# Prediction of hourly wind speed time series at unsampled locations using machine learning

Freddy Houndekindo[1*], Taha B.M.J. Ouarda[1]

[1]Canada Research Chair in Statistical Hydro-Climatology, Institut national de la recherche scientifique, Centre Eau Terre Environnement, INRS-ETE, 490 de la Couronne, Québec, QC, G1K 9A9, Canada

*Corresponding author: Freddy Houndekindo
490, Couronne street, Québec, QC, G1K 9A9, Canada
Tel: +1 418-654-3842
E-mail: freddy.houndekindo@inrs.ca

## Abbreviations

| | |
|---|---|
| **a.g.l** | Above ground level |
| **CDF** | Cumulative distribution function |
| **DEM** | Digital elevation model |
| **ECCC** | Environment and Climate Change Canada |
| **ERA5-WSQ** | Wind speed quantiles extracted from the ERA5 dataset (m/s) |
| **GWA** | Global wind atlas |
| **GWA-ERA5** | Bias-corrected ERA5 using GWA (m/s) |
| **IAV** | Interannual variability |
| **IDW** | Inverse distance weighting |
| **LGBM** | Light gradient-boosting machine |
| **LGBMQR** | Lightgbm for quantile regression |
| **LGBMSI** | LGBM for spatial interpolation |
| **LGBMSI-ERA5** | LGBMSI using the ERA5 wind data as covariates |
| **LGMBQR-ERA5** | LGBMQR using ERA5-WSQ as covariates |
| **MAE** | Mean absolute error (m/s) |
| **ME** | Mean error (m/s) |
| **MRMR** | Minimum redundancy maximum relevancy algorithm |
| **OP** | Overlap percentage (%) |
| **PC** | Pearson correlation |
| **PD** | Probability distribution |
| **QM** | Quantile mapping |
| **QM-ERA5** | Quantile mapping bias correction of ERA5 wind data |
| **QR** | Quantile regression |
| $\mathbf{R^2}$ | Coefficient of determination |
| **RCov** | Robust coefficient of variation |
| **RFSI** | Random forest for spatial interpolation |
| **RMSE** | Root-mean-squared error (m/s) |
| **TS** | time series |
| **WDC** | Wind Duration Curve method |
| **WRA** | Wind resource assessment |
| **WS** | Wind speed |
| **WSD** | Wind speed distribution |
| **WSNEP** | Wind speed non-exceedance probabilities |
| **WSQ** | Wind speed quantiles |
| **WSTS** | Wind speed time series |

**Abstract**

Various models for wind speed mapping have been developed, with increasing attention on models focusing on mapping wind speed distribution. This study extends these models to predict hourly wind speed time series at unsampled locations. A model based on the quantile mapping (QM) procedure was compared to a traditional and machine-learning model to interpolate wind speed spatially. These proposed models were also used with inputs from the ERA5 reanalysis dataset, enabling them to consider local variation in orography and large-scale wind fields. A widely used procedure for mean bias correction of reanalysis based on the Global Wind Atlas (GWA) was implemented and compared to the proposed models. It was found that the QM and machine learning model, both using input from ERA5, significantly outperformed GWA bias correction in terms of time series correlation and probability distribution. Despite being more computationally intensive than GWA bias correction, both models are recommended due to their significantly (in a statistical sense) superior performance.

**Keywords:** Bias-correction, ERA5, Light gradient-boosting machine, Quantile regression, Reanalysis, Wind resource assessment

## 1. Introduction

The past decades have witnessed a significant uptake of wind energy in various parts of the world [1]. This growth reflects a global shift toward more renewable energy sources, with wind power playing a prominent role in energy supply [2]. The intermittent nature of wind speed still poses some challenges to the development of the renewable energy source [3]. Due to the cubic relationship between wind speed and power output, inaccuracies in estimating wind speed are amplified when estimating the energy production, leading to suboptimal design of wind energy infrastructure and jeopardizing the profitability and sustainability of the project [4].

Prospective studies to evaluate the wind resource across a large region at a high spatial and temporal resolution provide valuable sources of information for the expansion of wind energy [5, 6]. In-situ wind speed (WS) data are generally accepted as the most reliable data source for wind resource assessment (WRA). However, measuring stations are often sparsely available in a given region and have limited record length for WRA. Several publicly available datasets exist that give access to wind data at the global scale with high temporal resolution and extensive record length. The European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis v5 [ERA5: 7], and the NASA's Modern Era Retrospective Analysis for Research and Applications-2 [MERRA-2: 8] have been used extensively to conduct WRA across large regions [9, 10]. Samal [11] evaluated the adequacy of MERRA-2 for WRA in India. The author compared the wind data from the reanalysis dataset with observed data collected at meteorological stations. The study found that the reanalysis dataset was more suitable for long-term than short-term planning. In another study, MERRA-2 was used to perform a preliminary evaluation of the wind resource in South Sudan [12]. The authors identified areas in the region with high wind potential. Five global reanalysis datasets including ERA5 and MERRA-2 were evaluated for WRA by comparing them with measured WS data from meteorological stations distributed worldwide [13]. The comparative study was based on estimated mean WS, variability, and trends. From the study results, the ERA5 dataset was recommended for wind energy applications.

Direct application of reanalysis datasets for WRA still has some drawbacks. Notably, the coarse spatial resolution of reanalysis datasets renders them unable to resolve local variations in orography and surface roughness influencing near-surface WS [14]. A review of the uncertainties associated with the application of reanalysis data for WRA was presented by Gualtieri [9]. Several studies endeavoured to increase the spatial resolution and bias-correct reanalysis datasets using ground measurements and other datasets with higher spatial resolution. The Global Wind Atlas (GWA) is a popular dataset used to correct the bias in reanalysis WS data [15]. In this procedure, the mean WS from the reanalysis dataset is corrected to match the GWA mean WS by applying a correction factor estimated during the overlapping period of both datasets.

Alternatively, to reanalysis datasets, spatial interpolation and machine learning models have been used to map wind data at a high spatial resolution using in-situ observations. The main advantage of this approach over the use of reanalysis data is its ability to account for the rapid change in the topography and surface roughness by using covariates extracted from DEM and land use maps. A comparative analysis of several spatial interpolation methods for hourly WS mapping was performed by Collados-Lara, et al. [16]. The authors found that the regression kriging model produced the best results and was selected to generate hourly wind speed time series (WSTS) between 1996 and 2016 in The Granada province, Spain.  In another study, Cellura, et al. [17] developed a machine-learning model to interpolate mean WS in Sicily, Italy. The author recommended the approach for its ease of application and transferability to other regions. A similar study was conducted in Venezuela [18] to create a regional mean WS map. It should be noted that wind speed distribution (WSD) is often skewed, and the mean is not a good representative of the most typical value of the distribution.

In recent studies, authors have been interested in mapping the entire WSD, allowing a better evaluation of the wind resource variability at unsampled locations of interest. For example, Veronesi, et al. [19] mapped the parameters of the Weibull distribution fitted to WS data across the United Kingdom (UK). Jung [20] mapped the parameter of the Wakeby distribution fitted to WS data to estimate the annual wind energy yield with a high spatial resolution in Germany. In another study, Jung and Schindler [21] developed a global model that estimates the parameters of the Kappa and Wakeby distribution for WS variability assessment using estimated L-moments. Houndekindo and Ouarda [22] recently proposed a nonparametric approach for WSD mapping. The approach does not restrict the region to a single WSD distribution family. The availability of methods to map the entire WSD is a crucial step forward compared to past studies where only aggregated values of WS were estimated. However, For the evaluation of WS variability at different temporal resolutions (ex., daily, seasonal, annual), WSTS with a high temporal resolution (e.g., ten min. or one hour) are still required.

This study proposes expanding upon previously developed techniques for mapping WSD to predict hourly WSTS at unsampled locations. The proposed method named the Wind Duration Curve (WDC) is inspired by an approach commonly used for environmental variables (see, for instance, Castellarin, et al. [23] and Requena, et al. [24] for application to streamflow data and Ouarda, et al. [25] for application to daily river temperature) and can be seen as an adaptation of the quantile mapping (QM) technique often used to downscale global circulation models and regional climate model outputs [26, 27]. A comprehensive evaluation of the WDC method is performed and the approach is compared to other methods for WSTS estimation at unsampled locations.

The paper is structured as follows: Section 2 describes the study area and the datasets. The methodology employed is presented in section 3. The results of the comprehensive evaluation of the different approaches are presented in section 4. The discussion follows in section 5, and section 6 gives the conclusions of the study.

## 2. Study area and dataset

Experimental data for the study were obtained from Environment and Climate Change Canada (ECCC) historical climate database (https://climate.weather.gc.ca/). Stations with less than 10% missing values between 2011 and 2021 (11 years of mean hourly WS) were selected from the database, resulting in 303 meteorological stations available for the study. WS data at the meteorological stations were typically collected at 10 m above ground level according to ECCC. The measured WS data was considered the most representative of the actual WS condition. Figure 1 illustrates the study area and the location of the 303 meteorological stations. In the figure, stations represented with circles were used during the training of the models and those represented with triangles were solely used as test samples.

Reanalysis WS data were obtained from ERA5 dataset. Wind speed data from ERA5 are provided in a grid format with a temporal resolution of 1 h available between 1980 and the present. The eastward and northward WS components at 10 m were obtained from the dataset (https://doi.org/10.24381/cds.adbb2d47), and the 10 m horizontal WS was calculated and interpolated at the 303 meteorological stations using nearest neighbor interpolation.

The WS covariates used in the study are presented in detail in Table S1 of the supporting material. Topographical covariates were calculated from the Advanced Land Observing Satellite (ALOS) Digital Elevation model (DEM) of 30m resolution [ALOS DEM: 28] obtained freely from the Japan Aerospace Exploration Agency. The surface roughness length was estimated from a 2015 land use map of Canada [29] obtained from Natural Resource Canada.

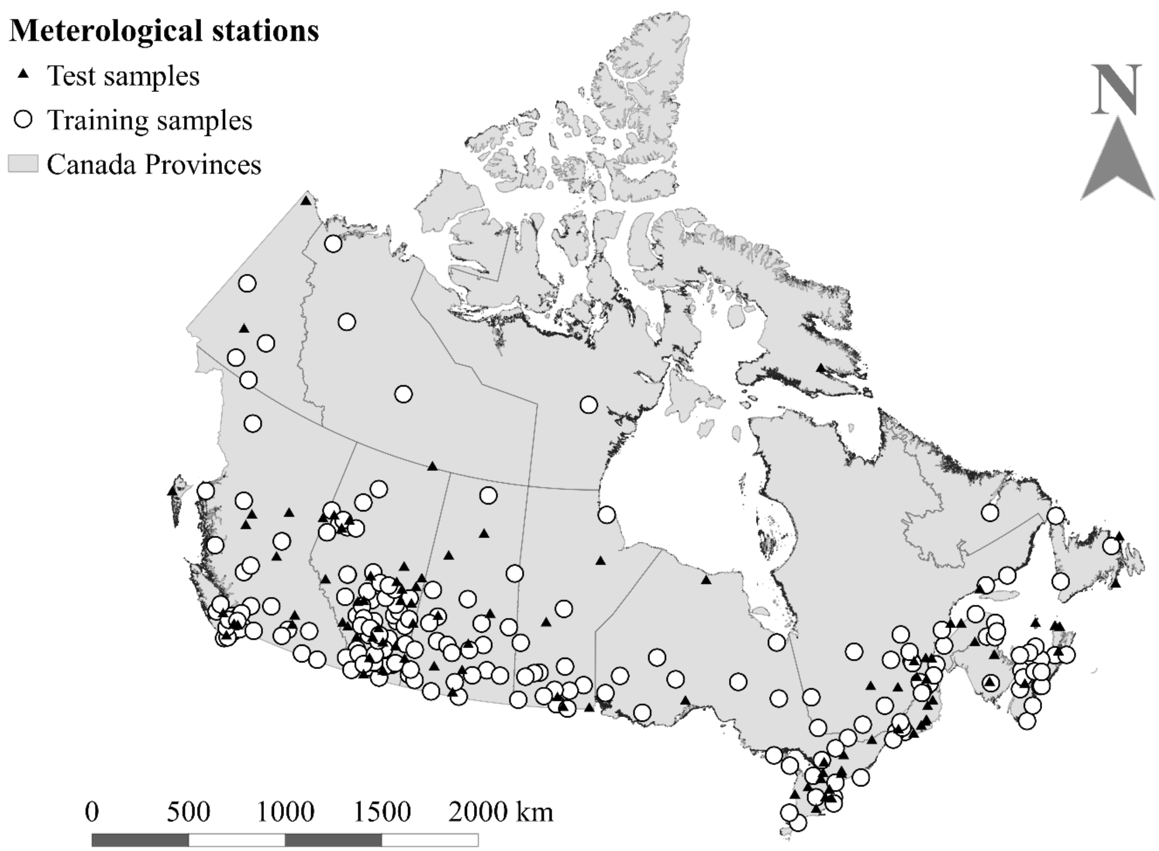



Figure 1: Study area and location of the 303 meteorological stations used in the study.

## 3. Methods

### 3.1. Wind speed distribution mapping

In recent studies, different methodologies to map WSD were introduced. Most of these approaches relied on mapping the parameters of a distribution function fitted to WS data. More recently, a nonparametric method was developed by Houndekindo and Ouarda [22] to map hourly WSD. The approach starts by mapping hourly wind speed quantiles (WSQ) using a machine learning model and WS covariates. Then, the estimated WSQ are used as input of an asymmetric kernel function to estimate the WS cumulative distribution function (CDF) at unsampled locations. The approach is flexible and does not restrict the region to a unique WSD family. In their study, Houndekindo and Ouarda [22] extracted 13 quantiles from observed WSTS and then built a regression model between the covariates and each WSQ. The present study proposes a quantile regression (QR) model to directly estimate 13 conditional WSQ. Although QR models have been used in previous studies for WS forecasting [30] and for the estimation of other hydro-climatic variables at unsampled locations [31], to the author's knowledge, it is the first time they are applied to estimate conditional WSQ at unsampled locations. As done by Houndekindo and Ouarda [22], WSQ at the following 13 percentile points were considered: 5.0% (P1), 12.5% (P2), 20.0% (P3), 27.5% (P4), 35.0% (P5), 42.5% (P6), 50.0% (P7), 57.5.0% (P8), 65.0% (P9), 72.5% (P10), 80.0% (P11), 87.5% (P12), and 95.0% (P13).

The Light Gradient-Boosting Machine [LGBM: 32] with the pinball loss function (Eq 1) was used as the QR model (herein referred to as LGBMQR). The LGBM was adopted based on its efficiency, scalability for large datasets, and proven high prediction accuracy [33-35]. The LGBM is a histogram-based gradient-boosting model that sequentially builds additive decision trees to minimize a loss function. By discretizing the continuous values of the covariates into a fixed number of bins, the LGBM can significantly reduce the training time and memory usage for large datasets (ex., N > 10,000) while maintaining good prediction accuracy. In addition, the LGMB adopts a leaf-wise tree expansion with a fixed maximum depth, improving the model's training performance. Table 1 shows the different model parameters that were tuned. Random search with 1000 iterations was used to select the best parameters for the QR model. Random search is not an optimal algorithm for parameter tuning but can still find suitable parameters when allocated a sufficient number of iterations [36]. LGBMQR is a single-output QR model. Thus, it needs to be trained separately for each conditional WSQ of interest. Also, parameter searches can be performed independently for each considered quantile. To reduce the computation burden associated with performing parameter tuning independently for every quantile of interest, the best parameters selected when training the model to predict the median (P7) were used for all quantiles.

$$\rho_\tau(w - w_\tau) = \begin{cases} (\tau - 1)|w - w_\tau| & (w - w_\tau) < 0 \\ \tau|w - w_\tau| & (w - w_\tau) \geq 0 \end{cases} \quad (1)$$

where $w_\tau$ is the τ-quantile defined as follows:

$$w_\tau = \inf\{w : F(w|X = x) \geq \tau\} \quad (2)$$

with $F(w|\, X = x)$ the conditional cumulative distribution function of the random variable $w$.

In addition to the covariates presented in Table S1 of the supporting material, hourly WSQ extracted from the ERA5 dataset (ERA5-WSQ) were assessed as covariates in the current study. As stated by Jung and Schindler [21], covariates from the ERA5 reanalysis dataset can represent the large-scale wind field unaffected by local surface properties. The LGBMQR that uses the ERA5-WSQ will be referred to as LGMBQR-ERA5, and the benefit of using the ERA5-WSQ as covariates will be evaluated and discussed in the following sections of the paper.

Furthermore, to select the optimal number of covariates to include in the model, the available covariates were ranked according to their relevance and redundancy using the minimum redundancy maximum relevancy algorithm [MRMR: 37]. Then, the number of covariates to use with LGBMQR and LGBMQR-ERA5 was treated as an additional hyperparameter during the implementation of the random search algorithm. The MRMR algorithm has already demonstrated good performance for WSQ mapping in a comparative study of covariate selection techniques [38].

The estimated conditional WSQs were used as input for the Birnbaum-Saunders asymmetric kernel estimator of CDF [39] to estimate the WS CDF at unsampled locations. For more details on fitting the Birnbaum-Saunders kernel using the WSQ as input, the readers are referred to Houndekindo and Ouarda [22].

Table 1: Parameters of LGBMQR and LGBMQR-ERA5. The same set of randomly selected parameters was tested for LGBMQR and LGBMQR-ERA5 to implement the random search.

| Model parameter | Description | Range |
|---|---|---|
| **learning_rate** | Learning rate | 0.02-0.1 |
| **max_depth** | Maximum depth of the regression trees | 3-8 |
| **feature_fraction** | Fraction of covariate to use to build each tree | 0.1-0.9 |
| **bagging_fraction** | Fraction of data to sample to build each tree | 0.1-0.9 |
| **extra_trees** | Use of extremely randomized trees [40] | True, False |
| **lambda_l2** | L2 regularization | 0-1000 |
| **lambda_l1** | L1 regularization | 0-1000 |
| **num_leaves** | maximum number of leaves per regression tree | 2-50 |
| **max_bin** | max number of bins for the discretization of the covariates | 50-400 |
| **min_data_in_leaf** | minimal amount of data in one leaf | 100-20000 |
| **num_boost_round** | Number of trees to build (boosting iteration) | 90-400 |
| **n_features** | Number of features to include in the model | 5-30 |

### 3.2. Prediction of wind speed time series at unsampled locations

It is proposed to adapt the QM [41] procedure to predict WSTS at unsampled locations using the following general formula [26]:

$$\widehat{w}_t(s_0) = \widehat{F}_{S_0}{}^{-1}\left[\widehat{F}(w_t)\right] \quad (3)$$

where: $\hat{w}_t(s_0)$ is the estimated WS at time $t$ and unsampled location $s_0$. $\hat{F}_{S_0}$ is the estimated WS CDF at the unsampled location $s_0$, and $\hat{F}_{S_0}{}^{-1}$ is its inverse. $\hat{F}(w_t)$ is the estimated wind speed non-exceedance probabilities (WSNEP) at time $t$. The methodology to estimate $\hat{F}_{S_0}$ at any unsampled location in the region was described in section 3.1. For the estimation of $\hat{F}(w_t)$ two approaches have been put forward in previous studies:

1. Some authors [25, 42] proposed using information from nearby locations to estimate $\hat{F}(w_t)$ at any unsampled location. This technique assumes that observed non-exceedance probabilities (or exceedance probabilities) between nearby locations are correlated. Thus, a spatial interpolation method could be applied to estimate the WSNEP at unsampled locations. The Inverse Distance Weighting (IDW) was used to interpolate the WSNEP. The method was named the flow duration curve and the temperature duration curve for streamflow and temperature modelling. Following this nomenclature, the technique was referred to as the Wind Duration Curve (WDC) in the context of WS modelling.
2. Jung and Schindler [43] derived the non-exceedance probabilities $\hat{F}(w_t)$ directly from a reanalysis dataset, thereby performing bias correction. This approach will be applied with ERA5 and named quantile mapping bias correction of ERA5 (QM-ERA5) in the following sections.

The Weibull plotting position was used to estimate the WSNEP from the WSTS as follows:

$$F_n(w_t) = {}^{i_t}/_{n+1} \tag{4}$$

where: $i_t = 1, 2, 3, \ldots, n$ is the rank of the WS value observed at time t ($w_t$) after sorting the time series in ascending order.

### 3.3. Spatial interpolation methods

Two spatial interpolation methods were selected and evaluated to interpolate the WSTS directly. The IDW technique was selected for its ease of application and set as the baseline method in the study. The general formula of the IDW methods is:

$$\widehat{w}_t(s_0) = \sum_{i=1}^{k} \lambda_i w_t(s_i) \tag{5}$$

where:

$$\lambda_i = {d_i^{-p}} \Big/ {\sum_{j=1}^{k} d_j^{-p}} \tag{6}$$

where: $w_t(s_{i=1:\,k})$ is the observed WS value at time t and the nearest location $s_i$, located at a distance $d_i$ from the target location $s_0$. The parameters p and $k$ are the exponents and the number of nearest neighbours to consider. It should be noted that the IDW was used in the study to interpolate observed WSTS and WSNEP (during the implementation of the WDC method). In both cases, the optimal number of nearest locations and the exponent were selected based on 1) the time series (TS) evaluation using the Pearson correlation coefficient between observed and estimated WSTS and 2) the probability distribution (PD) evaluation by calculating the coefficient of determination ($R^2$) between observed and estimated WSQ derived from the WSTS. The $R^2$ is presented in equation S4 of the supporting material. The results of the models were presented for each evaluation metric (TS and PD) separately.

The second spatial interpolation method implemented in the study was the Random Forest for Spatial Interpolation model [RFSI: 44]. The model uses nearby observations and their distance from the target location as covariates with a random forest regression model to interpolate at unsampled locations. The general formula of the model is [44]:

$$\widehat{w}_t(s_0) = f(x_1(s_0), \ldots, x_m(s_0), w_t(s_1), d_1, \ldots, w_t(s_k), d_k) \quad (7)$$

where: $x_{i=1:m}(s_0)$ are covariates available at the target location $s_0$, $f(.)$ is a regression function linking the covariates and the WS values at the unsampled location. A comparative analysis carried out by Sekulić, et al. [44] revealed that in real-world conditions, the RFSI model outperformed Space-time regression kriging, and the approach can scale and perform better than another spatial interpolation method based on the random forest model [45]. Furthermore, as RFSI does not require semi-variogram modelling, it is easier to implement than kriging methods with less restrictive assumptions (e.g., stationarity and linearity). In the original RFSI model, the authors used the random forest model to learn the regression function. Due to its efficiency and scalability for large datasets, the LGBM implementation of the gradient boosting algorithm was used in place of the random forest model, and the approach was renamed LGBMSI for this study. The tuned LGBMSI parameters were the same parameters presented in Table 1 of the present paper. These parameters were also tuned using a random search with 1000 iterations. As done for the QR model, the available covariates were ranked using the MRMR algorithm. The number of covariates to include in the model was treated as a parameter to be tuned during random search. Two versions of LGBMSI were tested: The version presented in equation 7 (it will be referred to as simply LGBMSI in the following sections) and a version which uses as additional covariate the WS values from the nearest ERA5 grid point to the unsampled location ($w_t(ERA5_{S_0})$). The LGBMSI model with the ERA5 covariates will be referred to as LGBMSI-ERA5 in the following sections of the paper and is presented in equation 8:

$$\widehat{w}_t(s_0) = f\left(x_1(s_0), \ldots, x_m(s_0), w_t(s_1), d_1, \ldots, w_t(s_k), d_k, w_t(ERA5_{S_0})\right) \quad (8)$$

### 3.4. Global Wind Atlas mean bias correction

The GWA version 3 (https://globalwindatlas.info/ ) feeds the output from a mesoscale atmospheric model into a microscale model to downscale the ERA5 wind data. The resulting wind data has a spatial resolution of 250m and accounts for the effect of the local topography and surface roughness. Several studies used the GWA to bias-correct reanalysis WS data [46-48]. The procedure involves applying a scaling factor to the reanalysis WS data to ensure that their mean matches the mean WS from GWA. The scaling factor is computed as the ratio between the mean WS from GWA and the reanalysis during the overlapping period of both datasets. The mean WS from GWA and ERA5 at 10 m estimated for the period between 2008 and 2017 were used to calculate the scaling factor. Nearest neighbour interpolation was used to interpolate the GWA data at locations of interest. The bias-corrected ERA5 using GWA will be referred to as GWA-ERA5 in the remainder of the paper.

### 3.5. Validation

The model validation strategy adopted in this study is aligned with the modelling procedure's primary task, which consisted of predicting WSTS at unsampled locations. During the models tuning, random k-fold cross-validation across the training locations was implemented to estimate the model's performance for prediction at (pseudo) unsampled locations. In 5-fold cross-validation, the training locations are randomly split into five groups. Training is carried out with the data of 4 groups, and the model is evaluated on the remaining group. This procedure was repeated five times, using each group once as the validation set. The final evaluation of the selected model was performed on a group of locations (test samples) held back and comprising approximately 30% (97 locations) of the available locations (303) for the entire study.

The estimated WSTS at locations of the test samples were evaluated according to the following criteria:

1. Time series evaluation: The Pearson correlation (PC), mean absolute error (MAE) and root-mean-squared error (RMSE) were calculated between observed and estimated WSTS. The PC, MAE and RMSE are presented in equations S1, S2, and S3 of the supporting material, respectively.
2. Probability distribution evaluation: Two approaches were used to evaluate the probability distribution of the estimate WSTS. First, quantiles with non-exceedance probabilities between 10% and 90% and a spacing of 10% were calculated from the WSTS using equation S8 in the supporting material. The $R^2$, MAE and RMSE were used to compare the observed and estimated WSQ. Lastly, the Overlap percentage [OP: 49] was used to assess the overlap between estimated and observed empirical probability distribution function (PDF). The OP is presented in equation S6 of the supporting material. For a review of criteria used for the selection of PD for WS data the reader is referred to [50].
3. Interannual variability (IAV) evaluation: The robust coefficient of variation [RCov: 51] of annual median WS was calculated to assess IAV. RCov serves as a robust and resistant measure of variability analogue to the coefficient of variation, which lacks robustness and resistance to outliers. The MAE and mean error (ME) between observed and estimated RCov were used to evaluate the performance of the models in reproducing the observed IAV. The RCov and the ME are presented in equations S7 and S5 of the supporting material, respectively.

# 4. Results

## 4.1. Quantile regression models

A thousand random combinations of the LGBM hyperparameters (Table 1) were tested with LGBMQR and LGBMQR-ERA5 models. Table S2 in the supporting material shows the best parameters found using a random search, including the number of selected covariates. Figure 2 illustrates the $R^2$, MAE, and RMSE between estimated and observed WSQ from the test samples. For reference, the same metrics between ERA5-WSQ and observed WSQ are also presented in Figure 2. Figure 3 shows boxplots of the metrics calculated over the different percentile points (P1 – P13) at each test sample. The Wilcoxon signed-rank test was used to test the statistical significance of these metrics between pairs of models (the test P-values are shown in Table S3 of the supporting material). The P-values associated with LGBMQR-ERA5 are all less than 0.05, and the P-values between LGBMQR and ERA5-WSQ are more significant than 0.05. LGBMQR and ERA5-WSQ had significantly lower median $R^2$ and higher median MAE and RMSE than LGBMQR-ERA5. LGBMQR underperformed compared to ERA5-WSQ, but the difference between the methods was not significant according to the Wilcoxon signed-rank test. LGBMQR outperformed ERA5-WSQ for WSQ with low exceedance probabilities (P1, P2, P3) while ERA5-WSQ were more accurate in the middle and upper tail of the distributions. It is evident from these results that the inclusion of the ERA5-WSQ improves the QR model performance; thus, WSQs from LGBMQR-ERA5 were used in subsequent analyses of the study.

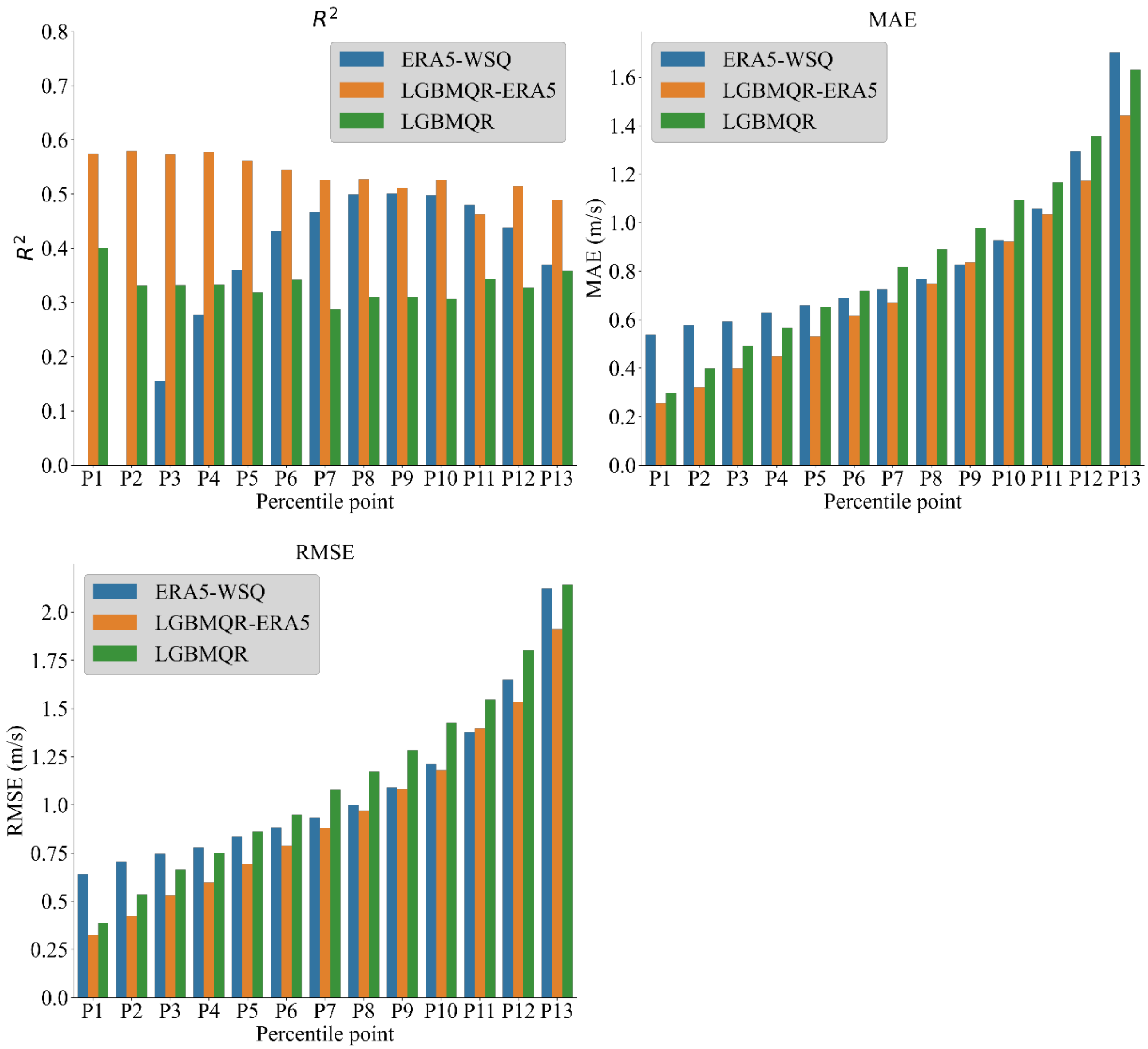



Figure 2: Results of the $R^2$, RMSE and MAE between estimated and predicted WSQ at various percentile points (P1-P13). The metrics were calculated across the test samples for each

percentile point.

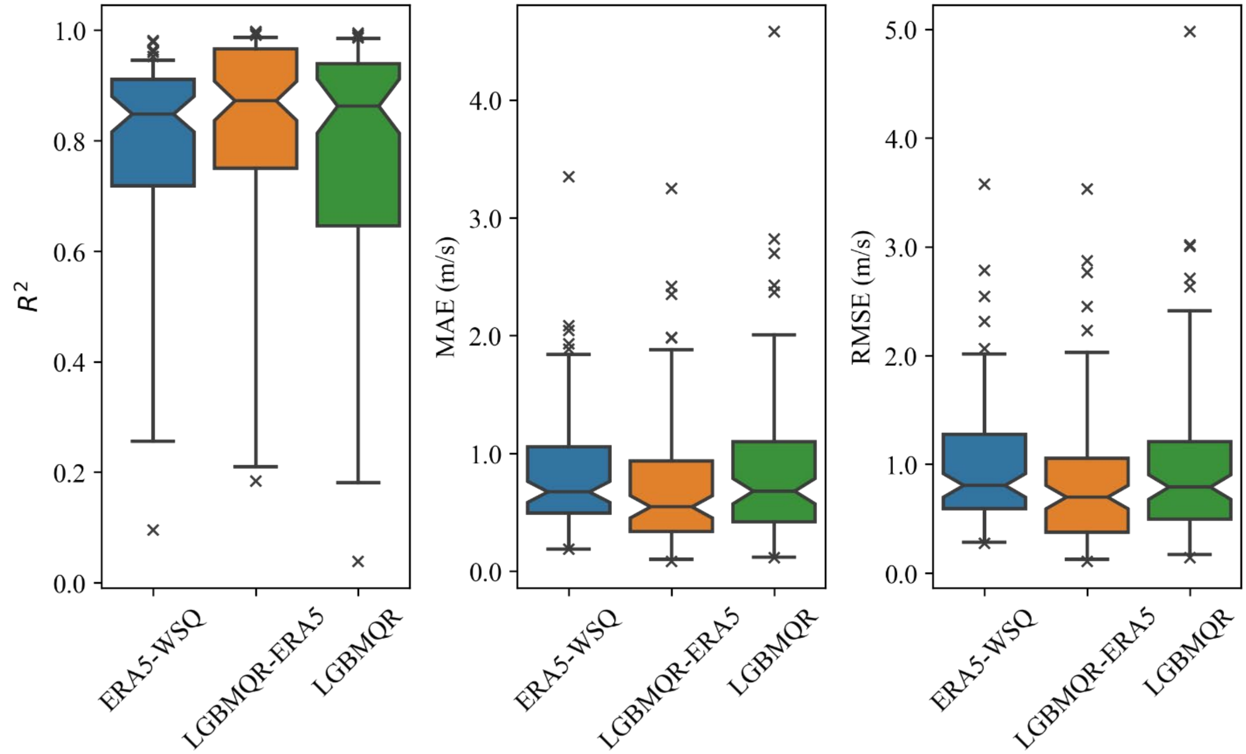



Figure 3: Result of the $R^2$, MAE and RMSE between observed and estimated WSQ. The metrics were calculated across the percentile points (P1-P13) at each location in the test samples.

### 4.2. Inverse distance weighting parameters

Table 2 shows the optimal parameters (p and k) for IDW based on the TS and PD evaluation. The selection of the best parameters was performed with the training set. The optimal k and p was contingent upon the evaluation criteria. For the interpolation of WSNEP, the optimal number of nearest neighbours (optimal k = 1) based on the PD evaluation is equated to the nearest neighbour interpolation. Generally, it was observed smaller values of k were optimal for the PD criteria. The results of the different evaluation criteria will be presented and discussed separately in the following section.

Table 2: Optimal parameters of the IDW for WSTS and WSNEP interpolation

| Interpolated variable | Evaluation criteria | Optimal k | Optimal p | Abbreviation used for the model herein |
|---|---|---|---|---|
| **WSTS** | PD | 6 | 0.3 | IDW-PD |
| | TS | 11 | 1.7 | IDW-TS |
| **WSNEP** | PD | 1 | - | WDC-PD |
| | TS | 9 | 1 | WDC-TS |

### 4.3. Time series evaluation

Figure 4 shows a boxplot of the PC, MAE and RMSE between observed and estimated WSTS from the test samples, while the median values of the metrics are given in Table 3. LGBMSI-ERA5 had the highest median PC alongside the lowest median MAE and RMSE. In contrast, WDC-PD had the lowest median PC and the highest median MAE and RMSE. WDC-TS performed better than WDC-PD, with performances comparable to the IDW model. The ERA5 WSTS showed a relatively high median PC and methods directly exploiting this dataset (GWA-ERA5, LGMBSI-ERA5, QM-ERA5) maintained a higher median PC with less variability in the distribution of the metric in comparison to methods solely using observations from nearby locations (WDC-PD, WDC-TS, LGBMSI, IDW-PD and IDW-TS). Despite QM-ERA5 showing a relatively high median PC, it also had a high median MAE and RMSE. Table S4 in the supporting material gives the P-value of the Wilcoxon signed-rank test between pairs of models for the different evaluation metrics. From the results of the Wilcoxon test, it was found that LGBMSI-ERA5 was the only method with an MAE and RMSE statistically inferior to IDW-TS. For the time series evaluation criteria, LGBMSI-ERA5 was the best-performing method, WDC-PD was the least effective method and most other models had performances comparable (in a statistical sense) to IDW-TS.

Table 3: Median PC, MAE and RMSE between observed and estimated WSTS

| Model | PC | MAE (m/s) | RMSE (m/s) |
|---|---|---|---|
| **WDC-TS** | 0.73 | 1.26 | 1.59 |
| **WDC-PD** | 0.64 | 1.62 | 2.19 |
| **ERA5** | 0.75 | 1.22 | 1.60 |
| **QM-ERA5** | 0.76 | 1.34 | 1.80 |
| **GWA-ERA5** | 0.75 | 1.32 | 1.68 |
| **IDW-TS** | 0.74 | 1.29 | 1.68 |
| **IDW-PD** | 0.72 | 1.31 | 1.66 |
| **LGBMSI** | 0.72 | 1.28 | 1.59 |
| **LGBMSI-ERA5** | 0.78 | 1.13 | 1.47 |

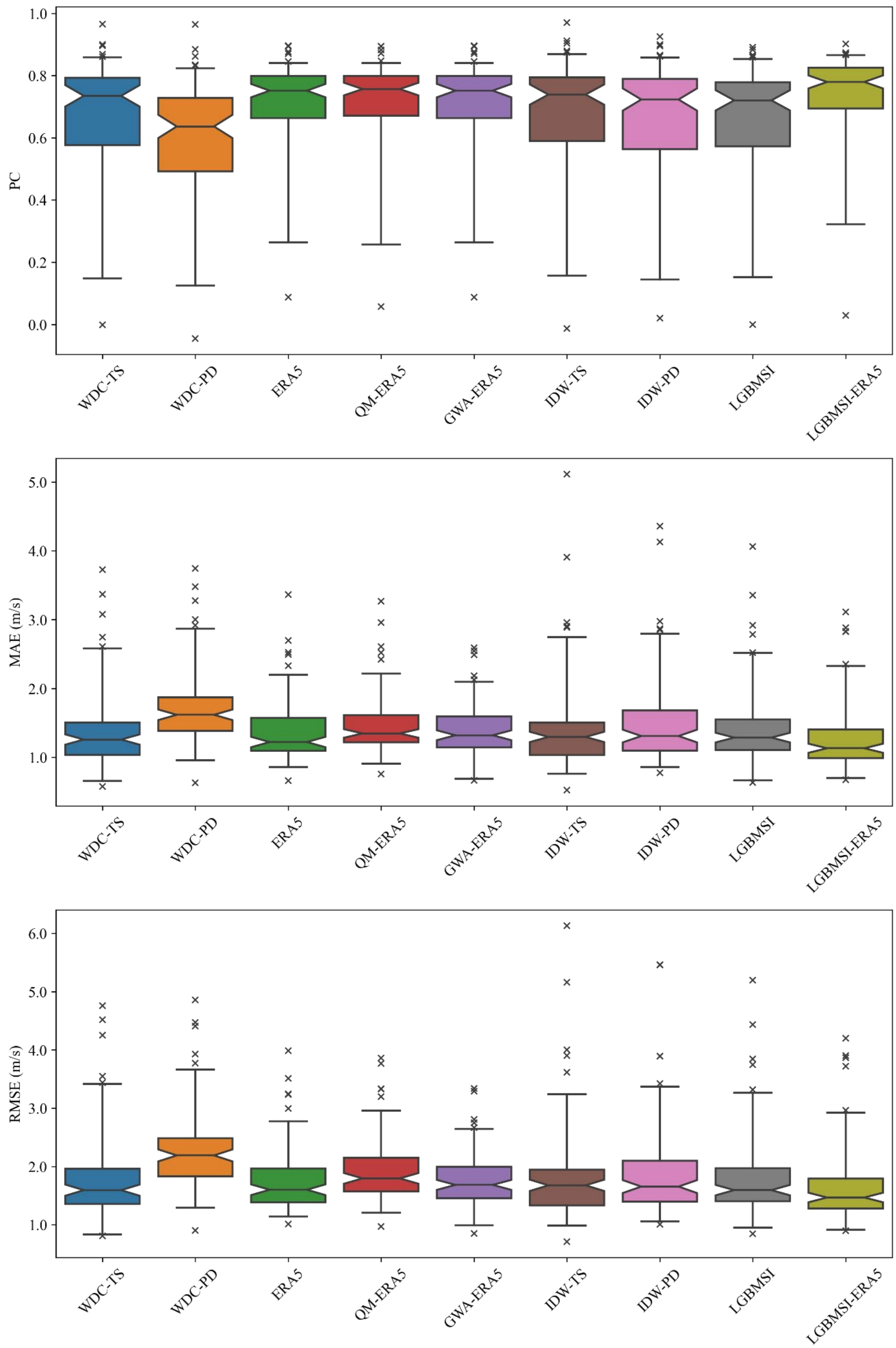

PC
MAE (m/s)
RMSE (m/s)
WDC-TS
WDC-PD
ERA5
QM-ERA5
GWA-ERA5
IDW-TS
IDW-PD
LGBMSI
LGBMSI-ERA5

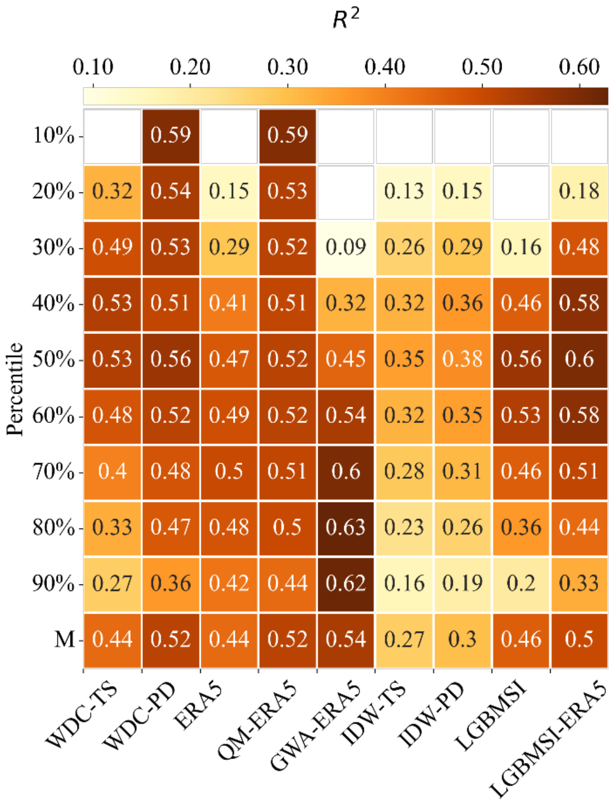
Figure 4: Result of the PC, MAE and RMSE between observed and estimated WSTS.

### 4.4. Probability distribution evaluation

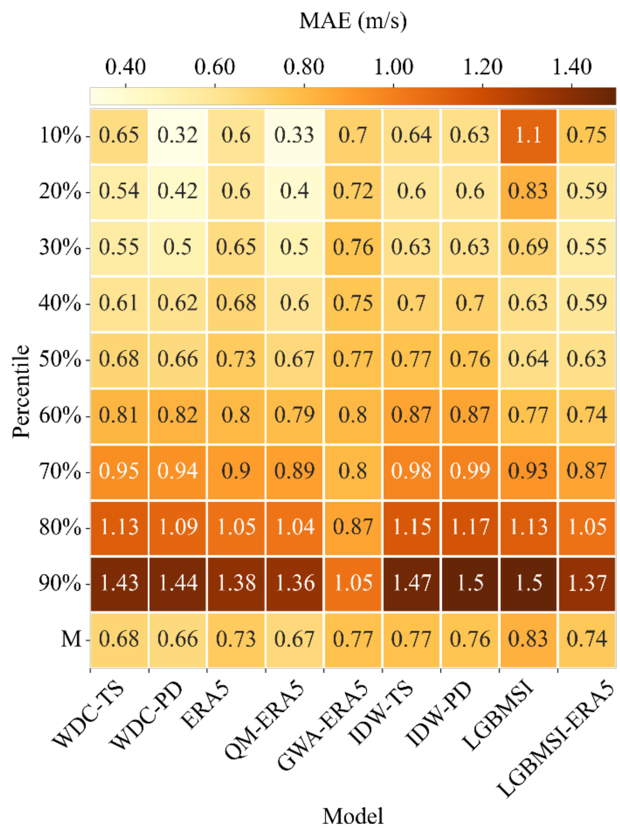
Figure 5 shows matrices detailing the $R^2$, MAE and RMSE calculated between the estimated and observed WSQ across various percentile points. The last row of these matrices (labelled M) presents the median value (calculated over the different percentile points). QM-ERA5 and WDC-PD were the top-performing methods overall, mainly due to their relatively strong performance in estimating WSQ in the lower and middle tail of the distribution. Both LGBMSI-ERA5 and LGBMSI performed relatively well in the middle of the distribution but were less effective in estimating WSQ in the lower tail. GWA-ERA5 was the best method for estimating WSQ in the upper tail of the distribution, yet it performed poorly for low exceedance probabilities WSQ. The IDW methods demonstrated an overall lack of effectiveness in estimating WSQ across the distribution.

The OP metric measured the overlap between the empirical PDF computed from the estimated and observed WSTS. Figure 6 presents boxplots illustrating the distribution of the OP metric. QM-ERA5 had the highest median OP at 80%, followed by ERA5 at 77%, GWA-ERA5 at 77% and WDC-PD at 76%. Also, QM-ERA5 and WDC-PD displayed less spread in the distribution of the metric compared to ERA5 and GWA-ERA5. LGBMSI and LGBMSI-ERA5 had the lowest median OP values at 65% and 72% respectively. The statistical significance of the results was tested with the Wilcoxon signed-ranked test between pairs of models (Table S5 of the supporting material). The P-values associated with QM-ERA5, LGBMSI, LGBMSI-ERA5 and WDC-TS were always small (ex.: less than 0.05). The differences between IDW, ERA5, GWA-ERA5 and WDC-PD were not statistically significant (P-values greater than 0.05) for the OP metric. Overall, QM-ERA5 was the top performer for the OP metric, followed by (listed in no particular order) IDW, ERA5, GWA-

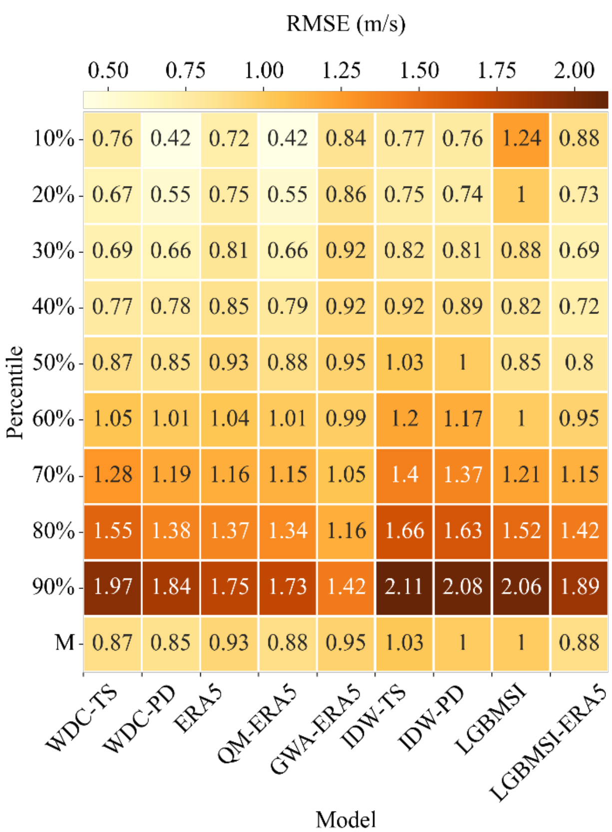
ERA5 and WDC-PD. WDC-TS performed slightly better than LGBMSI-ERA5, while LGBMSI was the least effective method.



$R^2$
Percentile
Model

MAE (m/s)
Percentile
Model

RMSE (m/s)
Percentile
Model


387

Figure 5: Result of the $R^{2,}$ MAE and RMSE between observed and estimated WSQ. The last row of the matrices gives the median of the metric calculated across the different percentile points. Values of $R^2$ less than 0 were omitted from the matrices.

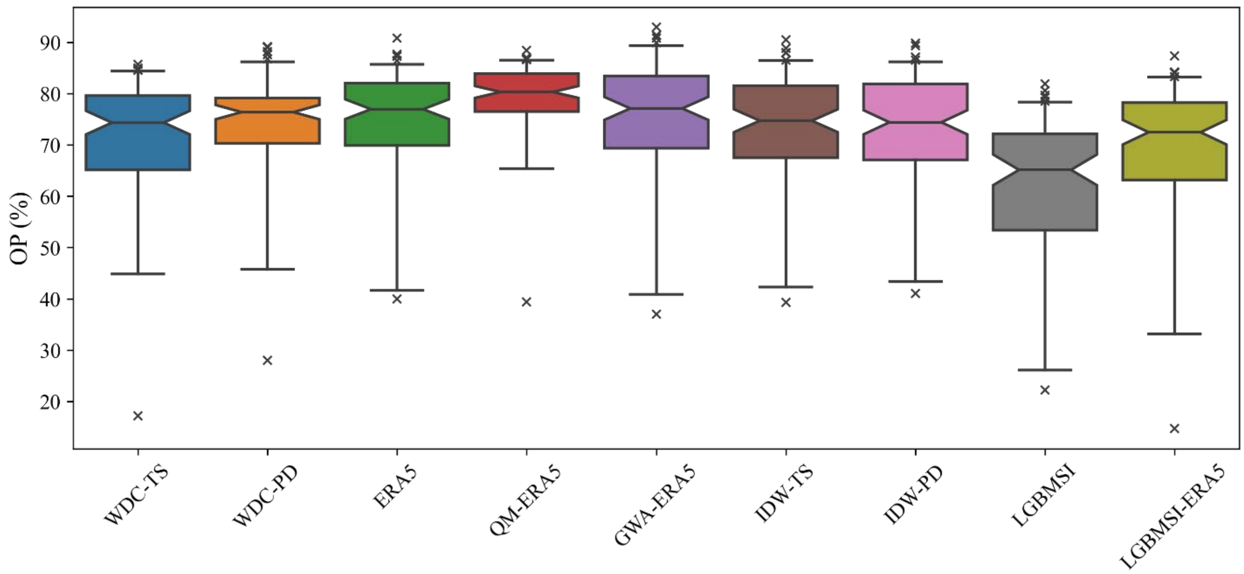



Figure 6: Boxplots of OP metrics calculated between observed and estimated empirical PDF.

### 4.5. Interannual variability evaluation

The IAV assesses the fluctuation of wind speed across multiple years. Studies have indicated that wind speed exhibits IAV in many parts of the world [52-54]. The IAV has been linked to atmospheric teleconnections [54-56] such as the El Niño-Southern Oscillation and the North Atlantic Oscillation. Accurately assessing the IAV of wind resources is essential for providing adequate information for the long term planning of wind energy projects [57]. Some attempts have been made to develop teleconnection-based long term forecasting models for wind speed that use low frequency atmospheric circulation patterns as covariates [58].

Figure 7 presents a bar plot representing the MAE and ME between observed and estimated RCov of median annual WS.

WDC-PD gave the smaller MAE at 2%, while the other methods gave a slightly higher MAE at 3%. Notably, WDC-PD was the only method that overestimated, on average, the IAV (positive ME). The other methods showed, on average, an underestimation of the IAV (negative ME). There was no substantial difference in the performance among the various methods based on the IAV.

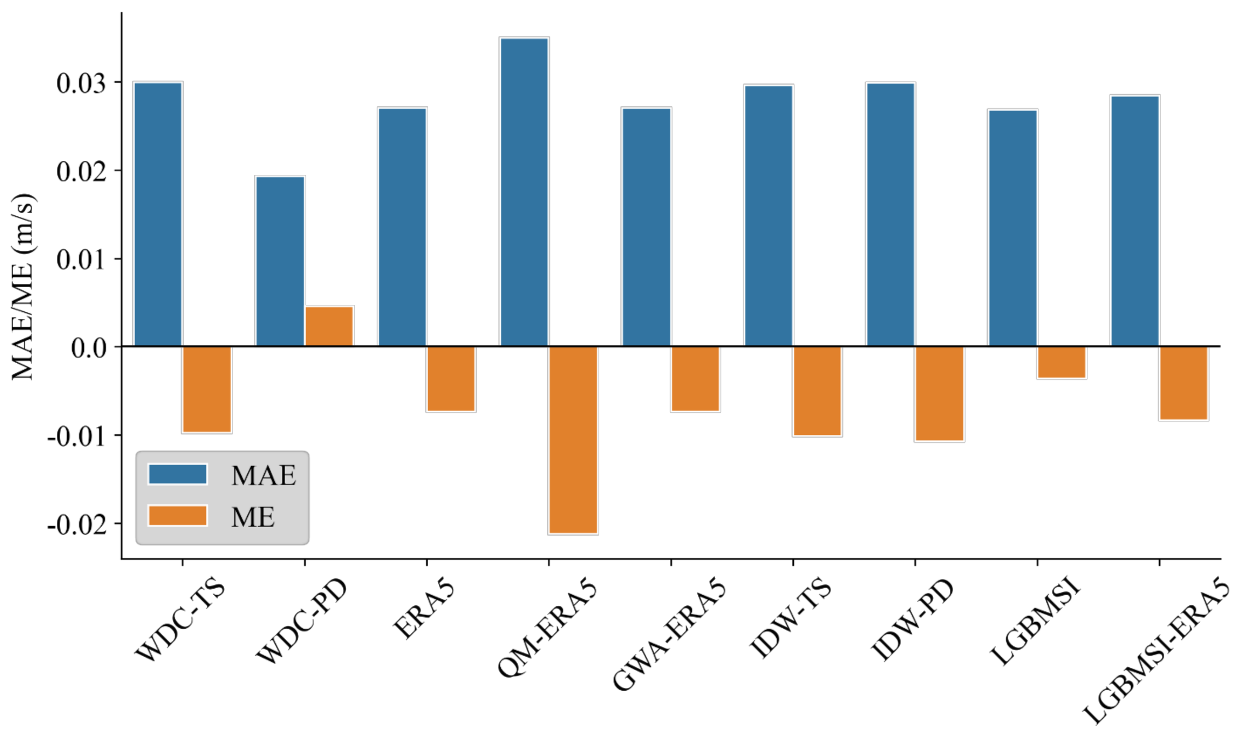



Figure 7: Result of MAE and ME between observed and estimated RCov of median annual WS.

## 5. Discussion

The study results indicate that no single method excelled according to all evaluation criteria, suggesting potential for improvement through combining specific methods. For instance, it was found that WSQ derived from the GWA-ERA5 time series was the most accurate in the upper tail of the distribution. Conversely, in the lower tail, WSQs from GWA-ERA5 were inaccurate compared to QM-ERA5 and WDC-PD. Based on these outcomes, future studies are recommended to explore using the mean WS from the GWA dataset as covariates of the QR model to potentially improve the estimation of the conditional WSQ in the upper tail of the distribution, thus enhancing the performance of QM-ERA5 and WDC-PD.

LGBMSI-ERA5 was the top performer based on the time series evaluation. In the case of the evaluation based on the PD, QM-ERA5 was the top performer. Generally, more complex methods yielded superior performances compared to the baseline model (IDW), suggesting some benefits in implementing complex methods in part due to their ability to integrate various WS covariates. The ERA5 dataset was a valuable covariate. For instance, ERA5 WSTSs are well correlated with ground measurements, and this correlation could be improved significantly (in a statistical sense) by using the dataset as a covariate with LGBMSI. Also, ERA-WSQ significantly improved (in a statistical sense) the performance of the QR model. It should be noted that other covariates used as input of the QR models demonstrated a higher ability to predict WSQ in the distribution's lower tail than ERA5-WSQ, which seemed less accurate in the lower tail.

QM-ERA5 improved the performance of ERA5 in most cases. The approach is relatively easy to implement and relies on a reasonable estimation of the WSD at unsampled locations. One reason that could explain the improved performance of QM-ERA5 is its higher accuracy in the lower tail of the distribution compared to ERA5 wind data. It was also revealed that the WDC method was competitive. However, the approach is sensitive to the evaluation criteria used to select the optimal parameters of the IDW for interpolating the WSNEP. Different evaluation criteria lead to different optimal parameters, which leads, in turn, to different performances during evaluation. For instance, WDC-PD performed relatively well based on the evaluation of PD, while it performed poorly based on the TS evaluation. In contrast, WDC-TS performed relatively well based on the TS evaluation and was less effective than WDC-PD based on the evaluation of the PD. In future studies, it is recommended that different methods to interpolate the WSNEP are explored to improve the performance of the WDC method. For instance, a more complex interpolation method, such as RFSI, could be applied to interpolate the WSNEP.

In this study, LGBM with the pinball lost function was used as the QR model (LGBMQR). Other quantile regression models could be viable alternatives, such as quantile regression forests [59] and quantile regression neural networks [60]. LGBMQR was adopted because it is efficient during training, and in general, gradient-boosting models have demonstrated superior performance on tabular data [61]. In upcoming research, a comparative analysis can be performed to evaluate the performance of different QR models for conditional WSQ mapping.

For practical reasons, the analysis in the present study was carried out at the World Meteorological Organization (WMO) recommended wind speed measurement height of 10 m. Modern wind turbines operate at hub heights of 100 m and beyond. It would be ideal to assess the wind resource directly at these hub heights. However, there is lack of extensive wind speed time series data at these heights and even when available, accessing such data from private wind farm operators can pose challenges. To account for this disparity, vertical wind profile equations such as the logarithmic and power law are employed to extrapolate the estimated wind speed from 10 m to the hub height [15, 21]. This procedure inevitably introduces additional uncertainty to the estimated wind resource. Future research should be conducted to evaluate and quantify this layer of uncertainty more comprehensively.

## 6. Conclusions and future research

This study conducted a comprehensive evaluation of various approaches for the prediction of wind speed time series at unsampled locations. It was found that no single method consistently outperformed the other methods according to all evaluation criteria. However, complex methods that include various covariates were more effective than the baseline method. Mainly, two approaches (QM-ERA5 and LGBMSI-ERA5) applied to bias-correct ERA5 wind speed data seemed promising and showed improved results compared to the most common ERA5 bias correction method (GWA-ERA5). It should be noted that both methods are more complex and computationally demanding than GWA-ERA5. However, LGBMSI-ERA5 significantly improved the accuracy of the ERA5 data when evaluating the time series correlations, while QM-ERA5 significantly improved the overlap percentage between the observed and estimated empirical PDF. In future studies, it is recommended that the performance of LGBMSI-ERA5 and QM-ERA5 be explored further in different regions with different wind regimes. Another promising research route is the potential to combine different approaches to produce a more accurate model across multiple evaluation criteria.

Also, with the QR model, there is a potential to account for the non-stationarity of the WSD by using related covariates. For instance, Ouarda and Charron [54] found that the North-Atlantic Oscillation and the Pacific North American indices of atmospheric circulation were good predictors of the IAV of WS in the province of Québec, Canada. In future studies, these climate indices can be used as covariates with a QR model in the region to map conditional WSQ that accounts for the resource's IAV. This analysis could lead to a better evaluation of the wind resources at unsampled locations, thus reducing the risk associated with future projects.

The comprehensive evaluation provided in the present study aims to assist practitioners in choosing the most suitable methodologies for their specific projects. Furthermore, it is anticipated that this research will inspire future studies to systematically evaluate various approaches for predicting wind speed time series at unsampled locations. This will foster in the long run a better understanding of the strengths and limitations of these approaches and encourage their refinement and the development of more robust techniques for the prediction of wind speed time series at unsampled locations.

**Acknowledgments**

The authors thank the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canada Research Chair Program for funding this research. The authors would like to extend their gratitude to the Editor, and two anonymous reviewers for their comments and suggestions, which significantly improved the quality of the paper.

**Data availability statement**

The data used in the study are available from public source. The measure wind speed data were acquired from Environment and climate change Canada (https://collaboration.cmc.ec.gc.ca/cmc/climate/Get_More_Data_Plus_de_donnees/ ), the digital elevation model from the Japan Aerospace Exploration Agency (https://www.eorc.jaxa.jp/ALOS/en/dataset/aw3d30/aw3d30_e.htm) , the ERA5 10 m wind components from ECMWF (https://doi.org/10.24381/cds.adbb2d47) , the land use map from

Natural Resources Canada (https://doi.org/10.3390/rs9111098), the GWA mean wind speed from the Technical University of Denmark (https://globalwindatlas.info/en).

## References

[1] A. Cherp, V. Vinichenko, J. Tosun, J.A. Gordon, J. Jewell. National growth dynamics of wind and solar power compared to the growth required for global climate targets. Nature Energy. 6 (2021) 742-54. 10.1038/s41560-021-00863-0

[2] R. Wiser, J. Rand, J. Seel, P. Beiter, E. Baker, E. Lantz, P. Gilman. Expert elicitation survey predicts 37% to 49% declines in wind energy costs by 2050. Nature Energy. 6 (2021) 555-65. 10.1038/s41560-021-00810-z

[3] G. Ren, J. Liu, J. Wan, Y. Guo, D. Yu. Overview of wind power intermittency: Impacts, measurements, and mitigation solutions. Applied Energy. 204 (2017) 47-65. https://doi.org/10.1016/j.apenergy.2017.06.098

[4] J.C.Y. Lee, M.J. Fields. An overview of wind-energy-production prediction bias, losses, and uncertainties. Wind Energ Sci. 6 (2021) 311-65. 10.5194/wes-6-311-2021

[5] A. Lopez, T. Mai, E. Lantz, D. Harrison-Atlas, T. Williams, G. Maclaurin. Land use and turbine technology influences on wind potential in the United States. Energy. 223 (2021) 120044. https://doi.org/10.1016/j.energy.2021.120044

[6] D. Niermann, M. Borsche, Andrea K. Kaiser-Weiss, F. Kaspar. Evaluating renewable-energy-relevant parameters of COSMO-REA6 by comparison with satellite data, station observations and other reanalyses. Meteorologische Zeitschrift. 28 (2019) 347-60. 10.1127/metz/2019/0945

[7] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, et al. The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society. 146 (2020) 1999-2049. https://doi.org/10.1002/qj.3803

[8] A. Molod, L. Takacs, M. Suarez, J. Bacmeister. Development of the GEOS-5 atmospheric general circulation model: evolution from MERRA to MERRA2. Geosci Model Dev. 8 (2015) 1339-56. 10.5194/gmd-8-1339-2015

[9] G. Gualtieri. Analysing the uncertainties of reanalysis data used for wind resource assessment: A critical review. Renewable and Sustainable Energy Reviews. 167 (2022) 112741. https://doi.org/10.1016/j.rser.2022.112741

[10] K. Gruber, P. Regner, S. Wehrle, M. Zeyringer, J. Schmidt. Towards global validation of wind power simulations: A multi-country assessment of wind power simulation from MERRA-2 and ERA-5 reanalyses bias-corrected with the global wind atlas. Energy. 238 (2022) 121520. https://doi.org/10.1016/j.energy.2021.121520

[11] R.K. Samal. Assessment of wind energy potential using reanalysis data: A comparison with mast measurements. Journal of Cleaner Production. 313 (2021) 127933. https://doi.org/10.1016/j.jclepro.2021.127933

[12] A. Ayik, N. Ijumba, C. Kabiri, P. Goffin. Preliminary wind resource assessment in South Sudan using reanalysis data and statistical methods. Renewable and Sustainable Energy Reviews. 138 (2021) 110621. https://doi.org/10.1016/j.rser.2020.110621

[13] J. Ramon, L. Lledó, V. Torralba, A. Soret, F.J. Doblas-Reyes. What global reanalysis best represents near-surface winds? Quarterly Journal of the Royal Meteorological Society. 145 (2019) 3236-51. https://doi.org/10.1002/qj.3616

[14] G. Gualtieri. Reliability of ERA5 Reanalysis Data for Wind Resource Assessment: A Comparison against Tall Towers. Energies2021.
[15] K. Gruber, C. Klöckl, P. Regner, J. Baumgartner, J. Schmidt. Assessing the Global Wind Atlas and local measurements for bias correction of wind power generation simulated from MERRA-2 in Brazil. Energy. 189 (2019) 116212. https://doi.org/10.1016/j.energy.2019.116212
[16] A.-J. Collados-Lara, L. Baena-Ruiz, D. Pulido-Velazquez, E. Pardo-Igúzquiza. Data-driven mapping of hourly wind speed and its potential energy resources: A sensitivity analysis. Renewable Energy. 199 (2022) 87-102. https://doi.org/10.1016/j.renene.2022.08.109
[17] M. Cellura, G. Cirrincione, A. Marvuglia, A. Miraoui. Wind speed spatial estimation for energy planning in Sicily: A neural kriging application. Renewable Energy. 33 (2008) 1251-66. https://doi.org/10.1016/j.renene.2007.08.013
[18] F. González-Longatt, H. Medina, J. Serrano González. Spatial interpolation and orographic correction to estimate wind energy resource in Venezuela. Renewable and Sustainable Energy Reviews. 48 (2015) 1-16. https://doi.org/10.1016/j.rser.2015.03.042
[19] F. Veronesi, S. Grassi, M. Raubal. Statistical learning approach for wind resource assessment. Renewable and Sustainable Energy Reviews. 56 (2016) 836-50. https://doi.org/10.1016/j.rser.2015.11.099
[20] C. Jung. High Spatial Resolution Simulation of Annual Wind Energy Yield Using Near-Surface Wind Speed Time Series. Energies. 9 (2016) 344. doi:10.3390/en9050344
[21] C. Jung, D. Schindler. Integration of small-scale surface properties in a new high resolution global wind speed model. Energy Conversion and Management. 210 (2020) 112733. https://doi.org/10.1016/j.enconman.2020.112733
[22] F. Houndekindo, T.B.M.J. Ouarda. A non-parametric approach for wind speed distribution mapping. Energy Conversion and Management. 296 (2023) 117672. https://doi.org/10.1016/j.enconman.2023.117672
[23] A. Castellarin, G. Botter, D.A. Hughes, S. Liu, T.B.M.J. Ouarda, J. Parajka, et al. Prediction of flow duration curves in ungauged basins. in: G. Blöschl, H. Savenije, M. Sivapalan, A. Viglione, T. Wagener, (Eds.), Runoff Prediction in Ungauged Basins: Synthesis across Processes, Places and Scales. Cambridge University Press, Cambridge, 2013. pp. 135-62.
[24] A.I. Requena, T.B.M.J. Ouarda, F. Chebana. Flood Frequency Analysis at Ungauged Sites Based on Regionally Estimated Streamflows. Journal of Hydrometeorology. 18 (2017) 2521-39. 10.1175/JHM-D-16-0143.1
[25] T.B.M.J. Ouarda, C. Charron, A. St-Hilaire. Regional estimation of river water temperature at ungauged locations. Journal of Hydrology X. (2022). 10.1016/j.hydroa.2022.100133
[26] A.J. Cannon, S.R. Sobie, T.Q. Murdock. Bias Correction of GCM Precipitation by Quantile Mapping: How Well Do Methods Preserve Changes in Quantiles and Extremes? Journal of Climate. 28 (2015) 6938-59. https://doi.org/10.1175/JCLI-D-14-00754.1
[27] M.A. Ben Alaya, F. Chebana, T.B.M.J. Ouarda. Multisite and multivariable statistical downscaling using a Gaussian copula quantile regression model. Climate Dynamics. 47 (2016) 1383-97. 10.1007/s00382-015-2908-3
[28] T. Tadono, H. Ishida, F. Oda, S. Naito, K. Minakawa, H. Iwamoto. Precise Global DEM Generation by ALOS PRISM. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences. II4 (2014) 71-6. 10.5194/isprsannals-II-4-71-2014
[29] R. Latifovic, D. Pouliot, I. Olthof. Circa 2010 Land Cover of Canada: Local Optimization Methodology and Product Development. Remote Sensing. 9 (2017) 1098.

[30] Y. He, H. Li. Probability density forecasting of wind power using quantile regression neural network and kernel density estimation. Energy Conversion and Management. 164 (2018) 374-84. https://doi.org/10.1016/j.enconman.2018.03.010

[31] D. Ouali, F. Chebana, T.B.M.J. Ouarda. Quantile Regression in Regional Frequency Analysis: A Better Exploitation of the Available Information. Journal of Hydrometeorology. 17 (2016). 10.1175/JHM-D-15-0187.1

[32] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Neural Information Processing Systems2017.

[33] J. Fan, X. Ma, L. Wu, F. Zhang, X. Yu, W. Zeng. Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. Agricultural Water Management. 225 (2019) 105758. https://doi.org/10.1016/j.agwat.2019.105758

[34] J. Park, J. Moon, S. Jung, E. Hwang. Multistep-Ahead Solar Radiation Forecasting Scheme Based on the Light Gradient Boosting Machine: A Case Study of Jeju Island. Remote Sensing. 12 (2020) 2271.

[35] E. Genov, C.D. Cauwer, G.V. Kriekinge, T. Coosemans, M. Messagie. Forecasting flexibility of charging of electric vehicles: Tree and cluster-based methods. Applied Energy. 353 (2024) 121969. https://doi.org/10.1016/j.apenergy.2023.121969

[36] M. Feurer, F. Hutter. Hyperparameter Optimization. in: F. Hutter, L. Kotthoff, J. Vanschoren, (Eds.), Automated Machine Learning: Methods, Systems, Challenges. Springer International Publishing, Cham, 2019. pp. 3-33.

[37] C. Ding, P. Hanchuan. Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol. 3 (2005) 185-205. 10.1142/s0219720005001004

[38] F. Houndekindo, T.B.M.J. Ouarda. Comparative study of feature selection methods for wind speed estimation at ungauged locations. Energy Conversion and Management. 291 (2023) 117324. https://doi.org/10.1016/j.enconman.2023.117324

[39] H.A. Mombeni, B. Mansouri, M. Akhoond. Asymmetric kernels for boundary modification in distribution function estimation. REVSTAT-Statistical Journal. 19 (2021) 463–84-–84.

[40] P. Geurts, D. Ernst, L. Wehenkel. Extremely randomized trees. Machine Learning. 63 (2006) 3-42. 10.1007/s10994-006-6226-1

[41] A.W. Wood, E.P. Maurer, A. Kumar, D.P. Lettenmaier. Long-range experimental hydrologic forecasting for the eastern United States. Journal of Geophysical Research: Atmospheres. 107 (2002) ACL 6-1-ACL 6-15. https://doi.org/10.1029/2001JD000659

[42] C. Shu, T.B.M.J. Ouarda. Improved methods for daily streamflow estimates at ungauged sites. Water Resources Research. 48 (2012). https://doi.org/10.1029/2011WR011501

[43] C. Jung, D. Schindler. Introducing a new wind speed complementarity model. Energy. 265 (2023) 126284. https://doi.org/10.1016/j.energy.2022.126284

[44] A. Sekulić, M. Kilibarda, G.B.M. Heuvelink, M. Nikolić, B. Bajat. Random Forest Spatial Interpolation. Remote Sensing2020.

[45] T. Hengl, M. Nussbaum, M.N. Wright, G.B.M. Heuvelink, B. Gräler. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ. 6 (2018) e5518. 10.7717/peerj.5518

[46] J.P. Murcia, M.J. Koivisto, G. Luzia, B.T. Olsen, A.N. Hahmann, P.E. Sørensen, M. Als. Validation of European-scale simulated wind speed and wind generation time series. Applied Energy. 305 (2022) 117794. https://doi.org/10.1016/j.apenergy.2021.117794

[47] S.C. de Aquino Ferreira, F.L. Cyrino Oliveira, P.M. Maçaira. Validation of the representativeness of wind speed time series obtained from reanalysis data for Brazilian territory. Energy. 258 (2022) 124746. https://doi.org/10.1016/j.energy.2022.124746
[48] G. Luzia, M.J. Koivisto, A.N. Hahmann. Validating EURO-CORDEX climate simulations for modelling European wind power generation. Renewable Energy. 217 (2023) 118989. https://doi.org/10.1016/j.renene.2023.118989
[49] S.E. Perkins, A.J. Pitman, N.J. Holbrook, J. McAneney. Evaluation of the AR4 Climate Models' Simulated Daily Maximum Temperature, Minimum Temperature, and Precipitation over Australia Using Probability Density Functions. Journal of Climate. 20 (2007) 4356-76. https://doi.org/10.1175/JCLI4253.1
[50] T.B.M.J. Ouarda, C. Charron, F. Chebana. Review of criteria for the selection of probability distributions for wind speed data and introduction of the moment and L-moment ratio diagram methods, with a case study. Energy Conversion and Management. 124 (2016) 247-65. https://doi.org/10.1016/j.enconman.2016.07.012
[51] S. Watson. Quantifying the Variability of Wind Energy. Advances in Energy Systems2019. pp. 355-68.
[52] C. Jung, D. Taubert, D. Schindler. The temporal variability of global wind energy – Long-term trends and inter-annual variability. Energy Conversion and Management. 188 (2019) 462-72. https://doi.org/10.1016/j.enconman.2019.03.072
[53] P.E. Bett, H.E. Thornton, R.T. Clark. Using the Twentieth Century Reanalysis to assess climate variability for the European wind industry. Theoretical and Applied Climatology. 127 (2017) 61-80. 10.1007/s00704-015-1591-y
[54] T.B.M.J. Ouarda, C. Charron. Non-stationary statistical modelling of wind speed: A case study in eastern Canada. Energy Conversion and Management. 236 (2021) 114028. https://doi.org/10.1016/j.enconman.2021.114028
[55] F. Zhou, Z. Zhao, C. Azorin-Molina, X. Jia, G. Zhang, D. Chen, et al. Teleconnections between large-scale oceanic-atmospheric patterns and interannual surface wind speed variability across China: Regional and seasonal patterns. Science of The Total Environment. 838 (2022) 156023. https://doi.org/10.1016/j.scitotenv.2022.156023
[56] M.S. Naizghi, T.B.M.J. Ouarda. Teleconnections and analysis of long-term wind speed variability in the UAE. International Journal of Climatology. 37 (2017) 230-48. https://doi.org/10.1002/joc.4700
[57] S.C. Pryor, T.J. Shepherd, R.J. Barthelmie. Interannual variability of wind climates and wind turbine annual energy production. Wind Energ Sci. 3 (2018) 651-65. 10.5194/wes-3-651-2018
[58] H. Woldesellasse, P.R. Marpu, T.B.M.J. Ouarda. Long-term forecasting of wind speed in the UAE using nonlinear canonical correlation analysis (NLCCA). Arabian Journal of Geosciences. 13 (2020) 962. 10.1007/s12517-020-05981-9
[59] N. Meinshausen, G. Ridgeway. Quantile regression forests. Journal of machine learning research. 7 (2006).
[60] A.J. Cannon. Quantile regression neural networks: Implementation in R and application to precipitation downscaling. Computers & Geosciences. 37 (2011) 1277-84. https://doi.org/10.1016/j.cageo.2010.07.005
[61] L. Grinsztajn, E. Oyallon, G. Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? Advances in Neural Information Processing Systems. 35 (2022) 507-20.