 Comparison of key attributes between RU-WRF and the WIND Toolkit

Feature	RU-WRF	WIND Toolkit
Atmospheric forcing	FNL 0.25° × 0.25° forecast	ERA-Interim 0.7° × 0.7° reanalysis
PBL scheme	MYNN 2.5	YSU
Surface layer scheme	Monin–Obukhov with Janjic Eta	Monin–Obukhov
Microphysics	Thompson	Ferrier
Longwave radiation	RRTMG	RRTM
Shortwave radiation	Goddard	Dudhia
Topographic data	USGS GTOPO30	USGS GTOPO30
Land-use data	NLCD 3 s	NLCD 3 s
Cumulus parameterization	Off	Kain–Fritsch
Number of Vertical levels	40	41
Near-surface-level heights (m)	53.7, 131.0, and 232.5	33.9, 67.9, 102.1, 136.4, and 170.8

(2 km for RU-COOL SST; $[1/12]°$ for NCEP RTG SST). In addition to these high-resolution products, we also consider the default products included in the ERAI and FNL atmospheric data products.[‡]

3. *WRF version*: At the time this analysis was conducted, the RU-WRF was based on WRF Version 3.9. The WIND Toolkit was produced in 2013 and was based on WRF Version 3.4; however, an updated WIND Toolkit to be released in 2021 or 2022 will be based on WRF Version 4.0 or greater. Therefore, we consider WRF Versions 3.9 and 4.0 in this analysis. Given considerable updates made between WRF Versions 3.9 and 4.0,[§] we expect the change in WRF version to result in some model sensitivity.

4. *WRF namelists*: The namelist.input file in WRF specifies the key physics and dynamical properties of the simulation (e.g., parameterization schemes, spatial resolution, and model time step). Given the large number of differences between the WIND Toolkit and RU-WRF namelists (summarized in Table 2), we limit this analysis to only considering the different WIND Toolkit and RU-WRF namelists as a whole when building the ensemble. We will therefore be unable to attribute any sensitivity to the different namelists to specific namelist parameters. However, we expect the different PBL schemes to be large drivers of model variability given the extensive literature on the subject (as described in Section 1).

In total, 24 ensemble members were constructed considering all possible combinations of model inputs and setups described previously.

3 | RESULTS

When presenting results in this section, we will usually compare modeled wind speeds at 130 m. This height is selected because offshore hub heights are anticipated to be around this height.[29] Furthermore, both the WIND Toolkit and RU-WRF model setups have a vertical level that is, on average, close to 130 m (136 m for WIND Toolkit and 131 m for RU-WRF). Choosing this height rather than a typical comparison height (e.g., 80 or 100 m) minimizes the impact of interpolation from different model levels and the creation of artificial spread between the WIND Toolkit and RU-WRF-based ensemble members. Given the small interpolation distance to 130 m, we use linear interpolation between model levels to calculate the 130-m wind speeds.

We also conduct most of the analysis on a single grid point located within a New York offshore wind energy lease area (Figure 1). We call this location the Principal Study Location, or PSL, throughout this paper.

3.1 | NWP model sensitivity

To get a sense for the sensitivity in modeled wind speeds under the 24 different WRF model setups, we first present a sample time series plot of wind speeds and estimated power production in Figure 3. Here we select a sample 4-day time period at the PSL. This specific period was selected because of a strong modeled ramp event. Estimated power production is based on the 130-m wind speeds and the NREL 15-MW power curve.[29] We see considerable variability in wind speeds and power over the 4-day period. The range in wind speeds (i.e., difference between highest and
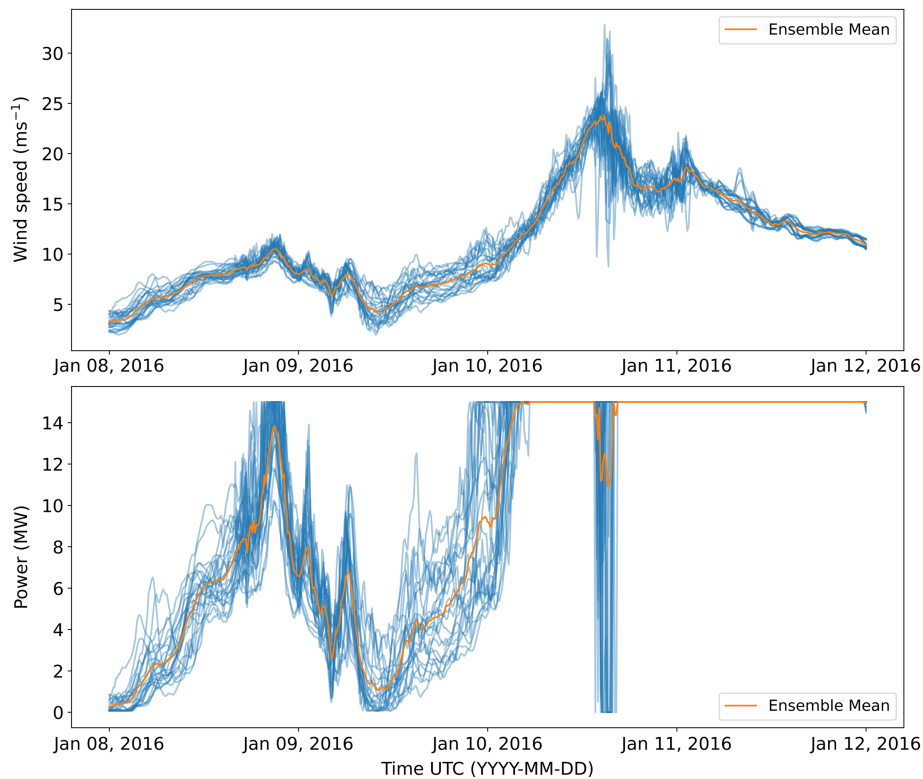
FIGURE 3 Snapshot of modeled 130-m wind speeds (top panel) and estimated wind power (bottom panel) from the 24-member WRF ensemble extracted from the closest grid cell to the PSL. Ensemble members are shown in blue, and the ensemble mean in orange. Estimated power is based on the NREL 15-MW power curve[29]

lowest ensemble member) often exceeds $4\,\mathrm{m\,s^{-1}}$. The range in power production is even more pronounced given the cubic relationship to wind speed. At several time steps, some ensemble members model 3 times the power of others. Furthermore, during the ramp event on 2016-01-10, some ensemble members predict a cutoff of wind power (i.e., wind speeds exceed the cutoff of $25\,\mathrm{m\,s^{-1}}$), while most predict peak power, resulting in an ensemble mean that drops as low as 12 MW.

## 3.2 | Best metric to quantify wind speed sensitivity

There are several metrics that could be used to quantify the high spread in the modeled wind speeds and power shown in Figure 3. Common measures of spread typically include the range (difference between the greatest and smallest value), interquartile range (IQR), and standard deviation (STD). These metrics are often normalized to express spread in percentage units; for example, the STD divided by the mean of the data is generally referred to as the coefficient of variation (CoV). Normalization by the spread of wind speeds (e.g., STD or range) is also common.

We are interested here in finding a single metric of sensitivity that can be disseminated along with wind speed data such that wind energy stakeholders can assign a level of confidence to that data. In this regard, the use of multiple quartiles is not ideal for this application. Furthermore, the range metric is highly sensitive to outliers and may tend to overestimate model variability in cases where we have a small number of data points (i.e., 24 ensembles). Given these considerations, the IQR and STD metrics seem best suited for quantifying sensitivity in modeled wind speeds from a 24-member ensemble.

In Figure 4, we explore the characteristics of the IQR and STD metrics over the same 4-day event in Figure 3. Here we consider the metrics applied to both 130-m wind speed and estimated power and consider the metrics when normalized by the wind speed at each time interval (i.e., IQR-norm and CoV). Figure 4 reveals several interesting comparisons between the different metrics:

1. IQR is generally higher than STD, both in normalized and non-normalized units. We generally expect this result given that IQR $\approx 1.34 \times$ STD for a normal distribution.¶The extent to which the 24-member ensembles resemble a normal distribution is investigated in Section 3.2.1.

¶For a normal distribution, the Q1 value falls on $-0.67 \times$ STD and the Q3 value falls on $0.67 \times$ STD, giving the total IQR = $1.34 \times$ STD
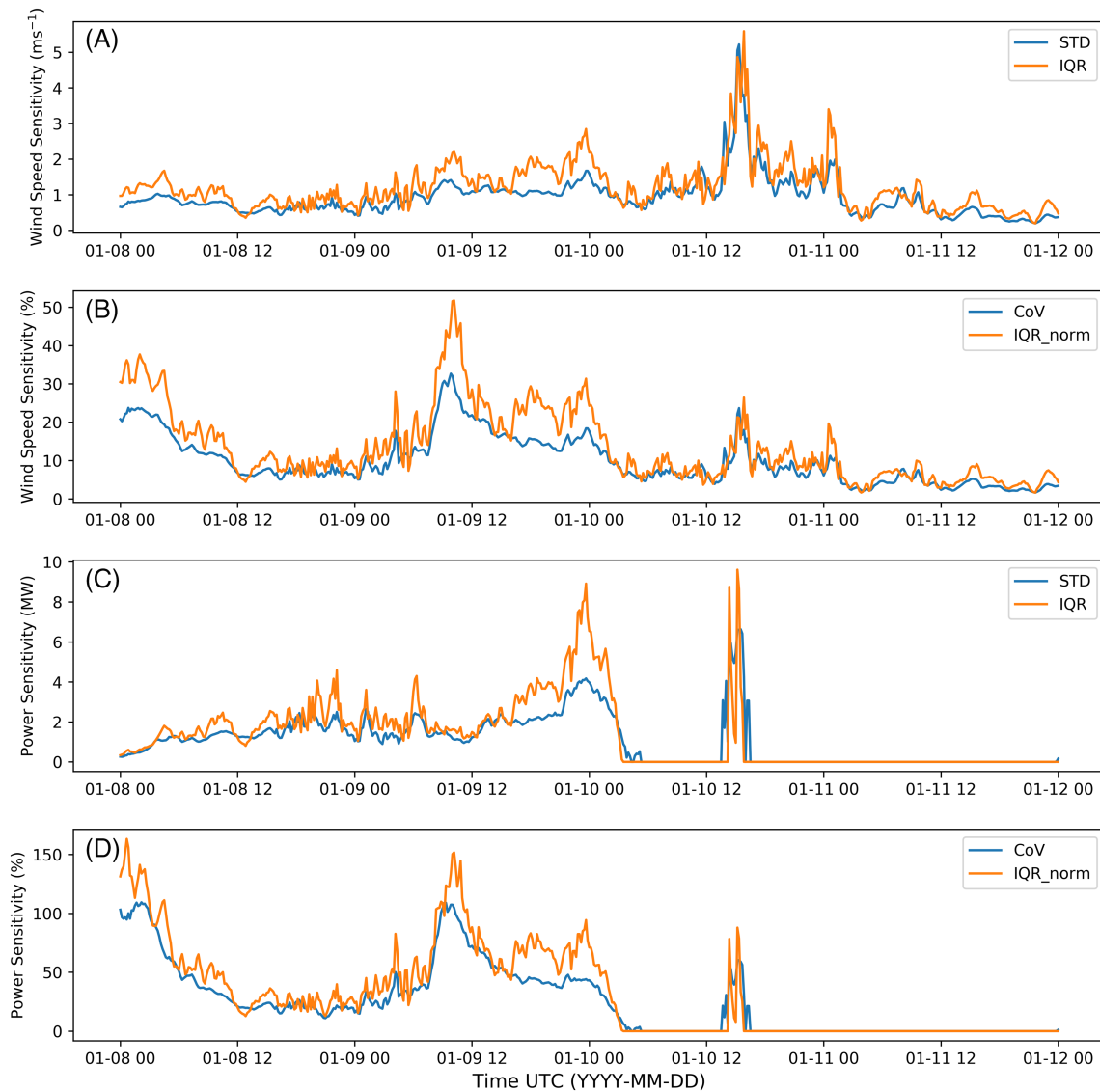
FIGURE 4 Exploring different metrics for wind speed sensitivity during the 4-day sample period in Figure 3. Here we contrast STD and IQR (A), CoV and normalized IQR (B) when applied to wind speed, and the same metrics when applied to estimated wind power (C) and (D)

2. The normalized and non-normalized metrics tend to be somewhat uncorrelated and quantify sensitivity differently for different events. For example, non-normalized wind speed metrics in Figure 4A show the highest wind speed sensitivity during the high wind speed peak. By contrast, the normalized metrics in Figure 4B show lower sensitivity relative to other events, on account of dividing by the high magnitude of wind speed. Instead, the normalized metric shows the highest sensitivity during an earlier event, which is mostly attributable to the moderate sensitivity and the lower relative wind speeds. More precisely, normalizing by mean wind speed can considerably impact the sensitivity analysis, leading to interpretations of higher sensitivity for low-wind-speed events and lower sensitivity for high-wind-speed events, relative to what is estimated from a non-normalized approach.

3. During the peak wind event, when some ensemble members exceed the cut-out wind speed and zero power is predicted, both the IQR and STD metrics predict wind power sensitivity (Figure 4C and 4D). We notice that the STD metric predicts sensitivity over a longer period during this peak wind event relative to IQR. We explain this observation by the fact that the IQR only considers the middle 50% of the data, while STD accounts for all ensemble members. Therefore, in cases where the middle 50% of data do not exceed the cut-out wind speed but some data outside the IQR range do, we see non-zero STD values but zero IQR values.

4. The IQR metric shows more hour-to-hour variability (i.e., noise) compared to the STD metric, while the STD metric tends to transition more smoothly from 1 h to the next.

We explore these observations in more detail in the following subsections by extending the analysis to all hourly data in the full model year.

3.2.1 | IQR and STD magnitudes and the normalcy of ensemble distributions

Figure 4 showed that IQR tended to be higher than STD over that sample 4-day period. Exploring this fact in more detail over the full model year, we find that the IQR is almost always greater than STD (in fact, 90% of the time) with a mean ratio between IQR and STD of 1.41. We
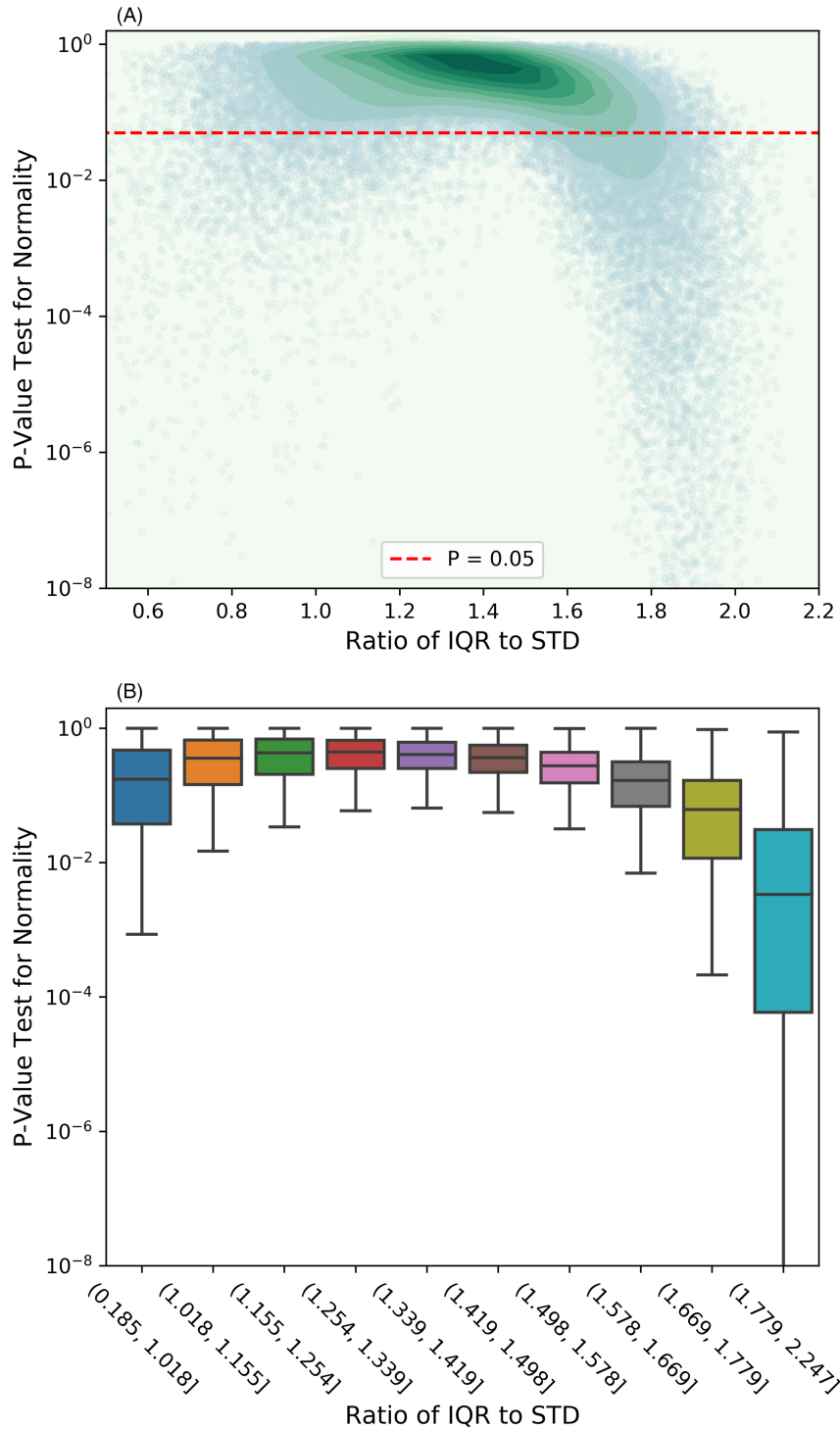
FIGURE 5 Comparing the hourly IQR to STD ratio as it depends on *p* value tests for normalcy in the 24-member hourly ensembles. Both 2-D kernel density estimates (A) and boxplots (B) are shown. The red line in subplot (A) denotes the 5% *p* value line, commonly used to identify a statistically meaningful relationship

hypothesize that the tendency of IQR to exceed STD is related to how closely the 24-member ensemble distributions of wind speeds for a given hour resembles a normal distribution.

We explore this hypothesis in Figure 5, where we compare the ratio of IQR/STD to the degree of normalcy in the distribution of wind speeds at each hour. Here we use the "normaltest" function in Python SciPy library,[30] which tests the null hypothesis that a data sample came from a normal distribution. The function returns a p value or probability of the hypothesis test: Larger p values (e.g., 0.05 to 1, or 5% to 100%) indicate a stronger likelihood that the data came from a normal distribution, while lower p values (e.g., less than 0.05) indicate that the sample data are likely not from a normal distribution.

Figure 5 shows that wind speed distributions from each 24-member hourly ensemble tend to be most normally distributed when the IQR/STD ratio is around the 1.34 range, as we would expect. As we move away from this value in either direction, we tend to see less normal distributions and notably highly skewed distributions when IQR is much greater than STD. Examining the distribution of p values, we see that the values tend to cluster in the 10^{-1} to 10^{0} range, or between 10% and 100%. This clustering indicates reasonable likelihoods that the 24-member wind speed ensembles tend to resemble normal distributions.

3.2.2 | IQR and STD hour-to-hour variability

Referring back to Figure 4, we see that the hourly IQR metric tended to be more "noisy" than the STD metric. We explore this fact for the whole data set in Figure 6 by comparing distributions of the magnitudes of hour-to-hour changes in each metric. As shown in the figure, the IQR metric is considerably more variable hour-to-hour than the STD metric over the whole data set. We attribute this finding to the fact that IQR is a range-based metric and is highly sensitive to the two ensemble members that define the IQR. By contrast, STD is a measure of the average distance to the mean for all 24 ensemble members and is therefore less sensitive to individual ensemble members—assuming, of course, that there are not egregious outliers. It is apparent for this specific application that an average-based metric such as STD is more stable and consistent than a range-based metric such as IQR.

3.2.3 | Most appropriate metric for quantifying sensitivity

Based on the analysis in this section, we conclude that the STD and IQR metrics can behave somewhat differently when used to quantify ensemble sensitivity. The STD metric is found to be more "stable" from hour to hour, while the IQR varied more substantially. The STD metric also captured longer time periods of sensitivity during the cut-out wind speed events. In such scenarios, a sensitivity metric such as STD that is more conservative and predicts sensitivity over longer time periods seems advantageous over a metric that does not capture that potential, such as IQR. However, being conservative may come at the cost of higher uncertainty and greater financial penalty in wind forecasts, for example, making the STD less desirable.

Overall, both metrics do a reasonable job at capturing ensemble sensitivity. Noting that the 24 ensembles tend to resemble normal distributions from hour to hour, then the STD and IQR metrics become somewhat interchangeable and should generally capture the ensemble spread.
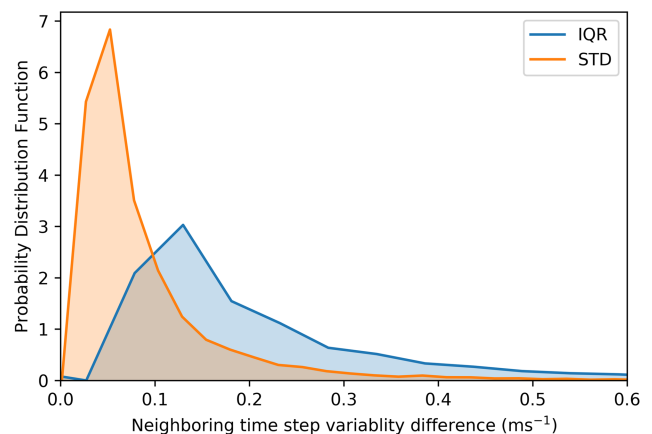
FIGURE 6 Distributions of the absolute difference in hour-to-hour STD and IQR based on 130-m wind speeds at the PSL

3.3 | Should the sensitivity metric be normalized?

As discussed in Section 3.1 and illustrated in Figure 4, normalizing the STD by wind speed to calculate CoV tended to put more emphasis on lower-wind-speed events and less emphasis on high-wind-speed events. In Figure 7, we explore this tendency in more detail using data from the full model year.

We see in Figure 7A and 7B that in relation to wind speed, the wind speed STD and CoV metrics trend in opposite directions: STD generally increases with wind speed and CoV generally decreases. This behavior is as expected; we tend to see higher fluctuations in wind speed when wind speeds are generally higher, and therefore, the STD metric is proportional to wind speed. However, when dividing by wind speed to calculate CoV, we enhance sensitivity when wind speeds are low and diminish sensitivity when wind speeds are high.

We further contrast STD and CoV when predicting wind power sensitivity in Figure 7C and 7D, again using the 15-MW NREL offshore power curve. We immediately see a very different trend compared to Figure 7A and 7B: Instead of monotonically increasing or decreasing relationships, we see that wind power sensitivity starts at 0 with low wind before increasing along with wind speed up to about $10 \, \mathrm{m \, s^{-1}}$ and then decreasing back down to 0. Comparing the STD- and CoV-based trends, we see that STD in wind power increases more slowly with wind speed up to peak sensitivity around $10 \, \mathrm{m \, s^{-1}}$ before sharply decreasing to 0. By contrast, the CoV in wind power increases very sharply to peak sensitivity at around $5 \, \mathrm{m \, s^{-1}}$ before gradually decreasing to 0. Similar to Figure 7A and 7B, we attribute these contrasting trends with the different regimes of emphasis: STD will tend to predict higher sensitivity for higher power values (up until peak power is reached, at least) while CoV predicts highest sensitivity in lower power ranges.

Given the different wind speed ranges emphasized by STD and CoV, we examine which is the most appropriate to use. We first note that in practical terms, the choice is not all that important. If the intention is to disseminate STD or CoV metrics along with hourly wind resource data, then it is very straightforward to calculate STD from CoV, or vice versa. Instead, the most important consideration then might simply be interpretability. For example, we could ask which metric is more easily understood and processed by a wind energy or financial analyst? In this context, the
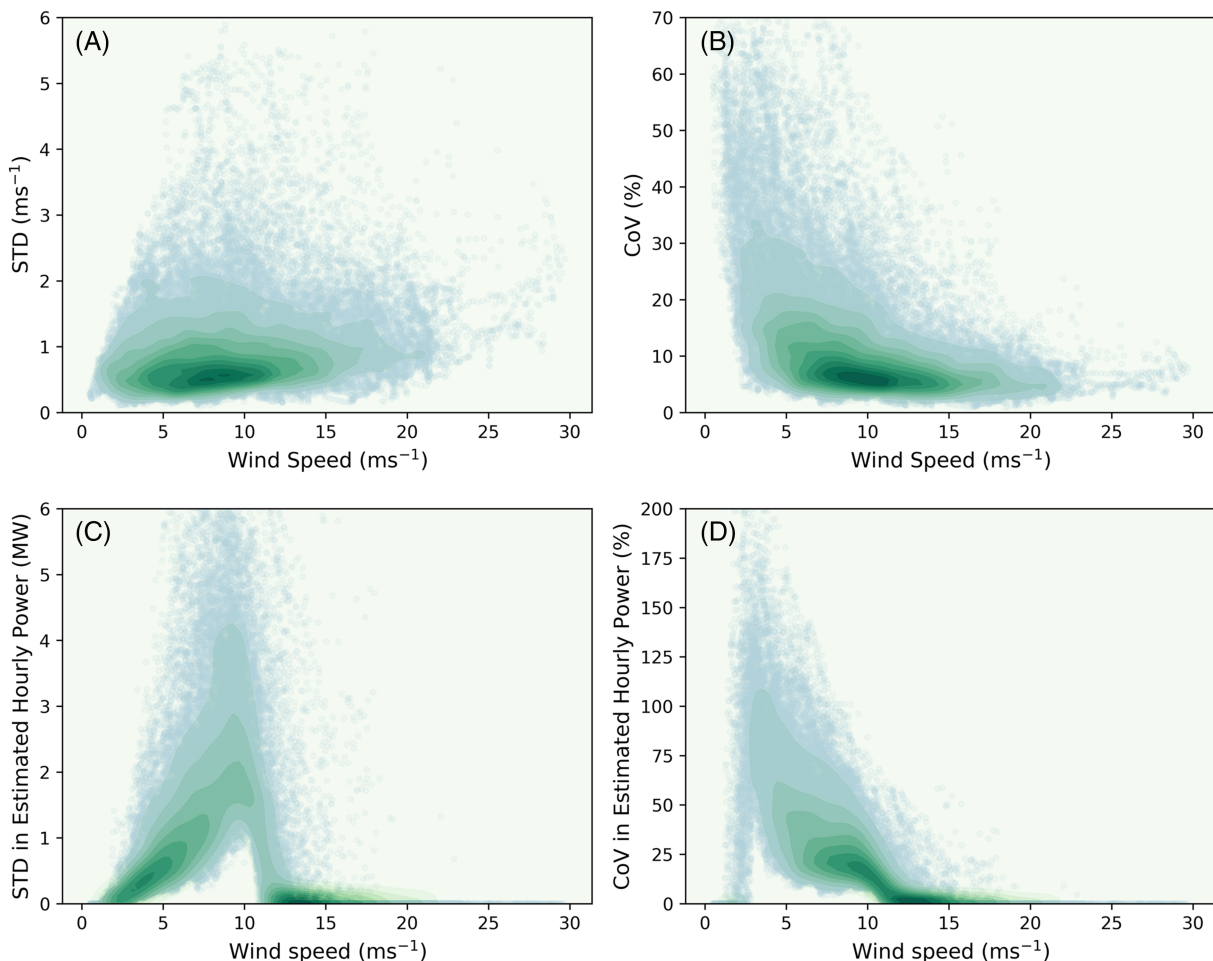

FIGURE 7 Two-dimensional kernel density estimates of wind speed against STD of wind speed (A), CoV of wind speed (B), STD of power (C), and CoV of power (D)

CoV expressed as a percentage likely has an advantage, as percentage units are commonly used and easily understood (e.g., turbulence intensity—defined as the fluctuations in wind speed divided by the mean—is the principal metric used to quantify turbulence in the wind industry). By contrast, the STD in units of $m\,s^{-1}$ also requires knowledge of the mean wind speed in order to put the magnitude of the sensitivity into context. As such, STD would probably be less interpretable.

Based on these considerations, we will move forward for the remainder of this study using CoV as the main metric for analysis. However, when presenting the remaining analyses, we will comment on any differences in results had a non-normalized sensitivity metric been used.

3.4 | Ensemble sensitivity by height and time scale

So far we have only analyzed sensitivity in wind speed and power for hourly 130-m wind speeds. In this section, we explore how this sensitivity changes with both height and time scale. We also continue to focus on the PSL for the bulk of this analysis and focus on the CoV metric. We focus here on mean values of CoV for the different time scales, which we calculate first by averaging the 10-min wind speed data by time scale (e.g., daily and monthly), computing the CoV for each averaged time step, and then presenting the mean of all CoV data across all time setups. For example, mean daily CoV is an average of 365 CoV values, and monthly mean CoV is the average of 12 CoV values, etc.
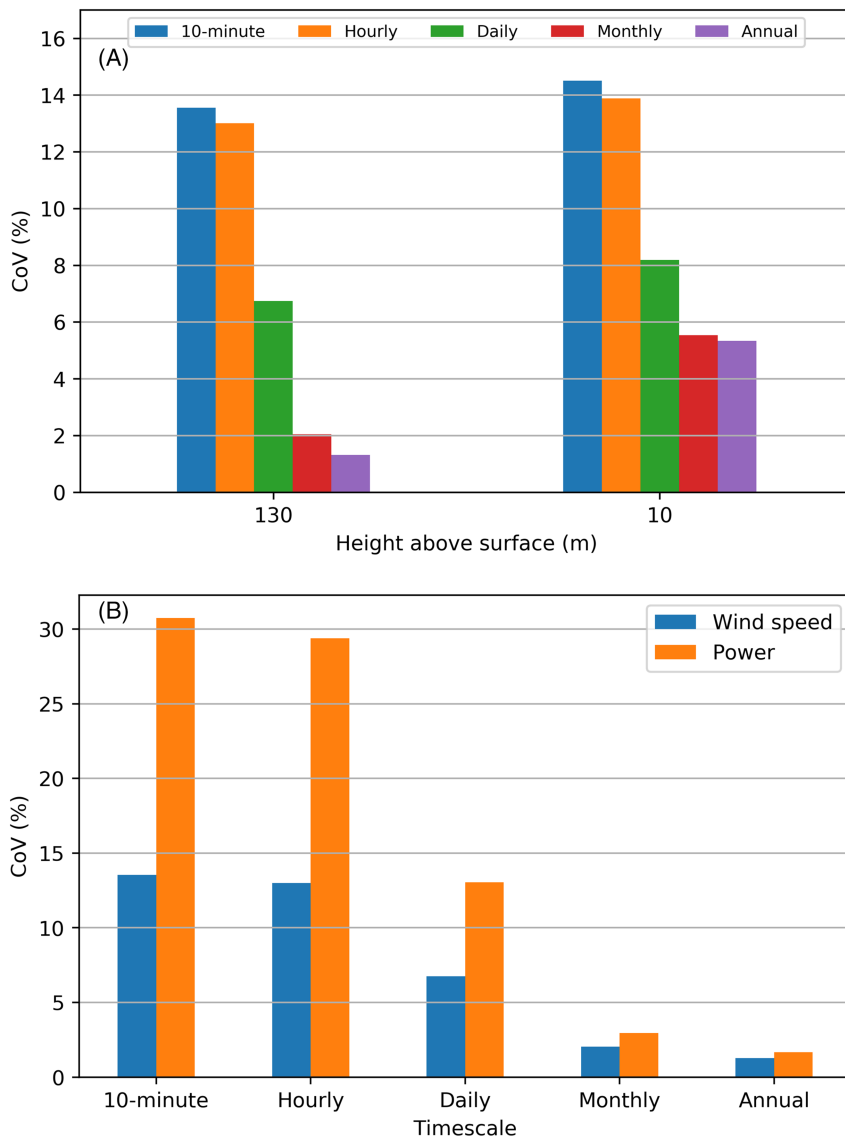
FIGURE 8 Mean coefficient of variation of 10- and 130-m modeled wind speeds (A) and 130-m wind speed and estimated power production (B) from 10-min to annual time scales. Modeled wind speeds from the PSL are used

Figure 8A presents CoV metrics for both the 130-m wind speeds and the diagnostic 10-m wind speed output from WRF. Here we see that the CoV is greatest on short time scales and smallest on long time scales, which we would expect because model differences tend to average out over longer time scales. The 130-m modeled wind speeds are subject to about 13% mean variability at the 10-min and hourly time scales and decrease to about 1.5% on an annual time scale. We also see that sensitivity is higher at 10 m compared to 130 m. We generally would expect this given the increased influence of surface physics (e.g., temperature changes and turbulent fluxes) at heights closer to the surface and the different surface physics representations across the ensembles (e.g., different PBL and surface layer schemes, as shown in Table 2). However, when reproducing this figure using non-normalized sensitivity metrics (i.e., in m s^{-1}), we see the opposite trend and lower sensitivities at 10 m. We attribute this fact to the lower magnitude wind speeds at 10 m relative to 130 m.

Less understood in Figure 8A, however, is the fact that 10- and 130-m sensitivity are comparable at short time scales but diverge considerably at higher time scales. As we will discuss in Section 3.6, this discrepancy can be attributed to how different ensemble components (e.g., reanalysis product and WRF namelist) contribute to total ensemble sensitivity at different time scales. In this case, the choice of reanalysis product dominates sensitivity at short time scales but is nearly negligible on annual scales. By contrast, the WRF namelist (in which surface physics options are specified) dominates sensitivity on the annual scale, therefore producing more sensitivity at 10 m compared to 130 m at the annual scale.

Figure 8B compares ensemble CoV between the 130-m wind speeds and estimated power using the 15-MW power curve. Because of the cubic relationship between wind speed and power, we see considerably higher sensitivity in power (about 35% or greater CoV for time scales of an hour or less). The CoV decreases to about 2% CoV on an annual time scale. Considering that AEP estimates have uncertainty on the order of 8% to 10%,[2] the impact of an additional 2% annual uncertainty from this model sensitivity under a sum of squares approach would lead to a modest but important 0.20%–0.25% additional AEP uncertainty.

3.5 | Spatial variability of CoV

Extending our analysis beyond the PSL, we examine in Figure 9 how sensitivity changes with location within the domain. Here we focus only on the annual wind speeds and their respective CoV metrics. The left map shows mean 130-m wind speeds, and the CoV is shown in the right map. The wind speed map is as expected, with greater wind speeds farther offshore with relative spatial homogeneity and lesser wind speeds onshore with more spatial variability. Interestingly, the offshore CoV is relatively low (up to 1.5%) compared to the onshore CoV (as high as 5%), which is likely caused by more complex terrain and greater variations and amplitudes of atmospheric stability relative to offshore. However, we note that the CoV by construction scales inversely with the mean wind speed, which is greater offshore; therefore, the contrast in CoV in Figure 9 is greater than what would be observed if the non-normalized STD metric was plotted.

Focusing just on the offshore sensitivity, we see that CoV is lowest close to the coast and increases farther offshore. This seems counterintuitive at first given that reduced influence of land–sea interaction should lead to lower variability in the modeled wind resource. Upon investigation, we find that this pattern is driven by the two atmospheric forcing products: ERAI and FNL. Far from the coast, the mean annual wind speeds from each product diverge. We speculate that this is because of sparse to no measurements far offshore, the decreased role of observational data assimilation in the NWP simulations, and therefore increased reliance on model physics. By contrast, close to the coast, where measurements are extensive and data assimilation more influential, we would expect the atmospheric forcing products to be in closer agreement.
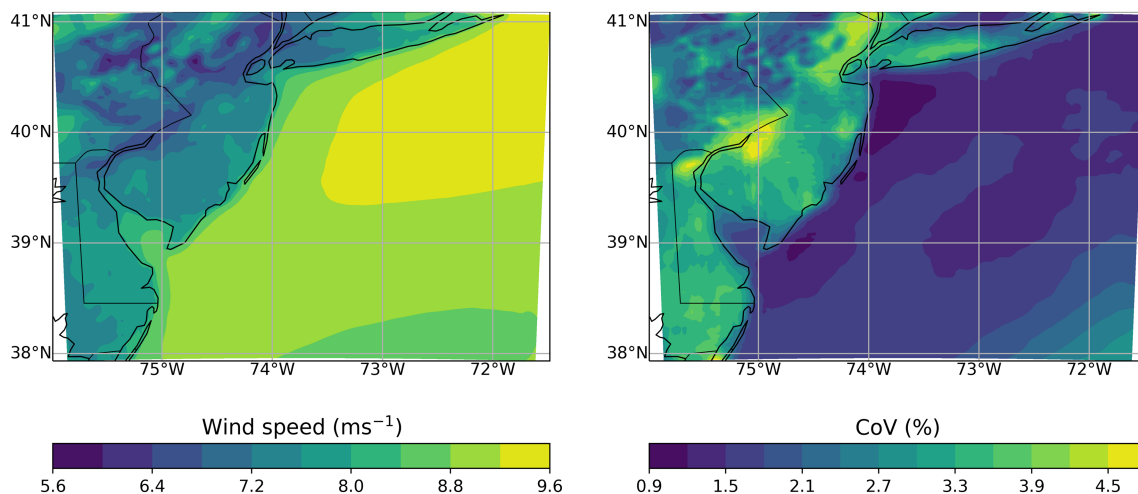
FIGURE 9 Maps of annual wind speeds (left) and associated CoV (right) over the 3-km simulation domain

3.6 | Attributing sensitivity to ensemble components

Our analysis to this point has focused on only the total sensitivity from the 24-member ensemble; however, the ability to attribute that total uncertainty to changes in different ensemble components is also important. For example, understanding the impact of changes to atmospheric or SST forcing would help to optimize ensemble selection and reduce the computational requirements or producing the ensemble runs.

We can think of attributing model sensitivity to ensemble components in two general ways:

1. *Contribution to total ensemble sensitivity*: In this context, the impact of an ensemble component is weighed against the impact of other ensemble components in relative terms (e.g., sensitivity of modeled wind speeds to SST product choice contributes to 10% of total ensemble variability on an annual scale).
2. *Sensitivity to single component*: In this context, the total sensitivity due to changes in a single ensemble component could be quantified (e.g., the choice of SST product results in a 1% CoV in modeled annual wind speeds).

Interpretations from these approaches may be quite different. For example, wind speeds may be somewhat sensitive to the choice of SST product; however, if that sensitivity is low compared to the sensitivity to the WRF version, then the contribution from SST may be low, or even near 0, even if its magnitude of sensitivity is moderate when considered in isolation.

We implement these two approaches to attributing sensitivity as follows:

Contribution to total ensemble sensitivity

1. Remove the target ensemble component from consideration (e.g., to assess atmospheric forcing contributions, first consider the set of ensemble members for which the forcing is set to the ERAI product)
2. Select a time scale and assess the sensitivity from that set of ensemble members resulting in Step 1 (e.g., the resulting ensemble yields a 3% sensitivity on an annual scale)
3. Calculate the difference between the total sensitivity for all 24 ensemble members to the sensitivity calculated in the previous step (e.g., 4% total sensitivity for all 24 ensemble members minus 3% when keeping ERAI product constant yields a 1% difference)
4. Repeat Steps 1 to 3 using the remaining choices of the target component (e.g., when fixing atmospheric forcing as the FNL product, Step 3 yields a difference value of 0.5%)
5. Take the average of the calculated differences and record the result (e.g., 1% and 0.5% differences yield a mean of 0.75% difference)
6. Repeat Steps 1–5 for all ensemble components
7. Once a table of differences has been produced, divide each value by the sum of differences across all ensemble components to arrive at a percentage contribution of each component to the total ensemble sensitivity
8. Repeat Steps 1–7 for all desired time scales.

An advantage of this approach is that we ensure a good number of ensemble members per distribution. For example, for the 24-member ensemble considered in this study, each set of sub-ensembles from which sensitivity is calculated will have at least 8 and as many as 12 members. Therefore, CoV metrics will generally be more robust than those calculated with fewer data points.

Sensitivity to single component

1. For a target ensemble component, select a time scale and construct a set of ensembles by allowing variation in the target component while keeping other components constant (e.g., to assess sensitivity from atmospheric product choice, keep RU-WRF namelist, WRF Version 4.0, and NCEP RTG SST product fixed while varying atmospheric forcing, thereby constructing a two-member ensemble)
2. Calculate and record the sensitivity metric for the ensemble constructed in Step 1
3. Repeat Steps 1 and 2 by cycling through every possible combination of the fixed ensemble components (e.g., assessing atmospheric product sensitivity would yield 12 different ensembles) (i.e., two WRF namelists × two WRF versions × three SST products, each with two members)
4. Calculate and record the mean sensitivity from all ensemble distributions created in Step 3
5. Repeat Steps 1 to 4, setting each ensemble component as the target component
6. Repeat Steps 1 to 5 for each desired time scale

We compare these two approaches in Figure 10 and make the following observations:

- Sensitivity for all ensemble components decreases with increasing time scale, as we would expect
- The WRF namelist ensemble component consistently and considerably contributes to ensemble sensitivity across all time scales
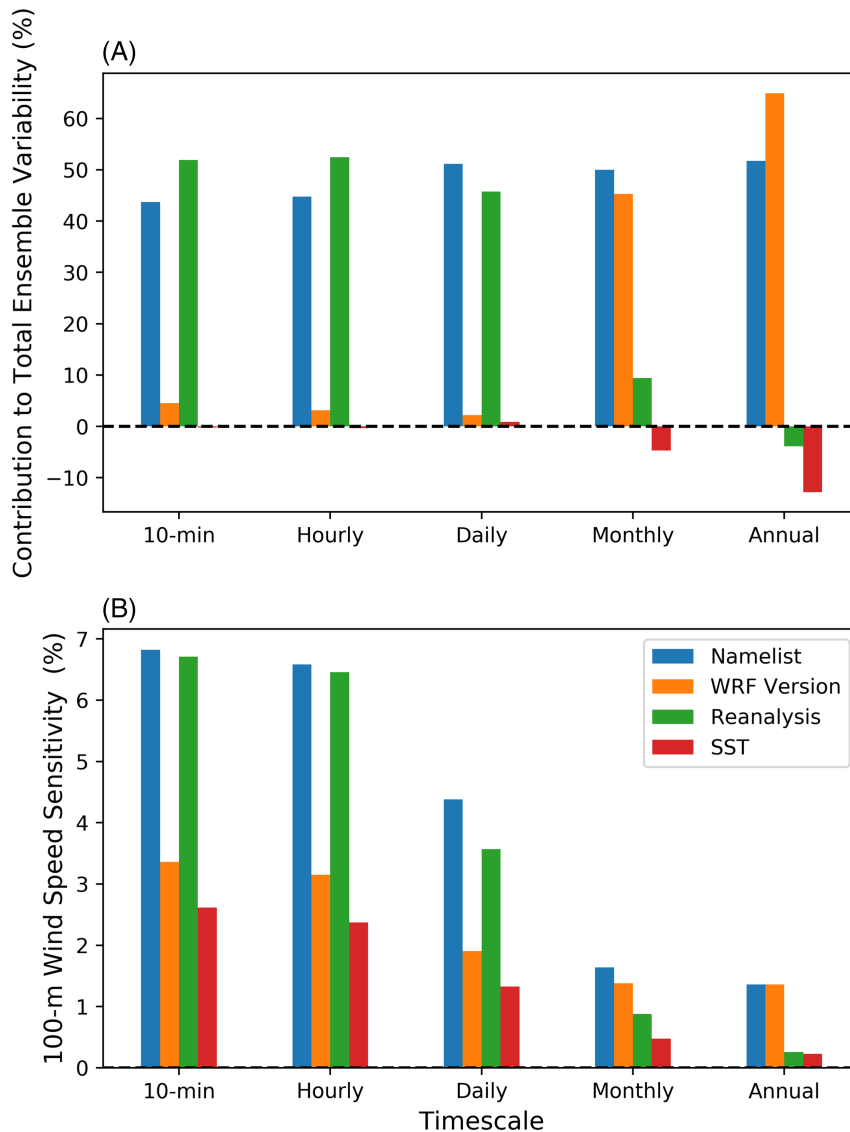
FIGURE 10 Impact of different sources of uncertainty on the CoV at different time scales. Results are based on data from the PSL

- The choice of atmospheric forcing drives larger sensitivity at short time scales; however, its contribution decreases at higher time scales and is negligible at annual scales. We speculate that this trend is because different atmospheric forcing products may often be "out of phase" (i.e., a relative delay in modeled events) on shorter time scales, which would likely contribute considerably to total ensemble sensitivity. However, on longer time scales, those phase differences would average out, and the different atmospheric forcing products would tend to make similar predictions
- The WRF version contributes less to total ensemble sensitivity on short time scales but becomes much more important on longer time scales. We speculate that this trend has more to do with the contributions from atmospheric forcing decreasing with longer time scales and the contributions from other components increasing as a result
- The choice of SST product has negligible impact across all time scales

    We notice in some cases that contributions are actually negative (e.g., SST product on monthly and annual time scales). These results are not realistic but are more a consequence of low sensitivity overall at longer time scales and limited data sample sizes. These impacts are highlighted in Figure 11, where we plot annual wind speeds from each ensemble member but color coded by the different namelist/WRF version/atmospheric forcing combinations. Each color, then, represents different SST sensitivity calculations when keeping all other ensemble components constant. We see that SST sensitivity is very low compared to sensitivity of the ensemble as a whole. Because of the limited number of data points, it turns out that keeping SST product constant actually leads to higher ensemble uncertainty (1.33%) compared to
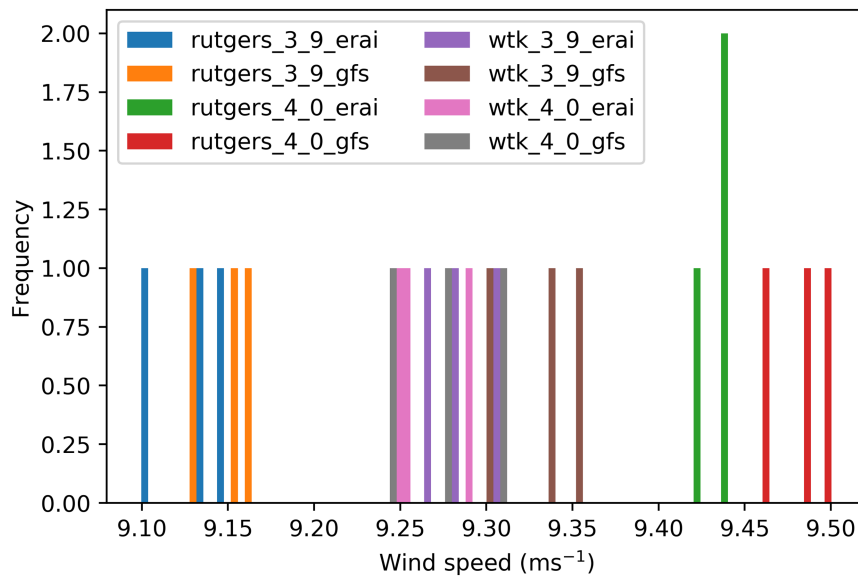
FIGURE 11 Histogram of mean wind speeds calculated from sub-ensembles where all but the SST product is kept constant. Different colors denote the different combinations of WRF namelist, WRF version, and atmospheric forcing products that are kept constant

sensitivity across the whole 24-member ensemble (1.29%). This small difference is sufficient to produce the negative contribution values shown in Figure 10A.

Finally, we note that using a non-normalized sensitivity metric (i.e., in m s^{-1} rather than percentage-based units) resulted in negligible differences in the relative contributions from different ensemble categories. Therefore, it appears that these contributions are not weighted substantially by wind speed regime but rather apply rather well across the whole wind speed distribution.

4 | CONCLUSIONS

We have explored in this analysis best practices for quantifying and disseminating the sensitivity in NWP-modeled wind speeds in the U.S. offshore North Atlantic. The basis of our analysis was a 24-member ensemble of WRF simulations capturing variations in the WRF namelist, WRF version, atmospheric forcing product, and SST product. Through this analysis we arrived at several key conclusions:

- The STD metric is more stable and less "noisy" hour to hour compared to the IQR
- Normalizing STD by mean wind speed to calculate the CoV results in a percentage-based metric that is generally more familiar and easily interpretable. Practically, the choice between STD and CoV is not all that important given that they can readily be calculated from one another provided wind speed data are available
- Sensitivity in wind speeds can vary considerably with height above the surface. Specifically, lower level winds are more sensitive than upper level winds, which we attribute to increased influence of model physics near the surface and variations of those physics within the considered ensemble
- Sensitivity decreases with increased time scale, as we would expect given the fact that model differences tend to average out over longer time scales
- We can attribute total ensemble sensitivity to different ensemble components using two main approaches: one that computes the importance of a model component relative to others, and another approach that simply assesses wind speed sensitivity from variations in individual ensemble components
- The relative attributions from different ensemble components vary with time scale; for example, the choice of atmospheric forcing is very influential at short time scales because of "out-of-phase" differences between the modeled data sets, whereas these differences tend to average out on longer time scales.

The fact that NWP-modeled wind speeds are sensitive to model setup is of course well established in the literature; this study advances the state of the art, however, by proposing methods to quantify and disseminate that sensitivity. Such dissemination may have considerable

implications for wind energy analyses that have historically unaccounted for this source of uncertainty but instead taken NWP modeled data as "truth." The propagation of the wind speed sensitivity into AEP estimates, grid integration studies, wind power forecasts, and levelized cost of energy analyses is a critical area of future work that was not explored in this study. Tracking NWP model sensitivity through these analysis pipelines and providing confidence bounds on key energy financial calculations will better inform decision-making across a range of stakeholders including policy makers, utilities, regulatory agencies, developers, and investors. We will explore such propagation of sensitivity in follow-up studies.

This sensitivity analysis was not exhaustive, and therefore, the actual magnitudes of sensitivity presented here should be interpreted cautiously. First, two WRF namelists were contrasted without a more detailed analysis of sources of variability within those namelists (e.g., PBL scheme and vertical resolution). Second, only two reanalysis products were considered, which are somewhat outdated; the impact on CoV from the use of more accurate, modern reanalysis products such as the European Centre for Medium-Range Weather Forecasts's ERA5 or the National Aeronautics and Space Administration's MERRA-2 is an important area of future work. Third, assessing NWP sensitivity in all U.S. offshore wind energy lease areas outside this study domain (and especially the PSL) will allow for more insight into how geography influences both the magnitude of variability on different time scales and the contribution of ensemble components to total variability.

Despite the non-exhaustive sensitivity analysis, it is our hope that the work presented here provides a foundation for quantifying and disseminating our confidence in NWP-modeled wind speeds in cases where we lack sufficient hub height observations and therefore cannot robustly validate model results. Observational data sparsity will always hinder accurate offshore wind resource assessments, despite the emergence of floating lidar. Furthermore, ensemble-based wind resource modeling is becoming increasingly common and more accessible given ever-increasing computing capacity. The ability to assess confidence in NWP-modeled wind speeds through an "observation-free" approach, as presented here, should provide a benefit to the emerging U.S. offshore wind industry.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/we.2611.

ORCID

Mike Optis https://orcid.org/0000-0001-5617-6134
Joseph Brodie https://orcid.org/0000-0003-1041-5077
Travis Miles https://orcid.org/0000-0003-1992-0248

REFERENCES

1. Bureau of Ocean Energy Management. Outer continental shelf renewable energy leases map book, Bureau of Ocean Energy Management; 2018. https://www.boem.gov/sites/default/files/renewable-energy-program/Mapping-and-Data/Renewable_Energy_Leases_Map_Book_March_2019.pdf
2. Clifton A, Smith A, Fields M. Wind plant preconstruction energy estimates, Current practice and opportunities; 2016. https://www.nrel.gov/docs/fy16osti/64735.pdf
3. Carbon Trust Offshore Wind Accelerator. Carbon trust offshore wind accelerator roadmap for the commercial acceptance of floating lidar technology, Carbon Trust; 2018. https://prod-drupal-files.storage.googleapis.com/documents/resource/public/Roadmap%20for%20Commercial%20Acceptance%20of%20Floating%20LiDAR%20REPORT.pdf
4. OceanTech Services/DNV GL. NYSERDA floating lidar buoy data. https://oswbuoysny.resourcepanorama.dnvgl.com
5. Durakovic A. USWind schedules met Mast Installation Offshore Maryland. https://www.offshorewind.biz/2019/05/20/us-wind-schedules-met-mast-installation-offshore-maryland/; 2019.
6. Brower M. *Wind resource assessment: a practical guide to developing a wind project*. Hoboken, New Jersey: John Wiley & Sons; 2012.
7. Mahoney WP, Parks K, Wiener G, et al. A wind power forecasting system to optimize grid integration. *IEEE Trans Sustain Energy*. 2012;3(4):670-682. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6237561
8. Jong P, Dargaville R, Silver J, Utembe S, Kiperstok A, Torres EA. Forecasting high proportions of wind energy supplying the Brazilian Northeast electricity grid. *Appl Energy*. 2017;195:538-555. http://www.sciencedirect.com/science/article/pii/S0306261917302726
9. Hasager CB, Madsen PH, Giebel G, et al. Design tool for offshore wind farm cluster planning. In: Proceedings of the EWEA Annual Event and Exhibition 2015. European Wind Energy Association (EWEA). Hamburg, Germany; 2015.

10. Hale E. The uncertainty of uncertainty. https://www.nrel.gov/wind/assets/pdfs/session_2_hale_uncertainty_of_uncertainty.pdf; 2015.

11. Ruiz JJ, Saulo C, Nogués-Paegle J. WRF model sensitivity to choice of parameterization over South America: Validation against surface variables. *Mon Weather Rev*. 2010;138(8):3342-3355. https://doi.org/10.1175/2010MWR3358.1

12. Carvalho D, Rocha A, Gómez-Gesteira M, Santos CS. Sensitivity of the WRF model wind simulation and wind energy production estimates to planetary boundary layer parameterizations for onshore and offshore areas in the Iberian Peninsula. *Appl Energy*. 2014a;135:234-246. https://doi.org/10.1016/j.apenergy.2014.08.082

13. Gómez-Navarro JJ, Raible CC, Dierer S. Sensitivity of the WRF model to PBL parametrisations and nesting techniques: evaluation of wind storms over complex terrain. *Geosci Model Dev*. 2015;8(10):3349-3363. https://doi.org/10.5194/gmd-8-3349-2015

14. Hahmann AN, Vincent CL, Peña A, Lange J, Hasager CB. Wind climate estimation using WRF model output: method and model sensitivities over the sea. *Int J Climatol*. 2015;35(12):3422-3439. https://doi.org/10.1002/joc.4217

15. Olsen BT, Hahmann AN, Sempreviva AM, Badger J, Jørgensen HE. An intercomparison of mesoscale models at simple sites for wind energy applications. *Wind Energy Scie*. 2017;2(1):211-228. https://doi.org/10.5194/wes-2-211-2017

16. Siuta D, West G, Stull R. WRF hub-height wind forecast sensitivity to PBL scheme, grid length, and initial condition choice in complex terrain. *Weather Forecast*. 2017;32(2):493-509. https://doi.org/10.1175/WAF-D-16-0120.1

17. Carvalho D, Rocha A, Gómez-Gesteira M, Silva Santos C. WRF wind simulation and wind energy production estimates forced by different reanalyses: comparison with observed data for Portugal. *Appl Energy*. 2014b;117:116-126. https://doi.org/10.1016/j.apenergy.2013.12.001

18. Carvalho D, Rocha A, Gómez-Gesteira M, Santos CS. Offshore wind energy resource simulation forced by different reanalyses: comparison with observed data in the Iberian Peninsula. *Appl Energy*. 2014c;134:57-64. https://doi.org/10.1016/j.apenergy.2014.08.018

19. Ulazia A, Saenz J, Ibarra-Berastegui G. Sensitivity to the use of 3DVAR data assimilation in a mesoscale model for estimating offshore wind energy potential. A case study of the Iberian northern coastline. *Appl Energy*. 2016;180:617-627. https://doi.org/10.1016/j.apenergy.2016.08.033

20. Hahmann AN, Sile T, Witha B, et al. The making of the New European Wind Atlas, Part 1: model sensitivity. *Geosci Model Dev Discuss*. 2020;2020: 1-33. https://doi.org/10.5194/gmd-2019-349

21. Dörenkämper M, Olsen BT, Witha B, et al. The making of the New European Wind Atlas—part 2: production and evaluation. *Geosci Model Dev*. 2020; 2020:5079-5102. https://doi.org/10.5194/gmd-13-5079-2020

22. Skamarock WC, Klemp JB, Dudhia J, et al. A description of the Advanced Research WRF Version 3. NCAR/TN-475+STR, Boulder, Colorado, USA, Mesoscale and Microscale Meteorology Division, National Center for Atmospheric Research; 2008.

23. Draxl C, Clifton A, Hodge B-M, McCaa J. The Wind Integration National Dataset (WIND) toolkit. *Appl Energy*. 2015;151:355-366. https://doi.org/10.1016/j.apenergy.2015.03.121

24. Glenn S, Dunk R. An Advanced Atmospheric/Ocean Assessment Program designed to reduce the risks associated with offshore wind energy development defined by the NJ Energy Master Plan and the NJ Offshore Wind Energy Economic Development Act. Final Report, Trenton, NJ, USA, New Jersey Board of Public Utilities; 2013. https://rucool.marine.rutgers.edu/media/bpu/RU-BPU-Offshore-Study-Final-Report-June21-FINAL.pdf

25. Dee DP, Uppala SM, Simmons AJ, et al. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q J R Meteorol Soc*. 2011;137(656):553-597. https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.828

26. White G, Yang F, Tallapragada V. The development and success of NCEP's Global Forecast System. National Center for Environmental Prediction. https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-forcast-system-gfs; 2018.

27. Glenn S, Schofield O. Observing the oceans from the COOL room: our history, experience, and opinions. *Oceanogr*. 2003;16(4):37-52. http://archipelago.uma.pt/pdf_library/Glenn&Schofield_2003_Oceanography.pdf

28. Glenn S, Arnone R, Bergmann T, et al. Biogeochemical impact of summertime coastal upwelling on the New Jersey Shelf. *J Geophys Res Oceans*. 2004; 109(12):1-15.

29. Gaertner E, Rinker J, Sethuraman L, et al. Definition of the IEA 15 MW offshore reference wind turbine. NREL/TP-75698, International Energy Agency; 2020.

30. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*. 2020;17:261-272.