A non-parametric approach for wind speed distribution mapping

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, 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

Table of Contents

List of abbreviations

BS	Birnbaum-Saunders
CANGRD	Canadian gridded temperature and precipitation anomalies dataset
CDF	Cumulative probability function
D	Kolmogorov–Smirnov statistic
DEM	digital elevation model
ECDF	Empirical cumulative probability function
FS	Feature selection
GBT	Gradient boosting trees
GG	Generalized Gamma
KCDF	Kernel estimator of cumulative distribution function
LN	Log-Normal
LR	Linear regression
LSE	Least Square Estimation
MI	Mutual information
MRMR	Minimum redundancy maximum relevance
MRMR-MI	Minimum redundancy maximum relevance with Mutual information
MRMR-PC	Minimum redundancy maximum relevance with Pearson correlation coefficient
PC	Pearson correlation coefficient
PDF	Probability distribution function
PP plot	Percentage probability plot
QWSM	Quantile-based wind speed probability distribution mapping
R	Rayleigh
RD	Regional distribution
SSE	Sum of the square error
W	Weibull
WPM	Weibull parameters mapping
WS	Wind speed
WSQ	Wind speed quantile
XGB	Extreme Gradient Boosting

Highlights

- A non-parametric approach for wind speed mapping is developed.
- A comparative analysis of parametric and non-parametric approaches is carried out.
- The non-parametric method slightly outperforms the parametric approach and avoids the hypothesis of a single distribution.
- The new method is recommended for regions having diverse wind regimes.

Abstract

Statistical methods to estimate wind resources at unsampled locations in a region can serve as an initial step to identify locations that warrant further investigation. There has been an ongoing effort to develop approaches for mapping the parameters of the wind speed distribution with statistical methods. This approach enables a comprehensive understanding of the wind resource variability across the entire region by considering the full wind speed distribution rather than focusing solely on mean values. The present study proposes a non-parametric approach to map the wind speed distribution. The method's main advantage is that it avoids constraining the region to a single distribution family and is thus more flexible than existing methods. In the proposed approach, a number of wind speed quantiles are first mapped in the region using machine learning techniques. Afterwards, the wind speed distribution is estimated by fitting an asymmetric kernel estimator to the estimated wind speed quantiles at unsampled locations. The new approach was compared to the standard statistical method based on mapping the regional wind speed distribution parameters. The results indicate that the non-parametric approach leads in the best scenario to a 9% and 6% drop in the Kolmogorov-Smirnov statistic on average during cross-validation and validation, respectively. The Birnbaum-Saunders and the Log-Normal kernels gave a better fit to the estimated wind speed quantiles than the Weibull kernel. The proposed approach is recommended in regions with high wind regime variability.

Keywords: Asymmetric kernel estimator, Non-parametric, Quantile, Wind speed distribution, wind variability, ungauged location, regional estimation.

1. Introduction

Wind energy has the potential to become a crucial source of power worldwide [1]. In 2021, worldwide wind energy installed capacity reached 837 GW, with an estimated offset of over 1.2 billion tons of CO2 [2]. However, more effort is needed to raise the contribution of wind energy in the world energy mix to achieve a more sustainable and low-carbon future [3].

One of the initial stages of building a wind farm involves finding a suitable location with sufficient wind resources to generate electricity. This objective typically involves conducting an in-depth assessment of the wind regime, which requires a long-term dataset of wind speed measurements. However, this data is often only available at irregular points in space rather than at the location of interest for wind energy production. It may not be feasible to install a monitoring station to gather sufficient data during the preliminary site selection due to time and financial constraints. Using methods that can estimate wind resources at unsampled locations is more suitable. Although these methods may not be as accurate as a monitoring station, they can help identify potential sites that warrant further investigation.

Numerous WS estimation studies have been conducted at unsampled locations, as detailed in the review by Houndekindo and Ouarda [4]. These studies typically estimate an aggregated WS value [5, 6], such as the mean and occasionally the WS distribution, via mapping the parameters of a theoretical probability distribution function. Both approaches have some downsides. First, using the mean WS for wind resource assessment may underestimate the long-term resource depending on the frequency distribution's shape [7]. Second, when estimating the WS distribution at unsampled locations, authors typically select a unique family of distributions with different parameters for the entire region (the regional distribution (RD)). For example, Veronesi, et al. [8] mapped WS distribution in the UK using random forests and assumed that the Weibull distribution (W) was adequate across the study region. Although the W is the most commonly used distribution for WS modelling, some studies have found that other types of distributions may provide a better fit depending on the wind regime at a location. For instance, the three-parameter W distribution (an additional location parameter) is better suited for calm wind regimes [9]. Tsvetkova and Ouarda [10] reported that the heavy-tailed Halphen distribution family provided a better fit than the two-parameter W distribution in all 125 WS stations considered in Eastern Canada.

In another study, Jung [11] mapped WS distribution parameters in Southwest Germany. First, the author evaluated the goodness of fit (GOF) of 67 theoretical distributions to select the RD. Then, a gradient-boosting model was employed to map the parameters of the selected distribution. Similarly, Laib and Kanevski [12] conducted a study in Switzerland for extreme WS. The authors used the quantiles plot to evaluate the GOF of three theoretical distributions and select a RD. Then, with a machine learning model, they mapped the parameters of the RD. This approach can be tedious, requiring the testing of multiple distributions, and there is no guarantee that the selected distribution would be adequate at the unsampled locations of interest. Previous studies evaluated the goodness of fit of different theoretical distributions for WS modelling in a given region [13-18] and found that no single distribution family provided the best fit at all locations in the region. Thus, using a single family of distributions may not be appropriate for characterizing the WS distribution in an entire region.

This work proposes a new approach for WS distribution mapping that does not constrain the region to a single distribution family (i.e.: a regional distribution). The proposed approach consists of estimating several WS quantiles (WSQ) at a location of interest. Then, a distribution function can be fitted to the estimated WS quantiles using the Least Square Estimation (LSE) method.

It can be tedious to test several distributions with the LSE method. Indeed, in most cases, the LSE method does not have an analytical solution. Thus, optimization algorithms may be required with an initial guess of the parameters, which can lead to suboptimal solutions. To address this issue, it is proposed to fit a kernel estimator of cumulative distribution function (KCDF) to the estimated WSQ.

Kernel estimators are, in general, rather flexible and do not require prior knowledge of the family of distributions of the data. The literature shows a growing interest in kernel estimators for WS distribution modelling [19]. In most of these studies, symmetric kernels (ex: gaussian) were used to estimate the probability distribution function. WS values are non-negative, while symmetric kernels have unbounded support leading to probability leakage below zero [20]. This is a well-known problem called the boundary effect, and several solutions have been proposed [21]. In this study, one of these solutions based on asymmetric kernel estimators [22] is adopted and introduced for WS distribution modelling. According to Hirukawa [22], asymmetric kernels are weight functions with support on the unit interval [0, 1] or the positive half-line. The effectiveness of the proposed approach was assessed by comparing it to another method based on mapping the W parameters in the study region.

The paper's novelty can be summarized as follows: First, a methodology to map WS distribution is proposed based on mapping WSQ. Quantiles are relatively easy to estimate from time series, while selecting an adequate RD can be tedious, requiring the fitting and evaluation of multiple distributions. Secondly, to the author's knowledge, this is the first study employing asymmetric kernels to model WS distribution. By combining the mapping of WSQ and asymmetric kernels, a fully non-parametric approach for WS distribution mapping is proposed in this study. The main advantage of the non-parametric approach is that it does not require specifying a unique distribution family to the region of interest. This allows to effectively combine all the available data in the region to build a more robust model in case the region does not have a homogenous wind regime which can be described by a single family of distribution functions.

The current paper is structured as follows. Section 2 illustrates the methodology of the proposed approach with the evaluation procedure. The study area and the dataset are presented in section 3. The results obtained are shown in section 4. In sections 5 and 6, the discussion of the findings and the conclusion are given, respectively.

2. Methodology

This study proposes a new approach for mapping WS distribution using regional information without constraining the region to a single distribution family. First, various WSQ are estimated at sampled locations in the region. Then, machine learning and WS covariates are used to map the quantiles, allowing the estimation of these WSQ at any unsampled location in the region. Finally, parametric, and non-parametric approaches are implemented to recover the WS distribution at unsampled locations from estimated quantiles. The proposed approach will be referred to as Quantile-based WS probability distribution Mapping (QWSM) in the next sections. The QWSM approach will be compared to another approach based on directly mapping the W parameters [8]. This method will be referred to as the W parameters mapping (WPM) in the next sections. A flowchart of the methodology is available in Figure 1.

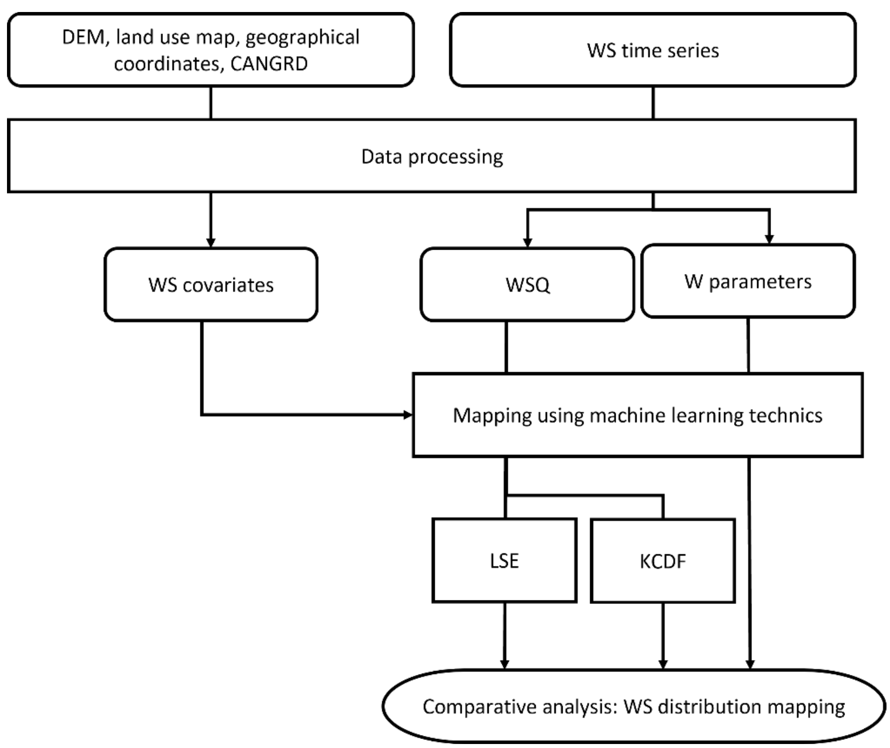


Figure 1: Methodology of the comparative analysis of WS probability distribution mapping approaches

2.1. Quantile-based WS probability distribution mapping

At the sampled locations in the region, WSQ at some fixed percentile points can be estimated from the sorted values of the hourly time series with the following general formula [23]:

$$W(P) = (1 - \gamma)X_{(j)} + \gamma X_{(j+1)} \quad (1)$$

Where P is the percentile point of interest, $X_{(j)}$ and $X_{(j+1)}$ are j-th order statistics. γ is a weight ($0 \leq \gamma \leq 1$) that is function of $j = floor(Pn + m)$, $m = \alpha + P(1 - \alpha - \beta)$ and $g = nP + m - j$. In case it is desired to obtain $W(P)$ as a continuous function of P, then $\gamma = g$ and selecting γ reduces to selecting α, β. Typical values of α, β are available in [23]. In this study, α, β were both set to 1/3 given quantiles that are approximately median-unbiased regardless of the WS true probability distribution [24]. Using equation 1, WSQ associated with the following 13 percentile points were estimated at the sampled locations: 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). Table 6 in Appendix I gives an overview of the distribution of the estimated WSQ.

These percentile points were chosen to cover the WS cumulative distribution functions (CDF) evenly, ensuring a representative estimation of the WSQ at various points along the distribution. In previous studies employing a similar modelling approach, varying numbers of percentile points have been modelled to estimate the probability distribution of a target variable. For instance, to forecast power load probability distribution, [25] modelled 20 percentiles evenly spaced between 1% and 96%. In another study, to map wind speed shear distribution, [26] estimated 11 percentiles evenly spaced between 1% and 99%. Additionally, to regionalize river temperature at ungauged locations, [27] estimated 17 percentiles non-evenly spaced between 0.05% and 99.95%. This diversity in the number of percentile point selections highlights a lack of consensus in the literature regarding the optimal number to ensure a comprehensive target distribution coverage. Nevertheless, it is worth noting that the number of percentiles selected in the current study falls within the range of those used in previous research.

A regression function was constructed between the observed WSQ and WS covariates. Two regression models were compared, the multilinear regression (LR) and the Gradient boosting trees [GBT: 28] model. Feature selection (FS) was performed using the minimum redundancy maximum relevance (MRMR) method [29] to reduce the complexity of the models and improve their performance. A comparative study of FS methods was carried out by Houndekindo and Ouarda [30]. They found that MRMR was among the most effective FS methods for WSQ estimation. Houndekindo and Ouarda [30] used MRMR with simple linear regression. However, the approach can be adapted to non-linear models such as tree-based gradient boosting. The FS method (MRMR) and the GBT model are presented in more detail in the following subsections.

2.1.1. MRMR approach for covariate selection

MRMR is a filter-based FS approach with the benefit of considering both the covariates' relevancy and redundancy during selection. Filter-based FS methods are computationally efficient algorithms and are agnostic to the regression model [31]. The MRMR algorithm uses an iterative approach to select the covariate (X_i) at each step with the best trade-off between its relevancy to the response variable (Y) and its redundancy relative to selected features from previous iterations. At the first step of the algorithm, the most relevant covariate is selected based on a measure of relevancy ($Rel(X_i, Y)$).

Let $Red(X_i, X_j)$ be a measure of the dependency between the covariates X_i and X_j and let S be the set of covariates selected during previous iterations. After the first step of the algorithm, S contains only the most relevant covariate ($\max_{X_i} [Rel(X_i, Y)]$) and the objective criterion at each subsequent iteration of the MRMR algorithm can be formulated in two ways:

$$\max_{X_i \notin S} \left[Rel(X_i, Y) / Red(X_i, X_j) \right] \quad (2)$$

Or

$$\max_{X_i \notin S} \left[Rel(X_i, Y) - \ Red(X_i, X_j) \right] \quad (3)$$

Several measures of relevancy and redundancy can be applied. In this study the following formulations of the MRMR objective criterion were compared:

$$MRMR - PC\colon \max_{X_i \notin S} \left[F(X_i, Y) / \left(\tfrac{1}{S} \textstyle\sum_{X_{j \in S}} \rho(X_i, X_j) \right) \right] \quad (4)$$

and

$$MRMR - MI\colon \max_{X_i \notin S} \left[I(X_i, Y) / \left(\tfrac{1}{S} \textstyle\sum_{X_{j \in S}} I(X_i, X_j) \right) \right] \quad (5)$$

Where $F(X_i, Y)$ is the F-statistic used to measure the relevancy, $\rho(X_i, X_j)$ is the Pearson correlation coefficient (PC) used to measure redundancy, $I(X_i, Y)$ is the mutual information (MI) used to measure relevancy and $I(X_i, X_j)$ is the MI used to measure redundancy. The MI between two random variables X and Y can be defined as follows:

$$I(X, Y) = \iint p(X, Y) \log(p(X, Y) / p(X) p(Y)) \ \ dxdy$$

(6)

The Python package scikit-learn [32] was used to calculate the MI between the variables.

2.1.2. Regression models

The LR model was implemented and used as a benchmark for the GBT model. Tree-based regression models such as GBT perform better than deep learning models on tabular data and often outperform other regression models [33]. The GBT algorithm works by fitting sequentially decision trees to the residuals from previous iterations. Contrary to the LR model, the GBT model can learn nonlinear relationships between the covariates and the response variable and is robust against non-informative covariates [34]. The GBT model is a popular regression model that has been successfully applied in studies for short-term wind power prediction [35], wind resource mapping [26], the selection of solar power plant location [36] and short-term prediction of solar irradiance [37].

The eXtreme Gradient Boosting package [XGB: 38] is a popular machine-learning library that implements the GBT algorithm efficiently. Several regularization strategies are available in XGB to improve the model performance and reduce computational time. To find adequate values for the parameters of XGB, a random search with 1000 iterations was implemented. Grid search and random search are popular algorithms used for hyperparameter tuning [39]. Grid search is a brute force algorithm that systematically tries all possible combinations of hyperparameter values within specified ranges. The algorithm can find the optimal hyperparameter values within the defined search space at the cost of increased computational resources and time. On the other hand, random search is a more efficient algorithm that does not guarantee the optimal solution but can find good hyperparameters [40]. Table 1 presents the hyperparameters of the XGB model that were tuned in the study.

Table 1: Hyperparameters of the XGB model

Hyperparameters used during training	**Search space (Min, Max, Step)**
Learning rate (Boosting learning rate)	(0.01, 0.1, 0.01)
Minimum loss reduction (gamma)	(0.0, 1.0, 0.1)
Maximum depth of the trees (max_depth)	(3, 10, 1)
Ratio of predictor to use during training (colsample_bytree)	(0.1, 0.7, 0.1)
Subsample ratio of the training data (subsample)	(0.1, 0.5, 0.1)

Number of trees (n_estimators)	(20, 300, 10)

2.1.3. Recovery of the WS distribution from WSQ

With estimated WSQ available at any non-sampled location, it is possible to fit different theoretical distribution functions using the LSE method. The LSE method is widely used for fitting WS probability distributions [41]. In their study, Jung and Schindler [26] applied the LSE method to recover the probability distribution of wind shear exponent from estimated quantiles of the same variable. LSE involves minimizing the sum of the square error (SSE) between the empirical cumulative probability (ECDF) and the theoretical CDF to determine the best-fitting parameters of the theoretical distribution function. Let $\widehat{W_i}$ be the predicted WSQ and $\hat{F}(W_i))$ their associated CDF, the SSE can be written as follows:

$$SSE = \sum_{i=1}^{13}\left[\hat{F}(W_i) - F(\widehat{W_i};\, \hat{\theta})\right]^2 \quad (7)$$

Where: $F(\widehat{W_i};\, \hat{\theta})$ corresponds to the cumulative probability function of $\widehat{W_i}$ with estimated parameter $\hat{\theta}$. The W, Log-Normal (LN), Rayleigh (R) and Generalized Gamma (GG) distribution were fitted to the estimated WSQ.

Additionally, it is proposed to recover the WS distribution at unsampled locations using asymmetric KCDF. The asymmetric kernels method represents one of the solutions to the boundary effects that appear when using symmetric kernels with bounded random variables (ex.: WS values are bounded on [0, ∞]). By combining WSQ mapping and asymmetric kernel fitting, this study proposes a fully non-parametric method for wind speed distribution mapping. Traditional parametric methods might introduce bias if the selected RD does not align with the data. The non-parametric approach can adapt to various WS distribution patterns without being restricted by specific parametric assumptions. This flexibility is necessary for a region with complex and diverse wind behaviors. In addition, combining the

WSQ mapping and asymmetric kernel fitting avoids the tedious process of testing and evaluating different probability distribution functions to model WS.

The general expression for the asymmetric KCDF is given by [21]:

$\hat{F}(w) = 1/n \sum_{i=1}^{n} \bar{K}_{w,b}(W_i),$

(8)

Where:

$b > 0$ is the bandwidth and $\bar{K}(\cdot)$ is the CDF of an asymmetric kernel function. In this work, the Birnbaum-Saunders (BS), the Log-Normal (LN) and W asymmetric kernel functions were tested [21, 42]:

$$\hat{F}^{BS}(w) = 1/n \sum_{i=1}^{n} \bar{K}_{BS}\left(W_i;\ w, \sqrt{b}\right), \quad (9)$$

$$\hat{F}^{LN}(w) = 1/n \sum_{i=1}^{n} \bar{K}_{LN}(W_i; \log w\,, \sqrt{b}), \quad (10)$$

$$\hat{F}^{WB}(w) = 1/n \sum_{i=1}^{n} \bar{K}_{WB}(W_i;\ w/\Gamma(1+b), 1/b), \quad (11)$$

Where:

$\bar{K}_{BS}(x;\ \beta, \alpha) = 1 - \Phi\left(\left(\sqrt{x/\beta} - \sqrt{\beta/x}\right)/\alpha\right),\ \beta, \alpha > 0,$

(12)

$\bar{K}_{LN}(x; \mu, \sigma) = 1 - \Phi((\log x - \mu)/\alpha),\ \mu, \sigma > 0,$

(13)

$$\bar{K}_{WB}(x; \alpha, \beta) = \exp(-(x/\beta)^{\alpha}),\ \ \alpha, \beta\ > 0, \quad (14)$$

$\Phi(\cdot)$ is the CDF of the standard normal distribution and $\Gamma(\cdot)$ is the gamma function.

The optimal bandwidths can be selected by minimizing the Mean Integrated Square Error (MISE)

$$MISE = \int_0^{\infty} MSE\left(\hat{F}(w)\right)\, dw \quad (15)$$

Where:

$$MSE\left(\hat{F}(w)\right) = E\left[\left(\hat{F}(w) - F(w)\right)^2\right] \quad (16)$$

Mombeni, et al. [21] derived the asymptotical optimal bandwidth of $\overline{K}_{BS}$ and $\overline{K}_{WB}$ with respect to the MISE:

$$b_{opt}^{BS} \approx \left\{\int_0^\infty xf(x)dx\right\}^{2/3} \left\{\pi^{\frac{1}{2}} \int_0^\infty \left(xf(x) + x^2 f'(x)\right)^2 dx\right\}^{-2/3} n^{-2/3} ,$$

(17)

$$b_{opt}^{WB} \approx \left\{36 \ln 2 \int_0^\infty xf(x)dx\right\}^{\frac{1}{3}} \left\{\pi^4 \int_0^\infty \left(x^2 f'(x)\right)^2 dx\right\}^{-\frac{1}{3}} n^{-\frac{1}{3}} , \quad (18)$$

Lafaye de Micheaux and Ouimet [42] proposed the following asymptotical optimal bandwidth with respect to the MISE for $\overline{K}_{LN}$:

$$b_{opt}^{LN} \approx \left\{\frac{1}{\sqrt{\pi}} \int_0^\infty x\, f(x)dx\right\}^{\frac{2}{3}} \left\{4 \int_0^\infty \frac{x^2}{4} \left(f(x) + xf'(x)\right)^2 dx\right\}^{-\frac{2}{3}} n^{-\frac{2}{3}} , \quad (19)$$

The optimal bandwidth with respect to the MISE was selected under the assumption that the W with parameters estimated using the predicted WSQ and the LSE method was the target distribution. The reason for employing the W distribution in the paper is two-fold: First, it is the parametric probability distribution function most commonly used to model WS; Secondly, it is convenient because its CDF can be linearized with respect to its parameters and the WSQ. As a result, finding the best-fitting parameters with the LSE method is equivalent to solving a linear equation and does not require an optimization algorithm.

2.2. Weibull parameter mapping

In previous studies, to estimate the WS probability distribution at unsampled locations, machine learning models were used to map the parameters of a RD. The approach selects a single distribution family for the entire region. Then, the distribution function parameters are fitted at the sampled locations, and a regression model is built between the parameters and WS covariates. Jung [11] selected the Wakeby distribution as the RD in southwest Germany based on two goodness of fit measures: Kolmogorov-Smirnov statistic and the coefficient of determination. For a review of criteria used for the identification of adequate WS distributions the reader is referred to Ouarda, et al. [43]. Veronesi, et al. [8] selected the W as the RD in the UK due to its widespread use in modelling WS, and convenience as it requires only two parameters to characterize the WS probability distribution. The W was also adopted as the RD in this study to evaluate the QWSM approach. The W parameters were estimated with the LSE method and the best-fitting parameters were mapped in the region using the WS covariates described in section 3 and the LR and XGB regression models described in section 2.1.2. The MRMR algorithm was also applied to identify the best set of covariates to include in the regression models.

2.3. Model validation

To evaluate the QWSM and the WPM, holdout and 5-fold cross-validation were implemented with the available samples. During the holdout procedure, parts of the samples were withheld (the validation set) before model training and parameter tuning and used to evaluate the final model generalization performance. During 5-fold cross-validation, the training samples were divided into five approximately equal subsets. Then, the holdout method was implemented five times by considering each subset as the validation set and training the model on the remaining subsets.

The following metrics were calculated based on the observed (y_i) and estimated $(\hat{y}_i)$ values:

$$R^2 = 1 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad (20)$$

$$RMSE = \sqrt{1/n \sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (21)$$

$$MAE = 1/n \sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (22)$$

The evaluation of the GOF of the estimated WS probability distribution was based on the percentage probability plot [PP plot: 44]. The PP plot compares the ECDF to the estimated CDF. During cross-validation and validation, the R^2, the RMSE and the MAE defined in equations 20, 21 and 22, respectively, were used to evaluate the degree of association between the ECDF and the CDF. Horst [45] noted that the PP plot has strong discriminatory power in high-density regions of the distribution (i.e.: the middle of a distribution), where the CDF changes more rapidly with the WS values compared to low-density regions (i.e.: the tails). Regions of the probability distribution with high density are the most crucial for wind energy production. Also, in their reviews on WS distribution selection, Jung and Schindler [9] observed that the most widely used GOF metrics were based on the PP plot.

The Kolmogorov–Smirnov statistic (D) is an alternative measure that was used to compare the ECDF and the CDF:

$$D = \max |F_n(W_i) - \hat{F}(W_i)| \quad (23)$$

Where $F_n(W_i)$ is the ECDF and $\hat{F}(W_i)$ is the estimated CDF.

The ECDF was calculated with the Weibull plotting position [46] giving unbiased non-exceedance probabilities regardless of the underlying distribution of the data [47]:

$$F_n(W_i) = i/(n+1) \quad (24)$$

Where $i = 1, \dots, n$ is the rank of the WS values after sorting them in ascending order.

3. Study area and dataset

The study was conducted on data from the whole Canada representing a total area of 9,984,670 square kilometers. Hourly WS data from 207 meteorological stations located throughout the country were used for the research. From Environment and Climate Change Canada (ECCC) historical climate database, stations with at least 20 years of recent WS record were selected. Additional filtering was performed to eliminate all stations with more than ten years of record having two months of missing data. Figure 2 illustrates the geographical location of the 207 stations that were selected after filtering. From the available stations, 155 (white triangles in figure 2) were used for FS, model training and cross-validation and the remaining stations (black dots in figure 2) were used to validate the final model as explained in section 2.3.

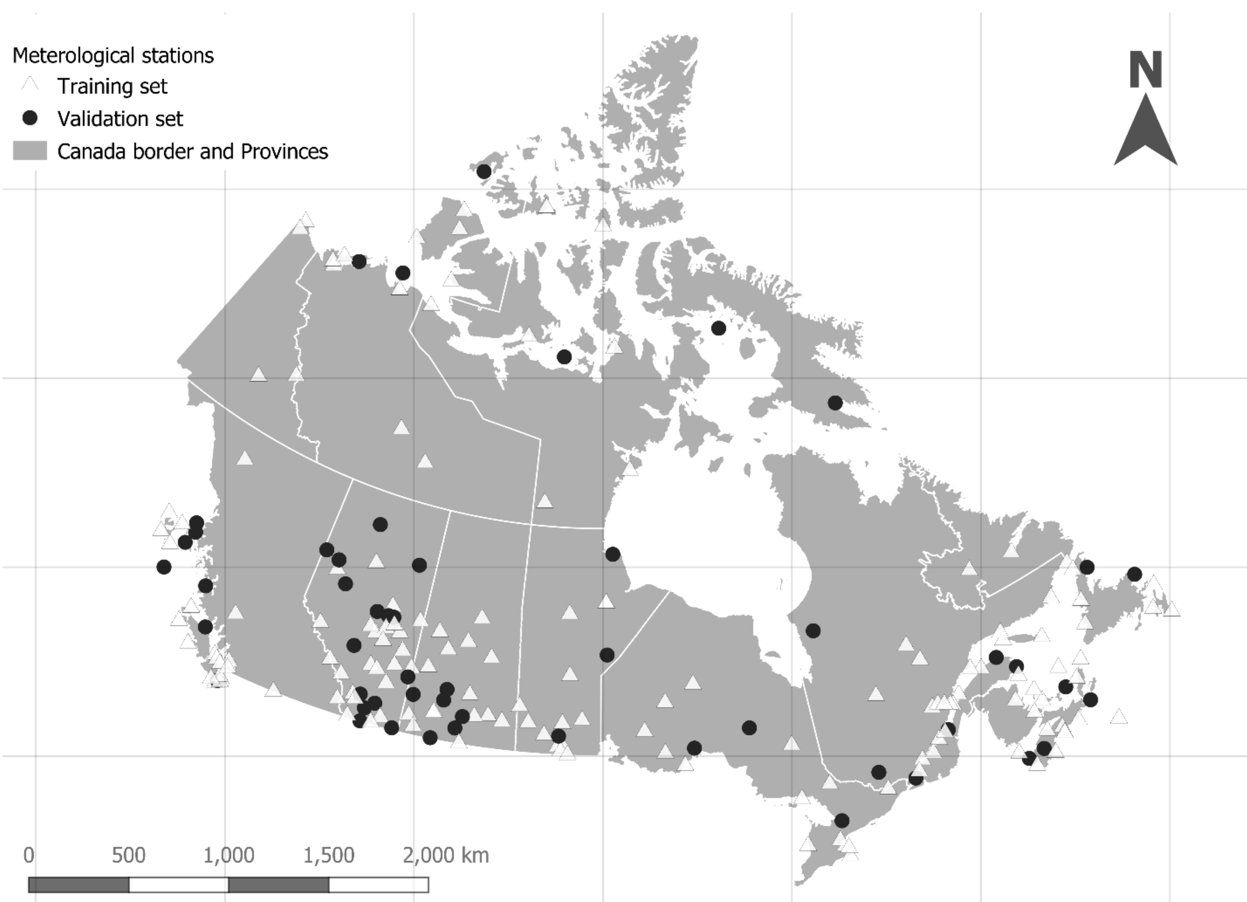


Figure 2: Spatial distribution of the training and validation stations used in this study.

The following four types of covariates were used with the regression models to either estimate the WSQ or the W parameters: topographic, climatic, geographic, and surface roughness length. The topographical covariates were created using the WhiteboxTools [48] and a 30m resolution global DEM

[49]. Seasonal and annual trends of mean temperature data were acquired from the Canadian gridded temperature and precipitation anomalies (CANGRD) dataset (available at https://climate-change.canada.ca/climate-data/#/historical-gridded-data). Surface roughness length was extracted from a 2015 Canada land use map [50] resampled at different spatial resolutions using majority resampling (i.e.: most popular value in a defined radius). A surface roughness length was associated with each land use type based on a lookup table proposed by Wierngа [51]. Table 7 in Appendix II provides more details about the covariates.

4. Results

4.1. Performance of regression models

The LR and the XGB models were fitted with covariates selected using MRMR-PC and MRMR-MI. The results of comparing the different combinations of regression models and FS methods are presented in Tables 3 and 4 for QWSM and the WPM, respectively. Figure 3 details the average R^2 for estimating the 13 WSQ and the two W parameters (shape and scale). The comparisons using cross-validation and validation lead to very similar results, indicating, in general, that XGB with MRMR-PC outperforms the other combinations of regression models and FS methods. Indeed, XGB gave better results than LR in most cases, and MRMR-PC was more effective than MRMR-MI for FS in the study. In the few cases where LR outperformed XGB, the performance difference was marginal and inconsistent during cross-validation and validation (see, for instance, P8 in Figures 3a and 3b). Tables 3 and 4 indicate that the improved performance of XGB with MRMR-PC is consistent across all metrics. Hereon, only the results obtained with estimations from the top-performing FS and regression model (MRMR-PC + XGB) will be presented.

Figure 4 displays the spatial distribution of the RMSE (WSQ) scaled by the actual WS median for the validation set. This representation allows for comprehensive visualization of the accuracy and variability of the model's predictions across different locations. Scaling the RMSE with the actual median provides a relative measure of error that can be compared and interpreted meaningfully. The spatial distribution of the scaled RMSE revealed that the model exhibited acceptable performances in estimating the WSQ in regions with sparse training samples highlighting its generalization capability.

Table 3: Average performance metrics for the estimation of WSQ

Validation Methods	**Regression model**	**MRMR**	**MAE**	**R^2**	**RMSE**
			km/h		km/h
Cross-validation	LR	MI	3.59	0.23	4.90
Cross-validation	LR	PC	3.40	0.26	6.11
Cross-validation	XGB	MI	3.24	0.42	4.30
Cross-validation	**XGB**	**PC**	**3.08**	**0.47**	**4.07**
Validation	LR	MI	3.64	0.36	4.48
Validation	LR	PC	3.24	0.46	4.19
Validation	XGB	MI	3.30	0.46	4.22
Validation	**XGB**	**PC**	**3.00**	**0.57**	**3.74**

Table 4: Average performance metrics for the estimation of the W parameters

Validation Methods	**Regression model**	**MRMR**	**MAE**	**R^2**	**RMSE**
Cross-validation	LR	MI	1.88	0.27	2.47
Cross-validation	LR	PC	2.02	-	4.79
Cross-validation	XGB	MI	1.83	0.45	2.27
Cross-validation	**XGB**	**PC**	**1.61**	**0.48**	**2.12**
Validation	LR	MI	2.07	0.32	2.42
Validation	LR	PC	1.76	0.37	2.27
Validation	XGB	MI	1.75	0.42	2.16
Validation	**XGB**	**PC**	**1.58**	**0.48**	**1.97**

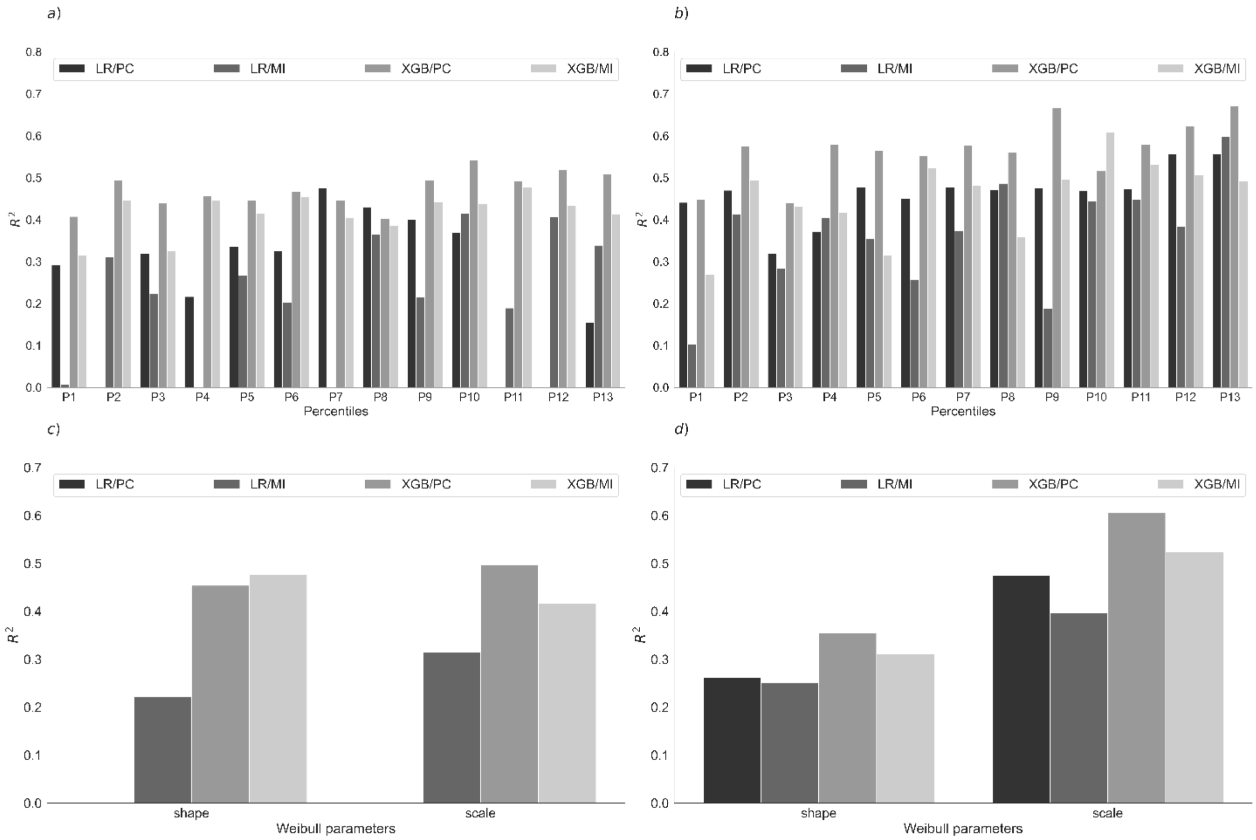


Figure 3: Performance of LR and XGB for the estimation of the WSQ (a and b) and the W parameters (c and d) during cross-validation (a and c) and validation (b and d). Note: Negative values of R^2 were set to zero

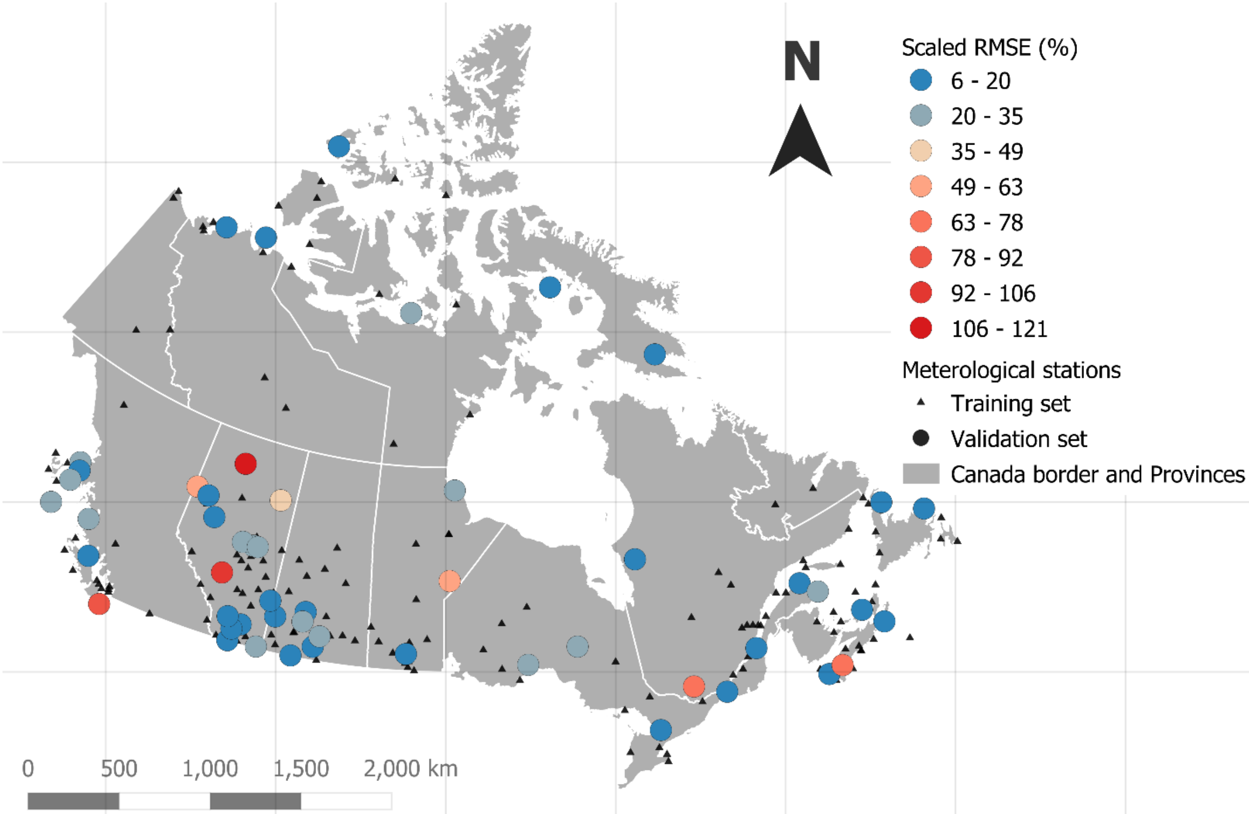


Figure 4: Spatial distribution of the scaled RMSE (WSQ) of the validation set

4.2. Wind speed distribution mapping

This section presents the results of the comparative analysis between the QWSM and WPM. Table 5 shows the mean values of the GOF metrics. In general, it is observed that the QWSM gave a better fit than WPM for the considered metrics. Also, QWSM/W gave better fit than WPM. According to the R^2, RMSE and MAE criteria, QWSM/W and QWSM/GG were the best-performing methods, and their performances are very similar to QWSM/KCDF/BS and QWSM/KCDF/LN. However, during cross-validation and validation, the Kolmogorov-Smirnov statistic (D) seemed to favor QWSM/KCDF/LN and QWSM/KCDF/BS. The distribution of the GOF measures was represented using boxplots in Figure 4. The most noticeable difference in the distribution of the GOF measures was observed with D when comparing the different approaches. The methods based on QWSM/KCDF/LN and QWSM/KCDF/BS resulted in smaller D values and less variability in the same GOF measure compared to other approaches.

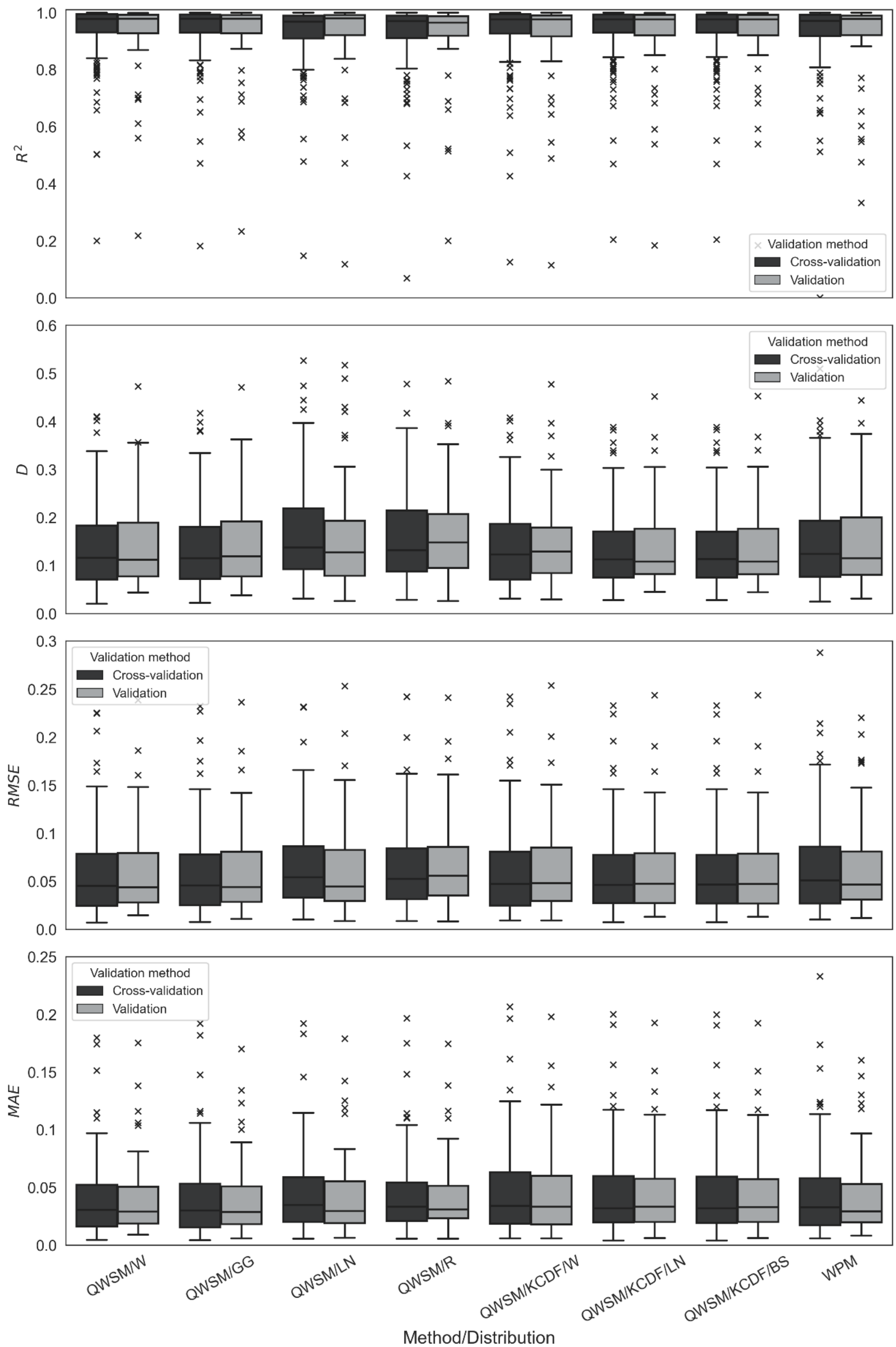
Furthermore, the different methods were evaluated by comparing the observed and estimated WSQ across ten equidistant percentiles ranging from 0.1 to 0.9. The outcome of this analysis (Figure 6) indicated that the QWSM methods often outperformed the WPM for the considered WSQ. Methods based on QWSM with the asymmetric kernels tend to give comparable performances to the parametric methods in the middle of the distribution (ex.: 0.4, 0.5, 0.6 percentiles). While in the tails (ex.: percentiles 0.1, 0.9) the parametric methods showcased a better performance than the non-parametric methods.

Table 5: Mean value of the GOF measures

Distribution	Validation Methods	D	MAE	R^2	RMSE
QWSM/GG	Cross-validation	0.137	**0.039**	0.938	**0.058**
QWSM/GG	Validation	0.147	0.041*	0.922*	0.062*
QWSM/KCDF/BS	Cross-validation	**0.131**	0.043	0.938	0.059
QWSM/KCDF/BS	Validation	0.143*	0.045	0.920	0.063
QWSM/KCDF/LN	Cross-validation	**0.131**	0.044	0.937	0.059
QWSM/KCDF/LN	Validation	0.143*	0.045	0.920	0.063
QWSM/KCDF/W	Cross-validation	0.137	0.046	0.932	0.061
QWSM/KCDF/W	Validation	0.150	0.046	0.911	0.064
QWSM/LN	Cross-validation	0.165	0.042	0.93	0.064
QWSM/LN	Validation	0.165	0.043	0.913	0.065
QWSM/R	Cross-validation	0.157	0.042	0.926	0.065
QWSM/R	Validation	0.168	0.044	0.908	0.069
QWSM/W	Cross-validation	0.136	**0.039**	**0.939**	**0.058**
QWSM/W	Validation	0.147	0.041*	0.921	0.062*
WPM	Cross-validation	0.144	0.042	0.93	0.062
WPM	Validation	0.152	0.043	0.910	0.065

Note: The best-performing methods are indicated in bold for the cross-validation and marked with * for the validation.

Table 5



Validation method
Cross-validation
Validation
R^2
D
$RMSE$
MAE
QWSM/W
QWSM/GG
QWSM/LN
QWSM/R
QWSM/KCDF/W
QWSM/KCDF/LN
QWSM/KCDF/BS
WPM
Method/Distribution

449

Figure 5: GOF of estimated WS probability distribution. Note: Negative values of R^2 were set to zero

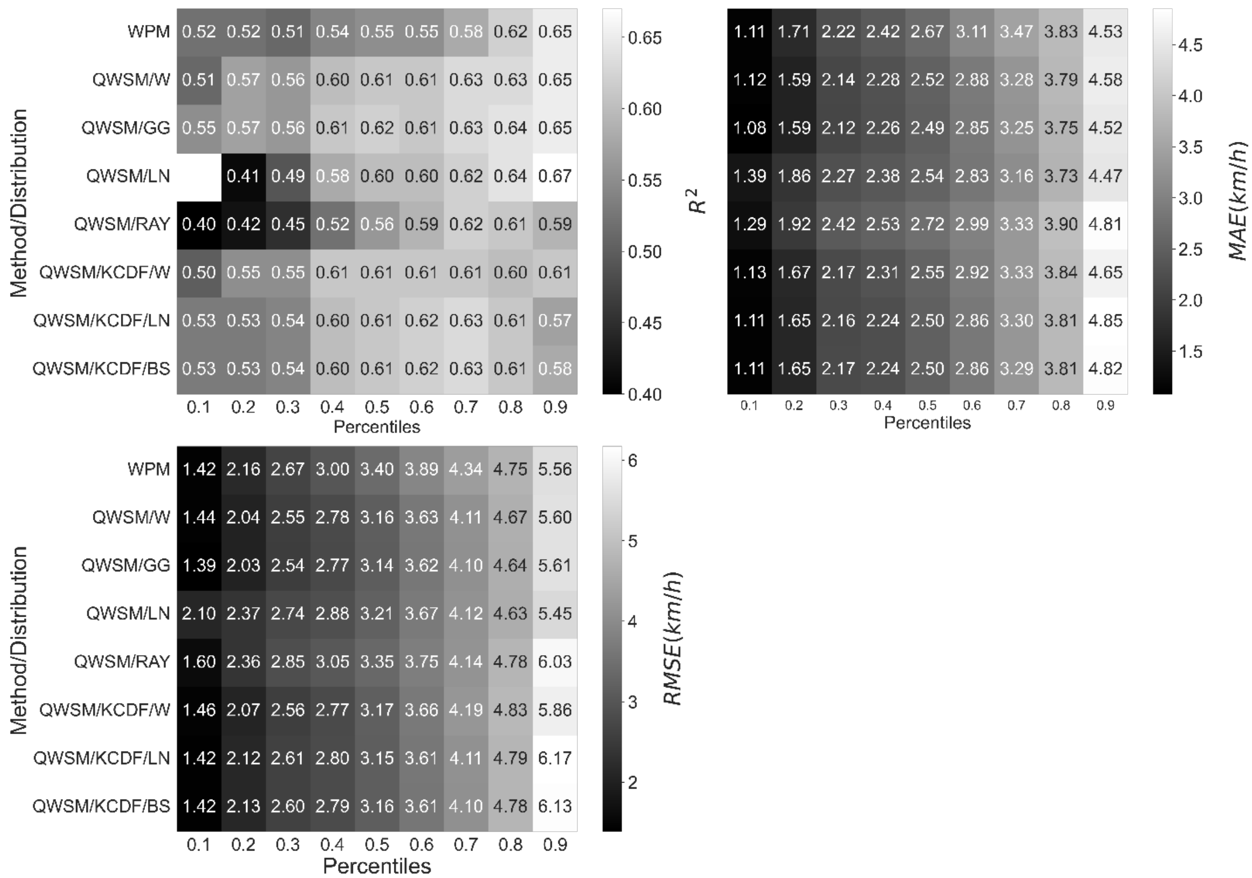


Figure 6: Performance metrics for observed WSQ and estimated WSQ using QWSM and WPM (validation set)

In Figure 7, the P-P plot, the CDF, and the probability density function (PDF) plot of 3 validation samples are presented for illustration purposes. These plots offer a comprehensive visual analysis of the actual and estimated WS distribution agreement. Recall that QWSM/W was selected as the target distribution to estimate the optimal bandwidth for all KCDF. However, it is observed that the kernel PDFs exhibited more flexibility than QWSM/W. The W kernel demonstrated more flexibility than the BS and LN kernels, while both gave an almost identical PDF.

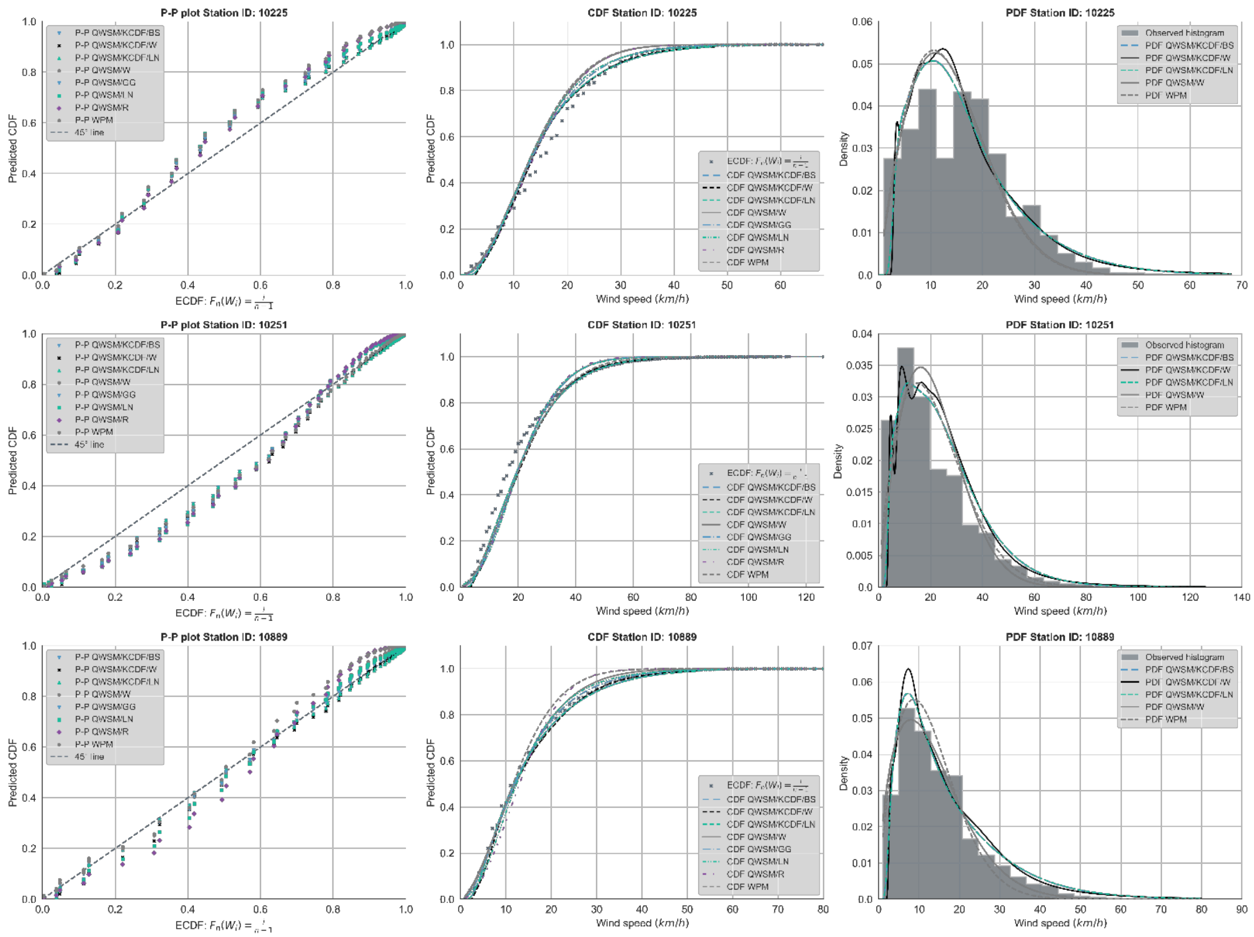


Figure 7: PP plot, CDF plot and PDF plot of estimated wind speed probability distributions

5. Discussion

The comparison of the regression models indicates that the non-linear model (XGB) outperformed the linear model (LR) for the estimation of WSQ and the W parameters. The superior performance of the XGB model suggests that there are non-linear associations and interactions between the covariates and the WS response variables (WSQ and W parameters). The XGB model can effectively capture these non-linear relationships, leading to more accurate and precise estimates than the linear model. There is potential for further improvement in the performance of the XGB model by conducting a more comprehensive hyperparameter tuning. A random search was employed for the XGB hyperparameter tuning and proved sufficient to demonstrate the superiority of the XGB model over the LR model. However, a more extensive hyperparameter tuning process, such as grid search or Bayesian optimization [52], could be conducted to thoroughly search for the optimal combination of hyperparameters that maximizes the model's performance.

The study also found that MRMR-PC was more effective for FS than MRMR-MI. MI can assess linear and nonlinear dependencies between variables, and it was initially expected that combining MRMR-MI with XGB would outperform the combination of MRMR-PC with XGB. However, similar results were observed by Ren, et al. [53] in the field of hydrology. The authors discovered that a FS method based on the partial Pearson correlation outperformed FS methods based on MI (including MRMR-MI) when applied with linear and nonlinear regression models for monthly streamflow forecasting. The study attributed these results to the possibility that the relationship between the covariates and the target variable in their models exhibited more linearity than nonlinearity. Similar conclusions may be formulated in this study, suggesting that the gain in performance achieved using the XGB could also be attributed to other characteristics of the models, such as its robustness against redundant features and collinearity within the features set. Despite these findings, it is still recommended to evaluate different FS methods. Different scenarios or datasets may yield different results.

It is well known that wind speed and other climatic variables like humidity, pressure, and temperature are interconnected. The main challenge in using climatic variables for estimating wind speed at unsampled locations is that those variables should also be unavailable. Gridded climate data can be used as an alternative source of climatic covariates. This study only used gridded climate data of long-term temperature trends as climatic covariates. Investigating the applicability of other gridded climate data as covariates for WS distribution mapping in future studies is recommended.

Veronesi, et al. [8] reviewed the performance of physical and statistical methods for wind resources assessment. They found that most studies applying statistical methods reported an RMSE of around 1 m/s on their validation set when considering the central tendency of the wind speed distribution (ex.: mean). In the current study, the average RMSE for estimating the median wind speed obtained was 3.28 km/h (0.87 m/s), and the average MAE was 2.62 km/h (0.69 m/s). These results seem to agree with previous studies. However, as was pointed out by Veronesi, et al. [8], results from different studies are generally difficult to compare as different datasets, regions and techniques were covered in these studies.

In general, based on the evaluation of the GOF, QWSM demonstrated a better fit compared to WPM. This result may be explained by the fact that the estimation of the WS distribution from WSQ may be less sensitive to mapping error compared to WPM. For instance, in the case of the WPM, minor errors in mapping the W parameter could have disproportionate effects on the overall resulting shape of the wind speed distribution. In contrast, with the QWSM, the implications of mapping errors are less severe, as inaccuracies in wind speed quantile mapping seemed to have a smaller impact on the overall distribution's shape. Consequently, the QWSM approach exhibits enhanced robustness against errors in mapping, rendering it a more dependable framework for wind speed distribution mapping.

The non-parametric approach with the BS and LN KCDF gave slightly better results than the parametric approach when considering the Kolmogorov-Smirnov statistic. The non-parametric method does not require fixing a regional distribution and can adequately recover the WS distribution from the estimated quantiles. Parametric methods require fitting the data to a specific probability distribution family, which may introduce bias if the assumed distribution does not align with the underlying distribution. Another potential source of bias common to both methods (i.e.: QWSM, WPM) is related to the regression models used to estimate either the WSQ or the RD parameters. It should be noted that the bulk of the bias of the QWSM + KCDF method arises from the regression model used to map the WSQ in the region.

Thus, the non-parametric approach can reduce potential biases by minimizing the assumptions. The proposed approach becomes particularly interesting in regions where the wind regime exhibits significant variations, and no single distribution family is suitable for all locations within the region. With their constraints, parametric methods may struggle to capture the diversity of complex patterns that can be present in such regions. In contrast, with its flexibility, the non-parametric approach can be more appropriate and should yield more accurate results. Alternatively, it is possible to segregate the regions into sub-regions and select a different RD for each sub-region. However, this would reduce the number of samples used to learn the relationship between the covariates and the RD parameters, potentially leading to a loss in performance. For WS values located in the distribution's tails (for instance, extreme values), opting for the QWSM method with parametric distribution functions would be more suitable. This recommendation is based on the finding that these parametric approaches exhibited superior performance compared to non-parametric approaches in this case.

Mapping the WSQ in this study involved extracting the quantiles from the time series and then using a regression model that estimates the conditional mean of the quantiles given the covariates. An alternative approach could be directly estimating the conditional quantiles using a quantile regression [54-57] model incorporating the covariates. Quantile regression is a statistical technique that allows estimating specific quantiles of the response variable rather than focusing solely on the conditional mean.

The main drawback of the QWSM approach is that the number of independent variables (quantiles) that need to be mapped to recover the WS distribution would often be superior to the number of the RD parameters that require mapping in the WPM approach. Fitting these individual regression models can become time-consuming and resource intensive. However, some quantile regression models can simultaneously estimate multiple quantiles [57, 58], providing a more efficient approach compared to building separate regression models for each quantile. Also, when estimating multiple quantiles simultaneously, additional constraints can be formulated to enforce monotonicity [59] and avoid the issue of quantile crossing that arises when estimating the quantiles independently. It is worth mentioning that a gradient-boosting model [60] was recently proposed to simultaneously estimate the parameters of a probability distribution conditioned on some covariates. This model could be used to estimate the parameters of a RD simultaneously rather than building an independent model for each parameter.

Modern wind turbine hub heights vary between 80m and 100m, while wind speed data are conventionally collected at 10m at meteorological stations. As a result, a technique for extrapolating wind speed data to hub height becomes necessary (ex.: the power law). Such techniques can extend the method proposed in this study to map wind speed distribution at hub height. Nevertheless, it is worth noting that such extrapolation introduces a notable increment in the uncertainty of the outcomes.

Jung and Schindler [26] proposed a technique for mapping wind shear distribution, allowing the wind speed distribution to be mapped at any standard hub height. Jung and Schindler [26] selected the Dagum family distribution to represent the wind shear distribution. In future research, the non-parametric approach proposed in this study could be adapted to map wind shear distribution without prior assumptions about its distribution. Also, future studies can explore the possibility of extending the proposed approach to other types of climatic variables, such as temperature and solar irradiation.

The approach proposed in this study can provide valuable information to estimate wind resources over a large area during a prospecting phase. Once an area that meets the necessary socio-economic requirements and showcases sufficient wind potential is identified, alternative methods are available to evaluate the wind flow at the microscale. An example of such an approach involves conducting wind flow simulations via Computational Fluid Dynamics (CFD), especially in complex terrain [61]. The implementation of a CFD model requires the provision of initial wind data, which can be sourced from outputs generated by Numerical Weather Prediction [NWP: 62, 63, 64]. NWP models entail considerable computational costs compared to statistical methodologies proposed herein. A compelling avenue of research would involve comparing the performance of NWP and statistical models for CFD model input and developing methods to combine statistical and CFD models to assess microscale wind flow dynamics.

6. Conclusion

A fully non-parametric approach was developed to map wind speed distribution. The new method was compared to a more traditional approach based on mapping the parameters of a regional distribution. The results of the comparative analysis highlighted the superiority of the proposed approach. The main conclusions of the paper are summarized as follow:

- The non-parametric approach is more practical as it does not require fitting and evaluating several distribution functions to the available wind speed data. In the proposed method, wind speed quantiles can be easily extracted from the time series and mapped using suitable machine-learning techniques. At any location in the study area, the entire wind speed distribution can be recovered from the estimated wind speed quantile by fitting asymmetric kernel estimators. The proposed approach is free from any assumption on the wind speed probability distribution family in the region that can bias the analysis. The non-parametric approach is recommended for mapping wind speed distribution in regions with a highly variable wind regime. The analysis indicates that the fully non-parametric approach improved the Kolmogorov-Smirnov statistic by 9% on average during validation.
- Compared to the regional distribution parameter mapping approach, quantile-based wind speed distribution mapping can be slower to implement as it requires the estimation of multiple wind speed quantiles. However, with the advancement in quantile regression models, it is possible to

build a single regression model to predict multiple quantiles. This type of quantile regression model should reduce the computational burden associated with the proposed approach.

- The Gradient boosting trees model outperformed the multilinear regression model for mapping wind speed quantiles and the Weibull parameters. At the same time, feature selection based on the Pearson correlation coefficient was more effective than the Mutual information. Utilizing the Gradient Boosting Trees model and feature selection based on the Pearson correlation coefficient resulted in a 23% improvement in R^2 during validation compared to the second-best model for estimating wind speed quantiles.
- It should be noted that symmetric kernels could also be fitted to the estimated wind speed quantiles, with some probabilities associated to small negative wind speed values. Using an asymmetric kernel effectively avoids probability leakage at the boundary of the lower tail of the wind speed probability distribution.
- The proposed approach is easily portable to regions with sparsely available wind speed measuring stations. The other data sources used in the study (ex.: DEM and land use map) are often freely accessible from global datasets covering most regions of the world.

7. 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 wind speed data was freely accessed from Environment and Climate Change Canada database, and the Japan Aerospace Exploration Agency provided free access to the digital elevation model. The authors would like to extend their gratitude to the Editor, Prof. Nesreen Ghaddar, and three anonymous reviewers for their comments and suggestions, which significantly improved the quality of the paper.

Appendix I Statistics of the estimated wind speed quantiles

Table 6: Statistics of the estimated wind speed quantiles

Percentile	Mean (km/h)	Std (km/h)	Min (km/h)	25% (km/h)	50% (km/h)	75% (km/h)	Max (km/h)
5	4.2	1.6	1	3	4	5	9
12.5	6.4	2.2	2	5	6	7	13
20	8.2	2.9	3	6	7	9.5	18
27.5	9.9	3.4	4	7	9	12	20
35	11.6	4.0	4	9	11	14.5	24
42.5	13.4	4.4	5	11	13	16.5	28
50	15.2	4.9	6	12	15	19	31
57.5	17.1	5.4	6	13	17	20	35
65	19.4	6.1	7	15	19	23.5	39
72.5	21.9	6.7	7	17	21	26	44
80	24.8	7.6	9	19	24	30	51
87.5	29.0	8.7	11	22.5	28	35	59
95	36.2	10.9	14	28.5	35	44	74

Appendix II. Wind speed covariates

Table 7: Overview of the WS covariates

Predictor	Description	Spatial scale
Altitude	Altitude of the location in meter.	
Aspect	Slope orientation in degree.	100m, 500m, 1000m, 1500m, 2000m
Deviation from mean elevation	Difference between the grid cell elevation and the mean of its neighbouring cells normalized by the standard deviation.	100m, 500m, 1000m, 1500m, 2000m
Difference from cell mean elevation	Difference between the grid cell elevation and the mean of its neighbouring cells.	100m, 500m, 1000m, 1500m, 2000m
Difference of Gaussian	Difference between two copies of the DEM smoothed with two different gaussian kernel. Measure land surface curvature.	(100m, 500m), (100m, 1000m), (500m, 1000m), (300m, 500m), (1000m, 2000m), (1000m, 1500m), (100m, 2000m), (500m, 2000m)
Distance to coast	The location distance to the coast	
Elevation percentile	Percentile of the grid cell elevation relative to the neighbouring cells.	100m, 500m, 1000m, 1500m, 2000m

Gaussian curvature	Product between the maximal and the minimal curvature. Measure of surface curvature [65].	100m, 500m, 1000m, 1500m, 2000m
Geographical coordinates	Geographical coordinates of the location.	
geomorphologic phonotypes (geomorphons)	Landform element classification with the geomorphons-based method [66].	
Laplacian of Gaussian	Derivative filter used to highlight location of rapid elevation change.	100m, 500m, 1000m, 1500m, 2000m
Maximal curvature	Measure of surface curvature [67].	100m, 500m, 1000m, 1500m, 2000m
Mean curvature	Measure of surface curvature [67].	100m, 500m, 1000m, 1500m, 2000m
minimal curvature	Measure of surface curvature [65].	100m, 500m, 1000m, 1500m, 2000m
Pennock landform class	Landform classification based on the slope and curvature of the grid cell [68].	
plan curvature	Measure of surface curvature [65].	100m, 500m, 1000m, 1500m, 2000m
Relative topographical position	Normalized measure of the grid cell elevation relative to its neighbouring cells.	100m, 500m, 1000m, 1500m, 2000m
Ruggedness index	A measure of the local terrain heterogeneity [66, 69]	100m, 500m, 1000m, 1500m, 2000m
Slope	Slope at the grid cell.	100m, 500m, 1000m, 1500m, 2000m
Standard deviation of slope	Measure of surface roughness [70].	100m, 500m, 1000m, 1500m, 2000m
Surface area ratio	Measure of the surface roughness [71].	100m, 500m, 1000m, 1500m, 2000m
Surface roughness length	Surface roughness length estimated from land use map.	100m, 500m, 1000m, 1500m, 2000m
tangential curvature	Measure of surface curvature [65].	100m, 500m, 1000m, 1500m, 2000m
Total curvature	Measure of surface curvature.	100m, 500m, 1000m, 1500m, 2000m
Temperature trend	Seasonal and annual trends of mean temperature change between 1948-2018.	

References

[1] Y. Zhou, P. Luckow, S.J. Smith, L. Clarke. Evaluation of Global Onshore Wind Energy Potential and Generation Costs. Environmental Science & Technology. 46 (2012) 7857-64. 10.1021/es204706m

[2] Global Wind Energy Council. GWEC Global Wind Report 2022. GWEC Global Wind Report. Global Wind Energy Council, Brussels, Belgium, 2022.

[3] C. Jung, D. Schindler, J. Laible. National and global wind resource assessment under six wind turbine installation scenarios. Energy Conversion and Management. 156 (2018) 403-15. https://doi.org/10.1016/j.enconman.2017.11.059

[4] F. Houndekindo, T.B.M.J. Ouarda. Statistical approaches for wind speed estimation at ungauged or partially gauged locations, review, and open questions (Under review). Institut national de la recherche scientifique, Centre Eau Terre Environnement. (2023).

[5] W. Luo, M.C. Taylor, S.R. Parker. A comparison of spatial interpolation methods to estimate continuous wind speed surfaces using irregularly distributed data from England and Wales. International Journal of Climatology. 28 (2008) 947-59. https://doi.org/10.1002/joc.1583

[6] W. Ye, H.P. Hong, J.F. Wang. Comparison of Spatial Interpolation Methods for Extreme Wind Speeds over Canada. Journal of Computing in Civil Engineering. 29 (2015) 04014095. doi:10.1061/(ASCE)CP.1943-5487.0000429

[7] V. Nelson, K.r.e. Starcher. Wind Energy: Renewable Energy and the Environment. CRC Press, Boca raton, Floride USA, 2018.

[8] 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

[9] C. Jung, D. Schindler. Wind speed distribution selection – A review of recent development and progress. Renewable and Sustainable Energy Reviews. 114 (2019) 109290. https://doi.org/10.1016/j.rser.2019.109290

[10] O. Tsvetkova, T.B.M.J. Ouarda. Use of the Halphen distribution family for mean wind speed estimation with application to Eastern Canada. Energy Conversion and Management. 276 (2023) 116502. 10.1016/j.enconman.2022.116502

[11] 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

[12] M. Laib, M. Kanevski. Spatial Modelling of Extreme Wind Speed Distributions in Switzerland. Energy Procedia. 97 (2016) 100-7. https://doi.org/10.1016/j.egypro.2016.10.029

[13] T.B.M.J. Ouarda, C. Charron. On the mixture of wind speed distribution in a Nordic region. Energy Conversion and Management. 174 (2018) 33-44. https://doi.org/10.1016/j.enconman.2018.08.007

[14] J. Zhou, E. Erdem, G. Li, J. Shi. Comprehensive evaluation of wind speed distribution models: A case study for North Dakota sites. Energy Conversion and Management. 51 (2010) 1449-58. https://doi.org/10.1016/j.enconman.2010.01.020

[15] B. Safari. Modeling wind speed and wind power distributions in Rwanda. Renewable and Sustainable Energy Reviews. 15 (2011) 925-35. https://doi.org/10.1016/j.rser.2010.11.001

[16] N. Aries, S.M. Boudia, H. Ounis. Deep assessment of wind speed distribution models: A case study of four sites in Algeria. Energy Conversion and Management. 155 (2018) 78-90. https://doi.org/10.1016/j.enconman.2017.10.082

[17] O. Alavi, K. Mohammadi, A. Mostafaeipour. Evaluating the suitability of wind speed probability distribution models: A case of study of east and southeast parts of Iran. Energy Conversion and Management. 119 (2016) 101-8. https://doi.org/10.1016/j.enconman.2016.04.039

[18] T.B.M.J. Ouarda, C. Charron, J.Y. Shin, P.R. Marpu, A.H. Al-Mandoos, M.H. Al-Tamimi, et al. Probability distributions of wind speed in the UAE. Energy Conversion and Management. 93 (2015) 414-34. http://doi.org/10.1016/j.enconman.2015.01.036

[19] Q. Han, S. Ma, T. Wang, F. Chu. Kernel density estimation model for wind speed probability distribution with applicability to wind energy assessment in China. Renewable and Sustainable Energy Reviews. 115 (2019) 109387. https://doi.org/10.1016/j.rser.2019.109387

[20] S. Węglarczyk. Kernel density estimation and its application. ITM Web Conf. 23 (2018).

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

[22] M. Hirukawa. Asymmetric Kernel Smoothing: Theory and Applications in Economics and Finance. Springer Nature Singapore, Singapore, 2018.

[23] R.J. Hyndman, Y. Fan. Sample quantiles in statistical packages. The American Statistician. 50 (1996) 361-5.

[24] R.D. Reiss. Approximate Distributions of Order Statistics: With Applications to Nonparametric Statistics. Springer New York1989.

[25] Y. He, R. Liu, H. Li, S. Wang, X. Lu. Short-term power load probability density forecasting method using kernel-based support vector quantile regression and Copula theory. Applied Energy. 185 (2017) 254-66. https://doi.org/10.1016/j.apenergy.2016.10.079

[26] C. Jung, D. Schindler. 3D statistical mapping of Germany’s wind resource using WSWS. Energy Conversion and Management. 159 (2018) 96-108. https://doi.org/10.1016/j.enconman.2017.12.095

[27] 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

[28] J.H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics. 29 (2001) 1189-232.

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

[30] 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. 10.1016/j.enconman.2023.117324

[31] I. Guyon, A. Elisseeff. An introduction to variable and feature selection. Journal of machine learning research. 3 (2003) 1157-82.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 12 (2011) 2825-30.

[33] L. Grinsztajn, E. Oyallon, G. Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? , Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track2022.

[34] T. Hastie, R. Tibshirani, J.H. Friedman, J.H. Friedman. The elements of statistical learning: data mining, inference, and prediction. Springer, New York, 2009.

[35] L. Ye, B. Dai, Z. Li, M. Pei, Y. Zhao, P. Lu. An ensemble method for short-term wind power prediction considering error correction strategy. Applied Energy. 322 (2022) 119475. https://doi.org/10.1016/j.apenergy.2022.119475

[36] Y. Sun, D. Zhu, Y. Li, R. Wang, R. Ma. Spatial modelling the location choice of large-scale solar photovoltaic power plants: Application of interpretable machine learning techniques and the national inventory. Energy Conversion and Management. 289 (2023) 117198. https://doi.org/10.1016/j.enconman.2023.117198

[37] J. Lee, W. Wang, F. Harrou, Y. Sun. Reliable solar irradiance prediction using ensemble learning-based models: A comparative study. Energy Conversion and Management. 208 (2020) 112582. https://doi.org/10.1016/j.enconman.2020.112582

[38] T. Chen, C. Guestrin. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, San Francisco, California, USA, 2016. pp. 785–94.

[39] R. Turner, D. Eriksson, M. McCourt, J. Kiili, E. Laaksonen, Z. Xu, I. Guyon. Bayesian Optimization is Superior to Random Search for Machine Learning Hyperparameter Tuning: Analysis of the Black-Box Optimization Challenge 2020. in: E. Hugo Jair, H. Katja, (Eds.), Proceedings of the NeurIPS 2020 Competition and Demonstration Track. PMLR, Proceedings of Machine Learning Research, 2021. pp. 3--26.

[40] J. Bergstra, Y. Bengio. Random search for hyper-parameter optimization. Journal of machine learning research. 13 (2012).

[41] 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

[42] P. Lafaye de Micheaux, F. Ouimet. A Study of Seven Asymmetric Kernels for the Estimation of Cumulative Distribution Functions. Mathematics2021.

[43] 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. http://dx.doi.org/10.1016/j.enconman.2016.07.012

[44] M.B. Wilk, R. Gnanadesikan. Probability plotting methods for the analysis for the analysis of data. Biometrika. 55 (1968) 1-17. 10.1093/biomet/55.1.1

[45] R. Horst. The Weibull Distribution: A Handbook. Chapman and Hall/CRC, New York, 2008.

[46] F.G. Akgül, B. Şenoğlu, T. Arslan. An alternative distribution to Weibull for modeling the wind speed data: Inverse Weibull distribution. Energy Conversion and Management. 114 (2016) 234-40. https://doi.org/10.1016/j.enconman.2016.02.026

[47] E.C. Morgan, M. Lackner, R.M. Vogel, L.G. Baise. Probability distributions for offshore wind speeds. Energy Conversion and Management. 52 (2011) 15-26. https://doi.org/10.1016/j.enconman.2010.06.015

[48] J.B. Lindsay. The Whitebox Geospatial Analysis Tools project and open-access GIS. GIS Research UK 22nd Annual Conference. The University of Glasgow, University of Glasgow, 2014.

[49] 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

[50] R. Latifovic, D. Pouliot, I. Olthof. Circa 2010 Land Cover of Canada: Local Optimization Methodology and Product Development. Remote Sensing. 9 (2017) 1098.

[51] J. Wiernga. Representative roughness parameters for homogeneous terrain. Boundary-Layer Meteorology. 63 (1993) 323-63. 10.1007/BF00705357

[52] J. Wu, X.-Y. Chen, H. Zhang, L.-D. Xiong, H. Lei, S.-H. Deng. Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimizationb. Journal of Electronic Science and Technology. 17 (2019) 26-40. https://doi.org/10.11989/JEST.1674-862X.80904120

[53] K. Ren, W. Fang, J. Qu, X. Zhang, X. Shi. Comparison of eight filter-based feature selection methods for monthly streamflow forecasting – Three case studies on CAMELS data sets. Journal of Hydrology. 586 (2020) 124897. https://doi.org/10.1016/j.jhydrol.2020.124897

[54] R. Koenker. Quantile Regression: 40 Years On. Annual Review of Economics. 9 (2017) 155-76. 10.1146/annurev-economics-063016-103651

[55] B. Nasri, T. Bouezmarni, A. St-Hilaire, T.B.M.J. Ouarda. Non-stationary hydrologic frequency analysis using B-spline quantile regression. J Hydrol. 554 (2017) 532-44. https://doi.org/10.1016/j.jhydrol.2017.09.035

[56] D. Ouali, F. Chebana, T. 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

[57] N. Meinshausen, G. Ridgeway. Quantile regression forests. Journal of machine learning research. 7 (2006).

[58] Y. Liu, Y. Wu. Simultaneous multiple non-crossing quantile regression estimation using kernel constraints. Journal of Nonparametric Statistics. 23 (2011) 415-37. 10.1080/10485252.2010.537336

[59] A.J. Cannon. Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. Stochastic Environmental Research and Risk Assessment. 32 (2018) 3207-25. 10.1007/s00477-018-1573-6

[60] T. Duan, A. Anand, D.Y. Ding, K.K. Thai, S. Basu, A. Ng, A. Schuler. Ngboost: Natural gradient boosting for probabilistic prediction. International Conference on Machine Learning. PMLR2020. pp. 2690-700.

[61] X.-Y. Tang, S. Zhao, B. Fan, J. Peinke, B. Stoevesandt. Micro-scale wind resource assessment in complex terrain based on CFD coupled measurement from multiple masts. Applied Energy. 238 (2019) 806-15. https://doi.org/10.1016/j.apenergy.2019.01.129

[62] P. Beaucage, M.C. Brower, J. Tensen. Evaluation of four numerical wind flow models for wind resource mapping. Wind Energy. 17 (2014) 197-208. https://doi.org/10.1002/we.1568

[63] R.E. Keck, N. Sondell. Validation of uncertainty reduction by using multiple transfer locations for WRF–CFD coupling in numerical wind energy assessments. Wind Energ Sci. 5 (2020) 997-1005. 10.5194/wes-5-997-2020

[64] T. Simões, A. Estanqueiro. A new methodology for urban wind resource assessment. Renewable Energy. 89 (2016) 598-605. https://doi.org/10.1016/j.renene.2015.12.008

[65] I.V. Florinsky. An illustrated introduction to general geomorphometry. Progress in Physical Geography: Earth and Environment. 41 (2017) 723-52. 10.1177/0309133317733667

[66] J. Jasiewicz, T.F. Stepinski. Geomorphons — a pattern recognition approach to classification and mapping of landforms. Geomorphology. 182 (2013) 147-56. https://doi.org/10.1016/j.geomorph.2012.11.005

[67] J.P. Wilson. Environmental applications of digital terrain modeling. John Wiley & Sons2018.

[68] D.J. Pennock, B.J. Zebarth, E. De Jong. Landform classification and soil distribution in hummocky terrain, Saskatchewan, Canada. Geoderma. 40 (1987) 297-315. https://doi.org/10.1016/0016-7061(87)90040-1

[69] S.J. Riley, S.D. DeGloria, R. Elliot. Index that quantifies topographic heterogeneity. intermountain Journal of sciences. 5 (1999) 23-7.

[70] C.H. Grohmann, M.J. Smith, C. Riccomini. Multiscale Analysis of Topographic Surface Roughness in the Midland Valley, Scotland. IEEE Transactions on Geoscience and Remote Sensing. 49 (2011) 1200-13. 10.1109/TGRS.2010.2053546

[71] J. Jenness. Calculating Landscape Surface Area from Digital Elevation Models. Wildlife Society Bulletin. 32 (2004) 829-39. 10.2193/0091-7648(2004)032[0829:CLSAFD]2.0.CO;2