https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

Methods for high-accuracy wind resource assessment to support distributed wind turbine siting

Kevin Menear[1], Sameer Shaik[1], Lindsay Sheridan[2], Dmitry Duplyakin[1], and Caleb Phillips[1]

[1]National Laboratory of the Rockies, Golden, CO, USA
[2]Pacific Northwest National Laboratory, Richland, WA, USA

Correspondence: Kevin Menear (kevin.menear@nlr.gov)

Abstract. Public wind resource datasets are central to wind energy planning, particularly for distributed wind installations where it may be infeasible to collect on-site measurements or run bespoke simulations. Yet despite their broad use, the site-level accuracy at hub height remains only partially quantified. This work addresses this gap in two steps: (i) we develop a unified, observation-based benchmark to evaluate the performance of the most common models used in industry, and (ii) we propose a new machine-learned ensemble approach that leverages multiple models to synthesize improved estimates that address the shortcomings of individual models. For each dataset and observation series we form long-term empirical wind speed quantiles. This quantile representation allows us to compare products with different periods of record without requiring temporal overlap and evaluate both wind speed distribution errors and site-level mean biases. Results show that the ensemble method reduces quantile-dependent mean bias to near zero across the distribution and lowers mean absolute bias in long-term mean wind speed by roughly one-third relative to the best-performing individual dataset. Finally, we use the trained model to produce a national, gridded set of wind speed quantiles for the publicly accessible WindWatts platform. Together, the benchmark, ensemble model, and deployment dataset demonstrate that machine learning can meaningfully correct and combine existing public datasets, providing more reliable, distributional wind resource information for early-stage assessment and planning.

1 Introduction

Accurate characterization of wind resources is fundamental for siting, financing, and operating wind energy projects. This is especially true for distributed and community-scale deployments, where projects are smaller, margins are tighter, and on-site measurement campaigns are often prohibitively expensive. In these settings, practitioners typically rely on public wind resource datasets to estimate long-term wind conditions at prospective sites. These products offer broad spatial coverage at low marginal cost, but they are not designed to be perfectly unbiased at arbitrary locations. Instead, they reflect a combination of model physics, parameterizations, historical forcings, and post-processing choices and therefore exhibit spatially structured errors at the site level.

These errors matter in practice. Biases of order 0.5 to $1\ \mathrm{m\,s^{-1}}$ in long-term mean wind speed are common in complex terrain and coastal zones (Sheridan et al., 2025), and even smaller systematic errors can translate into substantial changes in annual energy production and project economics through nonlinear turbine power curves. Moreover, the accuracy of public

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

datasets is not uniform: performance varies across regions, heights, and products. Developers and planners typically have limited guidance on which products are most reliable in a given regime or how to combine multiple datasets and ancillary information to obtain a more faithful site-level characterization. Prior work has also shown that the limitations of these models are not easily correctable with *ex post facto* accounting for impacts of built obstacles and vegetation (Phillips et al., 2024) and that bias correction with measurements is the most promising method to improve performance in practice (Phillips et al., 2022).

The current practice of desktop assessment has several limitations in this regard. First, many workflows reduce each dataset to a single scalar summary per site, typically a long-term mean wind speed or a coarse climatology (e.g., a 12×24 matrix of monthly hour-of-day means). Such summaries discard information about the full wind speed distribution, including tails and intermittency, which directly affect annual energy production. Mapping these reduced statistics through nonlinear turbine power curves can amplify biases, particularly in regions where the distribution is skewed or multi-modal. Second, assessments often rely on a single preferred dataset without a systematic, observation-based comparison against alternatives. Third, although many studies have evaluated individual datasets at subsets of towers or regions, there is no unified benchmark that compares the major publicly available wind resource products across a large, diverse set of U.S. sites and heights using a consistent methodology.

At the same time, the modeling community has developed increasingly powerful tools that could, in principle, learn systematic corrections to these products. Gradient-boosted trees and related machine-learning (ML) methods can flexibly combine multiple datasets, topographic descriptors, and other static fields. However, most existing work with such models has focused on corrections at a single dataset rather than learning the *full* wind speed distribution in a way that remains tightly anchored to observations and is directly usable for energy assessment. There is a gap between, on one hand, the wealth of public wind resource datasets and observational archives now available over the United States, and, on the other hand, the practical need for a validated, distribution-aware, site-level resource product suitable for both research and deployment.

This work addresses these gaps with two key contributions. First, we construct a comprehensive, observation-based benchmark of leading public wind resource datasets across the continental United States. We assemble a large observational archive that combines long-term 10 m Automated Surface Observing System (ASOS) stations from the Integrated Surface Database with a curated set of tall Gold Standard (GS) sites spanning typical hub-height regimes. For each observation site and height, we compute empirical wind speed quantiles from observations and from all time series–based products: European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5), (Hersbach et al., 2020), Wind Integration National Dataset (WIND) Toolkit (WTK) (Draxl et al., 2015), WIND Toolkit Long-term Ensemble Dataset Climate (WTK-LED Climate) (Draxl et al., 2024) and contiguous United States (CONUS) (WTK-LED CONUS) (Draxl et al., 2024), National Oceanic and Atmospheric Administration (NOAA) High-Resolution Rapid Refresh (HRRR) (Dowell et al., 2022), and Technical University of Denmark's Global Wind Atlas (GWA) (Davis et al., 2023). GWA provides only mean wind speeds, so it enters the benchmark as a static, single-value reference rather than a source of quantiles. Using numerical integration of quantile curves, we obtain consistent long-term mean wind speeds for all datasets. This unified representation supports direct, site-by-site comparison of bias, absolute bias, and error metrics across products, heights, and regions.

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

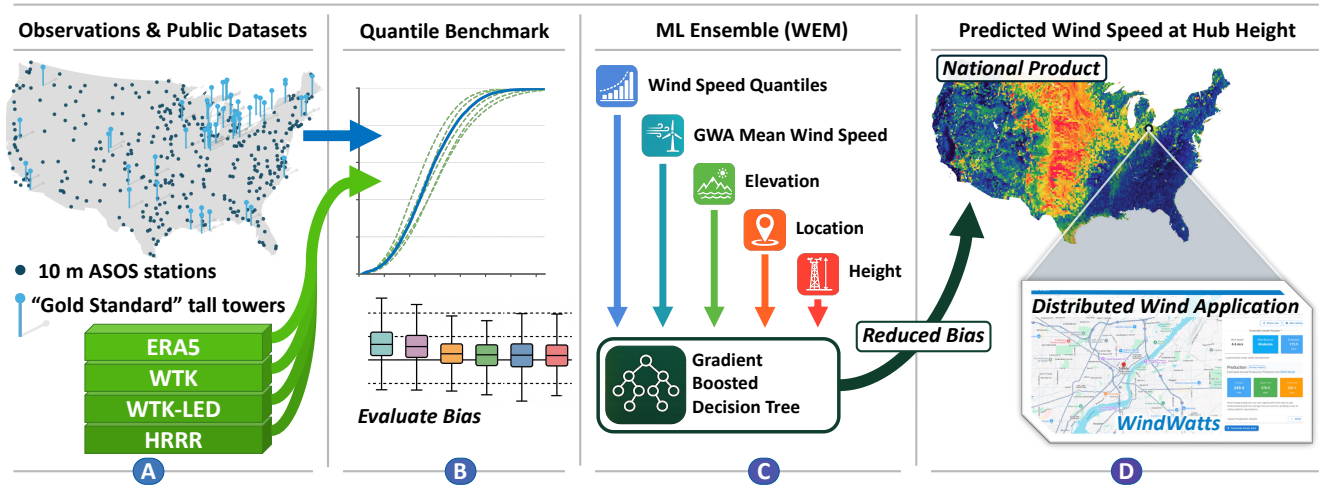


Figure 1. Overview of the workflow used in this study. *Panel A* shows the observational archive of 10 m ASOS stations and Gold Standard (GS) tall towers together with the public wind resource datasets (e.g., ERA5, WTK, WTK-LED variants, and HRRR). For clarity, only 20 % of the sites are included here. See Figs. 2 and 3 for full maps including all ASOS and GS sites. *Panel B* illustrates the quantile-based benchmark, in which empirical wind speed quantiles are formed for each dataset and site and used to evaluate long-term bias and distributional errors. *Panel C* depicts the WindWatts Ensemble Model (WEM), a gradient-boosted decision tree that takes wind speed quantiles, Global Wind Atlas (GWA) mean wind speed, elevation, location, and height as predictors and learns quantile distributions. *Panel D* shows the resulting national gridded product of predicted hub-height wind speed, which is served through the WindWatts web application for distributed wind siting and early-stage assessment.

Second, we leverage insights from this benchmark to develop the WindWatts Ensemble Model (WEM), an ML ensemble that predicts *quantile distributions* of wind speed at site level. Rather than predicting a single mean value, WEM learns a mapping from a feature set that includes multiple public datasets and latitude, longitude, height above ground, and topographic descriptors derived from the U.S. Geological Survey (USGS) 3D Elevation Program (3DEP) (elevation, slope, aspect) to the full wind speed quantile curve at each site and height. We design a geographically informed leave-one-out cross-validation scheme that withholds each GS site and all sites within 10 km from training, and we use this strict holdout to evaluate how well WEM and each dataset generalize to unseen locations. Finally, we retrain WEM on all available data and apply the model to every point on the ERA5 grid, producing a national gridded quantile field that can be explored via maps and a web application and used as a practical tool for desktop wind resource assessment. Figure 1 summarizes this end-to-end workflow, from observational data and public wind resource products through the quantile benchmark and WEM ensemble to the final distributed wind deployment.

The remainder of the paper is organized as follows. Section 2 reviews prior work on public wind resource datasets, bias correction, and distributional modeling and situates our contribution in that context. Section 3 describes the observational archive and proposed quantile data representation. Section 4.1 details the methodology for evaluating public wind resource

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

datasets, including temporal alignment of observations and model fields, height matching and vertical interpolation, and error metrics. Section 4.2 formulates the ML problem and outlines the WEM model architecture and cross-validation scheme. Section 5.1 summarizes the benchmark errors across datasets, heights, and site types and highlights spatial patterns of bias. Section 5.2 reports the performance of WEM relative to each baseline dataset at the quantile and site levels. Section 6 describes the full-grid application of the final WEM model over the ERA5 grid and the resulting national quantile field. Finally, Sect. 7 discusses the implications of our findings for wind resource assessment and limitations and opportunities for future work. We conclude with a summary of key results and a discussion of reproducibility and data availability.

2 Background and related work

Accurate wind resource assessment is the critical first step in the development of any wind energy project, governing everything from turbine selection to financial viability. In utility-scale development, this risk is typically mitigated through extensive on-site measurement campaigns using meteorological towers or lidar, which can cost upwards of $60,000 per site (Vasiljević et al., 2020). However, for distributed wind and community-scale projects, typically defined as individual or small clusters of turbines installed near the point of end use, such costs are often prohibitive relative to the total capital expenditure (Sheridan et al., 2024). Consequently, distributed wind developers rely disproportionately on "desktop" studies, utilizing virtual data from public models (Sheridan et al., 2024).

This reliance on modeled data introduces significant risk. The nonlinear relationship between wind speed and power generation, where power is proportional to the cube of the wind speed ($P \propto u^3$), means that seemingly minor systematic errors in mean wind speed estimates can result in disproportionately large errors in annual energy production estimates (Indra et al., 2014). While a bias of $0.5\ \mathrm{m\,s^{-1}}$ might be tolerable for initial screening, it can shift the levelized cost of energy enough to render a marginal distributed project financially unfeasible (Drew et al., 2015). Therefore, the availability of low-bias, site-specific, and distributionally accurate public datasets is a prerequisite for the scalable deployment of distributed wind.

To meet the need for broad spatial coverage, the wind energy community typically relies on a hierarchy of modeled datasets, ranging from global reanalyses to high-resolution mesoscale and microscale products. Global reanalysis products, such as the ERA5 dataset, serve as the standard boundary condition for the industry (Hersbach et al., 2020). ERA5 assimilates vast quantities of historical observations to produce a temporally consistent record of the atmospheric state (Hersbach et al., 2020). However, with a native horizontal resolution of approximately 31 km, ERA5 cannot resolve local topographic features, such as ridges or valleys, which frequently drive the wind resource at the site level (Olauson, 2018).

To bridge this scale gap, mesoscale numerical weather prediction models downscale global reanalyses using physics-based solvers like the Weather Research and Forecasting (WRF) model (Powers et al., 2017). Prominent examples include the WTK (Draxl et al., 2015) and NOAA's HRRR model (Dowell et al., 2022). By increasing resolution to the 2 to 4 km range, these models better capture mesoscale flow phenomena. Nevertheless, they remain prone to systematic biases arising from planetary boundary layer parameterization schemes, land-surface smoothing, and imperfect initial conditions (Carvalho et al., 2014). At the finest scale, microscale models such as the GWA employ linear flow models (e.g., WAsP) to downscale mesoscale

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

data to resolutions as fine as 250 m (Davis et al., 2023). Crucially, these models represent distinct physical formulations and parameterizations, often resulting in uncorrelated error structures. The diversity of their errors presents an unexploited opportunity: rather than selecting a single "best" model, the datasets can be treated as complementary signals of the underlying wind resource.

Given the inevitability of bias in gridded datasets, practitioners have long employed statistical correction methods. The industry standard has traditionally been the measure-correlate-predict (MCP) method, which applies linear regression to calibrate a single long-term reference (typically a reanalysis product or nearby station) against short-term on-site data (Rogers et al., 2005). In recent years, ML has emerged as a powerful alternative for this calibration task (Haq et al., 2025). Nonlinear algorithms, including gradient boosted machines (Xu et al., 2020), have demonstrated superior ability to map coarse model outputs and topographic features to site-specific wind speeds. However, the vast majority of these studies focus on *single-source bias correction*: they utilize ML to learn a transfer function f that maps a single input dataset (e.g., ERA5) to ground truth (Hu et al., 2023). While this improves upon linear MCP, it remains fundamentally limited by the information content and structural errors of the chosen base model (Vannitsem et al., 2021).

This work departs from the single-source paradigm by treating public wind datasets not as baselines to be corrected, but as input features for a *multi-source synthesis*. Multi-fidelity learning and ensemble stacking techniques have shown promise in other domains of computational physics (Peherstorfer et al., 2018), yet they are rarely applied to wind resource assessment beyond simple weighted averaging or basic ensembles (He et al., 2023). By feeding multiple distinct wind products (e.g., ERA5, WTK, HRRR) into a single regressor, the ML model can learn to dynamically weigh the inputs based on the regime (Abad-Santjago et al., 2025). This approach shifts the goal from "correcting a map" to "synthesizing a new map" that exploits the collective intelligence of the existing public data ecosystem.

A further limitation in many ML-based workflows is the reduction of the wind resource to a single scalar summary, typically the long-term mean wind speed (Amato et al., 2022), or the computationally expensive prediction of high-frequency time series (Stengel et al., 2020). Correcting the mean does not guarantee that the shape of the distribution is preserved, and predicting hourly time series is prone to "double penalty" timing errors that are irrelevant for long-term planning (Haben et al., 2014).

Quantile mapping, a technique used in climate science to bias-correct global climate models (Cannon et al., 2015), addresses this by adjusting the cumulative distribution function directly. However, its application in wind energy has largely been limited to univariate transformations rather than supervised learning targets (Zscheischler et al., 2019). By combining the multi-source fusion approach described above with a quantile-based objective, our framework learns to predict the full wind speed probability distribution. This allows the model to correct regime-dependent biases, such as the underprediction of extreme wind events, resulting in a synthesized dataset that is robust not just for mean wind speed, but for the energy-producing tails of the distribution.

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

3 Data

Our analysis combines: (i) an observational archive spanning near-surface and hub-height measurements across the United States, and (ii) a suite of public wind resource datasets and ancillary fields that are co-located with those observations and later applied over the full ERA5 grid. All sources are processed into a common quantile-based representation that underpins both the benchmarking and ML components of this work.

3.1 Observational archive

We assemble two complementary observational datasets that together span a wide range of locations, exposure conditions, and measurement heights: surface winds from ASOS stations at nominal 10 m height, and a curated set of tall GS sites closer to typical turbine hub heights.

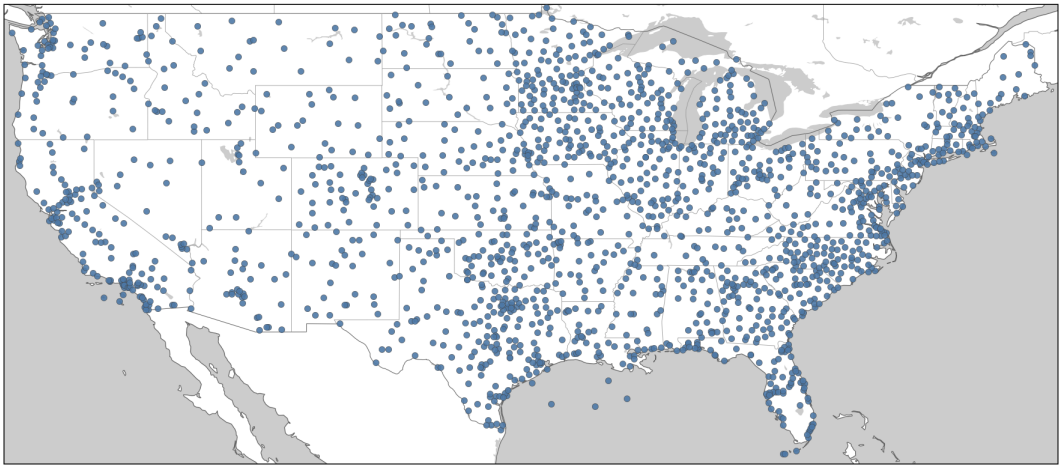


Figure 2. Spatial distribution of the 1,842 ASOS stations used in this study. Points show land-based stations with at least one year of sufficiently complete hourly wind records between 2007 and 2024 within the contiguous United States and adjacent coastal regions.

The first component of the observational archive consists of 10 m surface wind measurements from the ASOS, obtained from the NOAA National Centers for Environmental Information Integrated Surface Database (ISD). We retrieve all hourly U.S. wind records from 2007 to 2024 and restrict attention to land-based stations within the contiguous United States and nearby coastal regions. Each station's wind speed is parsed from the ISD `WND` field and converted to units of meters per second. The resulting network of 1,842 stations (Fig. 2) provides dense, quasi-uniform coverage of the CONUS domain, with especially high station density in the eastern U.S. and along major coastlines.

ISD reports wind speed in integer knots (kt), and this discretization is especially pronounced for light winds: any true wind speed in the interval $(0, 2]$ kt is recorded as *0 kt*. To avoid treating all low-speed conditions as identical, a distortion that would otherwise flatten the lower tail of the wind speed distribution and bias empirical quantiles, we apply a de-quantization step. Calm reports (0 kt) are replaced by samples drawn uniformly on $(0, 2]$ kt, reflecting the full range of physical speeds that map to

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

a 0 kt observation. For reported speeds $k \geq 3$ kt, we draw uniformly on $(k-1, k]$ kt, consistent with the integer-knot rounding in ISD. Converting from knots to meters per second produces a continuous, non-discretized distribution of hourly wind speeds.

To construct reliable long-term climatologies, we evaluate completeness at the station-year level. For each station, we identify the nominal sampling frequency and compute the fraction of expected observations present in each year. Only years with at least 95 % completeness are retained, and stations without any valid years over the analysis window are removed. After applying all filters, the ASOS archive provides a large set of consistent, long-term 10 m wind records that anchor the near-surface portion of our benchmark.

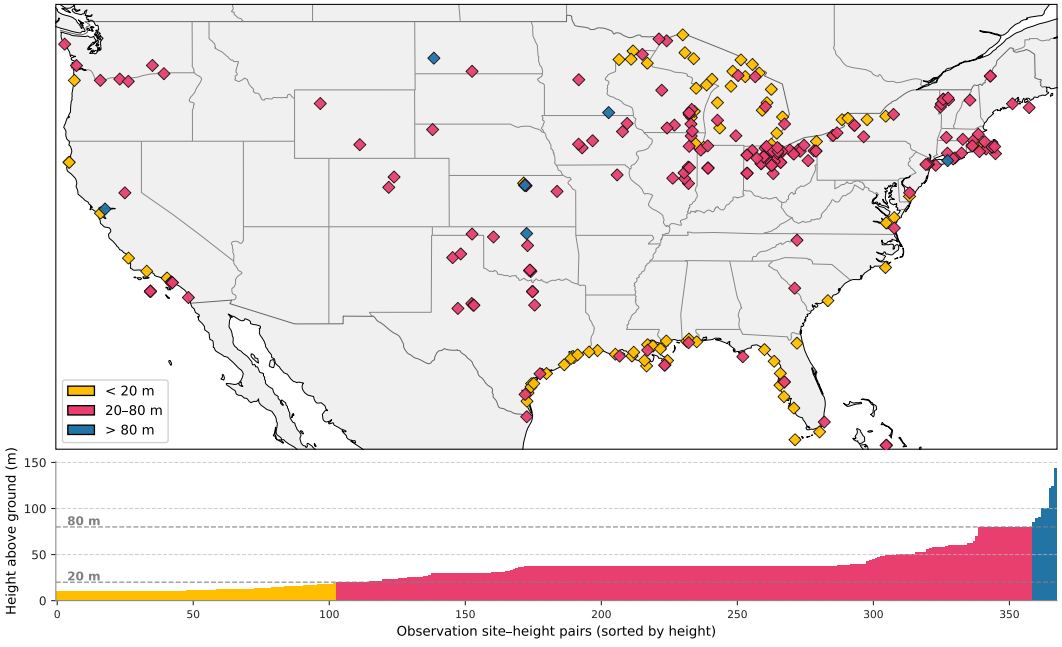


Figure 3. Spatial distribution and measurement heights of GS wind sites. Each diamond denotes a site contributing at least one quality-controlled wind speed time series at or above 10 m. Marker colors indicate the height of the highest anemometer at each site: yellow for < 20 m, magenta for 20 to 80 m, and blue for > 80 m. Relative to the dense ASOS network, GS sites are fewer in number but targeted toward wind-rich and development-relevant regions, and they collectively sample a mix of near-surface measurements, typical onshore hub heights, and very tall masts above 80 m.

The second component of the observational archive is a high-quality set of meteorological tower and commercial wind measurement sites, collectively referred to as GS sites. This dataset is designed to capture hub-height wind characteristics that cannot be inferred from 10 m ASOS stations alone. The GS collection comprises 178 meteorological towers and 128 de-identified commercial wind locations, yielding a total of 306 distinct sites and 368 unique height–location combinations. Compared with the dense ASOS network, GS sites are more sparsely distributed and tend to cluster in wind-relevant regions such as the central Plains, the Midwest, and selected coastal and complex-terrain areas, reflecting historical measurement campaigns and project development activity (Fig. 3). Most sites have measurement heights in the mid-range (20 to 80 m),

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

which are particularly relevant for distributed wind applications, whereas fewer sites sample only near-surface levels (< 20 m) or very tall levels (> 80 m).

All GS time series are subjected to the same completeness and integrity checks applied to ASOS, including year-level completeness thresholds and additional quality-control filters based on campaign metadata and automated screening. The combination of broader geographic coverage from ASOS and targeted hub-height measurements from the GS archive provides a robust observational foundation for both benchmarking public wind datasets and training the ML quantile model.

Table 1. Summary of wind resource datasets used in this study.

	Time Range		Resolution			
Model	**Start**	**End**	**Spatial**	**Temporal**	**Heights (m)**	**Coverage**
ERA5	1940	present	31 km	1 hour	10, 100	Global
WTK	2007	2013	2 km	5 minute	10, 40–200 in increments of 20	CONUS
WTK-LED Climate	2001	2020	4 km	1 hour	WTK heights & 250, 300, 500, 1,000	North America
WTK-LED CONUS	2018	2020	2 km	5 minute	WTK heights & 250, 300, 500, 1,000	CONUS
HRRR	2014	present	3 km	1 hour	10, 80	CONUS & Alaska
GWA	2008	2017	250 m	N/A	10, 50–250 in increments of 50	Global

3.2 Wind resource datasets

We evaluate six public wind resource products that span a range of spatial scales, modeling approaches, and reference periods: the global reanalysis ERA5, WTK, two WIND Toolkit Long-term Ensemble Datasets (WTK-LED Climate and WTK-LED CONUS), the HRRR model, and the GWA. Table 1 summarizes their time ranges, native resolutions, and geographic coverage. All gridded fields were obtained directly from the respective public archives (ECMWF Climate Data Store, National Laboratory of the Rockies data services, NOAA model archives, and the GWA portal) using scripted, reproducible downloads restricted to the spatial domain and periods relevant to this study. For HRRR, we used a version of the dataset that was regridded to the WTK grid (Buster et al., 2024).

3.3 Topography features

Local terrain strongly influences near-surface and hub-height winds in ways that are not fully resolved by gridded reanalyses or mesoscale models. To capture these effects, we attach static topographic attributes to every observational site and, later, to every point on the ERA5 grid used for full-domain inference. Terrain fields are sourced from the USGS 3DEP via the National Map Elevation ImageServer. For each unique site or grid-point coordinate, we query the service to obtain three quantities: *elevation*, defined as the bare-earth digital elevation model in meters above mean sea level; *slope*, in degrees; and *aspect*, also in degrees, giving the azimuth of steepest descent.

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

3.4 Quantile representation of observations and datasets

A central choice in this study is to represent both observational time series and gridded wind resource datasets using empirical wind speed quantiles at each site and height. Given a quality-controlled hourly series $\{u_{s,h,t}\}_{t=1}^{T}$ for site s and height h, we compute the empirical cumulative distribution function $F_{s,h}(u)$ and evaluate its inverse on a fixed percentile grid:

$$q_{s,h,p} = F_{s,h}^{-1}(p), \qquad p = 0, \ldots, 100. \tag{1}$$

These percentile values (`q000`–`q100`) form a unified, distribution-level representation shared by all data sources and used throughout the benchmarking and ML stages.

The quantile transformation is applied uniformly across all sources. For the ASOS and GS archives, the cleaned hourly time series at each site and height are converted directly into empirical quantiles. Each wind resource dataset is processed in the same manner: after extracting wind speeds at the observational site coordinates, the identical quantile grid is computed, ensuring that observations and modeled datasets share the same distributional summary. For the full-grid application, the wind resource datasets are similarly transformed to quantiles at every ERA5 grid point, and the resulting per-dataset tables are merged into a unified input structure for downstream model inference.

The empirical quantile representation also provides a consistent way to recover long-term mean wind speed for each site and height. Given the quantile curve $\{(p_k, q_k)\}_{k=0}^{100}$ evaluated on a regular probability grid $p_k = k/100$, the mean can be approximated by integrating the quantile function over the unit interval. Using a simple trapezoidal rule, the estimated mean is

$$\widehat{\mu} \approx \sum_{k=0}^{99} \frac{q_k + q_{k+1}}{2} (p_{k+1} - p_k) = \sum_{k=0}^{99} \frac{q_k + q_{k+1}}{2} \times 0.01. \tag{2}$$

This numerical integration provides a smooth, distribution-respecting estimate of the long-term mean without relying on parametric assumptions. These quantile-derived means are used throughout the study to compute bias and error metrics for the benchmarked datasets and to evaluate the ML predictions generated by WEM.

3.5 Interannual variability

To quantify interannual variability at each site and height, we compute annual-mean wind speeds $\overline{u}_{s,h,y}$ for all years y with sufficient coverage and summarize variability via the simple range

$$R_{s,h} = \max_y \overline{u}_{s,h,y} - \min_y \overline{u}_{s,h,y}. \tag{3}$$

Figure 4 shows $R_{s,h}$ for observed winds at GS sites. Across the network, long-term mean wind speeds span more than 7 $\mathrm{m\,s^{-1}}$, yet the year-to-year range at a given location is typically only a few tenths of a meter per second; the histogram inset demonstrates that most sites have $R_{s,h} < 1\ \mathrm{m\,s^{-1}}$. Thus, spatial differences in mean wind speed dominate over interannual fluctuations at a fixed site.

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

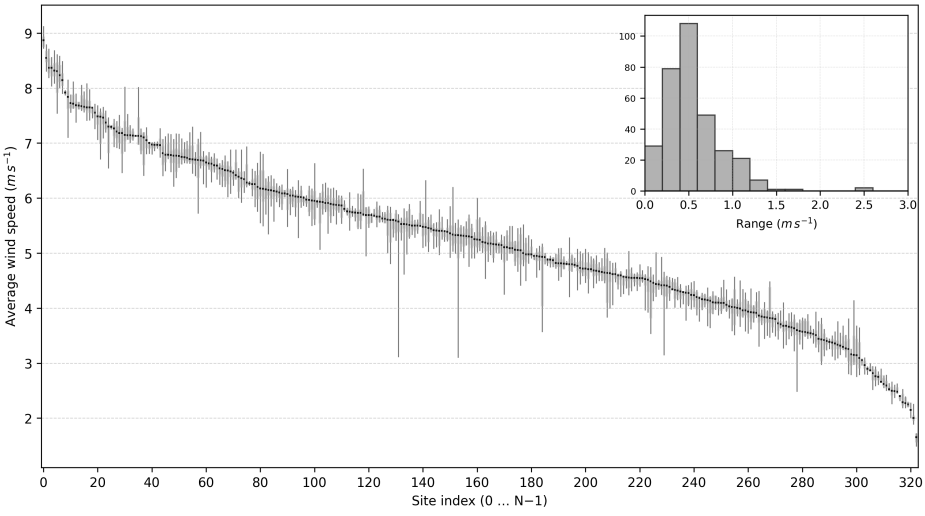


Figure 4. Interannual variability of observed annual mean wind speeds at GS sites. Sites are sorted by their long-term mean (black markers); for each site, vertical lines show the full range across years, and boxes show the interquartile range of annual means. The inset histogram summarizes the distribution of per-site ranges $R_{s,h}$ (Eq. 3). Most observational sites exhibit interannual ranges well below 1 m s^{-1}, even though long-term means span more than 7 m s^{-1} across the network.

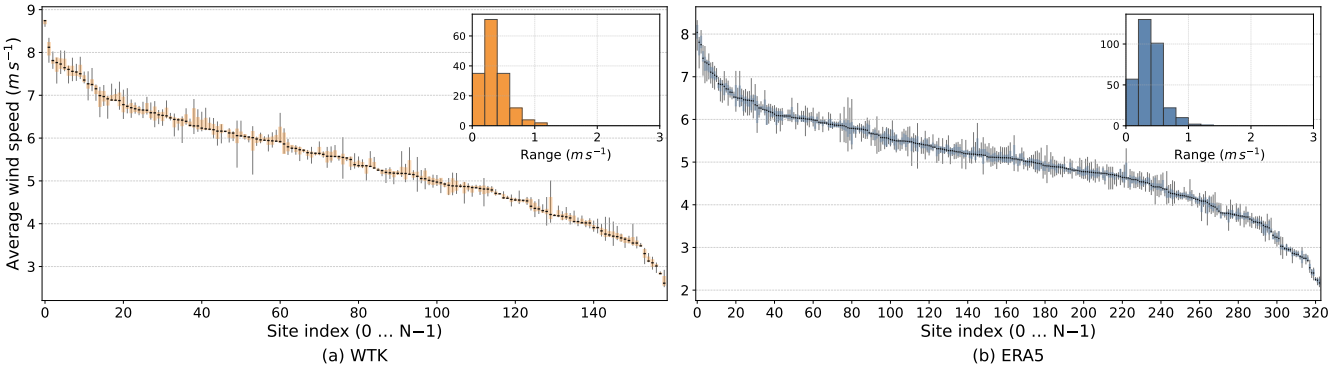


Figure 5. Interannual variability of annual mean wind speeds at GS sites for two representative wind resource datasets, WTK and ERA5. As in Fig. 4, both panels show site-level ranges and interquartile ranges of annual means, with an inset histogram of per-site ranges $R_{s,h}$. All products exhibit relatively modest interannual variability at individual sites, with typical ranges on the order of a few tenths of a meter per second and only a small minority of sites exceeding 1 m s^{-1}.

220 Figure 5 presents the same analysis for the modeled datasets. The qualitative picture is similar: within each product's own period of record, interannual ranges are modest (again mostly below 1 m s^{-1}), even though the absolute mean speeds differ across sites and datasets. The panels also highlight that not all products are evaluated at the same number of sites. For example, WTK-LED CONUS is only available for 2018 to 2020, so under a strict time-matching strategy it can be compared only at

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

GS sites with sufficient observations during those years; this naturally yields fewer stations than longer-record products such as WTK or ERA5.

The small interannual ranges in Figs. 4–5 play a specific role: they justify treating each product's multi-year time series at a site as a single long-term distribution for evaluation and learning. Given that year-to-year fluctuations are much smaller than spatial gradients and systematic dataset biases, a quantile summary over all available years is essentially as informative as a summary restricted to any particular subset of years. This allows us to exploit all available information from each dataset at all GS sites without being constrained by overlapping temporal windows.

4 Methods

4.1 Evaluation of public wind resource datasets

We evaluate public wind resource datasets by comparing their long-term wind speed statistics to those derived from ASOS and GS observations. All comparisons are performed in a harmonized quantile framework (Sect. 3.4), and metrics are computed at the site level, with separate analyses for near-surface (10 m) and hub-height regimes.

4.1.1 Horizontal and vertical interpolation

Wind resource datasets differ not only in the vertical levels at which they report wind speed but also in their native horizontal grids. To generate fair, site-specific comparisons against both ASOS and GS observations, we apply consistent horizontal and vertical matching procedures across all products.

Because observational sites rarely coincide exactly with model grid points, we horizontally interpolate modeled wind speeds to the station coordinates before performing any height adjustments. For all datasets, we use inverse-distance weighting applied to the four nearest model grid points surrounding each site. This approach provides a smooth, locally responsive estimate without imposing assumptions about the spatial structure of the wind field. Given grid-point values u_i at horizontal distances d_i from the station location, the interpolated wind speed is

$$u_{\mathrm{IDW}} = \frac{\sum_{i=1}^{4} d_i^{-1} u_i}{\sum_{i=1}^{4} d_i^{-1}}, \tag{4}$$

This same inverse-distance weighting procedure is used for all datasets and all heights.

For ASOS stations, the native observation height is nominally 10 m above ground, and all wind resource datasets used in this study provide a directly comparable 10 m diagnostic. After applying the horizontal interpolation described above, we take the modeled 10 m wind speeds as is, without further vertical adjustment. This allows for a clean, height-consistent comparison between observations and all gridded products at the near-surface level.

Measurement heights span 10–144 m for GS sites, while the wind resource datasets provide winds at a limited set of standard heights that may not coincide with observational levels. If a dataset reports wind speed at exactly the observed height, we use it directly. Otherwise, when the observational height lies between two modeled levels z_1 and z_2 with corresponding wind speeds

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

u_{z_1} and u_{z_2}, we estimate the wind speed at the observational height using a standard power-law profile

$$u(z) = Az^{\alpha}, \tag{5}$$

with α diagnosed from the two bracketing model levels in log–log space. This interpolation is applied consistently across datasets so that each gridded product is evaluated at the same effective measurement height before constructing quantiles and long-term means.

4.1.2 Bias and error metrics

We quantify dataset performance at each site using simple statistics based on long-term mean wind speed. For a given site s and dataset d, let $\overline{u}_s^{(d)}$ denote the dataset mean and $\overline{u}_s^{(\mathrm{obs})}$ the observed mean (from ASOS or GS). We define the site-level bias and absolute bias as

$$\mathrm{bias}_s^{(d)} = \overline{u}_s^{(d)} - \overline{u}_s^{(\mathrm{obs})}, \tag{6}$$

$$|\mathrm{bias}_s^{(d)}| = \left|\overline{u}_s^{(d)} - \overline{u}_s^{(\mathrm{obs})}\right|. \tag{7}$$

The sign of $\mathrm{bias}_s^{(d)}$ indicates whether the dataset overestimates (positive) or underestimates (negative) the observed mean; the magnitude $|\mathrm{bias}_s^{(d)}|$ is the absolute error at that site.

To summarize performance across sites, let $\mathcal{S}$ denote a site subset (e.g., all ASOS sites, all GS sites, or a regional subset) with $N = |\mathcal{S}|$. The mean signed bias

$$\overline{\mathrm{bias}}^{(d)} = \frac{1}{N}\sum_{s\in\mathcal{S}} \mathrm{bias}_s^{(d)} \tag{8}$$

captures the overall tendency of a dataset to over- or underestimate long-term mean wind speed, while the mean absolute bias

$$\overline{|\mathrm{bias}|}^{(d)} = \frac{1}{N}\sum_{s\in\mathcal{S}} \left|\mathrm{bias}_s^{(d)}\right| \tag{9}$$

quantifies the typical error magnitude in units of meters per second. Note that $\overline{|\mathrm{bias}|}^{(d)}$ is exactly the mean absolute error (MAE) of dataset d on $\mathcal{S}$; when reporting this metric we refer to it as mean absolute bias. We additionally report the median absolute bias $\mathrm{median}_{s\in\mathcal{S}}\left(|\mathrm{bias}_s^{(d)}|\right)$, which is less sensitive to outliers. Together, these statistics concisely characterize the distribution of site-level errors in long-term wind speed means.

4.2 Machine-learned quantile model

4.2.1 Problem formulation

The goal of the ML component is to predict the *entire* wind speed quantile curve at a given site and height, using multiple public wind resource datasets and static geographic information as predictors. We work in a long-format representation where

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

each row corresponds to a single combination of

$$(s,h,q) \in \{\text{site}\} \times \{\text{height}\} \times \{0,\ldots,100\}, \tag{10}$$

and the target is the observed wind speed at the q-th empirical percentile.

Let $u_{\text{obs}}(s,h,q)$ denote the empirical quantile derived from observations (either ASOS 10 m or GS tall sites), and let

$$\mathbf{z}(s,h,q) = \big(u_{\text{ERA5}}(s,h,q),\, u_{\text{WTK}}(s,h,q),\, u_{\text{WTK-LED Climate}}(s,h,q),\, u_{\text{WTK-LED CONUS}}(s,h,q),\, u_{\text{HRRR}}(s,h,q)\big) \tag{11}$$

denote the corresponding quantiles from the public datasets at the same (s,h,q) triple. We also define a vector of static site- and height-specific features,

$$\mathbf{t}(s,h) = \big(\text{elevation, slope, aspect, latitude, longitude, } h,\, u_{\text{GWA}}(s,h)\big), \tag{12}$$

where $u_{\text{GWA}}(s,h)$ is a power-law-interpolated mean wind speed from the GWA at height h (see below).

The full feature vector for a given row is then

$$\mathbf{x}(s,h,q) \;=\; \Big(q,\, \mathbf{z}(s,h,q),\, \mathbf{t}(s,h)\Big), \tag{13}$$

where q is the quantile index (treated as a continuous covariate, typically $q \in [0,100]$ stored as a float).

We formulate supervised regression at the quantile-row level:

$$f_\theta :\, \mathbf{x}(s,h,q) \;\longmapsto\; \widehat{u}_{\text{obs}}(s,h,q), \tag{14}$$

where θ are model parameters (gradient-boosted tree weights) and $\widehat{u}_{\text{obs}}$ is the predicted wind speed at the given quantile. The training dataset consists of all rows constructed from the combined long-format table

$$\mathcal{D} = \Big\{\big(\mathbf{x}(s,h,q),\, u_{\text{obs}}(s,h,q)\big)\Big\} \tag{15}$$

across all ASOS and GS sites, heights, and quantile levels that pass quality control.

4.2.2 Feature set

The feature vector $\mathbf{x}(s,h,q)$ is constructed by merging wind resource, topographic, and ancillary information for each quantile row.

Wind resource features

For each dataset

$$d \in \{\text{ERA5}, \text{WTK}, \text{WTK-LED Climate}, \text{WTK-LED CONUS}, \text{HRRR}\}, \tag{16}$$

we compute $u_d(s,h,q)$, the empirical quantile at level q for dataset d at site s and height h, using a fixed climatological window from 2007 through 2024. Unlike the observational records, which require completeness screening, the gridded products are

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

temporally continuous by construction, so we simply include all available hours within this window without additional filtering. The lower bound of 2007 is chosen to coincide with the start of the WTK archive and to avoid extending the climatology too far into the past, where conditions may be less representative of contemporary and future wind resources. The upper bound of 2024 reflects the timing of this study: observational archives do not yet contain a complete year for 2025, so restricting all datasets to this window ensures a consistent, fully populated comparison period.

Topographic features

Static terrain descriptors are derived from USGS 3DEP elevation products at each site location. The feature set includes elevation above mean sea level (in meters), denoted $\mathrm{elev}(s)$, the local slope magnitude computed from the 3DEP gradient, and aspect encoded via sine and cosine of the aspect angle,

$$\mathrm{northness}(s) = \cos\big(\mathrm{aspect}(s)\big), \quad \mathrm{eastness}(s) = \sin\big(\mathrm{aspect}(s)\big), \tag{17}$$

with special handling for nearly flat terrain, where the slope falls below a small threshold. These features are constant across quantiles for a given (s, h) and enter the model unchanged (no standardization is required for tree-based methods).

GWA-derived features

We incorporate the GWA as a high-level characterization of the local wind climate. Specifically, we first sample GWA mean wind speed at reference heights of 10, 50, 100, and 150 m on a regular grid of unique $(\mathrm{lon}, \mathrm{lat})$ pairs. For each site, we then fit a simple power-law profile $u(z) = Az^{\alpha}$ by regressing $\log u$ against $\log z$ using these four heights and evaluate the fitted profile at the target height h to obtain $u_{\mathrm{GWA}}(s, h)$. The resulting interpolated mean wind speed $u_{\mathrm{GWA}}(s, h)$ is used as a scalar feature for all quantiles at that (s, h).

Geographic and height features

We append three simple descriptors to the feature set: latitude and longitude in degrees, treated as continuous inputs; height above ground h (in meters); and the quantile index q represented as a continuous variable. The inclusion of q allows a single model to learn a smooth functional dependence of bias corrections on the quantile level rather than training separate models for each quantile.

4.2.3 Geographic cross-validation and neighbor exclusion

To rigorously assess generalization to new sites, we use a leave-one-out cross-validation (LOOCV) scheme over the GS stations, augmented with a 10 km geographic exclusion radius. Using great-circle distances between all observational sites (both ASOS and GS), we define for each site s the set of neighbors within 10 km,

$$\mathcal{N}_{10}(s) \;=\; \big\{ s' : d(s, s') < 10\ \mathrm{km} \big\}. \tag{18}$$

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

The outer LOOCV loop iterates over GS sites $s^\star$. For each held-out site, the test set consists of all quantile rows $(s^\star, h, q)$ at that site, while the training set excludes *all* rows from $s^\star$ and from any neighbor $s' \in \mathcal{N}_{10}(s^\star)$; the remaining rows from GS and ASOS sites form the training pool. This 10 km buffer avoids overly optimistic estimates that would arise if nearby, strongly correlated sites (for example, multiple turbines instrumented within the same wind plant) were included in the training set when evaluating a given location. In practice, this is especially important for vendor-supplied projects in which multiple turbines are instrumented within the same wind plant: each turbine is treated as a distinct site, often separated by only a few hundred meters, and the 10 km neighbor exclusion ensures that all other turbines in the same project are removed from the training set whenever one of them is used for testing.

Balancing ASOS and GS rows

Because the number of 10 m ASOS rows far exceeds that of tall GS rows, naive training would overweight the near-surface regime. To mitigate this, we apply a simple down-sampling strategy in each training fold: rows are partitioned into a "GS" group and an "ASOS" group (10 m stations), and the ASOS group is then randomly down-sampled so that the two groups have comparable total weight in the training loss. This ensures that the model pays comparable attention to hub-height and 10 m regimes and learns a unified mapping that performs well across both.

4.2.4 Model choice and training procedure

We use gradient-boosted regression trees as implemented in XGBoost (`XGBRegressor`). This class of models is well suited to tabular data with nonlinear interactions and has the practical advantage of requiring minimal feature preprocessing.

The base model is an ensemble of decision trees trained with a squared-error objective:

$$\mathcal{L}(\theta) = \sum_i \big(y_i - f_\theta(\mathbf{x}_i)\big)^2 + \Omega(\theta), \tag{19}$$

where $\Omega(\theta)$ is the standard XGBoost regularization term (penalizing the number of leaves and leaf weights). We monitor both root-mean-square error and MAE on held-out data but optimize the training loss in terms of root-mean-square error. Because quantiles must be non-decreasing in the quantile index q, we treat q as a special feature and impose a monotonicity constraint in XGBoost so that predictions are required to monotonically increase as q increases.

For a given LOOCV fold, we first construct the training set after applying the 10 km exclusion and ASOS/GS balancing, then train an `XGBRegressor` model with a fixed maximum number of boosting rounds (trees), and finally save the fitted model for that fold and generate predictions for the held-out site $s^\star$ across all its heights and quantiles.

4.2.5 Feature sweeps and hyperparameter optimization

To understand which inputs are most valuable and to avoid overfitting to a particular combination of features, we perform systematic sweeps over the feature space and then tune model hyperparameters within the selected feature set. These experiments are implemented in a set of scripts that all wrap the same LOOCV training-and-evaluation pipeline.

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

Wind feature combination sweeps

We first perform an exhaustive combinatorial sweep over the choice of wind resource datasets included in $\mathbf{z}(s,h,q)$ while keeping topographic and ancillary features fixed. Let

$$\mathcal{D}_{\text{wind}} = \{\text{ERA5}, \text{WTK}, \text{WTK-LED Climate}, \text{WTK-LED CONUS}, \text{HRRR}\} \tag{20}$$

denote the candidate set of gridded wind products. For every non-empty subset $\mathcal{S} \subseteq \mathcal{D}_{\text{wind}}$, we construct feature vectors using only $\{u_d(s,h,q) : d \in \mathcal{S}\}$ as wind inputs, together with a fixed auxiliary feature set consisting of elevation, slope, aspect encoding, measurement height h, and geographic coordinates (lat, lon). We then run the full LOOCV procedure, recording quantile-level and mean wind error metrics on the held-out GS sites, and finally aggregate these metrics into summary statistics (for example, bias-based and absolute bias–based scores) for that particular subset $\mathcal{S}$. This stage yields a complete performance surface over all possible combinations of wind datasets and is used to select a preferred wind feature subset for the production model.

Auxiliary feature sweeps

With the wind feature subset fixed from the first stage, we then carry out a second exhaustive sweep over combinations of auxiliary predictors. The auxiliary pool includes topographic variables (elevation, slope, and aspect encoding) and geometric variables (measurement height h and geographic coordinates (lat, lon)). We define a set of binary inclusion/exclusion switches over these auxiliary features and evaluate *all* resulting combinations by reconstructing the feature vectors using the fixed wind feature subset together with each chosen auxiliary subset, re-running the LOOCV protocol, and computing the mean wind error metric on the held-out GS sites.

Two-stage optimization

We do not perform a single, joint combinatorial sweep over all wind and auxiliary features together because the number of configurations grows exponentially in the total number of predictors, making a full factorial exploration computationally infeasible given the LOOCV setup. Instead, the two-stage design (exhaustive over wind subsets, then exhaustive over auxiliary subsets conditional on the chosen wind set) provides full coverage within each feature family while keeping the overall experiment count tractable. Finally, in light of the benchmark results in Sect. 5.1, which show that GWA is the best-performing dataset for the GS observation sites, we treat the GWA mean wind speed as a required auxiliary predictor in the final feature set. That is, for the production WEM configuration, GWA is included alongside the selected combination of gridded wind products and auxiliary variables.

Hyperparameters

We tune the standard XGBoost hyperparameters rather than fixing them a priori, namely, the learning rate (`learning_rate`), maximum tree depth (`max_depth`), minimum child weight (`min_child_weight`), row subsampling fraction (`subsample`),

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

Table 2. Dataset performance on ASOS and GS sites (long-term mean wind speed). For each metric, entries are reported as ASOS/GS, computed over station-level climatological means. The quantity $\overline{|\text{bias}|}$ is the site-level mean absolute bias, i.e., the MAE defined in Sect. 4.1.2, and the percentage column reports MAE expressed as a fraction of the observed mean wind speed. Bold text signifies the top-performing dataset for each pair of metric and site-type.

Dataset	$\overline{\text{bias}}$ ASOS/GS (m s^{-1})	$\overline{\lvert\text{bias}\rvert}$ ASOS/GS (m s^{-1})	Median $\lvert\text{bias}\rvert$ ASOS/GS (m s^{-1})	% $\overline{\lvert\text{bias}\rvert}$ ASOS/GS
ERA5	0.17 / -0.24	0.64 / 0.99	0.53 / 0.74	23 / 19
WTK	0.28 / 0.29	**0.54** / 0.70	**0.44** / 0.62	**20** / 15
WTK-LED Climate	0.68 / 0.87	0.84 / 1.09	0.74 / 0.91	32 / 25
WTK-LED CONUS	0.57 / 1.13	0.74 / 1.20	0.64 / 1.01	29 / 28
HRRR	0.18 / 0.21	**0.54** / 0.74	0.47 / 0.67	**20** / 16
GWA	**0.14** / **0.05**	0.59 / **0.63**	0.48 / **0.48**	21 / **14**

column subsampling fraction (`colsample_bytree`), and number of boosting rounds (`n_estimators`). With a fixed feature set, we use Optuna to sample candidate hyperparameter vectors, run the full LOOCV procedure on GS sites, and evaluate the site-level MAE of reconstructed mean wind speed (i.e., $\overline{|\text{bias}|}$ from Sect. 4.1.2) as the objective. After a specified budget of trials, we select the configuration with the best LOOCV performance.

Final model for deployment

After completing LOOCV analysis and feature/hyperparameter selection, we retrain a single "final" model on *all* available rows (with no geographic holdout) using the chosen feature set and hyperparameters. This model is used to generate predictions over the full ERA5 grid.

5 Results

5.1 Performance of public wind resource datasets

5.1.1 Summary metrics across sites

We first quantify how closely each wind resource dataset reproduces observed long-term mean wind speeds at the ASOS and GS sites. For every station, we compute the multi-year climatological mean over the years 2007 to 2024 (subject to the completeness rules in Sect. 4.1), and then form bias and absolute bias statistics.

Table 2 summarizes error statistics for each dataset over the ASOS (10 m) and GS (hub-height) subsets, and Fig. 6 shows the corresponding distributions of site-level bias and absolute bias for the ASOS stations and the GS sites. Across the ASOS 10 m subset, all datasets exhibit modest positive bias, indicating a tendency to overestimate near-surface winds, but the magnitude and spread of this bias vary substantially by product. The smallest errors are obtained for WTK and HRRR, which have

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

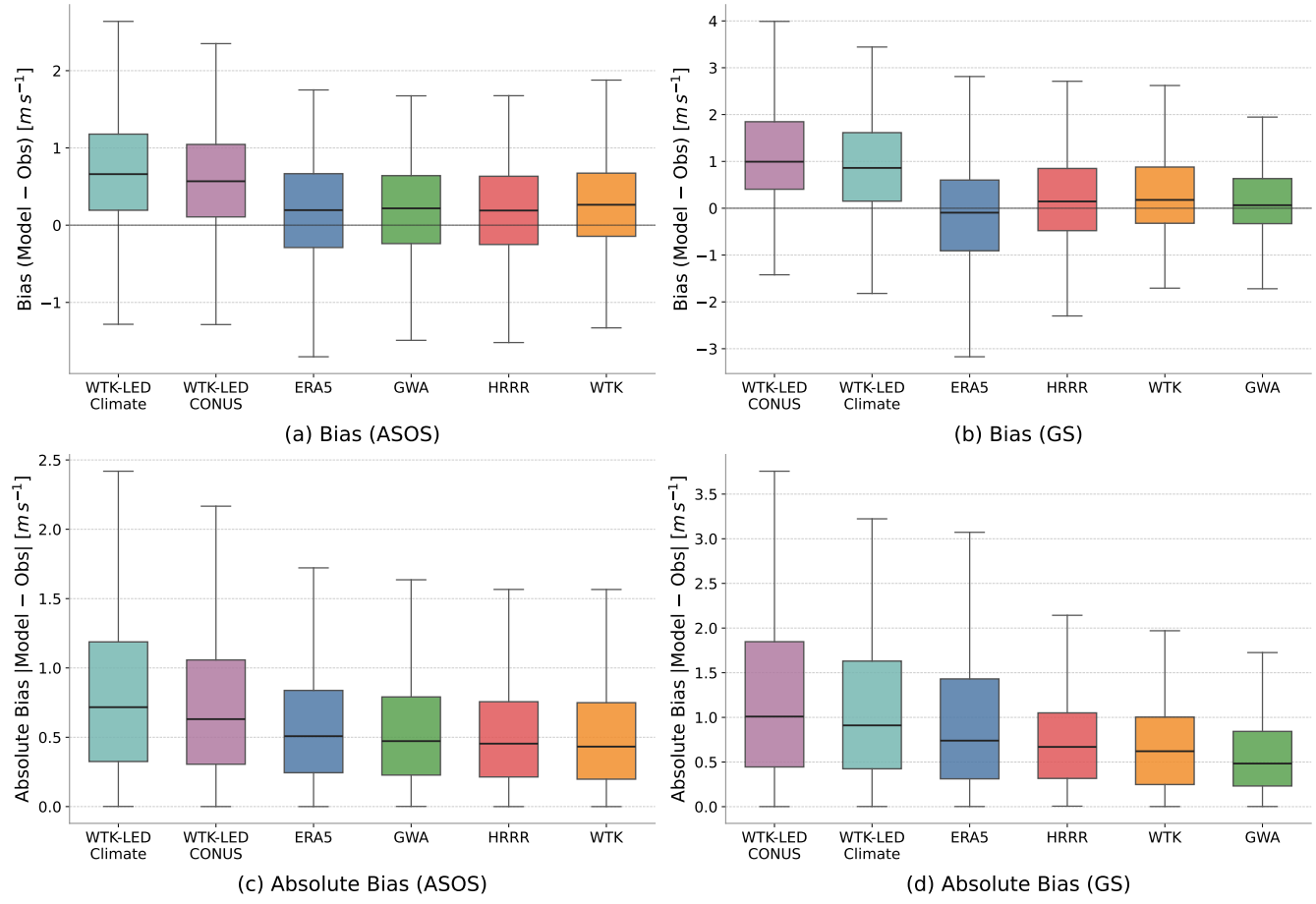


Figure 6. Distributions of site-level bias (top row) and absolute bias (bottom row) for the ASOS 10 m sites (left column) and GS sites (right column).

median absolute biases of approximately $0.44\ \mathrm{m\,s^{-1}}$ and $0.47\ \mathrm{m\,s^{-1}}$, respectively, and mean absolute biases near $0.54\ \mathrm{m\,s^{-1}}$. The ASOS boxplots highlight this behavior: both datasets have medians close to zero and relatively narrow interquartile ranges, in contrast to the broader, more positively biased distributions for the WTK-LED products. GWA and ERA5 perform slightly worse than WTK and HRRR but still fall in a similar regime, with median absolute biases around $0.48\ \mathrm{m\,s^{-1}}$ and $0.53\ \mathrm{m\,s^{-1}}$ and mean absolute biases near $0.59\ \mathrm{m\,s^{-1}}$ and $0.64\ \mathrm{m\,s^{-1}}$. WTK-LED Climate and WTK-LED CONUS show the largest errors in the ASOS subset, with mean biases of roughly $0.68\ \mathrm{m\,s^{-1}}$ and $0.57\ \mathrm{m\,s^{-1}}$ and mean absolute biases of about $0.84\ \mathrm{m\,s^{-1}}$ and $0.74\ \mathrm{m\,s^{-1}}$, and their boxplots show both higher medians and wider spreads.

Over the GS hub-height subset, the absolute error levels increase, but the relative ranking of datasets becomes even clearer. GWA shows the best overall agreement with observations, with a mean bias close to zero ($0.05\ \mathrm{m\,s^{-1}}$), a mean absolute bias of about $0.63\ \mathrm{m\,s^{-1}}$, and the smallest median absolute bias ($0.48\ \mathrm{m\,s^{-1}}$) among all products. The GS boxplots confirm this: GWA

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

maintains a tight, nearly unbiased distribution, whereas the WTK-LED products remain strongly and consistently positively biased. WTK and HRRR form the next performance tier, with mean absolute biases of roughly $0.70\ \mathrm{m\,s^{-1}}$ and $0.74\ \mathrm{m\,s^{-1}}$ and median absolute biases of $0.62\ \mathrm{m\,s^{-1}}$ and $0.67\ \mathrm{m\,s^{-1}}$, respectively, and generally symmetric error distributions. ERA5 exhibits a negative mean bias (about $-0.24\ \mathrm{m\,s^{-1}}$) but larger absolute errors at hub height (mean absolute bias near $0.99\ \mathrm{m\,s^{-1}}$, median absolute bias around $0.74\ \mathrm{m\,s^{-1}}$), and the corresponding boxplots show both under- and overestimation with a wide interquartile range. Comparing the 10 m and hub-height subsets, median absolute biases are generally larger for GS sites than for ASOS, consistent with the increased difficulty of representing tall-tower winds in more complex terrain; but the overall ranking of WTK, HRRR, and especially GWA as the best-performing products, and the WTK-LED variants as the most biased, remains broadly consistent across heights.

5.1.2 Height dependence

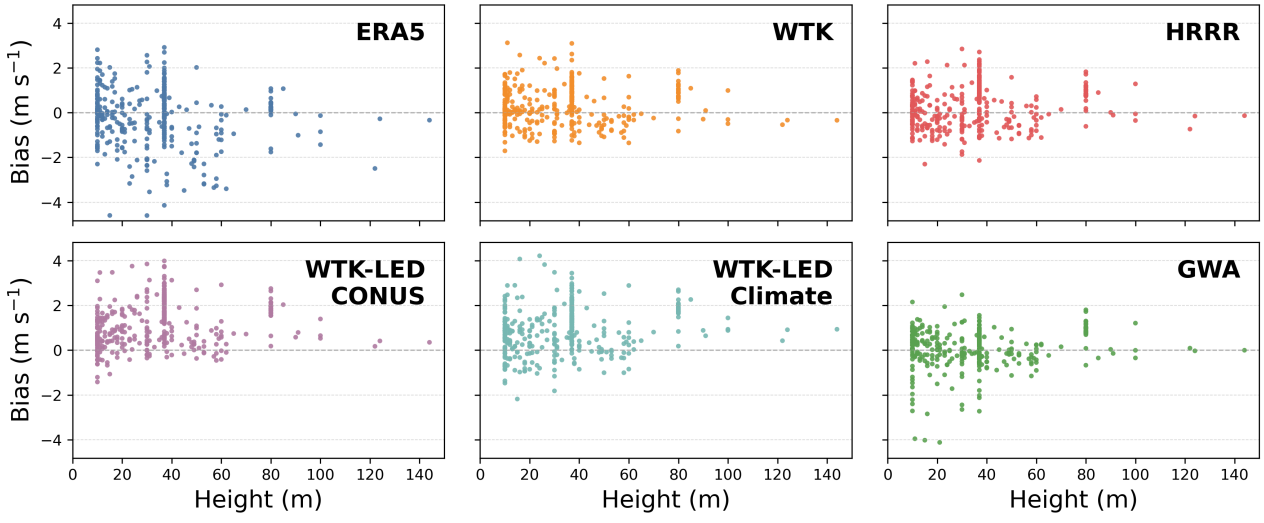


Figure 7. Height-dependent performance of all datasets at GS sites. Each panel shows the bias in mean wind speed (dataset minus observations) for individual site–height combinations as a function of measurement height, with common axes across panels.

The GS sites span 10 to 144 m above ground, providing a direct view of how model bias varies with height (Fig. 7). Figure 7 shows the bias for each site–height combination, revealing no simple or monotonic dependence of bias on height for any dataset. Instead, all products exhibit substantial spread at a given height, with particularly wide and noisy bias distributions at lower levels, indicating larger site-to-site variability near the surface.

5.1.3 Spatial patterns of bias

Maps of station-level bias show that errors in the wind datasets are spatially structured rather than random, and that different products express distinct regional patterns. For two representative datasets (ERA5 and GWA), we map $\mathrm{bias}_s^{(d)}$ at 10 m over the

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

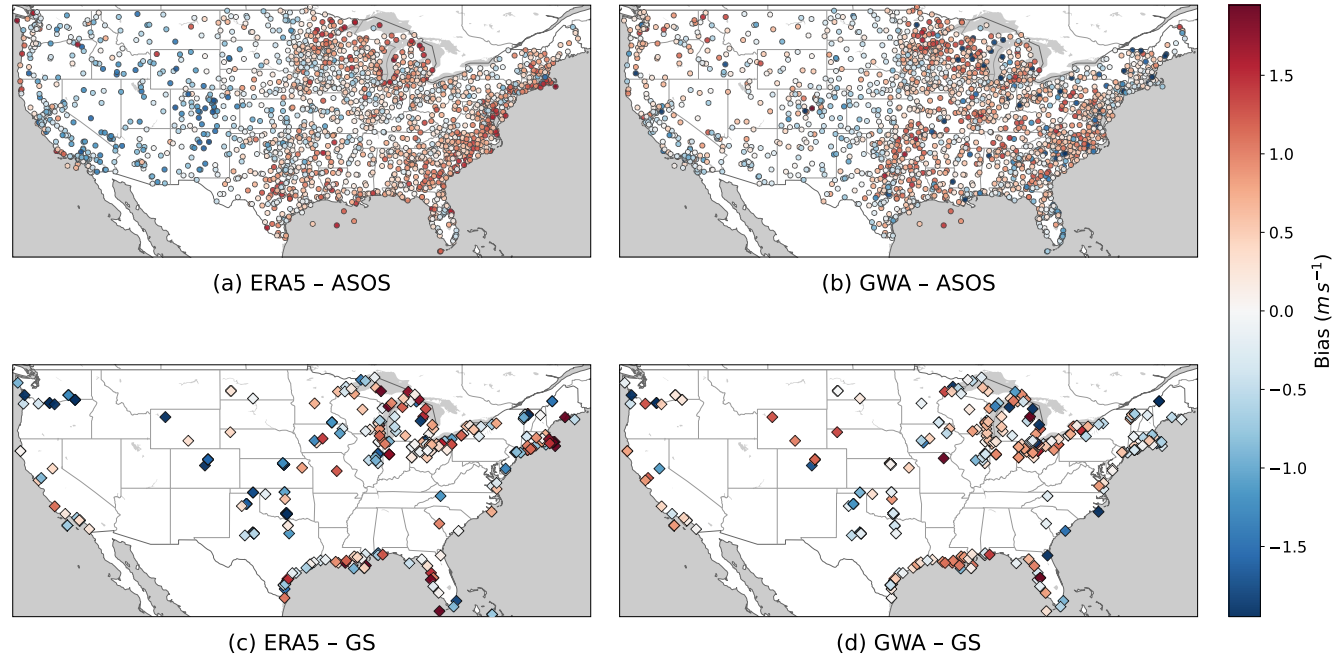


Figure 8. Spatial patterns of ERA5 and GWA bias relative to observations. *Top row:* bias at 10 m across the ASOS network. *Bottom row:* bias across the GS sites. In all panels, blue indicates underestimation and red indicates overestimation, with common color limits applied across all four panels.

ASOS network and at hub heights over the GS towers (Fig. 8), using a common diverging color scale so that differences across panels directly reflect differences between the two datasets.

On the ASOS network, both ERA5 and GWA exhibit clear geographic organization: neighboring stations typically share the same sign of bias and comparable magnitude, indicating that each product has regionally coherent tendencies rather than independent, site-by-site noise. At the same time, the two products often disagree in both sign and magnitude at individual stations and across broader regions, with ERA5 and GWA alternately over- or underestimating the observed winds. The hub-height maps over GS towers, though based on a sparser and more clustered sample, reinforce this picture: at many towers ERA5 and GWA have noticeably different biases, even when their large-scale spatial patterns are harder to discern.

These examples illustrate a broader theme that also appears in the station-level statistics: the various datasets provide related but non-redundant views of the wind climate. The joint vector of model estimates at a site therefore encodes valuable information about the true mean wind speed. In our later modeling, we exploit these systematic cross-dataset differences as a source of diversity, allowing a single ML model to learn robust corrections without explicitly defining geographic regions or prescribing fixed weights for each product.

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

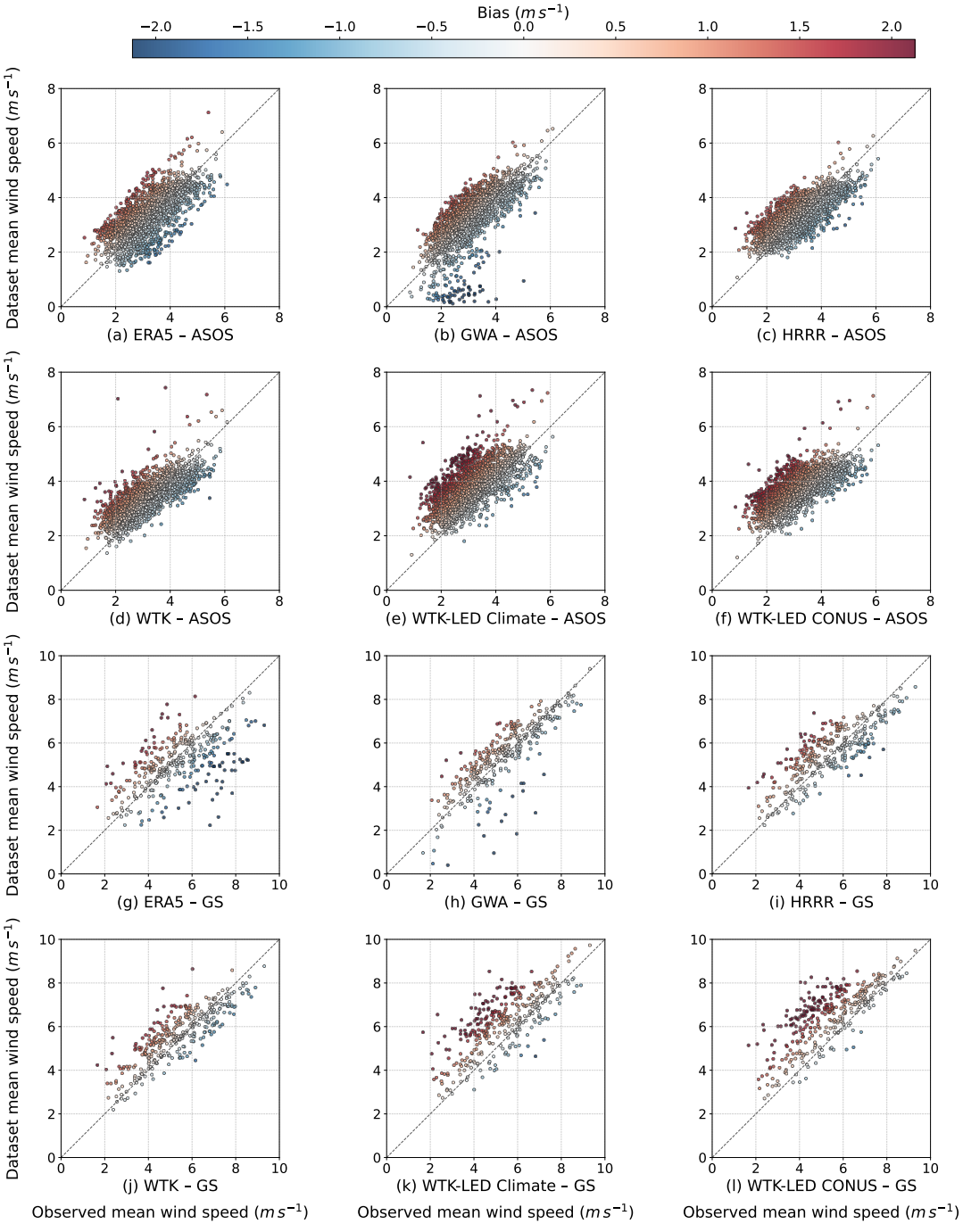


Figure 9. Dataset–observation scatterplots for ASOS 10 m stations (top six panels) and GS hub-height stations (bottom six panels). Each panel shows station-level climatological mean wind speed from the dataset versus observations, with point color indicating bias.

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

5.1.4 Dataset–observation scatterplots

To further examine how the datasets behave across the range of mean wind speeds, we plot station-level climatological means from each product against the corresponding observed mean, coloring each point by its bias (Fig. 9). For easy comparison across models and observation types, all panels share the same axis limits.

Across all products and both networks, the points lie along a clear diagonal band, confirming that each dataset captures the broad ordering of sites from low to high mean wind speed. For HRRR and WTK, a consistent pattern appears: low-wind sites tend to lie above the line and are colored warm (positive bias), whereas high-wind sites fall below the line and are colored cool (negative bias). This implies a compressed effective slope relative to the observations. These datasets systematically overestimate weaker regimes and underestimate stronger ones rather than merely adding random noise. By contrast, GWA generally remains closer to the one-to-one line and exhibits a weaker compression effect, with its largest deviations arising from a small subset of sites where wind speeds are substantially underestimated, all of which are at heights of 37 m or lower.

The WTK-LED Climate and WTK-LED CONUS ensembles behave differently. In both the ASOS and GS panels, their point clouds are shifted upward relative to the 1:1 line. This indicates a predominantly additive bias: the ensembles are "too energetic" almost everywhere. ERA5 shows a different pattern. On the ASOS network, its points are roughly centered on the 1:1 line but display a larger vertical spread than WTK, HRRR, or GWA, indicating a mix of positive and negative biases across sites rather than a consistently high bias. On the GS network, the ERA5 panel appears as a relatively diffuse cloud rather than a narrow diagonal band, consistent with weaker correspondence to observed mean wind speed at hub height and larger site-to-site variability in bias.

5.1.5 Key findings from the benchmark

The benchmark analysis points to several conclusions that directly shape the design of the ML ensemble. First, climatological means are a stable target: interannual variability in annual mean wind speed is modest at both ASOS and GS sites, with most stations exhibiting year-to-year ranges below $1\ \mathrm{m\,s^{-1}}$, so multi-year averages provide a reliable reference, even when datasets cover different sub-periods. Second, typical errors are moderate and height-dependent: the better-performing datasets have median absolute biases of roughly 0.4–$0.5\ \mathrm{m\,s^{-1}}$ at 10 m, increasing to about 0.6–$0.8\ \mathrm{m\,s^{-1}}$ at hub height, with WTK and HRRR performing best on the ASOS network and GWA performing best on the tall towers. The WTK-LED ensembles consistently exhibit the largest bias at both heights. Third, bias is structured in space and by wind speed regime: spatial maps reveal geographically coherent, dataset-specific patterns, and scatterplots show systematic behavior across the distribution, with WTK and HRRR tending to overestimate low-wind sites and underestimate high-wind sites (a compressed slope relative to observations), WTK-LED behaving more like an additive positive offset, and ERA5 exhibiting substantial variability across both networks. Finally, the datasets provide complementary, non-redundant information: products often disagree in both sign and magnitude of bias, and these contrasts vary systematically by region, so the multi-dataset vector at a site carries a characteristic signature rather than a single consensus signal. These findings show that public wind resource datasets offer several

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

partially independent perspectives on local wind climatology and motivate the use of the ML ensemble that learns how these multi-dataset signatures map to observed mean wind speed.

5.2 Machine-learned quantile ensemble

5.2.1 Feature sweep outcomes

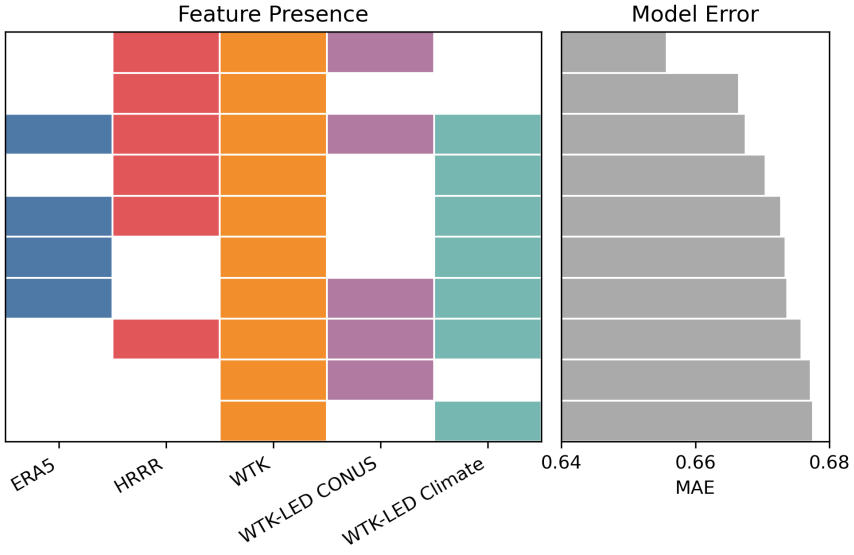


Figure 10. Top-performing wind–feature combinations from the sweep. Left: presence (colored bars) or absence (white) of each wind dataset in a given model. Right: corresponding cross-validated row-level MAE on GS quantile rows (Sect. 5.2.3), with lower values indicating better performance. Rows are ordered from best (top) to worse (bottom), illustrating that combinations including WTK and HRRR consistently yield the lowest errors.

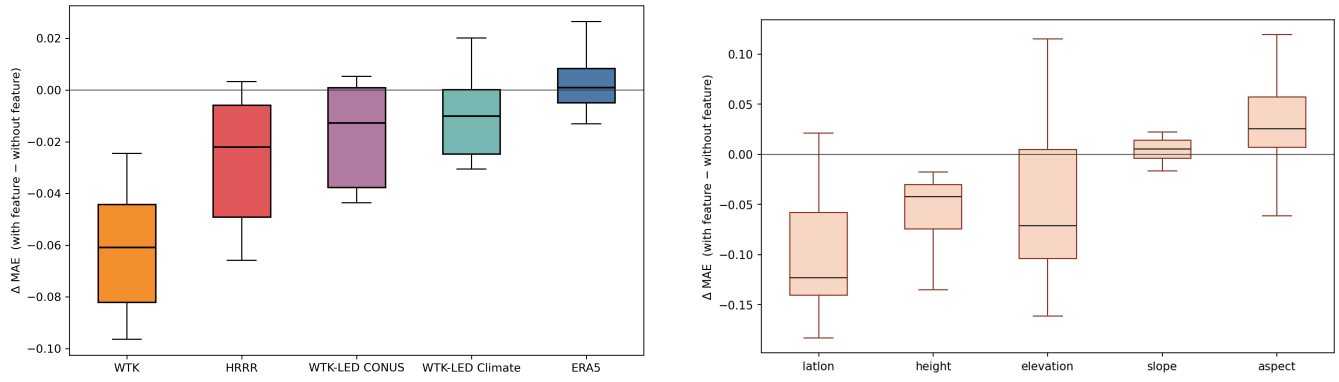
Section 4.2 described the combinatorial sweeps used to identify an effective but compact feature set. Figure 10 summarizes the best-performing wind–feature combinations: each row shows a particular subset of wind datasets (left panel), with the corresponding cross-validated MAE on GS quantile rows (right panel). The lowest-error rows almost always include both WTK and HRRR, and the results from using these two products are significantly improved by the addition of WTK-LED CONUS. This pattern reinforces the idea that a small set of complementary wind products is more valuable than either a single dataset or an indiscriminate "all products" configuration.

Figure 11 quantifies these tendencies through marginal effects on MAE. Each boxplot shows the distribution of

$$\Delta\mathrm{MAE}_{\mathrm{row}} = \mathrm{MAE}_{\mathrm{row,\ with\ feature}} - \mathrm{MAE}_{\mathrm{row,\ without\ feature}} \tag{21}$$

across all combinations in which the feature is toggled on and off. Negative values therefore indicate that including the dataset improves performance. WTK yields the largest and most consistent reduction in error, followed by HRRR, and

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026



(a) Marginal MAE impact of wind datasets.

(b) Marginal MAE impact of auxiliary features.

Figure 11. Marginal contributions of (a) wind datasets and (b) auxiliary predictors.

Table 3. Representative XGBoost hyperparameters for the final WEM model.

	Learning rate	Max tree depth	Min child weight	Subsample	Colsample_bytree	Boosting rounds
Value	0.0222	20	4.28	0.610	0.976	500

WTK-LED CONUS also provides a modest but generally beneficial contribution. WTK-LED Climate has a smaller and more variable effect, and ERA5 is nearly neutral on average once the other predictors are present. The auxiliary-feature sweep shows that geographic coordinates (latitude/longitude), hub height, and elevation each produce clear median reductions in row-level MAE, while local slope and aspect generally degrade performance. Taken together, these results justify the final feature set used in WEM: GWA mean wind speed, retained as a required climatological predictor because it showed the strongest stand-alone hub-height performance; wind quantiles from WTK, HRRR, and WTK-LED CONUS; and a compact group of auxiliary predictors consisting of latitude/longitude, hub height, elevation, and the quantile index q, which allows a single model to learn smooth, quantile-dependent corrections rather than training separate models at each percentile.

5.2.2 Hyperparameter optimization results

The XGBoost hyperparameters for WEM are selected via an Optuna search nested within the LOOCV framework, using average site-level MAE over GS means as the objective. Across the search, the best-trial loss improved by only about 3% relative to the default configuration, indicating modest but consistent benefits from tuning. The resulting hyperparameter values for the selected configuration are summarized in Table 3.

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

5.2.3 Row-level error metrics

We first evaluate WEM in the same space in which it is trained: individual site–height–quantile rows. For each GS site and hub height, we hold the site out under the 10 km geographic exclusion scheme, predict all empirical quantiles $\hat{u}(q)$ from the trained ensemble, and compare them to the observed quantiles $u_{\text{obs}}(q)$. For a given site–height pair, this yields $n_q = 101$ paired values on the $q = 0, \ldots, 100$ grid. Row-level errors are then aggregated over all GS site–height–quantile combinations.

For a model or dataset m, we define row-level bias and absolute error at a given row r (corresponding to a specific site s, height h, and quantile q) as

$$e_r^{(m)} = \hat{u}_r^{(m)} - u_r^{(\text{obs})}, \tag{22}$$

$$|e_r^{(m)}| = \left|e_r^{(m)}\right|. \tag{23}$$

Aggregating across all R rows in the GS subset, we obtain

$$\text{MAE}_{\text{row}}^{(m)} = \frac{1}{R}\sum_{r=1}^{R}\left|e_r^{(m)}\right|, \tag{24}$$

$$\text{Bias}_{\text{row}}^{(m)} = \frac{1}{R}\sum_{r=1}^{R} e_r^{(m)}. \tag{25}$$

These row-level MAE and bias metrics are used throughout the feature-sweep analysis (Figs. 10–11) and in the comparison to baseline datasets in Table 4.

Table 4 summarizes these row-level metrics for the main public datasets and for WEM on the held-out GS rows. Because the comparison is carried out at the level of individual site–height–quantile entries, the table reflects how well each product reproduces the full empirical quantile curve, not just the mean. GWA is omitted here because only long-term mean wind speeds are available; without quantile information, row-level errors e_r cannot be defined.

Several patterns emerge from Table 4. First, WEM achieves the lowest row-level MAE_{row} by a wide margin, reducing typical absolute error to about $0.52\ \text{m s}^{-1}$ and driving the global mean bias Bias_{row} essentially to zero. The best-performing baseline datasets, WTK and HRRR, have substantially larger row-level MAE (roughly $0.78\ \text{m s}^{-1}$ and $0.81\ \text{m s}^{-1}$, respectively) and retain positive mean biases on the order of 0.2-0.3 m s^{-1}. ERA5 exhibits larger error and negative mean bias, while the WTK-LED Climate and WTK-LED CONUS ensembles show both the largest absolute errors and the strongest positive biases.

Taken together, these row-level results indicate that the ML ensemble is not merely interpolating within the spread of the existing datasets: it simultaneously reduces typical quantile-level errors and removes most of the systematic bias present in the individual products. Learning directly in quantile space allows WEM to exploit complementary information across datasets and terrain features, rather than inheriting the tendencies of any single wind resource dataset.

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

Table 4. Baseline datasets versus ML model on GS quantile rows (leave-one-station-out with 10 km buffer).

	ERA5	WTK	WTK-LED Climate	WTK-LED CONUS	HRRR	**WEM**
MAE_{row} ($m\,s^{-1}$)	1.076	0.778	1.150	1.268	0.814	**0.520**
$Bias_{row}$ ($m\,s^{-1}$)	-0.252	0.287	0.883	1.134	0.214	**-0.006**

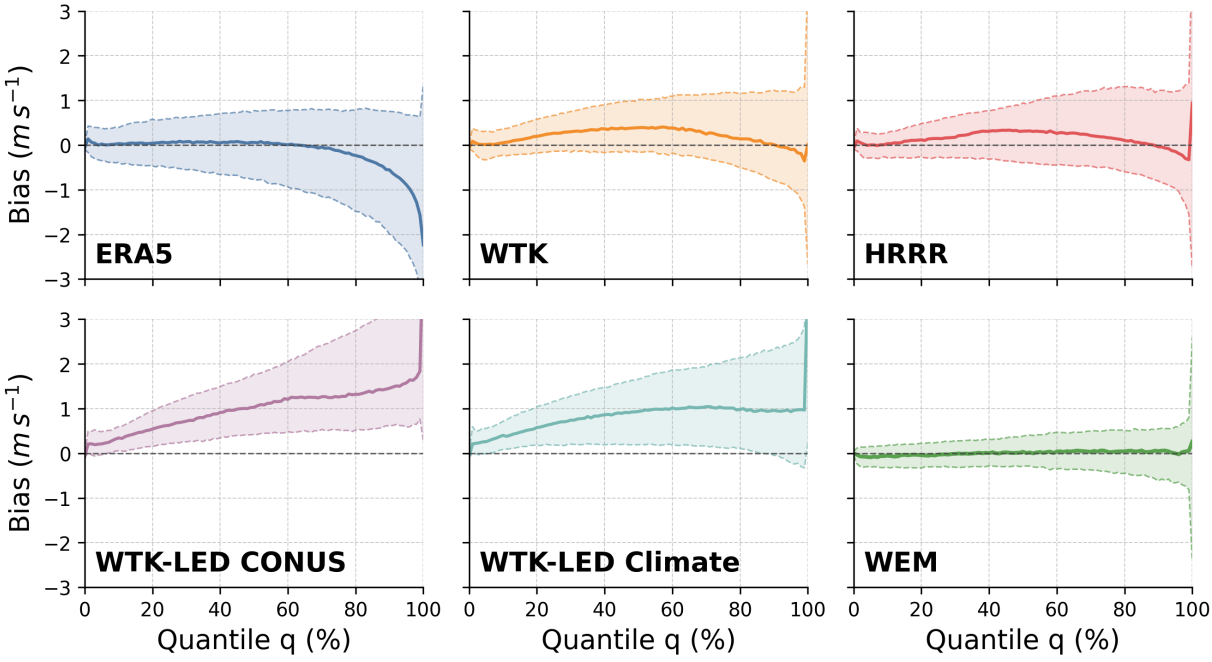


Figure 12. Quantile-dependent bias at GS sites for all datasets with available quantile information (ERA5, HRRR, WTK, WTK–LED CONUS, WTK–LED Climate) and for WEM. Each panel shows the distribution of row-level errors $e_r^{(m)} = x_r^{(m)} - y_r$ as a function of quantile level q: the solid line gives the median bias, dashed lines denote the 25th and 75th percentiles, and the shaded region marks the interquartile range. The horizontal dashed line indicates zero bias.

5.2.4 Quantile-dependent bias structure

Row-level metrics aggregate errors across all quantiles. To examine how performance varies across the wind speed distribution, we compute quantile-dependent error distributions. For a model or dataset m and quantile level q, we collect the row-level errors $e_r^{(m)}$ over all GS site–height rows with $q_r = q$ and summarize them by their median

$$\tilde{e}^{(m)}(q) \;=\; \mathrm{median}_{r:\,q_r=q}\, e_r^{(m)}, \tag{26}$$

along with the 25th and 75th percentiles. Figure 12 shows these summaries for all datasets with quantile information and for WEM: each panel reports $\tilde{e}^{(m)}(q)$ (solid line) together with its interquartile range (shaded band).

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

The baselines exhibit strong and systematic regime-dependent structure. For ERA5, the median bias is close to zero at low and moderate quantiles but becomes increasingly negative toward the upper tail, reflecting underestimation of the strongest winds and a widening spread of errors. For WTK and HRRR, median bias is modestly positive through the middle of the distribution, with a gradual return toward zero at the very highest quantiles. The WTK-LED ensembles exhibit a nearly monotonic increase in positive median bias from low to high quantiles and interquartile ranges that remain largely above zero, indicating systematic overestimation across most of the distribution and particularly strong positive bias in the upper tail.

In contrast, WEM's panel shows a median bias curve that is nearly flat and centered close to zero across almost the entire quantile range, with a narrower interquartile range than the baselines. Small residual deviations are concentrated at the very highest quantile, where some towers exhibit sharp upticks in their maximum observed winds. This flattening of the quantile-dependent error profile indicates that the ensemble has largely removed the regime-dependent biases present in the input datasets: quantiles that are systematically over- or underestimated in individual products are brought back toward neutrality. Because the upper quantiles contribute disproportionately to expected energy production when combined with turbine power curves, reducing bias across the full quantile spectrum – and especially in the strong-wind tail – is a key benefit of learning directly in quantile space rather than only correcting mean wind speeds.

5.2.5 Site-level bias and bias differences

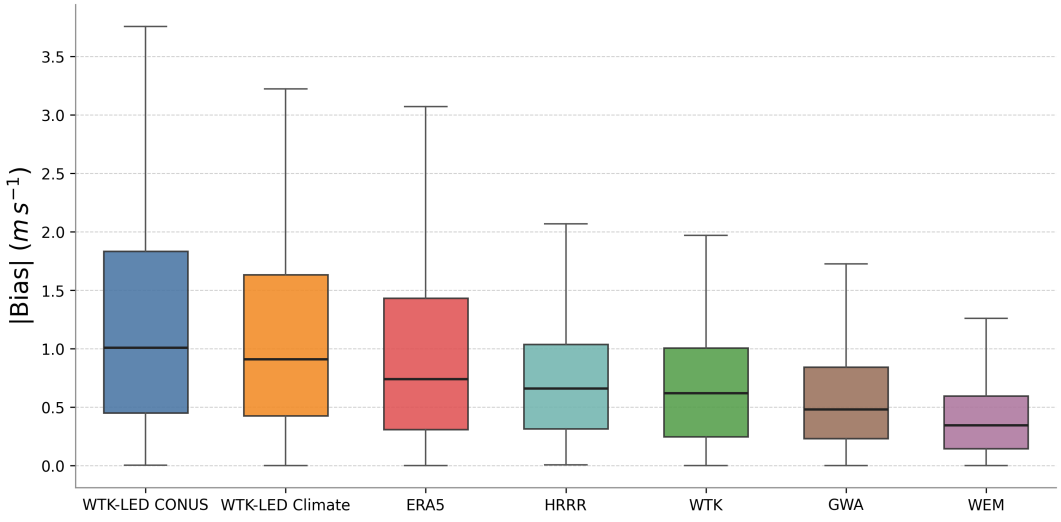


Figure 13. Distribution of site-level absolute biases $|\mathrm{bias}_s^{(m)}|$ for each model or dataset m, evaluated over all GS sites and hub heights. Lower values indicate better agreement with observed long-term mean wind speeds.

To evaluate performance at the scale most relevant for planning, site-level long-term mean wind speed, we integrate the predicted quantile curves to obtain a mean at each site and height. For a model or dataset m, the mean at site s is denoted $\widehat{\mu}_s^{(m)}$,

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

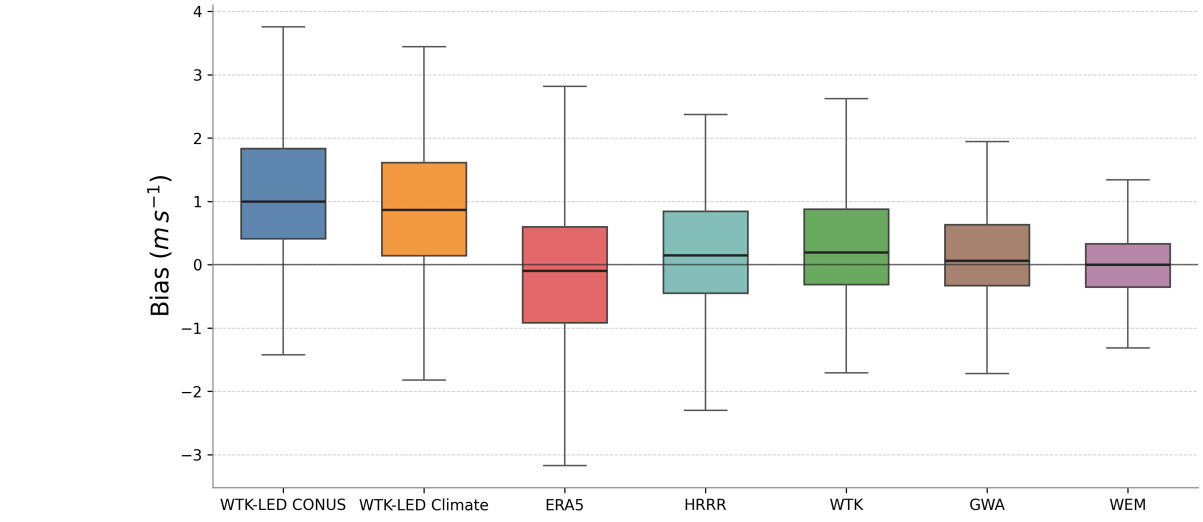


Figure 14. Distribution of site-level signed biases $\mathrm{bias}_s^{(m)} = \widehat{\mu}_s^{(m)} - \widehat{\mu}_s^{(\mathrm{obs})}$ over all GS sites and hub heights. Positive values indicate overestimation of mean wind speed.

Table 5. Site–height mean and median absolute bias in long-term mean wind speed at GS locations. Biases are computed for each dataset or model relative to observed climatological means at each site–height combination and summarized across all GS site–height combinations. Smaller values indicate better agreement with observations.

Dataset / model	Mean $\lvert\mathrm{bias}\rvert$ (m s^{-1})	Median $\lvert\mathrm{bias}\rvert$ (m s^{-1})
WTK-LED CONUS	1.207	1.009
WTK-LED Climate	1.093	0.912
ERA5	0.990	0.740
HRRR	0.732	0.662
WTK	0.691	0.621
GWA	0.633	0.483
WEM	**0.422**	**0.346**

and we define

$$\mathrm{bias}_s^{(m)} = \widehat{\mu}_s^{(m)} - \widehat{\mu}_s^{(\mathrm{obs})} \tag{27}$$

565 Unlike the row-level quantile analysis, GWA can be included here because only mean wind speed is required.

Figures 13 and 14 reproduce the GS boxplots from the benchmark section, now augmented with WEM. The behavior of the individual datasets has already been discussed in Sect. 5.1; here we focus on how the ML ensemble compares to that baseline. Two features stand out. First, WEM achieves the smallest median absolute bias and the narrowest interquartile range

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

among all products, indicating that it reduces both typical error and site-to-site variability relative to even the best-performing individual datasets. Second, the signed bias distribution for WEM is tightly centered on zero and nearly symmetric, in contrast to the skewed distributions of most underlying products. In other words, the ensemble does not simply track one dataset with a constant offset: it systematically corrects both positive and negative biases across sites.

To complement the distribution plots, Table 5 summarizes GS site–height mean and median absolute biases for each dataset and for WEM. The ensemble attains a mean absolute bias of $0.43\ ms^{-1}$ and a median absolute bias of $0.34\ \mathrm{m\,s^{-1}}$, substantially lower than the best-performing individual datasets: GWA ($0.63/0.48\ \mathrm{m\,s^{-1}}$ for mean/median) and WTK ($0.69/0.62\ \mathrm{m\,s^{-1}}$). The WTK-LED variants are the most biased, with mean absolute biases exceeding $1.0\ \mathrm{m\,s^{-1}}$. These numbers quantify the improvements visible in Fig. 13 and indicate that WEM reduces typical GS site–height errors by roughly one-third relative to the strongest single dataset.

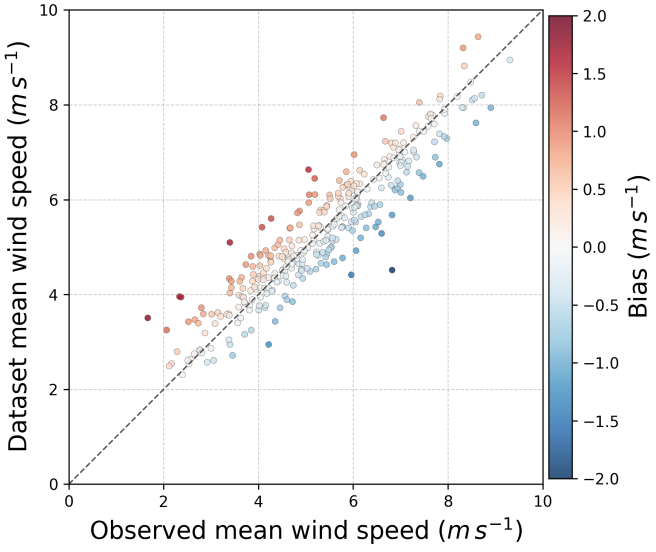


Figure 15. Parity plot of ML site-level mean wind speeds $\widehat{\mu}_s^{(\mathrm{ML})}$ versus observations $\widehat{\mu}_s^{(\mathrm{obs})}$. Marker colors encode $\mathrm{bias}_s^{(\mathrm{ML})}$ using a common color scale, so deviations from zero are directly comparable to the bias map in Fig. 16.

Parity and spatial bias

A site-level parity plot of $\widehat{\mu}_s^{(\mathrm{ML})}$ versus $\widehat{\mu}_s^{(\mathrm{obs})}$ (Fig. 15) reinforces this picture: most sites lie in a tight band around the 1:1 line, indicating that remaining mean wind errors are generally small, with only a modest number of clear over- or underestimation outliers. The corresponding ML bias map over GS sites (Fig. 16) shows a mix of weak positive and negative residuals with no large region dominated by a single sign, in contrast to the broad coherent patterns seen for several individual datasets. Together with the row-level results, this demonstrates that the quantile-based ensemble substantially reduces both spatially organized biases and site-level mean errors, while still leaving a few challenging locations where further refinement may be possible.

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

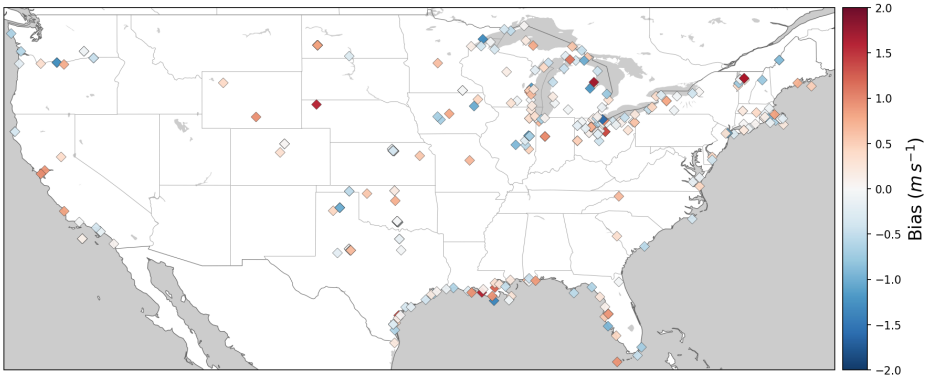


Figure 16. Spatial distribution of ML site-level biases $\mathrm{bias}_s^{(\mathrm{ML})}$ across GS sites, highlighting spatially incoherent residual biases.

Relative performance across datasets

To highlight differences relative to a specific dataset d, we define a site-level change in absolute bias,

$$\Delta|\mathrm{bias}|_s^{(\mathrm{ML}-d)} = |\mathrm{bias}_s^{(\mathrm{ML})}| - |\mathrm{bias}_s^{(d)}|, \tag{28}$$

so that negative values indicate that WEM has lower absolute bias than dataset d at site s. Figure 17 summarizes these differences for all baselines considered. Within each panel, site–height combinations are sorted from the largest improvement (most negative $\Delta|\mathrm{bias}|$) on the left to the largest degradation on the right, and the dashed vertical line marks the proportion of cases where WEM outperforms that dataset. Across the six baselines, WEM reduces absolute bias at 65 to 79 % of GS site–height combinations, with the largest gains relative to the two WTK-LED ensembles and substantial improvements relative to ERA5, HRRR, WTK, and GWA as well.

It is important to note that $\Delta|\mathrm{bias}|$ is not expected to be negative at every single site. The ensemble is trained to minimize average error across all towers, heights, and quantiles under a strict leave-one-site-out protocol with a 10 km buffer, not to reproduce the single best dataset at each individual location. As a result, a baseline product can occasionally "win" at a particular site simply by chance. In the WTK-LED CONUS panel, for example, most bars are negative, reflecting the dataset's strong positive bias and the resulting gains from correction, but a minority of bars on the right-hand side are positive. At those sites, the WTK-LED CONUS bias happens to compensate for errors in other inputs, placing it closer to the observations than the learned combination and yielding $\Delta|\mathrm{bias}|_s^{(\mathrm{ML}-\mathrm{WTK\text{-}LED\ CONUS})} > 0$ despite the dataset's poor performance overall. From a statistical perspective this behavior is expected: an essentially unbiased, variance-reducing ensemble cannot simultaneously beat every biased predictor at every point, especially when some predictors are occasionally accurate for the "wrong" reasons.

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

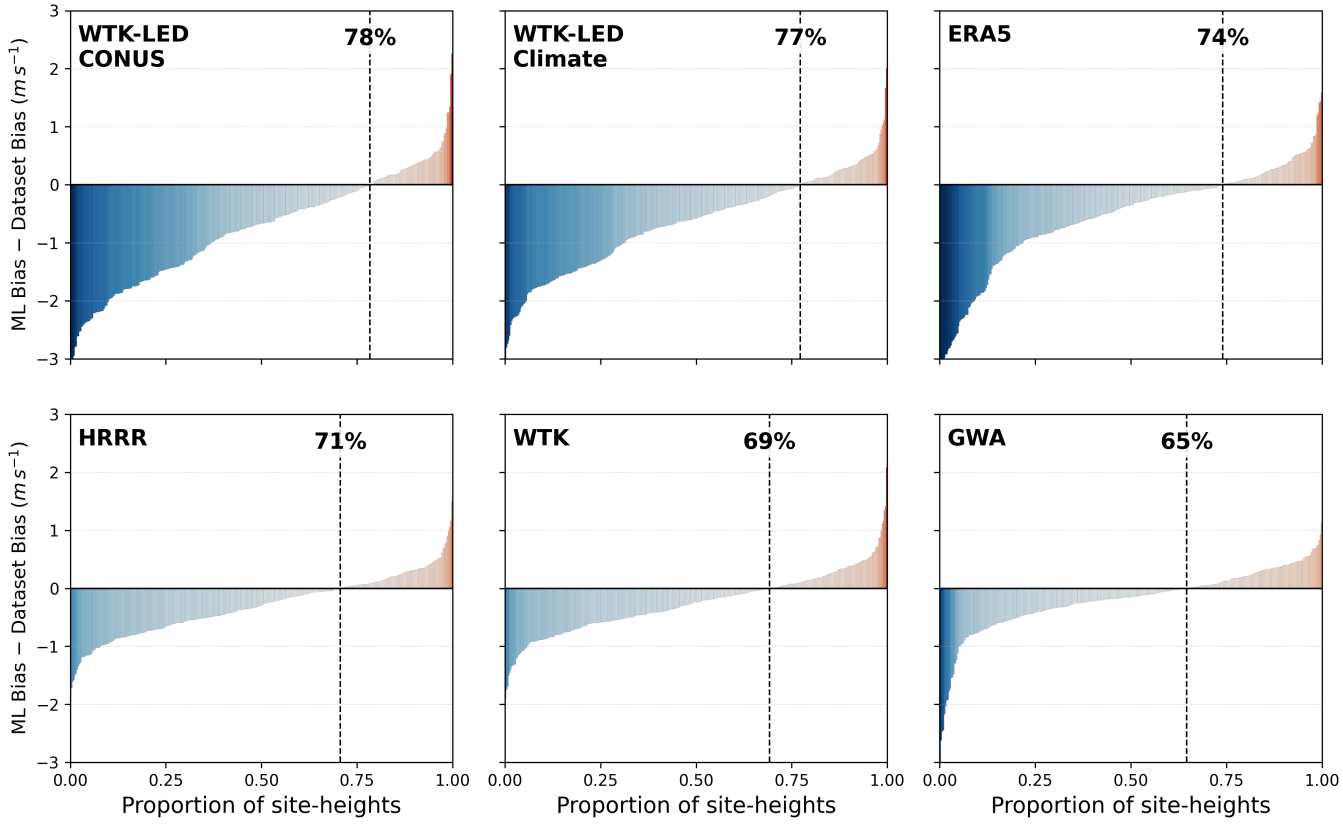


Figure 17. Site-level change in absolute bias for WEM relative to each baseline dataset. For a given dataset d, the quantity $\Delta|\mathrm{bias}|_s^{(\mathrm{ML}-d)} = |\mathrm{bias}_s^{(\mathrm{ML})}| - |\mathrm{bias}_s^{(d)}|$ is computed for every GS site–height combination and sorted along the horizontal axis. Negative values (blue) indicate that WEM has lower absolute bias than the baseline at that site–height combination, while positive values (red) indicate higher absolute bias. The dashed vertical line and percentage in each panel mark the fraction of site–height combinations where WEM improves upon the corresponding dataset, ranging from 65% (relative to GWA) to 78% (relative to WTK-LED CONUS).

5.2.6 CONUS mean wind maps from public datasets and the ensemble

While the station-based evaluation focuses on point locations, many applications require spatially continuous estimates of mean wind speed. To illustrate how the public datasets and the learned ensemble behave at the grid scale, we compute mean winds on the CONUS grid for each product and visualize them at a hub height of 60 m (Fig. 18).

Across all products, the familiar high-wind corridor from the central Great Plains into the northern interior West is evident, along with weaker winds across much of the Southeast and mid-Atlantic. The maps differ mainly in amplitude and texture. ERA5 shows a relatively smooth, large-scale pattern due to its lower resolution; GWA, HRRR, and WTK introduce finer structure tied to topography; and WTK-LED CONUS is clearly more energetic over broad regions, consistent with its positive bias in the station-based benchmark. The WEM map retains the major spatial features of the underlying datasets but moderates

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

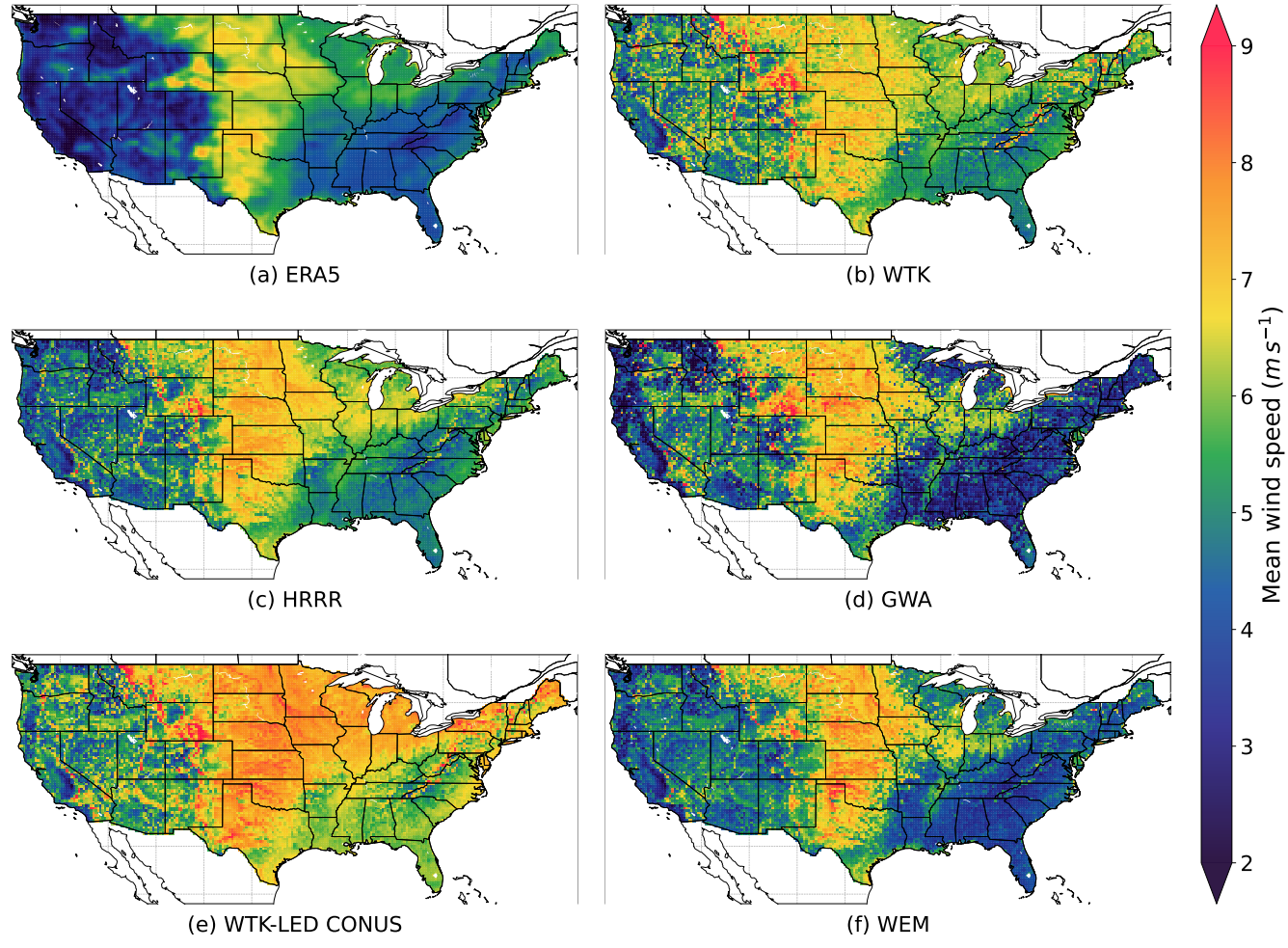


Figure 18. Full-grid mean wind speed at 60 m for six datasets: ERA5, WTK, HRRR, GWA, WTK–LED CONUS, and WEM. All maps share a common color scale for direct comparison.

the extremes, particularly in regions where individual products disagree, yielding a visually plausible blend of the large-scale patterns from reanalyses and the small-scale detail from high-resolution resource grids.

615 To show how the learned climatology varies with height, Fig. 19 presents WEM mean wind maps at 30, 40, 50, 60, 80, and 100 m. The spatial pattern is vertically coherent: high-wind regions strengthen and expand with height, especially across the central Plains and interior mountain west, whereas low-wind areas in the Southeast and parts of the Northeast remain comparatively weak. This progression reflects the expected increase in wind speed with altitude and demonstrates that the ensemble produces a smooth, physically consistent vertical structure suitable for hub-height extrapolation.

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

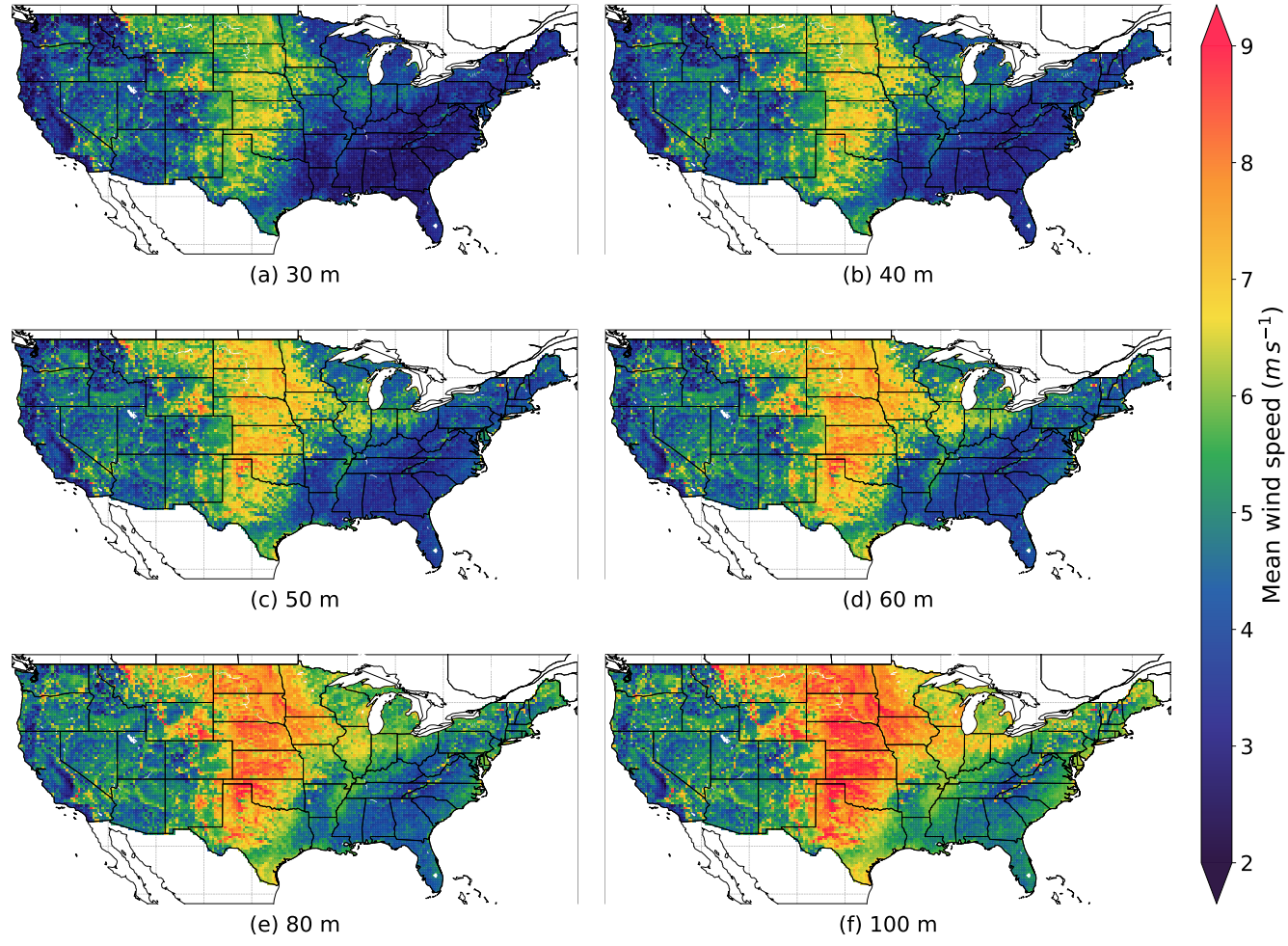


Figure 19. Vertical structure of the ML ensemble's full-grid mean wind speed field across six hub heights (30 to 100 m). The ensemble produces smooth, physically consistent height dependence across the domain.

6 Deployment, full-grid inference, and data products

To produce maps that are useful beyond the observational network, we apply the final WEM to every land grid point in the ERA5-based CONUS domain. For deployment, the model is retrained on the full set of ASOS and GS sites (without geographic holdouts), and then evaluated on a prediction table that mirrors the structure used during training.

We adopt the ERA5 grid as a pragmatic compromise between coverage and computational cost. Running the system on a much finer mesh (e.g., the native WTK grid) would require loading, processing, and running inference on millions of points, whereas the ERA5 grid comprises only tens of thousands of land cells over CONUS, making it feasible to generate multi-height, multi-quantile products with modest computing resources. Importantly, nothing in the methodology ties it to this

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

resolution: the features used by WEM (quantiles from the driving datasets, elevation, height, and latitude–longitude) can be evaluated on any grid, and the predicted quantile fields can be horizontally and vertically interpolated where needed. Extending the system to higher-resolution grids is therefore a natural direction for future work.

For each ERA5 grid cell and each target height (30, 40, 50, 60, 80, and 100 m), we summarize the time series from the driving wind datasets into empirical quantiles on the same $q = 0, \ldots, 100$ grid used in the benchmark. In practice this yields, for every grid–height–quantile combination, a set of long-term wind speed quantiles from WTK, HRRR, and WTK-LED CONUS, together with the quantile index itself q. Elevation is sampled from the USGS 3D Elevation Program (3DEP) Digital Elevation Model at each grid cell, and latitude, longitude, and height are included explicitly so that the model can express broad-scale geographic gradients and vertical structure.

The GWA is incorporated as an additional large-scale climatological predictor. Mean wind speeds from the GWA tiles at several reference heights (10, 50, 100, 150 m) are sampled at each grid cell and interpolated in height to the target levels, producing a single vertically consistent GWA mean for each grid–height pair. Joining these fields with the quantile summaries and terrain variables yields a uniform inference table with one row per (grid cell, height, quantile), containing exactly the feature set used in training.

Running WEM on this table produces a predicted quantile curve $\widehat{u}(q)$ at each ERA5 land grid point and height. From these curves we derive the full set of data products described below, including maps of mean wind speed obtained by integrating the predicted quantiles and, optionally, turbine-specific energy estimates obtained by integrating power curves against the same distributions.

6.1 Data representation for scalable access

To make the full-grid WEM data available for scalable, low-latency access, we store it in a location-indexed layout in Amazon S3, with one compressed file per grid cell keyed by a unique integer identifier. Each object holds the data for that location in a compressed tabular format. This layout avoids mixing unrelated locations in the same file and preserves spatial locality in the storage layer. Interactive queries first map user-specified coordinates to a location index and then read only the corresponding object-level partition. Combined with Athena's partition projection, this design reduces both scanned data volume and metadata overhead and allows partitions to be pruned without explicit registration. As a result, query cost and latency scale with the length of a single location's record rather than the size of the global dataset, making the system well suited for interactive, high-concurrency geospatial applications, including the WindWatts web interface described below.

Real-time access to this layout is provided by the open-source Python package `windwatts-data` (https://github.com/NatLabRockies/windwatts-data), which wraps S3 object storage and the Athena serverless query engine behind a simple API for querying, filtering, and downloading wind speed data. The package handles spatial lookup (mapping coordinates to the nearest grid point and index), executes the underlying SQL queries, and returns results as analysis-ready tables, without requiring users to manage storage layout or table definitions. To control cost under repeated interactive use, query outputs are cached in a user-managed S3 results bucket; repeated calls with identical parameters within a 7-day window reuse stored results and avoid additional Athena charges. A specialized implementation underpins the WindWatts web application, while the public

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

release is aimed at researchers working in local environments such as Jupyter notebooks and is accompanied by configuration files, table definitions, and usage examples to support reproducibility and community extension.

6.2 Access via the WindWatts web application

These gridded products are exposed to end users through the public WindWatts web application (https://windwatts.nrel.gov), which sits on top of the location-indexed storage and `windwatts-data` access layer described above. The interface is built on the same ERA5 grid used in this study: users interact with maps of long-term wind climatology, select a location of interest, and view the corresponding distributional summary at standard heights.

By default, WindWatts displays the baseline ERA5 climatology. In the settings panel, users can enable the "ensemble" option, which switches the map and site-level summaries to the WEM predictions evaluated on the ERA5 grid. In both modes, the tool reports statistics derived from the underlying quantile distributions, allowing practitioners to compare the raw reanalysis and ML-corrected products in a consistent framework and to incorporate the ensemble results directly into their own resource assessment workflows.

7 Discussion

The benchmark shows that existing public wind resource datasets already contain substantial information for U.S. wind assessment, but none is uniformly reliable across heights, regions, and wind speed regimes. At 10 m, WTK and HRRR perform best on the ASOS network, and at hub height GWA performs best on the GS towers. Across both networks, the WTK-LED products are strongly and consistently positively biased, and ERA5 exhibits larger spread and mixed signs of bias.

The datasets are also not redundant. Station-level bias maps reveal geographically coherent but dataset-specific structures, and scatterplots versus observed means show systematic differences across the wind speed range: for example, WTK and HRRR tend to overestimate low-wind sites and underestimate high-wind sites, whereas the WTK-LED products behave more like an additive positive offset. Because these patterns differ from product to product, the joint vector of model outputs at a site forms a characteristic "fingerprint" rather than a single consensus estimate. This motivates treating the datasets as complementary predictors in a learning system rather than selecting a single "best" map.

WEM is designed to exploit this complementarity. Trained in quantile space on multiple wind datasets together with elevation, height, and geographic coordinates, it learns how biases vary with both regime and location. On held-out GS rows, WEM achieves the lowest row-level mean absolute error and essentially zero mean bias among all products, and its mean bias as a function of quantile is nearly flat across the distribution (Sect. 5.2.4), in contrast to the strongly regime-dependent curves of individual datasets. When predicted quantiles are integrated to site-level means, the ensemble again outperforms all individual datasets: boxplots over GS sites show the smallest median absolute bias and the narrowest interquartile range, and the signed bias distribution is tightly centered on zero.

Spatial diagnostics support the same interpretation. Parity plots of ensemble mean wind speed versus observations show most GS sites lying close to the 1:1 line, and the ensemble bias map lacks the broad, coherent error patterns visible in several

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

of the baselines. Site-level comparisons of absolute bias indicate that WEM improves upon each dataset at a majority of towers while leaving a minority of locations where a given baseline happens to agree more closely with the observations. This behavior is expected for a data-driven model optimized for performance *on average* across a finite, noisy network.

Taken together, these findings suggest a clear division of roles. Public wind resource datasets provide the physically grounded, spatially complete representations of the flow, each with its own strengths and weaknesses. The ML quantile ensemble does not replace these products; instead, it acts as a calibrator that combines their partially independent information with topography to reduce systematic bias and tighten uncertainty, particularly in regimes where individual datasets disagree or show strong, structured errors.

7.1 Role of ASOS and GS in training

The combined use of 10 m ASOS data and tall GS measurements is central to the design and performance of WEM, because the two archives occupy complementary niches. The ASOS network provides a dense, nationally distributed set of near-surface measurements with relatively homogeneous instrumentation and long, consistent records, anchoring the model at 10 m and constraining large-scale horizontal gradients across most of the CONUS domain, including many regions that lack any tall towers. The GS collection, by contrast, consists of targeted measurements at or near turbine hub heights, often in wind-exposed or complex terrain, and therefore directly constrains the regime in which projects operate. Relying exclusively on one source would be unsatisfactory: training only on GS sites would tightly tune hub-height behavior but leave near-surface gradients and much of the national domain underconstrained, and training only on ASOS would yield good 10 m calibration but little information about vertical structure or turbine-level winds. By combining the two, with explicit inclusion of height as a feature, the model simultaneously learns horizontal and vertical patterns.

7.2 Advantages of quantile-based modeling

A central design choice in WEM is to model the full wind speed distribution via quantiles rather than predict only a mean. The diagnostics in Sect. 5.2.4 show strong quantile-dependent biases in the baseline datasets: ERA5 tends to underestimate the upper tail while WTK and HRRR overestimate the mid-range quantiles, and the WTK-LED products become increasingly positively biased at higher quantiles. By contrast, the WEM bias curve is nearly flat across $q = 0$ to 100, with mean bias close to zero at every quantile. This behavior cannot be reproduced by a single additive or multiplicative correction to a baseline mean; it requires a model that can adjust the entire distribution.

The quantile representation also provides a direct bridge between wind speed distributions and energy-relevant quantities. Long-term mean wind speed is recovered by integrating the quantile function, and the same construction applies to any turbine power curve by integrating the power curve against the predicted quantiles. In practice, this allows a single model output to support both climatological bias analysis and turbine-specific energy estimates. The strong reductions in high-quantile bias are particularly important, because upper-tail winds contribute disproportionately to expected power. Including the quantile index as a feature enables WEM to learn how corrections should vary across the distribution, conditioned on height, terrain, and the joint behavior of the input datasets, rather than applying a single global adjustment.

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

7.3 Limitations and future work

Despite the encouraging performance of WEM and the breadth of the benchmark, several limitations remain. First, observational coverage is uneven. The GS sites, while invaluable for hub-height validation, are geographically sparse and concentrated in existing or prospective wind development regions, and important regimes (e.g., some mountainous, forested, and offshore areas) are underrepresented. The dense ASOS network samples near-surface winds at airports rather than wind-optimal exposures. Generalization in data-sparse regimes is therefore constrained by the quality and diversity of the public wind datasets themselves, and expanded archives that include more tall towers, offshore platforms, and remote-sensing campaigns would improve calibration and regional assessment.

Second, the quantile targets and predictions are treated as deterministic, even though empirical quantiles from finite time series are noisy, especially in the tails. We do not propagate this sampling uncertainty through training or provide formal uncertainty bands beyond cross-validation spread. Incorporating uncertainty-aware loss functions, ensemble or Bayesian methods, or parametric distribution fits informed by the learned quantiles would allow more explicit characterization of predictive uncertainty.

Third, although the feature suite already combines multiple wind datasets, topography, and location, additional predictors (e.g., land cover, vegetation, high-resolution roughness, or explicit coastal and complex-terrain indicators) and alternative model families (such as spatial neural networks, quantile regression forests, or hybrid physics-informed approaches) could further improve performance in challenging regimes. Similarly, the current cross-validation design focuses on geographic generalization around GS sites but does not explicitly test temporal transferability or transfer to regions far outside the training envelope, and the ensemble remains anchored in the shared limitations of the underlying public datasets.

Overall, the present study demonstrates that a carefully designed, quantile-based ML ensemble can substantially reduce site-level bias and improve distributional fidelity relative to leading public wind datasets while remaining transparent about its assumptions and limitations. The resulting national quantile field is best interpreted as an informed synthesis of existing resources and observations, not a replacement for local measurement campaigns when project-critical decisions are at stake. As observational networks expand, public datasets evolve, and additional predictors become available, the framework presented here can be updated to provide progressively more accurate and informative tools for wind resource assessment and planning.

8 Conclusions

Accurate, site-resolved wind climatology is fundamental for resource assessment, yet public datasets vary widely in performance and have lacked a unified, observation-based evaluation framework. Using a large, quality-controlled archive, we developed both a consistent benchmark of major public wind products and an ML quantile ensemble, WEM, that fuses multiple datasets with static predictors. The benchmark reveals pronounced, dataset-specific spatial and regime-dependent biases. By learning directly in quantile space under strict leave-one-site-out validation, WEM outperforms every publicly available dataset, substantially reducing these structured errors: quantile-dependent biases flatten, site-level mean biases contract toward zero, and regional over- and underestimation patterns are strongly diminished across tall-tower sites. Deploying the ensemble

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

760 on the ERA5 grid yields a national, multi-height climatology now accessible through the WindWatts application. This combination of a rigorous benchmark, a distribution-focused learning framework, and a reproducible large-scale product advances the accuracy and practical utility of public wind resource information.

Code and data availability. All data from atmospheric models used in this study are publicly available. Due to the proprietary nature of measurement data from commercial meteorological towers, our benchmark dataset cannot be released. Ensemble model results are available
765 for use through https://windwatts.nrel.gov API and web interface. Final model training and testing software will be made available at https://github.com/NatLabRockies/windwatts-ensemble following publication.

Author contributions. All authors contributed to concept development, writing and editing. KM led core research tasks including model development, software development, testing and benchmarking. LS, CP and DD contributed to problem formulation, metrics, and statistical design. SS led implementation tasks including backend software development.

770 *Competing interests.* The authors do not have competing interests to report.

Acknowledgements. This work was authored in part by the National Laboratory of the Rockies for the U.S. Department of Energy (DOE), operated under Contract No. DE-AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Wind Energy Technologies Office. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S.
775 Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes. A portion of the research was performed using computational resources sponsored by the Department of Energy and located at the National Laboratory of the Rockies.

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

References

Abad-Santjago, Á., Peláez-Rodríguez, C., Pérez-Aracil, J., Sanz-Justo, J., Casanova-Mateo, C., and Salcedo-Sanz, S.: Hybridizing Machine Learning Algorithms With Numerical Models for Accurate Wind Power Forecasting, Expert Systems, 42, e13 830, 2025.

Amato, F., Guignard, F., Walch, A., Mohajeri, N., Scartezzini, J.-L., and Kanevski, M.: Spatio-temporal estimation of wind speed and wind power using extreme learning machines: predictions, uncertainty and technical potential, Stochastic Environmental Research and Risk Assessment, 36, 2049–2069, 2022.

Buster, G., Pinchuk, P., Lavin, L., Benton, B., and Bodini, N.: Bias Correcting NOAA's High-Resolution Rapid Refresh (HRRR) Wind Resource Data for Grid Integration Applications, Tech. rep., National Renewable Energy Laboratory (NREL), Golden, CO (United States), 2024.

Cannon, A. J., Sobie, S. R., and Murdock, T. Q.: Bias correction of GCM precipitation by quantile mapping: how well do methods preserve changes in quantiles and extremes?, Journal of Climate, 28, 6938–6959, 2015.

Carvalho, D., Rocha, A., Gómez-Gesteira, M., and Santos, C. S.: 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, Applied Energy, 135, 234–246, 2014.

Davis, N. N., Badger, J., Hahmann, A. N., Hansen, B. O., Mortensen, N. G., Kelly, M., Larsén, X. G., Olsen, B. T., Floors, R., Lizcano, G., et al.: The Global Wind Atlas: A high-resolution dataset of climatologies and associated web-based application, Bulletin of the American Meteorological Society, 104, E1507–E1525, 2023.

Dowell, D. C., Alexander, C. R., James, E. P., Weygandt, S. S., Benjamin, S. G., Manikin, G. S., Blake, B. T., Brown, J. M., Olson, J. B., Hu, M., et al.: The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part I: Motivation and system description, Weather and Forecasting, 37, 1371–1395, 2022.

Draxl, C., Clifton, A., Hodge, B.-M., and McCaa, J.: The wind integration national dataset (wind) toolkit, Applied Energy, 151, 355–366, 2015.

Draxl, C., Wang, J., Sheridan, L., Jung, C., Bodini, N., Buckhold, S., Aghili, C. T., Peco, K., Kotamarthi, R., Kumler, A., et al.: Wtk-led: The wind toolkit long-term ensemble dataset, Tech. rep., National Renewable Energy Laboratory (NREL), Golden, CO (United States), 2024.

Drew, D. R., Barlow, J. F., Cockerill, T. T., and Vahdati, M. M.: The importance of accurate wind resource assessment for evaluating the economic viability of small wind turbines, Renewable Energy, 77, 493–500, 2015.

Haben, S., Ward, J., Greetham, D. V., Singleton, C., and Grindrod, P.: A new error measure for forecasts of household-level, high resolution electrical energy consumption, International Journal of Forecasting, 30, 246–256, 2014.

Haq, I. U., Kumar, A., and Rathore, P. S.: Machine learning approaches for wind power forecasting: a comprehensive review, Discover Applied Sciences, 7, 1139, 2025.

He, J., Li, Q., Chan, P., and Zhao, X.: Assessment of future wind resources under climate change using a multi-model and multi-method ensemble approach, Applied Energy, 329, 120 290, 2023.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.: The ERA5 global reanalysis, Quarterly journal of the royal meteorological society, 146, 1999–2049, 2020.

Hu, W., Scholz, Y., Yeligeti, M., von Bremen, L., and Deng, Y.: Downscaling ERA5 wind speed data: a machine learning approach considering topographic influences, Environmental Research Letters, 18, 094 007, 2023.

https://doi.org/10.5194/wes-2026-75
Preprint. Discussion started: 20 May 2026

Indra, T., Bain, D., and Acker, T. L.: Wind energy modeling for residential-scale wind power and sensitivity of economic valuation to errors in wind speed estimates, in: 43rd ASES National Solar Conference 2014, SOLAR 2014, Including the 39th National Passive Solar Conference and the 2nd Meeting of Young and Emerging Professionals in Renewable Energy, pp. 621–625, American Solar Energy Society, 2014.

Olauson, J.: ERA5: The new champion of wind power modelling?, Renewable energy, 126, 322–331, 2018.

Peherstorfer, B., Willcox, K., and Gunzburger, M.: Survey of multifidelity methods in uncertainty propagation, inference, and optimization, Siam Review, 60, 550–591, 2018.

Phillips, C., Sheridan, L., Conry, P., Fytanidis, D. K., Duplyakin, D., Zisman, S., Duboc, N., Nelson, M., Kotamarthi, R., Linn, R., et al.: Evaluation of obstacle modelling approaches for resource assessment and small wind turbine siting: case study in the northern Netherlands, Wind Energy Science Discussions, 2022, 1–27, 2022.

Phillips, C., Duplyakin, D., Sheridan, L., Ruzekowicz, J., Nelson, M., Fytanidis, D., Linn, R., Kotamarthi, R., and Tinnesand, H.: Resource Assessment for Distributed Wind Energy: An Evaluation of Best-Practice Methods in the Continental US, in: Journal of Physics: Conference Series, vol. 2767, p. 092005, IOP Publishing, 2024.

Powers, J. G., Klemp, J. B., Skamarock, W. C., Davis, C. A., Dudhia, J., Gill, D. O., Coen, J. L., Gochis, D. J., Ahmadov, R., Peckham, S. E., et al.: The weather research and forecasting model: Overview, system efforts, and future directions, Bulletin of the American Meteorological Society, 98, 1717–1737, 2017.

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.

Sheridan, L. M., Duplyakin, D., Phillips, C., Tinnesand, H., Rai, R. K., Flaherty, J. E., and Berg, L. K.: Evaluating the potential of short-term instrument deployment to improve distributed wind resource assessment, Wind Energy Science Discussions, 2024, 1–28, 2024.

Sheridan, L. M., Wang, J., Draxl, C., Bodini, N., Phillips, C., Duplyakin, D., Tinnesand, H., Rai, R. K., Flaherty, J. E., Berg, L. K., et al.: Performance of wind assessment datasets in United States coastal areas, Wind Energy Science, 10, 1551–1574, 2025.

Stengel, K., Glaws, A., Hettinger, D., and King, R. N.: Adversarial super-resolution of climatological wind and solar data, Proceedings of the National Academy of Sciences, 117, 16 805–16 815, 2020.

Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G. R., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., Atencia, A., et al.: Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world, Bulletin of the American Meteorological Society, 102, E681–E699, 2021.

Vasiljević, N., Harris, M., Tegtmeier Pedersen, A., Rolighed Thorsen, G., Pitter, M., Harris, J., Bajpai, K., and Courtney, M.: Wind sensing with drone-mounted wind lidars: proof of concept, Atmospheric Measurement Techniques, 13, 521–536, 2020.

Xu, W., Ning, L., and Luo, Y.: Wind speed forecast based on post-processing of numerical weather predictions using a gradient boosting decision tree algorithm, Atmosphere, 11, 738, 2020.

Zscheischler, J., Fischer, E. M., and Lange, S.: The effect of univariate bias adjustment on multivariate hazard estimates, Earth system dynamics, 10, 31–43, 2019.