Centre Eau Terre et Environnement

DEVELOPPEMENT DE METHODES D'APPRENTISSAGE AUTOMATIQUE POUR L'ESTIMATION A LONG TERME DE LA VITESSE DU VENT DANS DES SITES NON ECHANTILLONNES

Par

Freddy Houndekindo

Thèse présentée pour l'obtention du grade de

Philosophiae Doctor (Ph.D.)

en sciences de la Terre

Jury d'évaluation

Président du jury et examinateur interne	Amadou Barry Institut National de la Recherche Scientifique
Examinatrice externe	Sara C. Pryor Department of Earth and Atmoshperic Sciences Cornell University
Examinatrice externe	Xiaoli Guo Larsén DTU Wind Technical University of Denmark
Directeur de recherche	Taha B.M.J. Ouarda Institut National de la Recherche Scientifique

DÉDICACE

En hommage à mon père, Gervais Houndekindo, qui m'a toujours soutenu et qui m'a inculqué des valeurs qui resteront à jamais gravées dans mon cœur.

À ma mère, Aminata, pour son amour inconditionnel et les nombreux sacrifices consentis pour notre famille, je continue d'apprendre beaucoup de toi.

À mon frère bien-aimé, Lionel, qui est parti beaucoup trop tôt.

À ma sœur Suzanne et mon frère Habib, Je suis si reconnaissant envers Dieu de vous avoir dans ma vie.

À toutes mes nièces, pour qui j'ai tant d'affection

REMERCIEMENTS

Tout d'abord, je tiens à exprimer ma profonde gratitude à mon directeur de thèse, Taha B.M.J Ouarda, pour l'occasion qu'il m'a offerte de m'engager dans un nouveau domaine de recherche. Je le remercie sincèrement pour ses précieux conseils qui m'ont guidé tout au long de ma thèse et pour sa bienveillance.

J'exprime ma profonde gratitude aux membres du jury qui ont accepté d'évaluer cette thèse de doctorat.

Je tiens également à remercier chaleureusement tous les membres de l'équipe de recherche de Taha avec qui j'ai eu de nombreuses conversations enrichissantes. Cette aventure n'aurait pas été la même sans l'appui de mes collègues et amis, qui m'ont encouragé et inspiré.

Enfin, je voudrais adresser mes plus sincères remerciements à ma famille, sans qui cette réussite n'aurait pas été possible.

RÉSUMÉ

L'énergie éolienne représente une ressource durable, essentielle à la transition énergétique en cours. Son exploitation efficace repose sur une estimation précise de sa disponibilité à long terme, ce qui nécessite des données fiables sur les vitesses du vent. Lorsque ces données sont manquantes, il est nécessaire de recourir à des méthodes d'estimation basées sur des approches physiques, statistiques, ou hybrides.

Plusieurs méthodes statistiques d'estimation de la vitesse ont été développées au fil des années. Il était donc utile de procéder à une revue exhaustive de ces approches afin d'identifier leurs forces, leurs limites et les axes d'amélioration possibles. Dans notre premier article, une revue de littérature offrant un aperçu complet de ces méthodes a été réalisée. Cette synthèse permet d'établir un cadre de référence pour le développement de nouvelles méthodologies pour l'estimation de la vitesse du vent.

Le deuxième article compare six méthodes de sélection de variables explicatives pour l'interpolation spatiale de plusieurs quantiles de vitesse de vent au Canada. En considérant des quantiles associés à différentes probabilités de dépassement, cette étude permet non seulement d'identifier les variables explicatives les plus pertinentes, mais aussi d'analyser leur influence relative sur les différentes plages de vitesse du vent (faibles, moyennes ou élèves).

Dans le troisième article de cette thèse, nous proposons une nouvelle approche non paramétrique pour estimer la distribution de probabilité des vitesses du vent aux sites non échantillonnés. Cette approche est recommandée dans les régions où la variabilité spatiale des régimes du vent est importante et où une seule famille de distribution peut ne pas être suffisamment flexible pour représenter cette variabilité.

Les vitesses de vents issues des données de réanalyses sont couramment utilisées pour estimer le potentiel éolien aux sites non échantillonnés. Cependant, la résolution spatiale grossière de ces données, les rendent incapables de représenter avec précision les variations locales du relief et leur influence sur les vitesses de vent près du sol. Le quatrième article de cette thèse propose une étude comparative des méthodes de correction statistiques des vitesses du vent issues des données de réanalyse, incluant une nouvelle approche que nous avons développée. Des recommandations ont été formulées quant aux méthodes les plus adaptées selon les contextes d'application et les objectifs spécifiques des études réalisées.

Un défi majeur des méthodes de correction existantes demeure l’amélioration de la variabilité temporelle des vitesses de vents estimées à partir des données de réanalyses. Le cinquième article introduit une nouvelle approche de correction basée sur l’apprentissage profond (*Deep Learning*). Cette méthode permet non seulement de corriger les biais systématiques, mais aussi d’améliorer significativement la variabilité temporelle des séries.

Mots-clés : Apprentissage automatique ; Correction de biais ; Estimation par noyau ; Potentiel éolien ; Réanalyse atmosphérique ; Topographie, Variabilité temporelle.

ABSTRACT

Wind energy represents a sustainable resource essential to the ongoing energy transition. Its efficient exploitation relies on accurately estimating its long-term availability, which requires reliable data on wind speeds. When such data are missing, estimation methods based on physical, statistical, or hybrid approaches become necessary.

Several statistical methods for wind speed estimation have been developed over the years. Therefore, a comprehensive review of these approaches was necessary to identify their strengths, limitations, and potential areas for improvement. In our first paper, a literature review providing an extensive overview of these methods was conducted. This synthesis establishes a reference framework for the development of new methodologies for wind speed estimation.

The second article compares six feature selection methods for interpolating multiple wind speed quantiles spatially across Canada. Considering quantiles associated with different exceedance probabilities allowed the study to identify the most relevant explanatory variables and analyze their relative influence across different wind speed ranges (e.g., low, medium, or high).

In the third article of this thesis, we propose a nonparametric approach to estimating the probability distribution of wind speeds at unsampled sites. This approach is recommended in regions where the spatial variability of wind regimes is significant and where a single distribution family may not be flexible enough to represent this variability.

Wind speeds derived from reanalysis data are commonly used to estimate wind potential at unsampled sites. However, the coarse spatial resolution of these datasets limits their ability to accurately represent local terrain variations and their influence on near-surface wind speeds. The fourth article of this thesis presents a comparative study of statistical correction methods for wind speeds derived from reanalysis data, including a new approach we have developed. Based on the application context and specific objectives of the studies conducted, recommendations are provided regarding the most suitable methods.

A major challenge with existing correction methods remains improving the temporal variability of wind speeds estimated from reanalysis data. The fifth article introduces a new correction approach based on deep learning. This method not only corrects systematic biases but also significantly enhances the temporal variability of the time series.

Keywords: Machine Learning; Bias Correction; Kernel Estimation; Wind Potential; Atmospheric Reanalysis; Topography; Temporal Variability.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

NOTATIONS

Abréviations

CFD	Computational Fluid Dynamics
CNN	Convolutional Neural Network
ECCC	Environnement et Changement climatique Canada
ECMWF	European Centre for Medium-Range Weather Forecasts
ERA5	ECMWF Reanalysis v5
GBT	Gradient boosting trees
GCM	General circulation model
GOSS	Gradient-based One-Side Sampling
GRU	Gated Recurrent Unit
GWA	Global Wind Atlas
JMA	Japan Meteorological Agency
LASSO	Least Absolute Shrinkage and Selection Operator
LES	Large Eddy Simulation
LightGBM	Light Gradient-Boosting Machine
LSTM	Long Short-Term Memory
MERRA-2	Modern-Era Retrospective Analysis for Research and Applications, version 2
MNEMR	Modèle numérique d'élévation de moyenne résolution du Canada
MNT	Modèle numérique de terrain
MRMR	Maximum relevance minimum redundancy
NASA	National Aeronautics and Space Administration
NWA	Numerical Weather Prediction
p. ex.	Par exemple
Q-Q	Méthode quantile-quantile
RANS	Reynolds-Averaged Navier-Stokes
RF	Random forest
RNN	Recurrent neural network
WAsP	Wind Atlas Analysis and Application Program
WRF	Weather Research and Forecasting model
XGBoost	Extreme Gradient Boosting

Symboles

A, B	Paramètres liés à la stabilité de l'atmosphère
AIC	Critère d'information d'Akaike
α	Exposant de cisaillement
$\frac{\partial}{\partial t}$	Dérivée partielle par rapport à t
D_{KS}	Statistique de Kolmogorov-Smirnov
$F(\cdot)$	Fonction de répartition théorique
$\hat{F}(\cdot)$	Estimateur de la fonction de répartition à noyau asymétrique
$F^{-1}(\cdot)$	Fonction quantile
f_c	Paramètre de Coriolis
$f_m(\cdot)$	Prédiction du modèle de gradient boosting à l'itération m
$F_n(\cdot)$	Fonctions de repartition empirique
F_x, F_y, F_z	Composantes de la force de frottement dans chaque direction
G	Vitesse du vent géostrophique
g	Accélération de la pesanteur
$\Gamma(\cdot)$	Fonction gamma
γ	Paramètre de forme d'une distribution de probabilité
h	Paramétre de lissage d'une fonction de noyau
$h_m(\cdot)$	Estimateur du modèle de gradient boosting à l'itération m
$K(\cdot)$	Fonction de répartition à noyau asymétrique
k	Constante de von Kármán (environ 0,4)
K_{BS}	Fonction de répartition à noyau asymétrique de Birnbaum-Saunders
K_W	Fonction de répartition à noyau asymétrique de Weibull
L	Longueur d'Obukhov
$lik(\cdot)$	Fonction de vraisemblance
M	Nombre de paramètres libres d'un modèle
μ	Paramètre de position d'une distribution de probabilité
$\vec{\nabla}$	Nabla
P	Puissance instantanée
p	Pression atmosphérique
$\Phi(\cdot)$	Fonction de répartition de la loi normal centrée réduite
$\propto$	Proportionnel à

$\psi\left(\frac{z}{L}\right)$	Fonction de correction liée à la stabilité atmosphérique selon la théorie de Monin-Obukhov
ρ	Masse volumique de l’air
R_m	Sous-ensemble de l’espace des variables explicatives résultant de la partition successives d’un arbre de décision
RSS	Somme des carrés des résidus
σ	Paramètre d’échelle d’une distribution de probabilité
S_x	Indice d’exposition directionnel
$\boldsymbol{\tau}$	Taux d'apprentissage du modèle gradient boosting trees
$\widehat{\boldsymbol{\theta}}$	Vecteur des paramètres estimés
$\boldsymbol{\theta}$	Paramètres d’une fonction de distribution de probabilité
$\boldsymbol{\theta_i}$	Les paramètres associés à la composante i d'une distribution mixte
θ_m	Paramètres de l’estimateur $h_m(\cdot)$ du modèle de gradient boosting à l’itération m
$\vec{U}$	Vecteur vent
U	Vitesse du vent
u	Composante zonale de la vitesse cartésienne
u_g	Composante zonale du vent géostrophique
u_*	Vitesse de frottement
v	Composante méridionale de la vitesse cartésienne
v_g	Composante méridionale du vent géostrophique
w	Composante verticale de la vitesse cartésienne
w_i	Poids associé à la composante i d'une distribution mixte
z_0	Longueur de rugosité de la surface

1 INTRODUCTION

1.1 Mise en contexte

La disponibilité de source d'énergie renouvelable est primordiale à la réduction de la dépendance aux combustibles fossiles et la diminution des émissions de gaz à effet de serre (Hassan et al., 2024). Parmi elles, l'énergie éolienne se distingue comme une solution « propre » et durable, qui a connu une forte croissante ces dernières années (Global Wind Energy Council, 2024). Cette progression est due à plusieurs facteurs, dont la disponibilité mondiale abondante (Archer et al., 2005; Jung et al., 2022b), des avancées technologiques significatives qui améliorent les performances des turbines, ainsi qu'une réduction continue des coûts de production et d'exploitation (Veers et al., 2019; Wiser et al., 2021). Cette combinaison de facteurs rend cette ressource de plus en plus compétitive par rapport aux sources d'énergie traditionnelles.

La variabilité de la production éolienne est principalement influencée par la vitesse du vent (U), la puissance instantanée générée (P) par une éolienne étant proportionnelle au cube de celle-ci ($P \propto U^3$). Cette relation non linéaire accentue l'impact des fluctuations de la vitesse du vent sur la production énergétique. l'intermittence de la ressource pose certains défis à son intégration à grande échelle dans les réseaux électriques (Ren et al., 2017). Cela met en évidence la nécessité de poursuivre les recherches pour affiner l'évaluation du potentiel éolien et améliorer les modèles prédictifs.

La disponibilité de longues séries de données sur la vitesse du vent (p. ex., 30 ans) avec une haute résolution spatiale et temporelle est nécessaire pour une estimation adéquate du potentiel éolien (Pelser et al., 2024). Or, ces données sont souvent limitées dans le temps et l'espace, requérant le développement de méthodes physiques, statistiques et hybride pour estimer les vitesses du vent aux sites non échantillonnés (Zhang et al., 2015). Un site non échantillonné fait référence à un emplacement géographique pour lequel aucune mesure directe de la vitesse du vent n'est disponible.

Lors de l'évaluation des ressources éoliennes sur un vaste territoire, les méthodes physiques sont moins appropriées en raison de leur coût en temps de calcul élevé (Dupuy et al., 2023; Jung et al., 2023b). En revanche, les modèles empiriques (statistiques et d'apprentissage automatique), qui exploitent les données disponibles pour prédire les conditions du vent dans des zones non échantillonnées, offrent une solution plus pratique et abordable (Dujardin et al., 2022; Veronesi et al., 2016). Elles permettent d'intégrer des données issues de diverses sources, notamment les mesures in situ, les réanalyses, les données topographiques et de couverture du sol.

Les méthodes d’estimation des vitesses du vent aux sites non échantillonnés peuvent être classées selon le type de statistique à prédire : moyenne, valeur extrême, distribution de probabilité ou séries temporelles. Pour l’estimation de la moyenne, des méthodes traditionnelles d’interpolation spatiale, comme le krigeage, sont fréquemment employées (Lee, 2022; Luo et al., 2008). Les valeurs extrêmes de vitesse du vent peuvent être estimées par une analyse fréquentielle régionale (Campos et al., 2018). Elles permettent d’évaluer les risques que les vents violents posent aux infrastructures (Pryor et al., 2021).

Les études récentes accordent une attention accrue à l'estimation de la distribution complète des probabilités ou à la reconstruction de longues séries temporelles de vitesses du vent (Jung et al., 2023a). Ces méthodes permettent une analyse plus approfondie de la variabilité de la ressource, un aspect essentiel pour assurer la viabilité économique des projets éoliens sur le long terme (Millstein et al., 2019).

Les méthodes paramétriques sont souvent employées pour estimer la distribution des probabilités des vitesses du vent dans des zones non échantillonnées (Jung et al., 2020; Veronesi et al., 2016). Elles reposent sur l’hypothèse selon laquelle les vitesses du vent dans l’ensemble d’une région donnée suivent une seule loi de distribution dont les paramètres varient dans l’espace géographique. Toutefois, dans les régions où la variation spatiale du régime des vents est significative, ces méthodes présentent des limites. Il est donc crucial d’adopter des techniques plus flexibles, comme les méthodes non paramétriques, qui ne contraignent pas l’ensemble de la région à une forme particulière de la loi de probabilité.

Les progrès récents dans le développement des modèles de prévision numérique du temps (Numerical Weather Prediction, NWP), des méthodes d’assimilation de données climatiques et l’accroissement des capacités de calcul ont entraîné des améliorations significatives en termes de résolution spatiale, temporelle et de précision des données de réanalyse atmosphérique (Valmassoi et al., 2023).

Par conséquent, l’interpolation spatiale des données de réanalyse, telles que ERA5 (cinquième génération de réanalyse atmosphérique du climat mondial du Centre européen pour les prévisions météorologiques à moyen terme; Hersbach et al. (2020)) et MERRA-2 (Modern-Era Retrospective Analysis for Research and Applications, version 2 ; Gelaro et al. (2017)) demeure une méthode privilégiée pour reconstruire les séries temporelles de vitesse du vent dans les zones non échantillonnés (Olauson, 2018).

Malgré ces progrès, les données brutes de réanalyse restent affectées par des biais systématiques en raison de leur résolution spatiale relativement grossière. Cette limitation entrave la représentation précise de phénomènes locaux, notamment ceux liés aux interactions entre le vent, le relief et d'autres phénomènes météorologiques à petite échelle, associés aux gradients de température (Gualtieri, 2022). Ces biais peuvent engendrer des écarts significatifs entre les vitesses du vent mesurées et celles estimées à partir des données brutes de réanalyse, notamment dans la couche limite atmosphérique, où sont installées les turbines éoliennes. Pour remédier à ces limitations, des méthodes de correction statistique ont été développées pour ajuster les estimations issues de ces modèles physiques aux observations locales (Dujardin et al., 2022; Winstral et al., 2017).

Par ailleurs, des biais dans la variabilité temporelle des données de réanalyse ont également été signalés dans la littérature (Davidson et al., 2022; Ramon et al., 2019). La correction de ce type de biais reste encore peu explorée, mais le développement de telles approches pourrait significativement améliorer la corrélation temporelle à différentes échelles (p. ex., diurne, saisonnière et interannuelle) entre les données de réanalyse corrigées et les vitesses de vent observées. Cela permettrait d'obtenir une évaluation plus précise du potentiel éolien et une meilleure compréhension de sa variabilité temporelle, qui est cruciale pour la planification de la ressource.

1.2 Généralité sur la dynamique du vent

1.2.1 Équations régissant la dynamique du vent

Le gradient de pression est la principale force responsable du mouvement de l'air dans l'atmosphère. L'inégale distribution du rayonnement solaire en fonction de la latitude, de la topographie et de la saison, provoque des variations spatiales de température. Les masses d'air chaudes et légères (moins denses) s'élèvent, ce qui entraîne une baisse de la pression à la surface. À l'inverse, les masses d'air plus froides et moins denses tendent à descendre, augmentant ainsi la pression au sol. Ces différences de pression donnent naissance à un gradient de pression horizontal, ce qui entraîne un déplacement de l'air des zones de haute pression vers celles de basse pression.

Les mouvements d'air causés par la variation de pression sont influencés par d'autres forces, telles que la force de Coriolis, la gravité et les frottements près de la surface terrestre. La force de Coriolis, résultant de la rotation de la Terre, dévie les masses d'air vers la droite dans l'hémisphère Nord et vers la gauche dans l'hémisphère Sud. À proximité du sol, la force de frottement, due à la rugosité de la surface, ralentit la vitesse du vent. De plus, la gravité agit comme une force verticale opposée au soulèvement de l'air.

L'équation de bilan de quantité de mouvement pour l'écoulement d'une masse d'air dans l'atmosphère est donnée par (Andrews, 2010):

Équation 1.1

$$\frac{\partial u}{\partial t} + (\vec{U} \cdot \vec{\nabla})u - f_c v + \frac{1}{\rho}\frac{\partial p}{\partial x} - F_x = 0$$

$$\frac{\partial v}{\partial t} + (\vec{U} \cdot \vec{\nabla})v + f_c u + \frac{1}{\rho}\frac{\partial p}{\partial y} - F_y = 0$$

$$\frac{\partial w}{\partial t} + (\vec{U} \cdot \vec{\nabla})w + g + \frac{1}{\rho}\frac{\partial p}{\partial z} - F_y = 0$$

où u, v et w sont respectivement les composantes zonale (est-ouest), méridionale (nord-sud) et verticale du vecteur de vitesse $\vec{U}$, f_c est le paramètre de Coriolis, ρ est la densité de l'air, p est la pression atmosphérique, g est l'accélération gravitationnelle, F_x, F_y, F_z sont les composantes de la force de frottement dans chaque direction, et $\vec{\nabla}$ est le vecteur nabla.

En fonction des hypothèses physiques et des échelles spatiales et temporelles retenues, des simplifications peuvent être apportées aux équations de quantité de mouvement. Ces simplifications permettent d'isoler les mécanismes dominants à une échelle donnée, ce qui facilite l'analyse du phénomène.

Dans l'atmosphère libre, où l'influence des frottements devient négligeable, le vent tend vers un état d'équilibre entre la force du gradient de pression et la force de Coriolis. Cet équilibre, nommé équilibre géostrophique, est décrit par les équations suivantes :

Équation 1.2

$$u_g = -\frac{1}{\rho f_c}\frac{\partial p}{\partial y}$$

$$v_g = \frac{1}{\rho f_c}\frac{\partial p}{\partial x}$$

où u_g et $v_g$sont les composantes zonale et méridionale du vent géostrophique.

Cette relation permet d'estimer la vitesse du vent dans l'atmosphère libre à partir des champs de pression atmosphérique. Cependant, cet équilibre ne s'applique pas dans les couches basses de l'atmosphère, notamment la couche limite, où les effets de friction perturbent significativement l'écoulement.

1.2.2 Effet de la topographie et de la rugosité du sol

À proximité de la surface terrestre, les vents résultant de l'équilibre géostrophique constituent le forçage de grande échelle. La friction induite par la rugosité de surface (végétation, relief, bâtiments, etc.) modifie les vents géostrophiques en réduisant leur vitesse.

En terrain plat, homogène et sans obstacle à proximité, la vitesse du vent près du sol, $U(z)$, peut être estimée à partir de la vitesse géostrophique G, de la vitesse de friction et du profil logarithmique de la vitesse du vent. À partir du gradient de pression, on peut calculer la vitesse du vent géostrophique G, puis en déduire la vitesse de friction u_* grâce à l'Équation 1.3 (Zilitinkevich *et al.*, 1974). Une fois que u_* est déterminé, le profil logarithmique de la vitesse du vent (Équation 1.4) peut être utilisé pour estimer la vitesse moyenne du vent à une hauteur z donnée :

Équation 1.3

$$G = \frac{u_*}{k}\sqrt{\left(\ln\left(\frac{u_*}{f_c z_0}\right) - A\right)^2 + B^2}$$

Équation 1.4

$$U(z) = \frac{u_*}{k}\left[\ln\left(\frac{z}{z_0}\right) - \psi\left(\frac{z}{L}\right)\right]$$

Où, u_* représente la vitesse de frottement (qui mesure le degré de cisaillement du vent près du sol dû à la rugosité), k correspond à la constante de von Kármán (environ 0,4), f_c désigne le paramètre de Coriolis, z_0 correspond à la longueur de rugosité de la surface (hauteur moyenne du sol où la vitesse moyenne du vent est nulle), A et B sont des paramètres liés à la stabilité de l'atmosphère, $\psi\left(\frac{z}{L}\right)$ est une fonction de correction liée à la stabilité atmosphérique selon la théorie de Monin-Obukhov, et L désigne la longueur d'Obukhov, qui caractérise les conditions de stabilité thermique dans la couche limite.

Dans les régions dont la topographie est irrégulière, le flux de l'air devient plus variable et il ne peut plus être décrit de manière adéquate par la méthode analytique ci-dessus. En effet, les variations du relief, les changements d'occupation du sol, ainsi que la présence d'obstacles, tels que les bâtiments, ou les forêts, influencent le mouvement de l'air près du sol (Petersen *et al.*, 1998).

L'occupation du sol influence la dynamique du vent, notamment parce qu'elle entraine une friction avec les éléments de surface, comme la végétation ou les bâtiments. Cette friction se traduit par un ralentissement du vent, ainsi qu'une modification de la structure verticale du profil de vitesse caractérisé par le profil logarithmique du vent (Équation 1.4) ou la loi exponentielle de cisaillement du vent :

Équation 1.5

$$U(z_2) = U(z_1)\left(\frac{z_2}{z_1}\right)^{\alpha}$$

Où $U(z_2)$ est la vitesse du vent à la hauteur z_2, $U(z_1)$ est la vitesse mesurée à une hauteur de référence z_1 et α est l'exposant de cisaillement, qui dépend de la rugosité de surface et de la stabilité atmosphérique.

La rugosité de surface est mesurée par la longueur de rugosité (z_0), qui est proportionnelle à la hauteur des obstacles présents à la surface terrestre, comme la végétation et les bâtiments. Des valeurs élevées de longueur de rugosité indiquent généralement des surfaces entraînant un ralentissement plus important des vitesses du vent. Les milieux urbains et les zones végétalisées, comme les forêts, présentent généralement des longueurs de rugosité plus élevées que les surfaces couvertes de neige ou les terrains nus et peu accidentés. Le Tableau 1.1 présente des valeurs de la longueur de rugosité z_0 et de l'exposant de cisaillement α pour une stabilité atmosphérique neutre et pour différents types de surface.

Tableau 1.1 Valeur de la longueur de rugosité z_0 et de l'exposant de cisaillement α pour différents types de surface

Type de surface	Longueur de rugosité z_0	Exposant de cisaillement α
Eaux	0,001	0,11
Herbes	0,01-0,05	0,16
Arbustes	0,1-0,2	0,20
Forêts	0,5	0,28
Villes	1-2	0,4

Les valeurs de α sont données pour une stabilité atmosphérique neutre. Adapté de Emeis (2018).

L'interaction entre le relief et le vent peut être de nature mécanique ou thermique (régime de brise) (Whiteman, 2000). L'effet mécanique résulte de la modification du flux de vent synoptique par les caractéristiques orographiques locales.

Lorsqu'un flux d'air rencontre un obstacle orographique tel qu'une colline ou une chaîne de montagnes, plusieurs scénarios peuvent se produire selon les caractéristiques du relief (notamment sa hauteur, sa largeur et sa pente), la stabilité de l'atmosphère et la vitesse du vent (Jackson *et al.*, 2013). La masse d'air peut s'élever et franchir l'obstacle, être déviée latéralement, canalisée à travers des vallées ou des passes, ou encore bloquée en amont. Une atmosphère stable tend à limiter les mouvements verticaux, ce qui favorise le contournement horizontal des reliefs. À l'inverse, une atmosphère instable facilite les mouvements ascendants, permettant à l'écoulement de franchir verticalement les obstacles. Cependant, même en conditions stables, une masse d'air dotée d'une forte inertie (vitesse suffisante) peut être capable de franchir certains reliefs.

Les régimes de brises sont principalement induits par les contrastes de chauffage de surface générés par le relief, qui donnent lieu à des circulations thermiques locales, telles que les brises de pente et de vallée (Zardi *et al.*, 2013). Ces régimes sont caractérisés par une inversion de la direction du vent au cours du cycle diurne.

Pendant la journée, le rayonnement solaire réchauffe la surface du sol, entrainant un transfert d'énergie vers l'air situé au-dessus du sol sous forme de flux de chaleur sensible. Ce réchauffement favorise les mouvements ascendants du vent le long des pentes ou dans les vallées. À l'inverse, pendant la nuit, le sol se refroidit rapidement, ce qui refroidit l'air en contact avec sa surface, ce qui entraine des écoulements d'air descendants.

Les régimes de brises, qui découlent de ces contrastes thermiques, peuvent se produire à différentes échelles spatiales. Il peut s'agir de phénomènes locaux, comme ceux qui se forment autour d'une colline isolée, ou bien de circulations régionales impliquant de grands systèmes montagneux (Whiteman, 2000).

1.2.3 Simulation de l'écoulement atmosphérique

La variation locale de la vitesse du vent est le fruit de l'interaction de divers processus opérants à différentes échelles. À l'échelle planétaire, les systèmes de pression synoptiques dominent et déterminent les régimes de vent dominants. À l'échelle régionale, des gradients de température dus aux caractéristiques orographiques ou à la proximité de masses d'eau (océans, lacs) peuvent générer des circulations thermiques, telles que les brises de mer, de terre, de vallée ou de montagne. À l'échelle locale, la rugosité du sol, la présence d'obstacles naturels ou artificiels exercent une influence prépondérante sur l'écoulement du vent. L'ensemble de ces processus contribuent à une grande variabilité temporelle (avec des fluctuations allant de la minute à plusieurs jours) et spatiale du vent.

En général, les équations qui régissent la dynamique du vent n'ont pas de solution analytique et sont généralement résolues numériquement à l'aide de modèles de prévision ou de simulation climatique. Ces derniers reposent sur des méthodes de discrétisation spatiale et temporelle.

En fonction du phénomène étudié, différentes approches de modélisation numérique sont utilisées pour simuler le vent. Les modèles de circulation générale (GCMs) décrivent la dynamique atmosphérique à l'échelle globale, mais leur résolution spatiale relativement grossière (supérieure à 100 km) les rend inadéquats pour la simulation de phénomènes régionaux ou locaux (Jung *et al.*, 2022c)

Les modèles à méso-échelle, tels que le "Weather Research and Forecasting model" (WRF ; Powers *et al.* (2017)), sont largement utilisés avec l'assimilation des observations météorologiques pour produire des prévisions atmosphériques à haute résolution. Grâce à une résolution spatiale plus fine (de l'ordre de 10 à 100 km), ces modèles permettent de mieux simuler les circulations régionales induites par la topographie et les gradients thermiques. Ils intègrent aussi des schémas de paramétrisation représentant les processus non résolus à l'échelle de la grille.

Cependant, ces modèles ne résolvent pas adéquatement les turbulences de l'écoulement de petite échelle générées par les variations topographiques et la présence d'obstacles (Zajaczkowski *et al.*, 2011). Pour pallier ces limitations, des approches issues de la mécanique des fluides assistée par ordinateur (Computational Fluid Dynamics, CFD) sont utilisées. Elles permettent de représenter explicitement les turbulences à micro-échelle (Castorrini *et al.*, 2021). L'approche de Reynolds (*Reynolds-Averaged Navier-Stokes*, RANS) repose sur une modélisation statistique de la turbulence en séparant les variables en une valeur moyenne et les fluctuations autour de cette moyenne. Pour mieux caractériser les vitesses instantanées, on peut recourir à la simulation directe des grandes échelles (*Large Eddy Simulation*, LES). Le LES permet de simuler directement la majorité des turbulences et de modéliser des structures dont l'échelle est inférieure à la taille des mailles (Mehta *et al.*, 2014).

Toutefois, en raison de leur coût computationnel particulièrement élevé, l'utilisation de ces modèles reste généralement limitée à des domaines spatiaux restreints et à des fenêtres temporelles réduites. Ils sont donc principalement utilisés durant les phases avancées des projets, par exemple pour des analyses approfondies des sillages turbulents et des zones de séparations de l'écoulement (Thé *et al.*, 2017).

En alternative aux approches basée sur la mécanique des fluides assistée par ordinateur, il existe des modèles plus simples et moins exigeants en ressources de calcul, tels que WAsP (Wind Atlas Analysis and Application Program). Ces outils permettent de représenter certains effets de petite échelle à partir d'une formulation linéaire des équations de l'écoulement (Jackson *et al.*, 1975). Leur rapidité de calcul en fait des outils bien adaptés aux premières phases de l'évaluation du potentiel éolien sur un vaste territoire (Dörenkämper *et al.*, 2020). Néanmoins, ces modèles linéaires présentent des limites notables en terrain complexe, notamment lorsque les pentes sont abruptes (Ayotte *et al.*, 2004).

1.3 Données de réanalyses

Les données de réanalyse, largement utilisées pour l'évaluation du potentiel éolien à l'échelle régionale, illustrent une application concrète des modèles à méso-échelle. Elles permettent de reconstituer de manière cohérente et continue l'état de l'atmosphère sur plusieurs décennies.

Les données de réanalyses, telles que ERA5 et MERRA-2, sont produites à partir de modèles de prévision numérique du temps couplés à des systèmes d'assimilation d'observations météorologiques de diverses sources. Ces jeux de données présentent plusieurs avantages : ils sont disponibles sur de longues périodes, leur couverture spatiale est (quasiment) globale et leur fréquence temporelle est suffisante pour de nombreuses applications. Ils constituent donc une source importante de données pour l'analyse climatique (Gutiérrez et al., 2024), l'évaluation du potentiel éolien à grande échelle (Gualtieri, 2022; Olauson, 2018), et le forçage de modèles à méso-échelle (Lorenz et al., 2016). Le Tableau 1.2 présente une comparaison des principales caractéristiques de trois jeux de données de réanalyse couramment utilisés dans les études sur la ressource éolienne.

Tableau 1.2 Aperçu des données de réanalyse utilisées dans les études de potentiel éolien à grande échelle

Nom complet	Abréviation	Institution	Résolution spatiale	Résolution temporelle (heure)	Période disponible
European Centre for Medium-Range Weather Forecasts Reanalysis 5	ERA5	European Centre for Medium-Range Weather Forecasts (ECMWF)	0,25° × 0,25°	1	1940 à présent

Modern Era Retrospective-Analysis for Research and Applications version 2	MERRA-2	National Aeronautics and Space Administration (NASA)	0,5° latitude × 0,625° longitude	1	1980 à présent
Japanese 55-year Reanalysis	JRA-55	Japan Meteorological Agency (JMA)	1,25° × 1,25°	6	1978 à présent

De nombreuses études ont été consacrées à l'évaluation des performances des jeux de données de réanalyse en ce qui a trait à l'estimation du potentiel éolien (Gualtieri, 2021; Jourdier, 2020; Molina *et al.*, 2021; Olauson, 2018). Ces travaux s'appuient généralement sur la comparaison des vitesses du vent issues des réanalyses avec des observations in situ. Les résultats de ces comparaisons montrent que, même si les réanalyses permettent de caractériser les régimes de vent à l'échelle régionale, elles présentent souvent des biais au niveau local, en particulier dans les régions à topographie complexe ou côtière (Gualtieri, 2022). Ces biais peuvent entraîner une surestimation ou une sous-estimation significative de la vitesse du vent et, par le fait même, du potentiel éolien.

La Figure 1.1 illustre des séries temporelles de vitesses du vent interpolées à partir de données ERA5 pour deux stations de mesure d'Environnement et Changement climatique Canada (ECCC). Cette figure met en évidence l'influence de l'environnement sur la précision des données ERA5. La station numéro 504K0NM, située au bord du lac Manitoba, se trouve dans une région où la topographie est relativement simple, avec peu de variation brusque du relief. En revanche, la station numéro 7041166, située en bordure de la rivière Saguenay, est implantée dans une zone caractérisée par une variabilité du relief plus marquée. Les différences de corrélation observées entre les séries temporelles mesurées et ERA5 des deux stations montrent que la variation du relief influence la précision des vitesses de vent issues des données de réanalyse. Dans les régions à relief accidenté, les différences entre les données ERA5 et les mesures directes sont habituellement plus marquées (Gualtieri, 2021).

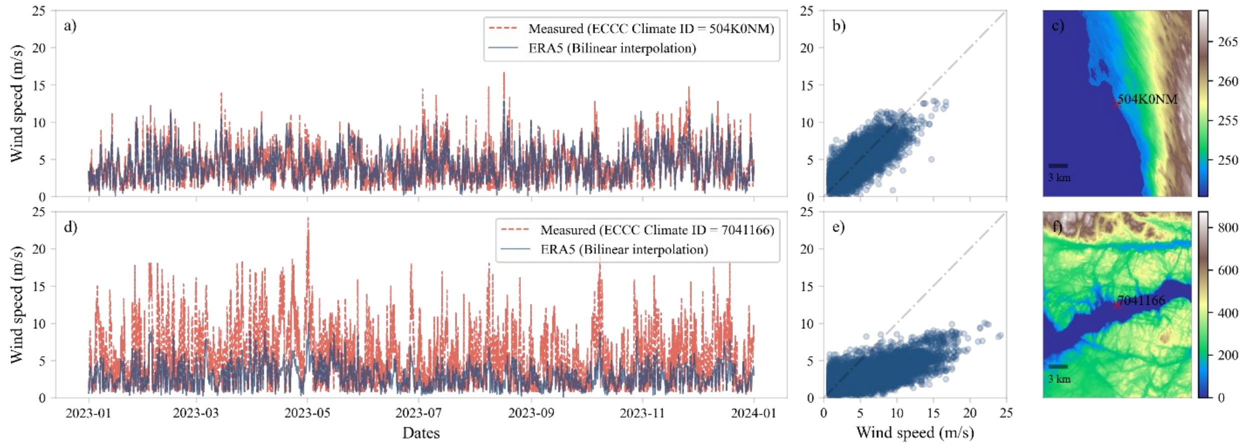


Figure 1.1 Vitesses du vent mesurées et issues d'ERA5

Séries temporelles des vitesses du vent horaires de l'année 2023 à la station numéro 504K0NM (panneau a) et à la station numéro 7041166 (panneau b). Les données mesurées sont représentées en rouge, tandis que les données ERA5 interpolées sont en bleu. Les nuages de points illustrant la relation entre les données mesurées et celles d'ERA5 aux deux stations sont présentés dans les panneaux d (504K0NM) et e (7041166). Enfin, les données d'élévation autour des stations sont présentées dans les panneaux c et f.

La résolution spatiale des données de réanalyse, relativement grossière, constitue un facteur limitant à leur utilisation directe pour évaluer le potentiel éolien, notamment dans les régions à terrain complexe et où les régimes de vent peuvent varier considérablement. La correction des vitesses du vent issues de réanalyses améliore la représentation des phénomènes locaux à petite échelle.

La correction du biais des vitesses du vent issues des réanalyses repose principalement sur l'Atlas mondial des ressources éoliennes (Global Wind Atlas, GWA), qui fournit des conditions de vent statiques (p. ex., moyennes et paramètres de distribution) à l'échelle quasi globale. Ces données intègrent les caractéristiques du terrain et de la couverture du sol à une résolution spatiale de 250 mètres carrés. Dans la littérature, on a utilisé deux principales approches pour corriger à partir du GWA les biais des vitesses du vent issues de réanalyse. La première est une méthode de type quantile-quantile (Q-Q; Cannon et al. (2015)) qui permet d'ajuster les quantiles de la distribution des vitesses de vent issues de réanalyse pour qu'ils correspondent à ceux des données de référence fournies par le GWA (González-Aparicio et al., 2017). La deuxième méthode de correction ajuste la moyenne des vitesses du vent de réanalyse pour la faire correspondre à celle issue du GWA (Gruber et al., 2019).

La Figure 1.2 présente un exemple d'application des deux méthodes de correction aux données d'ERA5 interpolées à la station numéro 7041166 d'ECCC. Le Tableau 1.3 présente les vitesses moyennes et la corrélation avec les données mesurées avant et après les corrections. Les vitesses du vent brutes d'ERA5 sous-estiment systématiquement les valeurs mesurées. Cette sous-estimation est partiellement corrigée par les méthodes de correction des biais, avec des résultats variables. La méthode Q-Q a tendance à surestimer les vents forts, tandis que la correction de la moyenne entraine une sous-estimation des vents forts. Il est important de souligner que ces observations ne peuvent pas être généralisées, car les performances des méthodes de correction dépendent fortement des caractéristiques locales du vent et des biais spécifiques des données de réanalyse.

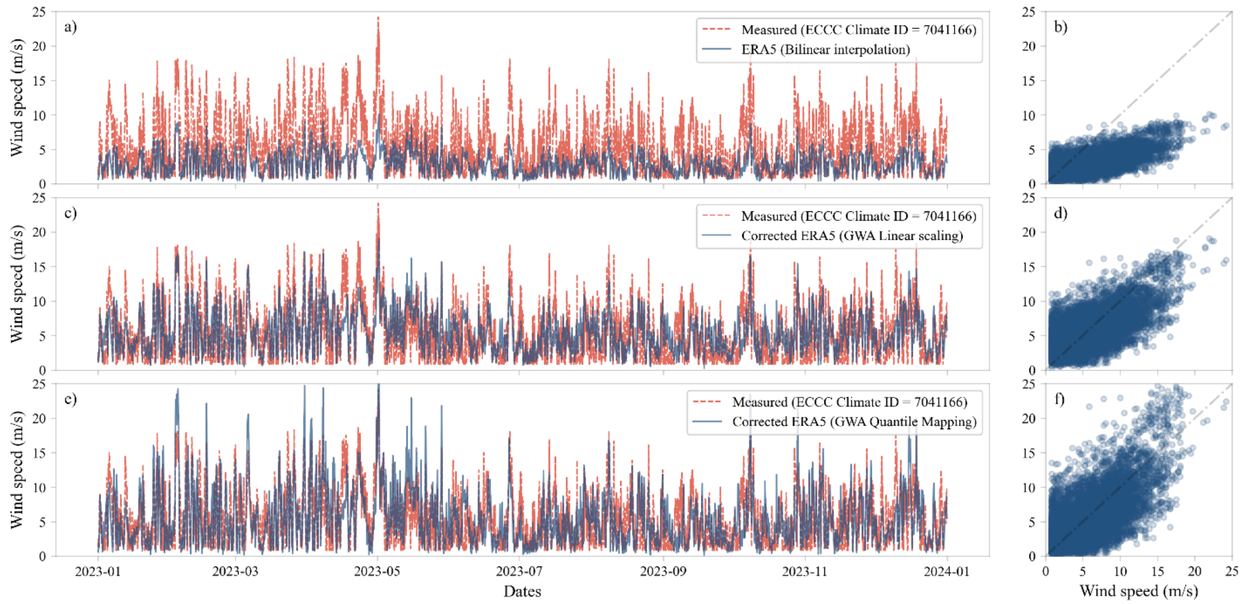


Figure 1.2 Correction des vitesses du vents d'ERA5 à partir du GWA

Application des méthodes de correction de biais des vitesses du vents horaires de 2023 d'ERA5 à la station numéro 7041166 d'ECCC. Les panneaux (a) et (b) montrent la comparaison des vitesses du vent mesurées à la station numéro 7041166 avec les données d'ERA5 interpolées. Les panneaux (c) et (d) illustrent l'effet de la correction de la moyenne, tandis que les panneaux (e) et (f) montrent les résultats de la correction par la méthode quantile-quantile.

Tableau 1.3 Correction des vitesses du vents d'ERA5 à partir du GWA

Données	Vitesses moyennes	Coefficient de corrélation de Pearson
Données mesurées	5,74	1,00
Données brutes d'ERA5	2,98	0,67
ERA5 + GWA (moyenne)	5,63	0,67
ERA5 + GWA (Q-Q)	5,65	0.68

Vitesses moyennes du vent à la station numéro 7041166 (ECCC) et corrélation entre les données ERA5 et celles mesurées, avant et après les corrections à partir du GWA.

Les données du GWA sont particulièrement pertinentes pour corriger les biais des données de réanalyse, en raison de leur intégration de caractéristiques topographiques et de couverture terrestre à une résolution fine (Bosch et al., 2017). Ces caractéristiques permettent de capturer les variations locales des régimes de vent, souvent mal représentées dans les données de réanalyse en raison de leur résolution plus grossière. Ces caractéristiques peuvent être estimées directement à partir d'un modèle numérique de terrain (MNT) et de cartes d'occupation du sol de haute résolution. Elles peuvent ensuite être utilisées comme variables explicatives pour ajuster un modèle de régression entre les vitesses du vent issues de réanalyse et celles mesurées directement. Ce modèle ajusté devrait permettre de corriger les données de réanalyse aux sites non échantillonnés. Cette méthode a l'avantage d'être plus flexible en matière de résolution spatiale du produit final par rapport à la résolution fixe imposée par l'utilisation du GWA.

Par exemple, une méthode développée par Jung et al. (2023b) repose sur l'application d'une correction Q-Q aux données ERA5. Premièrement, les auteurs ont ajusté un modèle de régression prenant en compte les caractéristiques topographiques locales, la longueur de rugosité et les L-moments (Hosking, 1990) des vitesses du vent issues d'ERA5, pour estimer la distribution des vitesses du vent aux sites non échantillonnés. Ce modèle permet de prédire les L-moments corrigés des vitesses du vent aux sites non échantillonnés à partir desquels les paramètres des distributions statistiques sont estimés. Une fois la distribution estimée au site non échantillonné, les séries temporelles des vitesses du vent d'ERA5 sont corrigées à l'aide de la relation :

Équation 1.6

$$\widehat{U}_t = F_{\widehat{\boldsymbol{\theta}}}^{-1}\left[F_n\left(U_t^{ERA5}\right)\right]$$

où $F_{\widehat{\boldsymbol{\theta}}}^{-1}$ est la fonction quantile de la distribution estimée, $\widehat{\boldsymbol{\theta}}$ est le vecteur des paramètres estimés au site non échantillonné par la méthode des L-moments et $F_n\left(U_t^{ERA5}\right)$ est la fonction de répartition empirique des données ERA5.

1.4 Données in situ de vitesse du vent

La disponibilité de longues séries de mesures de vitesses du vent obtenues à partir de stations météorologiques ou de tours instrumentées est essentielle pour évaluer rigoureusement la qualité des données simulées, telles que celles issues des jeux de réanalyse. En effet, ces mesures in situ constituent la référence la plus fiable pour caractériser les conditions locales du vent à un emplacement donné.

De plus, les données in situ remplissent deux fonctions essentielles dans la modélisation empirique (statistique et apprentissage automatique) de la vitesse du vent. Premièrement, elles permettent d'ajuster les paramètres des modèles pour mieux représenter les conditions observées. Deuxièmement, elles servent de référence pour évaluer la performance des données modélisées.

Cependant, même si les données in situ sont très utiles, elles ont aussi des limites. En effet, leur représentativité spatiale est restreinte, puisque chaque station ou tour instrumentée ne fournit des informations qu'à un endroit très spécifique. De plus, leur disponibilité est souvent restreinte par les coûts élevés d'installation, l'entretien et la calibration régulière des instruments de mesure. La qualité des données peut aussi être altérée par divers facteurs, notamment le changement d'instrumentation, les effets de site non corrigés, ou les discontinuités dans les séries temporelles. Par ailleurs, la plupart des mesures disponibles sont prises à des hauteurs standards (souvent 10 mètres), ce qui nécessite l'application de modèles de transfert vertical pour les adapter aux hauteurs caractéristiques des éoliennes modernes (100 m et plus).

Dans le cadre de cette thèse, les données in situ de vitesse du vent analysé sont issues de la base de données climatiques historiques d'ECCC. Cette base de données est soumise à des procédures de contrôle de qualité qui permettent de détecter et de corriger les éventuelles erreurs ou incohérences.

Les données de vitesse du vent de cette base sont généralement collectées à une hauteur standard de 10 mètres au-dessus du sol. Elles représentent des moyennes de vitesse calculées sur une période d'une à deux minutes, qui se terminent à l'heure d'observation. Pour minimiser les effets de perturbation, les stations de mesure sont implantées sur des sites plats, dégagés, et éloignés d'obstacles, comme les arbres, les bâtiments ou les collines (Environnement et Changement climatique Canada, 2025).

1.5 Distributions de probabilités des vitesses du vent

On compte plusieurs types de lois de distribution paramétriques dans la littérature pour représenter la distribution de la vitesse du vent (Jung et al., 2019b). Ces distributions jouent un rôle clé dans la modélisation statistique du vent, car elles permettent d'en caractériser la variabilité.

Lorsque les composantes orthogonales (composante nord-sud et est-ouest) de la vitesse du vent remplissent certaines conditions, telles que l'indépendance statistique et suivent une distribution normale bivariée centrée, la distribution de Rayleigh émerge comme la candidate idéale pour la modélisation de la distribution du vent (Hennessey, 1978). En réalité, cette distribution à un seul paramètre manque de flexibilité et les circonstances qui la rendent appropriée ne sont pas souvent remplis, principalement en raison de l'influence de la topographie sur le vent (Tuller et al., 1985). La fonction de densité de probabilité associée à la distribution de Rayleigh s'écrit :

Équation 1.7

$$p(U;\sigma) = \frac{U}{\sigma^2}\exp\left[-\frac{1}{2}\left(\frac{U}{\sigma}\right)^2\right]$$

Où U est la vitesse du vent et σ est un paramètre d'échelle.

La distribution de Weibull à deux paramètres, une généralisation de la Rayleigh, dispose également de fondements théoriques pour son application (Tuller et al., 1984). Grâce à ses deux paramètres, elle est plus flexible et est devenue la distribution la plus populaire pour les vitesses du vent (Carta et al., 2009). La fonction de densité de probabilité associée à la distribution de Weibull à deux paramètres est donnée par :

Équation 1.8

$$p(U;\sigma,\gamma) = \frac{\gamma}{\sigma}\left(\frac{U}{\sigma}\right)^{\gamma-1}\exp\left[-\left(\frac{U}{\sigma}\right)^{\gamma}\right]$$

Où σ est le paramètre d'échelle et γ est le paramètre de forme.

Dans les régimes de vent calme où la probabilité de valeurs nulles est significative, la distribution de Weibull à deux paramètres présente certaines limites (Wais, 2017). La distribution de Weibull à trois paramètres permet de pallier ces limites en intégrant un paramètre additionnel correspondant au seuil minimal. Ce paramètre permet de considérer les valeurs nulles ou très faibles. La fonction de densité de probabilité associée à cette distribution est exprimée comme suit :

Équation 1.9

$$p(U;\sigma,\gamma,\mu) = \frac{\gamma}{\sigma}\left(\frac{U-\mu}{\sigma}\right)^{\gamma-1}\exp\left[-\left(\frac{U-\mu}{\sigma}\right)^{\gamma}\right]$$

Où σ est le paramètre d'échelle, γ est le paramètre de forme et μ est le paramètre de position.

Plusieurs études recommandent l'adoption d'approches axées sur les données pour la sélection de la distribution adéquate (Dong et al., 2022; Lins et al., 2023). Ces méthodes s'appuient sur des techniques statistiques d'évaluation de l'ajustement, telles que le test de Kolmogorov-Smirnov et le critère d'information d'Akaike (AIC). Dans le cas univarié, le test de Kolmogorov-

Smirnov permet de mesurer la plus grande différence absolue entre les fonctions de répartition des données observées $F_n(U)$ et la fonction de répartition théorique $F(U)$ associée à la distribution testée (par exemple, la loi de Weibull). La statistique du test Kolmogorov-Smirnov est définie par :

Équation 1.10

$$D_{KS} = \sup_U |F_n(U) - F(U)|$$

où D_{KS} est la statistique de Kolmogorov-Smirnov et $\sup$ représente le supremum (la borne supérieure) de l'ensemble des différences absolues.

Le critère AIC permet de comparer les ajustements des distributions candidates en tenant compte de leurs parcimonies, le nombre de paramètres de la distribution:

Équation 1.11

$$AIC = -2 \log lik(U; \boldsymbol{\theta}) + 2M$$

Où $lik(U; \theta) = \prod_{i=1}^{n} p(U_i; \theta)$ est la vraisemblance du modèle évaluée sur les données observées U avec les paramètres θ et M représente le nombre de paramètres libres du modèle.

Plusieurs types de distributions offrant une flexibilité accrue ont été explorées pour modéliser les vitesses du vent (Shi et al., 2021; Tsvetkova et al., 2023). Cependant, dans les contextes où le régime des vent est hétérogène et présente un comportement bimodal ou multimodal, les distributions unimodales, comme la Weibull, montrent leurs limites (Ouarda et al., 2018). Les distributions mixtes, qui sont des combinaisons linéaires convexes de plusieurs distributions unimodales, permettent une meilleure représentation de la variabilité des vitesses du vent en tenant compte de la présence de plusieurs sous-populations ou régimes distincts. La fonction de densité de probabilité d'une distribution mixte à K composantes peut s'écrire comme suit :

Équation 1.12

$$p(U) = \sum_{i=1}^{K} w_i p_i(U; \boldsymbol{\theta}_i)$$

Où $p_i(U; \theta_i)$ est la densité de probabilité de la composante i, θ_i représente les paramètres de cette composante et w_i est le poids associé à la composante i, avec $w_i \geq 0$ et $\sum_{i=1}^{K} w_i = 1$.

Jung et al. (2017b) ont évalué l'ajustement de 24 distributions unimodales et 21 distributions mixtes sur des données de vitesse du vent à l'échelle mondiale. Les résultats de l'étude ont révélé que les distributions mixtes offrent un meilleur ajustement dans la partie centrale de la distribution que les distributions unimodales. Pour un ajustement optimal dans les queues de distribution, les auteurs recommandent la distribution de Wakeby, qui s'est avérée particulièrement efficace pour modéliser les valeurs extrêmes des vitesses du vent.

En complément des distributions paramétriques, des méthodes non paramétriques, comme l'estimation par noyau, sont également utilisées pour capturer des formes de distributions atypiques (Qin et al., 2011).

Dans le cas des vitesses du vent, des fonctions de noyau symétriques, comme le noyau gaussien, ont souvent été utilisées (Han et al., 2019; Zhang et al., 2019). Toutefois, ce type de noyau est connu pour entraîner un biais aux limites lorsque le support de la distribution est défini sur un ensemble borné $[0, \infty]$, comme c'est le cas pour les vitesses du vent. Pour résoudre ce problème, des noyaux asymétriques ont été proposés (Chen, 2000). Récemment, Mombeni et al. (2021) ont proposé le noyau asymétrique Weibull et Birnbaum-Saunders pour l'estimation de la fonction de répartition et Lafaye de Micheaux et al. (2021) ont proposé le noyau Log-Normal.

La formule générale de l'estimateur de la fonction de répartition à noyau asymétrique s'écrit comme suit :

Équation 1.13

$$\hat{F}(U) = \frac{1}{n}\sum_{i=1}^{n} K(U_i; U, h)$$

Où n est le nombre d'observations, $K(U_i; U, h)$ est le noyau asymétrique évalué aux observation U_i et h est le paramétre de lissage.

Le noyau asymétrique de Weibull est donné par :

Équation 1.14

$$K_W\ (U_i; \sigma\ , \gamma\ ) = \ \exp\left(-\left(\frac{U_i}{\sigma}\right)^{\gamma}\right)$$

Où $\sigma = U\Gamma(1 + h)$, $\Gamma(\cdot)$ est la fonction gamma et $\gamma = 1/h$.

Le noyau asymétrique de Birnbaum-Saunders est donné par :

Équation 1.15

$$K_{BS}\,(U_i; U, \sigma) = 1 - \Phi\left(\frac{1}{\sigma}\left(\sqrt{\frac{U_i}{U}} - \sqrt{\frac{U}{U_i}}\right)\right)$$

Où $\boldsymbol{\sigma} = \sqrt{\boldsymbol{h}}$ et $\boldsymbol{\Phi}(\cdot)$ représente la fonction de répartition de la loi normal centrée réduite

Le noyau asymétrique de Log-Normal est donné par :

Équation 1.16

$$K_{LN}\,(U_i; \mu\,, \sigma) = 1 - \Phi\left(\frac{\log U_i - \mu}{\sigma}\right)$$

Où $\boldsymbol{\mu} = \log \boldsymbol{U}$ et $\boldsymbol{\sigma} = \sqrt{\boldsymbol{h}}$.

1.6 Variables explicatives liées à la vitesse du vent

1.6.1 Variables topographiques

L'intégration des variables topographiques dans la modélisation empirique des vitesses du vent permet de mieux capter l'influence des éléments de la surface du sol sur la variabilité spatiale du vent. Des facteurs, tels que l'altitude, la pente, l’orientation du relief, la rugosité de surface ou encore la forme du terrain, jouent un rôle déterminant dans la modification des flux d’air, en favorisant par exemple leur canalisation, leur accélération ou leur atténuation.

Dans les régions où le relief est peu accidenté, l’interpolation spatiale permet généralement d’estimer avec une précision acceptable les vitesses du vent à des sites non échantillonnés (Fick et al., 2017). Toutefois, dans les zones où le relief est marqué, l’influence de la topographie sur les vitesses de vent devient significative et ne peut pas être négligée (González-Longatt et al., 2015; Petersen et al., 1998; Raupach et al., 1997).

Par exemple, l’intégration de l’altitude comme variable explicative dans les approches d’interpolation spatiale a permis d’améliorer significativement la précision des estimations par rapport aux méthodes basées uniquement sur les distances entre les sites (Collados-Lara et al., 2022; Lee, 2022; Palomino et al., 1995). Cependant, l’altitude seule ne suffit pas à capturer de manière exhaustive les interactions complexes entre le vent et le relief (Robert et al., 2013). En effet, divers autres facteurs liés à la topographie ont une influence significative sur les mouvements de masses d’air au sein de la couche limite atmosphérique (Whiteman, 2000).

Les avancées en géomorphométrie et la disponibilité de MNT de haute résolution ont permis l’extraction de paramètres topographiques plus détaillés (Maxwell et al., 2022), améliorant ainsi l'évaluation des interactions entre le vent et le relief. Par exemple, Foresti et al. (2011) et Robert et al. (2013) ont appliqué l'algorithme de différence de Gaussiennes à un MNT pour mesurer la convexité du terrain dans le cadre de la modélisation des vitesses du vent en Suisse. De même, Jung et al. (2023b) ont mis en évidence l'importance de l'élévation relative (Lapen et al., 1993), une mesure de l'exposition d'un site au vent, dans l'interpolation spatiale des vitesses du vent en Allemagne.

L'exposition d'un site peut également être évaluée selon la direction du vent. Winstral et al. (2002) ont proposé l'indice d'exposition directionnel S_x, qui mesure le niveau d'exposition ou d'abri d'un site en fonction des obstacles topographiques présents dans une direction spécifique, jusqu'à une distance maximale donnée. La Figure 1.3 illustre cet indice pour une même zone, en prenant en compte deux directions de vent distinctes : nord-sud et est-ouest. Cette illustration montre comment la topographie locale affecte l'exposition ou la protection au vent selon sa direction.

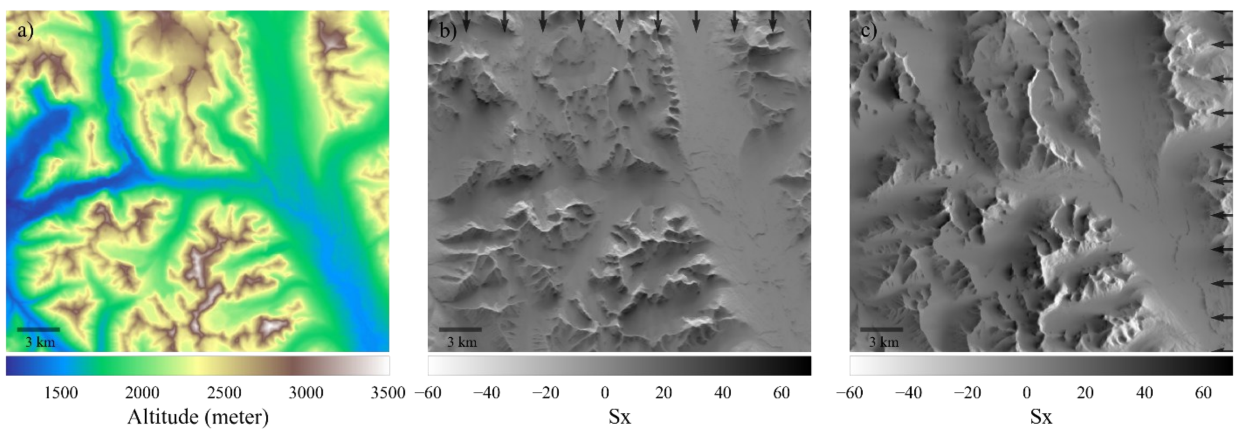


Figure 1.3 Indice d'exposition directionnel (S_x)

L'indice d'exposition directionnel a été calculé à partir d'un MNT (a) pour une direction du vent nord-sud (b), et est-ouest (c). Les flèches indiquent la direction du vent. Les valeurs négatives de l'indice (zones claires) signalent une exposition au vent provenant de la direction considérée, tandis que les valeurs positives (zones foncées) reflètent un effet d'abri dû à des obstacles topographiques. L'indice est calculé en considérant une distance maximale de 10 km. Les données d'élévation proviennent du Modèle numérique d'élévation de moyenne résolution (MNEMR) du Canada.

L'échelle spatiale est un facteur clé à considérer lors de l'extraction des variables topographiques à partir d'un MNT (Etienne et al., 2010; Grohmann, 2015). L'échelle spatiale optimale pour la modélisation d'un processus donnée est souvent inconnue à priori (Maxwell et al., 2022; Newman et al., 2022). Une échelle trop fine peut capturer des détails non pertinents et négliger des caractéristiques du terrain à une plus grande échelle qui influencent les processus étudiés. Tandis qu'une échelle trop large peut lisser les variations locales importantes et réduire la capacité du modèle à représenter des phénomènes critiques liés au relief. Pour surmonter cette difficulté, des approches multiéchelles sont fréquemment employées dans la modélisation des vitesses du vent (Etienne et al., 2010; Jung et al., 2020; Winstral et al., 2017). Ces approches consistent à extraire les variables topographiques à différentes résolutions spatiales et à analyser leur influence sur la variable cible.

À titre illustratif, la Figure 1.4 présente l'élévation relative estimée à deux résolutions spatiales distinctes. La cellule marquée en rouge a une élévation relative normalisée de 0,59 à une résolution de 10 km, ce qui indique qu'elle est relativement plus élevée que les cellules voisines et est donc plus exposée au vent. Cependant, lorsque la résolution spatiale est plus fine à 1,5 km, ladite cellule présente une élévation relative normalisée plus faible (0,18), indiquant une exposition au vent significativement moins élevée. Cette différence souligne l'importance de la résolution spatiale dans l'analyse de la topographie.

En outre, Petersen et al. (1998) classent la rugosité de surface parmi les éléments topographiques influençant les vitesses du vent près du sol. La longueur de rugosité z_0 est souvent dérivée de cartes d'occupation du sol à l'aide de relations empiriques établies entre les types d'occupation du sol et les valeurs de longueur de rugosité correspondantes (Davis et al., 2023; Wiernga, 1993).

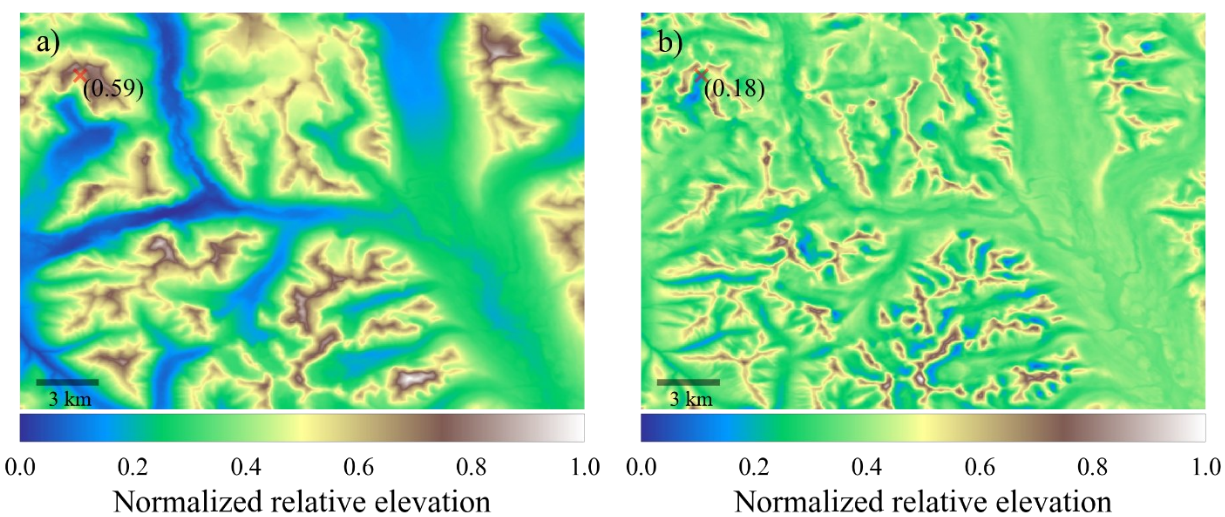


Figure 1.4 Élévation relative normalisée

L'élévation relative a été calculé à partir d'un MNT à une résolution spatiale de 10 km (panneau a) et de 1,5 km (panneau b).

1.6.2 Variables météorologiques

Le vent correspond aux mouvements de masse d'air résultant d'un gradient de pression, sa vitesse est donc étroitement liée aux variations de pression et de température dans l'atmosphère

(Ambach et al., 2017; Carrega, 2008). L'objectif des modèles d'estimation des vitesses du vent aux sites non échantillonnés est de fournir des prédictions dans des régions dépourvues de données de mesure directe. Dans ce contexte, l'utilisation de variables météorologiques comme prédicteur peut s'avérer inappropriée, car elles peuvent également être sujettes à des lacunes ou à des incertitudes similaires sur le plan de la disponibilité aux sites d'intérêts. Alternativement, des variables météorologiques issues de modèles numériques, notamment les données de réanalyse, sont souvent utilisées comme substituts. Les vitesses du vent issues de ces réanalyses sont généralement considérées comme des estimations fiables à l'échelle synoptique (Largeron et al., 2015).

Par exemple, Jung et al. (2020) ont utilisé des caractéristiques statistiques, notamment les L-moments, estimées à partir de séries temporelles de vitesses du vent issues de réanalyses pour représenter les vitesses du vent à l'échelle synoptique. Ces caractéristiques statistiques ont ensuite servi de variables explicatives pour l'estimation de la distribution de probabilité des vitesses du vent à des sites non échantillonnés.

1.6.3 Variables temporelles

Les variables temporelles, telles que l'heure de la journée et le mois de l'année, ont été utilisées pour modéliser les séries temporelles de vitesse du vent (Fadare, 2010; Jung et al., 2022a). Ces variables permettent de capturer les variations cycliques des régimes de vent, qui résultent de processus atmosphériques liés à l'ensoleillement, aux gradients de température et aux phénomènes climatiques locaux ou régionaux. Par exemple, l'heure de la journée est un facteur crucial pour représenter les variations diurnes, comme les brises terrestre et marine, qui se produisent dans les zones côtières.

1.7 Modèles d'apprentissage automatique

Les interactions entre le vent et la topographie forment un phénomène complexe, caractérisé par des dynamiques non linéaires (Whiteman, 2000). Ces interactions dépendent de plusieurs facteurs, notamment la rugosité de surface, le relief, et la stabilité atmosphérique, qui interagissent à différentes échelles. Cette complexité rend les approches simples, telles que la régression linéaire, inadaptées pour représenter la variabilité du processus.

L'apprentissage automatique est une discipline qui vise à concevoir des systèmes informatiques capables d'améliorer automatiquement leurs performances à partir de l'expérience (Jordan *et al.*, 2015). Ces systèmes ne sont pas programmés de manière explicite pour chaque tâche, mais ils apprennent directement à partir des données. En analysant les relations et les structures présentes dans les données, les algorithmes d'apprentissage construisent des modèles capables de faire des prédictions ou détecter des tendances.

Les approches d'apprentissage automatique peuvent être classées selon leur complexité et la profondeur de leur architecture en deux grandes catégories : l'apprentissage automatique classique (ou *shallow learning*) et l'apprentissage profond (*deep learning*). Les méthodes d' apprentissage automatique classique comprennent des algorithmes tels les forêts aléatoires (*random forest*, RF) (Breiman, 2001) et les algorithmes de boosting des arbres de décision (*gradient boosting trees*, GBT) (Hastie et al., 2009). Ces approches se basent généralement sur un ensemble limité d'étapes de transformation des données et nécessitent souvent en amont une ingénierie manuelle des variables les plus pertinentes à partir des données brutes. Contrairement aux techniques traditionnelles, les méthodes d'apprentissage profond reposent sur des architectures constituées de plusieurs couches hiérarchiques de traitement, permettant à chaque niveau d'extraire des représentations de plus en plus abstraites à partir des données brutes (Chollet, 2021).

Les techniques d'apprentissage automatique classique, telles que les algorithmes RF et GBT, ont été employées pour la modélisation de la vitesse du vent en raison de leur aptitude à modéliser des relations non linéaires et des interactions complexes entre les variables explicatives favorables à la prédiction de la variable cible.

Par exemple, Veronesi et al. (2016) ont appliqué un modèle RF pour estimer la distribution de la vitesse et de la direction du vent à des sites non échantillonnés au Royaume-Uni. La validation des résultats du modèle a révélé des performances supérieures à celles obtenues avec les méthodes d'interpolation spatiale classiques (Luo et al., 2008). Hu et al. (2023) ont proposé un modèle GBT intégrant des variables topographiques et météorologiques issues d'ERA5 pour prédire les vitesses du vent observées. Ce modèle a amélioré la précision des estimations, notamment dans les zones à topographie complexe, par rapport aux données brutes d'ERA5. Cependant, son influence s'est avérée moins prononcée dans les zones où la précision des données d'ERA5 était déjà élevée. Jung et al. (2020) ont indiqué que le modèle GBT surpassait la régression linéaire simple et le modèle RF pour la modélisation spatiale des L-moments de vitesse du vent à l'échelle mondiale.

En régression, les modèles RF et GBT reposent sur des ensembles d'arbres de décision, appelés arbre de régression. La construction d'un arbre de décision consiste à partitionner récursivement l'espace des variables explicatives en sous-ensembles de manière à minimiser un critère de variabilité au sein de chaque sous-ensemble résultant de la partition (Breiman et al., 1984). En régression, un critère couramment utilisé est la somme des carrés des résidus (Residual Sum of Squares, RSS) (Genuer et al., 2017):

Équation 1.17

$$RSS = \sum_{i \in R_m} \left(U_i - \widehat{U}_{R_m}\right)^2$$

où U_i, $i \in R_m$ représente les valeurs cibles des échantillons appartenant au sous-ensemble R_m et $\widehat{U}_{R_m}$est la prédiction de la variable cible dans ce sous-ensemble, généralement obtenue par la moyenne empirique des observations.

Les arbres de décision sont des modèles simples et faciles à interpréter, mais ils présentent une grande variabilité dans les prédictions. En effet, une petite modification des variables explicatives peut entraîner des changements significatifs dans la structure de l'arbre, ce qui rend les prédictions peu fiables (Hastie et al., 2009). Pour pallier ces limites, on privilégie des méthodes d'ensemble, comme les modèles RF et GBT. Le modèle RF permet de réduire la variance en combinant les prédictions de plusieurs arbres. Chaque arbre est construit à partir d'un échantillon aléatoire de données et de variables explicatives. L'échantillonnage aléatoire vise à minimiser la corrélation entre les arbres de la forêt.

Les modèles GBT sont construits à l'aide d'arbres de régression peu profonds, appelés « apprenants de base ». Ils corrigent successivement les erreurs de prédiction des arbres antérieurs :

Équation 1.18

$$f_m(x) = f_{m-1}(x) + \tau h_m(x; \theta_m)$$

où $f_m(x)$ représente la prédiction du modèle à l'itération m, x est le vecteur des variables explicatives, τ est un facteur de pondération (taux d'apprentissage) entre 0 et 1, qui contrôle la contribution de l'estimateur à l'itération m et $h_m(x; \theta_m)$ est l'estimateur de correction à l'itération m.

L'algorithme peut être initialisé avec la moyenne empirique de la variable cible. Aux itérations suivantes, les paramètres θ_m de l'estimateur $h_m(x; \theta_m)$ sont déterminés en minimisant le gradient de la fonction de perte $L\big(y_i, f(x_i)\big)$ estimé à partir des prédictions de l'itération précédente :

Équation 1.19

$$\theta_m = \arg\min_{\theta} \sum_{i=1}^{n} \left(\frac{\partial L\big(y_i, f_{m-1}(x_i)\big)}{\partial f_{m-1}(x_i)} - h_m(x_i;\theta) \right)^2$$

Au fil des années, plusieurs variantes du modèle GBT ont été développées afin d'améliorer ses performances, son efficacité et sa flexibilité. Notamment, XGBoost (Chen et al., 2016) et LightGBM (Ke et al., 2017) ont démontré de très bonne performance dans divers domaines (Bentéjac et al., 2021; Tyralis et al., 2021). Ces versions optimisées introduisent diverses améliorations, notamment la construction d'histogrammes pour les variables explicatives continues lors de l'identification des points de division optimaux dans les arbres de régression. Cette approche consiste à regrouper les valeurs d'une variable explicative en classes, réduisant ainsi le nombre d'itérations nécessaires pour déterminer les seuils de partition les plus pertinents.

XGBoost se distingue par l'ajout de termes de régularisation à la fonction de perte qui rendent le modèle plus robuste face au bruit dans les données, limitant ainsi le risque de surapprentissage. En outre, LightGBM introduit plusieurs autres optimisations pour le traitement efficace de données volumineuses et de haute dimension. Une de ces optimisations est le *Gradient-based One-Side Sampling* (GOSS), qui concentre les calculs sur les échantillons présentant les gradients les plus élevés, tout en incluant aléatoirement certains échantillons à faible gradient. Cet algorithme réduit considérablement la quantité de données à traiter à chaque itération. Le modèle LightGBM apporte également une solution au défi de la haute dimensionnalité des données en exploitant le regroupement des variables explicatives qui ne prennent pas simultanément des valeurs non nulles. Lorsque l'espace des variables explicatives est parcimonieux, il arrive souvent que de nombreuses variables soient mutuellement exclusives (p. ex., une variable associée à la présence ou l'absence d'un indicateur) et elles peuvent donc être groupées sans qu'il y ait une perte significative d'information (Ke et al., 2017).

Dans le traitement de vastes quantités de données issues de diverses sources, les approches d'apprentissage profond se sont révélées particulièrement performantes (Karpatne et al., 2019). L'un des principaux avantages de ces modèles est leur capacité à intégrer des biais inductifs dans l'apprentissage de la fonction de prédiction. Ils identifient et extraient automatiquement les caractéristiques les plus pertinentes, sans nécessiter l'ingénierie manuelle des variables (Reichstein et al., 2019). Par exemple, les réseaux convolutifs (CNN) exploitent les structures spatiales des données, facilitant ainsi l'intégration d'informations topographiques et météorologiques à haute résolution pour la prédiction du vent (Dujardin et al., 2022; Höhlein et al., 2020). De même, les réseaux de neurones récurrents (RNN) et leurs variantes, comme les

LSTM (*Long Short-Term Memory*) et GRU (*Gated Recurrent Unit*), sont particulièrement adaptés à l'analyse des séries temporelles, permettant l'extraction automatique des caractéristiques temporelles de variable explicative pertinente à la prédiction de la variable cible (Wang et al., 2021; Zhang et al., 2021).

Les RNN traditionnels sont confrontés à des problèmes, tels que la disparition et l'explosion des gradients, ce qui limite leur capacité à conserver des informations sur des séquences étendues. En incorporant un mécanisme de régulation qui régule le flux d'informations, les LSTM peuvent conserver et utiliser efficacement les caractéristiques temporelles pertinentes sur des périodes plus longues (Yu *et al.*, 2019).

Le Tableau 1.4 présente une synthèse des différentes approches d'apprentissage automatique abordées précédemment, en soulignant leurs principaux avantages ainsi que leurs limites.

Tableau 1.4 Synthèse des approches d'apprentissage automatique

Type de modèle	Algorithme	Atouts	Limites
Régression linéaire	Modèle paramétrique simple qui ajuste une relation linéaire entre les variables.	Interprétable et facile à entraîner.	Restreint aux relations linéaires simples.
Arbres à décision	Modèle fondé sur une série de règles de décision.	Interprétable, facile à entraîner et capable de modéliser des relations non linéaires.	Performance limitée et très sensible aux variations des données d'entrée.
Random forest	Ensemble d'arbres de décision construits en parallèle, puis agrégés pour produire une prédiction plus robuste.	Capable de modéliser des relations non linéaires, faciles à configurer et robustes face aux variables redondantes (Hastie *et al.*, 2009).	Interprétabilité réduite par rapport aux modèles linéaires ou aux arbres de décision simples.
Gradient boosting trees	Ensemble d'arbres de décision construits séquentiellement, où chaque nouvel arbre corrige les erreurs commises par des arbres antérieurs.	Modèle très flexible, notamment en ce qui a trait au choix de la fonction de perte.	Temps d'entraînement plus long et faible capacité de parallélisation par rapport au modèle Random Forest (Natekin *et al.*, 2013).

XGBoost	Implémentation optimisée de l'algorithme de gradient boosting avec régularisation de la fonction de perte.	Plus rapide à entraîner et très efficace pour la classification et la régression sur données tabulaires (Grinsztajn *et al.*, 2022a).	Configuration complexe comparée à l'algorithme classique de gradient boosting.
LightGBM	Implémentation optimisée de l'algorithme de gradient boosting pour des jeux de données volumineux et à hautes dimensions.	Très rapide à entraîner sur des jeux de données volumineux et à hautes dimensions (Bentéjac *et al.*, 2021).	Configuration complexe comparée à l'algorithme classique de gradient boosting.
LSTM	Réseau de neurones récurrents optimisé pour réduire le problème de disparition ou d'explosion du gradient dans les architectures récurrentes traditionnelles.	Adapté aux données séquentielles, comme les séries temporelles.	Nécessite de grandes quantités de données et reste difficile à interpréter.
CNN	Réseau de neurones convolutifs utilisant des couches de convolution pour extraire automatiquement des caractéristiques locales.	Adapté au traitement des données structurées en grille.	Long à entraîner, nécessite de grandes quantités de données et reste difficile à interpréter.

1.8 Objectifs et structure de la thèse

1.8.1 Objectifs de la thèse

Dans le cadre d'étude sur l'estimation des vitesses du vent à grande échelle, les modèles empiriques sont mieux adaptés, car elles nécessitent moins de ressources computationnelles que les approches physiques et sont généralement plus simples à mettre en œuvre. L'objectif principal de cette thèse est de développer des méthodes d'apprentissage automatique pour l'estimation à long terme de la vitesse du vent aux sites non échantillonnés. Les méthodes proposées s'appuient sur celles qui sont existantes afin de les améliorer et de réduire les incertitudes liées à l'évaluation des ressources éoliennes. Ces avancées contribueront, dès les premières phases, à affiner la planification des projets éoliens.

Les objectifs spécifiques de cette thèse sont présentés ci-dessous.

Objective n°1 : Revue de la littérature

Le premier objectif de cette thèse est de recenser et d'analyser les méthodes actuelles. Cet objectif a été abordé dans le premier article de cette thèse, intitulé « *Machine learning and statistical approaches for wind speed estimation at partially sampled and unsampled locations; review and open questions* ».

Dans un premier temps, nous avons procédé à une mise à jour de la revue des approches d'estimation à long terme des vitesses du vent aux sites disposant d'une courte série de données. Ensuite, nous avons passé en revue les méthodes d'estimation pour les sites non échantillonnés. Ces différentes méthodes ont été regroupées en fonction du type de variable cible pour une analyse plus structurée et une comparaison plus facile des approches. Enfin, la question de l'estimation des vitesses du vent dans un contexte non stationnaire a été abordée, en tenant compte des effets du changement climatique et des variations à long terme des régimes de vent. Cette étude nous a aidés à comprendre les avantages et les limites des méthodes actuelles. Ces limites soulignent le besoin de développer des méthodologies plus adaptées et plus précises pour mieux répondre aux défis liés à la variabilité spatiale et temporelle des régimes de vent.

Objective n°2 : Identification de méthodes performantes pour la sélection des variables explicatives pertinentes dans la modélisation empirique de la vitesse du vent

Le deuxième objectif de cette thèse est d'identifier les méthodes performantes pour la sélection des variables explicatives les plus pertinentes dans la modélisation empirique de la vitesse du vent. Cet objectif a été traité dans le deuxième article de cette thèse, intitulé « *Comparative study of feature selection methods for wind speed estimation at ungauged locations* ».

Le nombre de variables explicatives utilisées dans la modélisation du vent a considérablement augmenté en raison de la disponibilité de MNT de haute résolution et de diverses autres données issues de la télédétection. De plus, l'intégration de ces variables à différentes échelles spatiales a amplifié ce phénomène. Bien que cette abondance d'informations puisse améliorer la précision des modèles, elle soulève aussi des questions d'interprétabilité et de risque de surajustement des modèles.

Pour répondre à ces défis, des techniques de sélection de variables sont généralement employées comme étape de prétraitement afin de réduire la complexité des modèles tout en conservant les variables les plus pertinentes. Dans l'article 2, nous avons évalué les performances de plusieurs méthodes de sélection de variables explicatives, notamment la sélection pas à pas (*forward stepwise regression*), l'algorithme *Least Absolute Shrinkage and Selection Operato*r (LASSO), le modèle *Elastic Net*, l'algorithme *maximum relevance minimum redundancy* (MRMR), l'algorithme génétique et l'élimination récursive de variable (*recursive feature elimination*).

En outre, pour analyser l'influence des variables explicatives sur les différentes plages de vitesse du vent (faible moyenne et forte), plusieurs quantiles de vitesse ont été considérés avec des probabilités de dépassement variées. Cet aspect est crucial pour la modélisation de la distribution complète des vitesses du vent, car il permet une sélection plus adaptée des variables influentes en fonction du régime du vent.

Objective n°3 : Développement d'une approche non paramétrique d'interpolation spatiale de la distribution des vitesses de vent

Le troisième objectif de cette thèse est le développement d'une méthode non paramétrique pour l'estimation de la distribution de probabilité des vitesses du vent aux sites non échantillonnés. Cet objectif a été traité dans le troisième article, intitulé « *A non-parametric approach for wind speed distribution mapping* ».

Les méthodes actuelles d'estimation de la distribution de probabilité du vent reposent sur l'hypothèse qu'une seule loi de probabilité s'applique à l'ensemble de la région étudiée. Cette hypothèse, souvent trop restrictive, peut réduire la précision des estimations, particulièrement dans les régions où les régimes de vents présentent une forte variabilité spatiale.

Dans la méthodologie proposée dans l'article 3, on procède à une interpolation spatiale de plusieurs quantiles de vitesse du vent à l'aide d'un modèle de régression XGBoost. Ce modèle prend en entrée diverses variables liées à la topographie, aux conditions climatiques locales et à la longueur de rugosité. Une fois le modèle de régression établi, les quantiles de vitesse estimés aux sites non échantillonnés servent à estimer la fonction de répartition par noyaux asymétriques. Cette méthode permet d'obtenir une distribution de probabilité de la vitesse du vent aux sites non échantillonnés qui est suffisamment flexible pour s'accommoder des divers régimes de vent présents dans la région étudiée.

Objective n°4 : Développement de méthodes de reconstruction de séries temporelles de vitesse de vent aux sites non échantillonnés.

Le quatrième objectif de cette thèse vise à proposer de nouvelles méthodes statistiques pour la correction des biais des données de vitesse du vent issues de réanalyses. Cet objectif a été traité dans les articles 4 et 5 de cette thèse.

Dans l'article 4, intitulé « *Prediction of hourly wind speed time series at unsampled locations using machine learning* », nous proposons d'adapter la méthode quantile-quantile (quantile mapping) pour reconstruire, aux sites non échantillonnés, des séries temporelles de vitesse du vent à partir de séries temporelles de probabilité de non-dépassement. Ces dernières sont obtenues soit à partir de données de réanalyse, soit par interpolation spatiale d'observations in situ.

La méthode quantile-quantile repose sur l'interpolation non paramétrique de la distribution du vent, développée dans l'article 3, pour estimer la distribution des vitesses du vent aux sites non échantillonnées. Cette approche permet de reconstruire des séries temporelles dont les quantiles estimés par régression représentent mieux les conditions locales des vitesses du vent. Une analyse comparative de la méthode proposée et d'autres techniques d'interpolation spatiale directe des observations in situ a été réalisée. Plusieurs critères ont été considérés, dont la distribution de probabilité et la variabilité temporelle.

Enfin, l'article 5 intitulé « *LSTM and Transformer-based framework for bias correction of ERA5 hourly wind speeds* » introduit une nouvelle approche de correction des données de vitesse du vent issues de réanalyse, basée sur l'apprentissage profond. La plupart des études existantes utilisent une méthode de correction de biais consistant à appliquer un facteur constant, calculé à partir des vitesses du vent moyennes fournies par le GWA. Bien que cette méthode permette de réduire le biais systématique dans les données de réanalyse, elle ne tient pas compte de leur variabilité temporelle. Le modèle proposé s'appuie sur les architectures LSTM et Transformer. Il prend comme entrée des variables météorologiques dynamiques issues de réanalyse, ainsi que des variables statiques liées à la topographie locale et à la longueur de rugosité, pour prédire un facteur de correction dynamique. Le facteur de correction estimé permet d'ajuster la variabilité temporelle des séries de vitesses du vent issues de réanalyse, tout en réduisant le biais systématique.

1.8.2 Structure de la thèse

Le premier chapitre de cette thèse présente le contexte de la recherche et une revue de littérature générale. Il énonce également les objectifs et l'organisation de l'ensemble de la thèse. Le deuxième chapitre porte sur l'article 1, qui passe en revue les méthodes statistiques et d'apprentissage automatique pour estimer les vitesses du vent aux sites partiellement échantillonnés (disposant d'une courte série de mesures) et non échantillonnés. Ces travaux représentent une étape importante pour évaluer les méthodologies existantes. Ils permettront de proposer de nouvelles approches dans les articles ultérieurs.

Le troisième chapitre porte sur l'article 2, qui compare six méthodes de sélection de variables explicatives pour l'estimation des quantiles de vitesse du vent au Canada. Les variables ainsi identifiées serviront d'entrée pour l'ensemble des modèles développés dans les travaux suivants.

Le quatrième chapitre est dédié à l'article 3, qui propose une méthode non paramétrique d'interpolation spatiale de la distribution du vent. Le cinquième chapitre se concentre sur l'article 4, qui s'appuie sur la méthode développée dans l'article 3 pour reconstruire des séries temporelles de vitesse du vent aux sites non échantillonnés. De plus, une analyse comparative détaillée des méthodes actuelles et de celle proposée dans l'étude a été réalisée.

Le sixième chapitre est consacré à l'article 5, qui se concentre sur l'ajustement de la variabilité temporelle des séries de vitesse du vent issues de réanalyses.

Enfin, le septième chapitre offre une discussion générale et une conclusion des travaux réalisés dans cette thèse. Il souligne les limites de l'étude et suggère des pistes de recherche pour des travaux futurs.

2 MACHINE LEARNING AND STATISTICAL APPROACHES FOR WIND SPEED ESTIMATION AT PARTIALLY SAMPLED AND UNSAMPLED LOCATIONS ; REVIEW AND OPEN QUESTIONS

Titre de l'article : Approches d'apprentissage automatique et statistiques pour l'estimation de la vitesse du vent aux sites partiellement échantillonnés et non échantillonnés : revue et avenues de recherche

Auteurs :

Freddy Houndekindo[1] (F.H.), Taha B.M.J. Ouarda[1] (T.B.M.J.O)

[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

Titre de la revue :

Article publié le 1 mars 2025, dans la revue **Energy Conversion and Management**, volume 327, DOI : 10.1016/j.enconman.2025.119555

Contribution des auteurs :

Conceptualisation, T.B.M.J.O. ; Méthodologie, F.H., T.B.M.J.O. ; Rédaction de la version originale, F.H. ; Révision, T.B.M.J.O. ; Supervision, T.B.M.J.O. ; Acquisition du financement, T.B.M.J.O.

Abstract

Wind resource assessment (WRA) depends on the availability of accurate, long-term wind speed data. In locations where such data is limited (partially sampled locations, PSL) or completely missing (unsampled locations, USL), various physical, statistical, and machine-learning methods have been developed to address these gaps. This paper presents a comprehensive and up-to-date review of statistical and machine-learning methods for estimating long-term wind speed at PSLs and USLs.

It was found that the “Measure Correlate Predict” (MCP) is still the method of choice for estimation at PSL. However, this approach has evolved with the adoption of machine learning, especially Artificial Neural Networks, and reanalysis wind data as the reference site. In general, reanalysis datasets have seen growing adoption for WRA at both PSLs and USLs due to their global coverage, high temporal resolution, and demonstrated accuracy. At USLs, uncorrected and bias-corrected reanalysis wind speed data are used for WRA, with the Global Wind Atlas predominantly used to correct reanalysis wind speed data. There is also a growing effort to develop machine learning models, including deep learning models for reanalysis bias correction at unsampled locations using explanatory variables derived from high-resolution topographic and land use datasets.

Challenges to estimating long-term wind speed at PSLs and USLs are identified and discussed: data uncertainties, disparity in the accuracy of reanalysis wind data, model transferability, and nonstationary conditions. Finally, recommendations for future research and development directions are presented, including techniques that consider documented non-stationarity in wind speed data.

Keywords: Machine learning; Measure correlate predict; Nonstationary; Wind resource assessment; Partially sampled; Unsampled

2.1 Introduction

The availability of clean energy is critical for meeting several of the United Nations’ sustainable development goals (Fuso Nerini *et al.*, 2018). Present energy systems still have a high environmental footprint and are major contributors to greenhouse gas emissions (Ahmad *et al.*, 2020). Developing and adopting renewable energy sources are central to addressing the current climate crisis and fulfilling Goal seven of the United Nations’ Sustainable Development Goals, “Ensure access to affordable, reliable, sustainable and modern energy for all” (UN

General Assembly, 2015, p. 19). Wind energy represents a sustainable alternative to conventional energy sources and is projected to play a significant role in meeting global energy demands. By 2030, wind resources are expected to contribute nearly 24% of global electricity generation (Gielen *et al.*, 2021), up from just 8% of the total installed capacity worldwide in 2018 (International Renewable Energy Agency, 2022). Meeting this target will require accurate resource assessment (McKenna *et al.*, 2022).

In an ideal scenario, the assessment of the wind potential would rely on long-term (e.g., 30 years) wind speed data with high temporal resolution (Bosch *et al.*, 2017). However, such extensive data are frequently unavailable. Accurate wind speed data is essential for a reliable assessment of the energy potential, given the cubic relationship between wind speed and wind power. Several studies have proposed statistical and machine learning (ML) methods to address wind resource assessment (WRA) in locations with limited (herein called partially sampled location, PSL) or no wind speed measurements (herein called unsampled location, USL). At PSLs, the preferred method is the "Measure Correlate Predict" (MCP) approach. This method extends the available data at a location by correlating it with measurements from a nearby station with overlapping and more extensive records. Carta *et al.* (2013) provided a review of the MCP approach. Over the past decade (2014 – 2023), advancements have been made in the MCP methodologies. These include the widespread adoption of reanalysis wind speed data as reference (Basse *et al.*, 2021; Miguel *et al.*, 2019; Yue *et al.*, 2019), the application of ML models (Díaz *et al.*, 2017; Kristianti *et al.*, 2023; Schwegmann *et al.*, 2023), and the use of multiple reference sites to improve accuracy (Carta *et al.*, 2015; Zhang *et al.*, 2014). In addition, the MCP technique has been extended to estimate other wind-related variables, such as wind power (Díaz *et al.*, 2018) and peak gust (Kartal *et al.*, 2023).

Given these advancements, the literature review on wind speed estimation at PSL needs to be updated. This paper aims to provide a comprehensive review of these recent developments, highlighting their implications for the MCP approach and its broader applications.

At USLs, installing a meteorological station to collect sufficient data (e.g., for one year) to implement the MCP approach is possible. However, alternative statistical and ML methods must be employed when such installations are impossible due to time or financial constraints. Over the years, various types of wind speed data have been estimated at USLs at different time scales, including mean wind speed (MWS), wind speed probability distribution (WSPD), extreme wind speed (EWS), and wind speed time series (WSTS). The choice of the estimation method depends on the specific wind speed variable of interest.

A growing body of research has relied on reanalysis data and ML models to interpolate wind speed at USLs, demonstrating significant potential with varying reported degrees of success. Despite these advances, to the best of the authors’ knowledge, no review systematically examined statistical and ML methods for estimating long-term wind speed at USLs. This literature review aims to bridge this gap by providing a detailed evaluation of the state-of-the-art methods used for long-term wind speed estimation at USLs.

This paper comprehensively reviews statistical and ML methods for estimating wind speed at PSLs and USLs. It also examines comparative studies of existing methods and explores approaches that allow addressing documented non-stationarities in wind speed data. Additionally, the paper presents software and tools available for wind speed estimation.

This review can have broad implications for the research community and the wind energy industry. By providing a comprehensive review of state-of-art methods and identifying research gaps, it encourages further development of methodologies for wind speed estimation at PSLs and USLs. These advancements can lead to several key benefits, such as the elaboration of improved methodologies for the reliable estimation of wind resources and the development of tailored techniques for various project phases, such as large-scale assessment during early exploration and small-scale assessment in the later development stage when higher accuracy is required. This review advances the understanding of wind speed estimation methodologies and contributes to accelerating the adoption of wind energy as a sustainable solution to global energy and climate challenges.

The review is structured as follows: section 2.2 discusses various preprocessing procedures for wind speed data. Section 2.3 presents methods for long-term wind speed estimation at PSLs, while section 2.4 focuses on methods for long-term estimation at USLs. Section 2.5 explores nonstationary wind speed estimation. Section 2.6 outlines the software and tools available. Section 2.7 discusses current open questions and outlines future research directions. The last section concludes the paper.

2.2 Wind speed data preprocessing

Several issues affect data collected at meteorological stations and must be addressed prior to analysis. According to Pryor et al. (2009), common issues that affect the quality of near-surface wind speed measurements include station relocation, aging or malfunctioning of measurement instruments, upgrades to equipment or changes in measurement height, and alterations in the surrounding environment, which can induce a modification of surface roughness and wind exposure. While metadata can provide helpful information for identifying these issues, they are not always exhaustive and reliable (Aguilar et al., 2003). Several statistical methods are available to address these concerns. Wind data preprocessing includes the following steps:

2.2.1 Quality control and change point detection

Quality control of wind speed data reported in the literature relies on one or several of the following methods: 1) Measurements that significantly deviate from those recorded at adjacent stations are flagged as inconsistent records (Azorin-Molina et al., 2018; Zahradníček et al., 2019). When an adjacent station with a strong correlation is unavailable, Numerical Weather Prediction (NWP) and reanalysis datasets are also used for quality control. 2) Measurements that deviate significantly from neighboring values in the time series are flagged as suspicious (Wan et al., 2010). 3) Extreme values, such as daily gust or region-specific physical maxima, serve as thresholds to detect anomalies (Azorin-Molina et al., 2014; Minola et al., 2016; Wan et al., 2010).

Discontinuities in wind speed data due to non-climatic factors are common, especially with lengthier records (Wan et al., 2010; World Meteorological Organization, 2020). Several statistical methods are available for detecting time series discontinuities (change points). Table 2.1 presents software/packages for change point detection and homogenization applied to wind speed data. Several other efficient techniques for change point detection and data homogenization are available in the literature and are commonly applied to other meteorological variables (see, for instance, Seidou et al. (2007a); Xiong et al. (2015)).

Table 2.1 Software and packages for change point detection with application to wind speed data

Software/package	Reference	Example(s) of application to wind speed data
AnClim (Štěpánek *et al.*, 2009)	Standard normal homogeneity test (Alexandersson, 1986)	(Azorin-Molina *et al.*, 2014; Minola *et al.*, 2016)
Bayesian Changepoint Detection Procedure codes (Seidou *et al.*, 2007b)	Recursion-based multiple changepoint detection procedure (Seidou *et al.*, 2007b)	(Naizghi *et al.*, 2017)
Climatol (Guijarro, 2018)	Standard normal homogeneity test (Alexandersson, 1986).	(Azorin-Molina *et al.*, 2016; Azorin-Molina *et al.*, 2019; Zhang *et al.*, 2022a)
HOMER (Mestre *et al.*, 2013)	Penalized maximum likelihood. The package uses the modified Bayes information criterion (MBIC) (Zhang *et al.*, 2007) to penalize the likelihood	(Laapas *et al.*, 2017)
Multiple Analysis of Series for Homogenization (MASH) (Szentimrey, 1999)	Relative breakpoint detection using non-homogeneous reference sites.	(Németh *et al.*, 2014)

RHtestV4 package in R	Penalized maximal t-test (Wang *et al.*, 2007) Penalized Maximal f-test (Wang, 2008)	(Cui *et al.*, 2018; Jung *et al.*, 2015; Si *et al.*, 2018)

2.2.1 Measurement height adjustment

Wind speed measurements are typically collected at a height of 10 m above ground level (a.g.l) following the World Meteorological Organization (WMO) guidelines. However, measurements taken at turbine hub heights (e.g., 100 m a.g.l) are essential for accurately assessing wind resource potential. The power law (Equation 2.1) and the logarithmic model (Equation 2.2) are widely used for wind speed vertical extrapolation:

Equation 2.1

$$U_z = U_{z_r}\left(\frac{z}{z_r}\right)^{\alpha}$$

where U_{z_r} is the wind speed at a reference height z_r (e.g., 10 m a.g.l), U_z is the extrapolated wind speed at a higher height z, α is the power law exponent, and z_0 is the roughness length.

Equation 2.2

$$U_z = U_{z_r}\left(\frac{\ln(z/z_0)}{\ln(z_r/z_0)}\right)$$

where U_{z_r} is the wind speed at a reference height z_r (e.g., 10 m a.g.l), U_z is the extrapolated wind speed at a higher height z, α is the power law exponent, and z_0 is the roughness length.

Gualtieri (2019) reviewed these models, indicating that the logarithmic model is inadequate for modern turbine heights, while the power law offers ease of use and greater reliability. Typically, a constant power law exponent is used, $\alpha = \frac{1}{7}$. This simplification does not account for the spatial and temporal variability of α, which limits the model's accuracy (Crippa *et al.*, 2021; Li *et al.*, 2018; Yang *et al.*, 2024). Some studies have suggested modeling time-dependent α to improve the precision of wind speed extrapolation (Basse *et al.*, 2020; Crippa *et al.*, 2021; Jung *et al.*, 2021).

A significant challenge with this approach is the requirement for wind speed measurements at two different heights. To address this limitation, Jung *et al.* (2017a) used reanalysis wind speed data at 1000 m a.g.l to estimate the distributions of α at 397 meteorological stations in Germany. They then used the Gaussian copula function to model the joint distribution of α and wind speed measurement at 10 m a.g.l, enabling the extrapolation of 10 m wind speeds to turbine hub heights while accounting for the spatiotemporal variability of the α.

The logarithmic and power law models are considered parametric approaches for vertical wind speed extrapolation. In recent years, nonparametric approaches such as gradient boosting (GB) (Yu *et al.*, 2022), symbolic regression (Valsaraj *et al.*, 2020), and Artificial neural networks (ANNs) (Vassallo *et al.*, 2020) have been explored as alternatives. These models establish a regression function between wind speeds at different heights.

However, applying these regression models for vertical wind speed extrapolation at PSLs and USLs poses challenges due to the limited availability of wind speed measurement at multiple heights at these locations. Reanalysis data may be an alternative, but it introduces an additional layer of uncertainty. Future studies could examine the effectiveness of regression models using reanalysis wind speed data, compared to traditional methods (e.g., power law with $\alpha = \frac{1}{7}$) for the vertical extrapolation of wind speed at PSLs and USLs. This evaluation would help identify the trade-offs between model accuracy, complexity, and data requirements, ultimately guiding the selection of the most appropriate extrapolation methods at PSLs and USLs.

2.3 Long-term wind speed estimation at partially sampled locations

A PSL refers to a location where measurements are insufficient to accurately evaluate the long-term wind potential, whereas USLs lack wind speed measurements. Methods for estimating the long-term wind potential at PSLs rely on high-quality, long-term wind data (reference data) from nearby meteorological stations or reanalysis datasets. Historically, the MCP approach was the standard method for estimating the long-term wind conditions at PSLs (Carta *et al.*, 2013). This approach has evolved with the adoption of ML models and reanalysis wind data as reference data.

The following subsections provide a review of studies published between 2014 and 2023 on long-term wind speed estimation at PSLs, exploring the latest developments and comparative studies.

2.3.1 Overview of studies on estimation at partially sampled locations

Table 2.2 summarizes the reviewed studies (published between 2014 and 2023) on long-term wind speed estimation at PSLs. This section provides an overview of these studies and highlights new developments.

Record length. A key factor in accurately estimating long-term wind resources at PSLs is the length of the available short-term wind records. Longer records typically lead to less uncertainty in MCP (Landberg *et al.*, 1993; Miguel *et al.*, 2019; Rogers *et al.*, 2005b). Collecting at least one year of wind data at the PSL is generally recommended to capture the resource seasonal variability. Figure 2.1 shows the distribution of minimum record lengths used at PSLs, as reported in the reviewed studies. In line with the recommendations, most studies (68%) reported using at least 12 months of wind speed data.

Given the financial and time constraints of collecting year-long wind speed data at PSLs, some studies have investigated the uncertainty associated with shorter record lengths. Findings indicate that with shorter record lengths, the season covered by the data influences the model performance (Basse *et al.*, 2021; Weekes *et al.*, 2014b). For example, collecting data during a low wind speed season could result in underestimating the long-term wind potential during higher wind speed seasons.

Reference site. The correlation between wind speed at the PSL and a reference station during the overlapping period is another key factor in long-term wind resource estimation. To ensure a high level of accuracy, MCP traditionally requires a high Pearson correlation coefficient between wind speed data at the reference station and PSL. This requirement was mainly due to using linear regression (LR) as the prediction model. However, with non-linear regression and ML models, more complex relationships between the overlapping wind speed data could be accommodated (Díaz *et al.*, 2018; Kristianti *et al.*, 2023; Schwegmann *et al.*, 2023).

Adopting ML models has also enabled the development of methods that use multiple reference sites. Two such methods were identified in the literature. The most common method involves fitting a single regression model using wind speed data from multiple reference stations (Carta *et al.*, 2015; Deo *et al.*, 2018; Díaz *et al.*, 2018). An alternative method proposed by Zhang *et al.* (2014) involves fitting separate models for each reference site, with the final prediction determined by a weighted average of these individual predictions. The weights are based on the distance and difference in elevation between the PSL and the reference stations.

Wind direction in MCP. With the LR model, a standard step involved partitioning the overlapping wind speed based on wind direction at the reference site and fitting separate LR models for each bin. This method helps relax the linearity assumptions and better capture directional variability in wind patterns. However, the effectiveness of this technique depends on having sufficient data in each wind direction bin to estimate the regression coefficients reliably. The model accuracy may be hindered when data are sparse in specific wind directions (Addison *et al.*, 2000; Dinler, 2013). Also, determining the optimal number of bins poses additional challenges (Rogers *et al.*, 2005b). With the introduction of ML techniques in MCP, greater flexibility has been achieved, eliminating the need for manual binning. These models allow the wind direction to be used directly as an explanatory variable (Carta *et al.*, 2015; Mifsud *et al.*, 2018), simplifying the process and potentially improving performance.

Adoption of reanalysis wind speed data. As the availability of a suitable reference station located near the PSL is uncertain, reanalysis wind speed data are gaining attention as alternatives (Basse *et al.*, 2021; Gottschall *et al.*, 2021; Kim *et al.*, 2022; Kim *et al.*, 2016; Lee *et al.*, 2019; Miguel *et al.*, 2019; Saarnak *et al.*, 2014; Schwegmann *et al.*, 2023). The main advantages of modern reanalysis data include their extensive record lengths, typically spanning over 30 years of hourly measurements, completeness, global coverage, and the availability of wind speed data at wind turbine hub heights.

In recent years, the accuracy of reanalysis data has significantly improved due to advancements in NWP models and data assimilation techniques that integrate satellite, ground-based, and remote sensing observations (Gualtieri, 2022; Hersbach *et al.*, 2020). These advancements have enhanced the reliability of reanalysis data for WRA.

Table 2.2 Summary of the reviewed studies on wind speed estimation at PSLs

Study	Study area	Location type	Minimum available record length at the PSL (months)	Reference data source	Prediction model
(Ali *et al.*, 2018)	South Korea	onshore	24	Measurements	LR, VR
(Basse *et al.*, 2021)	Germany	onshore	3	Reanalysis	LR, VR
(Brune *et al.*, 2022)	Germany	Onshore/offshore	6	Reanalysis	ANN
(Carta *et al.*, 2015)	Spain	Onshore	12	Measurements	ANN
(Cavaiola *et al.*, 2023)	Italy	onshore	12	Reanalysis	RF
(Díaz *et al.*, 2017)	Spain	onshore	12	Measurements	SVR
(Díaz *et al.*, 2018)	Spain	Onshore	12	Measurements	ANN, SVR, RF
(Gottschall *et al.*, 2021)	North Sea and Baltic Sea	offshore	24	Reanalysis	LR
(Kang *et al.*, 2015)	South Korea	onshore	12	Measurements	LR
(Kartal *et al.*, 2023)	United-States of America	onshore	12	Reanalysis	RF, Others
(Kim *et al.*, 2022)	South Korea	offshore	12	Reanalysis	LR
(Ko *et al.*, 2015)	South Korea	onshore	12	Measurements	LR

(Kristianti *et al.*, 2023)	Switzerland	Onshore	3	Measurements	ANN
(Mifsud *et al.*, 2018)	Malta	Onshore	12	Measurements	ANN, SVR, LR, RF
(Mifsud *et al.*, 2020)	Malta	Onshore	12	Measurements	ANN, SVR, LR, RF
(Miguel *et al.*, 2019)	Brazil	Onshore	24	Reanalysis	Others
(Morales-Ruvalcaba *et al.*, 2020)	Mexico	Onshore	12	Reanalysis	Others
(Salehi Borujeni *et al.*, 2021)	USA	onshore	12	Reanalysis	ANN
(Schwegmann *et al.*, 2023)	Germany, France	Offshore/onshore	24	Reanalysis	LR, RF, ANN
(Weekes *et al.*, 2014b)	UK	onshore	3	Measurements	LR, VR
(Weekes *et al.*, 2014a)	UK	onshore	1	Measurements	Others
(Weekes *et al.*, 2014c)	UK	onshore	3	Measurements	LR, VR
(Weekes *et al.*, 2015)	UK	Onshore	1	Operational forecast data	LR, VR
(Yue *et al.*, 2019)	Taiwan	Offshore	12	Reanalysis / measurements	LR
(Zhang *et al.*, 2014)	USA	onshore	3	measurements	LR, VR, ANN, SVR, Others

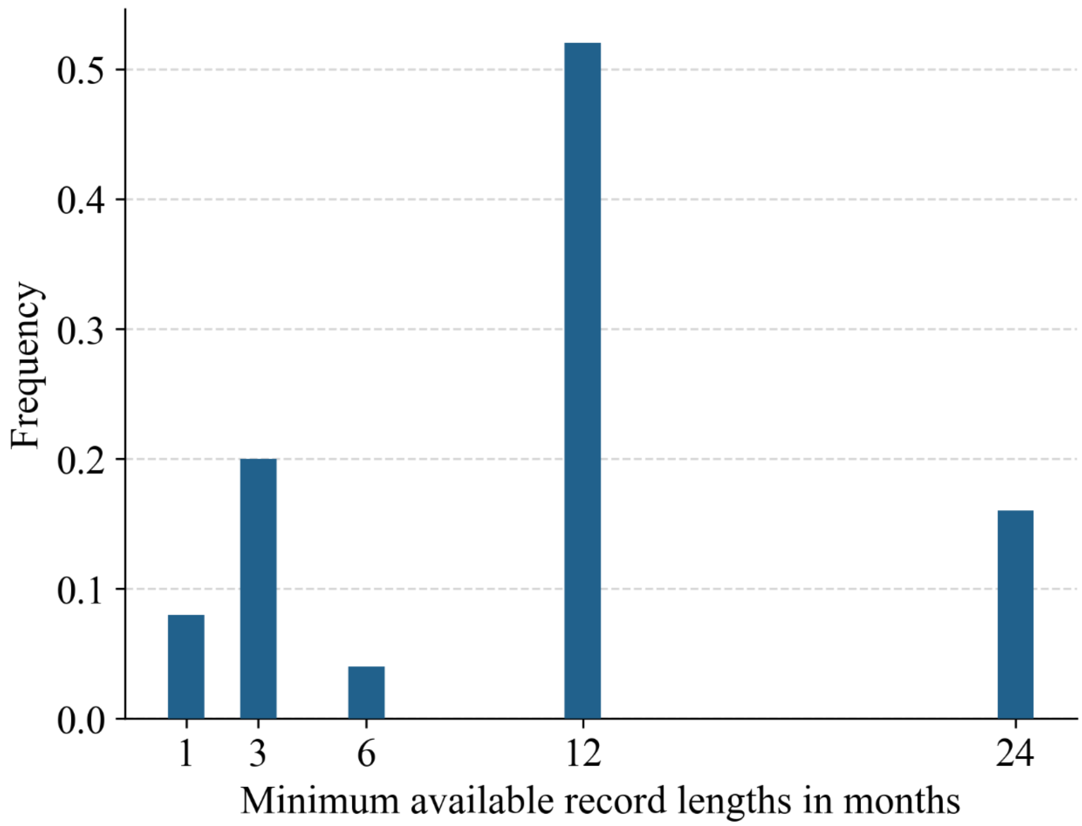


Figure 2.1 **Distribution of minimum record lengths used for wind speed estimation at PSLs in the reviewed studies**

2.3.1 Estimation methods at partially sampled locations

Figure 2.2 shows the proportion of the different prediction models used for long-term wind speed estimation at PSLs. The LR model has been widely applied in MCP (Carta *et al.*, 2013) and continues to serve as a benchmark in recent studies (Díaz *et al.*, 2017; Mifsud *et al.*, 2018; Weekes *et al.*, 2014a; Zhang *et al.*, 2014). Another linear model, the variance ratio regression (VR), was introduced by Rogers *et al.* (2005a) after they observed that LR tends to underestimate the prediction variance. The equation of the VR is expressed as follows:

Equation 2.3

$$\hat{y} = \left({}^{\sigma_y}/_{\sigma_x} \right) x + \left(\mu_y - \left({}^{\sigma_y}/_{\sigma_x} \right) \mu_x \right)$$

where $\hat{y}$ is the predicted wind speed, σ_y, and σ_x are the standard deviation of the wind speed at the PSL and the reference site, respectively, μ_y, and μ_x are the PSL's and reference site's mean wind speeds respectively, and x is the observed wind speed at the reference site

ML models such as ANN (Carta *et al.*, 2015; Díaz *et al.*, 2018; Koo *et al.*, 2015; Kristianti *et al.*, 2023; Mifsud *et al.*, 2018; Mifsud *et al.*, 2020; Salehi Borujeni *et al.*, 2021; Schwegmann *et al.*, 2023; Zhang *et al.*, 2014), support vector regression (SVR) (Díaz *et al.*, 2017; Díaz *et al.*, 2018; Mifsud *et al.*, 2018), random forest (RF) (Díaz *et al.*, 2018; Kartal *et al.*, 2023; Mifsud *et al.*, 2018; Schwegmann *et al.*, 2023) and GB (Kartal *et al.*, 2023) have been employed in the reviewed studies.

The multilayer perceptron (MLP) neural network was the most used ANN architecture. The MLP consists of an input layer, one or more hidden layers, and an output layer. Each layer contains hidden units (neurons), followed by a non-linear activation function used to model non-linear relationships between inputs and outputs. In matrix notation, the predictions from an MLP model are given by:

Equation 2.4

$$\hat{y} = g^{(l)}\left(\boldsymbol{b}^{(l)} + \boldsymbol{W}^{(l)}\boldsymbol{h}^{(l-1)}\right)$$

where the hidden vector $h^{(l-1)}$ is defined as:

Equation 2.5

$$\boldsymbol{h}^{(l-1)} = g^{(l-1)}\left(\boldsymbol{b}^{(l-1)} + \boldsymbol{W}^{(l-1)}\boldsymbol{h}^{(l-2)}\right)$$

where $\boldsymbol{W}^{(l)}$, $\boldsymbol{b}^{(l)}$ and $\boldsymbol{g}^{(l)}$ represent the weight matrix, bias vector, and activation function for the l-th layer of the neural network, respectively. The vector $\boldsymbol{h}^{(0)} = x$ is the input of the MLP and the weight matrices $\boldsymbol{W}^{(l)}$ and the bias vectors $\boldsymbol{b}^{(l)}$ are the learnable parameters of the models, which are typically optimized using the backpropagation algorithm (Goodfellow et al., 2016).

Various MLP architectures were implemented across the reviewed studies. Some studies (Carta et al., 2015; Díaz et al., 2018; Mifsud et al., 2018) used MLPs with a single hidden layer and the sigmoid activation function. However, the use of the sigmoid activation function has declined in recent years due to the vanishing gradient problem, which impairs the training of deeper networks (Apicella et al., 2021). The selection of the adequate activation function is critical for model performance, and in modern ANN architectures, rectified-based activation functions (Apicella et al., 2021) have become more common, particularly as the number of hidden layers increases. For instance, Kristianti et al. (2023) used the Continuously Differentiable Exponential Linear Units (CELU) activation function to train an MLP with four hidden layers. The CELU activation function is a more recent development designed to maintain the advantages of rectified-based functions while offering smoother gradients, which enhances performance and convergence in deep networks.

Selecting a suitable loss function is critical in training ANNs, as it penalizes the difference between predicted outputs and actual target values. The mean squared error (MSE) was the most frequently used loss function in the reviewed studies. However, MSE has been reported to potentially distort the prediction distribution (Jung et al., 2013). To address this issue, Kristianti et al. (2023) scaled MSE by the inverse probability of occurrence of the actual target value in the training set. A similarly scaled MSE was proposed by Jung et al. (2013) to assign greater importance to less frequent wind speed values during training.

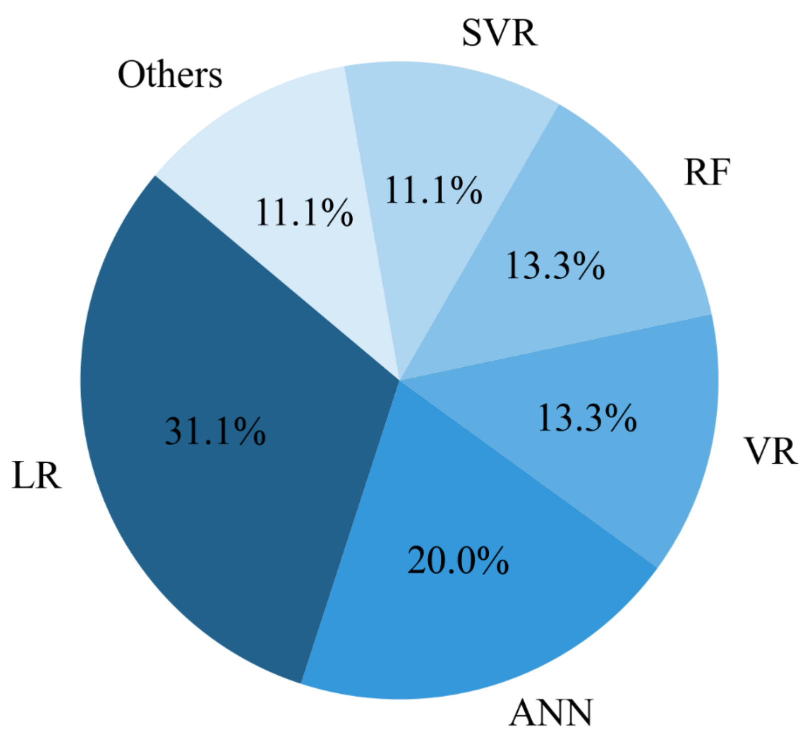


Figure 2.2 **Proportion of the different models used for long-term wind speed estimation at PSLs in the reviewed studies**

2.3.1 Comparative studies at partially sampled locations

Statistical and ML models exist in many varieties, and their performance can differ based on the specific prediction task at hand. Therefore, it is crucial to evaluate multiple models to identify the most suitable for a given application. In the context of wind speed estimation at PSLs, several comparative studies have analyzed the performance of statistical and ML models, as summarized in Table 2.3.

The criteria for evaluating model performance included standard regression metrics such as Mean Absolute Error (MAE), MSE, Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). MAE, MSE, and RMSE are scale-dependent metrics. To conduct evaluations that are independent of the magnitude of the target variable, Basse et al. (2021) used relative bias in mean (Err_{mean}) and variance (Err_{var}), while Díaz et al. (2018) employed the Mean Absolute Relative Error (MARE). These metrics are described in

Table 2.4.

Ultimately, the goal of the MCP method is to assess the wind energy potential of the PSL. Although wind speed is the main input for this analysis, it is crucial to consider the cubic relationship between wind speed and power output. Some researchers have directly evaluated the models by calculating metrics based on the derived power output (Basse et al., 2021; Díaz et al., 2018; Mifsud et al., 2020). While the approach may offer a more direct estimate of model performance, which is relevant to the project, it requires the selection of a power curve associated with the wind turbine at this stage. In their study, Basse et al. (2021) reported that variations in results due to the use of different power curves was marginal.

The findings from comparative studies (see Table 2.3) highlight the importance of testing various models, as no single model consistently outperforms the others. For example, Díaz et al. (2018) found that RF and SVR outperformed ANN, while Mifsud et al. (2020) reported the opposite, indicating that ANN surpassed both RF and SVR. Additionally, Schwegmann et al. (2023) indicated that the K-nearest neighbor (KNN) algorithm outperformed ANN, Gaussian process, RF, and SVR, particularly excelling in reproducing the wind speed distribution.

Differences in model configurations can explain the varying conclusions of these studies. Factors such as model hyperparameters, feature selection, data preprocessing techniques, and the length of training periods can all significantly influence model performance. The effectiveness of a particular model often depends on how well it is tuned to fit the specific characteristics of the dataset being used.

Table 2.3 Comparative studies of statistical and ML methods for long-term wind speed estimation at PSLs

Study	Performance metrics	Evaluation strategy	Results	Comments

(Basse *et al.*, 2021)	Err_{mean}, Err_{var}	Cross-validation	VR and LR produces different seasonal bias	Only three months of wind records at the PSL was used to evaluate seasonal biases
(Díaz *et al.*, 2017)	MAE, MARE, R^2		SVR > LR	The performance metrics were calculated from estimated power output
(Díaz *et al.*, 2018)	MAE, MARE, R^2	Cross-validation	[RF, SVR] > ANN	The performance metrics were calculated from estimated power output
(Kartal *et al.*, 2023)	Bias, MAE, MSE and R^2	Cross-validation	RF = GB	The peak wind gusts were estimated using the MCP approach.
(Mifsud *et al.*, 2018)	MAE, MSE, R^2	Test set	[ANN, RF] > [LR, SVR]	n.a
(Mifsud *et al.*, 2020)	MAE (normalized), MSE (normalized), and percentage error	Test set	[LR, ANN] > [SVR, RF]	The performance metrics were calculated from the estimated power output
(Schwegmann *et al.*, 2023)	MAE, R^2 , RMSE	Test set	K-NN > [LR, RF, ANN, Gaussian process]	n.a

Table 2.4 Evaluation metrics used in the review studies to compare different prediction models

Metric	Formula	Perfect score
Bias	$\frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)$	0
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert p_i - o_i\rvert$	0
MARE	$\frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{o_i - p_i}{o_i}\right\rvert$	0
MSE	$\frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2$	0
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2}$	0
R^2	$1 - \frac{\sum_{i=1}^{n}(o_i - p_i)^2}{\sum_{i=1}^{n}(o_i - \bar{o})^2}$	1
$\boldsymbol{Err_{mean}}$	$\frac{\bar{p} - \bar{o}}{\bar{o}}$	0

$\boldsymbol{Err_{var}}$	$\frac{\sum_{i=1}^{n}(p_i - \bar{p})^2 - \sum_{i=1}^{n}(o_i - \bar{o})^2}{\sum_{i=1}^{n}(o_i - \bar{o})^2}$	0

Note: o represents the observed variable, p the predicted variable, n the sample size

2.4 Long-term wind speed estimation at unsampled locations

Figure 2.3 presents an overview of important factors when estimating long-term wind speeds at USL. These key factors are covered in detail in the following subsections: a review of the prediction time scales identified in the literature, a presentation of commonly used explanatory variables, the estimation methods, and comparative studies.

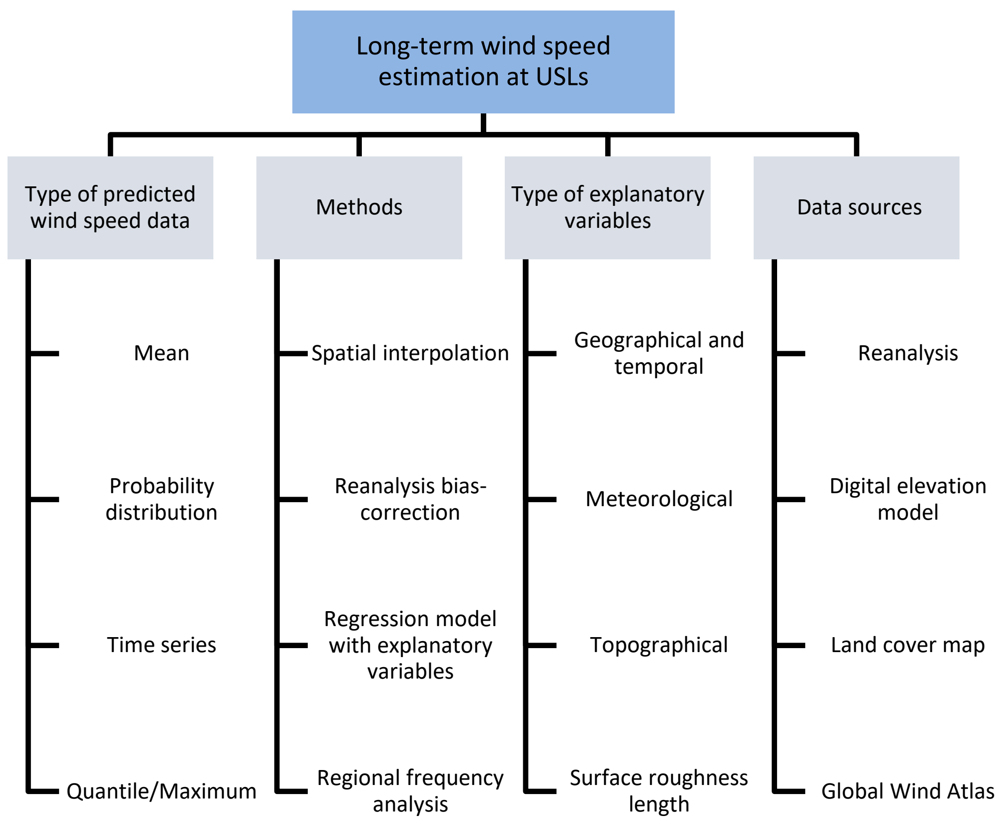


Figure 2.3 Important factors to consider when estimating long-term wind speeds at USLs

2.4.1 Prediction time scale

Wind speed is a continuous variable, and for practical reasons, researchers often work with aggregated values over specified time scales. The choice of the time scale and aggregation function depends on the study's objectives. In the literature, the time scales for predicted wind speeds at USLs range from hourly means (Cellura et al., 2008a; Cellura et al., 2008b; Cirrincione et al., 2009) to annual maximum values (Modarres, 2008). Some authors have predicted monthly

(Fadare, 2010) and daily (Jung et al., 2023a) mean wind speeds. The mean is the most frequently calculated statistic when aggregating wind speed time series. In studies on EWS, researchers have used wind speed quantiles (Etienne et al., 2010; Fischer et al., 2015; Jung et al., 2022a) and maximum wind speeds (Goel et al., 2004). The 50-year return period wind speed is an important design parameter for ensuring the safety of wind turbines (Palutikof et al., 1999).

Effenberger et al. (2024) examined how different temporal resolutions and aggregation methods (average vs. instantaneous) affect wind speed distribution compared to instantaneous data (e.g., 10-minute averages). Their findings indicated that using three- or six-hourly instantaneous values better preserved the observed 10-minute wind speed distribution than averaged values over a specified time scale. Petersen et al. (1981) discussed the implications of various averaging time scales and the resulting variance loss for WRA. They highlighted that variance loss is significant when averaging wind speed data over periods longer than a few hours.

High temporal resolution wind speed data is essential for understanding resource variability for WRA (Pfenninger et al., 2014). However, processing and storing large volumes of wind speed data can present some challenges. Mapping wind speed distributions can help summarize the data while retaining important information about variability (Houndekindo et al., 2023b; Jung, 2016; Veronesi et al., 2016).

2.4.1 Explanatory variables

In the literature, five categories of explanatory variables have been used for wind speed estimation at USLs: geographical variables (e.g., longitude, latitude), temporal variables (e.g., time of day, month of the year), meteorological variables (e.g., air temperature, atmospheric pressure, humidity), topographical variables (e.g., elevation, exposure) and surface roughness length (SRL).

Geographical and temporal variables. In spatial interpolation (SI) techniques, geographical coordinates are the primary variable for estimating wind speed at USLs. In the absence of complex terrain features, a high correlation is expected between neighboring sites, especially when aligned with prevailing wind direction, and this correlation should decrease with distance. With data-driven feature selection, Robert et al. (2013) identified geographical coordinates as important predictors of wind speed. Additionally, geographical coordinates could serve as proxies for other factors affecting wind speed, such as altitude or proximity to the sea. For example, higher latitudes (or longitudes) may correspond to higher (or lower) altitudes or proximity to the sea. Proximity to the sea significantly impacts the local wind climate through mechanisms such as sea and land breezes. Additionally, low surface friction over the sea typically results in stronger winds near coastal regions (Brinckmann et al., 2016). Houndekindo et al. (2023a) reported that the distance from the coast was important for mapping various wind speed quantiles across Canada.

Wind speed also exhibits diurnal, seasonal, and interannual variations. To account for these temporal variations, researchers have included time-related variables in their models (Fadare, 2010; Jung et al., 2022a) when predicting WSTS.

Meteorological variables. Wind speed is primarily driven by the movement of air resulting from pressure differences, which are caused mainly by variations in solar irradiation (Emeis, 2018). Meteorological variables, such as temperature and pressure, are typically well correlated with wind speed through underlying physical processes (Şahin et al., 2006), but these variables are also missing at USLs. Reanalysis and regional climate model (RCM) data serve as alternative sources of meteorological variables for wind speed estimation at USLs. Inaccuracies in reanalysis wind data are partly due to the limited ability of NWP models to accurately capture the interactions between large-scale winds (e.g., geostrophic winds) and local topographic features and surface roughness. However, reanalysis wind fields are generally considered representative of large-scale wind patterns, unaffected by surface friction (González-Aparicio et al., 2017; Houndekindo et al., 2024; Hu et al., 2023; Jung et al., 2020; Kirchmeier et al., 2014).

Topographic variables and surface roughness. Local topography influences near-surface wind speed through topography-induced acceleration, deceleration, and deflection (Emeis, 2018; Petersen et al., 1998; Raupach et al., 1997). Elevation is the most commonly used explanatory variable in estimating wind speed (Collados-Lara et al., 2022; Fick et al., 2017). However, elevation alone only captures part of the local topography's effect on near-surface wind speeds. The availability of high-resolution digital elevation models (DEM) allows for the extraction of additional topographic variables that more accurately describe terrain complexity, such as the topographic position index, which measures terrain ruggedness and topographic wind exposure, such as the maximum upwind slope (Winstral et al., 2017). These variables can be derived at multiple spatial scales, reflecting the varying degrees of influence that topographic features exert over different distances (Dujardin et al., 2022; Etienne et al., 2010; Foresti et al., 2011; Houndekindo et al., 2023a; Houndekindo et al., 2023b; Jung, 2016; Jung et al., 2018a; Jung et al., 2020; Jung et al., 2023b; Robert et al., 2013; Winstral et al., 2017).

For example, in Switzerland, Etienne et al. (2010) identified elevation from a 1 km resolution DEM and landform classes from a 2 km resolution DEM as significant predictors of wind speed at USLs. Similarly, in Germany, Jung (2016) found that a location’s relative elevation at a spatial scale of

2.5 km was the most important predictor of wind speed. The Relative elevation is the difference between the location's elevation and the mean elevation of surrounding cells within a specified radius. Recent advancements in ML, particularly Convolutional Neural Networks (CNNs), have shown promise in automatically extracting multiscale topographic features from high-resolution DEMs (Dujardin et al., 2022; Zhong et al., 2024).

Another crucial variable in wind speed estimation at USLs is the SRL. SRL values are typically derived from land cover maps (Davis et al., 2023; Etienne et al., 2010; Houndekindo et al., 2023a; Houndekindo et al., 2023b; Jung, 2016; Jung et al., 2018a; Jung et al., 2020; Jung et al., 2023b) by associating different land cover types (e.g., vegetation, urban infrastructure) with corresponding SRL values based on established empirical relationships (Wiernga, 1993). These SRL values reflect the influence of the terrain surface and its roughness elements on the frictional forces acting on wind near the ground. Higher SRL values typically indicate rougher surfaces, slowing wind speeds down (Petersen et al., 1998). For instance, vegetated areas such as forests typically exhibit higher SRL values, leading to more pronounced wind attenuation than smoother surfaces like open water or flat, barren land.

2.4.1 Estimation methods at unsampled locations

The methods for estimating wind speed at USLs can be categorized based on the type of predicted data:

1. Summary wind speed statistics, typically representing central tendencies (e.g., MWS), have been mapped using SI or regression models with explanatory variables.
2. Some studies have estimated the WSPD at USLs to capture the wind resource variability. Both parametric and nonparametric methods have been proposed for this purpose.
3. Bias-corrected and uncorrected reanalysis wind data are frequently used to obtain wind WSTS at USLs. Also, SI (Zhang *et al.*, 2022b), regression (Robert *et al.*, 2013), and hybrid models (Collados-Lara *et al.*, 2022) have been explored to interpolate WSTS at USLs.
4. EWS data are necessary for evaluating wind turbine suitability in a region (Pryor *et al.*, 2021). Uncorrected reanalysis wind data, regression models with explanatory variables, regional frequency analysis (RFA), and SI have been applied to estimate EWS at USLs.

Table 2.5 provides an overview of the various methods used for wind speed estimation at USLs, which are reviewed in detail in the following subsections.

Table 2.5 Summary of the different methods for wind speed estimation at USLs

Type of predicted wind speed data	Method/model	Study
MWS	Hybrid (SI and regression)	(Cellura *et al.*, 2008a)
	SI	(Baydaroğlu *et al.*, 2019; Fick *et al.*, 2017; Keskin *et al.*, 2017; Lee, 2022; Luo *et al.*, 2008; Sliz-Szkliniarz *et al.*, 2011; Van Ackere *et al.*, 2015)
WSPD	Parametric	(Jung, 2016; Laib *et al.*, 2016; Veronesi *et al.*, 2016)
		(Jung *et al.*, 2020; Jung *et al.*, 2023b)
	Nonparametric	(Houndekindo *et al.*, 2023b)
WSTS	Reanalysis wind speed mean bias correction	(Bosch *et al.*, 2018; González-Aparicio *et al.*, 2017; Gruber *et al.*, 2019; Gruber *et al.*, 2022; Murcia *et al.*, 2022; Nefabas *et al.*, 2021; Ryberg *et al.*, 2019; Schicker *et al.*, 2023)
	Quantile mapping (QM) with reanalysis wind speed	(Jung *et al.*, 2023b)
	Reanalysis wind speed bias correction with ML	(Dujardin *et al.*, 2022; Hu *et al.*, 2023)
	SI	(Zhang *et al.*, 2022b)
	SI and Hybrid (SI and regression)	(Brinckmann *et al.*, 2016; Collados-Lara *et al.*, 2022; Li *et al.*, 2014; Reinhardt *et al.*, 2018; Zhao *et al.*, 2022)
	Regression	(Philippopoulos *et al.*, 2012; Robert *et al.*, 2013)
EWS	Regression model with explanatory variables	(Etienne *et al.*, 2010; Jung *et al.*, 2022a)
	SI	(Ye *et al.*, 2015)
	RFA	(Ahmad *et al.*, 2024; Campos *et al.*, 2018; Fawad *et al.*, 2018; Goel *et al.*, 2004; Hong *et al.*, 2014)

2.4.1 Estimation methods at unsampled locations

Long-term MWS has been mapped across large regions for WRA at USLs using SI as the primary modeling approach (Baydaroğlu *et al.*, 2019; Cellura *et al.*, 2008a; Van Ackere *et al.*, 2015). Studies have demonstrated that incorporating anisotropy (Lee, 2022) and additional explanatory variables should improve model performance, with elevation being the most frequently used explanatory variable (Keskin *et al.*, 2017; Luo *et al.*, 2008; Sliz-Szkliniarz *et al.*, 2011).

Cellura *et al.* (2008a) proposed a hybrid approach combining MLP and SI models for MWS mapping. An MLP was used to establish a regression function between explanatory variables (e.g., geographical coordinates and elevation) and the observed MWS in Sicily, Italy. Kriging was then applied to the regression model residuals to refine the predictions. The model can be expressed as:

Equation 2.6

$$\widehat{W}(s_0) = f\big(x(s_0), y(s_0), z(s_0)\big) + \sum_{i=1}^{k} \lambda_i e(s_i)$$

where $\widehat{W}(s_0)$ is the predicted wind speed at location s_0, $x(s_0), y(s_0)$, and $z(s_0)$ are the location's longitude, latitude, and elevation, respectively. The function $f(\cdot)$ is the MLP-derived regression function, $e(s_i)$ represents the residuals at sampled locations s_i, and λ_i are the kriging weights.

2.4.1.1 Estimation of wind speed distribution

Relying exclusively on MWS to evaluate long-term energy potential may lead to an underestimation of the available resource's potential (Nelson *et al.*, 2018). To address this limitation, it is crucial to estimate the WSPD at USLs. This approach allows for a more accurate evaluation of the resource's long-term variability.

Some studies have been conducted to map WSPD across large regions (Houndekindo *et al.*, 2023b; Jung, 2016; Jung *et al.*, 2020; Jung *et al.*, 2023b; Laib *et al.*, 2016; Veronesi *et al.*, 2016). The statistical methods employed in these studies can broadly be categorized into parametric and nonparametric methods.

In the parametric method, the parameters of the WSPD are estimated at USLs using a regression model along with various explanatory variables. The Weibull distribution has been employed due to its widespread adoption in the field (Veronesi *et al.*, 2016). Some studies have explored the use of more flexible probability distributions. For example, Jung *et al.* (2020) proposed a parametric method that involves mapping L-moments of wind speed. They then used the L-moment method to estimate the parameters of the four-parameter Kappa and five-parameter Wakeby distributions at USLs.

L-moments introduce by Hosking (1990), serve as alternatives to traditional moments (such as mean and variance) for summarizing the shape of a probability distribution. As linear combinations of order statistics, L-moments exhibit more robustness to outliers than conventional sample moments when estimating from finite samples (Hosking, 1990). Similar to the method of moments, which is used to derive estimators of probability distribution parameters from sample moments, the L-moment method can also be used to estimate these parameters from sample L-moments. The sample L-Moments are given by (Hosking, 1990):

Equation 2.7

$$\hat{\lambda}_r = \frac{1}{r}\sum_{i=1}^{n}\left[\frac{\sum_{j=0}^{r-1}(-1)^j\binom{r-1}{j}\binom{i-1}{r-1-j}\binom{n-1}{j}}{\binom{n}{r}}\right]X_{i:n}$$

where r is the L-moment order and $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$ denote the sample order statistics.

In another study, Houndekindo *et al.* (2023b) proposed a nonparametric method for WSPD mapping. This method does not rely on the assumption of a specific parametric distribution, thus offering a more flexible and data-driven approach for estimating WSPD at USLs. The nonparametric approach uses a regression model to map multiple wind speed quantiles across a region. Subsequently, kernel estimators are fitted to the estimated quantiles at USLs to obtain the entire WSPD. This approach was recommended for large regions where a single family of probability distributions may fail to adequately represent the region's complex and diverse wind speed patterns. To address boundary effects that appear when using symmetric kernels for modeling bounded random variables like wind speeds, Houndekindo *et al.* (2023b) adopted asymmetric kernel estimators. The Birnbaum-Saunders (Mombeni *et al.*, 2021) and Log-Normal kernels (Lafaye de Micheaux *et al.*, 2021) provided a better fit among the evaluated kernels. The CDFs of the Log-Normal ($\hat{F}_{LN}(\cdot)$) and Birnbaum-Saunders ($\hat{F}_{BS}(\cdot)$) kernels are expressed as follows:

Equation 2.8

$$\hat{F}_{LN}(x) = \frac{1}{n}\sum_{i=1}^{n}K_{LN}\left(X_i; \log x\,, \sqrt{b}\right)$$

Equation 2.9

$$\hat{F}_{BS}(x) = \frac{1}{n}\sum\nolimits_{i=1}^{n} K_{BS}\left(X_i; x, \sqrt{b}\right)$$

where $b > 0$ is the kernel bandwidth, and the kernel functions are defined as:

Equation 2.10

$$K_{LN}\ (x; \mu\ , \sigma) = 1 - \ \Phi\left(\frac{\log x - \mu}{\sigma}\right)$$

Equation 2.11

$$K_{BS}\ (x;\ \beta, \sigma) = 1 - \ \Phi\left(\frac{1}{\sigma}\left(\sqrt{\frac{x}{\beta}} - \sqrt{\frac{\beta}{x}}\right)\right)$$

The function $\Phi(\cdot)$ denotes the CDF of the standard normal distribution.

2.4.1.2 Estimation of wind speed time series

The availability of WSPD at USLs provides valuable insights into the variability of the resource. However, to fully assess wind speed variability across different temporal scales, such as diurnal, seasonal, and inter-annual variability, time series data are essential.

To generate gridded WSTS, Zhang *et al.* (2022b) used a three-dimensional thin plate smoothing spline with longitude, latitude, and elevation as explanatory variables to interpolate observed daily wind speeds at a spatial resolution of 0.05 × 0.05° across Australia. In a similar study, Li *et al.* (2014) interpolated six-hourly WSTS in China at a spatial resolution of 1 km × 1 km. they used a two-dimensional thin-plate smoothing spine combined with kriging applied to the residuals. The authors further examined the effect of including reanalysis wind speed data as an additional explanatory variable but found no significant reduction in the cross-validation RMSE.

Collados-Lara *et al.* (2022) compared various kriging methods for hourly WSTS interpolation at 300 m × 300 m spatial resolution in Granada, Spain. Their results demonstrated that regression kriging outperformed ordinary kriging, kriging with external drift, and a simple regression model (without kriging of the residuals). In addition, the authors observed that the interpolated WSTS exhibited lower RMSE values compared to wind speeds derived from the European Center for Medium-Range Weather Forecasts Reanalysis v5 (ERA5; Hersbach *et al.*, 2020).

A key limitation of SI methods for estimating wind speeds at USL is their dependence on a dense network of meteorological stations in the region of interest. In regions with sparse station coverage, the accuracy of interpolated wind speeds tends to decline due to low station density (Fick *et al.*, 2017). Additionally, terrain complexity presents further challenges to interpolation accuracy. In areas with complex topography, standard interpolation methods that rely solely on geographical coordinates and elevation often fail to capture the spatial variability of wind speeds, leading to significant estimation errors.

To address these challenges, Robert *et al.* (2013) incorporated additional topographic explanatory variables, such as terrain convexity, slope, and wind exposure, to improve wind speed estimation in Switzerland's complex Alpine orography. The authors used the General regression neural network (GRNN) to model the complex, non-linear relationship between the explanatory variables and monthly wind speed values.

Recent advancements in data assimilation techniques and model physics (Hersbach *et al.*, 2020) have significantly improved the accuracy and spatial resolution of reanalysis wind speed data. Therefore, many recent studies have relied on reanalysis datasets to obtain WSTS at USLs (Ayik *et al.*, 2021; Davidson *et al.*, 2022; de Aquino Ferreira *et al.*, 2022; Gualtieri, 2021; Gualtieri, 2022; Jourdier, 2020; Olauson, 2018; Rabbani *et al.*, 2020; Ramon *et al.*, 2019; Thomas *et al.*, 2021).

Reanalysis datasets, such as ERA5 and the Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2; Gelaro *et al.*, 2017), provide global coverage of meteorological variables at hourly intervals over extended periods. However, their relative coarse spatial resolution can be a limiting factor for studies that require fine-scale analysis, particularly in regions with complex terrain where wind patterns can vary considerably.

To address this limitation, efforts have been made to downscale and bias-correct reanalysis wind speeds, enhancing their spatial resolution and addressing the discrepancies between reanalysis outputs and local measurements. The bias correction (BC) of reanalysis wind speeds has primarily relied on the Global Wind Atlas (GWA; Davis *et al.*, 2023), which provides quasi-global static wind conditions (e.g., MWS, Weibull scale, and shape parameters) integrating microscale terrain and land cover features.

González-Aparicio *et al.* (2017) applied a parametric QM method to correct the distribution of MERRA-2 wind speeds using the following equation:

Equation 2.12

$$\hat{y}_t = \alpha^{GWA} \left(\frac{y_t^{MERRA}}{\alpha^{MERRA}} \right)^{\frac{k^{MERRA}}{k^{GWA}}}$$

where $\hat{y}_t$ is the corrected wind speed at time t, α^{GWA} and k^{GWA} are the Weibull scale and shape parameters from the GWA, and α^{MERRA}, k^{MERRA}, and y_t^{MERRA} are the corresponding parameters and wind speed value from MERRA-2.

Other studies (Gruber *et al.*, 2019; Gruber *et al.*, 2022; Houndekindo *et al.*, 2024; Langer *et al.*, 2023; Murcia *et al.*, 2022; Ryberg *et al.*, 2019) used linear scaling for BC, where a multiplicative factor is applied to the reanalysis wind speeds:

Equation 2.13

$$\hat{y}_t = \frac{\overline{y^{GWA}}}{\overline{y^{MERRA}}} y_t^{MERRA}$$

where $\overline{y^{GWA}}$ and $\overline{y^{MERRA}}$ are the MWS from the GWA and MERRA, respectively.

Figure 2.4 shows an example of the linear scaling method applied to bias correct ERA5 wind speed data at a station in Canada.

Typically, a time-invariant scaling factor was applied, which improved the long-term MWS accuracy without improving the temporal properties of the reanalysis time series. Schicker *et al.* (2023) attempted to address this limitation by estimating hourly scaling factors to account for diurnal biases, but this approach degraded the performance of the uncorrected reanalysis data.

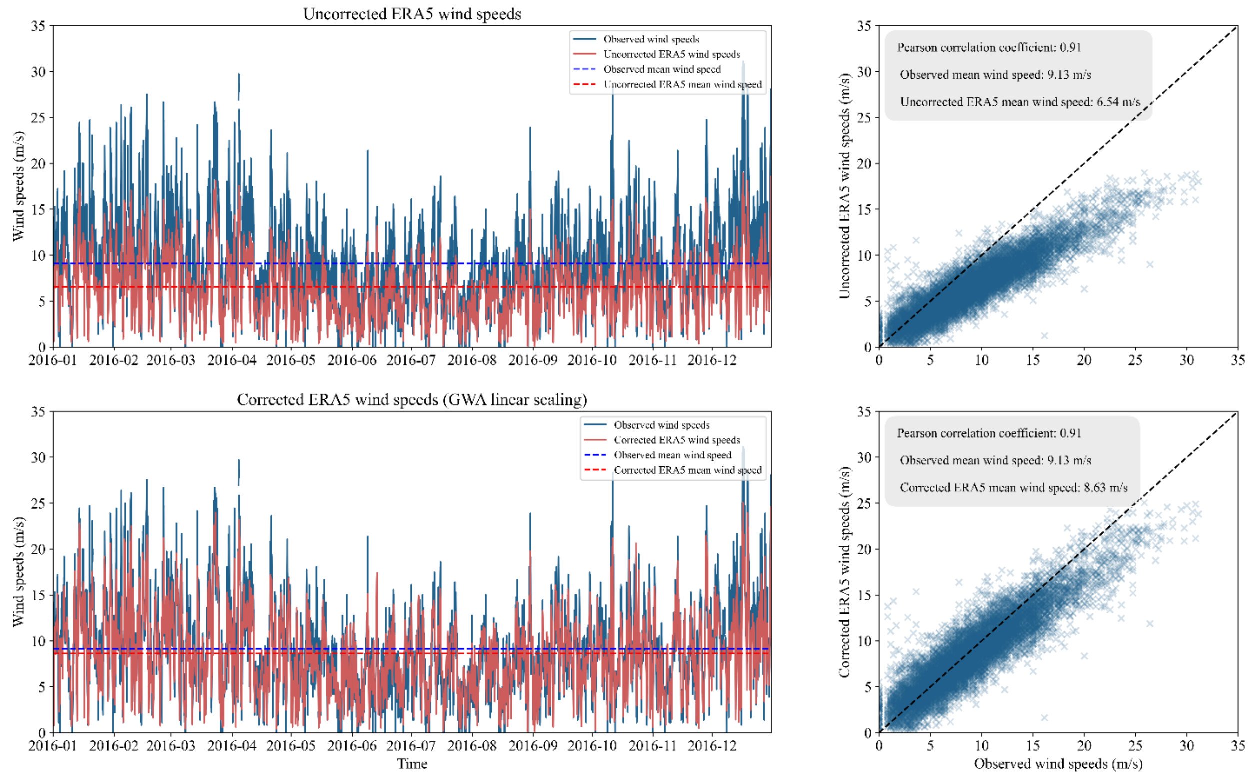


Figure 2.4 **Example of application of the linear scaling method for BC of ERA5 wind speeds using the GWA version 3.**

The measured hourly wind speed data were obtained from Environment and Climate Change Canada (ECCC) historical climate data archives (Climate ID: 8403255, period: 2016).

The GWA is commonly used for BC due to its incorporation of microscale topographic and land cover features, which are not fully resolved by reanalysis data. Alternatively, these features can be directly obtained from high-resolution DEM and land cover maps and used as additional explanatory variables to establish a regression function between reanalysis and observed wind speeds. This method offers greater flexibility in the spatial resolution of the downscaled wind speeds compared to the fixed resolution of 250 m × 250 m expected when bias-correcting with the GWA. For example, Hu *et al.* (2023) developed a regression model that integrated topographic and ERA5 meteorological variables to predict observed wind speeds. This model improved ERA5 accuracy in areas with complex topography, though it had a less pronounced impact in areas where ERA5 accuracy was already high.

Jung *et al.* (2023b) applied QM to correct ERA5 wind speeds. Their approach, illustrated in Figure 2.5, involved the following steps: First, they established a regression model to map wind speed L-moments in the study region using local topographic features, surface roughness length, and L-moments from ERA5 wind speeds. Then, the trained regression model was used to predict L-moments at USLs. Using the L-moment method, these predicted L-moments were used to estimate the four parameters of the Kappa distribution. Finally, the quantile function of the fitted Kappa distribution was used for QM. The QM equation is given by:

Equation 2.14

$$\hat{y}_t = Q_{\hat{\theta}}\left[F_n\left(y_t^{ERA5}\right)\right]$$

where $Q_{\hat{\theta}}(\cdot)$ is the quantile function of the Kappa distribution with parameters $\hat{\theta}$ estimated using the L-moment method, and $F_n(\cdot)$ is the empirical CDF of the ERA5 wind speed data.

Dujardin *et al.* (2022) proposed a novel approach that uses CNNs to interpolate hourly wind speeds in Switzerland. Their model operates at the grid level, using a CNN to predict observed wind components at meteorological stations. The input data for the CNN consist of patches of gridded meteorological and topographic variables centered on each station.

The results from the study indicated that the proposed model improved both the correlation and bias of COSMO-1 (Kruyt *et al.*, 2018) outputs when compared to measured wind speeds. Although the CNN architecture was not directly compared to standard regression models, it is expected to outperform traditional methods as it can effectively process multiscale spatial information related to topography and meteorological variables. For example, in standard regression models, gridded meteorological variables are typically interpolated at station locations using rigid methods, such as nearest neighbor or bilinear interpolation. In contrast, the CNN architecture employs a data-driven approach that interpolates gridded meteorological variables at station locations. Also, this method can potentially consider a broader spatial extent of these variables, which may enhance the accuracy of bias-corrected wind speed predictions. The CNN method presents a promising alternative for wind speed modeling in complex terrains.

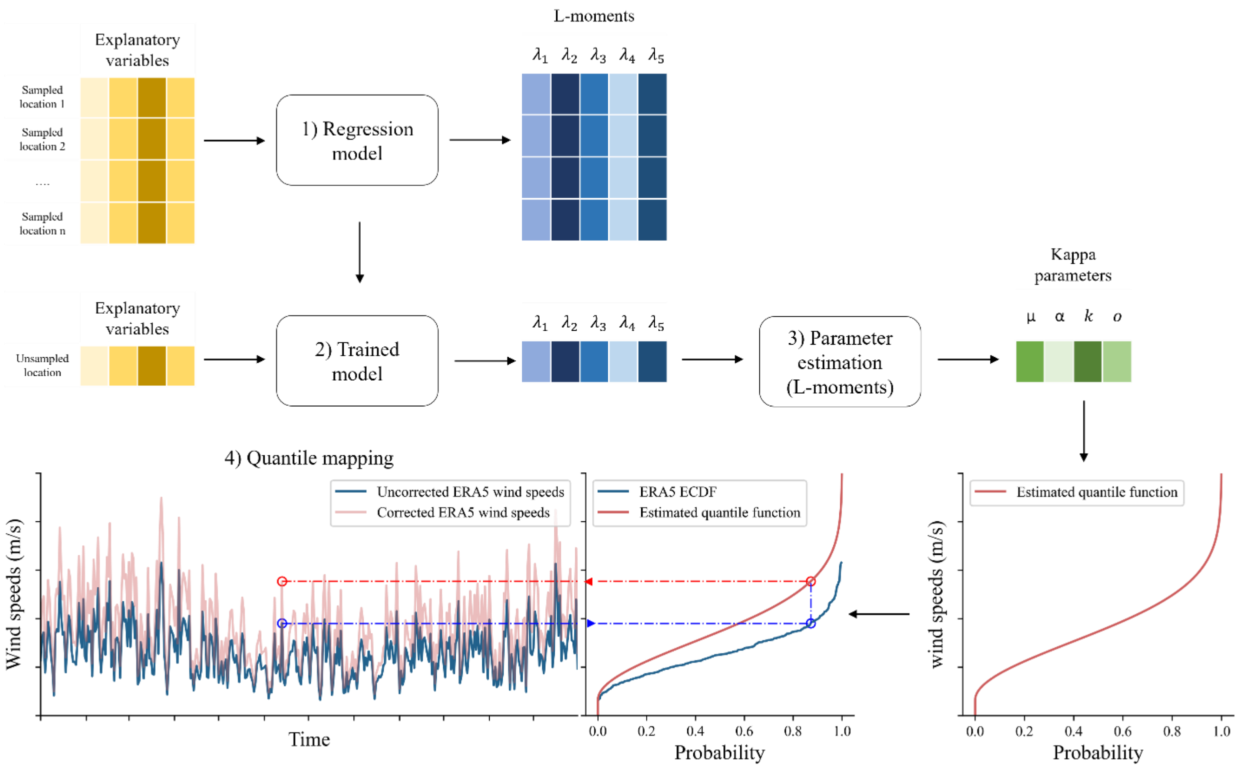


Figure 2.5 Illustration of the different steps of the QM process for ERA5 wind speeds' BC at USLs. This BC method was proposed by [121].

2.4.1.3 Extreme wind speed estimation

EWSs are evaluated to determine the suitability of specific wind turbines in a given region. This assessment considers factors such as structural integrity, safety, and performance in challenging weather conditions. It ensures that wind turbines are appropriately designed and located to withstand the severe wind conditions they may encounter during their operational lifespan.

Reanalysis wind speeds have been used to estimate EWSs on a global scale (Jung *et al.*, 2017c; Pryor *et al.*, 2021). However, researchers have employed alternative methods to estimate EWSs with higher spatial resolutions at national (Etienne *et al.*, 2010; Ye *et al.*, 2015) and continental scales (Jung *et al.*, 2022a).

For example, Etienne *et al.* (2010) mapped the 98th percentile of daily maximum wind speed across Switzerland. They used a generalized additive model (Hastie, 2017), which is a nonlinear regression model, to capture the relationship between topographic variables and observed EWS quantiles. Similarly, Jung *et al.* (2022a) developed a regression model to map the 90th percentile of monthly wind speed across North America and Europe. This model used the ERA5 90th percentile of monthly wind speeds, MWS from a global model (Jung *et al.*, 2020), along with geographical coordinates and the month of the year as input.

In other studies, RFA was applied to predict EWS at USLs. The RFA procedure involves two main steps, as shown in Figure 2.6: 1) formation of homogeneous regions, and 2) information transfer within the regions. Establishing homogeneous regions based on the wind generation mechanisms is essential for ensuring that information can be reliably transferred within the region. There are some statistical tools to evaluate the homogeneity of the regions. For EWS, studies have primarily relied on the statistical tests proposed by Hosking *et al.* (1997).

The "index flood" method (Hosking *et al.*, 1997), based on L-Moments, was the most commonly used RFA procedure, which has been adapted and referred to as the "local-index" method for EWS modeling (Campos *et al.*, 2018). The primary assumption of this method is that all sites within a homogenous region share the same wind speed distribution (the regional distribution), except for a site-specific scaling factor. This scaling factor, known as the local index, represents a statistical measure of the central tendency (e.g., mean, median) of EWS at each location within the region. The local index is usually estimated with a regression model using some explanatory variables (e.g., topographic and land cover features).

Determining a region’s adequate size presents some challenges. Generally, as the region’s size increases, its homogeneity tends to decrease. In contrast, a smaller region (limited sample size) may produce unreliable local index estimates using a regression model. For estimating EWS, Goel *et al.* (2004) recommended that a region include at least 5T station years of data, where T represents the return period of interest. A similar recommendation was provided for regional flood frequency analysis (Jakob *et al.*, 1999).

Another important factor to consider when forming regions is their identifiability (Burn *et al.*, 2000; Goel *et al.*, 2004). Identifiability refers to the ability to assign a new USL to a region based solely on specific explanatory variables independently of the target variable. Therefore, regions should not be defined using variables derived from observed wind speed data, as this would prevent the assignment of USLs to that region (Campos *et al.*, 2018; Hong *et al.*, 2014). However, it is acceptable to form regions using variables derived from measured wind speed data if the RFA is applied at PSLs to improve the estimation of EWS.

Several methods for region formation were used in the literature. Goel *et al.* (2004) and Fawad *et al.* (2018) used a pooling method based on geographical proximity. Campos *et al.* (2018) formed homogeneous regions using similarity in L-moment ratios estimated from observed wind speed data. Alternatively, Hong *et al.* (2014) used a clustering algorithm with some explanatory variables to form homogeneous regions.

Most studies have relied on a parametric distribution function to estimate EWS from the available wind speed time series. Table 2.6 summarizes the various parametric distribution functions used to estimate EWS at USLs.

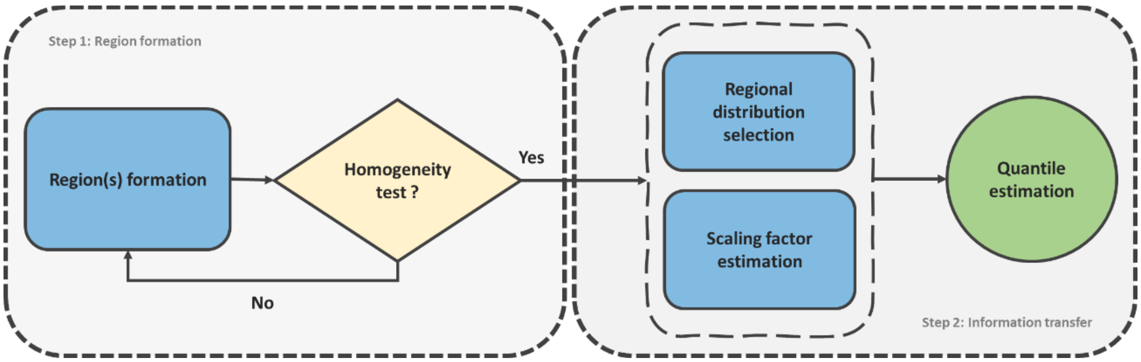


Figure 2.6 The different steps of the RFA

Table 2.6 Parametric distribution functions used for EWS estimation at USLs

Study	Selected parametric distribution function	Type of EWS
Ahmad *et al.* (2024)	Generalized logistic distribution	Daily Annual Maximum Wind Speed quantiles (2, 5, 10, 20, 50, 100, 500, 1000-year return periods)
Campos *et al.* (2018)	Weibull distribution (Three parameters)	Annual maximum wind speed quantiles (5, 10, 20-year return periods)
Fawad *et al.* (2018)	Generalized logistic distribution	annual maximum wind speed quantiles (2, 5, 10, 20, 50, 100, 500, 1000-year return periods)
Goel *et al.* (2004)	Generalized logistic distribution and generalized extreme value distribution	Annual maximum wind speed quantiles (10, 20, 50, 100, 500-year return periods)
Hong *et al.* (2014)	Generalized extreme value distribution	annual maximum wind speed quantiles (50, 500, 1000-year return periods)
Jung *et al.* (2017c)	Wakeby distribution	wind gust speed quantiles (30, 50 and 100 year30-, 50- and 100-year return period)
Jung *et al.* (2022a)	Generalized extreme value distribution and Wakey distribution	Monthly annual maximum wind speed quantiles (10-year return period)
Pryor *et al.* (2021)	Gumbel distribution	Wind speed quantiles (50-year return period)

2.4.1.4 Comparative studies at unsampled locations

Several studies have compared SI methods for mapping wind speed across large regions (Berndt *et al.*, 2018; Collados-Lara *et al.*, 2022; Lee, 2022; Luo *et al.*, 2008; Reinhardt *et al.*, 2018; Van Ackere *et al.*, 2015). The findings from these studies indicate that there is no clear choice between deterministic methods and geostatistical methods for SI of wind speeds. The relative performance of these methods varies based on several factors, such as the magnitude of the interpolated wind speed (e.g., EWS vs. MWS) and the specific region of study. Incorporating elevation in the kriging models (e.g., co-kriging) does not consistently improve its accuracy (Ye *et al.*, 2015). In some studies, regression kriging models outperformed standard SI methods (Collados-Lara *et al.*, 2022; Lee, 2022; Reinhardt *et al.*, 2018).

Some studies have reported that MLP models outperform SI methods for wind speed mapping (Cellura *et al.*, 2008a; Öztopal, 2006; Philippopoulos *et al.*, 2012). In contrast, Reinhardt *et al.* (2018) found that regression kriging with LR was more accurate than MLP and SVR for mapping wind speed components at various heights.

Houndekindo *et al.* (2023b) compared LR and GB for mapping various wind speed quantiles. Their results demonstrated that GB was more effective than LR in capturing the relationship between the explanatory variables and the wind speed quantiles. Similarly, Jung *et al.* (2020) reported that GB outperformed LR for mapping wind speed L-moments on a global scale. In their research, the authors evaluated a variant of the GB algorithm against LR, RF, SVR, and Gaussian process regression, concluding that GB exhibited the highest level of accuracy.

2.5 Long-term wind speed estimation under nonstationary conditions

The literature reports three causes of non-stationarity in WSTS. Several studies conducted in various countries have reported a declining trend in near-surface wind speeds (Azorin-Molina *et al.*, 2014; Cui *et al.*, 2018; Klink, 2002; McVicar *et al.*, 2008; Minola *et al.*, 2016; Vautard *et al.*, 2010a; Wan *et al.*, 2010; Zahradníček *et al.*, 2019). Roderick *et al.* (2007) introduced the term "stilling" to describe the observed trend. One of the identified causes of the stilling phenomenon is an increase in surface roughness (Klink, 2002; Pryor *et al.*, 2009; Vautard *et al.*, 2010b). However, Zeng *et al.* (2019) observed a reversal of the stilling trend by 2010 in a global study, concluding that the previously reported decreasing trend could be attributed to interannual, decadal ocean-atmosphere oscillations. Several authors have also studied the interannual variability of wind speeds due to ocean-atmosphere oscillations (Azorin-Molina *et al.*, 2016; Azorin-Molina *et al.*, 2018; Klink, 2007; Ouarda *et al.*, 2021; Pryor *et al.*, 2006a; Woldesellasse *et al.*, 2020).

Climate change is another significant factor contributing to the long-term variability of wind resources (Jung *et al.*, 2022c; Pryor *et al.*, 2010). Several studies have examined the impact of climate change on wind speeds (Bloom *et al.*, 2008; Jung *et al.*, 2019a; Pryor *et al.*, 2011; Pryor *et al.*, 2012; Pryor *et al.*, 2006b; Sailor *et al.*, 2008). Jung *et al.* (2022c) provided a comprehensive review of studies on the impact of climate change on wind energy.

Non-stationarity in WSTS can undermine the reliability of wind farm projects. Evaluating the resource's long-term variability during the project's initial phase can provide the data needed to build a more resilient energy system.

To evaluate the impact of climate change on wind resources, researchers primarily used Global climate models (GCM) and RCMs (Jung *et al.*, 2022c; Pryor *et al.*, 2010). GCM experiments are conducted globally to understand Earth's climate system dynamics and to project future climate conditions under various scenarios (Eyring *et al.*, 2016). However, one limitation of GCMs is their coarse spatial resolution, which limits their ability to represent mesoscale and microscale wind patterns (Shen *et al.*, 2022).

To address this limitation, downscaling techniques are often employed to refine the coarse-resolution output of GCMs into a finer spatial scale that is more relevant to specific regions of interest. RCM simulations, which represent the dynamical downscaling of GCMs, operate at a higher spatial resolution, enabling a more detailed representation of regional climate features. By nesting RCMs within GCMs, researchers capture finer-scale processes that influence local wind patterns, thus improving the accuracy of climate projections at the local levels (Tobin *et al.*, 2016).

As an alternative to RCM, statistical downscaling methods are also employed to bridge the gap between coarse-scale GCM outputs and local wind conditions. These methods involve establishing a relationship between the historical simulation outputs of GCMs and measured data (Schoof, 2013). The main advantage of using statistical downscaling over RCMs is its computational efficiency, making it particularly useful for large-scale studies. In some studies (Bartók *et al.*, 2019; Costoya *et al.*, 2020; Li *et al.*, 2020; Luzia *et al.*, 2023; Moemken *et al.*, 2018; Nabipour *et al.*, 2020; Vu Dinh *et al.*, 2022), statistical downscaling was applied as an additional step to post-process RCM simulations. However, a significant limitation of statistical downscaling is its reliance on long-term measured data for model training and validation (Jeong *et al.*, 2012).

At PSLs, MCP can be applied to extend the short-term wind speed records, which can serve as the measured data for statistical downscaling. Alternatively, outputs from GCM or RCM can be adjusted using wind speed data from the reference site and further corrected using MCP at the PSL. For example, Rajczak *et al.* (2016) proposed a two-step statistical approach for bias-correcting RCM outputs at PSLs. The RCM output is corrected at the reference site in the first step, followed by a subsequent BC at the PSL. For both BC steps, the authors used QM. This method relies on the availability of a meteorological station (reference site) with long-term observation data and a stationary relationship with the PSL, which is not guaranteed.

At USLs, reanalysis datasets often serve as the primary source of historical ‘quasi-observational’ data for statistical downscaling (Jung *et al.*, 2022c). These datasets provide consistent records of past weather conditions spanning several decades and typically offer finer spatial resolution than GCM simulations. However, it is important to recognize the potential limitations of reanalysis data as a substitute for observational data, primarily due to biases introduced during the assimilation process and the uncertainties inherent to the modeling techniques and assumptions (Gualtieri, 2022). To mitigate the uncertainties associated with individual datasets, it has been recommended to utilize multiple reanalysis datasets (Torralba *et al.*, 2017). In addition, Moradian *et al.* (2023) validated the reliability of the reanalysis data in their study region by comparing it with available observational data before conducting further analysis at USLs.

Using reanalysis datasets as the target, various statistical methods have been used to downscale wind speed data from GCMs and RCMs at USLs.

- QM (Bartók *et al.*, 2019; Costoya *et al.*, 2020; de Souza Ferreira *et al.*, 2024; Hdidouan *et al.*, 2017; Li *et al.*, 2020; Moemken *et al.*, 2018; Vu Dinh *et al.*, 2022)
- ML models (Lin *et al.*, 2023; Nabipour *et al.*, 2020; Zhang *et al.*, 2021)
- Copula functions (Moradian *et al.*, 2023)
- Linear scaling based on the GWA (Luzia *et al.*, 2023)

QM was the most widely used statistical downscaling technique, followed by ML models. Various forms of QMs were explored for wind speed downscaling. Li *et al.* (2019) conducted a comparative study of different QM techniques. Their findings indicated that parametric QM based on the Weibull distribution performed slightly better than empirical distribution QM and CDF transformation (Michelangeli *et al.*, 2009). The CDF transformation aims to improve QM by accounting for changes in WSPD between historical and future projection periods.

In the context of ML models, CNN and recurrent neural networks have been applied due to their ability to capture complex spatial and temporal patterns in climate data. Lin *et al.* (2023) demonstrated that CNNs can enhance the spatial resolution of RCM outputs by extracting spatial features from gridded wind speed data. Meanwhile, Zhang *et al.* (2021) showcased the suitability of recurrent neural networks in capturing the relationships between wind speed datasets with different spatial resolutions.

Copula functions offer a flexible way to model the joint distribution of multiple variables while allowing for different marginal distributions (Harry, 2014). Moradian *et al.* (2023) applied copula functions to model the dependency between GCM outputs and reanalysis wind speed data, demonstrating the superior performance of copula functions compared to the ensemble mean method, which combines multiple GCM outputs into a single estimate.

Despite their advantages, statistical downscaling methods have limitations. One significant constraint is the assumption of stationarity in the relationships between the GCM and RCM outputs and local climate variables. Addressing these limitations remains a crucial area of ongoing climate modeling and projection research.

2.6 Software and tools

This section presents the software and tools for wind speed estimation at PSLs and USLs. Table 2.7 lists the software available for the MCP method. The software and tools reported in the literature for wind speed estimation at USLs are either SI software or ML packages available in one of the programming languages commonly used in the scientific community. Table 2.8 lists the software used for wind speed estimation at USLs. No specific software for the application of RFA has been reported in the literature.

Table 2.7 Software and tools for MCP

Software tool	Methods implemented/features	Latest version	Availability
Continuum	Orthogonal Regression; VR; Method of Bins; Matrix; Reanalysis data (reference site)	Version 3	Open source https://www.continuumwind.com/
MINT	LR; quantile regression; orthogonal regression; multiple reference sites	Version 1.1	Licence https://www.sander-partner.com/en/products/mint.html
WindFarm	LR; Orthogonal regression	Version 5	Licence https://www.resoft.co.uk/English/index.htm
WindFarmer	SLR; PCA regression	n/a	Licence https://www.dnv.com/services/wind-resource-assessment-software-windfarmer-analyst-3766
WindoGrapher	LR; VR; Matrix Time Series method (Lambert *et al.*, 2012); Orthogonal least squares,	Version 5	Licence https://www.ul.com/services/windographer-wind-data-analytics-and-visualization-solution

	SpeedSort (King *et al.*, 2005); Vertical Slice (Leblanc *et al.*, 2009), ; Weibull fit (Van Lieshout, 2010); ERA5 and MERRA-2 reference data		
WindPro	Artificial neural network; Matrix MCP Modeling (Thøgersen *et al.*, 2007) LR; WSM	Version 3.5	Licence https://www.emd-international.com/windpro/
WindSim	LR; MLP	Version 11	Licence https://windsim.com/

Table 2.8 Software and tools for estimation at USLs

Package/extension	method	Programming language/ Software	Reference
Geostatistical analyst	SI	ArcGIS	(Ali *et al.*, 2012; Luo *et al.*, 2008; Sliz-Szkliniarz *et al.*, 2011; Van Ackere *et al.*, 2015; Ye *et al.*, 2015)
mgcv	Statistical	R	(Li *et al.*, 2014; Reinhardt *et al.*, 2018)
Surfer	SI		(Sapuan *et al.*, 2011)
Geostatistical Software Library (GSLIB)	SI	FORTRAN	(Berndt *et al.*, 2018)
gstat	SI	R	(Reinhardt *et al.*, 2018)
e1071	ML	R	(Reinhardt *et al.*, 2018)
nnet	ML	R	(Reinhardt *et al.*, 2018)
scikit-learn	ML	Python	(Höglund, 2020)
Neural Toolbox	ML	MATLAB	(Fadare, 2010)
NN Toolbox	ML	MATLAB	(Lawan *et al.*, 2016; Muhammad *et al.*, 2014)
nftool	ML	MATLAB	(Kumar *et al.*, 2016a)
Statistics Toolbox	ML	MATLAB	(Jung, 2016)
SVM and Kernel Methods	ML	MATLAB	(Foresti *et al.*, 2011)
GRASP	ML	Splus	(Etienne *et al.*, 2010)
XGBoost	ML	Python	(Houndekindo *et al.*, 2023b)
LSBoost	ML	Matlab	(Jung *et al.*, 2022a; Jung *et al.*, 2020; Jung *et al.*, 2023b)

2.7 Challenges and future directions

A significant barrier to increasing wind energy penetration in many regions is the lack of high-quality datasets for assessing the temporal and spatial variability of this resource (Pelser *et al.*, 2024). During the previous decades, progress has been made toward developing statistical and ML methods for wind speed estimation at PSLs and USLs. This paper reviewed these methods to highlight recent advances and identify areas where further research and development are necessary.

2.7.1 Challenges and advances in long-term estimation at partially sampled locations

The MCP approach has traditionally been used to extend short-term wind speed records at PSLs. One major challenge of this method is its reliance on a high-quality reference site near the PSL. In recent years, the availability of reanalysis data has mitigated this limitation by providing consistent, long-term wind speed data across broader geographic areas. However, using reanalysis data introduces other challenges: the data quality can vary significantly depending on the region (Miao *et al.*, 2020) and the complexity of the terrain (Gualtieri, 2021; Potisomporn *et al.*, 2023). Due to their coarse resolution, reanalysis datasets often struggle to accurately represent localized wind patterns, particularly in complex terrain, resulting in increased uncertainty in wind speed estimates.

At PSLs, using available wind speed records to correct and improve the quality of reanalysis data can effectively reduce uncertainty in long-term wind speed estimation. Additionally, the adoption of nonlinear models, such as ANNs, and the introduction of additional explanatory variables (e.g., wind direction and other meteorological variables) can further improve the models' ability to capture the complex, nonlinear relationship present in the overlapping data between the reference site and the PSL.

The MCP approach is based on the assumption that the relationship derived from the overlapping wind speed data remains stationary. This assumption may not always be valid, particularly in regions where the wind regime exhibits notable temporal variability due to teleconnections and climate change. This challenge is further amplified by the reliance on a single reanalysis wind speed dataset as reference data. Some studies have reported inconsistencies in interannual variability (Ramon *et al.*, 2019) and trends (Fan *et al.*, 2021; Miao *et al.*, 2020) across different reanalysis wind speed datasets.

In such a situation, extending the wind speed records at the PSL could help reduce the uncertainty in long-term estimations. However, due to time and financial constraints, collecting additional wind measurements at the PSL may not always be feasible. In such cases, multiple reanalysis datasets are recommended (Jung *et al.*, 2023a). This approach can provide multiple scenarios of the long-term wind resource or be combined into an ensemble to improve the robustness of the estimates. Additionally, adopting a probabilistic modeling technique can effectively quantify the uncertainties associated with long-term wind resource estimates.

2.7.2 Challenges and advances in long-term estimation at unsampled locations

At USLs, no direct wind speed records could be used to correct biases in reanalysis wind speed data. As a result, there is a growing reliance on corrected and uncorrected reanalysis data for WRA at USLs. Simple BC methods, such as the linear scaling using the GWA, have improved the accuracy of reanalysis wind speeds. While this method can improve the long-term MWS and is applicable to virtually any USL, it does not sufficiently address the discrepancies in the temporal variability of reanalysis wind speeds, as reported by several studies (Brune *et al.*, 2021; Davidson *et al.*, 2022; Jourdier, 2020; McKenna *et al.*, 2022; Ramon *et al.*, 2019).

Consequently, there is a need to develop more advanced techniques that can improve both the MWS and the temporal variability of reanalysis wind speed data. ML models were developed to meet this requirement in regions with a well-established network of meteorological stations. However, the global distribution of meteorological stations is uneven; even within a country, coverage can be inconsistent. This disparity in data availability highlights the need for developing reanalysis BC methods that can improve the long-term MWS and the temporal variability while being transferable across diverse regions and terrains. These models should generalize beyond specific geographic areas, ensuring that WRA in under-sampled or unsampled regions can benefit from accurate and reliable wind speed predictions.

2.7.3 Future directions

Both reanalysis and measured wind speed data contain inherent uncertainties, which the modeling processes can further compound. Accurately estimating these uncertainties is critical for a reliable WRA. Probabilistic modeling techniques, such as Bayesian inference and quantile regression, are often employed to quantify the uncertainties in model outputs. These techniques provide a range of possible outcomes or prediction intervals instead of a single deterministic prediction. This probabilistic information is essential for informed decision-making in wind energy projects, enabling stakeholders to consider the variability and risks associated with wind energy potential. Despite its importance, uncertainty quantification at PSLs and USLs has received limited attention in the literature.

Future research should focus on developing methods for uncertainty quantification at both PSLs and USLs. For example, at PSLs, ANN models could be designed to output parameters of a probability distribution function (Salinas *et al.*, 2020) rather than single-point predictions, providing a more comprehensive representation of the inherent uncertainty and variability in wind speed predictions. At USLs, GB models, which are widely used, can be adapted by altering the loss function to predict either parameters of a probability distribution function (Duan *et al.*, 2020) or conditional quantiles (Waldmann, 2018), allowing for the construction of prediction intervals.

Moreover, wind speed data is often subject to significant noise. Preprocessing the data with signal decomposition techniques, such as wavelet analysis and variational mode decomposition (Dragomiretskiy *et al.*, 2014), is expected to improve the modeling procedure. These decomposition techniques are increasingly applied in wind speed forecasting, with numerous studies demonstrating improved model performance (Wang *et al.*, 2021). By decomposing WSTS into different subseries, these methods can effectively filter out noise and capture both short-term fluctuations and long-term trends, leading to more accurate predictions.

Furthermore, the non-stationarity of WSTS poses a challenge for the long-term sustainability of wind farm projects. Despite the significance of this issue, few studies have focused on nonstationary wind speed estimation at PSLs and USLs. The availability of multiple reanalysis datasets and GCM outputs provides valuable insights into the impacts of several sources of non-stationarity. However, these datasets often come with high levels of uncertainty and coarse resolution. More research is necessary to address these challenges, particularly in light of climate change impacts on wind energy projections (Jung *et al.*, 2022c).

2.8 Conclusion

Wind energy development is crucial to achieving Goal 7 of the United Nations' Sustainable Development Goals, which focuses on affordable and clean energy. Accurate methods for estimating wind speed at partially sampled and unsampled locations, including assessments of its long-term variability, are critical for informed decision-making. These methods help build stakeholders' confidence in adopting wind energy as a reliable renewable resource. This paper reviews statistical and ML approaches for wind speed estimation at PSLs and USLs.

At PSLs, the MCP method is commonly used to extend the available short-term wind speed record, with most studies recommending at least a year of wind speed data at the PSL. Some studies have shown that longer records can significantly reduce uncertainty, whereas shorter records risk overestimating or underestimating resource potential due to seasonal variability. In cases where nearby reference stations are unavailable for the MCP method, many studies have turned to reanalysis wind speed data as a substitute. However, few of these studies validate the accuracy of reanalysis data for their specific region, and even fewer assess the interannual variability of the dataset. Discrepancies between the interannual variability of reanalysis data and the actual wind speed variability in a region can lead to errors in estimating the resource potential from hindcasted wind records at the PSL. Generally, more efforts are needed to evaluate the predictions' uncertainties at both PSL and USLs. While various uncertainty estimation methods exist that can be incorporated into current frameworks, their adoption remains limited in practice.

At USLs, the standard method for wind speed analysis has traditionally been based on the SI of observed wind speeds, which depends on the availability of dense meteorological station networks. In contrast, BC of reanalysis wind speeds is less reliant on such networks, particularly when utilizing the GWA. ML methods incorporating reanalysis wind speed data and high-resolution topographic and land use data have shown promising results in predicting wind speed at USLs. However, there remains a need to develop transferable models that can be applied in regions with a sparse network of stations.

The findings and recommendations outlined in this analysis have significant implications for the wind energy industry and research. Accurate wind speed estimation directly supports the optimization of site selection for wind farm projects, ensuring higher energy yields and improved project efficiency. This review can assist in selecting the most effective methods depending on different factors, such as the stage of the project, the data availability, and the complexity of the terrain.

In addition to aiding the energy industry, the findings of this paper have broader implications for research. Researchers can build on the reviewed methodologies to address current limitations, such as improving uncertainty estimation and model transferability and developing nonstationary wind speed estimation techniques at PSLs and USLs. Addressing these challenges can lead to the development of more robust and scalable models capable of capturing wind speed variability across different spatial and temporal scales. Furthermore, enhancing model interpretability and transparency will be critical to building stakeholder trust, particularly when decisions have significant economic and environmental impacts.

Nomenclature

Abbreviations

a.g.l	Above ground level
ANN	Artificial neural networks
BC	Bias correction
CDF	Cumulative distribution function
CELU	Continuously differentiable exponential linear units
CNN	Convolutional neural network
DEM	Digital elevation models
ERA5	European Center for Medium-Range Weather Forecasts Reanalysis v5
EWS	Extreme wind speeds
GB	Gradient boosting
GCM	Global climate models
GRNN	General regression neural network
GWA	Global wind atlas
K-NN	K-nearest neighbor
LR	Linear regression
MAE	Mean absolute error
MARE	Mean Absolute Relative Error
MCP	Measure correlate predict
MERRA-2	Modern-Era Retrospective Analysis for Research and Applications, Version 2
ML	Machine learning
MLP	Multilayer perceptron
MSE	Mean squared error
MWS	Mean wind speed
NWP	Numerical weather prediction
PSL	Partially sampled location
QM	Quantile mapping
R^2	Coefficient of determination
RCM	Regional climate model
RF	Random forest
RFA	regional frequency analysis

RMSE	Root mean squared error
SI	Spatial interpolation
SRL	Surface roughness length
SVR	Support vector regression
USL	Unsampled location
VR	Variance ratio regression
WMO	World Meteorological Organization
WRA	Wind resource assessment
WSPD	Wind speed probability distribution
WSTS	Wind speed time series

Symbols

α	Power law exponent
α^{GWA}	Weibull scale parameter from the GWA
α^{MERRA}	Weibull scale parameter from MERRA-2
b	Kernel bandwidth
$b^{(l)}$	Multilayer perceptron bias vector of the l-th layer
$e(s_i)$	Residuals at sampled locations s_i
Err_{mean}	Relative bias in the prediction mean
Err_{var}	Relative bias in the prediction variance
$\hat{F}_{BS}(\cdot)$	Birnbaum-Saunders CDF kernel
$\hat{F}_{LN}(\cdot)$	Log-Normal CDF kernel
$F_n(\cdot)$	Empirical Cumulative distribution function
$g^{(l)}(\cdot)$	Multilayer perceptron activation function of the l-th layer
$h^{(l)}$	Multilayer perceptron hidden vector of the l-th layer
$K_{BS}(\cdot)$	Birnbaum-Saunders kernel function
k^{GWA}	Weibull scale parameter from the GWA
$K_{LN}(\cdot)$	Log-Normal kernel function
k^{MERRA}	Weibull scale parameter from MERRA-2
λ_i	Kriging weights
$\hat{\lambda}_r$	Sample L-Moments of order r
μ	Mean wind speeds
$\bar{o}$	Mean of the observed values

o_i	Observed value
$\bar{p}$	Mean of the predicted values
$\Phi(\cdot)$	Cumulative distribution function of the standard normal distribution.
p_i	Predicted value
$Q_{\hat{\theta}}(\cdot)$	Quantile function of the Kappa distribution
r	L-moment order
σ	Standard deviation of the wind speed
$\hat{\theta}$	Estimated of the Kappa distribution
U_z	Wind speed at a height of z
U_{z_r}	Wind speed at a reference height z_r
$W^{(l)}$	Multilayer perceptron weight matrix of the l-th layer
$X_{i:n}$	Sample order statistics
$\hat{y}$	Estimated wind speed
y_t^{ERA5}	Wind speed value from ERA5
$\overline{y^{GWA}}$	Mean wind speed from the GWA
y_t^{MERRA}	Wind speed value from MERRA-2
$\overline{y^{MERRA}}$	Mean wind speed from MERRA-2
z_0	Roughness length

3 COMPARATIVE STUDY OF FEATURE SELECTION METHODS FOR WIND SPEED ESTIMATION AT UNGAUGED LOCATIONS

Titre de l'article : Étude comparative des méthodes de sélection de variables explicatives pour l'estimation de la vitesse du vent aux sites non échantillonnés

Auteurs :

Freddy Houndekindo[1] (F.H.), Taha B.M.J. Ouarda[1] (T.B.M.J.O)

[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

Titre de la revue :

Article publié le 1 septembre 2023, dans la revue **Energy Conversion and Management**, volume 291, DOI : 10.1016/j.enconman.2023.117324

Contribution des auteurs :

Conceptualisation, F.H., T.B.M.J.O. ; Méthodologie, F.H., T.B.M.J.O. ; Préparation des données, F.H. ; Logiciel, F.H. ; Rédaction de la version originale, F.H. ; Révision, T.B.M.J.O. ; Supervision, T.B.M.J.O. ; Acquisition du financement, T.B.M.J.O.

Au cours de la revue de littérature, nous avons constaté une absence de consensus sur les variables explicatives, en particulier topographiques, pour l'estimation des vitesses du vent aux sites non échantillonnés. Nous avons donc procédé, dans l'article 2, à une analyse comparative des variables explicatives identifiées dans la littérature, en analysant leur influence sur les différentes plages de vitesse du vent. En plus de répondre à l'une des pistes de recherche mises en évidence dans la revue de littérature, cet article représente une étape clé dans le développement des approches proposées dans les travaux suivants, car il permet d'identifier les facteurs les plus pertinents et a optimisé leur intégration dans les modèles d'estimation du vent aux sites non échantillonnés.

Abstract

Wind speed estimation at ungauged locations is one of the preliminary steps for wind resource assessment. With the availability of high-resolution Digital Elevation Models (DEM) and remote sensing data, the number of potential wind speed predictors has grown substantially. The adequate spatial scale of these predictors is unknown a priori, leading to the use of multiple spatial scales of predictors in wind speed estimation models. Implementing a feature selection method as a pre-processing step of the analysis is necessary to avoid overfitting and the resulting potential model underperformance. This paper evaluated six feature selection methods (forward stepwise regression, Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net, Maximum relevance Minimum redundancy (MRMR), Genetic algorithm, and recursive feature elimination using support vector regression) for the estimation of different wind speed quantiles across Canada. The selected features were used to fit a regression-kriging model, and the importance of the predictors was evaluated with their associated regression coefficients. The results of the study showed that LASSO and MRMR are the most efficient algorithms with the least number of parameters to tune and good generalization performance. The study found that some predictors were more important for specific exceedance probabilities. The most important predictors were the distance from the coast and surface roughness length, regardless of exceedance probability.

Keywords: Exceedance probability, Feature selection, Machine learning, Topographic feature, ungauged location, Wind speed.

3.1 Introduction

The global energy system significantly contributes to greenhouse gas emissions, with a share of approximately 34% (Lamb et al., 2021). Alternative energy sources, such as wind, can help mitigate the environmental footprint of our energy system (Jung et al., 2018b; Shin et al., 2016). Wind energy production has experienced substantial growth during the last decades, accounting for 8% (594 GW) of the 7 400 GW of installed generating capacity worldwide as of 2019 (International Renewable Energy Agency, 2022). Unlike conventional energy sources such as coal and nuclear energy, wind energy is intermittent and heavily reliant on wind speed (WS). A sound understanding of the WS variability at a location of interest for wind energy production is necessary to integrate the energy source effectively into the energy mix (Aries et al., 2018). A significant step in wind energy planning is identifying a good location for resource exploitation. Potential sites of interest often do not coincide with a location where extensive WS measurements are available. Therefore, it is helpful to implement approaches that estimate wind resources at ungauged locations.

The challenge of WS estimation at ungauged locations has initially been tackled with spatial interpolation models. In recent studies, machine learning models have gained more popularity, and some researchers have suggested combining spatial interpolation models and machine learning (see Houndekindo et al. (2023c)) for a detailed review of WS estimation at ungauged locations). These developments have led to the experimentation of new predictors, notably topographical features extracted from a Digital elevation model (DEM). Many topographical features can be used for WS modelling (Maxwell et al., 2022). One such feature is terrain curvature, which has been identified as one of the most effective WS predictors in regions with complex terrain, according to a study conducted in Switzerland by Robert et al. (2013). Several land surface parameters (e.g., plan curvature, gaussian curvature, minimum curvature) extracted from DEM can be used to describe the terrain curvature (Wilson, 2018), leading to several possible features to include in the model. Some of these features will undoubtedly be redundant (Maxwell et al., 2022). The selection of the spatial scales of the topographical features represents another significant challenge. Two potential downsides of incorporating too many features into the model are overfitting the model's parameters to the training data and compromising the model’s interpretability. To address this issue, feature selection (FS) can be used as a preprocessing step to build more accurate and concise models while minimizing computation time (Guyon et al., 2003).

FS methods are often categorized as filter-based, wrapper, or embedded methods (Guyon et al., 2003). Filter-based methods are more computationally efficient and less prone to overfitting compared to wrappers and embedded methods (Zhou et al., 2021b). A drawback of most filter methods compared to wrappers and embedded methods is their inability to consider feature interactions (Urbanowicz et al., 2018). The filter approach selects predictors based on their relevance to the dependent variable. In the case of regression, the correlation coefficient can be used to assess the relevance of features.

On the other hand, wrappers and embedded methods rely on the model performance to select an optimal set of features. The wrapper methods search for the feature subset, which gives the best performance with a predefined learning algorithm. Wrapper methods can be used with any model, while embedded methods rely on models that inherently rank the features' importance (e.g., random forest) or eliminate irrelevant features (e.g., penalization methods).

Most studies have applied a data-driven approach to solving the feature selection challenge for WS estimation. For example, Robert et al. (2013) applied a modified version of the general regression neural networks to select the best spatial scale and topographical features for monthly WS interpolation. Jung (2016) employed feature importance ranking with random forest and a forward stepwise feature selection to identify suitable predictors for WS estimation. In the second step, the author used the variance inflation factor to evaluate feature redundancy in the study. For extreme WS mapping, Etienne et al. (2010) used the linear correlation between predictors to evaluate their redundancy and backward elimination to retain the most important predictors in the model. Foresti et al. (2011) applied a multiple kernel learning model for FS in WS mapping. Veronesi et al. (2016) employed the Least Absolute Shrinkage and Selection Operator (LASSO) technique to select relevant features to implement a statistical model for estimating WS distribution at ungauged sites.

To the best of our knowledge, no studies compared the performance of FS methods for WS estimation at ungauged locations. Nevertheless, such comparison is necessary as the number of available WS predictors increases, and so is the risk of redundancy and overfitting. Comparative studies are essential as they allow for a systematic comparison of various approaches with diverse complexity and performance levels. They serve as a basis to identify the strengths and weaknesses of each approach and better understand their performance in different conditions. Several comparative studies of features selections methods have been conducted in studies related to environmental variables. For instance, Carta et al. (2015) compared a wrapper method to a filter approach for FS for long-term WS prediction at locations with a short record. The authors found that the filter method produced sparser feature subsets, while the wrapper method had a better predictive ability. In that study, FS increased the interpretability of the final model while improving its performance. Seven FS methods were compared for river flow quantile estimation in ungauged basins (Fouad et al., 2020). The authors found that the FS methods performed better than dimension reduction techniques (principal component analysis) to reduce multicollinearity in the feature subsets. The same study observed similar performance between FS using experts' knowledge and data-driven FS methods. Rodriguez-Galiano et al. (2018) evaluated the performance of various FS methods to predict the probability of the occurrence of nitrates above a threshold value in groundwater. The study revealed that FS helped isolate and identify the main drivers of nitrate pollution in groundwater. Chen et al. (2019) conducted a comparative study of statistical models with various FS methods to predict fine particles and nitrogen dioxide concentration across Europe. The study found that regularization algorithms such as LASSO and Elastic Net (ENET) efficiently selected relevant predictors despite high multicollinearity in the feature set. Also, the regularization algorithms had the additional benefit of model interpretability.

This study compared six different FS methods for WS quantile estimation. These methods included forward stepwise regression (FSWR), LASSO, ENET, Maximum relevance Minimum redundancy (MRMR), Genetic algorithm (GALG), and recursive feature elimination using support vector regression (RFES). The selected algorithms are composed of filter-based (e.g., MRMR), wrappers (e.g., FSWR, GALG), and embedded methods (e.g., LASSO, ENET, RFES). The selected predictors or features were used with a regression kriging (RK) model (Hengl et al., 2007) to estimate various WS quantiles. The RK model has previously shown promising results for WS estimation (Alsamamra et al., 2010; Lee, 2022). Reinhardt et al. (2018) also found that RK performed better than Artificial Neural Networks (ANN) and Support Vector Machines (SVM) for WS interpolation. RK is an attractive approach for interpolating environmental variables (Hengl et al., 2007). It allows the use of relevant predictors, and unlike universal kriging and kriging with external drift, RK can be adapted with various types of regression models (e.g., Random Forest, Generalized Additive Models).

The study also evaluated the importance of various predictors for estimating WS quantiles with different exceedance probabilities. Most features used in previous studies were derived and compared within the same framework. In addition, alternative features related to conventional WS predictors used in the literature were also evaluated. These alternative features may provide additional information and insights into WS behaviour at different exceedance probabilities and could improve the accuracy of WS predictions.

The paper is organized as follows. The dataset used is described in section 3.2. In section 3.3, the six FS methods evaluated are presented. Section 3.4 presents the results of the analysis. The discussion and the conclusion are given in sections 3.5 and 3.6, respectively.

3.2 Data

3.2.1 Wind speed data

The data analyzed in the study are hourly WS data at 10m above ground from measurement stations across Canada. The data were obtained from Environment and Climate Change Canada (ECCC) historical climate database. Stations with at least 20 years of record available until 2010 were selected, and only those with at least ten years of record with less than two months of missing data were used. Figure 3.1 shows the spatial distribution of the selected stations, which amounted to 207.

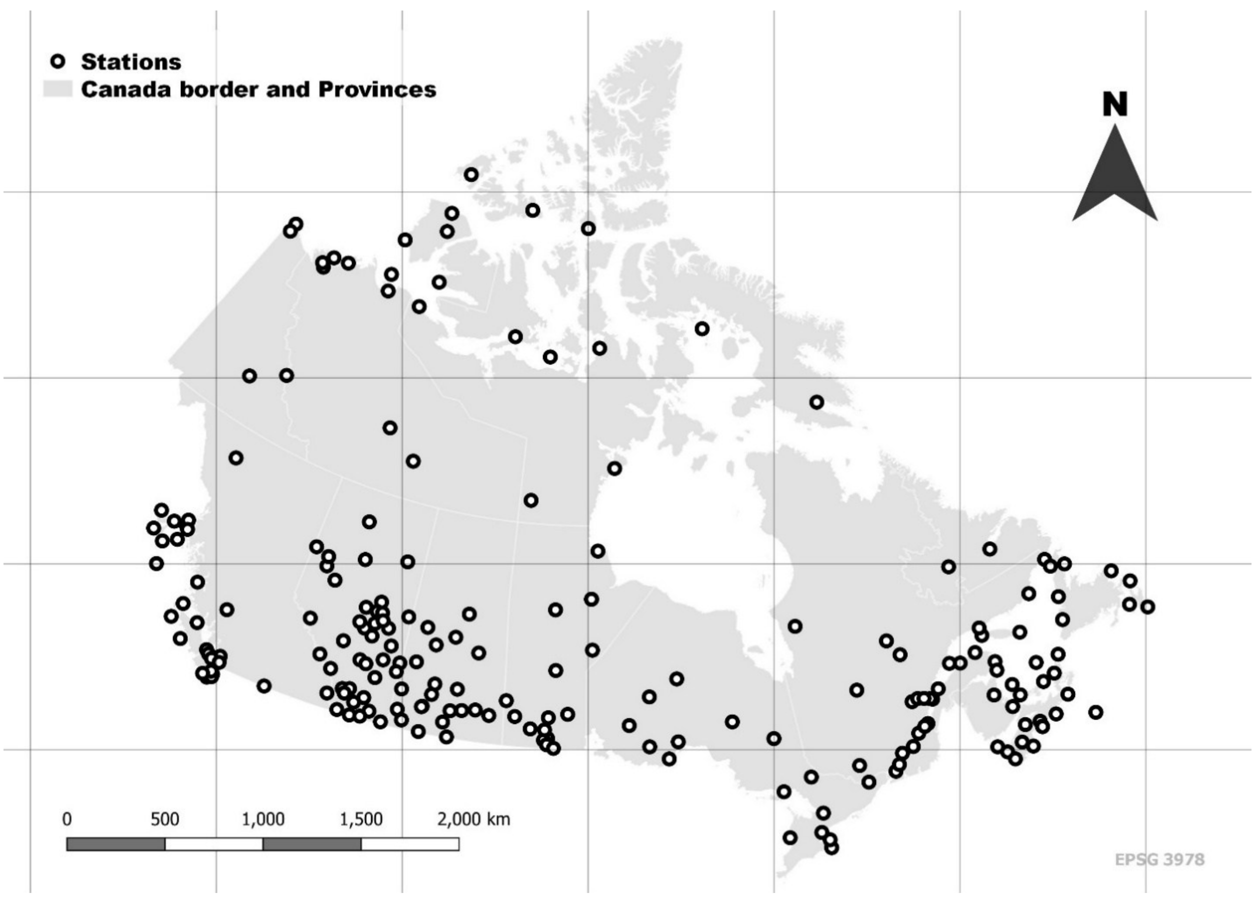


Figure 3.1 Study region and locations of the 207 selected stations

From the hourly records, empirical WS quantiles were estimated using the Weibull plotting position formula:

Equation 3.1

$$P_i = P(Ws > Ws_i) = \frac{i}{n+1}$$

where P_i is the probability of exceedance associated with the observed hourly wind speed (Ws_i), i is the rank of the observed wind speed Ws_i sorted in descending order, $i = 1$ corresponds to the highest observed WS, and $i = n$ corresponds to the lowest observed WS, with n the number of observations.

Monotonic decreasing penalized splines (P-Splines: Paul *et al.*, 1996; Pya *et al.*, 2015) were fitted between the exceedance probabilities and their associated observed WS quantiles to construct the empirical complementary cumulative distribution function (survival function). The fitted curve was used to estimate WS quantiles at 14 fixed percentile points at each location in the study area. The following 14 fixed percentile points were selected: p = 0.01%, 0.1%, 1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95% to cover an extensive range of WS quantiles. The P-Splines is a non-parametric model that allows fitting a smooth and flexible curve to data. Monotonic decreasing constraints were imposed on the P-Splines to respect the monotonic nature of complementary cumulative distribution functions.

3.2.2 Predictors

The predictors used in the study are topographical, surface roughness length, geographical coordinates, and the location distance from the coast. Table 3.1 provides more details on these predictors. The topographical variables were extracted from a resampled (100 m spatial resolution) ALOS DEM (Tadono et al., 2014) and computed with the WhiteboxTools (Lindsay, 2014) developed at the University of Guelph, Canada. Information on the land cover type obtained from a 2015 land use map of Canada (Latifovic et al., 2017) was used to estimate the surface roughness length according to Wiernga (1993). The land use map was resampled to produce multiple spatial resolutions, with majority resampling (mode) providing information on the most common land use type for the given spatial scale.

Some of the features selected for the study were previously studied because they describe physical processes that influence wind movement. This study also introduced alternative features describing similar physical processes. For instance, Jung (2016) used slope (SLPE), curvature, aspect (ASPC), roughness length (RGLH) and relative elevation for WS mapping in Germany. In the present study, relative elevation measures used were deviation and difference from mean elevation (DVME and DFME), relative topographic position (RTGP) and elevation percentile (ELVP). Also, seven surface curvature measures (gaussian, maximal, mean, minimal, plan, tangential, and total curvature) were extracted from the DEM and used as WS predictors. In Switzerland, Foresti *et al.* (2011) used altitude (ELVT), geographic coordinates (XGEO and YGEO), and Differences of Gaussians (DOGS) to map WS. DOGS serves as a measure of terrain convexity and approximates the Laplacian of Gaussian (LPGS: Lowe, 2004). In the current study, DOGS and LPGS were both evaluated. Veronesi *et al.* (2015) employed topographical surface roughness from a DEM to interpolate the parameters of the Weibull distribution for wind resource mapping. Alternative topographical surface roughness measures employed in the present study were the ruggedness index (RUGI), the surface area ratio (SART) and the standard deviation of the slope (STDS). Etienne *et al.* (2010) generated landform classes (e.g., canyons, ridges, valleys) from a DEM to model WS. Geomorphologic phenotypes (GMPG) and the Pennock landform class (PNCL) were two alternative landform classifications used in the present study. The distance from the coast (DSEA) was also used as a WS predictor in the current study, as done by Aniskevich *et al.* (2017).

Table 3.1 Description of the predictors and their spatial scale

Predictor	Abbreviation	Description	Spatial scale
Altitude	ELVT	Altitude of the location in m.	
Aspect	ASPC	Slope orientation in degree.	100m, 500m, 1000m, 1500m, 2000m
Deviation from mean elevation	DVME	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	DFME	Difference between the grid cell elevation and the mean of its neighbouring cells.	100m, 500m, 1000m, 1500m, 2000m
Difference of Gaussian	DOGS	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	DSEA	The location distance to the coast	
Elevation percentile	ELVP	Percentile of the grid cell elevation relative to the neighbouring cells.	100m, 500m, 1000m, 1500m, 2000m
Gaussian curvature	GSCV	Product between the maximal and the minimal curvature. Measure of surface curvature (Florinsky, 2017).	100m, 500m, 1000m, 1500m, 2000m
Geographical coordinates	XGEO, YGEO	Geographical coordinates of the location.	

Geomorphologic phenotypes (geomorphons)	GMPG	Landform element classification with the geomorphons-based method (Jasiewicz *et al.*, 2013).	
Laplacian of Gaussian	LPGS	Derivative filter used to highlight location of rapid elevation change.	100m, 500m, 1000m, 1500m, 2000m
Maximal curvature	MXCV	Measure of surface curvature (Wilson, 2018).	100m, 500m, 1000m, 1500m, 2000m
Mean curvature	MNCV	Measure of surface curvature (Wilson, 2018).	100m, 500m, 1000m, 1500m, 2000m
Minimal curvature	MICV	Measure of surface curvature (Florinsky, 2017).	100m, 500m, 1000m, 1500m, 2000m
Pennock landform class	PNCL	Landform classification based on the slope and curvature of the grid cell (Pennock *et al.*, 1987).	
Plan curvature	PLCV	Measure of surface curvature (Florinsky, 2017).	100m, 500m, 1000m, 1500m, 2000m
Relative topographical position	RTGP	Normalized measure of the grid cell elevation relative to its neighbouring cells.	100m, 500m, 1000m, 1500m, 2000m
Ruggedness index	RUGI	A measure of the local terrain heterogeneity (Jasiewicz *et al.*, 2013; Riley *et al.*, 1999).	100m, 500m, 1000m, 1500m, 2000m
Slope	SLPE	Slope at the grid cell.	100m, 500m, 1000m, 1500m, 2000m
Standard deviation of slope	STDS	Measure of surface roughness (Grohmann *et al.*, 2011).	100m, 500m, 1000m, 1500m, 2000m
Surface area ratio	SART	Measure of the surface roughness (Jenness, 2004).	100m, 500m, 1000m, 1500m, 2000m
Surface roughness length	RGLH	Surface roughness length estimated from land use map.	100m, 500m, 1000m, 1500m, 2000m
Tangential curvature	TGCV	Measure of surface curvature (Florinsky, 2017).	100m, 500m, 1000m, 1500m, 2000m
Total curvature	TLCV	Measure of surface curvature.	100m, 500m, 1000m, 1500m, 2000m

3.3 Materials and method

3.3.1 Feature selection methods

3.3.1.1 Forward stepwise regression

The stepwise regression is a greedy FS algorithm extensively covered in the literature. Three variants of the method exist backward, forward, and bi-directional stepwise regression. Backward stepwise regression builds a model with all potential predictors and eliminates the least relevant predictors at each iteration. Forward selection begins with a "null" model containing only a constant term and adds the most relevant predictors to the regression model at each iteration. Bi-directional stepwise regression combines backward and forward stepwise regression. Various criteria have been used in the literature to measure the predictors' relevancy (e.g., AIC, P-value, R^2-adjusted).

There is a thorough discussion in the literature about the shortcomings of stepwise regression (Whittingham *et al.*, 2006), with Smith (2018) advising against its use. The author found that stepwise regression underperformed as potential predictors increased. However, the method remains widely used in the scientific community. In this paper, a forward stepwise regression (FSWR) was applied as a benchmark. The algorithm was initiated with the null model, and potential predictors that led to the most significant increase in R^2-adjusted were added at each iteration. This procedure is repeated until no candidate variables left could improve the R^2-adjusted. A similar forward stepwise regression approach was implemented by Chen *et al.* (2019) and performed better than backward stepwise regression for annual average fine particle ($PM_{2.5}$) and nitrogen dioxide (NO_2) concentrations prediction.

3.3.1.2 Least Absolute Shrinkage and Selection Operator

LASSO algorithm is a penalty-based linear model developed by Tibshirani (1996), which imposes an L1-norm penalization on the regression coefficient forcing some coefficients to zero and thus producing a sparse solution. The LASSO regression coefficient estimates are given by:

Equation 3.2

$$\hat{\beta} = argmin_{\beta}(Y - X\beta)^T(Y - X\beta) + \alpha \sum_{j=1}^{p} \left|\beta_j\right|^1$$

where Y is the response vector, X is the matrix of predictors, $\boldsymbol{\beta}$ are the regression coefficients, p is the number of predictors, α is a tuning parameter that controls the degree of penalization, $\alpha \sum_{j=1}^{p} \left|\boldsymbol{\beta}_j\right|^1$ is the penalization term and $|.|^1$ represents the L1-norm of a vector.

Zou *et al.* (2005) discussed some limitations of LASSO regression, which renders the algorithm inappropriate for FS in some situations. A particularly relevant limitation in this study is the inferior prediction performance of LASSO regression compared to Ridge regression when there is a high correlation between the predictors.

3.3.1.3 Elastic Net

LASSO regression can be seen as a particular case of the Bridge regression introduced by Frank *et al.* (1993). In Bridge regression, the penalization term in Equation 3.2 becomes $\alpha \sum_{j=1}^{p} |\beta_j|^{\gamma} \; with \, \gamma \geq 0$. LASSO regression is equivalent to Bridge regression when $\gamma = 1$. Another well-known case of Bridge regression is Ridge regression with $\gamma = 2$. With Ridge regression, the regression coefficients are shrunk depending on the predictors' importance, but they are not set to zero if the variables are irrelevant to the regression.

The ENET model combines the Ridge and the LASSO penalty. The Elastic net algorithm minimizes the following equation:

Equation 3.3

$$min_{\beta}(Y - X\beta)^T(Y - X\beta) + \alpha\lambda \sum_{j=1}^{p} |\beta_j|^1 + \alpha(1-\lambda) \sum_{j=1}^{p} |\beta_j|^2$$

where $\alpha \; and \; \lambda$ ($0 \leq \lambda \leq 1$) are two hyperparameters of the model that can be selected using cross-validation.

3.3.1.4 Genetic Algorithms

GAGL is an optimization algorithm that emulates natural evolution and selection to find an optimal solution. It has been implemented in several studies for FS (Amini *et al.*, 2021; Eseye *et al.*, 2019; Gokulnath *et al.*, 2019; Leardi *et al.*, 1992). The algorithm starts with a population of solutions (individuals) initialized randomly. A fitness measure is defined to evaluate every solution in the population. A new population is formed by producing offspring from the best solutions of the old population (by reproduction and genetic mutation). This procedure is repeated until a stopping criterion is reached. Several variations of the algorithm control, among others, how the offspring of the population are bred. The different steps of the genetic algorithm implemented in this study are described as follows:

Step 1: A population was initialized randomly with 50 potential solutions. The solutions were encoded as a sequence of binary strings (the genes), with each gene associated with a particular feature among the candidate features. A selected gene (a feature) was represented by "1" and a none selected gene by "0". The population is represented by a binary matrix where the rows represent the potential solutions, and an entry represents a feature or a gene.

Step 2: The 50 solutions in the population were evaluated (fitness score), and the best solution was copied without modification to the next generation.

Step 3: The next generation's parents were selected with the roulette wheel selection method: the solutions with the highest fitness score have more chances to be selected as parents for reproduction to produce offspring. The reproduction process was performed through two genetic operators, uniform crossover and mutation.

Step 4: Step 3 was repeated until the new population size equalled the initial population size.

Step 5: Steps 2 to 4 were repeated until the maximum number of iterations was reached.

Table 3.2 presents different parameters of the algorithm used in this study. The performance of the solutions was evaluated with a 10-fold cross-validation root mean squared error (RMSE) estimated with a simple linear regression model:

Equation 3.4

$$RMSE(CV) = \frac{1}{10}\sum_{k=1}^{10}\sqrt{\frac{1}{n_k}\sum_{i=1}^{nk}(y_i - \hat{y}_i)^2}$$

where n_k is the size of the kth fold, and y_i and $\hat{y}_i$ are the observed and predicted WS values.

The fitness score was estimated as a weighted sum of the solution performance (RMSE) and its cardinality (Card) as follows:

Equation 3.5

$$F_i = \frac{w_1}{RMSE_i} + \frac{(1 - w_1)}{Card_i}$$

where $0 < w_1 < 1$

The probability of selection of a solution for the reproduction process was assigned based on:

Equation 3.6

$$Psel_i = \frac{F_i}{\sum_{i=1}^{50} F_i}$$

Table 3.2 Selected parameters of the genetic algorithm

GA parameter	Value/method
Initial population size	50

Crossover type	Uniform
Crossover probability	0.9
Mutation probability	0.05
Selection process	roulette wheel selection
Maximum number of iterations	100
$\boldsymbol{w_1}$	0.1, 0.5, 0.7

3.3.1.5 Minimum redundance – Maximum relevance

Filter-based FS approaches such as maximal relevancy (e.g., correlation) do not require the regression model to be evaluated multiple times (e.g., in cross-validation); they are relatively computationally efficient and less prone to overfitting. One of their drawbacks is their failure to ignore redundant predictors correlated to the response variable.

The MRMR algorithm is an iterative approach developed by Ding *et al.* (2005) to improve conventional filter-based FS approaches. MRMR benefits from the advantages of the filter-based FS approach while ignoring redundant features in the process. At each iteration of the MRMR algorithm, a function measuring the redundancy and relevancy is computed, and the feature that maximizes this function is selected. Several measures of relevancy and redundancy have been proposed in the literature depending on the type of variables (discrete vs. continuous), the desired level of trade-off between relevancy and redundancy (Zhao *et al.*, 2019), and the type of relationship (linear or nonlinear). In this study, the relevancy is measured with the F-statistic ($F(y, x_i)$). The redundancy of a non-selected feature is measured as the inverse of the sum of the correlation between the feature and the selected features (Ding *et al.*, 2005), and the MRMR optimization criterion function is:

Equation 3.7

$$f(x_i) = \frac{F(y, x_i)}{\frac{1}{s}\sum_{j=1}^{s} \rho(x_s, x_i)}$$

where $\rho(x_1, x_2)$ is the Pearson correlation coefficient between features x_1 and x_2, and s is the number of selected features.

Equation 3.8

$$F(y, x_i) = \frac{\rho(y, x_i)^2}{[1 - \rho(y, x_i)^2]} \times (n - 2)$$

where $n - 2$ is the degree of freedom of a simple linear regression model fitted with n samples, one predictor and a constant term

At each iteration, the algorithm seeks to find the feature (x_i) which maximizes $f(x_i)$. The stopping criterion of the algorithm (number of selected features to include in the model) is a hyperparameter that can be determined using cross-validation.

3.3.1.6 Recursive Feature Elimination Support Vector Regression

The RFES algorithm (Guyon *et al.*, 2002) is a backward elimination algorithm. The model is fitted to the data at each iteration, and the least important predictor is removed from the feature set. This process is repeated until a stopping criterion (e.g., minimum size of feature set) is reached. The stopping criterion can be determined through cross-validation. In the RFES algorithm, the importance of a predictor is measured by the square of its associated coefficient in the weight vector (w : Equation 3.17) using the epsilon-insensitive SVR formulation (Vapnik, 2000), with epsilon the maximum tolerable deviation between the predictions and the observed values.

Let $f(x)$ be the linear function used to approximate the relationship between the predictors (x) and the response variables y:

Equation 3.9

$$f(x) = \langle w.x \rangle + \ b$$

In the epsilon-insensitive SVR formulation (Vapnik, 2000), the loss function is defined as follows:

Equation 3.10

$$Loss = \begin{cases} 0 & if\ |y - f(x)| \le \varepsilon \\ |y - f(x)| - \ \varepsilon & otherwise \end{cases}$$

It is desirable to find a solution to Equation 3.9 having w with minimum norm to reduce the model complexity. The optimization problem can be re-written as follows:

Equation 3.11

$$minimize\ J(w) = \frac{1}{2} \|w\|^2$$

Subject to:

Equation 3.12

$$|y_i - \langle w, x_i \rangle + b| \leq \varepsilon$$

With noisy data, $f(x)$ may not satisfy the epsilon-insensitive constraint. Therefore, slack variables ($\xi_i \xi_i^*$) are introduced for each point to allow less restrictive constraints leading to the following formulation (Vapnik, 2000):

Equation 3.13

$$minimize\ J(w) = \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$

Equation 3.14

$$subject\ to: \begin{cases} y_i - \langle w. x_i \rangle - b & \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i & \leq \varepsilon + \xi_i^* \\ \xi_i \xi_i^* & \geq 0 \end{cases}$$

where C is a regularization parameter

From the objective function and the constraints (Equation 3.13 and Equation 3.14), a Lagrange function L is defined by introducing non-negative Lagrange multipliers $\alpha_i \alpha_i^*, \eta_i \eta_i^*$:

Equation 3.15

$$L = \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) - \sum_{i=1}^{n} (\eta_i \xi_i + \eta_i^* \xi_i^*) - \sum_{i=1}^{n} \alpha_i (\varepsilon + \xi_i - y_i + \langle w, x_i \rangle + b)$$

$$- \sum_{i=1}^{n} \alpha_i^* (\varepsilon + \xi_i^* + y_i - \langle w, x_i \rangle - b)$$

At the saddle point, the partial derivatives of L in all directions are null, giving the following equations in w direction:

Equation 3.16

$$\partial L / \partial w = w - \sum_{i=1}^{n} (\alpha_i + \alpha_i^*) x_i = 0$$

Equation 3.17

$$w = \sum_{i=1}^{n} (\alpha_i + \alpha_i^*) x_i$$

3.3.2 Performance evaluation

The RK model was implemented to estimate the WS quantiles using the selected predictors. The RK model can be expressed as follows (Hengl *et al.*, 2007):

Equation 3.18

$$\hat{y}(s_0) = \sum_{k=0}^{p} \beta_k \times x_k(s_0) + \sum_{i=1}^{n} \lambda_i \times \varepsilon(s_i)$$

Where $\hat{y}(s_0)$ is the estimated WS quantile at the target location (s_0), $x_k(s_0)$ are the values of the predictors at the target location, and β_k are the regression coefficients. λ_i are the ordinary kriging weights, and $\varepsilon(s_i)$ are the regression residuals at the sampled locations.

From the available data (207 samples), 155 samples (training set) were randomly selected for FS and fitting the RK model. The remaining 57 samples (test set) were used for the model evaluation. This procedure is a common practice in statistical modelling for the validation of the results (for instance, Qiu *et al.* (2022); Sun *et al.* (2023b)). It helps ensure unbiased assessment and generalization of the model's predictive capability. In addition, 10-fold cross-validation was performed on the training set, and the results were presented. Veronesi *et al.* (2016) used a similar validation procedure to validate their models for predicting WS distribution parameters at unsampled locations in the UK.

The coefficient of determination (R^2), the RMSE, the Relative Root Mean Squared Error (RRMSE), and the Mean Absolute error (MAE) were computed separately for each percent point considered in the study to evaluate the performance of the RK models during the cross-validation and with the test set:

Equation 3.19

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

Equation 3.20

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

Equation 3.21

$$RRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{\bar{y}}\right)^2}$$

Equation 3.22

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n}$$

3.4 Results

3.4.1 Wind speed quantiles

WS quantiles corresponding to 14 fixed percentile points for each location were estimated using shape-constrained P-Splines and the Weibull plotting position formula. Table 3.3 illustrates some statistics of the estimated WS quantiles in the training set.

Table 3.3 WS quantile statistics (P-Splines)

Percentile	Abbreviation	mean	std	min	25%	50%	75%	max
%		m/s	m/s	m/s	m/s	m/s	m/s	m/s
0.01	P1	19.68	5.72	7.92	15.49	19.54	23.29	45.58
0.1	P2	18.42	5.16	7.67	14.81	18.24	21.55	40.75
1	P3	12.72	3.92	5.75	9.90	12.17	15.07	29.75
5	P4	10.05	3.16	3.79	7.83	9.80	12.14	20.65
10	P5	8.63	2.65	3.15	6.84	8.43	10.48	17.51
20	P6	6.96	2.15	2.39	5.43	6.85	8.44	14.11
30	P7	5.85	1.84	2.00	4.61	5.73	7.08	11.88
40	P8	4.98	1.59	1.75	3.92	4.91	5.93	10.11
50	P9	4.24	1.39	1.56	3.33	4.17	5.12	8.61
60	P10	3.57	1.21	1.34	2.79	3.46	4.35	7.30
70	P11	2.92	1.02	1.07	2.28	2.77	3.53	6.08
80	P12	2.28	0.82	0.80	1.74	2.15	2.73	4.94
90	P13	1.55	0.58	0.51	1.19	1.45	1.87	3.41
95	P14	1.08	0.42	0.35	0.79	1.03	1.30	2.34

3.4.2 Model performances

The average R^2, RMSE, RRMSE, and MAE of the cross-validation with the training and the test set are listed in Table 3.4 and Table 3.5, respectively. When evaluated by cross-validation, the average R^2 ranges between 0.18 and 0.50, and the average RRMSE ranges between 22.4% and 33.1%. On the test set, the average R^2 ranges between 0.14 and 0.60, and the average RRMSE ranges between 20.7% and 35.5%. Model performance measured by cross-validation showed that GAGL was the best-performing FS algorithm, followed by MRMR, ENET and LASSO. On the test set, ENET, LASSO, and MRMR were the best-performing FS methods, and GALG and RFES had relatively medium performances. FSWR was the worst-performing FS method during cross-validation and with the test set.

A two-sample t-test ($H_0 : \mu_{\Delta RRMSE} \geq \mu_0$, $H_1 : \mu_{\Delta RRMSE} < \mu_0$) was conducted to assess the difference between the expected RRMSE ($\mu_{\Delta RRMSE} = \mu_{1RRMSE} - \mu_{2RRMSE}$) of pairs of FS methods on the test set. The results are presented in Table 3.6. The expected RRMSE of FSWR is significantly superior to the expected RRMSE of all the other FS methods. Also, ENET, LASSO, and MRMR performances were not significantly different when considering the RRMSE. However, ENET, LASSO, and MRMR performances were significantly superior (lower RRMSE) to GALG and RFES at the significance level of α = 0.05. There was no statistically significant difference between the expected RRMSE of GAGL and RFES.

Table 3.4 Performance of FS methods with cross-validation on the training set

FS method	R^2	RMSE	RRMSE	MAE
	-	m/s	-	m/s
ENET	0.410	1.668	0.246	1.238
FSWR	0.125	2.978	0.347	1.699
GALG	0.510	1.432	0.222	1.120
LASSO	0.408	1.659	0.246	1.239
MRMR	0.417	1.645	0.244	1.230
RFES	0.317	1.702	0.272	1.238

Table 3.5 Performance of FS methods on the test set

FS method	R^2	RMSE	RRMSE	MAE
	-	m/s	-	m/s
ENET	0.559	1.312	0.21	0.911
FSWR	0.137	2.307	0.353	1.555
GALG	0.491	1.438	0.233	1.002
LASSO	0.596	1.231	0.207	0.869

MRMR	0.602	1.226	0.211	0.847
RFES	0.459	1.506	0.24	1.053

Table 3.6 Results of the t-test between the expected RRMSE of pairs of FS methods

		μ_{2RRMSE}					
	FS method	ENET	FSWR	GALG	LASSO	MRMR	RFES
μ_{1RRMSE}	ENET		-4.69	-2.03	0.59	-0.18	-1.88
	FSWR	4.69*		3.46*	4.77*	4.51*	3.14*
	GALG	2.03*	-3.46		2.26*	2.21*	-0.78
	LASSO	-0.59	-4.77	-2.26		-0.91	-2.09
	MRMR	0.18	-4.51	-2.21	0.91		-1.99
	RFES	1.88*	-3.14	0.78	2.09*	1.99	

***: $\mu_{1RRMSE} - \mu_{2RRMSE}$ is significantly greater than 0 at α = 0.05**

The RRMSE of the FS methods is presented in Figure 3.2 for the standalone multilinear regression model (REG) and the regression-kriging model (ROK). The kriging of the regression model residuals led to a slight improvement in the performance metric. On average, the residual kriging decreased the RRMSE by 4%.

Figure 3.3 presents the RRMSE of the different WS quantiles. The model performance deteriorated as the probability of exceedance increased. For example, the mean RRMSE for the estimation of P1 is 17.0% (excluding FSWR), 21.4% for P9 (excluding FSWR), and 29.3% for P14 (excluding FSWR). FSWR performed relatively poorly for the estimation of P1 to P8 and improved for the estimation of P9 to P14.

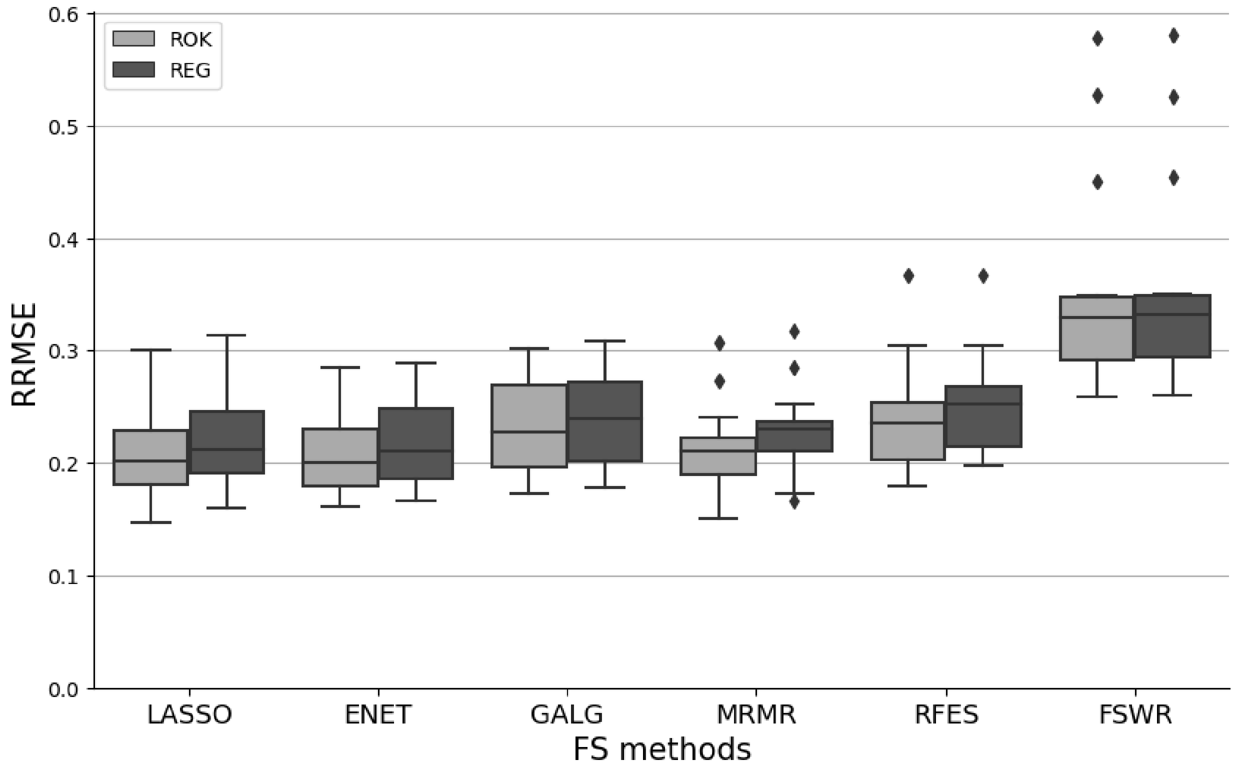


Figure 3.2 RRMSE of the standalone multilinear regression model (REG) and the regression-kriging model (ROK)

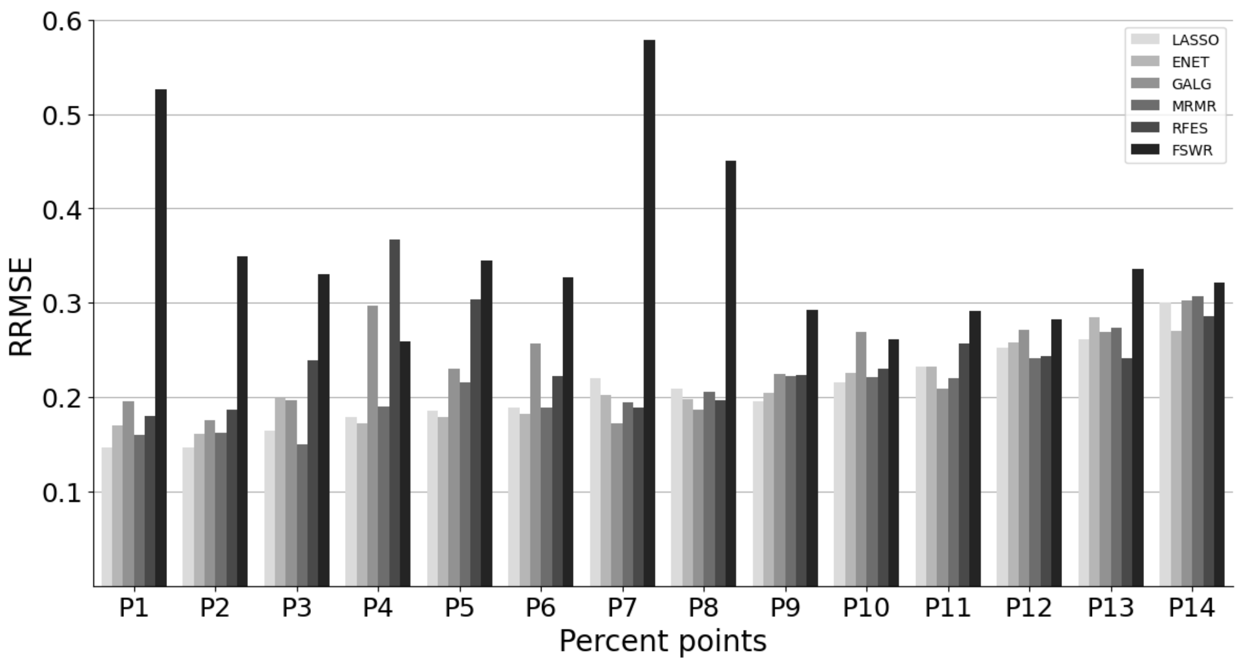


Figure 3.3 RRMSE of the FS methods for the estimation of different WS quantiles

3.4.3 Parsimony and multicollinearity

Figure 3.4 presents the mean number of selected features for each FS method. On average, the FSWR (44) method was the least sparse of the algorithm, followed by RFES (18) and GAGL (17). LASSO selected, on average, five features and was the sparsest FS method, followed by ENET (9) and MRMR (11). Figure 3.5 illustrates the mean number of selected features against the mean RRMSE. In general, the performance of the FS methods decreased (the RRMSE increased) as the number of selected features increased. Although FSWR selected many features, the model's performance remained relatively poor. As seen previously, the performance of ENET, LASSO and MRMR were not statistically different (two-sample t-test of the expected RRMSE), but LASSO was, on average, slightly more parsimonious.

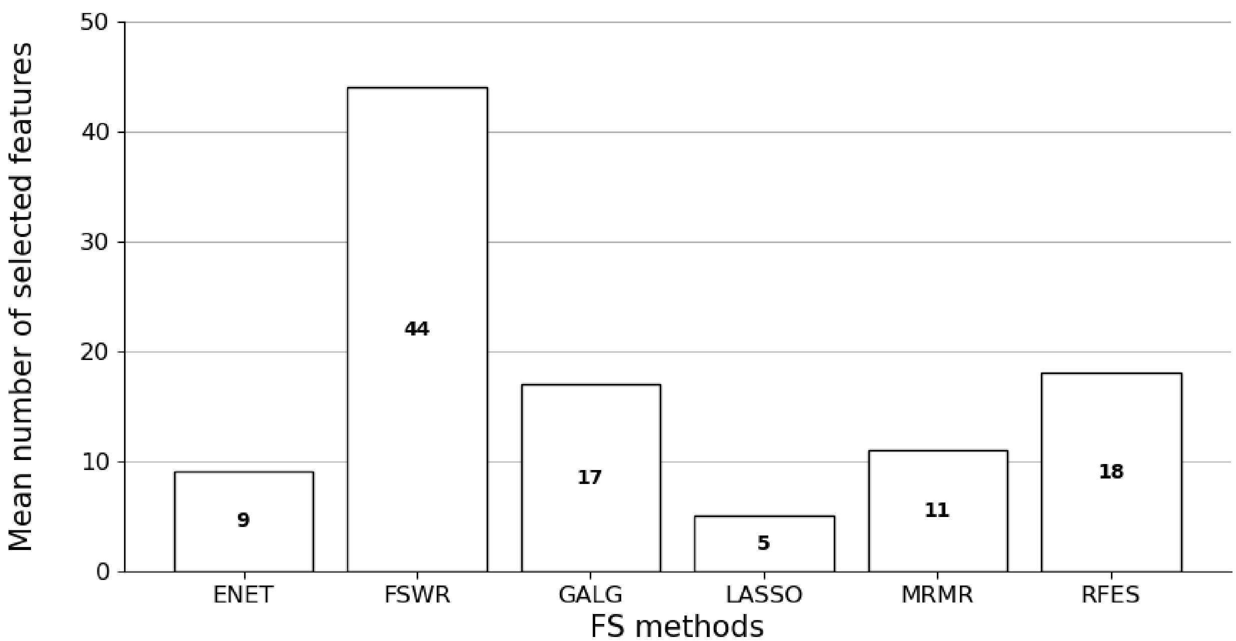


Figure 3.4 Mean number of selected features of each FS method

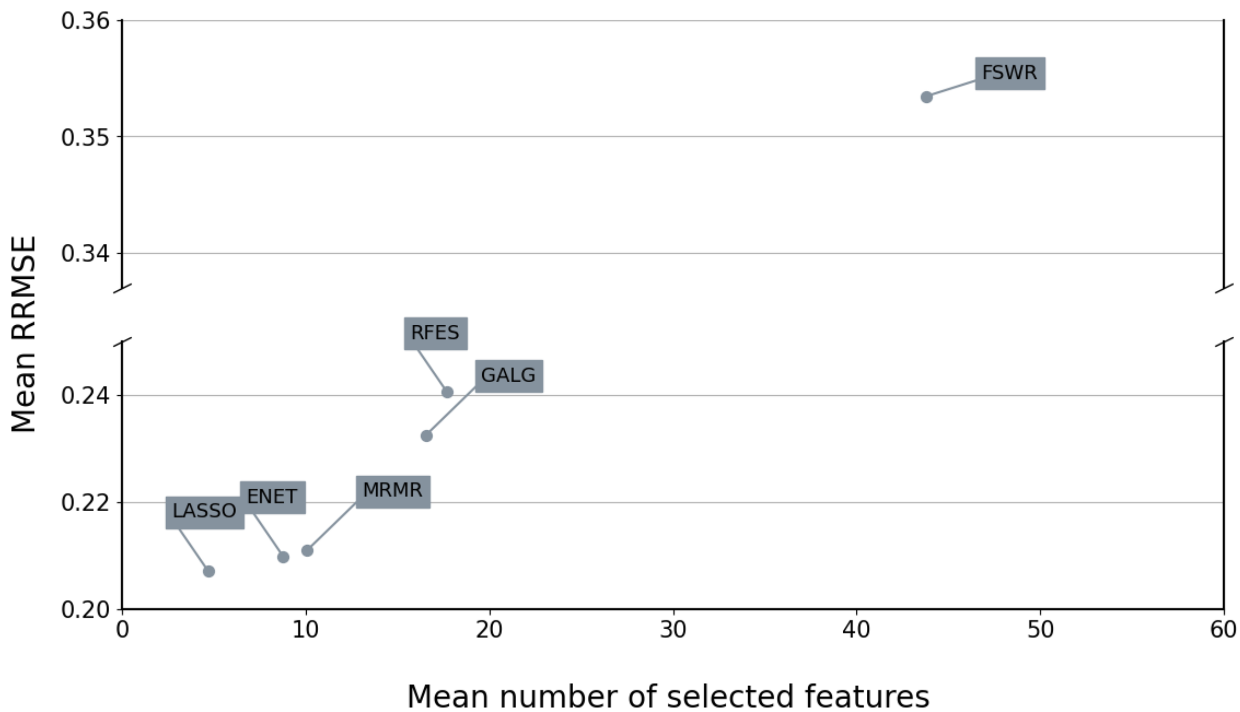


Figure 3.5 Mean number of selected features vs. mean RRMSE

The condition number (C) is a measure used to evaluate the presence of multicollinearity in a set of predictors. It is defined as the square root of the ratio between the maximum and the minimum eigenvalue of the predictor's correlation matrix. It is a single value summarising the likelihood of multicollinearity. Figure 3.6 shows the condition number estimated from the correlation matrix of the selected feature sets. From empirical observations, Chatterjee *et al.* (2013) suggested a cut-off of 15 to detect multicollinearity and recommended corrective action if C exceeds 30. All the feature sets estimated with LASSO had a condition number below 15. In the case of MRMR, the condition numbers were less than 15 in 13 cases out of 14 (92.8%) and were consistently below 30. For ENET and GAGL, the condition number was less than 15 in 11 cases out of 14 (78.6%). RFES condition numbers were inferior to 15 in 8 cases out of 14 (57.1%), and FSWR condition numbers consistently exceeded 15.

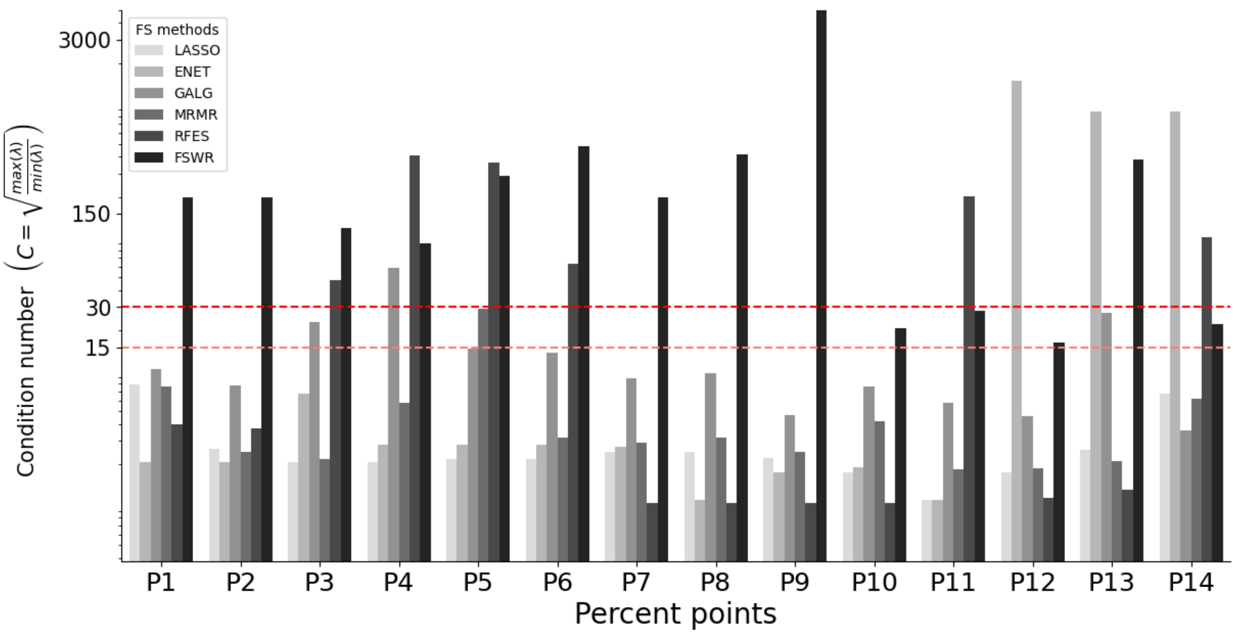


Figure 3.6 **Condition number (C) of the selected feature sets**

3.4.4 Residual analysis and visual inspection

Model residual analyses compare observed data with predicted values to evaluate a model's precision and reliability. Examining residuals can reveal patterns, outliers, and areas for improvement in the model's assumptions. Figure 3.7 compares the observed and predicted WS quantiles for the top-performing FS methods (MRMR and LASSO), indicating a strong agreement between the observed and estimated quantile for both methods with an R^2 of approximately 0.92. LASSO performed slightly better than MRMR, as indicated by the RRMSE. Two outliers were identified in the bottom-right section of the plots, with an underestimation of the WS quantiles for both outliers. The residual plot in Figure 3.8 confirmed that the models did not perform as well for high exceedance probabilities as they did for the lower ones.

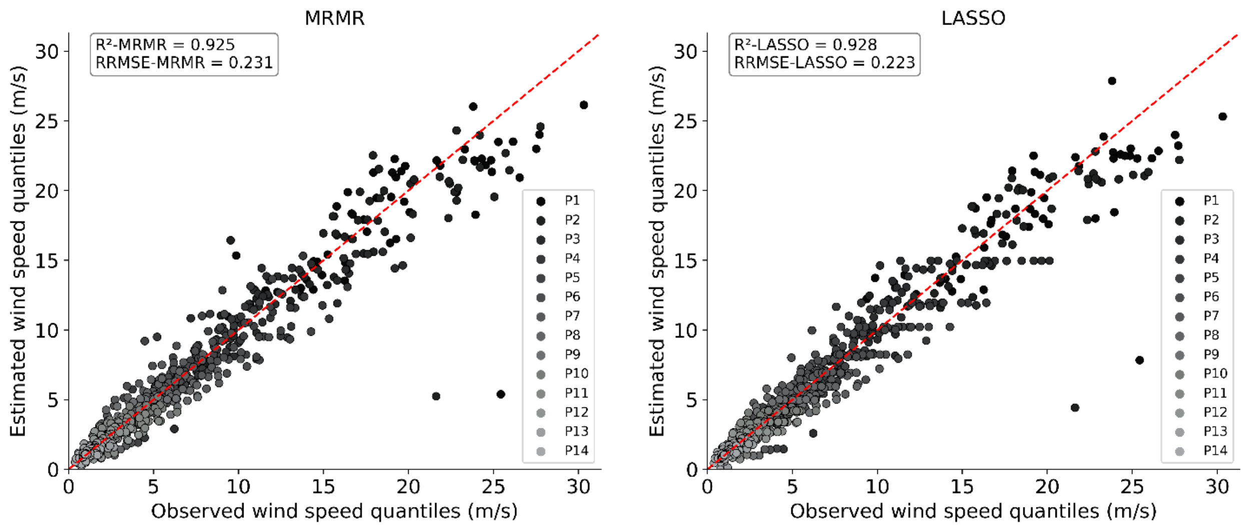


Figure 3.7 **Plot of observed vs. estimated WS quantiles for MRMR and LASSO. The R^2 was calculated without averaging across the percent points**

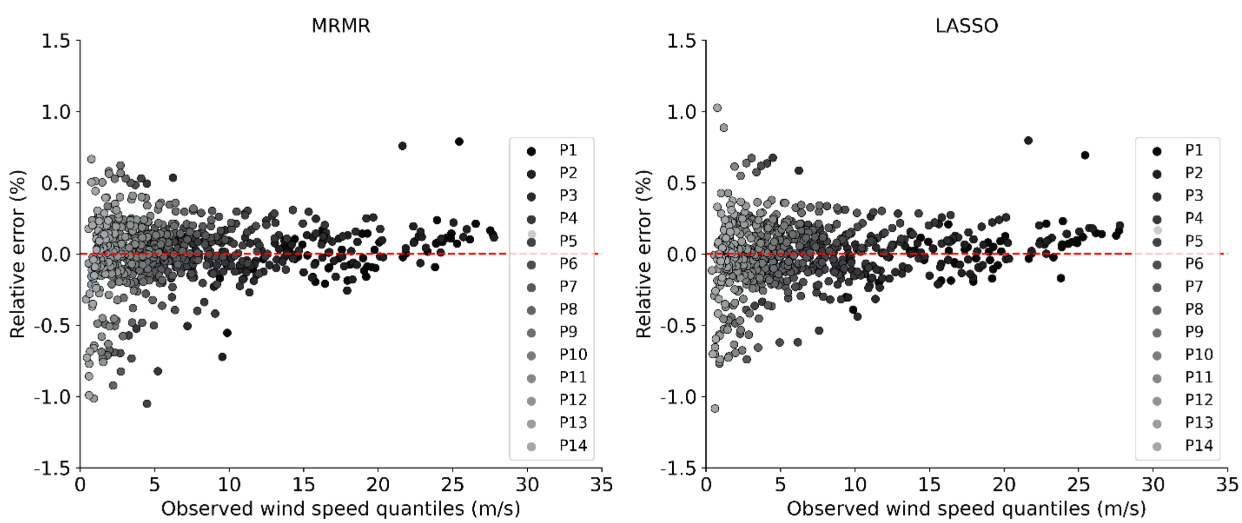


Figure 3.8 **Plot of observed WS quantiles vs. the residuals for MRMR and LASSO**

3.4.5 Predictor importance

Figure 3.9 shows the ten most selected features for each WS quantile and the number of times they were selected. Overall, the most selected features were RGLH and DSEA. DSEA was consistently selected by every FS method. For the surface roughness length (RGLH), 2000m and 1000m (RGLH_2000m and RGLH_1000m) were the most selected spatial scales. RGLH at 100m spatial scale (RGLH_100m) was mostly selected for medium to high exceedance probabilities. DVME at a spatial scale of 2000m (DVME_2000m) was often selected for high exceedance probabilities (P10 to P14) and less often selected for lower exceedance probabilities (P1 to P9). Predictors describing the land surface curvature (MXCV, MNCV, TLCV, TGCV, GSCV) seemed important for predicting WS quantiles corresponding to very low exceedance probabilities (P1 to P5) and less important for medium and high exceedance probabilities. PNCL was also among the most selected predictors, especially class 7 of PNCL (PNLC_7), which indicates a level terrain at the grid cell with a low slope gradient. The location coordinates (XGEO and YGEO) were also often selected for different WS quantiles in the region.

The predictors selected by the FS methods were used to fit a simple linear regression model. An advantage of the simple linear regression model is the interpretability of the model. Without multicollinearity, the regression coefficient magnitude and direction provide useful information to assess the relationship between the predictors and the dependent variable. Figure 3.10 shows the regression coefficient of the predictors selected with LASSO. The predictors were standardized to a zero mean and a unit variance prior to fitting the regression model. LASSO was the most parsimonious FS method with good predictive ability. In addition, the estimated condition numbers of all the feature sets selected by LASSO were below 15, indicating the absence of multicollinearity. It is observed that DSEA regression coefficients were often the strongest and were always negative. DSEA represents the location distance from the coast; the direction of the regression coefficient showed that WS quantiles, irrespective of their exceedance probabilities, were higher near the coast than inland. The surface roughness length (RGLH) showed relatively high regression coefficients with every WS quantile. The negative direction of the regression coefficient of RGLH is intuitive. An increase in surface roughness results in more friction between the land surface and the wind, decreasing WS near the ground. For P1 and P2, the maximum curvature (MXCV) had the second-highest regression coefficient with a positive direction. Note that higher values of MXCV correspond to elongated convex landforms such as ridges, and negative values are associated with concave landforms (Florinsky, 2017). The positive magnitude of the MXCV regression coefficient showed that the WS quantiles P1 and P2 were higher at locations where the landforms are convex and decreased as the landform concavity increased.

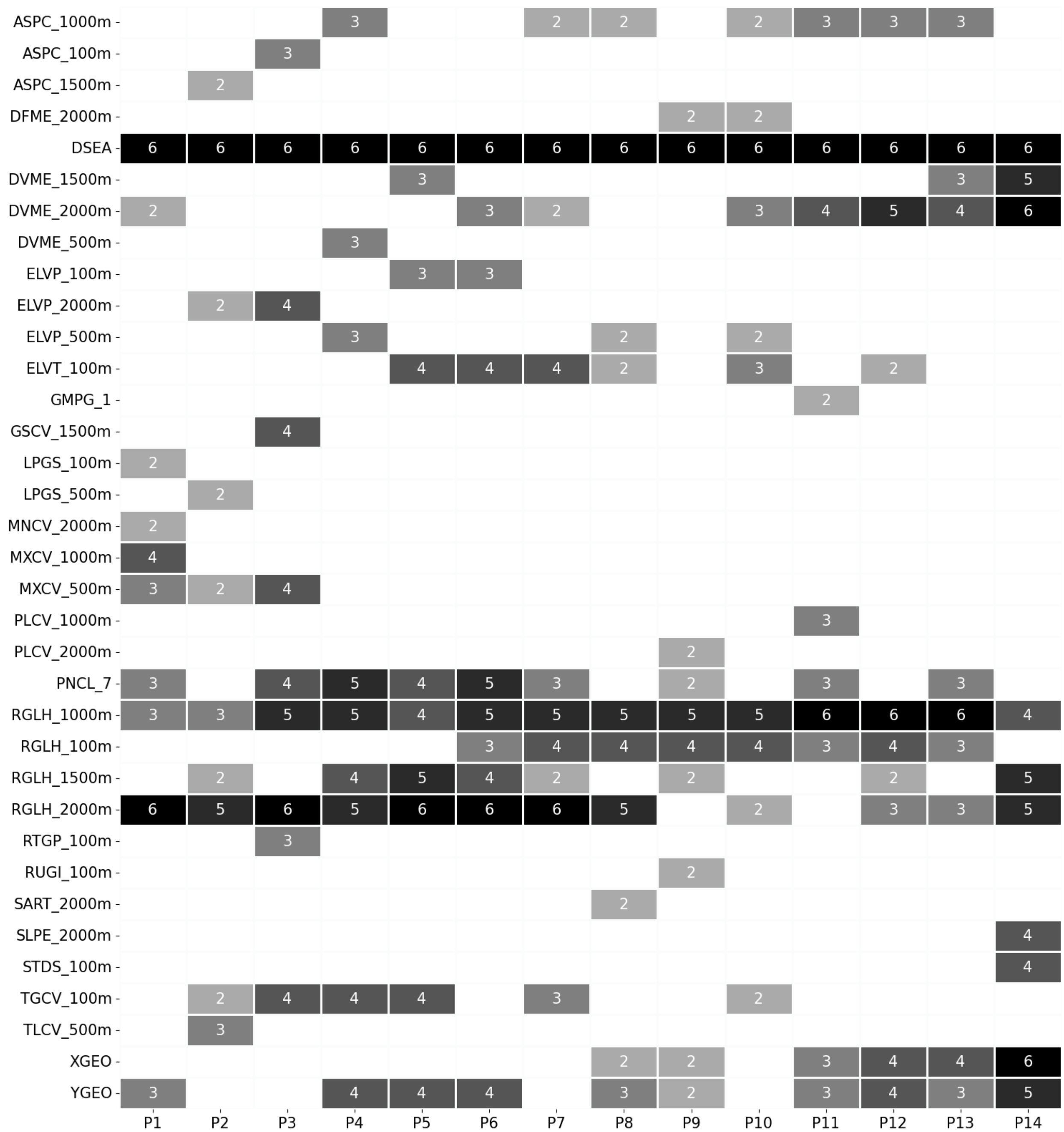


Figure 3.9 **Selected predictors for each WS quantile**

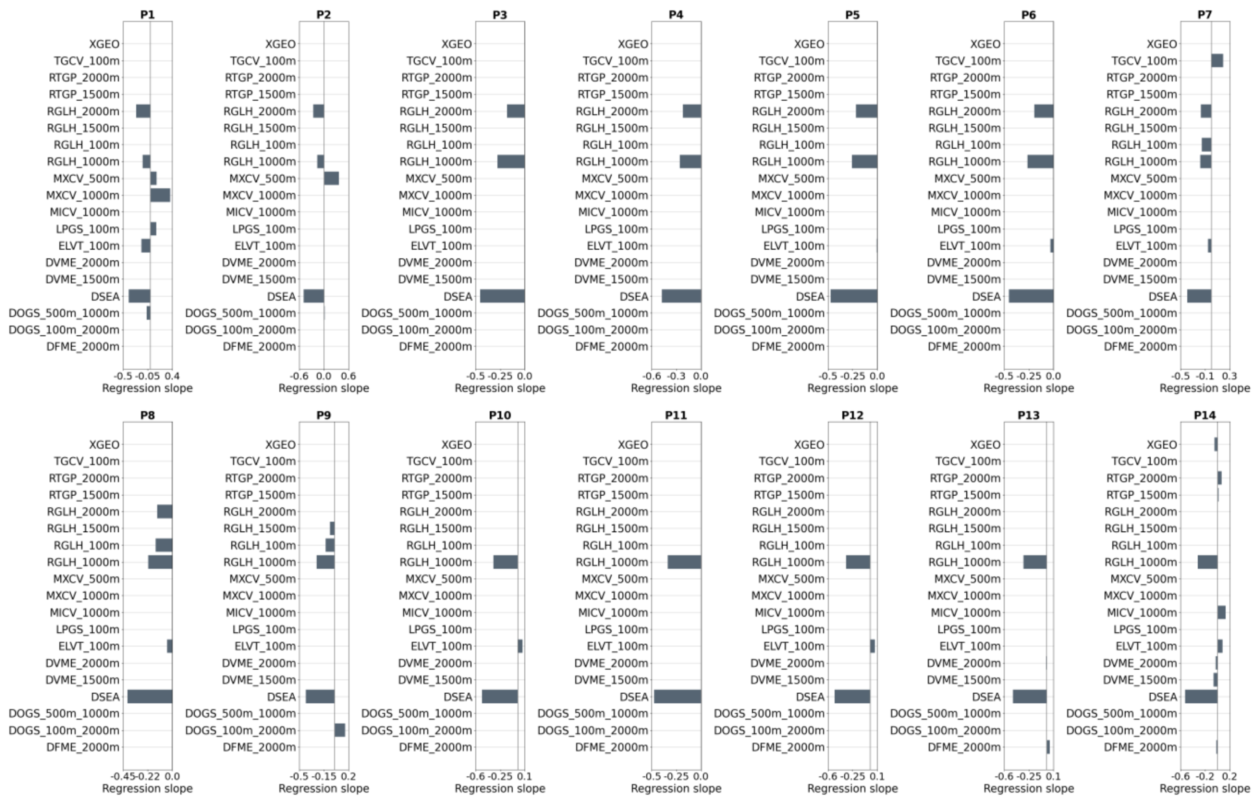


Figure 3.10 Regression coefficients of the WS quantile predictors

3.5 Discussion

This study compared six FS methods for WS quantile estimation in Canada. The results showed that LASSO, MRMR, and ENET had comparable performances on the test set and were the most effective FS methods. GAGL and RFES performed slightly worse than LASSO, MRMR, and ENET but outperformed FSWR. The FSWR method does not seem to ignore redundant features, leading to an unstable estimation of regression coefficients and poor performance during testing. This situation seems more pronounced for low than high exceedance probabilities (P10 to P14). There was less collinearity among the relevant predictors associated with high exceedance probabilities than for lower ones. Kriging of the regression residual slightly improved the model performances (4%), indicating that the selected predictors and the linear regression model could account for a significant portion of the spatial variability of WS quantiles in the region.

The models' performances were higher for low to medium exceedance probabilities and declined for high exceedance probabilities. This decline in performance could be attributed to several factors. One possible explanation is that there is a significant non-linear relationship between high exceedance probabilities WS and the predictors, requiring the implementation of non-linear models for improved performance. Another possible explanation is the exclusion of significant predictors of high exceedance probabilities from the models. For example, the models did not include climate-related predictors such as mean temperature or pressure. Climatic variables are often collected at meteorological stations where WS is also measured; thus, they should also be missing at locations with unavailable WS data. The results highlight the need for further research to enhance the performance of models in predicting high exceedance probability WS.

LASSO was found to produced, on average, the sparsest feature sets, followed by ENET and MRMR. In addition, LASSO could select relevant predictors without multicollinearity as evaluated by the feature set correlation matrix condition number. MRMR also eliminated multicollinearity in most cases (13 out of 14 cases), while ENET, RFES, GAGL, and FSWR were less effective at solving the issue of multicollinearity in their selected feature sets. These findings are consistent with existing literature on RFES, Xie *et al.* (2006) showed that this implementation does not consider feature redundancy. Overall, LASSO and MRMR were the most effective FS methods due of the following reasons:

- LASSO and MRMR exhibited high predictive ability, with no significant difference in performance between the two methods based on *t*-test results and residual analysis.
- Both FS methods could select relevant predictors while also reducing multi-collinearity within the feature subset.
- LASSO and MRMR are attractive because they are efficient to implement with a single parameter to tune, unlike ENET, which produced comparable performance. In the case of LASSO, the degree of penalization (α) is the only parameter that needs to be tuned. With MRMR, the number of features to select is the single tuning parameter of the algorithm. ENET requires the tuning of two parameters.

LASSO and MRMR have different approaches to feature selection. However, their good performance in the study could be explained by their inbuilt capability to select relevant features while ignoring redundant ones. LASSO is a penalization algorithm based on linear regression that promotes sparsity by imposing a penalty on the sum of the absolute values of the feature coefficients. In a group of redundant predictors, LASSO chooses one predictor among the group and shrinks towards zero the coefficients of the other predictors (Hammami *et al.*, 2012; Zou *et al.*, 2005), making it effective in dealing with collinear features.

On the other hand, MRMR ranks features from the most relevant and least redundant to the least relevant and most redundant, allowing for efficient selection of the smallest subset of the most relevant and least redundant features that provides the best cross-validation score. In addition, MRMR is a filter-based approach that is agnostic to any specific regression model, as it is based on the correlation coefficient. This coefficient is well suited for the linear regression model used in this study. However, other correlation metrics, such as mutual information, can be used for nonlinear models.

It is worth noting that GAGL showed superior performance during cross-validation on the training sets, and there was no significant decline in its performance on the test set. However, in some feature subsets selected by GAGL, the issue of multicollinearity remained unresolved. In addition, compared to LASSO and MRMR, GAGL has more parameters that require tuning, making it less efficient to implement.

In the present study, the location distance from the coast (DSEA) and the surface roughness length (RGHL) were the two most significant predictors of WS quantiles. The regression model coefficients for both DSEA and RGHL were physically consistent. In the case of DSEA, the regression coefficients were negative, indicating a decrease in the WS quantiles from coastal to inland areas. Few studies have used the distance from the coast to estimate WS, but it could be a valuable addition to models, particularly in larger study areas. For low exceedance probabilities (e.g., 1%), surface convexity (concavity) was a significant predictor of WS, but it was less relevant for higher exceedance probabilities.

There are some limitations to this study. The dataset contained only 207 samples (155 training and 52 testing samples), and some regions of Canada were naturally less densely represented (see Figure 3.1). Consequently, some results could be particular to the studied region or the analyzed dataset and may only be generalized after extensive analysis.

Among the various feature selection (FS) methods examined, the FSWR approach was the least effective. It is possible to improve the FSWR method performance by adding the variance inflation factor as a post-processing step. It should be noted that, the FSWR method in this study was mainly used as a benchmark for assessing the performance of other proposed FS methods as it remains one of the most common FS methods.

In the present study, the time series were considered stationary when estimating WS quantiles. Nevertheless, increased evidence points to non-stationarities in WS series and the importance of incorporating them in the analysis (see, for instance, Ouarda *et al.* (2021)). For instance, several authors observed significant correlations between low-frequency climate oscillation indices and annual mean WS in different regions of the world (see, for instance, Naizghi *et al.* (2017); Woldesellasse *et al.* (2020)); Including these climate oscillation indices in quantile estimation or regional transfer models could significantly improve their performances. Indeed, in a given region, WS stations are impacted by the same climate oscillation indices, and their incorporation in the models used to estimate WS at ungauged locations should lead to performance improvements. The issue of incorporation of teleconnections in WS estimation models is an important one but remains mainly unexplored in the literature. Future efforts should focus on incorporating non-stationarities in regional WS estimation models.

3.6 Conclusion

This paper evaluated six FS methods for WS quantile estimation. LASSO and MRMR were the most efficient algorithms in the study. It was found that the importance of some WS quantile predictors depends on their exceedance probability. The location distance from the coast and the surface roughness length were significant WS quantile predictors irrespective of the exceedance probability.

Future research should focus on the extrapolation of this study to other geographic regions, databases with different characteristics, and other FS methods. The diversity in the characteristics needs to be ensured to obtain guidelines for the relative performance and the applicability of different techniques based on such considerations as the number of sites, the length of the series, the number of features, the types of wind, the data variability and quality, etc.

Nomenclature

Abbreviations

ALOS	Advanced Land Observing Satellite
ANN	Artificial Neural Networks
ASPC	Aspect
DEM	Digital Elevation Models
DFME	Difference from mean elevation
DOGS	Differences of Gaussians
DSEA	Distance from the coast
DVME	Deviation from mean elevation
ECCC	Environment and Climate Change Canada
ELVP	Elevation percentile
ENET	Elastic Net
FS	Feature selection
FSWR	forward stepwise regression
GALG	Genetic algorithm
GMPG	Geomorphologic phenotypes
GSCV	Gaussian curvature
GW	Giga watt
LASSO	Least Absolute Shrinkage and Selection Operator
LPGS	Laplacian of Gaussian
MAE	Mean Absolute error
MICV	Minimal curvature
MNCV	Mean curvature
MRMR	Maximum relevance Minimum redundancy
MXCV	Maximal curvature
PLCV	Plan curvature
PNCL	Pennock landform class
P-Splines	Penalized splines
R^2	Coefficient of determination
REG	Standalone multilinear regression model
RFES	Recursive feature elimination using support vector regression
RGLH	Roughness length

RK	Regression kriging
RMSE	Root mean squared error
ROK	Regression-kriging model
RRMSE	Relative Root Mean Squared Error
RTGP	Relative topographic position
RTGP	Relative topographical position
RUGI	Ruggedness index
SART	Surface area ratio
SLPE	Slope
STDS	Standard deviation of the slope
SVM	Support Vector Machines
SVR	Support vector regression
TGCV	Tangential curvature
TLCV	Total curvature
vs.	Versus
WS	Wind speed
XGEO	Longitude
YGEO	Latitude

Symbols

α	Tuning parameter that controls the degree of penalization in the LASSO model
b	Bias in the epsilon-insensitive SVR model
β	Regression coefficients
C	Regularization parameter in the epsilon-insensitive SVR formulation
ε	Tolerance margin in the epsilon-insensitive SVR model
$F(x, y)$	F-statistic measuring the relationship between x and y
$f(\cdot)$	Function to optimize
F_i	Fitness score of the i-th solution of the genetic algorithm
$J(w)$	Norm of weight vector w in the epsilon-insensitive SVR model
L	Lagrange function in the epsilon-insensitive SVR formulation
$\alpha_i \alpha_i^*, \eta_i \eta_i^*$	Non-negative Lagrange multipliers in the epsilon-insensitive SVR formulation

λ	Tuning parameter in the elastic net model
$\mu_{\Delta RRMSE}$	Mean of the expected RRMSE differences in the t-test.
μ_0	Population mean in the t-test ($\mu_0 = 0$)
NO_2	Nitrogen dioxide
P1	Wind speed quantile at 0.01% percentile-level
P10	Wind speed quantile at 60% percentile-level
P11	Wind speed quantile at 70% percentile-level
P12	Wind speed quantile at 80% percentile-level
P13	Wind speed quantile at 90% percentile-level
P14	Wind speed quantile at 95% percentile-level
P2	Wind speed quantile at 0.1% percentile-level
P3	Wind speed quantile at 1% percentile-level
P4	Wind speed quantile at 5% percentile-level
P5	Wind speed quantile at 10% percentile-level
P6	Wind speed quantile at 20% percentile-level
P7	Wind speed quantile at 30% percentile-level
P8	Wind speed quantile at 40% percentile-level
P9	Wind speed quantile at 50% percentile-level
P_i	Probability of exceedance associated with wind speed value of rank i
$PM_{2.5}$	Fine particle matter
$Psel_i$	Probability of selecting the i-th solution in the genetic algorithm
$\rho(x, y)$	Pearson correlation coefficient between x and y
w	Weight vector in the epsilon-insensitive SVR model
w_1	Genetic algorithm weight parameter that controls the balance between performance and the number of selected features
$\xi_i \xi_i^*$	slack variables in the epsilon-insensitive SVR formulation

4 A NON-PARAMETRIC APPROACH FOR WIND SPEED DISTRIBUTION MAPPING

Titre de l'article : Une approche non paramétrique pour l'interpolation spatiale de la distribution des vitesses du vent

Auteurs :

Freddy Houndekindo[1] (F.H.), Taha B.M.J. Ouarda[1] (T.B.M.J.O)

[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

Titre de la revue:

Article publié le 15 novembre 2023, dans la revue **Energy Conversion and Management**, volume 296, DOI : 10.1016/j.enconman.2023.117672

Contribution des auteurs :

Conceptualisation, F.H., T.B.M.J.O. ; Méthodologie, F.H., T.B.M.J.O. ; Préparation des données, F.H. ; Logiciel, F.H. ; Rédaction de la version originale, F.H. ; Révision, T.B.M.J.O. ; Supervision, T.B.M.J.O. ; Acquisition du financement, T.B.M.J.O.

Dans la revue de littérature, nous avons constaté que les méthodes actuelles d'estimation de la distribution de probabilité de la vitesse du vent se basent sur l'utilisation d'un seul type de loi de distribution pour toute la région étudiée. Cette approche peut s'avérer moins fiable dans les régions où la variabilité spatiale du régime des vents est importante. L'article 3 propose une méthode plus flexible, basée sur l'interpolation spatiale des quantiles de vitesse du vent et l'estimation par noyau asymétrique, une méthode non paramétrique. Cette approche découle en partie de notre exploration, dans l'article 2, de la possibilité d'interpoler spatialement les quantiles de vitesse du vent à partir de différentes variables explicatives.

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.

4.1 Introduction

Wind energy has the potential to become a crucial source of power worldwide (Zhou *et al.*, 2012). In 2021, worldwide wind energy installed capacity reached 837 GW, with an estimated offset of over 1.2 billion tons of CO2 (Council, 2022). 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 (Jung *et al.*, 2018b).

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 wind speed (WS) estimation studies have been conducted at unsampled locations, as detailed in the review by Houndekindo *et al.* (2023c). These studies typically estimate an aggregated WS value (Luo *et al.*, 2008; Ye *et al.*, 2015), 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 (Nelson *et al.*, 2018). 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.* (2016) 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 (Jung *et al.*, 2019b). Tsvetkova *et al.* (2023) 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 (2016) 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 *et al.* (2016) 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 (Alavi *et al.*, 2016; Aries *et al.*, 2018; Ouarda *et al.*, 2018; Ouarda *et al.*, 2015; Safari, 2011; Zhou *et al.*, 2010) 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 (Han *et al.*, 2019). In most of these studies, symmetric kernels (e.g., 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 (Węglarczyk, 2018). This is a well-known problem called the boundary effect, and several solutions have been proposed (Mombeni *et al.*, 2021). In this study, one of these solutions based on asymmetric kernel estimators (Hirukawa, 2018) is adopted and introduced for WS distribution modelling. According to Hirukawa (2018), 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 4.2 illustrates the methodology of the proposed approach with the evaluation procedure. The study area and the dataset are presented in section 4.3. The results obtained are shown in section 4.4. In sections 4.5 and 4.6, the discussion of the findings and the conclusion are given, respectively.

4.2 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 (Veronesi et al., 2016). 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 4.1.

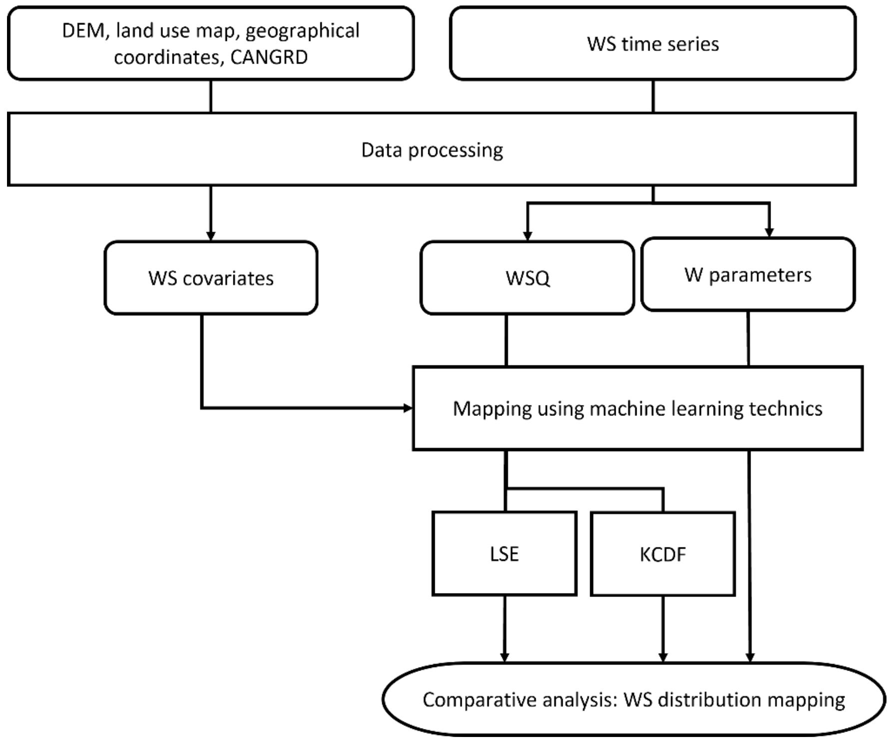


Figure 4.1 Methodology of the comparative analysis of WS probability distribution mapping approaches

4.2.1 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 (Hyndman *et al.*, 1996):

Equation 4.1

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

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 (Hyndman *et al.*, 1996). 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 (Reiss, 1989). Using Equation 4.1, 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 8.1 in the Appendix (Section 8) 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, (He *et al.*, 2017) modelled 20 percentiles evenly spaced between 1% and 96%. In another study, to map wind speed shear distribution, (Jung *et al.*, 2018a) estimated 11 percentiles evenly spaced between 1% and 99%. Additionally, to regionalize river temperature at ungauged locations, (Ouarda *et al.*, 2022) 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 Gradient boosting trees (GBT: Friedman, 2001) model. Feature selection (FS) was performed using the minimum redundancy maximum relevance (MRMR) method (Ding *et al.*, 2005) to reduce the complexity of the models and improve their performance. A comparative study of FS methods was carried out by Houndekindo *et al.* (2023a). They found that MRMR was among the most effective FS methods for WSQ estimation. Houndekindo *et al.* (2023a) 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.

4.2.1.1 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 (Guyon *et al.*, 2003). 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:

Equation 4.2

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

Equation 4.3

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

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

Equation 4.4

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

Equation 4.5

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

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:

Equation 4.6

$$I(X, Y) = \iint p(X, Y)\log\left(\frac{p(X, Y)}{p(X)p(Y)}\right)dxdy$$

The Python package scikit-learn (Pedregosa *et al.*, 2011) was used to calculate the MI between the variables.

4.2.1.2 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 (Grinsztajn *et al.*, 2022b). 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 (Hastie *et al.*, 2009). The GBT model is a popular regression model that has been successfully applied in studies for short-term wind power prediction (Ye *et al.*, 2022), wind resource mapping (Jung *et al.*, 2018a), the selection of solar power plant location (Sun *et al.*, 2023b) and short-term prediction of solar irradiance (Lee *et al.*, 2020).

The eXtreme Gradient Boosting package (XGB: Chen *et al.*, 2016) 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 (Turner *et al.*, 2021). 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 (Bergstra *et al.*, 2012b). Table 4.1 presents the hyperparameters of the XGB model that were tuned in the study.

Table 4.1 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)

4.2.1.3 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 (Ouarda *et al.*, 2021). In their study, Jung *et al.* (2018a) 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:

Equation 4.7

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

where $F\left(\widehat{W_i};\, \hat{\theta}\right)$ 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 (e.g., WS values are bounded on $[0, \infty]$). 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 (Mombeni *et al.*, 2021):

Equation 4.8

$$\hat{F}(w) = \frac{1}{n}\sum_{i=1}^{n} \overline{K}(W_i; w, b)$$

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 (Lafaye de Micheaux *et al.*, 2021; Mombeni *et al.*, 2021):

Equation 4.9

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

Equation 4.10

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

Equation 4.11

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

Equation 4.12

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

Equation 4.13

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

Equation 4.14

$$\bar{K}_{WB}(x; \alpha, \beta) = \exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right)\ \alpha, \beta\ > 0$$

Where $\Phi(\cdot)$ 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):

Equation 4.15

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

Equation 4.16

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

Mombeni *et al.* (2021) derived the asymptotical optimal bandwidth of $\overline{K}_{BS}$ and $\overline{K}_{WB}$ with respect to the MISE:

Equation 4.17

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

Equation 4.18

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

Lafaye de Micheaux *et al.* (2021) proposed the following asymptotical optimal bandwidth with respect to the MISE for $\overline{K}_{LN}$:

Equation 4.19

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

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.

4.2.2 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 (2016) 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.* (2016). Veronesi *et al.* (2016) 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 4.3 and the LR and XGB regression models described in section 4.2.1.2. The MRMR algorithm was also applied to identify the best set of covariates to include in the regression models.

4.2.3 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:

Equation 4.20

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

Equation 4.21

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

Equation 4.22

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

The evaluation of the GOF of the estimated WS probability distribution was based on the percentage probability plot (PP plot: Wilk *et al.*, 1968). 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 Equation 4.20, Equation 4.21, and Equation 4.22, respectively, were used to evaluate the degree of association between the ECDF and the CDF. Horst (2008) 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 *et al.* (2019b) 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:

Equation 4.23

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

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 (Akgül *et al.*, 2016) giving unbiased non-exceedance probabilities regardless of the underlying distribution of the data (Morgan *et al.*, 2011):

Equation 4.24

$$F_n(W_i) = \frac{i}{(n+1)}$$

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

4.3 Study area and dataset

The study was conducted on data from Canada with 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 stations with more than ten years of record having two months of missing data. Figure 4.2 illustrates the geographical location of the 207 stations that were selected after filtering. From the available stations, 155 (white triangles in Figure 4.2) were used for FS, model training and cross-validation and the remaining stations (black dots in Figure 4.2) were used to validate the final model as explained in 4.2.3.

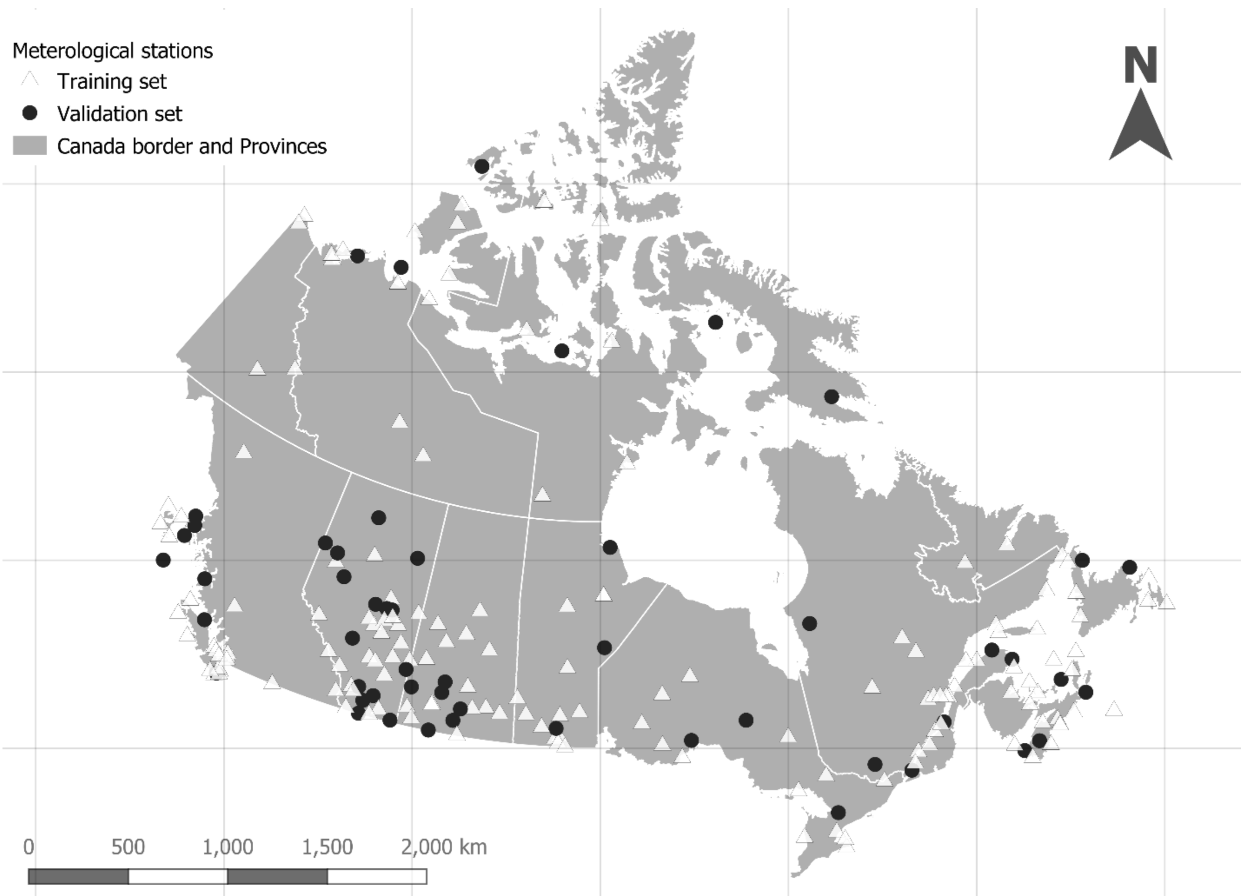


Figure 4.2 **Spatial distribution of the training and validation stations used in this study**

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 (Lindsay, 2014) and a 30 m resolution global DEM (Tadono *et al.*, 2014). 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 (Latifovic *et al.*, 2017) 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 Wiernga (1993). Table 8.2 in the Appendix (Section 8) provides more details about the covariates.

4.4 Results

4.4.1 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 Table 4.2 and Table 4.3 for QWSM and the WPM, respectively. Figure 4.3 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 Figure 4.3a and Figure 4.3b). Table 4.2 and Table 4.3 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 4.4 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 4.2 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 4.3 Average performance metrics for the estimation of the W parameters

Validation Methods	Regression model	MRMR	MAE	R²	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

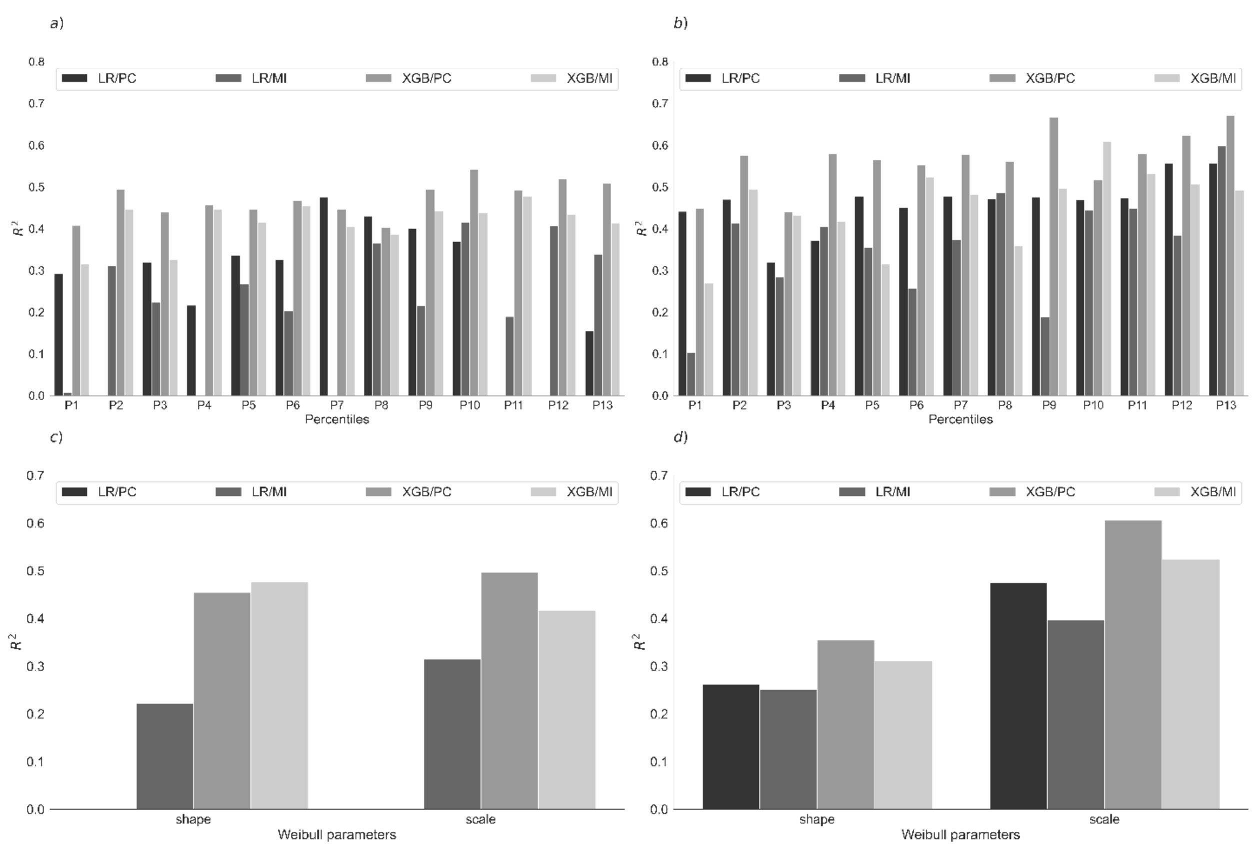


Figure 4.3 Performance of LR and XGB for the estimation of the WSQ (a and b) and the W parameters (c and d) during cross-validation (a and c) and validation (b and d).

Note: Negative values of R^2 were set to zero

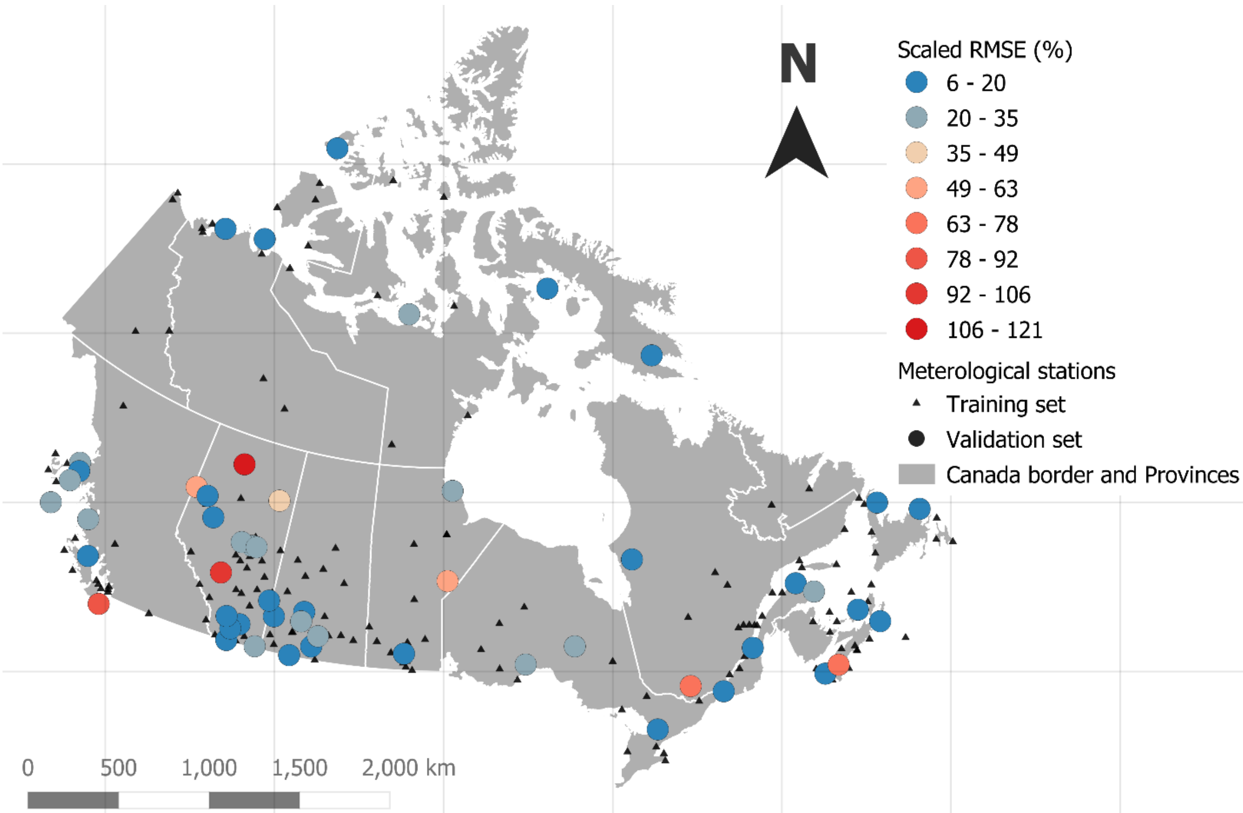


Figure 4.4 Spatial distribution of the scaled RMSE (WSQ) of the validation set

4.4.2 Wind speed distribution mapping

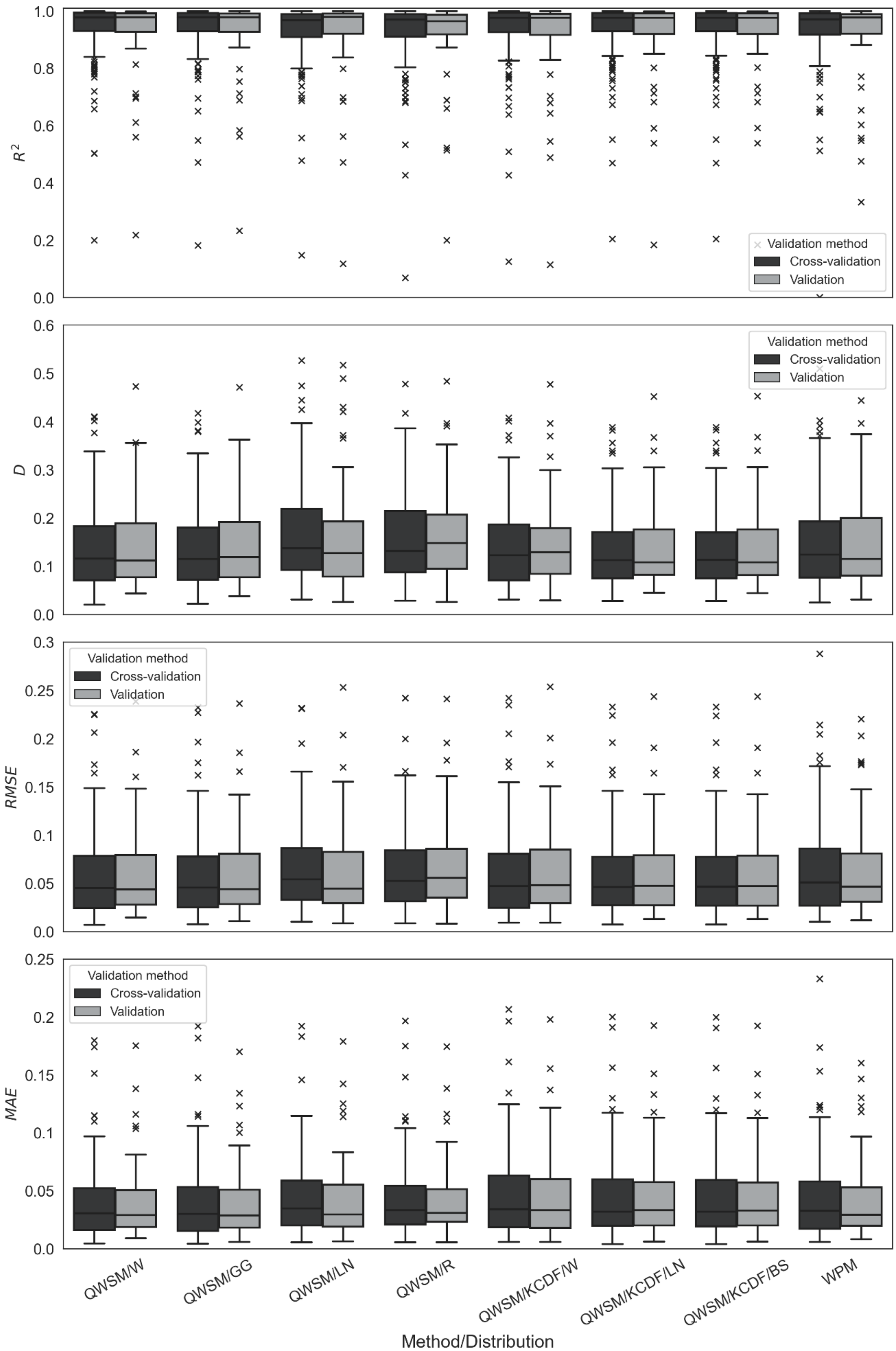
This section presents the results of the comparative analysis between the QWSM and WPM. Table 4.4 shows the mean values of the GOF metrics. In general, it is observed that the QWSM gave a better fit than WPM for the considered metrics. Also, QWSM/W gave better fit than WPM. According to the R^2, RMSE 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 4.5. 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 4.6) 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 (e.g., 0.4, 0.5, 0.6 percentiles). While in the tails (e.g., percentiles 0.1 and 0.9) the parametric methods showcased a better performance than the non-parametric methods.

Table 4.4 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.



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

Figure 4.5 GOF of estimated WS probability distribution. Negative values of R^2 were set to zero

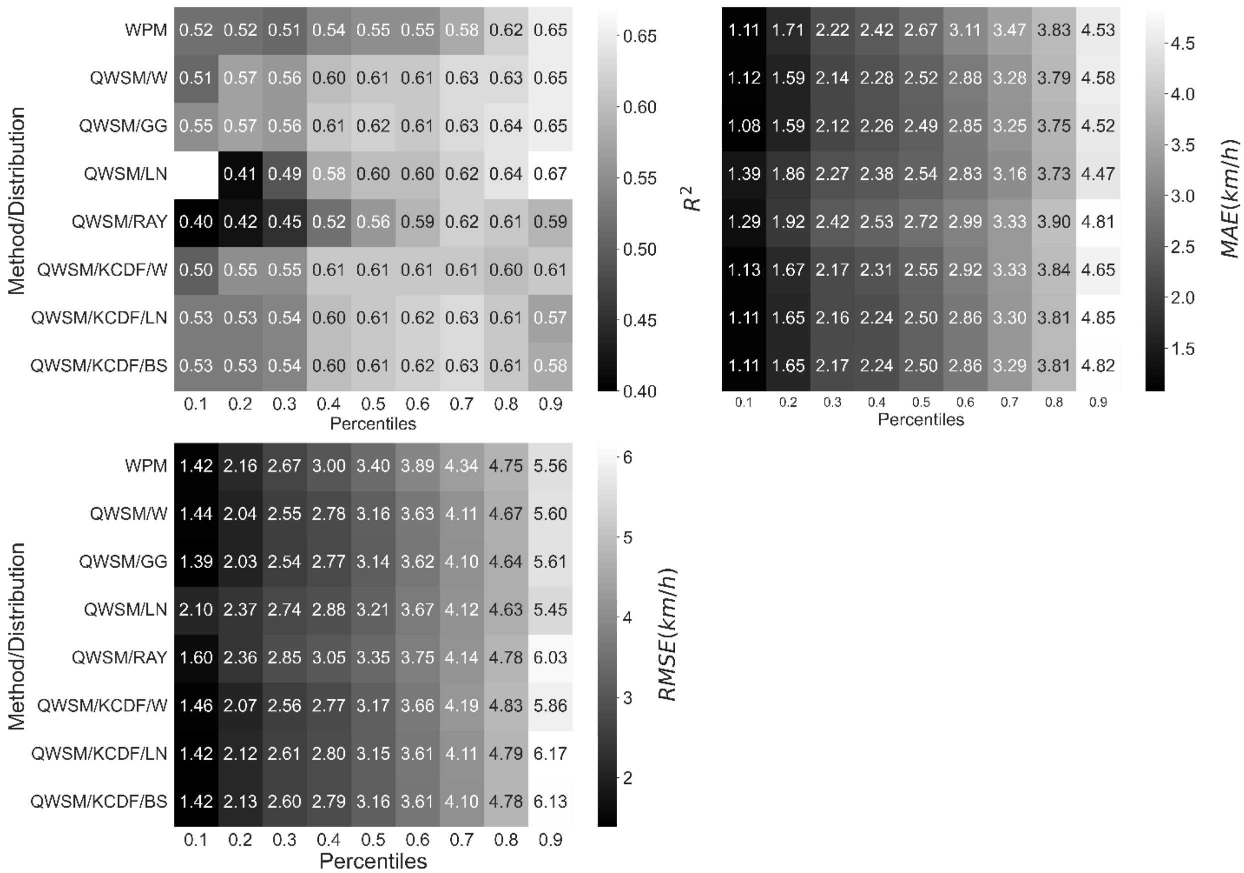


Figure 4.6 Performance metrics for observed WSQ and estimated WSQ using QWSM and WPM (validation set)

In Figure 4.7, 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.

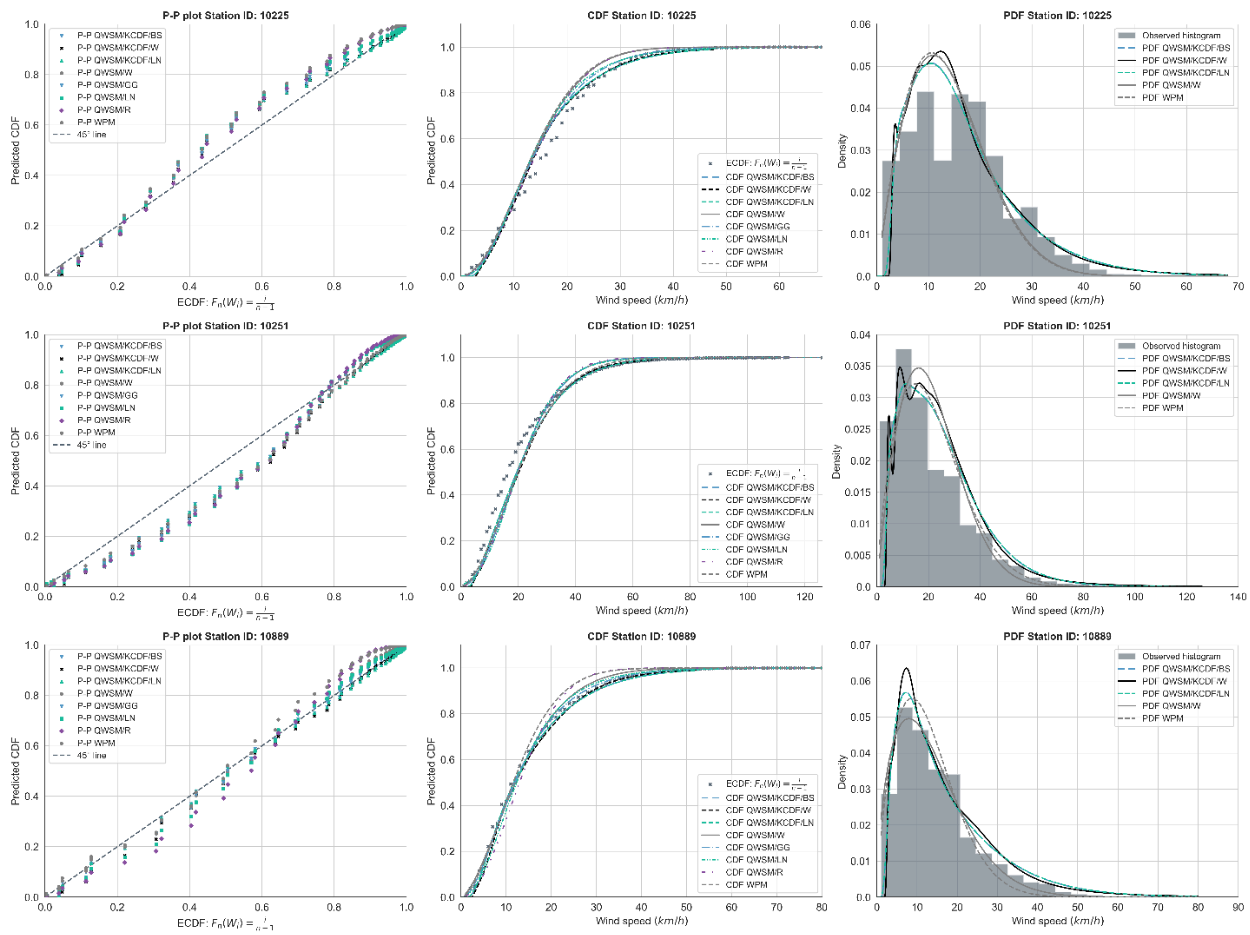


Figure 4.7 PP plot, CDF plot and PDF plot of estimated wind speed probability distributions

4.5 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 (Wu *et al.*, 2019), 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.* (2020) 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.* (2016) 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.* (2016), 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 (Koenker, 2017; Meinshausen *et al.*, 2006; Nasri *et al.*, 2017; Ouali *et al.*, 2016) 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 (Liu *et al.*, 2011; Meinshausen *et al.*, 2006), 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 (Cannon, 2018) and avoid the issue of quantile crossing that arises when estimating the quantiles independently. It is worth mentioning that a gradient-boosting model (Duan *et al.*, 2020) 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 80 m and 100 m, while wind speed data are conventionally collected at 10 m at meteorological stations. As a result, a technique for extrapolating wind speed data to hub height becomes necessary (e.g., 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 *et al.* (2018a) proposed a technique for mapping wind shear distribution, allowing the wind speed distribution to be mapped at any standard hub height. Jung *et al.* (2018a) 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 (Tang *et al.*, 2019). 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: Beaucage *et al.*, 2014; Keck *et al.*, 2020; Simões *et al.*, 2016). 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.

4.6 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 (e.g., DEM and land use map) are often freely accessible from global datasets covering most regions of the world.

Nomenclature

Abbreviations

BS	Birnbaum-Saunders
CANGRD	Canadian gridded temperature and precipitation anomalies dataset
CDF	Cumulative probability function
CFD	Computational Fluid Dynamics
CO2	Carbon dioxide
DEM	Digital elevation model
ECDF	Empirical cumulative probability function
FS	Feature selection
GBT	Gradient boosting trees
GG	Generalized Gamma
GOF	Goodness of fit
GW	Giga watt
KCDF	Kernel estimator of cumulative distribution function
LN	Log-Normal
LR	Linear regression
LSE	Least Square Estimation
MAE	Mean Absolute Error
MI	Mutual information
MISE	Mean Integrated Square Error
MRMR	Minimum redundancy maximum relevance
MRMR-MI	Minimum redundancy maximum relevance with Mutual information
MRMR-PC	Minimum redundancy maximum relevance with Pearson correlation coefficient
MSE	Mean squared error
NWP	Numerical Weather Prediction
PC	Pearson correlation coefficient
PDF	Probability distribution function
PP plot	Percentage probability plot
QWSM	Quantile-based wind speed probability distribution mapping
R	Rayleigh
R^2	Coefficient of determination
RD	Regional distribution

RMSE	root mean square error
SSE	Sum of the square error
UK	United Kingdom
W	Weibull distribution
WPM	Weibull parameters mapping
WS	Wind speed
WSQ	Wind speed quantiles
XGB	Extreme Gradient Boosting

Symbols

$\alpha, \beta, \mu, \sigma$	Parameters of asymmetric kernel function
b	Kernel function bandwidth
b_{opt}^{BS}	Asymptotical Birnbaum-Saunders Kernel optimal bandwidth
b_{opt}^{LN}	Asymptotical Log-Normal Kernel optimal bandwidth
b_{opt}^{WB}	Asymptotical Weibull Kernel optimal bandwidth
D	Kolmogorov–Smirnov statistic
$F(x, y)$	F-statistic between x and y
$\hat{F}(\cdot)$	Cumulative distribution function of estimated wind speed quantiles
$F(\cdot;\ \hat{\theta})$	Fitted cumulative distribution function, where $\hat{\theta}$ are the estimated parameters
$\hat{F}^{BS}(\cdot)$	Birnbaum-Saunders Kernel estimator of cumulative distribution function
$\hat{F}^{LN}(\cdot)$	Log-Normal Kernel estimator of cumulative distribution function
$F_n(\cdot)$	Empirical cumulative distribution function
$\hat{F}^{WB}(\cdot)$	Weibull Kernel estimator of cumulative distribution function
γ	Weight term in the empirical quantile estimation formula
$\Gamma(\cdot)$	Gamma function
$I(x, y)$	Mutual information between x and y
$\bar{K}_{w,b}(\cdot)$	Cumulative distribution function of an asymmetric kernel function
$\bar{K}(\cdot)$	Cumulative distribution function of an asymmetric kernel function
$\bar{K}_{BS}$	Cumulative distribution function of the Birnbaum-Saunders asymmetric kernel function
$\bar{K}_{LN}$	Cumulative distribution function of the Log-Normal asymmetric kernel function
$\bar{K}_{WB}$	Cumulative distribution function of the Weibull asymmetric kernel function

P	Percentile point
P1	5% percentile-level
P10	72.5% percentile-level
P11	80% percentile-level
P12	87.5% percentile-level
P13	95% percentile-level
P2	12.5% percentile-level
P3	20% percentile-level
P4	27.5% percentile-level
P5	35% percentile-level
P6	42.5% percentile-level
P7	50% percentile-level
P8	57.5% percentile-level
P9	65% percentile-level
$\Phi(\cdot)$	Standard normal distribution
$Rel(x, y)$, $Red(x, y)$	Measure of dependency between x and y
$\rho(x, y)$	Pearson correlation coefficient between x and y

5 PREDICTION OF HOURLY WIND SPEED TIME SERIES AT UNSAMPLED LOCATIONS USING MACHINE LEARNING

Titre de l'article : Prédiction des séries temporelles horaires des vitesses du vent aux sites non échantillonnés à l'aide de l'apprentissage automatique

Auteurs :

Freddy Houndekindo[1] (F.H.), Taha B.M.J. Ouarda[1] (T.B.M.J.O)

[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

Titre de la revue :

Article publié le 15 juillet 2024, dans la revue **Energy**, volume 299, DOI : 10.1016/j.energy.2024.131518

Contribution des auteurs :

Conceptualisation, F.H., T.B.M.J.O. ; Méthodologie, F.H., T.B.M.J.O. ; Préparation des données, F.H. ; Logiciel, F.H. ; Rédaction de la version originale, F.H. ; Révision, T.B.M.J.O. ; Supervision, T.B.M.J.O. ; Acquisition du financement, T.B.M.J.O.

Dans l'article 3, nous avons proposé une méthode non paramétrique d'estimation de la distribution du vent. Dans l'article 4, cette approche est étendue afin de reconstruire des séries temporelles de vitesse du vent aux sites non échantillonnés. La disponibilité de ces séries temporelles permet d'évaluer la variabilité temporelle des vitesses de vents à différentes échelles temporelles.

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

5.1 Introduction

The past decades have witnessed a significant uptake of wind energy in various parts of the world (Cherp *et al.*, 2021). This growth reflects a global shift toward more renewable energy sources, with wind power playing a prominent role in energy supply (Wiser *et al.*, 2021). The intermittent nature of wind speed still poses some challenges to the development of the renewable energy source (Ren *et al.*, 2017). 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 (Lee *et al.*, 2021).

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 (Lopez *et al.*, 2021; Niermann *et al.*, 2019). 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: Hersbach *et al.*, 2020), and the NASA's Modern Era Retrospective Analysis for Research and Applications-2 (MERRA-2: Molod *et al.*, 2015) have been used extensively to conduct WRA across large regions (Gruber *et al.*, 2022; Gualtieri, 2022). Samal (2021) 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 (Ayik *et al.*, 2021). 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 (Ramon *et al.*, 2019). 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 (Gualtieri, 2021). A review of the uncertainties associated with the application of reanalysis data for WRA was presented by Gualtieri (2022). 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 (Gruber *et al.*, 2019). 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.* (2022). 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.* (2008a) 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 (González-Longatt *et al.*, 2015) 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.* (2016) mapped the parameters of the Weibull distribution fitted to WS data across the United Kingdom (UK). Jung (2016) 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 *et al.* (2020) developed a global model that estimates the parameters of the Kappa and Wakeby distribution for WS variability assessment using estimated L-moments. Houndekindo *et al.* (2023b) 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 (e.g., 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.* (2013) and Requena *et al.* (2017) for application to streamflow data and Ouarda *et al.* (2022) 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 (Ben Alaya *et al.*, 2016; Cannon *et al.*, 2015). 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 5.2 describes the study area and the datasets. The methodology employed is presented in section 5.3. The results of the comprehensive evaluation of the different approaches are presented in section 5.4. The discussion follows in section 5.5, and section 5.6 gives the conclusions of the study.

5.2 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 5.1 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 9.1 of the supporting material (Section 9). Topographical covariates were calculated from the Advanced Land Observing Satellite (ALOS) Digital Elevation model (DEM) of 30 m resolution (ALOS DEM: Tadono *et al.*, 2014) obtained freely from the Japan Aerospace Exploration Agency. The surface roughness length was estimated from a 2015 land use map of Canada (Latifovic *et al.*, 2017) obtained from Natural Resource Canada.

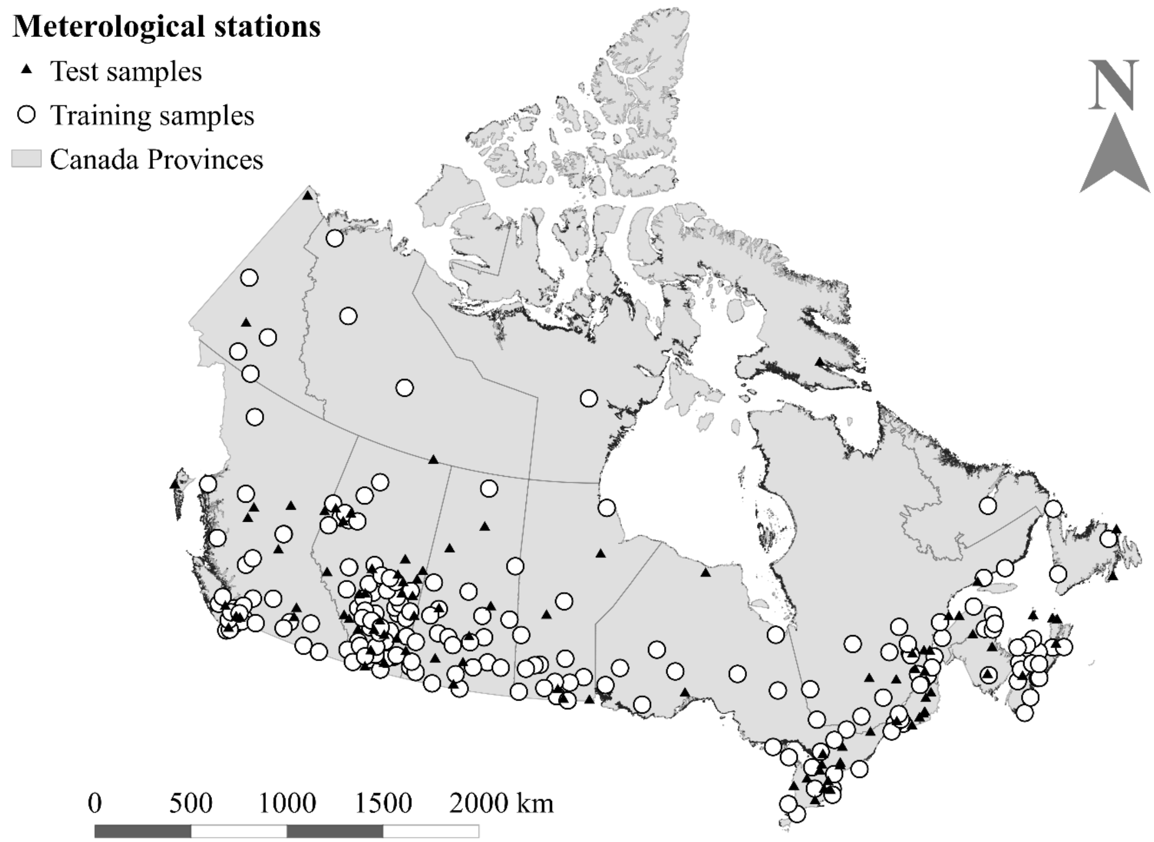


Figure 5.1 Study area and location of the 303 meteorological stations used in the study

5.3 Methods

5.3.1 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 *et al.* (2023b) 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 *et al.* (2023b) 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 (He *et al.*, 2018) and for the estimation of other hydro-climatic variables at unsampled locations (Ouali *et al.*, 2016), to the author's knowledge, it is the first time they are applied to estimate conditional WSQ at unsampled locations. As done by Houndekindo *et al.* (2023b), 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: Ke *et al.*, 2017) with the pinball loss function (Equation 5.1) 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 (Fan *et al.*, 2019; Genov *et al.*, 2024; Park *et al.*, 2020). 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 (e.g., N > 10,000) while maintaining good prediction accuracy. In addition, the LGBM adopts a leaf-wise tree expansion with a fixed maximum depth, improving the model's training performance. Table 5.1 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 (Feurer *et al.*, 2019). 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.

Equation 5.1

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

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

Equation 5.2

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

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

In addition to the covariates presented in Table 9.1 of the supporting material (Section 9), hourly WSQ extracted from the ERA5 dataset (ERA5-WSQ) were assessed as covariates in the current study. As stated by Jung *et al.* (2020), 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 LGBMQR-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: Ding *et al.*, 2005). 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 (Houndekindo *et al.*, 2023a).

The estimated conditional WSQs were used as input for the Birnbaum-Saunders asymmetric kernel estimator of CDF (Mombeni *et al.*, 2021) 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 *et al.* (2023b).

Table 5.1 Parameters of LGBMQR and LGBMQR-ERA5

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 (Geurts *et al.*, 2006)	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

The same set of randomly selected parameters was tested for LGBMQR and LGBMQR-ERA5 to implement the random search

5.3.2 Prediction of wind speed time series at unsampled locations

It is proposed to adapt the QM (Wood *et al.*, 2002) procedure to predict WSTS at unsampled locations using the following general formula (Cannon *et al.*, 2015):

Equation 5.3

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

where: $\widehat{w}_t(s_0)$ is the estimated WS at time t and unsampled location s_0. $\widehat{F}_{S_0}$ is the estimated WS CDF at the unsampled location s_0, and $\widehat{F}_{S_0}^{-1}$ is its inverse. $\widehat{F}(w_t)$ is the estimated wind speed non-exceedance probabilities (WSNEP) at time t.

The methodology to estimate $\widehat{F}_{S_0}$ at any unsampled location in the region was described in section 5.3.1. For the estimation of $\widehat{F}(w_t)$ two approaches have been put forward in previous studies:

1. Some authors (Ouarda *et al.*, 2022; Shu *et al.*, 2012) proposed using information from nearby locations to estimate $\widehat{F}(w_t)$ 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 *et al.* (2023b) derived the non-exceedance probabilities $\widehat{F}(w_t)$ 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:

Equation 5.4

$$F_n(w_t) = \frac{i_t}{n+1}$$

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.

5.3.3 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:

Equation 5.5

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

Equation 5.6

$$\lambda_i = \frac{d_i^{-p}}{\sum_{j=1}^{k} d_j^{-p}}$$

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 exponent 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 9.4 of the supporting material (Section 9). 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: Sekulić *et al.*, 2020). 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 (Sekulić *et al.*, 2020):

Equation 5.7

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

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.* (2020) 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 (Hengl *et al.*, 2018). 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 5.1 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 5.7 (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 5.8:

Equation 5.8

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

5.3.4 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 250 m × 250 m and accounts for the effect of the local topography and surface roughness. Several studies used the GWA to bias-correct reanalysis WS data (de Aquino Ferreira *et al.*, 2022; Luzia *et al.*, 2023; Murcia *et al.*, 2022). 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.

5.3.5 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 Equation 9.1, Equation 9.2, and Equation 9.3 of the supporting material (Section 9), 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 9.8 in the supporting material (Section 9). The R^2, MAE and RMSE were used to compare the observed and estimated WSQ. Lastly, the Overlap percentage (OP: Perkins et al., 2007) was used to assess the overlap between estimated and observed empirical probability distribution function (PDF). The OP is presented in Equation 9.6 of the supporting material (Section 9). For a review of criteria used for the selection of PD for WS data the reader is referred to (Ouarda et al., 2016).
3. Interannual variability (IAV) evaluation: The robust coefficient of variation (RCov: Watson, 2019) 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 Equation 9.7 and Equation 9.5 of the supporting material (Section 9), respectively.

5.4 Results

5.4.1 Quantile regression models

A thousand random combinations of the LGBM hyperparameters (Table 5.1) were tested with LGBMQR and LGBMQR-ERA5 models. Table 9.2 in the supporting material shows the best parameters found using a random search, including the number of selected covariates. Figure 5.2 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 5.2. Figure 5.3 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 9.3 of the supporting material (Section 9)). 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.

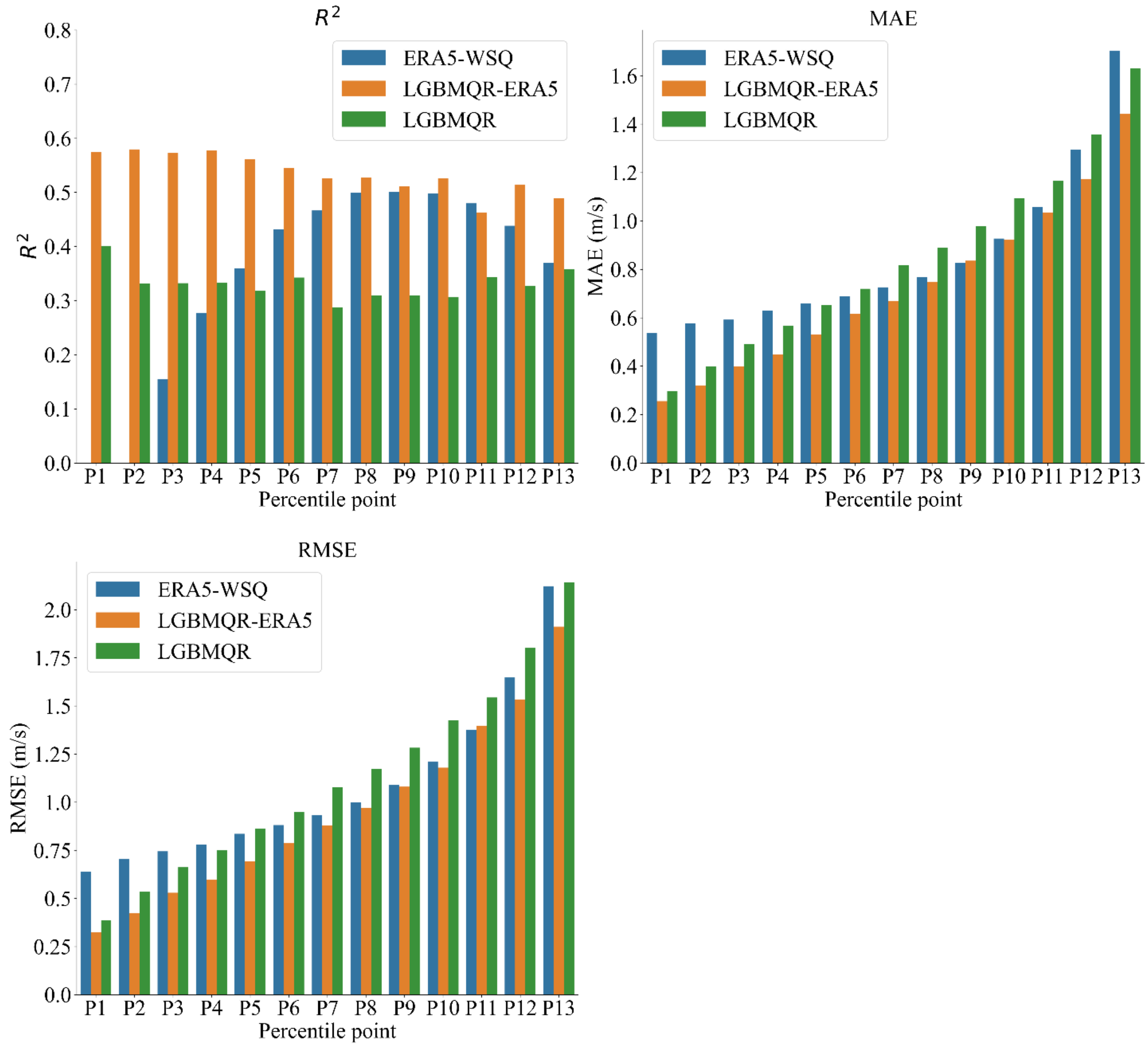


Figure 5.2 Results of the R2, RMSE and MAE between estimated and predicted WSQ at various percentile points (P1-P13)

The metrics were calculated across the test samples for each percentile point

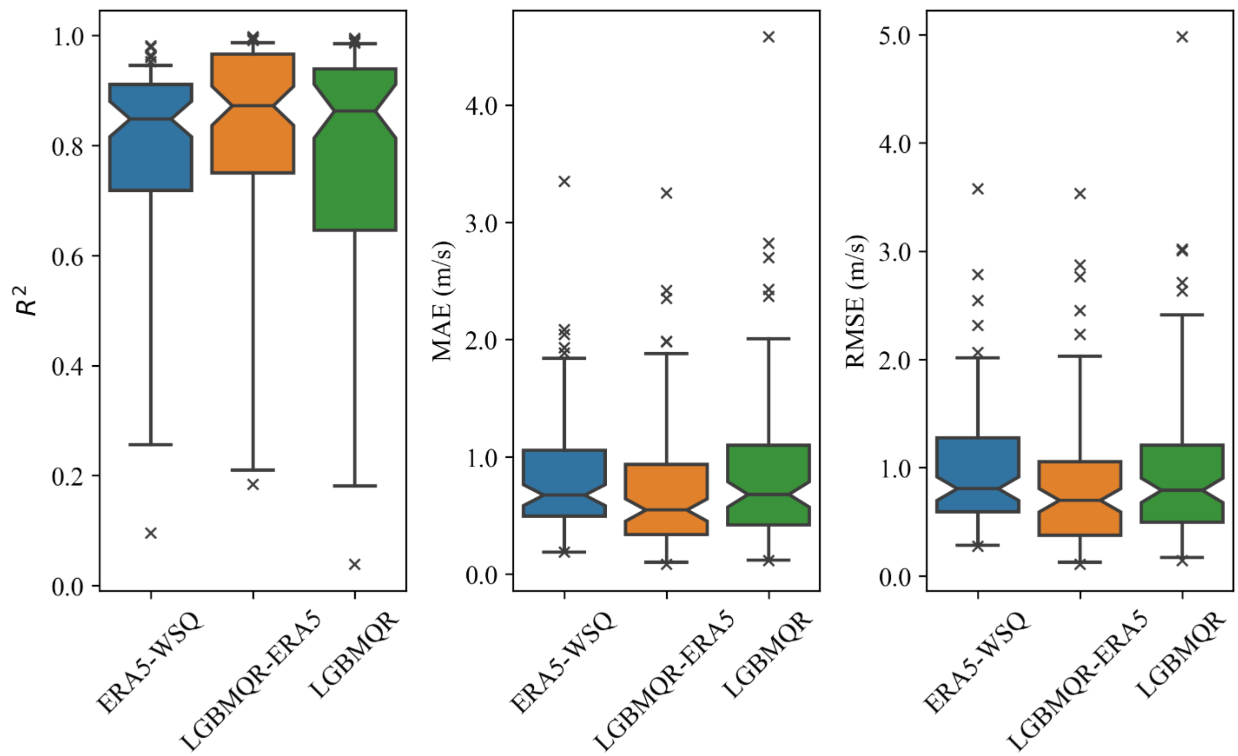


Figure 5.3 Result of the R^2, MAE and RMSE between observed and estimated WSQ
The metrics were calculated across the percentile points (P1-P13) at each location in the test samples

5.4.2 Inverse distance weighting parameters

Table 5.2 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 5.2 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

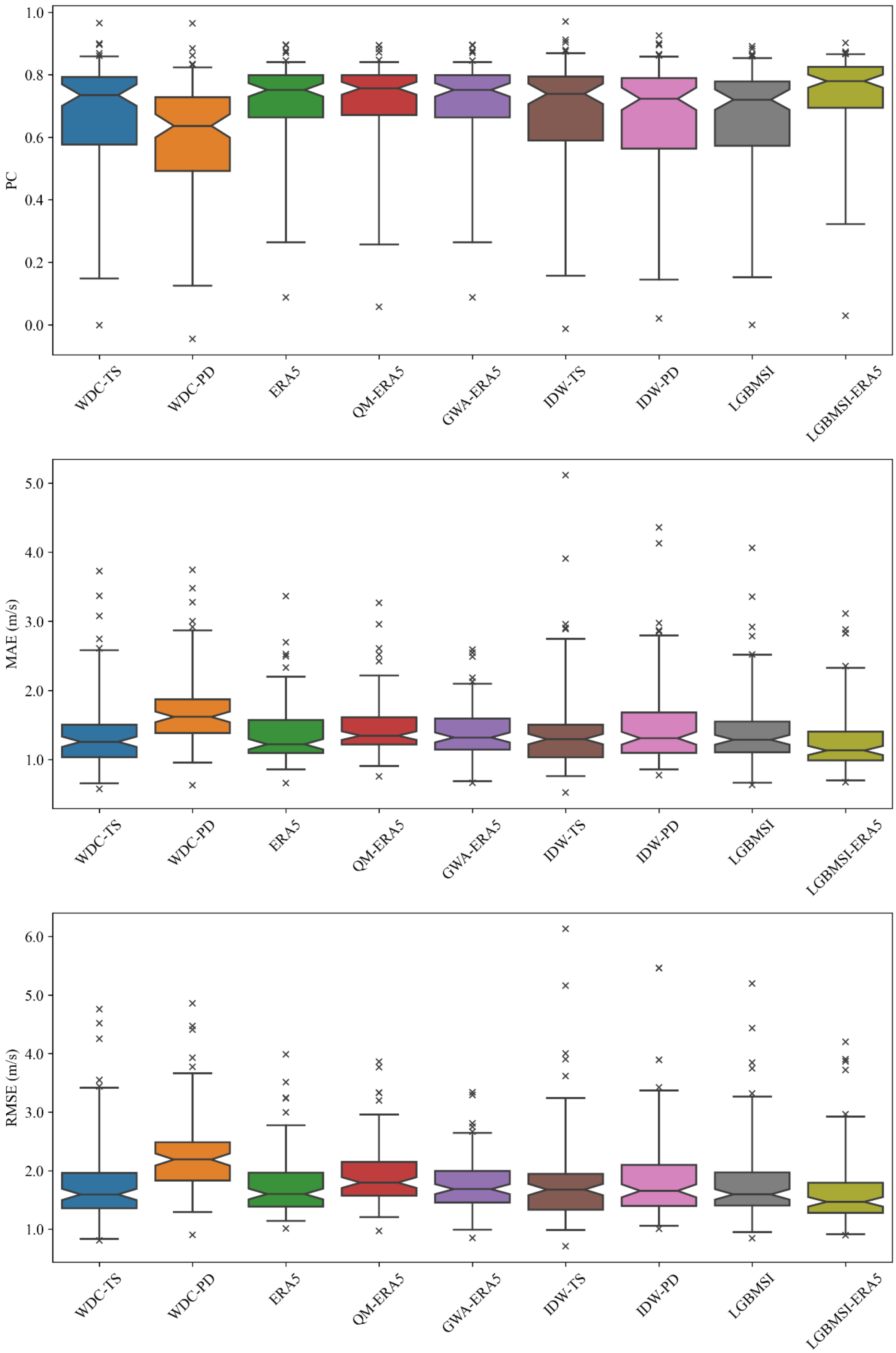
5.4.3 Time series evaluation

Figure 5.4 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 5.3. 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 9.4 in the supporting material (Section 9) 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 5.3 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



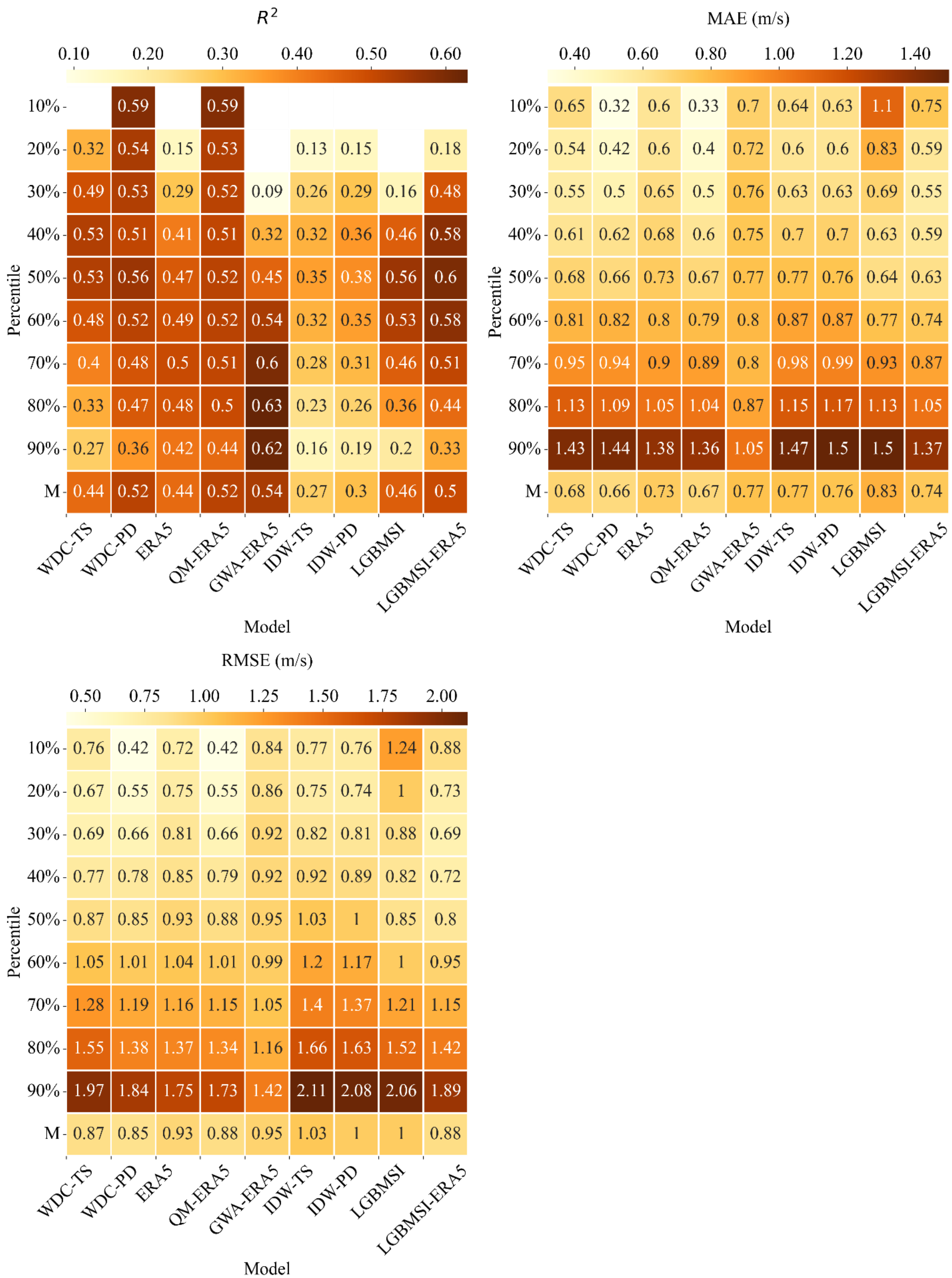
PC
MAE (m/s)
RMSE (m/s)
WDC-TS
WDC-PD
ERA5
QM-ERA5
GWA-ERA5
IDW-TS
IDW-PD
LGBMSI
LGBMSI-ERA5

Figure 5.4 Result of the PC, MAE and RMSE between observed and estimated WSTS

5.4.4 Probability distribution evaluation

Figure 5.5 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 5.6 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 9.5 of the supporting material (Section 9)). The P-values associated with QM-ERA5, LGBMSI, LGBMSI-ERA5 and WDC-TS were always small (e.g., 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
0.10 0.20 0.30 0.40 0.50 0.60
Percentile
10% 0.59 0.59
20% 0.32 0.54 0.15 0.53 0.13 0.15 0.18
30% 0.49 0.53 0.29 0.52 0.09 0.26 0.29 0.16 0.48
40% 0.53 0.51 0.41 0.51 0.32 0.32 0.36 0.46 0.58
50% 0.53 0.56 0.47 0.52 0.45 0.35 0.38 0.56 0.6
60% 0.48 0.52 0.49 0.52 0.54 0.32 0.35 0.53 0.58
70% 0.4 0.48 0.5 0.51 0.6 0.28 0.31 0.46 0.51
80% 0.33 0.47 0.48 0.5 0.63 0.23 0.26 0.36 0.44
90% 0.27 0.36 0.42 0.44 0.62 0.16 0.19 0.2 0.33
M 0.44 0.52 0.44 0.52 0.54 0.27 0.3 0.46 0.5
WDC-TS WDC-PD ERA5 QM-ERA5 GWA-ERA5 IDW-TS IDW-PD LGBMSI LGBMSI-ERA5
Model
MAE (m/s)
0.40 0.60 0.80 1.00 1.20 1.40
Percentile
10% 0.65 0.32 0.6 0.33 0.7 0.64 0.63 1.1 0.75
20% 0.54 0.42 0.6 0.4 0.72 0.6 0.6 0.83 0.59
30% 0.55 0.5 0.65 0.5 0.76 0.63 0.63 0.69 0.55
40% 0.61 0.62 0.68 0.6 0.75 0.7 0.7 0.63 0.59
50% 0.68 0.66 0.73 0.67 0.77 0.77 0.76 0.64 0.63
60% 0.81 0.82 0.8 0.79 0.8 0.87 0.87 0.77 0.74
70% 0.95 0.94 0.9 0.89 0.8 0.98 0.99 0.93 0.87
80% 1.13 1.09 1.05 1.04 0.87 1.15 1.17 1.13 1.05
90% 1.43 1.44 1.38 1.36 1.05 1.47 1.5 1.5 1.37
M 0.68 0.66 0.73 0.67 0.77 0.77 0.76 0.83 0.74
WDC-TS WDC-PD ERA5 QM-ERA5 GWA-ERA5 IDW-TS IDW-PD LGBMSI LGBMSI-ERA5
Model
RMSE (m/s)
0.50 0.75 1.00 1.25 1.50 1.75 2.00
Percentile
10% 0.76 0.42 0.72 0.42 0.84 0.77 0.76 1.24 0.88
20% 0.67 0.55 0.75 0.55 0.86 0.75 0.74 1 0.73
30% 0.69 0.66 0.81 0.66 0.92 0.82 0.81 0.88 0.69
40% 0.77 0.78 0.85 0.79 0.92 0.92 0.89 0.82 0.72
50% 0.87 0.85 0.93 0.88 0.95 1.03 1 0.85 0.8
60% 1.05 1.01 1.04 1.01 0.99 1.2 1.17 1 0.95
70% 1.28 1.19 1.16 1.15 1.05 1.4 1.37 1.21 1.15
80% 1.55 1.38 1.37 1.34 1.16 1.66 1.63 1.52 1.42
90% 1.97 1.84 1.75 1.73 1.42 2.11 2.08 2.06 1.89
M 0.87 0.85 0.93 0.88 0.95 1.03 1 1 0.88
WDC-TS WDC-PD ERA5 QM-ERA5 GWA-ERA5 IDW-TS IDW-PD LGBMSI LGBMSI-ERA5
Model

Figure 5.5 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.

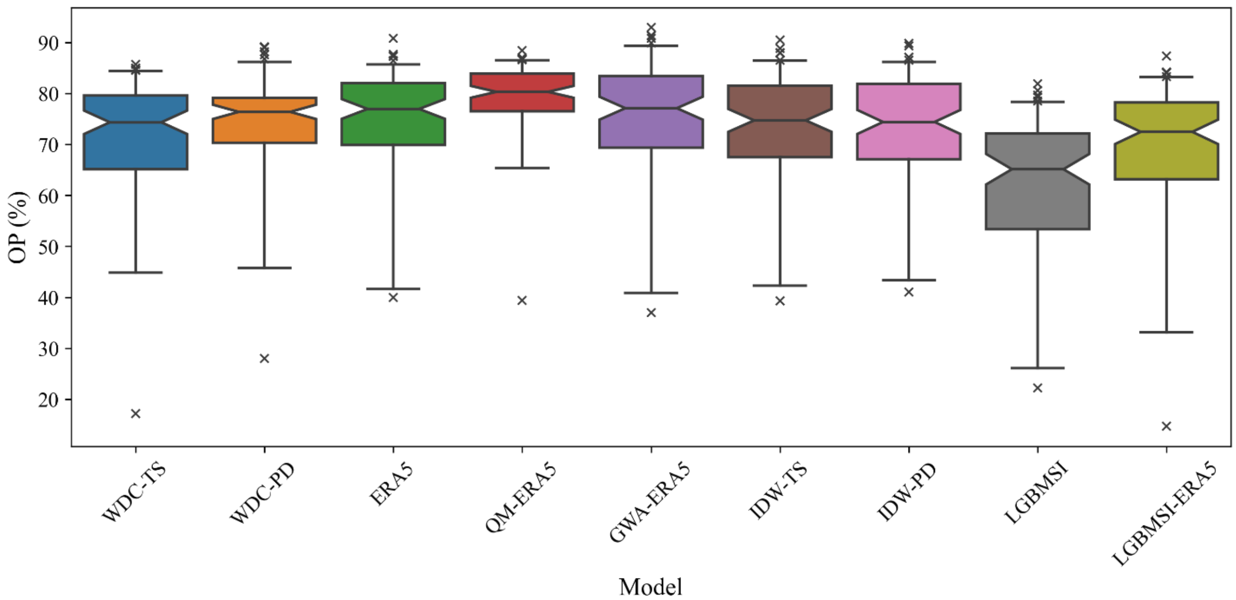


Figure 5.6 Boxplots of OP metrics calculated between observed and estimated empirical PDF.

5.4.5 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 (Bett *et al.*, 2017; Jung *et al.*, 2019c; Ouarda *et al.*, 2021). The IAV has been linked to atmospheric teleconnections (Naizghi *et al.*, 2017; Ouarda *et al.*, 2021; Zhou *et al.*, 2022) 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 (Pryor *et al.*, 2018). 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 (Woldesellasse *et al.*, 2020).

Figure 5.7 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.

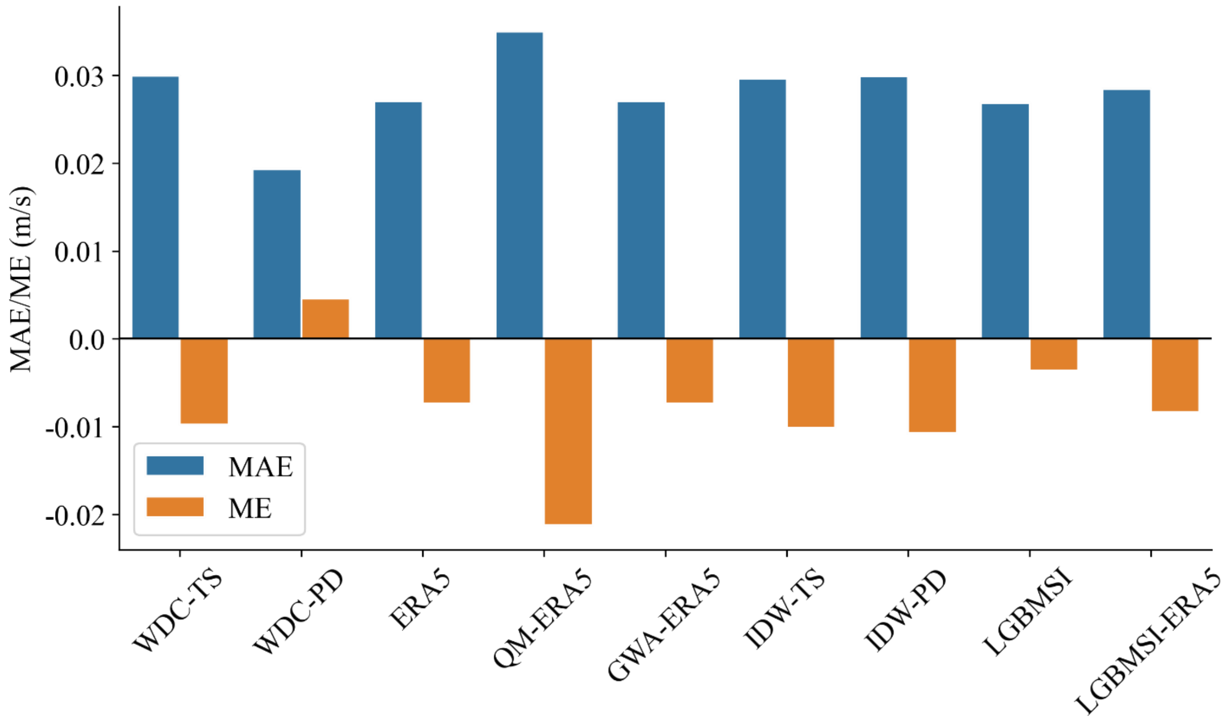


Figure 5.7 **Result of MAE and ME between observed and estimated RCov of median annual WS.**

5.5 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 (Meinshausen *et al.*, 2006) and quantile regression neural networks (Cannon, 2011). LGBMQR was adopted because it is efficient during training, and in general, gradient-boosting models have demonstrated superior performance on tabular data (Grinsztajn *et al.*, 2022a). 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 (Gruber *et al.*, 2019; Jung *et al.*, 2020). 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.

5.6 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 *et al.* (2021) 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.

Nomenclature

Abbreviations

a.g.l	Above ground level
ALOS	Advanced Land Observing Satellite
CDF	Cumulative distribution function
DEM	Digital elevation model
ECCC	Environment and Climate Change Canada
ECMWF	European Centre for Medium-Range Weather Forecasts
ERA5	European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis v5
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
LGBMQR-ERA5	LGBMQR using ERA5 wind speed quantiles as covariates
LGBMSI	LGBM for spatial interpolation
LGBMSI-ERA5	LGBMSI using the ERA5 wind data as covariates
MAE	Mean absolute error (m/s)
ME	Mean error (m/s)
MERRA-2	Modern Era Retrospective Analysis for Research and Applications-2
MRMR	Minimum redundancy maximum relevancy algorithm
OP	Overlap percentage (%)
PC	Pearson correlation
PD	Probability distribution
PDF	Probability distribution function
QM	Quantile mapping
QM-ERA5	Quantile mapping bias correction of ERA5 wind data
QR	Quantile regression
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
WMO	World Meteorological Organization
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

Symbols

d_i	Distance of location s_i from the target location s_0
$f(\cdot)$	Regression function
$F(\cdot \mid X = x)$	Conditional cumulative distribution function
$F_n(\cdot)$	Empirical cumulative distribution function
$\hat{F}_{S_0}(\cdot)$	Estimated cumulative distribution function at location S_0
$\hat{F}_{S_0}{}^{-1}(\cdot)$	Inverse of the estimated cumulative distribution function at location S_0
k	Number of nearest neighbours to consider in the inverse distance weighting method
λ_i	Interpolation weight associated with location s_i
p	Exponent in the inverse distance weighting formula
P1	5% percentile-level
P10	72.5% percentile-level
P11	80% percentile-level
P12	87.5% percentile-level
P13	95% percentile-level
P2	12.5% percentile-level
P3	20% percentile-level
P4	27.5% percentile-level
P5	35% percentile-level
P6	42.5% percentile-level
P7	50% percentile-level
P8	57.5% percentile-level

P9	65% percentile-level
$\rho_{\tau}(\cdot)$	Quantile loss function
τ	Quantile-level
$w_t\left(ERA5_{s_0}\right)$	ERA5 wind speed value at time t interpolated at location s_0
$\widehat{w}_t(s_0)$	Estimated wind speed value at time t and unsampled location s_0
$w_t(s_i)$	Observed wind speed value at time t and location s_i

6 LSTM AND TRANSFORMER-BASED FRAMEWORK FOR BIAS CORRECTION OF ERA5 HOURLY WIND SPEEDS

Titre de l'article : Application des LSTM et Transformer pour la correction des biais des vitesses de vent horaires d'ERA5

Auteurs :

Freddy Houndekindo[1] (F.H.), Taha B.M.J. Ouarda[1] (T.B.M.J.O)

[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

Titre de la revue :

Article publié le 1 août 2025, dans la revue **Energy**, volume 328, DOI : 10.1016/j.energy.2025.136498

Contribution des auteurs :

Conceptualisation, F.H., T.B.M.J.O. ; Méthodologie, F.H., T.B.M.J.O. ; Préparation des données, F.H. ; Logiciel, F.H. ; Rédaction de la version originale, F.H. ; Révision, T.B.M.J.O. ; Supervision, T.B.M.J.O. ; Acquisition du financement, T.B.M.J.O.

Cet article s'inscrit dans la continuité des travaux effectués dans les articles précédents, et vise à affiner les méthodes d'estimation des vitesses du vent aux sites non échantillonnés. Il met particulièrement l'accent sur la variabilité temporelle des estimations, un aspect crucial dans la planification des ressources éoliennes.

Abstract

Reanalysis-derived wind speeds are central to large-scale wind resource assessment (WRA). However, their coarse spatial resolution often introduces significant biases, particularly in complex terrains and coastal areas. A deep learning (DL) framework using LSTMs or Transformers was introduced to correct systematic biases and temporal variability in reanalysis-derived wind speeds by modeling a time-resolved scaling factor, which is then used to adjust ERA5 wind speeds. The proposed framework's spatiotemporal generalization capability was rigorously evaluated using a test set of 170 independent stations distributed throughout Canada in diverse environmental conditions. Results showed that the DL framework outperformed a standard bias correction method based on the Global Wind Atlas. It improved the median wind speed, the temporal variability, and the probability distributions of ERA5 wind speeds in coastal areas and complex terrains. Specifically, in coastal regions, the DL models increased the explained variability of median wind speed by over 70% relative to ERA5. In regions characterized by high surface roughness length, such as forests and urban areas, these models achieved average improvements of more than 10% in MAE and RMSE of the time series. While the DL models performed well in representing the probability distribution of the most typical wind speed values, some challenges remain in improving the distribution of extreme wind speeds. Overall, this framework represents a promising advancement in enhancing the accuracy of reanalysis-derived wind speeds in large-scale WRA. By reducing biases in ERA5 wind data, the DL framework supports more reliable site selection and estimation of long-term energy production.

Keywords: Bias correction, Deep learning, ERA5, High-resolution, Wind resource assessment.

6.1 Introduction

Addressing climate change and mitigating its risks to humanity requires significantly expanding renewable energy sources, such as wind and solar, to replace fossil fuel-based energy systems (Shang *et al.*, 2024). Wind energy, in particular, has emerged as a promising and scalable solution driven by technological progress (Kumar *et al.*, 2016b; McKenna *et al.*, 2016) and rapidly declining costs (Wiser *et al.*, 2021). Improving the accuracy of wind resource assessment (WRA) is crucial to increasing wind energy's contribution to the global energy mix (McKenna *et al.*, 2022; Pelser *et al.*, 2024). Preliminary assessment of wind energy resources typically involves evaluating their meteorological potential (Manwell *et al.*, 2010). In addition to considering the meteorological potential, comprehensive WRA studies also integrate economic, environmental, land-use, and technological constraints (Pelser *et al.*, 2024).

Accurate wind data is essential for WRA studies (Pelser *et al.*, 2024). The data required for large-scale meteorological wind potential assessments can be classified as time-invariant (static) or time-resolved (McKenna *et al.*, 2022). Time-invariant data provides an overview of long-term average wind conditions across a large region, typically with high spatial resolution (e.g., wind atlas). In contrast, time-resolved data (e.g., meteorological reanalysis) captures the temporal variability of wind resources, including diurnal, seasonal, and interannual variations.

The high temporal resolution of time-resolved data comes at the expense of relatively low spatial resolution, limiting its effectiveness for analysis that requires fine spatial detail (Gualtieri, 2022). Some studies have sought to integrate time-invariant and time-resolved wind data, enhancing spatial and temporal resolution (McKenna *et al.*, 2022). A commonly used method within this framework is a simple mean bias correction (BC), where the long-term mean wind speed from a reanalysis dataset is corrected to match the mean wind speed from a wind atlas.

The Global Wind Atlas (GWA) provides time-invariant wind data quasi-globally (excluding Antarctica) with a spatial resolution of 250 m × 250 m (Davis *et al.*, 2023). The dataset's global availability, relatively high spatial resolution, and free accessibility contribute to its widespread adoption for BC of reanalysis-derived wind speeds (Gualtieri, 2022). The dataset was generated by applying numerical weather prediction (NWP) models to downscale large-scale wind speeds from reanalysis datasets, followed by microscale modeling to account for high-resolution topography and land use.

Numerical models like NWP are often associated with high computational costs (Ben Bouallègue *et al.*, 2024; Schultz *et al.*, 2021). In contrast, machine learning (ML) and statistical models have emerged as effective alternatives to numerical models in some fields (Höhlein *et al.*, 2020; Kratzert *et al.*, 2018; Reichstein *et al.*, 2019) and typically require fewer computation resources. Given sufficient sample sizes, ML models can learn complex relationships between input variables and the target variable. For instance, in the context of WRA, ML has been applied to model mean wind speeds (Cellura *et al.*, 2008a), wind speed quantiles (Houndekindo *et al.*, 2023b), wind speed L-moments (Jung *et al.*, 2020), and parameters of wind speed distribution (Veronesi *et al.*, 2016).

To the authors' knowledge, no previous attempts have been made to use ML to directly predict a scaling factor that could adjust the long-term mean wind speed from reanalysis data. An ML-derived scaling factor could provide an alternative to the scaling factor typically estimated from the GWA, allowing for greater flexibility in the resulting spatial resolution of the downscaled wind data. Biases in reanalysis-derived wind speed partially stem from their coarse spatial resolutions, which limit their ability to resolve detailed terrain features and surface roughness (Petersen *et al.*, 1998). ML can effectively reduce these biases by establishing a relationship between high-resolution topography, land use, and the corresponding scaling factor. This approach leverages the capacity of ML algorithms to learn complex, nonlinear relationships between input variables and outputs.

Due to the static nature of data from wind atlases, time-invariant scaling factors have been used primarily for BC of reanalysis-derived wind speeds. While a time-invariant scaling factor can improve reanalysis data's long-term mean wind speed, it does not modify its temporal variability (Bosch *et al.*, 2018). Discrepancies have been reported between the temporal variations in reanalysis and measured wind speeds (Brune *et al.*, 2021; Davidson *et al.*, 2022; Jourdier, 2020; McKenna *et al.*, 2022; Ramon *et al.*, 2019). These discrepancies hinder understanding the data temporal variability, such as diurnal, seasonal, and inter- and intra-annual variations, essential for effective WRA. Developing a model that dynamically adjusts the scaling factor could address these issues. By leveraging deep learning (DL) architectures for sequence modeling, such as Long Short-Term Memory (LSTM; Hochreiter *et al.*, 1997) networks or Transformers (Vaswani *et al.*, 2017), time-resolved scaling factors can be modeled, enabling both the adjustment of the long-term mean wind speed and the improvement of the temporal correlations between reanalysis-derived and measured wind speeds.

This paper introduces an innovative ML framework for predicting time-invariant and time-resolved scaling factors for BC of reanalysis-derived wind speeds. The proposed framework uses covariates related to the topography, the surface roughness length (SRL), and reanalysis-derived past weather conditions to predict a scaling factor. The time-invariant scaling factor is modeled using a standard ML model, such as gradient boosting trees (GB). DL models suitable for sequence modeling, such as LSTM or Transformers, were used to model the time-resolved scaling factor.

The proposed framework was compared to a commonly used BC method based on the GWA. The GWA correction method relies on a time-invariant scaling factor that can only adjust the systematic bias of reanalysis-derived wind speeds. However, as mentioned above, the temporal variability may not be well represented in the reanalysis data. The time-resolved scaling factor proposed in this study can correct systematic bias in reanalysis-derived wind speeds while adjusting the temporal variability to better match the measured wind speeds. This novel approach significantly departs from conventional methods by introducing a time-resolved, data-driven framework capable of capturing complex wind-topography interaction and complex temporal dependencies to improve the accuracy of reanalysis-derived wind speeds.

To ensure a comprehensive evaluation, several criteria relevant to WRA, including median wind speed (to represent the typical wind speed condition), temporal variability, and probability distribution, were considered. Furthermore, the framework spatiotemporal generalization capabilities were assessed across 170 independent test stations located in diverse environmental conditions that impact reanalysis accuracy, such as coastal areas, complex terrain (hilly and mountainous), and regions with varying SRL. This comprehensive analysis is essential to demonstrate the framework's applicability in various environments and identify potential areas for refinement.

In the following sections, a survey of related work is presented (Section 6.2), followed by a description of the dataset (section 6.3), the methodology (Section 6.4), and the presentation of the results (Section 6.5), which are discussed in Section 6.6. Finally, the conclusion of the paper is provided in Section 6.7.

6.2 Related work

Various techniques for BC of reanalysis-derived wind speeds have been explored in the literature. A commonly used method is mean BC, which aligns the long-term mean wind speed from the reanalysis data with the GWA (Bosch *et al.*, 2017; Bosch *et al.*, 2018; de Aquino Ferreira *et al.*, 2022; Gruber *et al.*, 2019; Gruber *et al.*, 2022; Houndekindo *et al.*, 2024; Langer *et al.*, 2023; Luzia *et al.*, 2023; Murcia *et al.*, 2022; Ryberg *et al.*, 2019). This technique applies a time-invariant scaling factor estimated as the ratio between the reanalysis and the atlas mean wind speeds. The time-invariant scaling factor can adjust the reanalysis long-term mean wind speed while preserving its temporal characteristics.

Beyond mean BC, parametric quantile mapping using the Weibull scale and shape parameters from the GWA has been applied (González-Aparicio *et al.*, 2017; Nefabas *et al.*, 2021). This method assumes that the Weibull distribution adequately represents wind speed distributions across the region of interest. However, a growing body of literature advocates considering alternative or mixed distributions in wind speed modeling (Jung *et al.*, 2019b; Ouarda *et al.*, 2018; Ouarda *et al.*, 2015; Tsvetkova *et al.*, 2023).

Methods that rely on the GWA for reanalysis-derived wind speed BC are limited to the atlas' spatial resolution. Some studies have implemented ML for wind speed BC to achieve greater flexibility in the resulting spatial resolution. For instance, Jung *et al.* (2023b) applied a parametric quantile mapping procedure to downscale wind speeds from a regional reanalysis dataset across Germany. To estimate the wind speed distribution at unsampled locations, they developed a gradient-boosting regression model that uses high-resolution terrain and land use data to predict wind speed statistics in the form of L-Moments (Hosking, 1990). The predicted wind speed L-Moments were then used to derive the parameters of the Kappa distribution function at unsampled locations. This method successfully increased the spatial resolution of the reanalysis data from 11 km × 11 km to 25 m × 25 m.

Similarly, Hu *et al.* (2023) employed the eXtreme gradient boosting algorithm (XGBOOST, Chen *et al.* (2016)) to downscale hourly wind speed from the European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis version 5 (ERA5, Hersbach *et al.* (2020)). The model used meteorological variables (wind speed, wind direction, and gravity wave dissipation) from ERA5 and high-resolution topography data to predict observed wind speed. Houndekindo *et al.* (2024) compared various ML methods for the BC of ERA5 wind speeds. The findings demonstrated that quantile mapping and ML-based spatial interpolation, using ERA5 wind speed data, high-resolution topography, and SRL as covariates, improved the probability distribution and temporal variability of ERA5 wind speeds.

Dujardin *et al.* (2022) classified statistical downscaling methods based on input and output types, distinguishing between point-level (Director *et al.*, 2015) and gridded data. This classification resulted in three categories of methods: point-to-point, 2D-to-point (Dujardin *et al.*, 2022), and 2D-to-2D (Dupuy *et al.*, 2023; Höhlein *et al.*, 2020; Zhong *et al.*, 2024). The 2D-to-point and 2D-to-2D methods are designed with DL architectures for computer vision tasks such as convolutional neural networks (CNNs). These architectures use gridded meteorological and topography features as inputs to predict wind data either at a point location (e.g., the center of the grid) or at the grid level, i.e., super-resolution.

Given the conceptual alignment of DL architectures for computer vision with the problem of downscaling wind fields (Baño-Medina *et al.*, 2020; Höhlein *et al.*, 2020), these methods are expected to outperform point-to-point techniques. However, this improvement often requires significantly more computational and memory resources, which can present significant challenges for practical applications, particularly when analyzing large regions and extensive time series data (e.g., 30 years of hourly data for WRA).

A limited number of studies have applied DL architectures for sequence modeling to the problem of reanalysis-derived wind speed BC. However, models like LSTM and Transformers (Vaswani *et al.*, 2017) were often used to model the temporal characteristics of meteorological data. The LSTM was introduced to address the limitations of vanilla recurrent neural networks (RNNs) in capturing long-term dependencies (Goodfellow *et al.*, 2016). Several studies have shown that LSTMs outperform traditional statistical methods for wind speed forecasting (Chen *et al.*, 2021a; Elsaraiti *et al.*, 2021; Gao *et al.*, 2023). Zhang *et al.* (2021) applied the Gated Recurrent Unit (GRU) to downscale wind speed from global climate models (GCMs). GRUs have a similar architecture to LSTMs but with fewer parameters, making them more computationally efficient. Beyond wind speed forecasting, LSTMs were also successfully applied in hydrology for rainfall-runoff modeling (Kratzert *et al.*, 2018), a task traditionally handled by physics-based models.

Despite the improvements introduced by the LSTM architecture in sequence modeling, it still faces limitations due to its sequential processing, which can be time-consuming for long sequences. Researchers have explored several alternatives and enhancements to LSTMs to address this challenge. One prominent alternative is the Transformer architecture, which replaces the sequential nature of LSTMs with a parallelizable self-attention mechanism (Vaswani *et al.*, 2017). It allows Transformers to process entire sequences simultaneously, improving computational efficiency (Dao *et al.*, 2022) and effectively capturing long-range dependencies (Lim *et al.*, 2021b). Recent studies have applied Transformers for wind speed forecasting (Bentsen *et al.*, 2023; Jiang *et al.*, 2024; Nascimento *et al.*, 2023; Wu *et al.*, 2022), wind power forecasting (Sun *et al.*, 2023a), and global solar radiation prediction (Zhou *et al.*, 2023). Nascimento *et al.* (2023) reported that Transformers outperformed LSTMs, offering better generalization in wind speed forecasting. In contrast, Jiang *et al.* (2024) found that LSTMs and GRUs outperformed Transformers in their proposed DL framework for wind speed forecasting.

This study proposes using standard ML to model a time-invariant scaling factor for adjusting the long-term mean wind speed from reanalysis data. This approach offers an alternative to the time-invariant scaling factor typically derived from the GWA. The key advantage of using ML lies in its flexibility, enabling finer spatial resolution in the output compared to the fixed spatial resolution offered by the GWA.

Additionally, a novel DL framework is introduced for BC of reanalysis-derived wind speeds. This framework uses a DL model to learn the temporal correlation between meteorological variables from ERA5 dataset and a time-resolved scaling factor. The time-resolved scaling factor is used to adjust ERA5 wind speeds.

6.3 Dataset

6.3.1 Measured wind speed data

For the experiments conducted in this study, hourly wind speed measurements at 10 m above ground level (a.g.l) were obtained from Environment and Climate Change Canada (ECCC) historical climate data archives. Stations included in this study were selected based on the following criteria. 1) Stations with at least one valid year between 2008 and 2023 were selected. A valid year has at least 80% data availability, and each day has a complete 24-hour record (no missing values). 2) A minimum distance of 1 km between the stations was used to filter out potential duplicate stations, and only the station with the longest data record was retained.

The 10 years from 2008 to 2017 were used to train the models to match the temporal coverage of the GWA version 3 [GWA3; 9], while testing was conducted with data from 2018 to 2023. Figure 6.1 shows a map of the selected stations in Canada and their median wind speed from 2008 to 2023. The measured wind speeds were considered as the reference for model evaluation.

6.3.2 Dynamic covariates

Various meteorological variables from ERA5 hourly data on single levels were downloaded (https://doi.org/10.24381/cds.adbb2d47) and used as time-resolved covariates. ERA5 hourly data are consistent and available from 1940 to present with no missing values and a spatial resolution of 0.25 degrees on a regular latitude and longitude grid. The 10 m wind speed was derived from the 10 m u- and 10 m v-components (i.e., $\sqrt{u^2 + v^2}$). The time-resolved covariates employed in this study included 10 m u- and 10 m v-wind components, 10 m wind speed, 2 m temperature (2t), boundary layer height (blh), and surface pressure (sp). They were selected based on a review of previous studies on wind speed downscaling (Dujardin *et al.*, 2022; Dupuy *et al.*, 2023; Höhlein *et al.*, 2020). Additionally, wind speed quantiles (5th, 50th, and 95th percentiles) estimated from ERA5 hourly wind speeds between 2008 and 2017 were used as covariates to represent long-term wind conditions. Houndekindo *et al.* (2024) found that incorporating wind speed quantiles as covariates improved model performance for wind speed distribution mapping in Canada.

6.3.3 Static covariates

Static covariates derived from the Advanced Land Observing Satellite (ALOS) Digital Elevation Model (DEM) (Tadono *et al.*, 2014) and Canada 2020 land cover map (Latifovic *et al.*, 2017) were also used as predictors. These covariates provide valuable information on fine-scale terrain features and SRL, which may not be captured in coarse-resolution reanalysis data. Table 10.2 in the supplementary material (Section 10) provides the assigned SRL based on the land cover class. A detailed list of the static covariates is presented in Table 10.1 in the supplementary materials (Section 10), and their importance for wind speed modeling is discussed in detail by Houndekindo et al. (2023a).

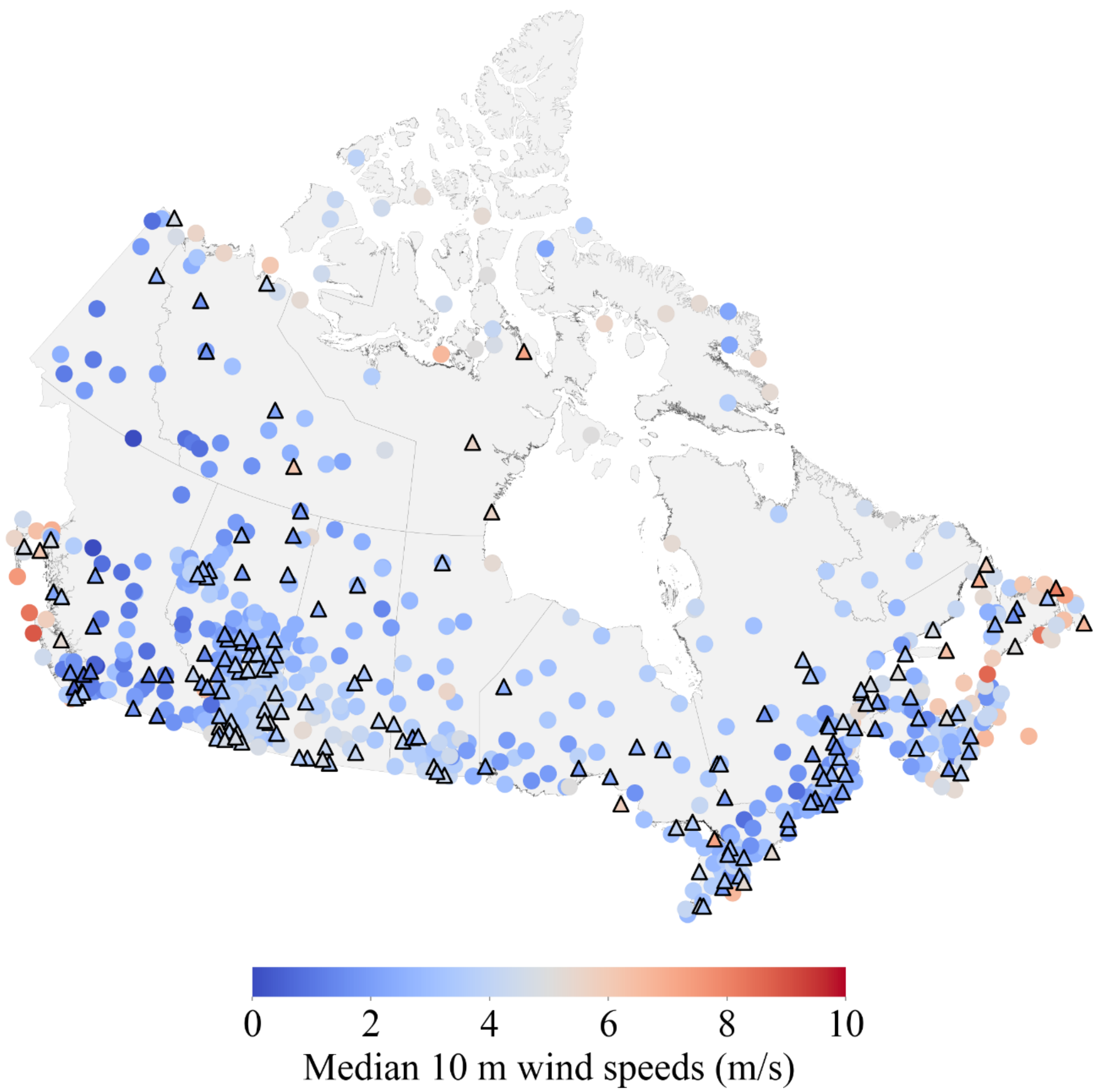


Figure 6.1 Locations of ECCC meteorological stations included in the study.

All the station are located onshore in Canada. The stations were randomly split into training (circle symbol) and test (triangle symbol) sets. The colors indicate the median 10 m wind speeds from 2008 –2023.

6.4 Method

The study flow diagram is illustrated in Figure 6.2 and includes the following main steps: (1) data acquisition, as detailed in section 6.3; (2) data processing, also described in section 6.3; (3) development and training of time-invariant and time-resolved models, outlined in section 6.4.1 and 6.4.2; (4) prediction of test data; and (5) evaluation of models, detailed in section 6.4.3.

This section is structured into three parts. First, the methodology for modeling the time-invariant scaling factor using standard ML is presented. Second, the procedure for modeling the time-resolved scaling factor using DL is outlined. Finally, the procedure for comprehensively evaluating the proposed methods is described.

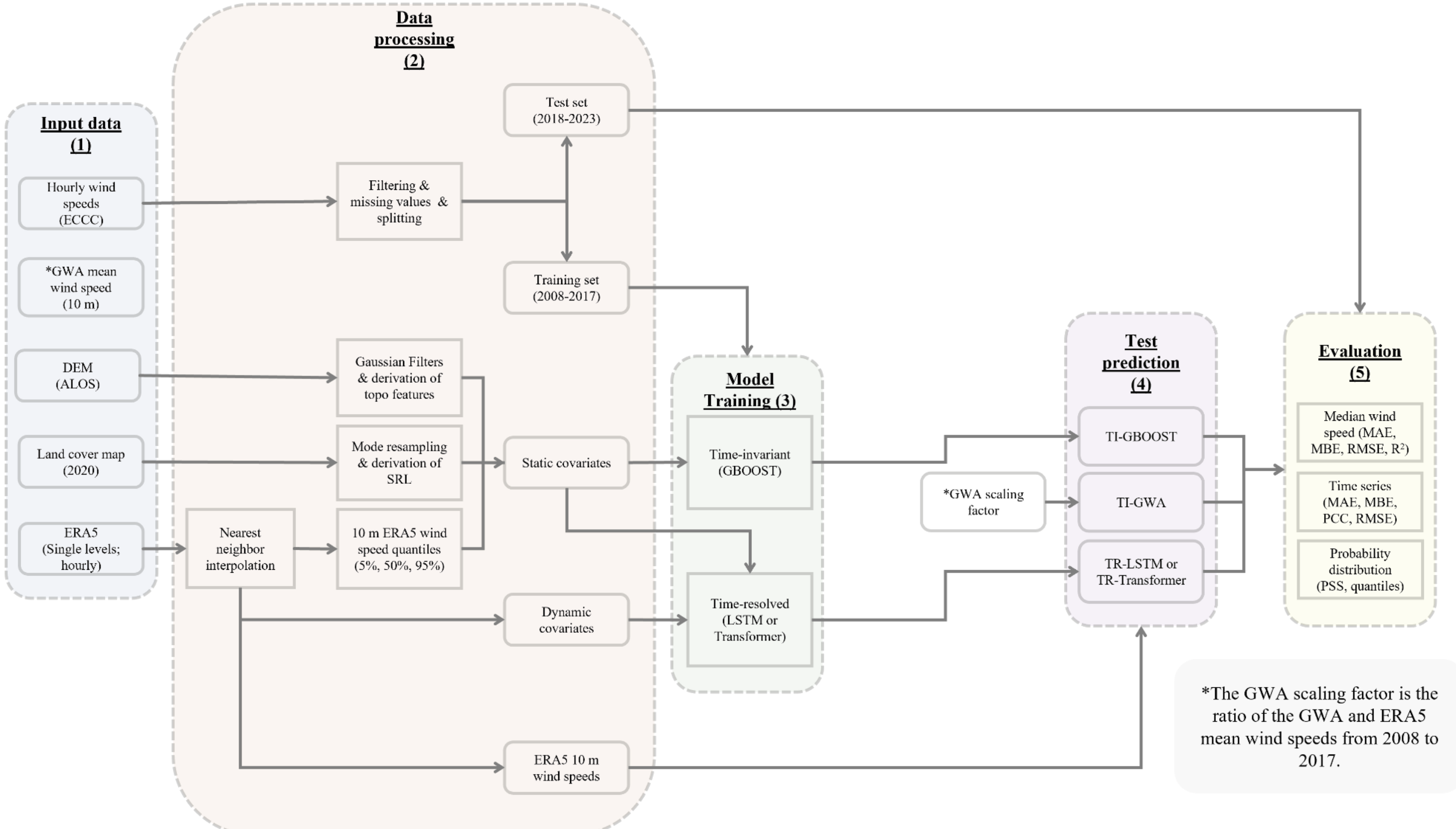


Figure 6.2 Study workflow.

The input data are preprocessed to derive static and dynamic covariates. The static covariates are derived from the DEM, land cover map, and ERA5 10 m wind speed quantiles, while the dynamic covariates are obtained from the ERA5 dataset. ERA5 data are interpolated to the station locations using nearest neighbor interpolation. Measured wind speeds from ECCC historical climate data archives serve as the target variable. The ECCC stations are randomly split into training and test sets. The training set is used to train the time-invariant and time-resolved model, which are evaluated on the test set. The GWA bias-correction method serves as a baseline during the evaluation. The

bias-correction methods predict scaling factors —either time-invariant (TI-GWA, TI-GBOOST) or time-resolved (TR-LSTM, TR-Transformer)—which are applied to ERA5 10 m wind speeds before the evaluation.

6.4.1 Time-invariant scaling factor

A common method for BC of reanalysis wind speed involves estimating a time-invariant scaling factor at the target location s_i using the following equation:

Equation 6.1

$$\hat{y}(s_i) = \frac{\overline{U}^{GWA}(s_i)}{\overline{U}^{ERA5}(s_i)}$$

where $\overline{\boldsymbol{U}}^{ERA5}(\boldsymbol{s_i})$ and $\overline{\boldsymbol{U}}^{GWA}(\boldsymbol{s_i})$ represent the mean wind speed from ERA5 and GWA dataset, respectively.

This scaling factor is then used to adjust the long-term mean wind speed from ERA5 as follows:

Equation 6.2

$$\widehat{U}_t(s_i) = U_t^{ERA5}(s_i) \times \hat{y}(s_i)$$

where $\boldsymbol{U_t^{ERA5}}(\boldsymbol{s_i})$ and $\widehat{\boldsymbol{U}}_{\boldsymbol{t}}(\boldsymbol{s_i})$ denote the wind speed from ERA5 and the adjusted wind speed at time $\boldsymbol{t}$, respectively.

In this study, it is proposed to model the time-invariant scaling factor as a function of static covariates (denoted $\boldsymbol{k}(s_i)$, throughout the remainder of the paper, vectors and matrices in equations are represented by boldface letters) related to the location topography and SRL, along with the long-term wind speed condition derived from ERA5:

Equation 6.3

$$\hat{y}(s_i) = f\big(\boldsymbol{k}(s_i)\big)$$

where $\boldsymbol{f}(\cdot)$ represents the regression function that maps the static covariates to the time-invariant scaling factor.

The gradient boosting regressor (herein referred to as GBOOST) from the Scikit-learn package (Pedregosa *et al.*, 2011) was employed to model the regression function. Gradient boosting is an ensemble technique that iteratively combines weak learners (regression trees) to form a robust predictive model. In this approach, each regression tree is fitted to the negative gradient of the loss function from the preceding iteration. The final prediction is obtained by summing the outputs of all regression trees (Jerome, 2001). Various studies have applied gradient-boosting models for wind speed modeling (Houndekindo *et al.*, 2023b; Houndekindo *et al.*, 2024; Hu *et al.*, 2023; Jung *et al.*, 2023b). Moreover, these models have demonstrated superior performance than DL models when applied to tabular data and require less tuning (Grinsztajn *et al.*, 2022a; Shwartz-Ziv *et al.*, 2022). The parameters of GBOOST (Table 6.1) were optimized using random search (Bergstra *et al.*, 2012a). Other optimisation algorithms are available in the literature (Abdollahzadeh *et al.*, 2024; El-kenawy *et al.*, 2024; Mona Ahmed Yassen *et al.*, 2024) and may be explored in future studies. The covariates included in the final regression model were selected based on the permutation feature importance technique (Breiman, 2001).

Table 6.1 List of GBOOST parameters that were tuned.

Parameters	Description	Tested values
learning_rate	Weighting factor applied to the contribution of each individual tree	0.01, 0.02, 0.05, 0.1, 0.2
subsample	Ratio of the training data to randomly sample for training	0.5, 0.6, 0.7, 0.8, 0.9
max_depth	Maximum depth of tree	2, 3, 4, 5, 6, 8
max_features	Maximum number of covariates to consider when looking for the best split	0.1, 0.2, 0.3, 0.4, 0.5
n_estimators	Number of boosting iterations	50, 100, 200, 400, 800, 1000
n_iter_no_change	The number of iterations after which training is stopped if the model's performance on the validation set does not improve.	10, 20, 50, 100
min_samples_split	The minimum number of samples required to split an internal node of the tree	0.001, 0.01, 0.02, 0.05, 0.1

6.4.2 Time-resolved scaling factor

A time-invariant scaling factor can only correct systematic biases in reanalysis-derived wind speed data without improving its temporal variability. To overcome this limitation, a DL framework is proposed to predict a time-resolved scaling factor. This framework aims to correct systematic biases and enhance the temporal correlation between reanalysis-derived and actual wind speed. The task of predicting a time-resolved scaling factor using static and time-resolved covariates can be formulated as follows. Let $y_t(s_i)$ represent the scaling factor between observed wind speeds $U_t(s_i)$ and ERA5 wind speeds $U_t^{ERA5}(s_i)$ at time t and location s_i:

Equation 6.4

$$y_t(s_i) = \frac{U_t(s_i)}{U_t^{ERA5}(s_i)}$$

The modeling of the time-resolved scaling factor $y_t(s_i)$ can then be expressed as:

Equation 6.5

$$\hat{y}_t(s_i) = f\left(\boldsymbol{k}(s_i), \boldsymbol{v}_{t-p:t}(s_i)\right)$$

where $\hat{y}_t(s_i)$ is the estimated scaling factor at time t and location s_i, p represents the length of the finite look-back window for the time-resolved reanalysis covariates $v_t(s_i)$, and $f(\cdot)$ denotes the prediction function.

6.4.2.1 Model architecture

The prediction function of the time-resolved models was parameterized using a two-branch DL architecture (Figure 6.3). One branch processes the static covariates (herein referred to as the static branch), while the other processes the time-resolved covariates (herein referred to as the dynamic branch).

The static covariates $\boldsymbol{k}(s_i)$ are encoded in the static branch through a two-layer feed-forward neural network (FFN) with the Gaussian Error Linear Unit (GELU) activation function:

Equation 6.6

$$\boldsymbol{\chi}(s_i) = FFN\left(\boldsymbol{k}(s_i)\right)$$

where $FFN(x) = GELU(xW_1 + b_1)W_2 + b_2$, with learnable parameters W_1, W_2, b_1, b_2, and $\chi(s_i)$ represents the hidden vector from the static branch.

In the dynamic branch, the time-resolved meteorological covariates are concatenated with the static covariates, which are replicated through time. This integration of static covariates into the dynamic branch provides location-specific context for the temporal model. The concatenated static and time-resolved covariates are then linearly projected and combined with embedded hour, day, and month timestamps, which provide the temporal context necessary for effective time series modeling (Wen et al., 2023). This is expressed as:

Equation 6.7

$$\boldsymbol{\zeta}'_t(s_i) = LP_1([\boldsymbol{k}(s_i), \boldsymbol{v}_t(s_i)]) + TEmbed(\boldsymbol{\tau_t})$$

where $TEmbed(\cdot)$ denotes the temporal embedding layer that maps the timestamp vector τ_t into hidden features, and $LP_1(x) = xW_3 + b_3$ is a linear projection with learnable parameters W_3 and b_3.

Although timestamp embeddings were initially proposed for Transformer architectures (Zhou *et al.*, 2021a), they were also found to improve the performance of the LSTM in this study. For the Transformers, in addition to the embedded timestamps, fixed sinusoidal positional encoding

(Vaswani *et al.*, 2017) was also added to the projected covariates, enabling the model to account for sequence order.

In the next step of the dynamic branch, the hidden features $\boldsymbol{\zeta}'_t(s_i)$ are fed into the temporal model:

Equation 6.8

$$\boldsymbol{\zeta}_t(s_i) = TModel\left(\boldsymbol{\zeta}'_{t-p:t}(s_i)\right)$$

where $\boldsymbol{TModel}(\cdot)$ represents the temporal model that learns the temporal patterns within the encoded time-revolved covariates that influence the target variable, subsections 6.4.2.2 and 6.4.2.3 describe in detail the different architectures of the temporal model explored.

Finally, the outputs from the static and dynamic branches are concatenated, and the model's prediction is obtained through a linear projection (denoted as LP_2), followed by the Softplus activation function to ensure that the predicted scaling factors are non-negative:

Equation 6.9

$$\hat{y}_t(s_i) = Softplus\left(LP_2([\boldsymbol{\chi}(s_i), \boldsymbol{\zeta}_t(s_i)])\right)$$

where $\hat{\boldsymbol{y}}_t(\boldsymbol{s}_i)$ denotes the predicted scaling factors, and S$\boldsymbol{oftplus}(\cdot)$ is the Softplus activation function.

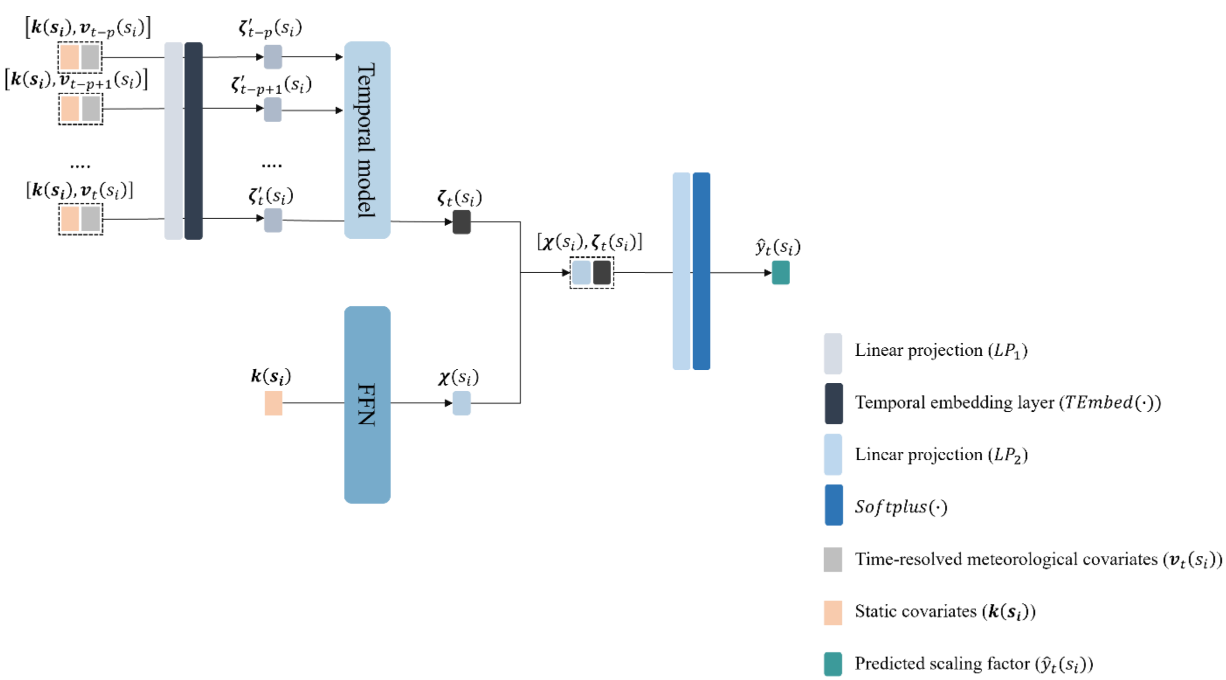


Figure 6.3 Overview of the DL framework.

The static covariates are embedded using a feed-forward neural network (FFN). The time-resolved meteorological covariates are combined with the static covariates and processed by a

temporal model, which is either an LSTM or a Transformer. The output from the static covariates FFN and the temporal model are combined, passed through a linear projection layer, and finally transformed using a softplus activation function.

6.4.2.2 LSTM

Traditional RNNs face issues such as vanishing and exploding gradients, which limits their ability to retain information over extended sequences (Goodfellow *et al.*, 2016). By incorporating a gating mechanism that regulates the flow of information, LSTMs can effectively maintain and use relevant data over more extended periods. The equations describing LSTMs are given by:

Equation 6.10

$$\boldsymbol{f}_t = \sigma\left(\boldsymbol{W}_f[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_f\right)$$

Equation 6.11

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_i)$$

Equation 6.12

$$\tilde{\boldsymbol{c}}_t = \tanh(\boldsymbol{W}_c[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_c)$$

Equation 6.13

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot c_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t$$

Equation 6.14

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_o)$$

Equation 6.15

$$\boldsymbol{h}_t = \tanh \boldsymbol{c}_t \odot \boldsymbol{o}_t$$

where $\boldsymbol{f}_t$, $\boldsymbol{i}_t$, $\tilde{c}_t$ and o_t represents the forget, input, cell and output gates, respectively. The vector x_t is the LSTM input, the matrices W_j and vectors b_j (where $j \in \{f, i, c, o\}$) are the set of learnable parameters, h_{t-1} is the previous hidden states, and h_t is the updated hidden states. The vector c_{t-1} is the previous cell state, and c_t is the updated cell state. The element-wise functions $\sigma(\cdot)$ and $\tanh(\cdot)$ denotes the sigmoid and hyperbolic tangent activation functions, respectively, and the notation $\odot$ denotes the element-wise product.

6.4.2.3 Transformer model

The original Transformer architecture with an encoder-decoder setup was designed for sequence-to-sequence tasks, such as translation. In this study, an encoder-only Transformer architecture was implemented (Figure 6.4). This model is composed of multiple identical encoder layers stacked with residual connections to ensure the smooth flow of information between inputs and outputs. Each encoder layer has two main components: (1) a multi-head self-attention (MHA) mechanism and (2) a FFN. The MHA module enables the model to attend to different parts of the input sequence by simultaneously performing the attention mechanism across multiple subspace representations. The scaled dot-product attention was used as the attention mechanism and is defined as follows:

Equation 6.16

$$H_i = Attention(Q_i, K_i, V_i) = \mathrm{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_{attn}}}\right) V_i$$

where $\boldsymbol{Q_i = QW_i^Q}$, $\boldsymbol{K_i = KW_i^K}$, and $\boldsymbol{V_i = VW_i^V}$ represent the queries, keys and values associated with the attention head $\boldsymbol{H_i}$ (where $\boldsymbol{i = 1, 2, \ldots, nh}$). In the self-attention, $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$ are all equal to the Transformer input.

The final output of the MHA module is obtained by linearly projecting the concatenated attention heads H_i:

Equation 6.17

$$\mathrm{MHA}(Q, K, V) = [H_1, \ldots, H_{nh}]W^O$$

The matrices $\boldsymbol{W_i^Q}$, $\boldsymbol{W_i^K}$, $\boldsymbol{W_i^V}$ and $\boldsymbol{W^O}$ are the learnable parameters of the MHA module.

The output from the MHA module is passed through a FFN with a GELU activation function. The Pre-norm Transformer architecture (Wang *et al.*, 2019) was implemented in this study. In this configuration, the input to each encoder sub-layer undergoes layer-normalization (LN) before being passed through the sub-layer, and a residual connection is added to the output of the sub-layer:

Equation 6.18

$$\boldsymbol{h} = \boldsymbol{x} + \mathcal{F}\big(LN(\boldsymbol{x})\big)$$

where $\mathcal{F}(\cdot)$ represents one of the sub-layers of the Transformer encoder, x is the input of the sub-layer, and h is the output.

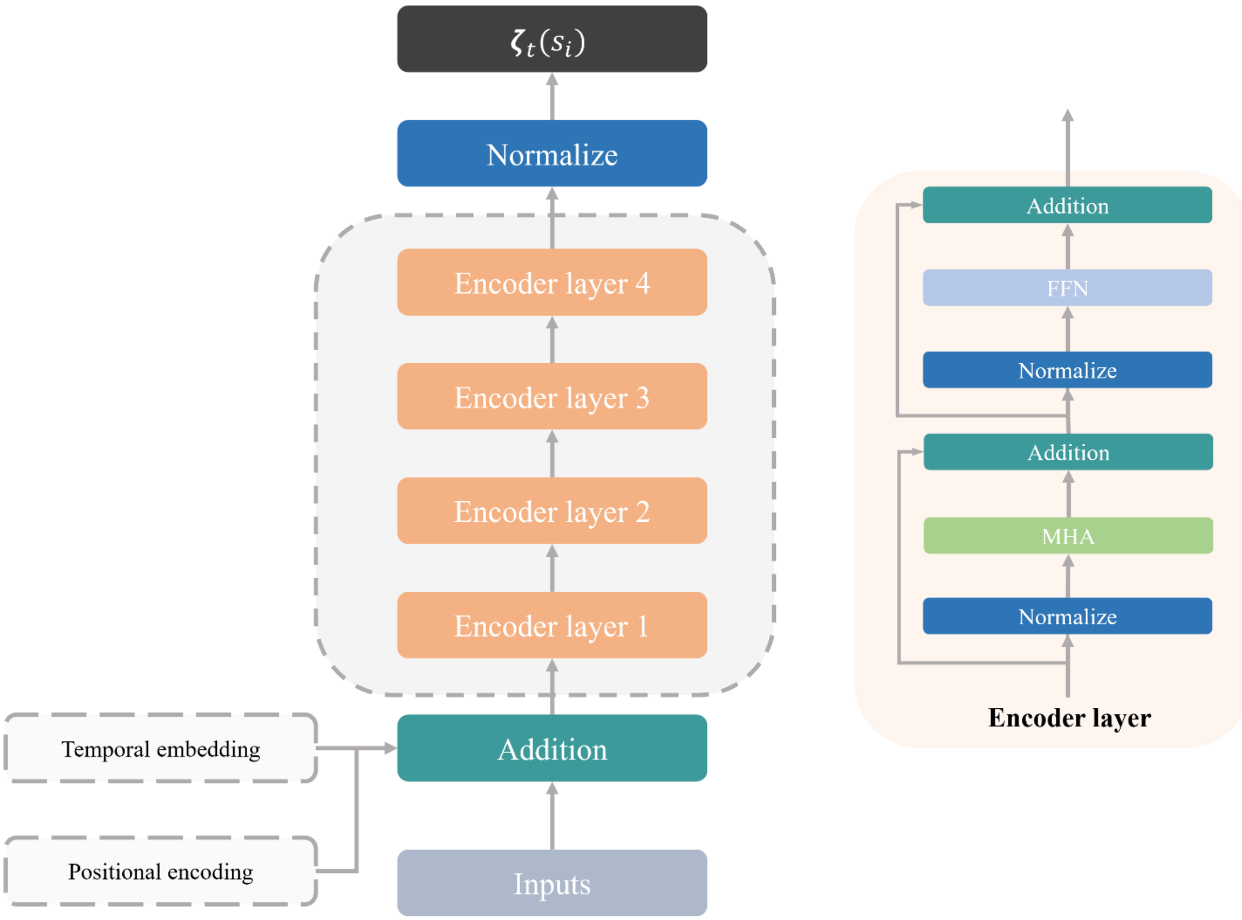


Figure 6.4 Overview of the Transformer architecture.

Positional encoding and temporal embedding are first added to the inputs, which are then passed through four encoder layers. Each encoder layer consists of multi-head self-attention (MHA), a feed-forward neural network (FFN), and skip connections. Layer normalization is applied to the input of each component within the encoder layers.

6.4.2.4 Data preprocessing and model training

The static and time-resolved covariates used as inputs to the model were preprocessed to improve training performance. The static covariates were normalized to a range between 0 and 1 using the min-max scaling, and the time-resolved covariates were standardized using a mean and standard deviation estimated separately for each grid point of the reanalysis dataset, except for the wind components (u, v) which were normalized by dividing by their magnitude to preserve information about wind direction. The observed scaling factors, which serve as the target variable, were derived from ECCC and ERA5 wind speeds ratio. ERA5 data were interpolated to each station's location using nearest-neighbor interpolation.

The deep learning models were implemented with PyTorch (Paszke *et al.*, 2019) and optimized using the Adam optimizer (Kingma *et al.*, 2015). The Adam optimizer is widely used for neural network optimization, and it is relatively robust for hyperparameter selection (Goodfellow *et al.*, 2016). Early stopping was employed to mitigate overfitting, halting training when the validation loss did not improve for ten consecutive epochs. The parameters used to train the models are presented in Table 6.2, which were selected based on best practices and manual tuning.

The proposed deep learning models were trained by minimizing the Huber loss function (Huber, 1964) between the model prediction and the target variable. The Huber loss function is defined as:

Equation 6.19

$$L_{\alpha}(y,\hat{y}) = \begin{cases} \frac{1}{2}(y-\hat{y})^2 & |y-\hat{y}| \leq \alpha \\ \alpha\left(|y-\hat{y}| - \frac{1}{2}\alpha\right) & |y-\hat{y}| > \alpha \end{cases}$$

where α is a parameter that controls the transition between the quadratic and linear behavior of the loss function.

Standard loss functions for regression tasks include mean square error (MSE) and mean absolute error (MAE). The MSE loss function is differentiable everywhere but sensitive to outliers, while the MAE is less sensitive to outliers but non-differentiable at zero. The Huber loss function balances MSE and MAE by behaving quadratically for minor errors and linearly for significant errors. it offers a level of robustness against outliers while being differentiable everywhere.

Table 6.2 List of selected parameters of the DL models

Parameters	Temporal model	
	LSTM	Transformer
Learning rate	2e-5	1e-5
Static branch dropout rate	0.05	0.05
Dynamic branch dropout rate	0.2	0.1
Look-back window	120 hours	120 hours
Static branch hidden size	32	32
Dynamic branch hidden size	256	256
Number of heads	n.a	4
Number of layers	1	4
Batch size	128	128
Huber loss parameter ($\boldsymbol{\alpha}$)	1.5	1.5
Early stopping round	10	10
Adam parameters ($\boldsymbol{\beta_1}, \boldsymbol{\beta_2}$)	0.9, 0.999	0.9, 0.999

6.4.3 Evaluation

The performance of reanalysis-derived wind speeds varies depending on the type of location (onshore, coastal, or offshore) and the complexity of the terrain (flat, hilly, or mountainous). Evaluations were conducted across different area types to comprehensively assess model performance, including coastal, hilly, and mountainous areas and regions with low or high SRL. Table 10.3 in the supplementary material lists the criteria used to classify the different ECCC stations within the study area.

Furthermore, the available stations were randomly divided into training (565 stations) and test (170 stations) sets, with the training set represented by circle symbols and the test set by triangle symbols in Figure 6.1. The test set was reserved exclusively for the final evaluation of the models. Model training and parameter tuning were performed using k-fold cross-validation with the training set with data from 2008 to 2017. The selected models were then evaluated with the test sets from 2018 to 2023 to assess their spatiotemporal generalization capabilities.

The analysis included evaluating the median wind speed, temporal variability, and probability distribution, using various scoring metrics to ensure a comprehensive evaluation of the models in the context of WRA. The median wind speed was selected as the measure of central tendency because it is robust and resistant, unlike the mean (Wilks, 2011). For skewed distributions, like wind speed, the median offers a more accurate representation of the typical wind speed value.

Table 6.3 presents the equations for the different scoring metrics used. The coefficient of determination (R^2), MAE, mean bias error (MBE), and root mean squared error (RMSE) were used to compare the predicted and observed median wind speeds. The R^2 served as a more interpretable metric, with values between 0 (indicating the worst performance) and 1 (indicating the best performance). It measures the model's effectiveness in capturing the spatial variability of the median wind speeds.

The MAE, MBE, Pearson correlation coefficient (PCC), and RMSE were used to compare the predicted and observed wind speed time series. The MAE and RMSE metrics are commonly used for time series model evaluation (Hodson, 2022). They complement each other in assessing different aspects of model performance (Chai *et al.*, 2014). The MAE directly measures the average error magnitude, making it valuable in assessing the overall bias correction effectiveness. The RMSE, which heavily penalizes more significant errors, highlights extreme deviations in the wind speed predictions. The MBE provides insight into the systematic tendency of the model to overestimate or underestimate. Meanwhile, the PCC quantifies the linear relationship between predicted and observed wind speeds, measuring how well the model captures the up-down trends (temporal variation) of the observed wind speeds (Liemohn *et al.*, 2021).

In addition, to assess the relative accuracy of the BC methods in comparison to the ERA5 wind speeds, the percentage improvement or skill scores for the time series MAE and RMSE were calculated. The percentage improvement is defined by Wilks (2011) as follows:

Equation 6.20

$$\text{Percentage Improvement} = \frac{SM_{model} - SM_{ERA5}}{SM_{perf} - SM_{ERA5}} \times 100\%$$

where SM_{model} and SM_{ERA5} are the scoring metrics of the model and ERA5 (the reference), respectively, and SM_{perf} denotes the value of the scoring metric for a perfect prediction (0 in the case of MAE and RMSE).

The Perkins skill score (PSS; Perkins *et al.*, 2007) was employed to evaluate the model performance regarding probability density functions (PDFs). The PSS was used in various wind speed modeling and BC studies (Belušić *et al.*, 2018; Carvalho *et al.*, 2021; Costoya *et al.*, 2020; Houndekindo *et al.*, 2024; Jung *et al.*, 2023a; Molina *et al.*, 2021). This metric can evaluate the PDF across the entire range of wind speeds and in specific parts of the distribution (e.g., lower or upper tails) (Keellings, 2016; Zhang *et al.*, 2024). Given that data in the distribution's tails are less frequent, the PSS calculated across the full range of wind speeds may not accurately represent model performance in these tails. Therefore, in addition to computing the PSS across the entire wind speed range (herein referred to as PSS-ALL), the PSS was also computed separately for wind speed values below the 10th percentile in the lower tail (herein referred to as PSS-LWT) and above the 90th percentile in the upper tail (herein referred to as PSS-UPT). The percentiles were estimated from the distribution of observed wind speeds and have been used in previous studies to characterize high and low wind speeds (see, e.g., Torralba, et al. [96]).

In addition to the PSS metrics, wind speed quantiles at percentile levels ranging from 10% to 90%, in 10% increments, were computed from the predicted and measured wind speed time series. These quantiles were compared using MAE, MBE, RMSE, and R^2. This evaluation provides a detailed assessment of the model performance across different wind speed levels, offering insight on how accurately the models represent low, moderate, and high wind conditions.

Table 6.3 Equations of the different scoring metrics used to evaluate the BC methods.

Scoring metric	Equation	Evaluation component
Coefficient of determination	$R^2 = 1 - \frac{\sum_{i=1}^{n}(o_i - p_i)^2}{\sum_{i=1}^{n}(o - \bar{o})^2}$	Wind speed quantiles

Mean absolute error	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert o_i - p_i\rvert$	Wind speed quantiles and time series
Mean bias error	$MBE = \frac{1}{n}\sum_{i=1}^{n} o_i - p_i$	Wind speed quantiles and time series
Pearson correlation coefficient	$PCC = \frac{\sum_{i=1}^{n}(o_i-\bar{o})(p_i-\bar{p})}{\sqrt{\sum_{i=1}^{n}(o_i-\bar{o})^2}\sqrt{\sum_{i=1}^{n}(p_i-\bar{p})^2}}$	Time series
Perkins' skill score	$PSS = \sum_{i=1}^{B}\min(\hat{Z}_i, Z_i) \times 100$	Probability distribution
Root mean squared error	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(o_i - p_i)^2}$	Wind speed quantiles and time series

Note: o represents the observed variable, p the predicted variable, B the number of bins, n the sample size, Z the observed wind speed frequency, and $\hat{Z}$ the predicted wind speed frequency

6.5 Results

This section presents the results of the evaluation of the BC methods. The distributions of the scoring metrics are visualized using boxplots, while their summary statistics (median values) are listed in tables for clarity. To ensure a comprehensive assessment, the model performance is evaluated across various regions, including coastal areas, regions with high and low SRL, and hilly and mountainous areas.

In the following subsections, the following abbreviations are used to refer to the evaluated models: Time-invariant BC with GBOOST (TI-GBOOST), Time-invariant BC with GWA3 (TI-GWA3), Time-resolved BC with LSTM (TR-LSTM), Time-resolved BC with Transformer (TR-Transformer), and uncorrected ERA5 wind speed data (UC-ERA5).

6.5.1 Mean wind speed evaluation

The median wind speed indicates the central tendency and represents typical wind speed conditions at a given location. Table 6.4 lists the scoring metrics (MAE, MBE, R^2, and RMSE) used to evaluate the median wind speed. Figure 6.5 presents scatter plots comparing the observed and predicted median wind speeds.

The ML-based BC methods achieved the lowest MAE, RMSE, and the highest R^2. In coastal areas, TR-Transformer was the best-performing model, slightly outperforming TR-LSTM. In regions with high and low SRL and hilly and mountainous areas, the TI-GBOOST model emerged as the top performer. The TR-LSTM model exhibited the lowest overall MBE, though all models tended to overestimate the median wind speeds (positive MBE) across most regions. However, in regions with High SRL, the TR-LSTM and TR-Transformer models tended to underestimate the median wind speeds. These patterns are also evident in Figure 6.5.

Table 6.4 Evaluation metrics for median wind speed.

Metric	Area	UC-ERA5	TI-GWA3	TI-GBOOST	TR-LSTM	TR-Transformer
MAE	Coastal (n=45)	1.13	0.91	0.91	0.85	**0.83**
	High SRL (n=36)	0.64	0.74	**0.40**	0.43	0.47
	Hilly and mountainous (n=29)	1.03	1.16	**0.92**	0.96	0.99
	Low SRL (n=60)	0.61	0.56	**0.50**	0.51	0.54
	Overall (n=170)	0.83	0.79	**0.66**	**0.66**	0.68
MBE	Coastal (n=45)	0.55	0.64	0.47	0.26	**0.25**
	High SRL (n=36)	0.40	0.58	**0.04**	-0.19	-0.12
	Hilly and mountainous (n=29)	0.40	0.50	0.40	**0.00**	0.07
	Low SRL (n=60)	-0.12	0.43	0.16	**0.07**	**0.07**
	Overall (n=170)	0.25	0.53	0.26	**0.05**	0.08
R^2	Coastal (n=45)	0.11	0.43	0.43	0.46	**0.47**
	High SRL (n=36)	0.12	-0.13	**0.56**	0.25	0.43
	Hilly and mountainous (n=29)	0.57	0.46	**0.64**	0.62	0.57
	Low SRL (n=60)	0.40	0.46	**0.57**	**0.57**	0.52
	Overall (n=170)	0.39	0.45	**0.58**	0.56	0.56
RMSE	Coastal (n=45)	1.44	1.15	1.14	1.12	**1.10**
	High SRL (n=36)	0.77	0.87	**0.54**	0.71	0.61
	Hilly and mountainous (n=29)	1.26	1.40	**1.15**	1.18	1.25
	Low SRL (n=60)	0.81	0.77	**0.68**	**0.68**	0.72
	Overall (n=170)	1.08	1.03	**0.89**	0.92	0.92

The table presents the scoring metrics for each area type, along with the overall performance. The best results are highlighted in bold, and n represents the sample size for each area.

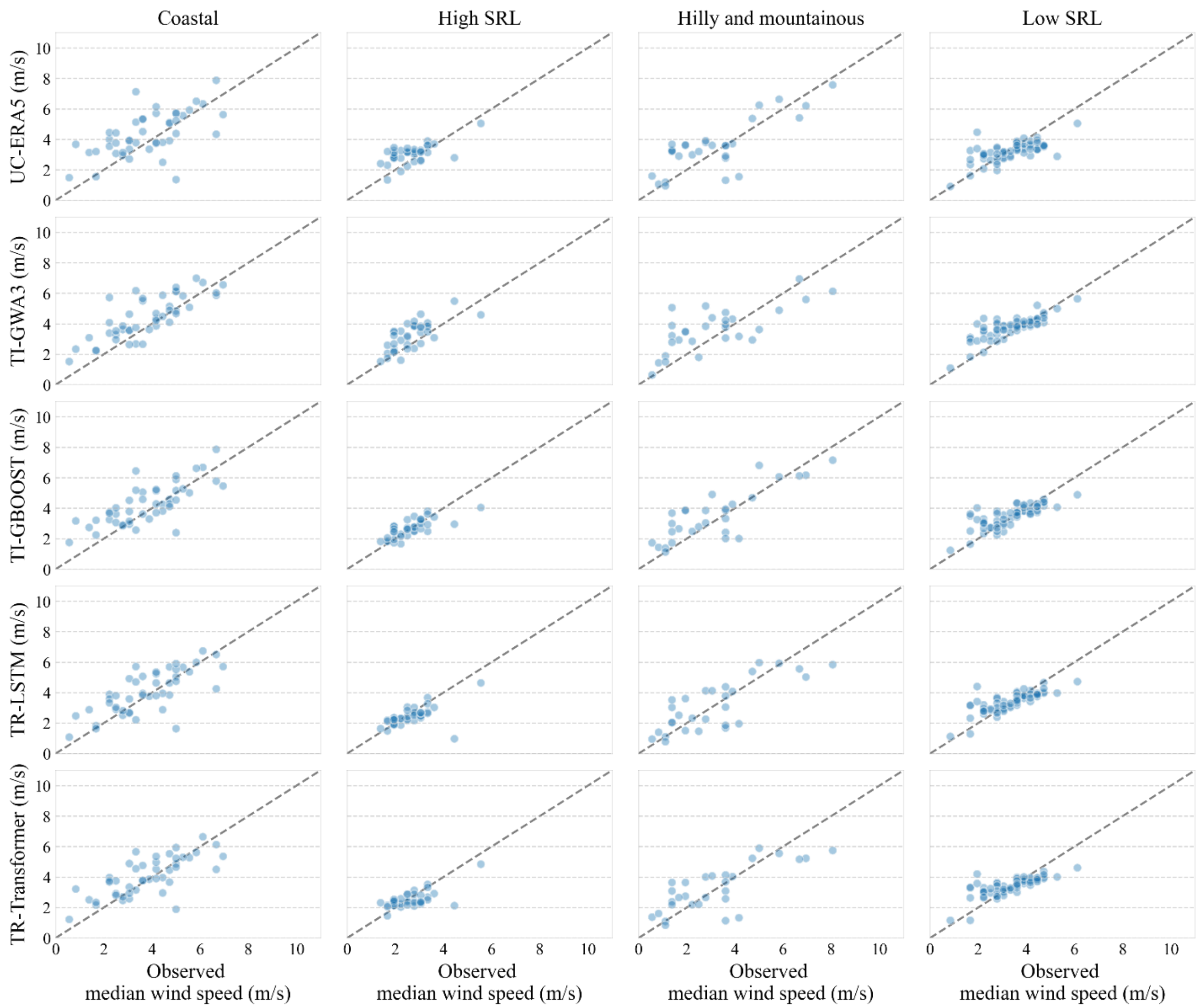


Figure 6.5 Scatter plots between observed and predicted median wind speeds.

The scatter plots are presented across different models (rows) and area types (columns). The 45° line is included for reference, indicating perfect agreement between observed and predicted values. Deviations from this line highlight the discrepancy between the predicted and actual wind speeds. These plots provide a visual representation of how well each model captures the central tendency of wind speeds in diverse settings. Overall, the best agreement between the predicted and observed median wind speeds are in regions with high and low SRL.

6.5.2 Time series evaluation

Figure 6.6 displays boxplots of the time series metrics (rows) for different area types (columns), while Table 6.5 lists the median metrics calculated across the test stations. The TR-LSTM and TR-Transformer emerged as the top performers, achieving the lowest median MAE and RMSE values and the highest median PCC. The ML-based BC methods consistently improved the overall median metrics compared to UC-ERA5. In contrast, TI-GWA3 often showed lower overall performance relative to UC-ERA5. In addition, the spatial variability of the metrics, as shown by the interquartile range in Figure 6.6, was generally more pronounced in coastal, hilly, and mountainous regions compared to inland regions with low or high SRL.

To further examine the performance of the time-resolved models (TR-LSTM and TR-Transformer), Figure 6.7 presents maps showing the spatial distribution of the percentage improvements over UC-ERA5 for MAE (leftmost column) and RMSE (rightmost column), with the corresponding boxplots provided in Figure 10.2 of the supplementary material (Section 10). The percentage improvements in time series metrics exhibit spatial heterogeneity without a distinct spatial pattern. In some instances, UC-ERA5 offered more reliable estimates, highlighting the importance of identifying conditions in which time-resolved models outperform UC-ERA5.

A further analysis was conducted to examine the Spearman correlations, which is a nonparametric measure of statistical dependence between two variables, between the most important static covariates identified through permutation feature importance (see Figure 10.1 in the supplementary material, Section 10) and percentage improvements in time series MAE and RMSE over UC-ERA5 for TR-LSTM. The leftmost correlation matrix in Figure 6.8 displays the correlations for TR-LSTM, while the rightmost matrix shows those for TR-Transformer.

Three groups of static covariates —standard deviation of slope (SDS), SRL, and UC-ERA5 quantiles —showed a relatively high correlation with percentage improvements in MAE and RMSE. The SRL was the covariate most strongly correlated with the performance metrics. The direction of the correlation indicates that higher SRL values are associated with more significant improvements in time series MAE and RMSE. SDS showed a consistent mild correlation with improvements in these time series metrics.

Section 10.1 of the supplementary material presents additional figures (Figure 10.3 to Figure 10.10) for visual inspection of the time-resolved model. These figures present the uncorrected and corrected wind speed time series at four selected stations of the test set, each representing a different area type.

Table 6.5 Summary of the time series evaluation metrics (MAE, MBE, PCC, and RMSE).

Metric	Area	UC-ERA5	TI-GWA3	TI-GBOOST	TR-LSTM	TR-Transformer
MAE	Coastal (n=45)	1.74	1.61	1.63	**1.55**	1.56
	High SRL (n=36)	1.13	1.20	1.05	**0.90**	0.94
	Hilly and mountainous (n=29)	1.71	1.60	1.57	1.51	**1.44**
	Low SRL (n=60)	1.25	1.31	1.22	**1.17**	1.19
	Overall (n=170)	1.32	1.35	1.29	**1.19**	1.23
MBE	Coastal (n=45)	0.37	0.44	0.48	0.24	**0.09**

	High SRL (n=36)	0.32	0.46	-0.04	-0.11	**0.00**
	Hilly and mountainous (n=29)	0.54	0.47	0.17	0.19	**0.15**
	Low SRL (n=60)	-0.32	0.17	-0.03	**0.00**	-0.01
	Overall (n=170)	0.10	0.35	**0.04**	0.09	0.06
PCC	Coastal (n=45)	0.77	0.77	0.77	**0.79**	**0.79**
	High SRL (n=36)	0.74	0.74	0.74	**0.79**	**0.79**
	Hilly and mountainous (n=29)	0.72	0.72	0.72	**0.74**	**0.74**
	Low SRL (n=60)	0.78	0.78	0.78	**0.80**	**0.80**
	Overall (n=170)	0.76	0.76	0.76	**0.79**	**0.79**
RMSE	Coastal (n=45)	2.25	2.08	2.06	**1.97**	2.03
	High SRL (n=36)	1.41	1.50	1.31	**1.19**	1.23
	Hilly and mountainous (n=29)	2.01	2.04	2.03	1.96	**1.91**
	Low SRL (n=60)	1.61	1.68	1.56	**1.51**	1.54
	Overall (n=170)	1.69	1.74	1.65	**1.54**	1.60

The table presents the median value of the scoring metrics computed across the test stations for each area type, along with the overall performance. The best results are highlighted in bold, and n represents the sample size for each area.

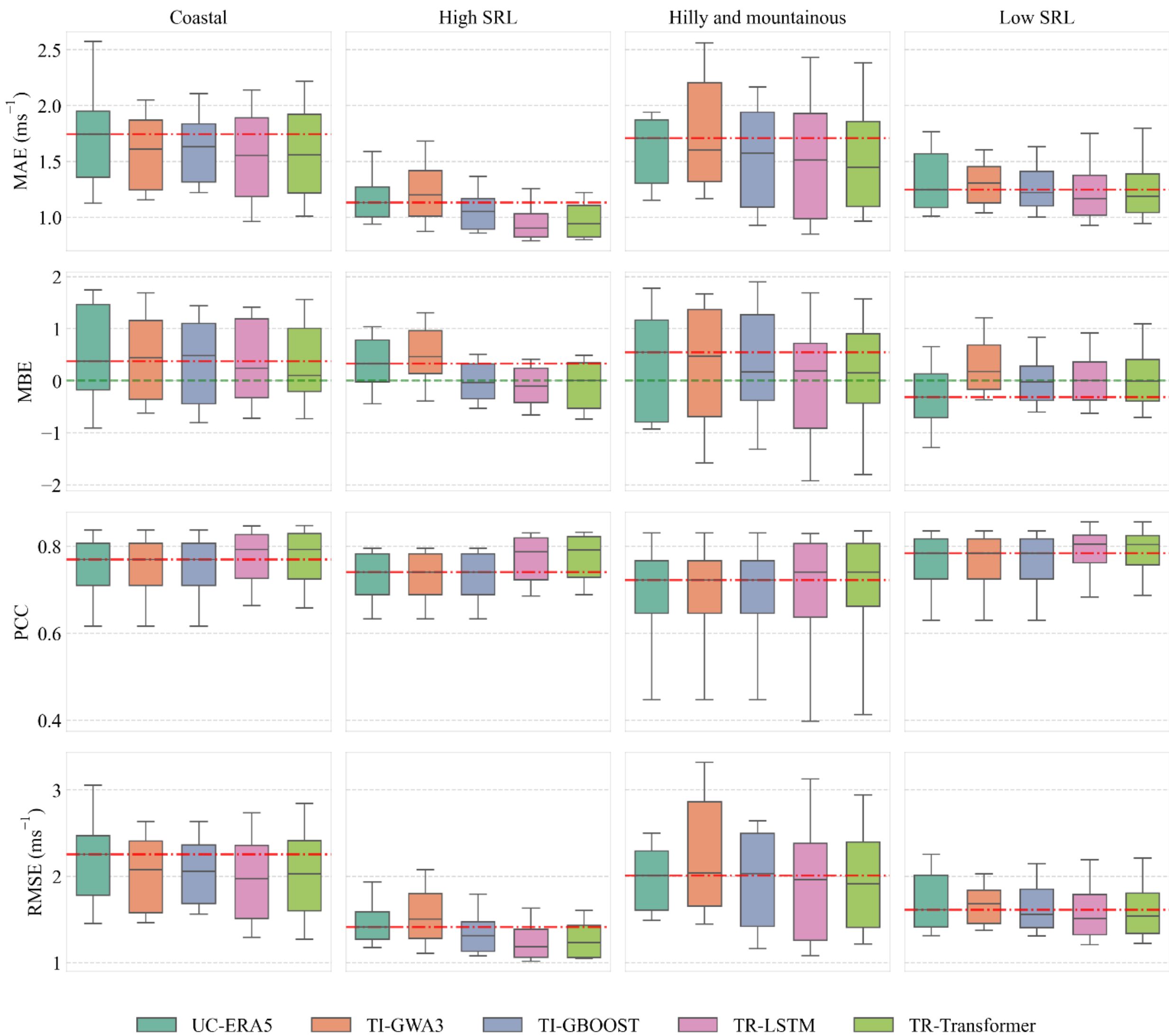


Figure 6.6 Boxplots of the time series metrics.

The boxplots are displayed separately for different types of environments and models. The boxes represent distribution quartiles, while the whiskers indicate the 10-90% range. The horizontal red dotted lines mark the median metrics of UC-ERA5, serving as a reference, while the horizontal green dotted line represents the zero MBE threshold.

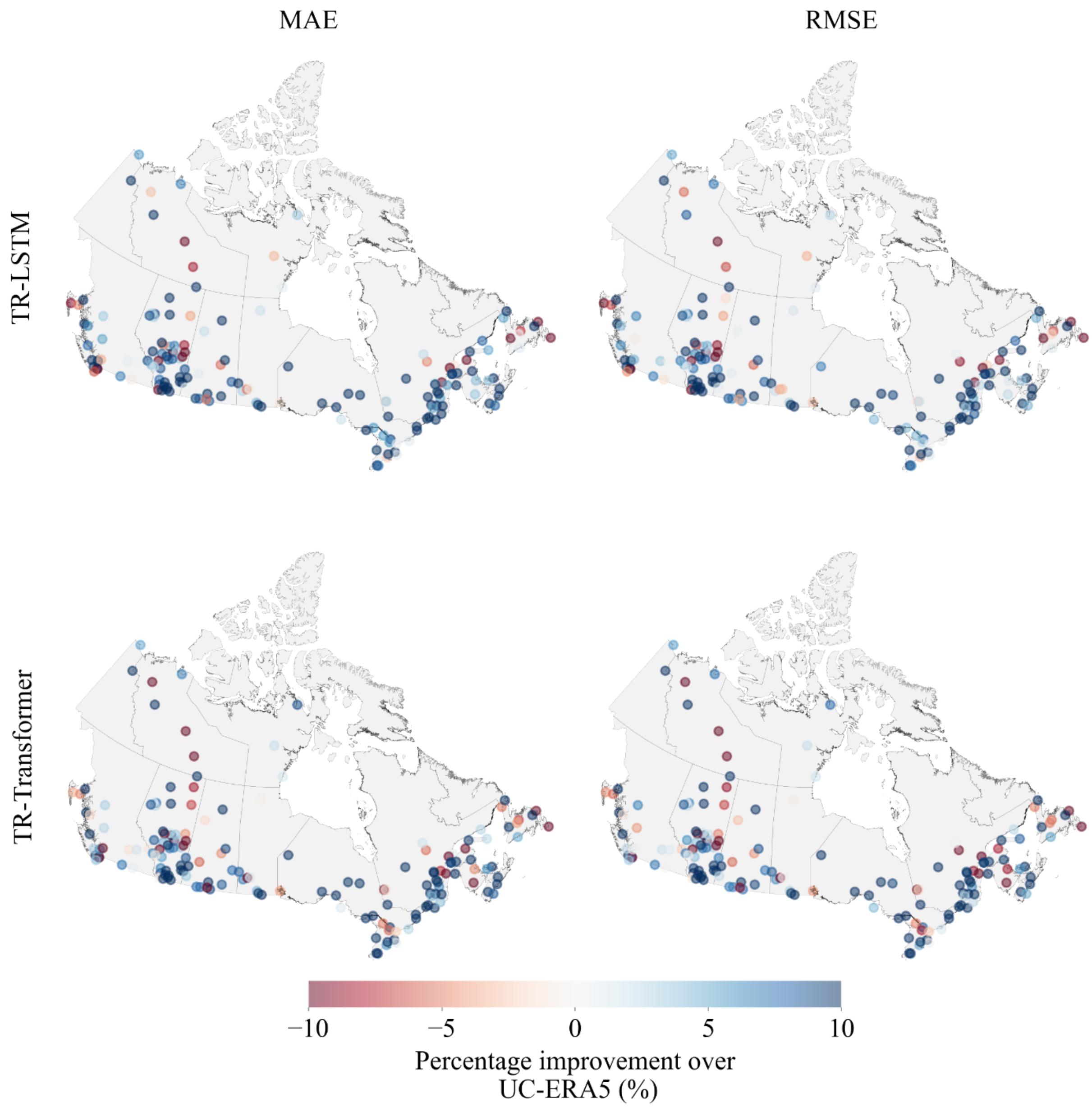


Figure 6.7 Maps illustrating the percentage improvement over UC-ERA5 (skill scores) at the test locations.

The leftmost column shows the MAE skill metric and the rightmost column shows the RMSE skill metrics. The top row displays the spatial distribution of the skill metrics for the TR-LSTM model, while the bottom row presents the same information for the TR-Transformer model. Red shades indicate a decline in the skill metric, while blue shades represent an improvement compared to UC-ERA5. No spatial patterns in the skill scores were observed, and the models were able to improve UC-ERA5 performance in sparsely sampled locations.

Figure 6.8 Spearman correlation between static covariates and percentage improvement over UC-ERA5 for time series MAE and RMSE.

The static covariates were the most important covariates identified using permutation feature importance. The abbreviations used for the static covariates are as follows. ELV-100m: Altitude estimated from a DEM with 100 m spatial resolution, DME-1km: deviation from mean elevation in a neighborhood of 1 km, ERA5-P5%: UC-ERA5 5th percentile, ERA5-P50%: UC-ERA5 50th percentile, ERA5-P95%: UC-ERA5 95th percentile; SDS-100m: standard deviation of slope in a neighborhood of 100 m, SDS-10km: standard deviation of slope in a neighborhood of 10 km, SDS-200m: standard deviation of slope in a neighborhood of 200 m, SRL-1km: dominant SRL in a neighborhood of 1 km, SRL-500m: dominant SRL in a neighborhood of 500 m, TACV-1km: tangential curvature in a neighborhood of 1 km.

6.5.3 Probability distribution evaluation

Figure 6.9 presents the distributions of PSS across the full range of wind speeds (top row, PSS-ALL), as well as for values in the lower (middle row, PSS-LWT) and upper tails (bottom row, PSS-UPT) across different area types (columns). Table 6.6 lists the median PSS metrics computed across all test locations.

Overall, the PSS-ALL values were higher than both PSS-LWT and PSS-UPT values. This indicates that all models, including UC-ERA5, were more effective at representing the frequency of typical wind speeds rather than extreme values. The spatial variability of PSS-ALL as shown by the interquartile range in Figure 6.9, was particularly pronounced in hilly and mountainous regions compared to other areas. Overall, the TR-LSTM model emerged as the top performer for PSS-ALL, followed by the TI-GBOOST model.

For the frequency of low wind speed values (PSS-LWT), none of the models showed an improvement over UC-ERA5 performance in regions with low SRL. In regions with high SRL and hilly and mountainous areas, the time-resolved models significantly improved UC-ERA5 PSS-LWT. In coastal areas, the TI-GBOOST model was the best-performing model, followed closely by the TR-Transformer model. For the frequency of high wind speed values (PSS-UPT), most models were ineffective in regions with high SRL, but the TR-LSTM model emerged as the best-performing model overall, followed closely by TI-GWA3.

The evaluation of wind speed quantiles across different percentile levels is presented in Figure 6.10. TI-GWA3 demonstrated the highest accuracy at the upper percentile levels (80% and 90%). Meanwhile, ML-based methods performed best for quantiles in the central part of the distribution (40%, 50%, and 60% percentile levels), with TR-LSTM and TR-Transformer exhibiting the least bias, as indicated by an MBE close to zero. Overall, the BC methods tended to overestimate lower quantiles while underestimating higher quantiles.

Table 6.6 Summary of the Perkins' skill score (PSS).

PSS	Area	UC-ERA5	TI-GWA3	TI-GBOOST	TR-LSTM	TR-Transformer
PSS-ALL	Coastal (n=45)	81.90	82.38	**83.32**	81.38	81.57
	High SRL (n=36)	78.42	77.71	79.74	**83.25**	80.86
	Hilly and mountainous (n=29)	**76.69**	73.30	73.26	75.77	75.34
	Low SRL (n=60)	80.54	**82.99**	82.73	82.85	81.90
	Overall (n=170)	79.75	80.42	81.66	**82.47**	80.93
PSS-LWT	Coastal (n=45)	32.16	27.80	**35.06**	29.88	32.82
	High SRL (n=36)	19.05	22.30	29.66	**43.78**	37.38
	Hilly and mountainous (n=29)	29.22	27.61	25.58	**43.72**	39.89
	Low SRL (n=60)	**42.04**	29.43	32.89	30.27	31.21
	Overall (n=170)	33.34	27.56	32.80	**35.15**	33.87
PSS-UPT	Coastal (n=45)	55.67	54.08	50.31	**55.87**	53.63
	High SRL (n=36)	**61.26**	55.61	40.93	50.61	54.50
	Hilly and mountainous (n=29)	40.10	39.07	44.90	**53.27**	52.09
	Low SRL (n=60)	36.02	**59.58**	50.19	56.31	54.41
	Overall (n=170)	44.48	54.26	48.10	**54.82**	53.82

The table presents the median value of the PSS metrics computed across the test stations for each area type, along with the overall performance. The best results are highlighted in bold, and n represents the sample size for each area.

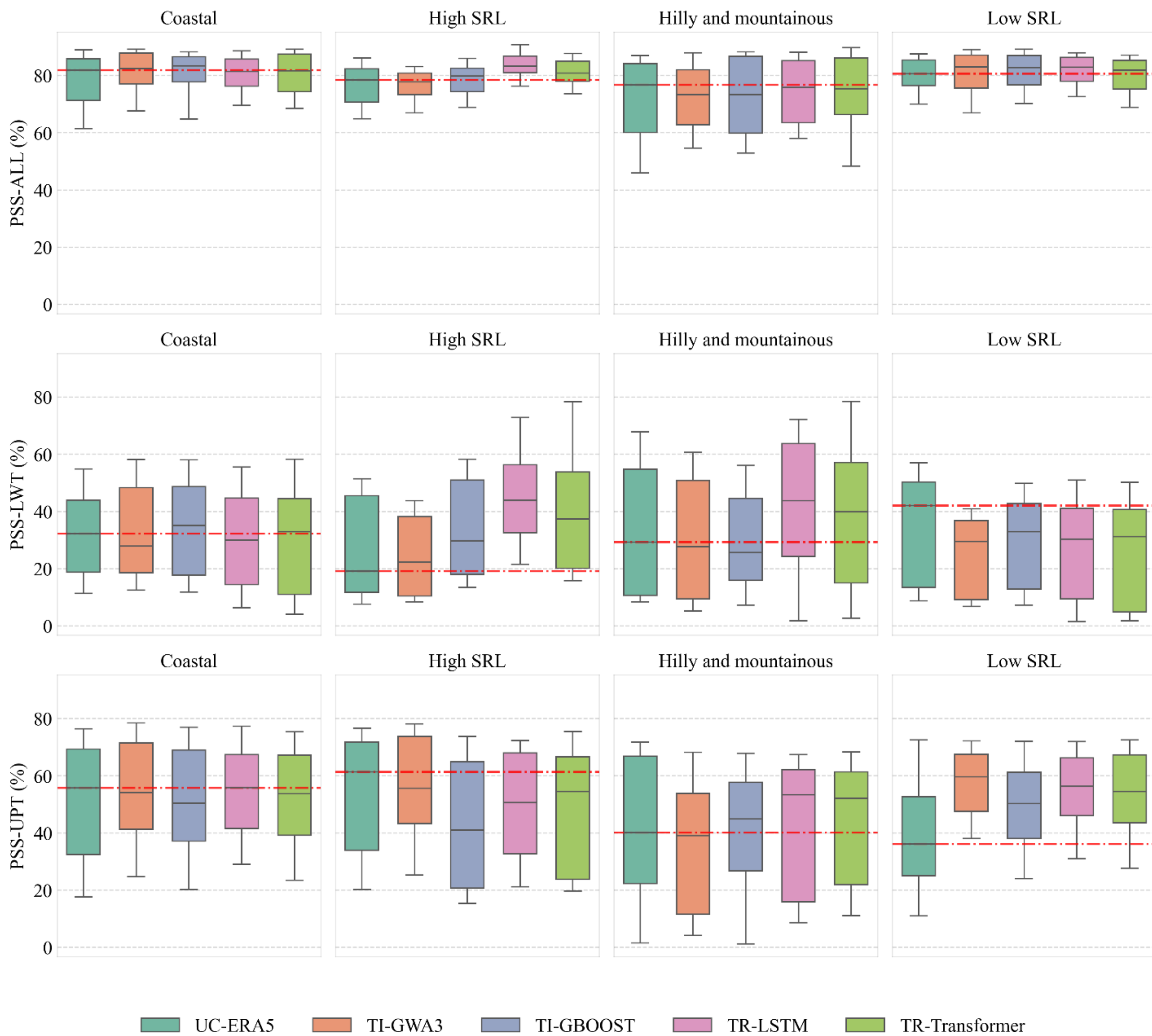


Figure 6.9 Boxplots of Perkins' skill scores (PSS).

The PSS metric was computed across the full range of wind speeds (PSS-ALL) for values below the 10th percentile (PSS-LWT) and above the 90th percentile (PSS-UPT). The boxplots are displayed separately for different types of environments and models. The boxes represent distribution quartiles, while the whiskers indicate the 10-90% range. The horizontal red dotted line marks the median metrics of UC-ERA5, serving as a reference.

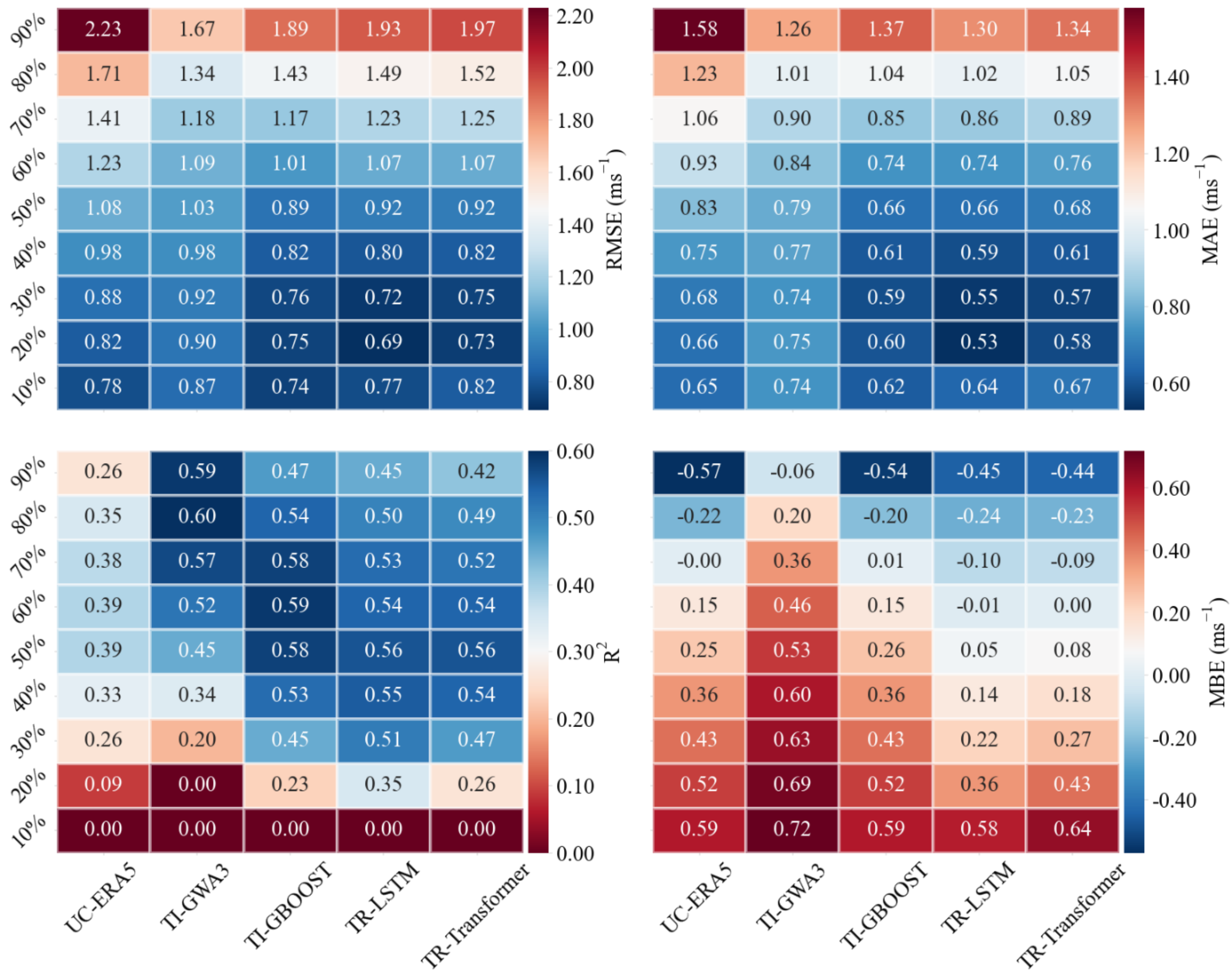


Figure 6.10 . Wind speed quantiles evaluation.

The panels display RMSE (top-left), MAE (top-right), R^2 (bottom-left), and MBE (bottom-right) between predicted and observed wind speed quantiles. The quantiles were computed at intervals between the 10th and the 90th percentiles, with 10% increments (columns).

6.6 Discussion

Modern reanalysis products like ERA5 are increasingly used in large-scale WRA studies. This work introduced a DL framework to correct systematic bias and temporal variability in ERA5 hourly wind speeds. The proposed models are applied to learn a time-resolved scaling factor between reanalysis and observed wind speeds.

6.6.1 Limitations of time-invariant methods

The evaluation of UC-ERA5 performance in the study region revealed that scale-dependent metrics, such as MAE and RMSE, for median wind speed and time series, were notably higher in coastal, hilly, and mountainous areas than in other regions. UC-ERA5 struggled to explain median wind speed spatial variability (as measured by R^2) across coastal areas and showed poor performance in terms of temporal variability (as measured by PCC) in hilly and mountainous regions. Additionally, in hilly and mountainous regions, PSS-ALL for UC-ERA5 showed significant spatial variability (as measured by the interquartile range) and a lower median than in other regions. These results highlight the challenges of ERA5 in accurately representing wind speed in regions with predominant complex terrain features and localized atmospheric dynamics. They align with previous studies' findings (e.g., Gualtieri (2022) for a review).

The evaluation of the time-invariant scaling factor estimated through the GWA (TI-GWA3) and an alternative time-invariant scaling factor estimated using ML (TI-GBOOST) revealed that, although these methods can improve the reanalysis median wind speed, they perform relatively poorly on time series metrics. TI-GWA3 and TI-GBOOST rely on time-invariant scaling factors, which do not account for diurnal and seasonal variation in the reanalysis biases.

To address this limitation, Schicker *et al.* (2023) estimated an hourly scaling factor between reanalysis-derived wind speeds and the GWA. Their findings indicated that this approach did not improve the performance of the reanalysis-derived wind speeds and, in some cases, even degraded it. While this method considers the diurnal variability of reanalysis mean wind speed, it still depends on a static mean wind speed from the GWA to compute the hourly scaling factor.

An extension of TI-GBOOST could be explored to account for diurnal biases by incorporating the hour of the day as an additional covariate, with the hourly observed scaling factor as the target variable. This method provides a balanced solution between fully time-resolved models, such as TR-LSTM and TR-Transformer, and the complete time-invariant methods. It can be beneficial when computational resources or data availability are limited while allowing for moderate improvements in capturing temporal variability.

6.6.2 Effective and scalable time-resolved bias correction framework

The time-resolved models (TR-LSTM and TR-Transformer) demonstrated strong performance in enhancing median wind speed and temporal variability of ERA5 across various regions. TR-LSTM and TR-Transformer showed significant improvements in coastal areas, explaining 46-47% of the spatial variability in median wind speed, compared to just 11% achieved by UC-ERA5. Additionally, TR-LSTM and TR-Transformer significantly reduced MAE and RMSE of median wind speed in these areas by more than 20% relative to UC-ERA5. In hilly and mountainous areas, the time-resolved models marginally increased UC-ERA5's median PCC, demonstrating their ability to address some of the challenges posed by complex terrain. The time-resolved models were particularly effective in regions with high SRL, such as forests and urban areas, where TR-Transformer achieved a median PCC of 0.79 (compared to just 0.74 for UC-ERA5) and an average reduction of 12% in MAE and 10% in RMSE relative to UC-ERA5. These findings indicate that UC-ERA5's coarse spatial resolution limits its ability to represent temporal variability in regions with high SRL. Incorporating small-scale SRL, topographic information, and reanalysis-derived past weather conditions into the time-resolved models mitigated this limitation, improving the representation of the temporal variability of wind speed in these regions.

Accurately estimating the wind speed probability distribution is essential for WRA. Results from the evaluation of PDFs across the full range of wind speeds indicate that the ML-based models were overall more effective than TI-GWA3 and UC-ERA5. Both time-resolved models significantly improved PSS-ALL in high SRL regions at most test stations. However, the time-resolved models did not consistently outperform UC-ERA5 for extreme wind speeds across all areas and exhibited a trade-off between improvements in low and high-wind speed PDFs. There were few locations (between 14.82% and 18.82% of test stations) where the time-resolved models improved low- and high-wind speed PDFs.

There is still potential to improve the performance of the time-resolved models, as their implementation has resulted in a decline in UC-ERA5 scoring metrics in some cases. A spatial analysis of model performances revealed no consistent trends that could explain this deterioration. However, SRL was strongly correlated with how well the time-resolved models performed, while the topographic covariates exhibited a weaker correlation. The direction of the correlation suggests that time-resolved models are likely to improve UC-ERA5's performance in areas with high SRL.

Analyzing the nonlinear interactions between terrain complexity, SRL, and proximity to the coast could improve the understanding of the model's performance. It would ensure that the models are applied in scenarios where improvements are likely while avoiding their use in cases where the UC-ERA5 may provide more reliable results. Additionally, increasing data availability —especially with broader spatial coverage —could improve the model's performance across all regions. For example, hilly and mountainous areas are underrepresented in the training dataset compared to coastal areas and region with high or low SRL. This imbalance may have contributed to the model's relative underperformance in hilly and mountainous areas. Addressing this data imbalance could improve model robustness and generalizability in complex terrains.

While the proposed deep learning (DL) framework, particularly the TR-Transformer architecture, exhibits higher computational demands than the time-invariant methods, its implementation remains feasible on standard personal computing hardware. The TR-Transformer, which has approximately 3 million parameters, required around 30 minutes to complete one training epoch when trained on an Nvidia T600 GPU (with 4GB of VRAM) using a dataset of about 17 million samples and a batch size of 128. In contrast, the TR-LSTM, which has around 600,000 parameters, finished a training epoch in just 6 minutes under the same conditions. While the computational costs for inference are significantly lower than during training, the scalability of both training and inference times indicates that the approach can be applied to larger datasets. In situations where resource constraints are a concern, the TR-LSTM provides a more computationally efficient alternative.

6.6.3 Limitations and future directions

The proposed framework has certain limitations that could be addressed in future studies. Additionally, this section highlights several research directions.

Various topographic covariates were identified from the literature and extracted from a DEM. To account for the scale dependency of these covariates, they were extracted at multiple spatial scales, and the most relevant covariates were identified using the permutation feature importance algorithm. This cumbersome approach may not reliably select the most appropriate covariates and associated spatial scales, especially in hilly and mountainous regions with sharp variations in the topography. Future studies could explore CNNs for automatically extracting topographic covariates from high-resolution DEMs, enabling end-to-end training, and extracting more effective topographic features that could significantly enhance model performance, particularly in complex terrains. However, it is essential to recognize that this approach requires increased computational resources and larger datasets to optimize the CNN trainable parameters.

In addition, other machine learning models, such as XGBOOST (Chen *et al.*, 2016) and LightGBM (Ke *et al.*, 2017), as well as deep learning architectures like Temporal Fusion Transformer (TFT; Lim *et al.* (2021a)) and Mamba (Wang *et al.*, 2025), could be explored to improve the performance of the proposed framework.

The evaluation of the framework indicates that there is still room for performance improvement, particularly for extreme wind speeds. Due to their limited representation in the training set, extreme values may be neglected as the model prioritizes more frequently occurring values (Schultz *et al.*, 2021; Wilson *et al.*, 2022). It is a well-known issue when applying deep learning models for downscaling and forecasting environmental variables (Dujardin *et al.*, 2022; Reichstein *et al.*, 2019). Prior studies have proposed modifications to the loss function to mitigate this issue

(Dujardin *et al.*, 2022; Jung *et al.*, 2013; Kristianti *et al.*, 2023); however, there is no consensus on the optimal approach. Future research should focus on testing various loss functions and exploring new loss functions and techniques, such as data augmentation and sample balancing strategies, that may enhance model performance for extreme wind speeds. Additionally, incorporating probabilistic methods, such as quantile regression (Olivier *et al.*, 2024), could help capture the uncertainty associated with extreme wind speed predictions. Future studies could explore using extreme value theory (EVT)-based methods (Wilson *et al.*, 2022) to better characterize the tails of the wind speed distribution, ensuring that rare but critical events are well estimated.

Developing BC models, such as the time-resolved models in this work, requires a robust network of meteorological stations in the study region that provide high-quality wind speed records. This network is essential for learning the relationship between reanalysis-derived wind speeds and ground-truth observations, and in regions where such a network is unavailable, employing transfer learning from data-rich areas could serve as a practical alternative. This approach would enable the application of models developed in well-monitored regions to areas with limited ground-truth data. However, it remains crucial to conduct careful validation to ensure that the model generalizes well in the new environment, as variations in SRL, terrain complexity, and meteorological conditions could significantly influence generalization performance.

Future studies could explore the application of the proposed framework to other reanalysis datasets, such as the Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) and to improve solar irradiance data from reanalysis datasets. Since reanalysis-derived solar radiation often suffer from biases due to cloud cover misrepresentation, implementing a BC framework could enhance the assessment of solar energy potential (Frank *et al.*, 2018). Improving the temporal variability of wind speed and solar irradiance can allow a more accurate assessment of the co-variability between these two renewable energy sources, which is crucial for hybrid energy system design (Pedruzzi *et al.*, 2023). Future studies could explore adapting the framework to additional meteorological variables, such as temperature and precipitation, which are also known to exhibit biases in reanalysis datasets (Gleixner *et al.*, 2020), particularly in regions with complex terrain (Cavalleri *et al.*, 2024; Chen *et al.*, 2021b). Enhancing temperature and precipitation data accuracy could have significant implications for climate modeling, hydrological studies, and disaster risk assessments.

Operational forecasts are crucial for the wind energy sector (Liu *et al.*, 2022). Future studies could investigate the application of the proposed framework to bridge the gap between NWP forecasts and observations (Sweeney *et al.*, 2020). This integration can enhance the accuracy of wind power forecasting for grid stability and energy market operations.

Due to computational constraints, this study performed model parameter optimization using best practices and manual tuning. Future research could explore the implementation of more advanced hyperparameter optimization techniques, such as Bayesian optimization or genetic algorithms, in operational settings. These methods could systematically refine model parameters, improving predictive performance.

Although the proposed approach was validated across a large region (Canada) with diverse wind patterns and topography complexity, further research is still encouraged to assess its generalizability performance across various climatic conditions and geographical settings. Expanding validation efforts to include additional regions would provide a more comprehensive assessment of the framework's robustness and adaptability. The framework is built upon globally accessible datasets like ERA5 and ALOS DEM; its application in other regions is feasible and could be explored in future studies.

Furthermore, the study only considered inland regions, where reanalysis wind speeds are more prone to bias due to the topography and land surface heterogeneity. However, future studies could explore the effectiveness of the BC framework in offshore regions influenced by seabed topography, straits, and islands. Unique factors in these regions may introduce different biases compared to inland regions.

6.7 Conclusion

Wind speeds derived from modern reanalysis datasets have become central to large-scale WRA studies due to their global coverage, extensive record length, and consistent records. However, the coarse spatial resolution of these datasets and limitations in data assimilation techniques and NWP models lead to varying accuracy across different locations (e.g., onshore, offshore, and coastal) and terrain complexities (e.g., flat, hilly, and mountainous). Therefore, developing and refining bias correction techniques for reanalysis-derived wind speeds are essential to ensure their accuracy across a broader range of environments, enabling more reliable resource assessments in diverse geographic and topographic settings.

A novel bias correction method using DL models for sequence modeling was proposed and evaluated across Canada. This method advances the commonly used mean BC technique, based on the GWA, by enhancing median wind speed and temporal variability accuracy across diverse environments. It leverages optimized DL architectures, making it a scalable option for large-scale applications.

Results from comprehensive evaluations indicate that this framework significantly improves ERA5 wind speeds at most test stations, performing particularly well in high SRL areas such as forest and urban regions. Notable improvements were observed in coastal, hilly, and mountainous areas, where reanalysis datasets typically perform poorly. Several potential enhancements to the DL framework have been proposed for future studies. Overall, this framework represents a promising advancement in improving the accuracy of reanalysis-derived wind data for WRA studies, helping to reduce uncertainties in predicting energy yields. In addition, the proposed framework can enhance the downscaling of Global Climate Models (GCMs). This improvement is important for understanding future wind patterns and developing effective climate adaptation strategies. Furthermore, the framework's ability to improve wind speed estimates could be beneficial for conducting risk assessments related to extreme wind events.

.

Nomenclature

Abbreviations

a.g.l	Above ground level
ALOS	Advanced land observing satellite
BC	Bias correction
CNN	Convolutional Neural Network
DEM	Digital Elevation Model
DL	Deep learning
ECCC	Environment and Climate Change Canada
ECMWF	European Centre for Medium-Range Weather Forecasts
ERA5	ECMWF reanalysis version 5
FFN	Feed-forward neural network
GB	Gradient boosting
GBOOST	Gradient boosting regressor from the Scikit-learn package
GCM	Global Climate Model
GELU	Gaussian error linear unit
GRU	Gated recurrent unit
GWA	Global Wind Atlas
GWA3	GWA version 3
hpa	Hectopascal
LSTM	Long short-term memory
MAE	Mean absolute error
MHA	Multi-head self-attention
ML	Machine learning
MSE	Mean square error
NWP	Numerical Weather Prediction
PCC	Pearson correlation coefficient
PDF	Probability density function
PSS	Perkins skill score
PSS-ALL	PSS computed for the full range of wind speed values
PSS-LWT	PSS computed for wind speed values below the 10th percentile
PSS-UPT	PSS computed for wind speed values above the 90th percentile
R^2	Coefficient of determination

RMSE	Root mean squared error
RNN	Recurrent neural network
SRL	Surface roughness length
TI-GBOOST	Time-invariant BC with GBOOST
TI-GWA3	Time-invariant BC with GWA3
TR-LSTM	Time-resolved BC with LSTM
TR-Transformer	Time-resolved BC with Transformer
UC-ERA5	Uncorrected ERA5 wind speed data
WRA	Wind resource assessment
XGBOOST	eXtreme gradient boosting algorithm

Symbols

α	Parameter that controls the transition between the quadratic and linear behavior of the Huber loss function
b	Bias vector
β_1, β_2	Adam optimizer parameters
$\tilde{c}_t$	LSTM's cell gate
c_t	LSTM's cell state
$f(\cdot)$	Regression function
$\mathcal{F}(\cdot)$	Sub-layers of the Transformer encoder
f_t	LSTM's forget gate
$\odot$	Element-wise product
H_i	Attention head
i_t	LSTM's input gate
$k(s_i)$	Static covariates estimated at location s_i
K_i	Keys matrix associated with the attention head H_i
$L_\alpha(\cdot,\cdot)$	Huber loss function
$LN(\cdot)$	Layer-normalisation
LP	Linear projection layer
o_t	LSTM's output gate
p	Length of the finite look-back window
χ, ζ_t', ζ_t	Hidden vectors of the deep learning model
Q_i	Queries matrix associated with the attention head H_i

$\sigma(\cdot)$	Sigmoid activation function
SM_{ERA5}	Scoring metrics value of ERA5 data
SM_{model}	Scoring metric value of the model predictions
SM_{perf}	Scoring metric value for a perfect prediction
$\mathrm{softmax}(\cdot)$	Softmax activation function
$Softplus(\cdot)$	Softplus activation function
$\tanh(\cdot)$	Hyperbolic tangent activation functions
τ_t	Timestamp vector
$TEmbed(\cdot)$	Temporal embedding layer
$TModel(\cdot)$	Deep learning temporal model (LSTM or Transformer)
u	Zonal component of wind speed
$\overline{U}^{ERA5}$	Mean wind speed from ERA5
$U_t^{ERA5}(s_i)$	ERA5 wind speeds at time t and location s_i
$\overline{U}^{GWA}$	Mean wind speed from the Global Wind Atlas
v	Meridional component of wind speed
V_i	Values matrix associated with the attention head H_i
$v_t(s_i)$	Reanalysis meteorological covariates at time t, and interpolated at location s_i
W	Weights matrix
$\hat{y}(s_i)$	Estimated static scaling factor at location s_i
$y_t(s_i)$	Scaling factor at time t and location s_i

7 DISCUSSION GENERALE ET CONCLUSION

Ce dernier chapitre, qui conclut cette thèse, se concentre sur la synthèse des principaux résultats présentés dans les articles, suivie par la présentation des limites de l'étude et des pistes pour de futurs travaux de recherche.

7.1 Synthèse

La croissance de la part des énergies éoliennes dans les systèmes énergétiques actuels requiert la disponibilité de données de vitesse du vent de haute qualité pour l'évaluation de la variabilité temporelle et spatiale de la ressource (McKenna *et al.*, 2022). Cette thèse vise à proposer de nouvelles méthodes d'estimation des vitesses du vent aux sites non échantillonnés basées sur l'apprentissage automatique.

La revue de la littérature (Article 1) a mis en évidence un intérêt croissant pour le développement d'approches d'estimation de la distribution de probabilité complète et de reconstruction de séries temporelles de vitesse du vent aux sites non échantillonnés. Cette tendance se justifie par la nécessité de tenir davantage compte de la variabilité des vitesses du vent qui reste un défi majeur à l'évaluation précise de la ressource éolienne (Pelser *et al.*, 2024). De plus, les défis additionnels introduits par la non-stationnarité des vitesses du vent lors de l'estimation aux sites non échantillonnés ont été soulevés dans cette revue. Par exemple, lors de la mise à l'échelle des sorties des modèles climatiques globaux, les données de réanalyse sont souvent utilisées comme référence aux sites non échantillonnés (Jung *et al.*, 2022c). On observe un manque de validation de ces données dans ces études, ce qui peut compromettre la validité des résultats. Par exemple, il existe des écarts entre les variabilités temporelles présentes dans les données de réanalyses et les observations in situ (Ramon *et al.*, 2019). Les conclusions de la revue de littérature ont démontré la nécessité de poursuivre la recherche dans le développement de méthodes plus précises et plus flexibles pour l'estimation des vitesses du vent aux sites non échantillonnés.

La littérature ne présente pas de consensus quant aux variables explicatives les plus pertinentes pour la modélisation empirique des vitesses de vent. Les variables utilisées ainsi que les méthodes de sélection varient d'un auteur à l'autre (voir, par exemple, Etienne *et al.* (2010) ; Jung (2016)). Dans l'article 2, on a effectué une étude comparative des principales variables explicatives identifiées dans la littérature. Cette étude a été accompagnée d'une analyse des performances de six méthodes de sélection de variables. Les résultats de cette étude ont révélé que l'influence des variables explicatives varie selon les différentes plages de vitesse du vent.

Par exemple, pour les vitesses élevées, la convexité du terrain a un impact plus important que pour les vitesses plus faibles. En revanche, la distance par rapport à la côte maritime et la longueur de rugosité ont un impact significatif sur l'ensemble des plages de vitesses de vent. Les algorithmes LASSO et MRMR ont été identifiés comme les techniques de sélection de variables les plus performantes en termes de précision des prédictions, de parcimonie et de réduction de la multicolinéarité. De plus, ces deux méthodes présentent l'avantage d'être assez simples à mettre en œuvre, puisqu'elles nécessitent uniquement l'optimisation d'un seul paramètre.

Les méthodes d'estimation de la distribution complète des vitesses du vent sont plus attrayantes que celles axées sur la prédiction de statistiques récapitulatives, telles que la moyenne ou un quantile spécifique. La distribution complète permet une analyse plus approfondie de la variabilité de la ressource, un facteur crucial pour une meilleure gestion des risques liés aux fluctuations à long et à court terme, qui peuvent affecter la rentabilité des projets.

Dans ce contexte, les approches existantes reposent généralement sur l'hypothèse restrictive selon laquelle une seule loi de probabilité s'appliquerait à l'ensemble de la région d'étude (voir, par exemple, Veronesi *et al.* (2016) ; Jung *et al.* (2023b)). Le développement d'une approche non paramétrique d'estimation de la distribution du vent aux sites non échantillonnés (article 3) a permis de surmonter cette limitation. Cette nouvelle approche offre une plus grande flexibilité pour représenter la variabilité spatiale des régimes de vent en s'ajustant aux particularités des conditions climatiques et topographiques locales. Cette approche s'appuie sur l'apprentissage automatique pour l'interpolation spatiale de plusieurs quantiles de la vitesse du vent, qui sont ensuite utilisés pour reconstruire la distribution de probabilité complète à l'aide de méthodes à noyau asymétrique. L'utilisation de l'apprentissage automatique, notamment des modèles de régression comme XGBoost, améliore la capacité de l'approche à gérer des relations non linéaires complexes entre les variables explicatives et les quantiles de vitesse du vent.

L'analyse comparative a révélé que la méthode non paramétrique fondée sur les noyaux asymétriques est supérieure aux approches paramétriques, selon le critère d'ajustement de Kolmogorov-Smirnov. En outre, l'approche non paramétrique est plus simple à implémenter, car elle ne nécessite pas un processus complexe de sélection d'une distribution régionale, qui impliquerait l'évaluation de plusieurs lois de distribution selon différents critères d'ajustement (voir, par exemple, Jung (2016)). Les noyaux asymétriques Birnbaum-Saunders et Log-Normale ont donné des résultats similaires et se sont avérés plus efficaces que le noyau asymétrique de Weibull. Cette nouvelle approche est particulièrement recommandée pour les régions présentant une forte variabilité spatiale du régime des vents.

Les approches de reconstruction des séries temporelles de vitesse du vent aux sites non échantillonnés reposent sur l'interpolation de données de réanalyse corrigées ou non corrigées (Gualtieri, 2022; Niermann *et al.*, 2019). Les limites des données de réanalyse non corrigées pour l'évaluation du potentiel éolien ont été soulignées dans plusieurs études (Davidson *et al.*, 2022; Staffell *et al.*, 2016). Il parait donc pertinent de recourir à des approches de correction de biais de ces données pour améliorer leur précision (Langer *et al.*, 2023).

Les méthodes de correction actuelles s'appuient soit sur le GWA (Bosch *et al.*, 2018; Gruber *et al.*, 2022), soit sur des modèles de régression (Hu *et al.*, 2023; Jung *et al.*, 2020). Ces modèles de régression, qui s'appuient sur l'apprentissage automatique, utilisent comme entrée les vitesses de vent issues des données de réanalyse, ainsi que des variables explicatives liées à la topographie et à la longueur de rugosité pour prédire les vitesses de vent observées. Dans l'article 4, plusieurs des méthodes proposées dans la littérature ont été rigoureusement comparées selon plusieurs critères essentiels pour l'évaluation du potentiel éolien, notamment la distribution de probabilité, la variabilité temporelle et le biais systématique.

Aucune des méthodes examinées ne s'est démarquée de manière uniforme en fonction de tous les critères d'évaluation. Par exemple, l'évaluation de la distribution de probabilité a révélé une préférence pour la méthode quantile-quantile appliquée aux séries temporelles de probabilités de non-dépassement issues des données de réanalyse. Toutefois, cette même méthode s'est avérée moins efficace lorsqu'elle a été appliquée à des séries temporelles de probabilités de non-dépassement interpolées à partir d'observations de sites voisins. L'analyse de la variabilité temporelle a révélé une préférence pour une méthode d'interpolation spatiale basée sur l'apprentissage automatique, qui combine les données de réanalyses et les observations mesurées sur des sites voisins. La combinaison de différentes approches semble donc être une stratégie prometteuse pour améliorer la précision de la reconstruction des séries temporelles de vitesse du vent. Effectivement, l'intégration des méthodes de type « quantile-quantile » et des techniques d'interpolation fondées sur l'apprentissage automatique permettrait de compenser les limites de chacune.

Comme cela a été mentionné précédemment, de nombreuses études ont utilisé les données du GWA pour corriger les biais dans les données de réanalyse. Bien que cette approche contribue à réduire les biais systématiques, elle ne permet pas de modifier ni d'améliorer la représentation de la variabilité temporelle des vitesses du vent (Bosch *et al.*, 2017). En effet, le GWA fournit des statistiques climatiques à long terme, telles que la vitesse moyenne entre 2008 et 2017, qui ne capturent pas les fluctuations temporelles à des échelles plus résolues (p. ex., diurne et saisonnière).

L'article 5 propose un cadre méthodologique intégrant des architectures LSTM et Transformer pour corriger les données de réanalyses à partir d'un facteur de correction dynamique. L'examen de cette méthode a révélé une amélioration significative par rapport aux méthodes fondées sur un facteur de correction invariant dans le temps. En outre, dans les zones côtières et les régions à relief marqué, où les performances des données de réanalyses sont souvent dégradées (Gualtieri, 2021), l'approche proposée permet d'améliorer sensiblement la qualité des estimations. Une amélioration notable de la représentation de la distribution de probabilité a été observée, en particulier dans la partie centrale de la distribution. Cela reflète une meilleure adéquation avec les observations in situ. Les architectures LSTM et Transformer se sont avérées efficaces pour capturer la dynamique temporelle complexe des vitesses du vent, grâce à leur capacité à modéliser des dépendances à long terme dans les séries temporelles issues des données de réanalyse. Ces résultats soulignent le potentiel des approches basées sur l'apprentissage profond pour surmonter les limitations des méthodes de correction de biais traditionnelles.

Enfin, les différents modèles développés dans le cadre de cette thèse contribuent à une meilleure compréhension des processus qui gouvernent la dynamique du vent. Ils permettent notamment une meilleure compréhension des interactions entre les facteurs météorologiques et topographiques, d'identifier les principaux déterminants de la variabilité spatio-temporelle du vent, et d'améliorer les capacités de prédiction dans des zones non échantillonnées.

7.2 Limites et perspectives

La non-stationnarité des séries temporelles constitue un défi majeur pour les modèles empiriques (Manuca et al., 1996). En ce qui concerne la vitesse du vent, la non-stationnarité des séries peut être attribuée à divers facteurs, tels que la variabilité interannuelle liée aux oscillations climatiques (Pryor et al., 2020; Zeng et al., 2019), les tendances à la baisse parfois causées par l'accroissement de la rugosité de la surface (Vautard *et al.*, 2010a), ainsi que les effets du changement climatique (Martinez et al., 2024).

L'incorporation de signaux de non-stationnarité dans les divers modèles élaborés dans cette thèse représente un axe de recherche prometteur pour accroître la fiabilité des estimations de la ressource éolienne. Par exemple, on peut intégrer aux modèles de régression utilisés pour l'interpolation spatiale des quantiles de vitesses du vent des variables liées aux oscillations climatiques, comme l'oscillation nord-atlantique (NAO). Les quantiles conditionnels estimés à partir de ces modèles tiendront compte non seulement des caractéristiques topographiques locales, mais aussi des variations interannuelles attribuables aux oscillations climatiques.

En outre, la rugosité de surface, telle qu'estimée dans nos travaux, est considérée comme étant constante durant la période d'analyse. Cette hypothèse peut limiter la précision des estimations, car la rugosité de surface peut évoluer au fil du temps en raison de changements dans l'utilisation des sols, de l'expansion urbaine, de la croissance de la végétation ou de la déforestation (Petersen *et al.*, 1998; Vautard *et al.*, 2010a). L'intégration de la rugosité de surface comme une variable dynamique dans les modèles pourrait ainsi améliorer la représentation des conditions locales de vent. Par exemple, l'utilisation de séries temporelles de données provenant de la télédétection, telles que Landsat (Zhu, 2017), permettrait de capturer les fluctuations temporelles de la rugosité de surface.

L'incertitude associée aux estimations des vitesses du vent est un sujet qui reste peu abordé dans la littérature (McKenna et al., 2022). Les méthodes proposées dans ces travaux peuvent être adaptées pour intégrer des mécanismes d'évaluation et de quantification de cette incertitude. Par exemple, les fonctions objectives utilisées dans les modèles d'apprentissage, tels que les LSTM et les Transformers, peuvent être adaptées pour obtenir des estimations probabilistes par l'intermédiaire de quantiles conditionnels ou de paramètres d'une loi de distribution (Lim *et al.*, 2021b). Ces approches fourniraient des informations additionnelles pour la gestion des risques dans le cadre du développement de projets éoliens. Effectivement, la prise en compte de l'incertitude permettrait aux décideurs d'évaluer non seulement les prévisions moyennes, mais aussi les marges d'erreur associées. L'intégration de l'incertitude dans les approches développées représente donc une piste prometteuse pour améliorer la fiabilité des modèles d'évaluation de la ressource éolienne.

Les données in situ des vitesses du vent utilisées dans cette thèse ont été collectées à 10 mètres au-dessus du sol. Toutefois, les éoliennes modernes fonctionnent à des hauteurs beaucoup plus élevées (p. ex., 80-150 mètres). La vitesse du vent variant selon l'altitude, il serait préférable d'évaluer directement la ressource éolienne à ces hauteurs. Toutefois, il y a un manque de séries temporelles des vitesses du vent à ces hauteurs (Ramon et al., 2020). Pour contourner cette limite, des méthodes d'extrapolation verticale, telles que la loi de puissance ou la loi logarithmique, sont couramment utilisées (Gruber et al., 2022). Ces approches sont fondées sur des hypothèses simples qui peuvent ne pas être valables dans certaines conditions (Gualtieri, 2019). Par conséquent, l'extrapolation verticale introduit inévitablement des incertitudes additionnelles dans l'estimation de la ressource éolienne. Des recherches futures devraient être menées pour évaluer et quantifier de manière plus approfondie cette source d'incertitude. De plus, des approches plus avancées d'extrapolation verticale, qui s'appuient sur l'apprentissage automatique, ont été proposées (p. ex., Yu et al. (2022); Vassallo et al. (2020)). Dans leur forme actuelle, ces approches plus complexes requièrent néanmoins la disponibilité de mesures de vitesse du vent à au moins deux hauteurs, ce qui est problématique à des sites non échantillonnés.

Le développement des méthodes d'estimation proposées dans nos travaux dépend de la disponibilité d'un réseau dense de stations de mesure dans la région d'étude. Ce réseau est essentiel pour formuler la relation entre les variables explicatives et les vitesses de vents à prédire. Toutefois, la répartition des stations de mesure à l'échelle mondiale est très inégale, et même au sein d'une même région, des disparités locales peuvent exister. Le développement de modèles transférables d'une région à l'autre représente donc une solution potentielle à ce défi. Ces modèles, développés dans des régions disposant d'un réseau dense de stations de mesure, pourront être appliqués à des zones où le réseau de stations de mesures est limité, voire inexistant. Des recherches futures devraient explorer l'optimisation de ces modèles transférables, en évaluant leur performance dans différents milieux, comme les zones côtières ou les régions montagneuses, où les conditions climatiques et topographiques sont particulièrement hétérogènes.

8 ANNEXE DE L'ARTICLE 3

Table 8.1 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

Table 8.2 Overview of the WS covariates

Predictor	Description	Spatial scale
Altitude	Altitude of the location in m.	
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 (Florinsky, 2017).	100m, 500m, 1000m, 1500m, 2000m
Geographical coordinates	Geographical coordinates of the location.	
geomorphologic phonotypes (geomorphons)	Landform element classification with the geomorphons-based method (Jasiewicz *et al.*, 2013).	
Laplacian of Gaussian	Derivative filter used to highlight location of rapid elevation change.	100m, 500m, 1000m, 1500m, 2000m
Maximal curvature	Measure of surface curvature (Wilson, 2018).	100m, 500m, 1000m, 1500m, 2000m
Mean curvature	Measure of surface curvature (Wilson, 2018).	100m, 500m, 1000m, 1500m, 2000m

minimal curvature	Measure of surface curvature (Florinsky, 2017).	100m, 500m, 1000m, 1500m, 2000m
Pennock landform class	Landform classification based on the slope and curvature of the grid cell (Pennock *et al.*, 1987).	
plan curvature	Measure of surface curvature (Florinsky, 2017).	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 (Jasiewicz *et al.*, 2013; Riley *et al.*, 1999)	100m, 500m, 1000m, 1500m, 2000m
Slope	Slope at the grid cell.	100m, 500m, 1000m, 1500m, 2000m
Standard deviation of slope	Measure of surface roughness (Grohmann *et al.*, 2011).	100m, 500m, 1000m, 1500m, 2000m
Surface area ratio	Measure of the surface roughness (Jenness, 2004).	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 (Florinsky, 2017).	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.	

9 INFORMATION SUPPLÉMENTAIRE POUR L'ARTICLE 4

Supporting material for "Prediction of hourly wind speed time series at unsampled locations using machine learning"

Contents

Table 9.1 Wind speed 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.	100m, 500m, 1000m, 1500m, 2000m
Geographical coordinates	Geographical coordinates of the location.	
geomorphologic phonotypes (geomorphons)	Landform element classification with the geomorphons-based method.	
Laplacian of Gaussian	Derivative filter used to highlight location of rapid elevation change.	100m, 500m, 1000m, 1500m, 2000m
Maximal curvature	Measure of surface curvature.	100m, 500m, 1000m, 1500m, 2000m
Mean curvature	Measure of surface curvature.	100m, 500m, 1000m, 1500m, 2000m
minimal curvature	Measure of surface curvature.	100m, 500m, 1000m, 1500m, 2000m

Pennock landform class	Landform classification based on the slope and curvature of the grid cell.	
plan curvature	Measure of surface curvature.	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	100m, 500m, 1000m, 1500m, 2000m
Slope	Slope at the grid cell.	100m, 500m, 1000m, 1500m, 2000m
Standard deviation of slope	Measure of surface roughness.	100m, 500m, 1000m, 1500m, 2000m
Surface area ratio	Measure of the surface roughness.	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.	100m, 500m, 1000m, 1500m, 2000m
Total curvature	Measure of surface curvature.	100m, 500m, 1000m, 1500m, 2000m

Evaluation metrics

The Pearson correlation coefficient was used to evaluate the correlation between observed and estimated time series:

Equation 9.1

$$PC(w_t, \hat{w}_t) = \frac{\sum_{i=1}^{n}(w_{t_i} - \bar{w})\sum_{i=1}^{n}(\hat{w}_{t_i} - \bar{\hat{w}})}{\sqrt{\sum_{i=1}^{n}(w_{t_i} - \bar{w})^2 \sum_{i=1}^{n}(\hat{w}_{t_i} - \bar{\hat{w}})^2}}$$

In addition, the following metric were used during evaluation:

Equation 9.2

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Equation 9.3

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

Equation 9.4

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

Equation 9.5

$$ME = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{y}_i$$

The OP (Equation 9.6) is used to evaluate the overlap between two probability distribution function (PDF). An OP value equal to 100% implies a perfect match between the PDFs and an OP close to zero implied a poor overlap between the PDFs.

Equation 9.6

$$OP = \sum_{i=1}^{B} \min\,(\widehat{Z_\iota}, Z_i)$$

Where B is the number of bins used to empirically estimate the PDF, $\widehat{Z_\iota}$ and Z_i are the frequency of the wind speed values in i^{th} bin from the estimated and observed wind speed data respectively. A bin width of 0.5 m/s was selected as done by Jung *et al.* (2023a) study.

To assess the IAV of annual median wind speed, the RCov was calculated as follows:

Equation 9.7

$$RCov = \frac{median(\left|X_i - \tilde{X}\right|)}{\tilde{X}}$$

where $\tilde{X} = median(X)$

Estimation of wind speed quantiles at defined percentile points

Quantiles at the fixe percentile point p were calculated from observed and estimated WSTS using the following general formula (Hyndman *et al.*, 1996):

Equation 9.8

$$w_p = (1 - \gamma)X_{(j)} + \gamma X_{(j+1)}$$

Where $X_{(j)}$ 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 $g = np + m - j$. γ was set equal to g and $\alpha = \beta = 1/3$ given quantiles that are approximately median-unbiased regardless of the WS true probability distribution (Reiss, 1989).

Table 9.2 Best parameters for LGBMQR and LGBMQR-ERA5 found during Random Search.

Model parameter	LGBMQR	LGBMAR-ERA5
learning_rate	0.03	0.04
max_depth	4	3
feature_fraction	0.4	0.5
bagging_fraction	0.3	0.8
extra_trees	False	True
lambda_l2	296.8	525
lambda_l1	292.8	469.3

num_leaves	162	22
max_bin	160	140
min_data_in_leaf	600	16800
num_boost_round	230	200
n_features	19	26

Note: The same sets of randomly selected parameters were tested for LGBMQR and LGBMAR-ERA5

Table 9.3 P-values from Wilcoxon signed-rank test for the quantile regression models.

R^2			
	ERA5-WSQ	LGBMQR-ERA5	LGBMQR
ERA5-WSQ		2.38E-02	7.39E-01
LGBMQR-ERA5	2.38E-02		5.24E-03
LGBMQR	7.39E-01	5.24E-03	
MAE			
	ERA5-WSQ	LGBMQR-ERA5	LGBMQR
ERA5-WSQ		8.85E-03	7.91E-01
LGBMQR-ERA5	8.85E-03		1.51E-03
LGBMQR	7.91E-01	1.51E-03	
RMSE			
	ERA5-WSQ	LGBMQR-ERA5	LGBMQR
ERA5-WSQ		4.60E-03	5.11E-01
LGBMQR-ERA5	4.60E-03		2.55E-03
LGBMQR	5.11E-01	2.55E-03	

Table 9.4 P-values from Wilcoxon signed-rank test for the time series evaluation.

PC									
	WDC-TS	WDC-PD	ERA5	QM-ERA5	GWA3-ERA5	IDW-TS	IDW-PD	LGBMSI	LGBMSI-ERA5
WDC-TS		4.09E-14	1.04E-02	5.56E-03	1.04E-02	1.14E-01	1.96E-02	3.82E-05	5.20E-11
WDC-PD	4.09E-14		1.03E-12	5.20E-13	1.03E-12	1.86E-15	3.13E-11	4.15E-12	1.60E-16
ERA5	1.04E-02	1.03E-12		9.21E-06	8.06E-01	1.98E-02	1.59E-03	4.36E-04	2.23E-05
QM-ERA5	5.56E-03	5.20E-13	9.21E-06		9.21E-06	1.06E-02	8.24E-04	1.66E-04	9.96E-05
GWA3-ERA5	1.04E-02	1.03E-12	8.06E-01	9.21E-06		1.98E-02	1.59E-03	4.36E-04	2.23E-05
IDW-TS	1.14E-01	1.86E-15	1.98E-02	1.06E-02	1.98E-02		3.37E-06	2.98E-11	3.29E-10
IDW-PD	1.96E-02	3.13E-11	1.59E-03	8.24E-04	1.59E-03	3.37E-06		2.00E-01	4.44E-13
LGBMSI	3.82E-05	4.15E-12	4.36E-04	1.66E-04	4.36E-04	2.98E-11	2.00E-01		6.79E-15
LGBMSI-ERA5	5.20E-11	1.60E-16	2.23E-05	9.96E-05	2.23E-05	3.29E-10	4.44E-13	6.79E-15	
MAE									
	WDC-TS	WDC-PD	ERA5	QM-ERA5	GWA3-ERA5	IDW-TS	IDW-PD	LGBMSI	LGBMSI-ERA5
WDC-TS		3.28E-16	6.78E-01	8.46E-05	5.16E-01	1.91E-01	2.49E-02	1.24E-01	1.28E-05
WDC-PD	3.28E-16		1.26E-11	2.98E-13	4.15E-10	1.75E-10	2.85E-08	1.17E-11	2.04E-14
ERA5	6.78E-01	1.26E-11		3.91E-04	1.84E-01	2.91E-01	5.27E-02	3.85E-01	1.78E-03
QM-ERA5	8.46E-05	2.98E-13	3.91E-04		1.63E-03	1.02E-01	3.77E-01	5.27E-02	3.71E-11

GWA3-ERA5	5.16E-01	4.15E-10	1.84E-01	1.63E-03		9.70E-01	4.55E-01	7.02E-01	7.15E-03
IDW-TS	1.91E-01	1.75E-10	2.91E-01	1.02E-01	9.70E-01		1.80E-02	9.30E-01	1.20E-05
IDW-PD	2.49E-02	2.85E-08	5.27E-02	3.77E-01	4.55E-01	1.80E-02		2.94E-01	4.25E-07
LGBMSI	1.24E-01	1.17E-11	3.85E-01	5.27E-02	7.02E-01	9.30E-01	2.94E-01		3.10E-08
LGBMSI-ERA5	1.28E-05	2.04E-14	1.78E-03	3.71E-11	7.15E-03	1.20E-05	4.25E-07	3.10E-08	
RMSE									
	WDC-TS	WDC-PD	ERA5	QM-ERA5	GWA3-ERA5	IDW-TS	IDW-PD	LGBMSI	LGBMSI-ERA5
WDC-TS		4.55E-16	8.14E-01	2.41E-04	9.70E-01	4.28E-01	6.21E-02	5.21E-01	8.02E-07
WDC-PD	4.55E-16		2.98E-13	6.36E-14	6.10E-13	6.38E-12	1.26E-09	1.99E-13	1.39E-15
ERA5	8.14E-01	2.98E-13		9.21E-06	3.15E-01	1.41E-01	1.36E-02	3.22E-01	9.62E-04
QM-ERA5	2.41E-04	6.36E-14	9.21E-06		7.04E-06	4.01E-02	2.30E-01	5.21E-03	1.21E-09
GWA3-ERA5	9.70E-01	6.10E-13	3.15E-01	7.04E-06		6.72E-01	1.96E-01	5.93E-01	1.01E-02
IDW-TS	4.28E-01	6.38E-12	1.41E-01	4.01E-02	6.72E-01		1.49E-02	6.75E-01	1.52E-06
IDW-PD	6.21E-02	1.26E-09	1.36E-02	2.30E-01	1.96E-01	1.49E-02		9.39E-02	3.16E-08
LGBMSI	5.21E-01	1.99E-13	3.22E-01	5.21E-03	5.93E-01	6.75E-01	9.39E-02		1.05E-08
LGBMSI-ERA5	8.02E-07	1.39E-15	9.62E-04	1.21E-09	1.01E-02	1.52E-06	3.16E-08	1.05E-08	

Table 9.5 P-values from Wilcoxon signed-rank test for the OP metric.

OP									
	WDC-TS	WDC-PD	ERA5	QM-ERA5	GWA-ERA5	IDW-TS	IDW-PD	LGBMSI	LGBMSI-ERA5
WDC-TS		9.83E-03	6.60E-04	1.12E-12	9.11E-05	7.31E-03	9.04E-03	1.59E-14	2.75E-02
WDC-PD	9.83E-03		4.05E-01	1.17E-12	1.03E-01	9.90E-01	9.81E-01	6.49E-16	5.94E-06
ERA5	6.60E-04	4.05E-01		1.49E-05	4.13E-01	4.89E-01	5.18E-01	1.89E-13	8.76E-06
QM-ERA5	1.12E-12	1.17E-12	1.49E-05		4.16E-03	6.91E-07	5.04E-07	3.38E-17	5.40E-15
GWA-ERA5	9.11E-05	1.03E-01	4.13E-01	4.16E-03		1.17E-01	1.13E-01	2.93E-14	2.73E-06
IDW-TS	7.31E-03	9.90E-01	4.89E-01	6.91E-07	1.17E-01		4.80E-01	1.84E-13	1.34E-04
IDW-PD	9.04E-03	9.81E-01	5.18E-01	5.04E-07	1.13E-01	4.80E-01		1.06E-12	2.27E-05
LGBMSI	1.59E-14	6.49E-16	1.89E-13	3.38E-17	2.93E-14	1.84E-13	1.06E-12		6.38E-12
LGBMSI-ERA5	2.75E-02	5.94E-06	8.76E-06	5.40E-15	2.73E-06	1.34E-04	2.27E-05	6.38E-12	

10 INFORMATION SUPPLÉMENTAIRE POUR L'ARTICLE 5

Supplementary material for "LSTM and Transformer-based framework for bias correction of ERA5 hourly wind speeds"

Contents

Table 10.1 Covariates used in the study.

Type of covariate	Covariates	Description	Scaling or interpolation method	Spatial scale
Static	Distance from the coast (Dcoast)	Distance from the coast	n.a	n.a
	Surface roughness length (SRL)	Dominant Surface roughness length	Mode resampling	100 m, 1 km, 5 km, 10 km 20 km,
	Elevation percentile (ELP)	Percentile of the grid cell elevation relative to the neighbouring cells.	Gaussian filter	100 m, 200 m, 1 km, 2 km, 5 km, 10 km
	Elevation ((ELV)	Elevation above sea level at the location of the meteorological station	n.a	100 m
	Ruggedness index (RGN)	Measure the local terrain heterogeneity	Gaussian filter	100 m, 200 m, 1 km, 2 km, 5 km, 10 km
	Standard deviation of slope (SDS)	Measure of surface roughness.	Gaussian filter	50 m, 200 m, 1 km, 2 km, 5 km, 10 km
	Deviation from mean elevation (DME)	Difference between the grid cell elevation and the mean of its neighbouring cells normalized by the standard deviation.	Gaussian filter	50 m, 200 m, 1 km, 2 km, 5 km, 10 km
	Aspect (ASP)	Slope orientation	Gaussian filter	50 m, 200 m, 1 km, 2 km, 5 km, 10 km
	Slope (SLP)	Slope at the grid cell.	Gaussian filter	50 m, 200 m, 1 km, 2 km, 5 km, 10 km
	tangential curvature (TAC)	Measure of surface curvature.	Gaussian filter	50 m, 200 m, 1 km, 2 km, 5 km, 10 km
	Total curvature (TOC)	Measure of surface curvature.	Gaussian filter	50 m, 200 m, 1 km, 2 km, 5 km, 10 km
	Wind speed quantiles	5th, 50th, and 95th ERA5 wind speed percentiles estimated from 2008 to 2017	Nearest neighbor	n.a
Time resolved	blh	Boundary layer height	Nearest neighbor	n.a
	sp	Surface pressure	Nearest neighbor	n.a

	10u	U component of wind at 10 m	Nearest neighbor	n.a
	10v	V component of wind at 10 m	Nearest neighbor	n.a
	$\sqrt{u^2 + v^2}$	wind speed at 10 m	Nearest neighbor	n.a
	2t	2 m temperature	Nearest neighbor	n.a

Two types of covariates were used: static and time resolved. The time-resolved covariates are reanalysis meteorological variables from ERA5.

Table 10.2 Assigned surface roughness length (SRL) to land cover classes.

Land cover class	**Pixel ID**	**SRL (m)**
Temperate or sub-polar needleleaf forest	1	0.9
Sub-polar taiga needleleaf forest	2	0.9
Temperate or sub-polar broadleaf deciduous forest	5	0.9
Temperate or sub-polar Shrubland	8	0.01
Temperate or sub-polar grassland	10	0.01
Sub-polar or polar shrubland-lichen-moss	11	0.01
Sub-polar or polar grassland-lichen-moss	12	0.01
Sub-polar or polar barren-lichen-moss	13	0.01
Wetland	14	0.04
Cropland	15	0.1
Barren lands	16	0.005
Urban	17	0.8
Water	18	0.0002
Snow and ice	19	0.0002

The table presents the land cover classes derived from the 2020 Canada land cover dataset, along with the assigned

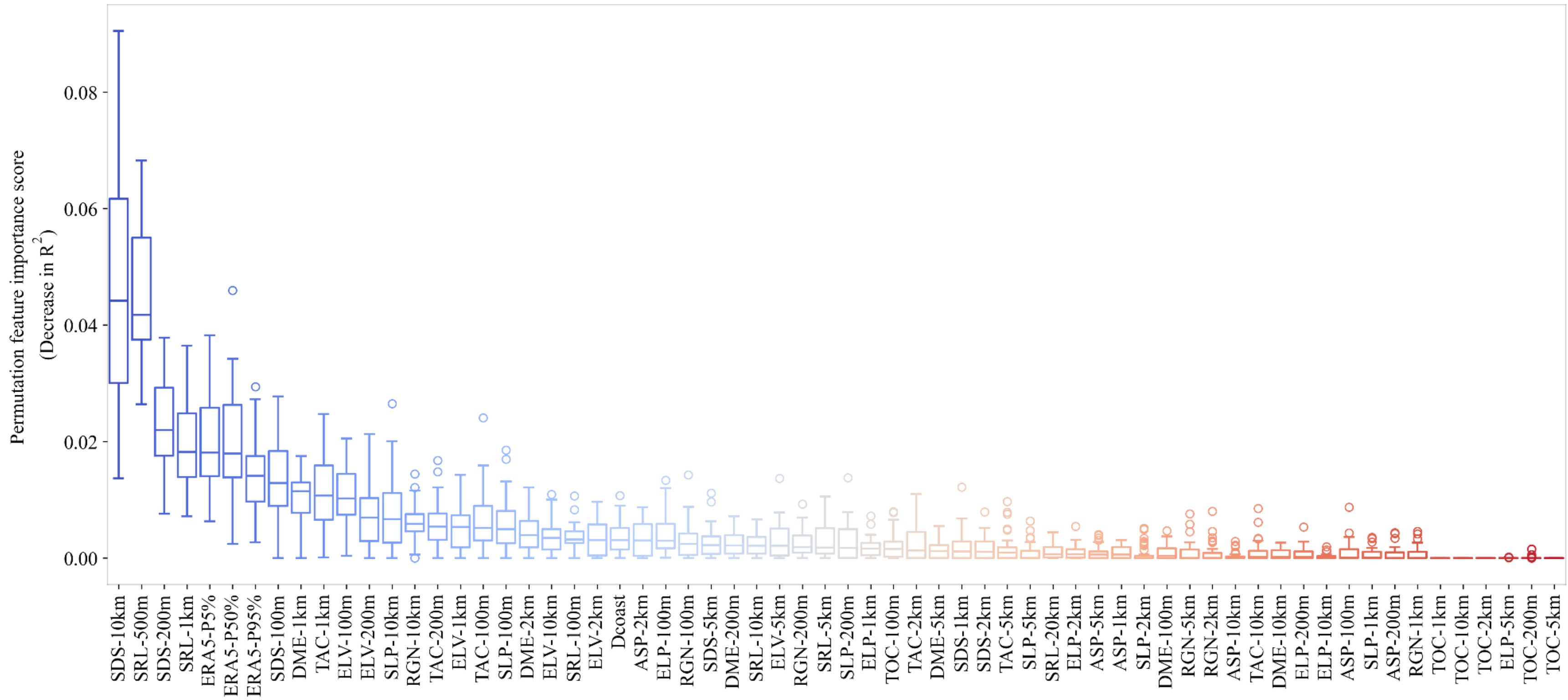


Figure 10.1 **Covariate importance based on permutation features importance (PFI, (Breiman, 2001)).**

The PFI score distribution was estimated by computing the PFI scores for the most optimal hyperparameter configurations of the gradient boosting algorithm using a cross-validation approach with the training set. The vertical axis represents the PFI score, while the horizontal axis lists the covariates. Refer to Table S1 for covariate name abbreviations, followed by the corresponding spatial scale.

Table 10.3 Criteria used to classify test stations in coastal, hilly, mountainous, low SRL and high SRL areas.

Area type	Area delimitation criterions	Data source
Hilly and mountainous	Located in one of Canada ecodistricts dominated (more than 80%) by hills and mountains	https://open.canada.ca/data/en/dataset/546f1a67-5f22-4af9-8618-b94e1d33c52f (accessed 29 October 2024)
Coastal	Located less than 10 km from Canada coastal waters (excluding mountainous and hilly areas)	https://open.canada.ca/data/en/dataset/6c78fb2f-d23b-45b4-b3af-cc6f6cc4fff8 (accessed 29 October 2024)
High SRL	Surface roughness length greater than 0.5 m in a radius of 1 km (excluding mountainous, hilly, and coastal areas)	https://open.canada.ca/data/en/dataset/ee1580ab-a23d-4f86-a09b-79763677eb47 (accessed 29 October 2024)
Low SRL	Surface roughness length less than 0.5 m in a radius of 1 km (excluding mountainous, hilly, and coastal areas)	https://open.canada.ca/data/en/dataset/ee1580ab-a23d-4f86-a09b-79763677eb47 (accessed 29 October 2024)

The link to the data source used for the classification are provided in the last column of the table.

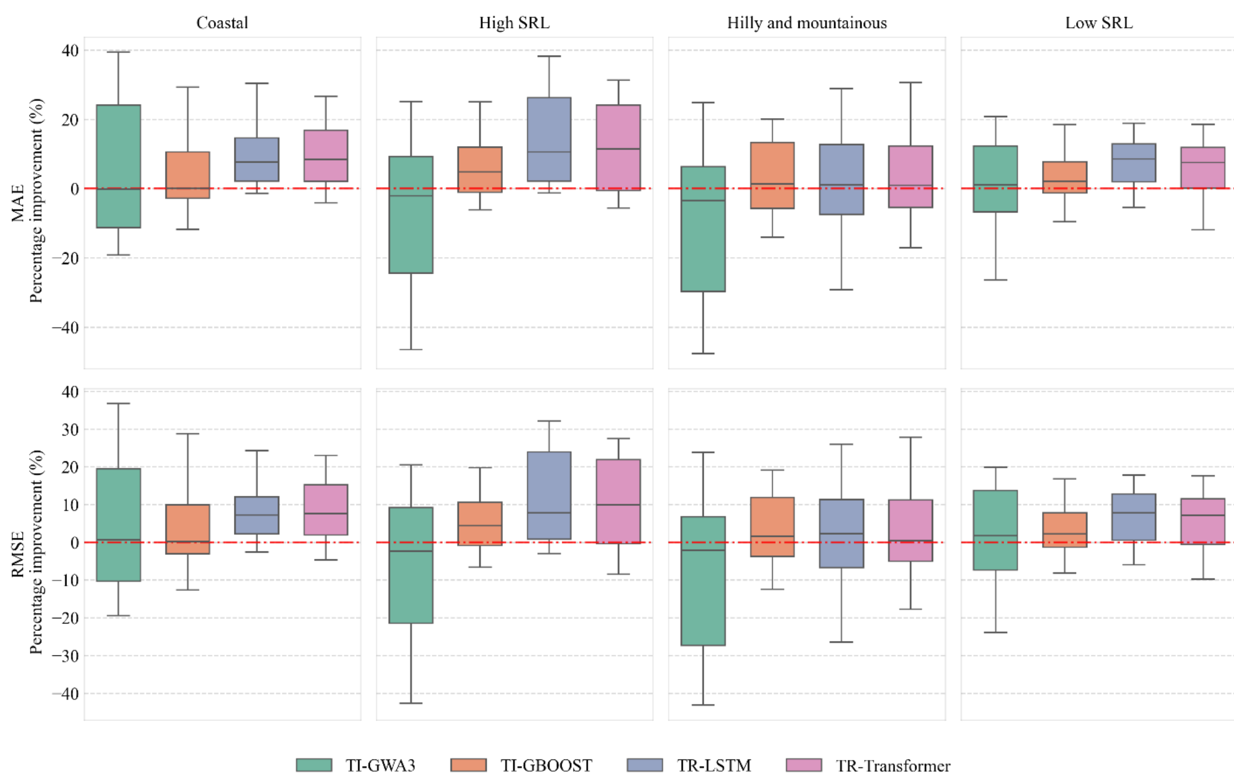


Figure 10.2 Boxplot of the skill scores.

The skills scores are the percentage improvements relative to UC-ERA5 in time series MAE and RMSE. The boxplots are displayed separately for different types of environments and models. The boxes represent distribution quartiles, while the whiskers indicate the 10-90% range. The horizontal red dotted lines mark the 0% percentage improvement over UC-ERA5.

10.1 Additional results on selected stations

Four stations, each representing a different area type, were selected from the test set to illustrate the uncorrected and corrected wind speed time series through multiple perspectives, including time series, 2D kernel density, and distribution plots.

Station ECCC Climate ID No. 7056202 is located in a coastal area. The uncorrected ERA5 wind speeds (UC-ERA5) tend to overestimate the observed wind speeds across the entire distribution (Figure 10.3 and Figure 10.4). The corrected wind speed using the time-resolved LSTM model (TR-LSTM) show improvements by reducing the bias; However, the model tends to overcorrect high wind speeds, leading to an underestimation of extreme wind events.

Station ECCC Climate ID No. 8202592 is located in a high surface roughness length (SRL) area. UC-ERA5 tends to substantially overestimate the observed wind speeds (Figure 10.5 and Figure 10.6). TR-LSTM shows improvement across the entire distribution.

Station ECCC Climate ID No. 3030200 is situated in a low SRL area. UC-ERA5 generally underestimates the observed wind speeds, particularly in the mid-to-upper range of the distribution (Figure 10.7 and Figure 10.8). TR-LSTM effectively reduces this bias with a less pronounced effect in the upper part of the distribution where an underestimation tendency is still observable.

Station ECCC Climate ID No. 7091299 is located in a hilly and mountainous area. UC-ERA5 exhibits significant overestimation across the entire distribution, likely due to the complex topography that ERA5 struggles to resolve at its coarse resolution (Figure 10.9 and Figure 10.10). TR-LSTM improved the accuracy by reducing the overall bias and better aligning with observed wind speed variations. However, despite this improvement, TR-LSTM still shows some biases, particularly in capturing extreme wind speeds, where it tends to slightly overestimate peak values.

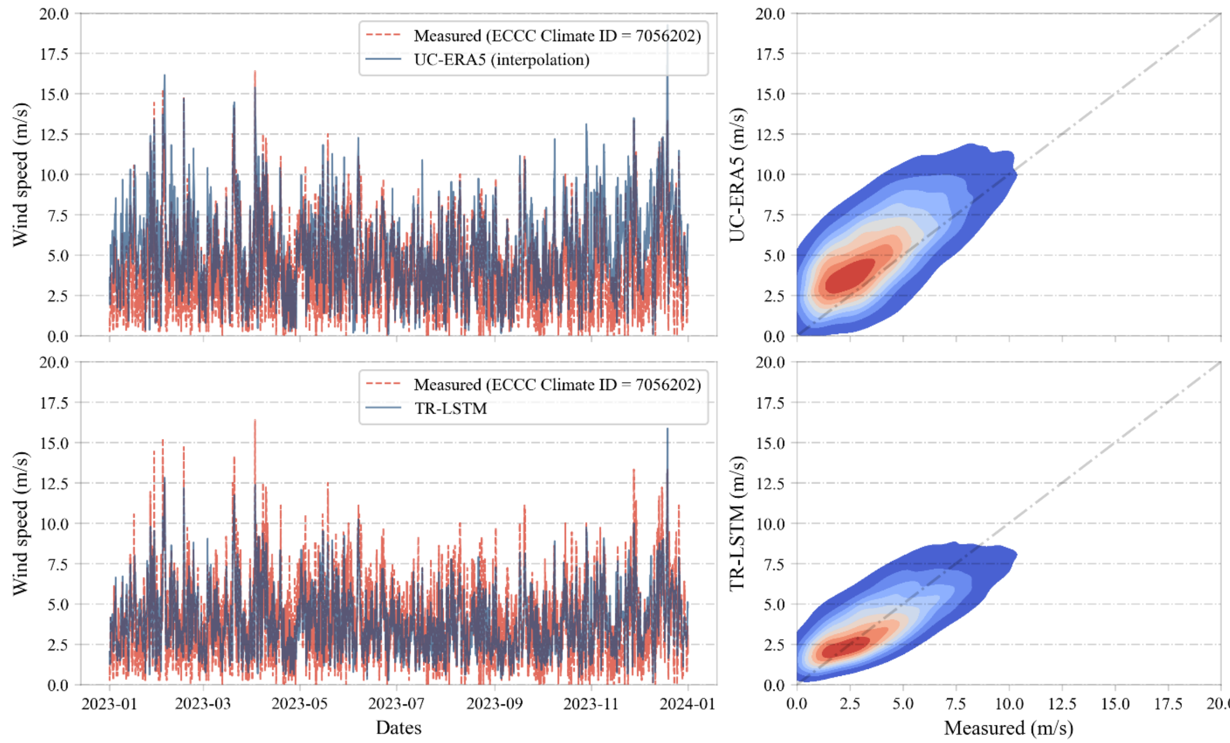


Figure 10.3 Time series and 2D kernel density plot for station ECCC Climate ID No. 7056202 for 2023. The station is located in a coastal area. The top-left and bottom-left panels present the time series of UC-ERA5, and TR-LSTM overlaid on the measured hourly wind speeds, respectively. The top-right and bottom-right panels display the 2D kernel density plot comparing measured wind speeds with UC-ERA5 and TR-LSTM, respectively.

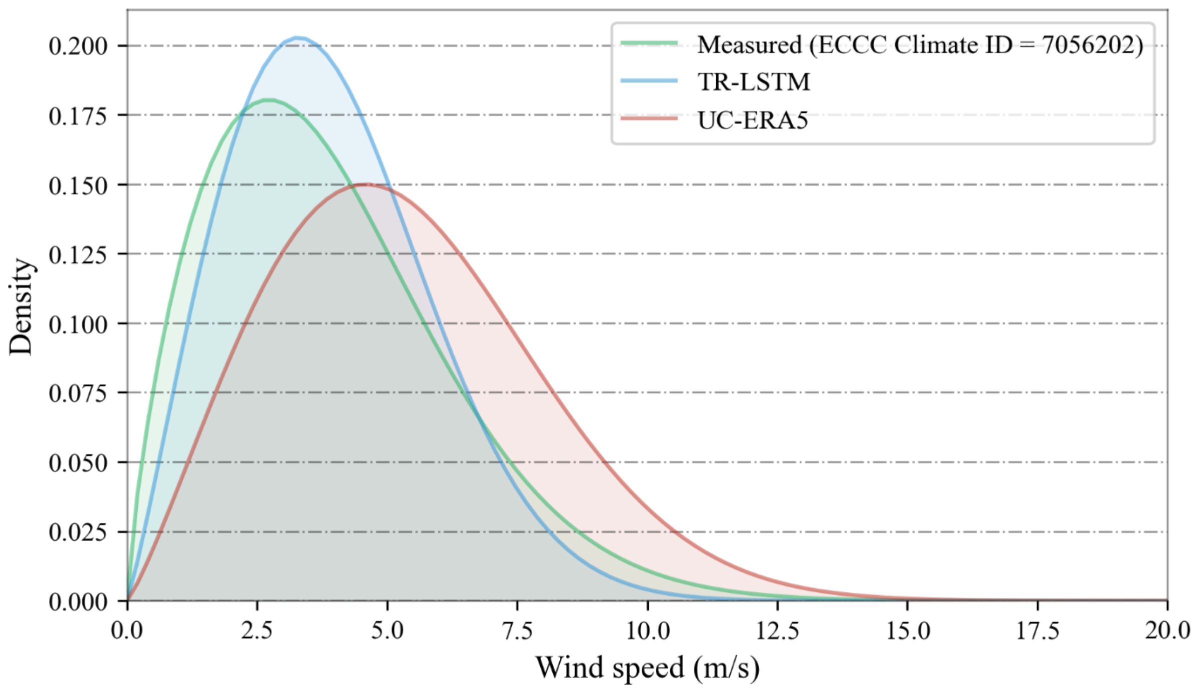


Figure 10.4 Density plot for station ECCC Climate ID No. 7056202.

The station is located in a coastal area. The probability density function (pdf) was estimated by fitting the two parameters of the Weibull distribution to the data using the maximum likelihood estimation method.

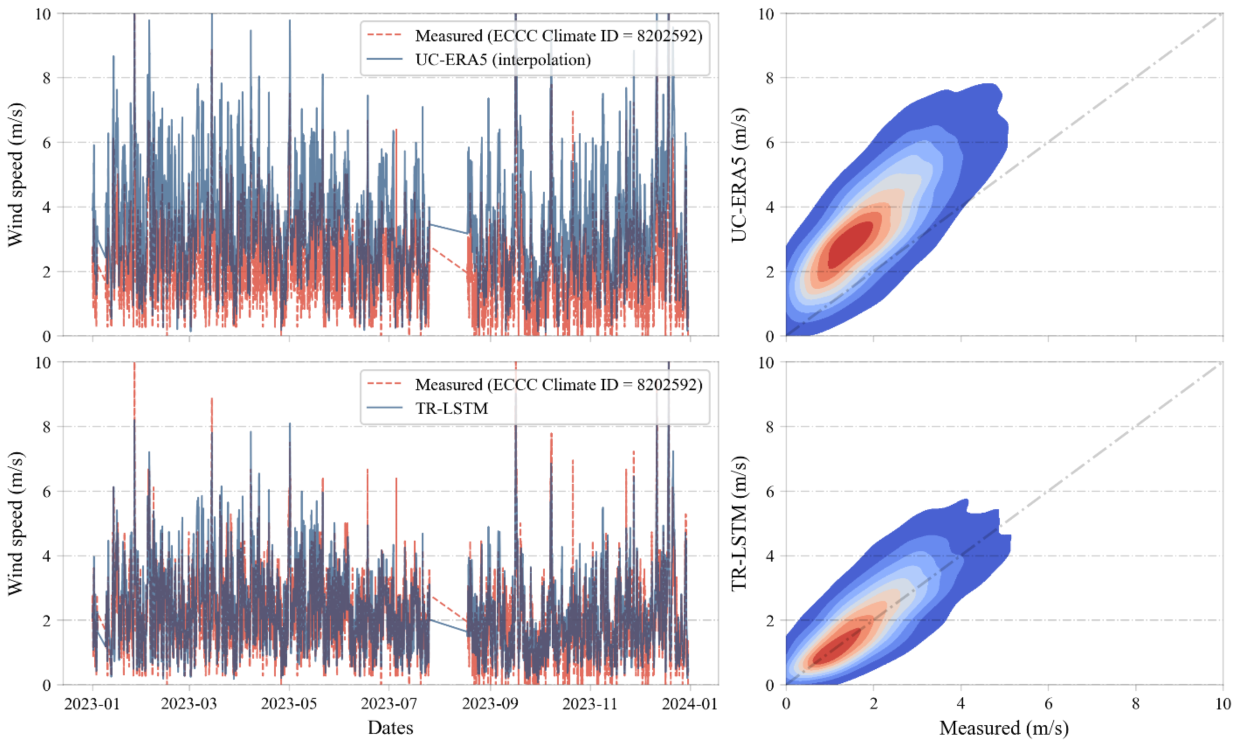


Figure 10.5 Time series and 2D kernel density plot for station ECCC Climate ID No. 8202592 for 2023. The station is located in a region with high surface roughness length. The top-left and bottom-left panels present the time series of UC-ERA5, and TR-LSTM overlaid on the measured hourly wind speeds, respectively. The top-right and bottom-right panels display the 2D kernel density plot comparing measured wind speeds with UC-ERA5 and TR-LSTM, respectively.

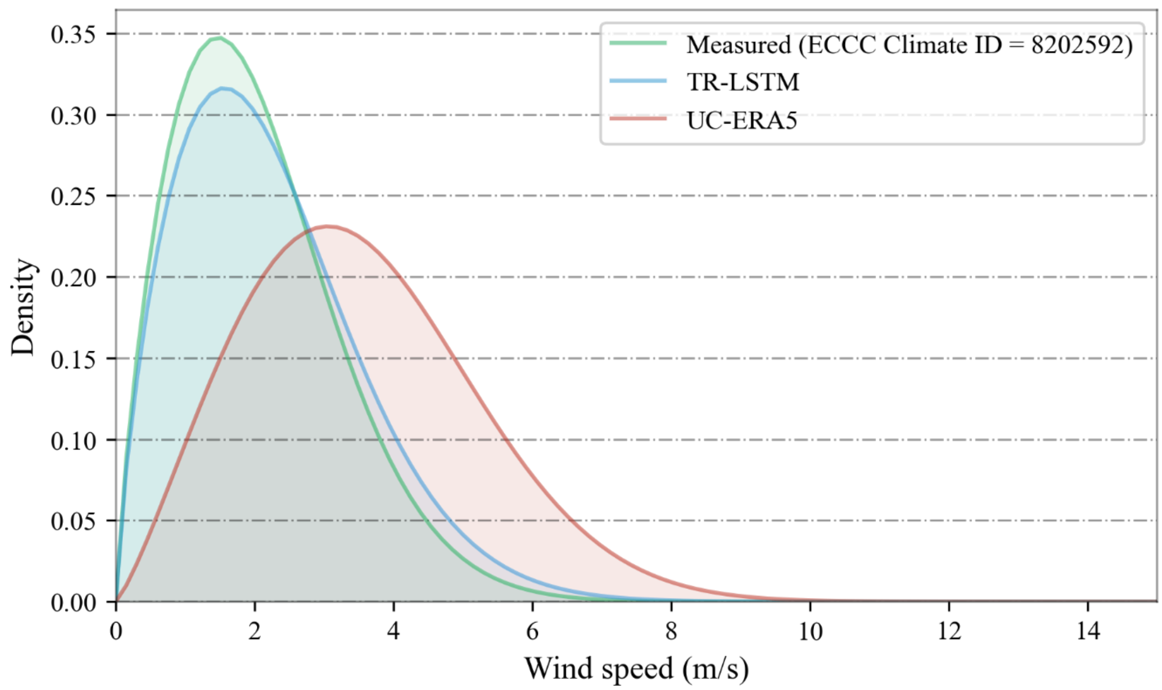


Figure 10.6 Density plot for station ECCC Climate ID No. 8202592.

The station is located in a region with high surface roughness length. The probability density function (pdf) was estimated by fitting the two parameters of the Weibull distribution to the data using the maximum likelihood estimation method.

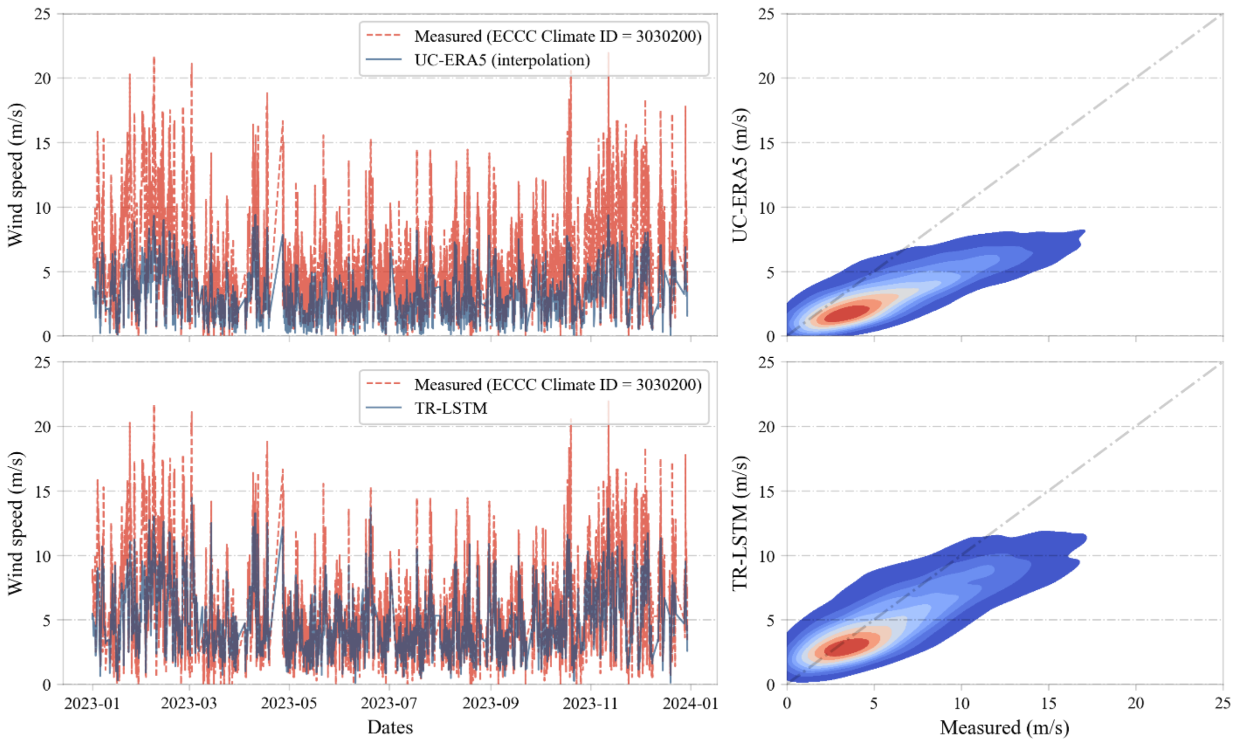


Figure 10.7 Time series and 2D kernel density plot for station ECCC Climate ID No. 3030200 for 2023. The station is located in a region with low surface roughness length. The top-left and bottom-left panels present the time series of UC-ERA5, and TR-LSTM overlaid on the measured hourly wind speeds, respectively. The top-right and bottom-right panels display the 2D kernel density plot comparing measured wind speeds with UC-ERA5 and TR-LSTM, respectively.

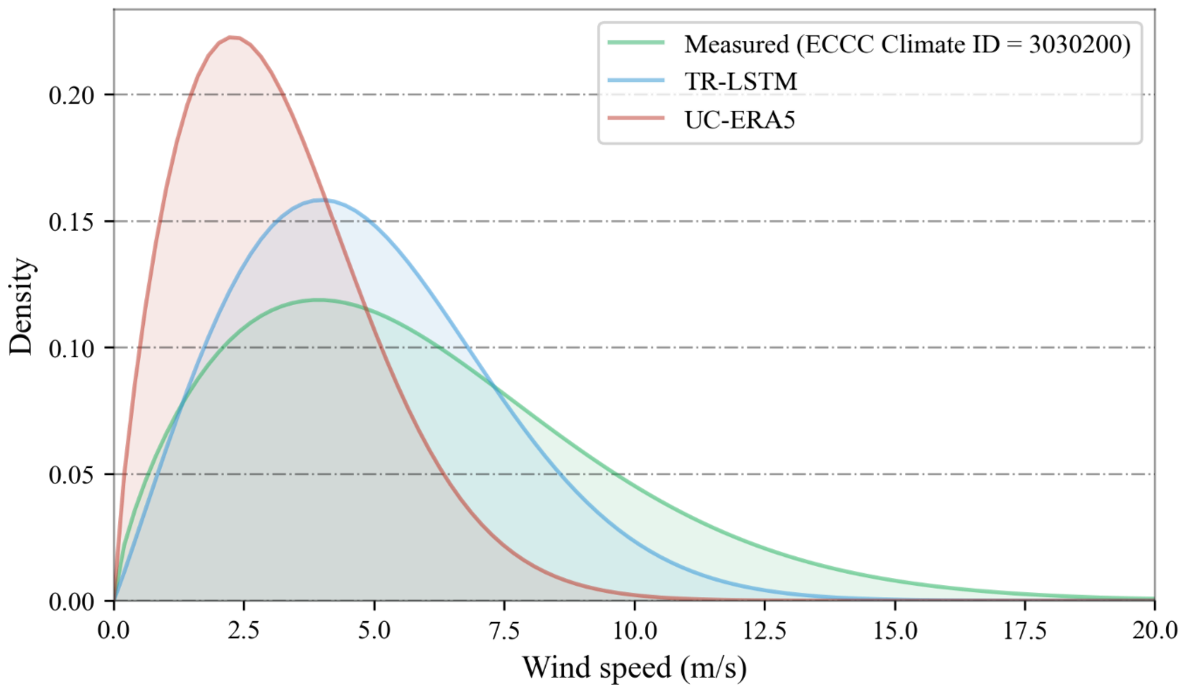


Figure 10.8 Density plot for station ECCC Climate ID No. 3030200.

The station is located in a region with low surface roughness length. The probability density function (pdf) was estimated by fitting the two parameters of the Weibull distribution to the data using the maximum likelihood estimation method.

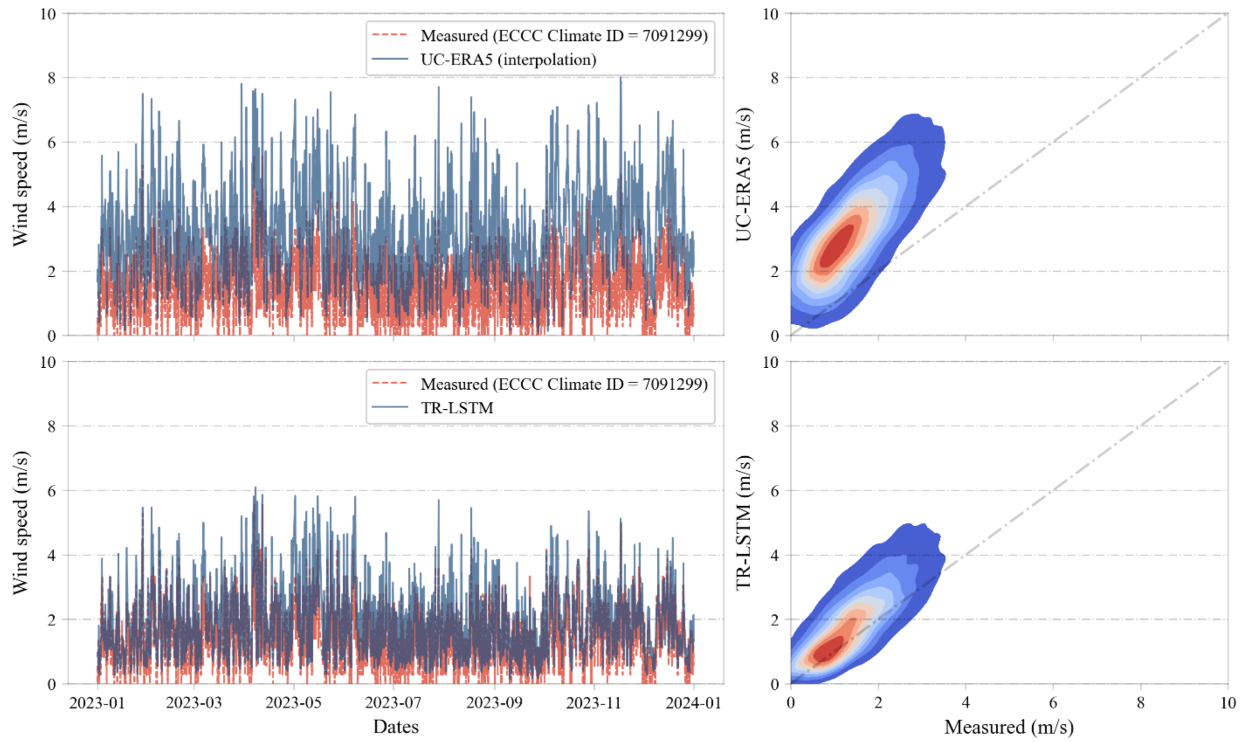


Figure 10.9 Time series and 2D kernel density plot for station ECCC Climate ID No. 7091299 for 2023. The station is located in a hilly and mountainous area. The top-left and bottom-left panels present the time series of UC-ERA5, and TR-LSTM overlaid on the measured hourly wind speeds, respectively. The top-right and bottom-right panels display the 2D kernel density plot comparing measured wind speeds with UC-ERA5 and TR-LSTM, respectively.

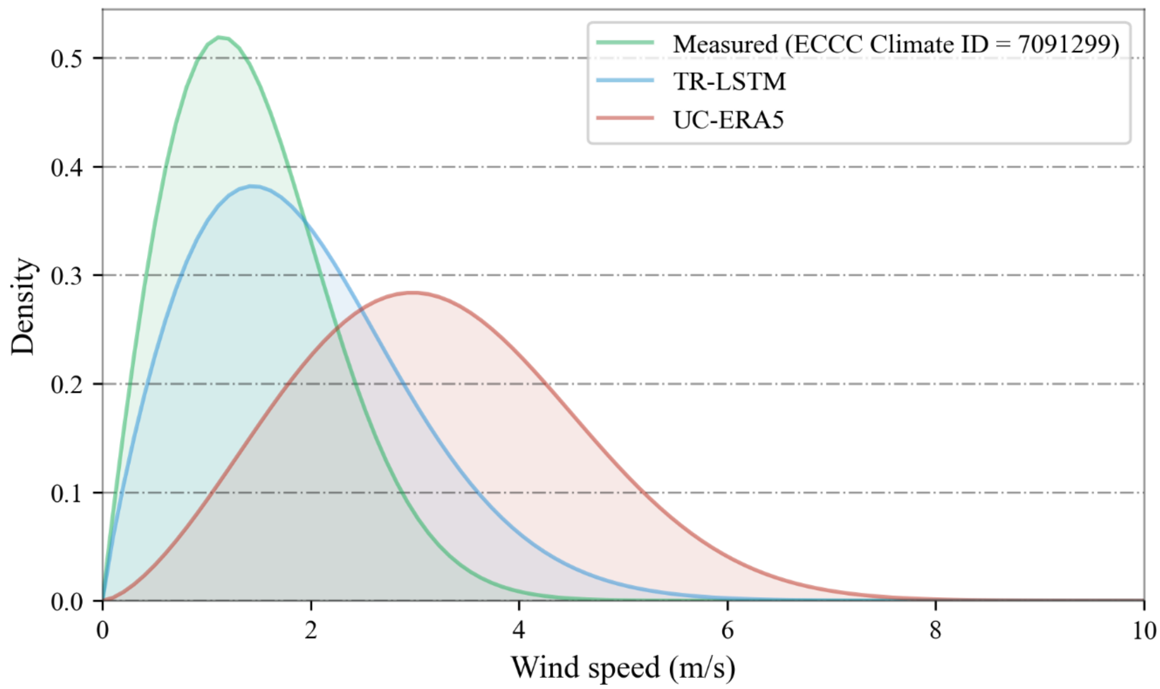


Figure 10.10 Density plot for station ECCC Climate ID No. 3030200.

The station is located in a hilly and mountainous area. The probability density function (pdf) was estimated by fitting the two parameters of the Weibull distribution to the data using the maximum likelihood estimation method.

11 BIBLIOGRAPHIE

Abdollahzadeh B, Khodadadi N, Barshandeh S, Trojovský P, Gharehchopogh FS, El-kenawy E-SM, Abualigah L, Mirjalili S (2024) Puma optimizer (PO): a novel metaheuristic optimization algorithm and its application in machine learning. *Cluster Computing* 27(4):5235-5283.

Addison JD, Hunter A, Bass J, Rebbeck M (2000) A neural network version of the measure correlate predict algorithm for estimating wind energy yield. Édit Centre for Adaptive Systems SOCEaT, University of SunderlandSunderland, UK), p 4.

Aguilar E, Auer I, Brunet M, Peterson TC, Wieringa J (2003) Guidance on metadata and homogenization. *Wmo Td* 1186(January 2003):1-53.

Ahmad I, Ahmad T, Shahzad U, Ameer MA, Emam W, Tashkandy Y, Badar Z (2024) Estimation of regional and at-site quantiles of extreme winds under flood index procedure. *Heliyon* 10(1):e23388.

Ahmad T, Zhang D (2020) A critical review of comparative global historical energy consumption and future demand: The story told so far. *Energy Reports* 6:1973-1991.

Akgül FG, Şenoğlu B, Arslan T (2016) An alternative distribution to Weibull for modeling the wind speed data: Inverse Weibull distribution. *Energy Conversion and Management* 114:234-240.

Alavi O, Mohammadi K, Mostafaeipour A (2016) Evaluating the suitability of wind speed probability distribution models: A case of study of east and southeast parts of Iran. *Energy Conversion and Management* 119:101-108.

Alexandersson H (1986) A homogeneity test applied to precipitation data. *Journal of Climatology* 6(6):661-675.

Ali S, Lee S-M, Jang C-M (2018) Forecasting the Long-Term Wind Data via Measure-Correlate-Predict (MCP) Methods. *Energies* 11(6):1541.

Ali S, Mahdi A, Shaban A (2012) Wind Speed Estimation for Iraq using several Spatial Interpolation Methods. *The British Journal for the Philosophy of Science* 7:48-55.

Alsamamra H, Pozo-Vazquez D, Santos-Alamillos FJ, Ruiz-Arias JA, Lara-Fanego V, Luzon-Cuesta R, Tovar-Pescador J (2010) Mapping surface wind speed in Andalusia (Southern Spain) based on residual kriging. *10th EMS Annual Meeting.*September 01, 2010), p EMS2010-2124.

Ambach D, Schmid W (2017) A new high-dimensional time series approach for wind speed, wind direction and air pressure forecasting. *Energy* 135:833-850.

Amini F, Hu G (2021) A two-layer feature selection method using Genetic Algorithm and Elastic Net. *Expert Systems with Applications* 166:114072.

Andrews DG (2010) *An Introduction to Atmospheric Physics.* Cambridge University Press, Cambridge, 2. https://www.cambridge.org/core/product/3F0120AB397D07A552C821158930ACB8

Aniskevich S, Bezrukovs V, Zandovskis U, Bezrukovs D (2017) Modelling the Spatial Distribution of Wind Energy Resources in Latvia. *Latvian Journal of Physics and Technical Sciences* 54(6):10-20.

Apicella A, Donnarumma F, Isgrò F, Prevete R (2021) A survey on modern trainable activation functions. *Neural Networks* 138:14-32.

Archer CL, Jacobson MZ (2005) Evaluation of global wind power. *Journal of Geophysical Research: Atmospheres* 110(D12).

Aries N, Boudia SM, Ounis H (2018) Deep assessment of wind speed distribution models: A case study of four sites in Algeria. *Energy Conversion and Management* 155:78-90.

Ayik A, Ijumba N, Kabiri C, Goffin P (2021) Preliminary wind resource assessment in South Sudan using reanalysis data and statistical methods. *Renewable and Sustainable Energy Reviews* 138:110621.

Ayotte KW, Hughes DE (2004) Observations of Boundary-Layer Wind-Tunnel Flow over Isolated Ridges of Varying Steepness and Roughness. *Boundary-Layer Meteorology* 112(3):525-556.

Azorin-Molina C, Guijarro J-A, McVicar TR, Vicente-Serrano SM, Chen D, Jerez S, Espírito-Santo F (2016) Trends of daily peak wind gusts in Spain and Portugal, 1961–2014. *Journal of Geophysical Research: Atmospheres* 121(3):1059-1078.

Azorin-Molina C, Guijarro JA, McVicar TR, Trewin BC, Frost AJ, Chen D (2019) An approach to homogenize daily peak wind gusts: An application to the Australian series. *International Journal of Climatology* 39(4):2260-2277.

Azorin-Molina C, Rehman S, Guijarro JA, McVicar TR, Minola L, Chen D, Vicente-Serrano SM (2018) Recent trends in wind speed across Saudi Arabia, 1978–2013: a break in the stilling. *International Journal of Climatology* 38(S1):e966-e984.

Azorin-Molina C, Vicente-Serrano SM, McVicar TR, Jerez S, Sanchez-Lorenzo A, López-Moreno J-I, Revuelto J, Trigo RM, Lopez-Bustins JA, Espírito-Santo F (2014) Homogenization and Assessment of Observed Near-Surface Wind Speed Trends over Spain and Portugal, 1961–2011. *Journal of Climate* 27(10):3692-3712.

Baño-Medina J, Manzanas R, Gutiérrez JM (2020) Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geosci. Model Dev.* 13(4):2109-2124.

Bartók B, Tobin I, Vautard R, Vrac M, Jin X, Levavasseur G, Denvil S, Dubus L, Parey S, Michelangeli P-A, Troccoli A, Saint-Drenan Y-M (2019) A climate projection dataset tailored for the European energy sector. *Climate Services* 16:100138.

Basse A, Callies D, Grötzner A, Pauscher L (2021) Seasonal effects in the long-term correction of short-term wind measurements using reanalysis data. *Wind Energ. Sci.* 6(6):1473-1490.

Basse A, Pauscher L, Callies D (2020) Improving Vertical Wind Speed Extrapolation Using Short-Term Lidar Measurements. *Remote Sensing* 12(7).

Baydaroğlu Ö, Koçak K (2019) Spatiotemporal analysis of wind speed via the Bayesian maximum entropy approach. *Environmental Earth Sciences* 78(1):17.

Beaucage P, Brower MC, Tensen J (2014) Evaluation of four numerical wind flow models for wind resource mapping. *Wind Energy* 17(2):197-208.

Belušić A, Prtenjak MT, Güttler I, Ban N, Leutwyler D, Schär C (2018) Near-surface wind variability over the broader Adriatic region: insights from an ensemble of regional climate models. *Climate Dynamics* 50(11):4455-4480.

Ben Alaya MA, Chebana F, Ouarda TBMJ (2016) Multisite and multivariable statistical downscaling using a Gaussian copula quantile regression model. *Climate Dynamics* 47(5):1383-1397.

Ben Bouallègue Z, Clare MCA, Magnusson L, Gascón E, Maier-Gerber M, Janoušek M, Rodwell M, Pinault F, Dramsch JS, Lang STK, Raoult B, Rabier F, Chevallier M, Sandu I, Dueben P, Chantry M, Pappenberger F (2024) The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context. *Bulletin of the American Meteorological Society* 105(6):E864-E883.

Bentéjac C, Csörgő A, Martínez-Muñoz G (2021) A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review* 54(3):1937-1967.

Bentsen LØ, Warakagoda ND, Stenbro R, Engelstad P (2023) Spatio-temporal wind speed forecasting using graph networks and novel Transformer architectures. *Applied Energy* 333:120565.

Bergstra J, Bengio Y (2012a) Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* 13:281–305.

Bergstra J, Bengio Y (2012b) Random search for hyper-parameter optimization. *Journal of Machine Learning Research* 13(2).

Berndt C, Haberlandt U (2018) Spatial interpolation of climate variables in Northern Germany—Influence of temporal resolution and network density. *Journal of Hydrology: Regional Studies* 15:184-202.

Bett PE, Thornton HE, Clark RT (2017) Using the Twentieth Century Reanalysis to assess climate variability for the European wind industry. *Theoretical and Applied Climatology* 127(1):61-80.

Bloom A, Kotroni V, Lagouvardos K (2008) Climate change impact of wind energy availability in the Eastern Mediterranean using the regional climate model PRECIS. *Nat. Hazards Earth Syst. Sci.* 8(6):1249-1257.

Bosch J, Staffell I, Hawkes AD (2017) Temporally-explicit and spatially-resolved global onshore wind energy potentials. *Energy* 131:207-217.

Bosch J, Staffell I, Hawkes AD (2018) Temporally explicit and spatially resolved global offshore wind energy potentials. *Energy* 163:766-781.

Breiman L (2001) Random Forests. *Machine Learning* 45(1):5-32.

Breiman L, Friedman J, Olshen RA, J. Stone C (1984) *Classification and regression trees.* Chapman and Hall/CRC., 1st ed.

Brinckmann S, Krähenmann S, Bissolli P (2016) High-resolution daily gridded data sets of air temperature and wind speed for Europe. *Earth Syst. Sci. Data* 8(2):491-516.

Brune S, Keller JD (2022) Statistical post-processing of reanalysis wind speeds at hub heights using a diagnostic wind model and neural networks. *Wind Energ. Sci.* 7(5):1905-1918.

Brune S, Keller JD, Wahl S (2021) Evaluation of wind speed estimates in reanalyses for wind energy applications. *Adv. Sci. Res.* 18:115-126.

Burn DH, Goel NK (2000) The formation of groups for regional flood frequency analysis. *Hydrological Sciences Journal* 45(1):97-112.

Campos RM, Guedes Soares C (2018) Spatial distribution of offshore wind statistics on the coast of Portugal using Regional Frequency Analysis. *Renewable Energy* 123:806-816.

Cannon AJ (2011) Quantile regression neural networks: Implementation in R and application to precipitation downscaling. *Computers & Geosciences* 37(9):1277-1284.

Cannon AJ (2018) Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. *Stochastic Environmental Research and Risk Assessment* 32(11):3207-3225.

Cannon AJ, Sobie SR, Murdock TQ (2015) Bias Correction of GCM Precipitation by Quantile Mapping: How Well Do Methods Preserve Changes in Quantiles and Extremes? *Journal of Climate* 28(17):6938-6959.

Carrega P (2008) Le vent: importance, mesures, modélisation et tribulations. *Bulletin de la Société Géographique de Liège* 51:17-29.

Carta JA, Cabrera P, Matías JM, Castellano F (2015) Comparison of feature selection methods using ANNs in MCP-wind speed methods. A case study. *Applied Energy* 158:490-507.

Carta JA, Ramírez P, Velázquez S (2009) A review of wind speed probability distributions used in wind energy analysis: Case studies in the Canary Islands. *Renewable and Sustainable Energy Reviews* 13(5):933-955.

Carta JA, Velázquez S, Cabrera P (2013) A review of measure-correlate-predict (MCP) methods used to estimate long-term wind characteristics at a target site. *Renewable and Sustainable Energy Reviews* 27:362-400.

Carvalho D, Rocha A, Costoya X, deCastro M, Gómez-Gesteira M (2021) Wind energy resource over Europe under CMIP6 future climate projections: What changes from CMIP5 to CMIP6. *Renewable and Sustainable Energy Reviews* 151:111594.

Castellarin A, Botter G, Hughes DA, Liu S, Ouarda TBMJ, Parajka J, Post DA, Sivapalan M, Spence C, Viglione A, Vogel RM (2013) Prediction of flow duration curves in ungauged basins. *Runoff Prediction in Ungauged Basins: Synthesis across Processes, Places and Scales*, Blöschl G, Savenije H, Sivapalan M, Viglione A,Wagener T (Édit.) Cambridge University Press, CambridgeDOI: 10.1017/CBO9781139235761.010. p 135-162.

Castorrini A, Gentile S, Geraldi E, Bonfiglioli A (2021) Increasing spatial resolution of wind resource prediction using NWP and RANS simulation. *Journal of Wind Engineering and Industrial Aerodynamics* 210:104499.

Cavaiola M, Tuju PE, Ferrari F, Casciaro G, Mazzino A (2023) Ensemble Machine Learning greatly improves ERA5 skills for wind energy applications. *Energy and AI* 13:100269.

Cavalleri F, Viterbo F, Brunetti M, Bonanno R, Manara V, Lussana C, Lacavalla M, Maugeri M (2024) Inter-comparison and validation of high-resolution surface air temperature reanalysis fields over Italy. *International Journal of Climatology* 44(8):2681-2700.

Cellura M, Cirrincione G, Marvuglia A, Miraoui A (2008a) Wind speed spatial estimation for energy planning in Sicily: A neural kriging application. *Renewable Energy* 33(6):1251-1266.

Cellura M, Cirrincione G, Marvuglia A, Miraoui A (2008b) Wind speed spatial estimation for energy planning in Sicily: Introduction and statistical analysis. *Renewable Energy* 33(6):1237-1250.

Chai T, Draxler RR (2014) Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* 7(3):1247-1250.

Chatterjee S, Hadi AS (2013) *Regression analysis by example.* John Wiley & Sons, Hoboken, New Jersey, USA. 424 p

Chen J, de Hoogh K, Gulliver J, Hoffmann B, Hertel O, Ketzel M, Bauwelinck M, van Donkelaar A, Hvidtfeldt UA, Katsouyanni K, Janssen NAH, Martin RV, Samoli E, Schwartz PE, Stafoggia M, Bellander T, Strak M, Wolf K, Vienneau D, Vermeulen R, Brunekreef B, Hoek G (2019) A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. *Environment International* 130:104934.

Chen SX (2000) Probability Density Function Estimation Using Gamma Kernels. *Annals of the Institute of Statistical Mathematics* 52(3):471-480.

Chen T, Guestrin C (2016) XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, San Francisco, California, USA), p 785–794.

Chen Y, Dong Z, Wang Y, Su J, Han Z, Zhou D, Zhang K, Zhao Y, Bao Y (2021a) Short-term wind speed predicting framework based on EEMD-GA-LSTM method under large scaled wind history. *Energy Conversion and Management* 227:113559.

Chen Y, Sharma S, Zhou X, Yang K, Li X, Niu X, Hu X, Khadka N (2021b) Spatial performance of multiple reanalysis precipitation datasets on the southern slope of central Himalaya. *Atmospheric Research* 250:105365.

Cherp A, Vinichenko V, Tosun J, Gordon JA, Jewell J (2021) National growth dynamics of wind and solar power compared to the growth required for global climate targets. *Nature Energy* 6(7):742-754.

Chollet F (2021) *Deep Learning with Python, Second Edition.* Manning. https://books.google.ca/books?id=XHpKEAAAQBAJ

Cirrincione G, Marvuglia A (2009) A Novel Self-organizing Neural Technique for Wind Speed Mapping. in *Sustainability in Energy and Buildings*, Édit Howlett RJ, Jain LC,Lee SH (Springer Berlin Heidelberg, Berlin, Heidelberg), p 209-217.

Collados-Lara A-J, Baena-Ruiz L, Pulido-Velazquez D, Pardo-Igúzquiza E (2022) Data-driven mapping of hourly wind speed and its potential energy resources: A sensitivity analysis. *Renewable Energy* 199:87-102.

Costoya X, Rocha A, Carvalho D (2020) Using bias-correction to improve future projections of offshore wind energy resource: A case study on the Iberian Peninsula. *Applied Energy* 262:114562.

Council GWE (2022) *GWEC Global Wind Report 2022.* Global Wind Energy Council, Brussels, Belgiump

Crippa P, Alifa M, Bolster D, Genton MG, Castruccio S (2021) A temporal model for vertical extrapolation of wind speed and wind energy assessment. *Applied Energy* 301:117378.

Cui X, Dong Z, Sun H, Li C, Xiao F, Liu Z, Song S, Li X, Xiao N, Xiao W (2018) Spatial and temporal variation of the near-surface wind environment in the dune fields of northern China. *International Journal of Climatology* 38(5):2333-2351.

Dao T, Fu D, Ermon S, Rudra A, Ré C (2022) Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems* 35:16344-16359.

Davidson MR, Millstein D (2022) Limitations of reanalysis data for wind power applications. *Wind Energy* 25(9):1646-1653.

Davis NN, Badger J, Hahmann AN, Hansen BO, Mortensen NG, Kelly M, Larsén XG, Olsen BT, Floors R, Lizcano G, Casso P, Lacave O, Bosch A, Bauwens I, Knight OJ, Potter van Loon A, Fox R, Parvanyan T, Krohn Hansen SB, Heathfield D, Onninen M, Drummond R (2023) The Global Wind Atlas: A High-Resolution Dataset of Climatologies and Associated Web-Based Application. *Bulletin of the American Meteorological Society* 104(8):E1507-E1525.

de Aquino Ferreira SC, Cyrino Oliveira FL, Maçaira PM (2022) Validation of the representativeness of wind speed time series obtained from reanalysis data for Brazilian territory. *Energy* 258:124746.

de Souza Ferreira GW, Reboita MS, Ribeiro JGM, Carvalho VSB, Santiago MEV, Silva PLLS, Baldoni TC, de Souza CA (2024) Assessment of the wind power density over South America simulated by CMIP6 models in the present and future climate. *Climate Dynamics* 62(3):1729-1763.

Deo RC, Ghorbani MA, Samadianfard S, Maraseni T, Bilgili M, Biazar M (2018) Multi-layer perceptron hybrid model integrated with the firefly optimizer algorithm for windspeed prediction of target site using a limited set of neighboring reference station data. *Renewable Energy* 116:309-323.

Díaz S, Carta JA, Matías JM (2017) Comparison of several measure-correlate-predict models using support vector regression techniques to estimate wind power densities. A case study. *Energy Conversion and Management* 140:334-354.

Díaz S, Carta JA, Matías JM (2018) Performance assessment of five MCP models proposed for the estimation of long-term wind turbine power outputs at a target site using three machine learning techniques. *Applied Energy* 209:455-477.

Ding C, Hanchuan P (2005) Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol* 3(2):185-205.

Dinler A (2013) A new low-correlation MCP (measure-correlate-predict) method for wind energy forecasting. *Energy* 63:152-160.

Director H, Bornn L (2015) Connecting Point-Level and Gridded Moments in the Analysis of Climate Data. *Journal of Climate* 28(9):3496-3510.

Dong Z, Wang X, Zhu R, Dong X, Ai X (2022) Improving the accuracy of wind speed statistical analysis and wind energy utilization in the Ningxia Autonomous Region, China. *Applied Energy* 320:119256.

Dörenkämper M, Olsen BT, Witha B, Hahmann AN, Davis NN, Barcons J, Ezber Y, García-Bustamante E, González-Rouco JF, Navarro J, Sastre-Marugán M, Sīle T, Trei W, Žagar M, Badger J, Gottschall J, Sanz Rodrigo J, Mann J (2020) The Making of the New European Wind Atlas – Part 2: Production and evaluation. *Geosci. Model Dev.* 13(10):5079-5102.

Dragomiretskiy K, Zosso D (2014) Variational Mode Decomposition. *IEEE Transactions on Signal Processing* 62(3):531-544.

Duan T, Anand A, Ding DY, Thai KK, Basu S, Ng A, Schuler A (2020) Ngboost: Natural gradient boosting for probabilistic prediction. *International Conference on Machine Learning.* PMLR, p 2690-2700.

Dujardin J, Lehning M (2022) Wind-Topo: Downscaling near-surface wind fields to high-resolution topography in highly complex terrain with deep learning. *Quarterly Journal of the royal meteorological society* 148(744):1368-1388.

Dupuy F, Durand P, Hedde T (2023) Downscaling of surface wind forecasts using convolutional neural networks. *Nonlin. Processes Geophys.* 30(4):553-570.

Effenberger N, Ludwig N, White RH (2024) Mind the (spectral) gap: how the temporal resolution of wind data affects multi-decadal wind power forecasts. *Environmental Research Letters* 19(1):014015.

El-kenawy E-SM, Khodadadi N, Mirjalili S, Abdelhamid AA, Eid MM, Ibrahim A (2024) Greylag Goose Optimization: Nature-inspired optimization algorithm. *Expert Systems with Applications* 238:122147.

Elsaraiti M, Merabet A (2021) A Comparative Analysis of the ARIMA and LSTM Predictive Models and Their Effectiveness for Predicting Wind Speed. *Energies* 14(20).

Emeis S (2018) *Wind energy meteorology: atmospheric physics for wind power generation.* Springer International Publishing AG, Berlin, Germany. 198 p

Environnement et Changement climatique Canada (2025) *Au sujet des données/Glossaire.* https://climat.meteo.gc.ca/glossary_f.html

Eseye AT, Lehtonen M, Tukia T, Uimonen S, Millar RJ (2019) Machine Learning Based Integrated Feature Selection Approach for Improved Electricity Demand Forecasting in Decentralized Energy Systems. *IEEE Access* 7:91463-91475.

Etienne C, Lehmann A, Goyette S, Lopez-Moreno J-I, Beniston M (2010) Spatial Predictions of Extreme Wind Speeds over Switzerland Using Generalized Additive Models. *Journal of Applied Meteorology and Climatology* 49(9):1956-1970.

Eyring V, Bony S, Meehl GA, Senior CA, Stevens B, Stouffer RJ, Taylor KE (2016) Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.* 9(5):1937-1958.

Fadare DA (2010) The application of artificial neural networks to mapping of wind speed profile for energy application in Nigeria. *Applied Energy* 87(3):934-942.

Fan J, Ma X, Wu L, Zhang F, Yu X, Zeng W (2019) Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. *Agricultural Water Management* 225:105758.

Fan W, Liu Y, Chappell A, Dong L, Xu R, Ekström M, Fu T-M, Zeng Z (2021) Evaluation of Global Reanalysis Land Surface Wind Speed Trends to Support Wind Energy Development Using In Situ Observations. *Journal of Applied Meteorology and Climatology* 60(1):33-50.

Fawad M, Ahmad I, Nadeem FA, Yan T, Abbas A (2018) Estimation of wind speed using regional frequency analysis based on linear-moments. *International Journal of Climatology* 38(12):4431-4444.

Feurer M, Hutter F (2019) Hyperparameter Optimization. *Automated Machine Learning: Methods, Systems, Challenges*, Hutter F, Kotthoff L,Vanschoren J (Édit.) Springer International Publishing, Cham10.1007/978-3-030-05318-5_1. p 3-33.

Fick SE, Hijmans RJ (2017) WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* 37(12):4302-4315.

Fischer PF, Etienne C, Tian J, Krauss T (2015) Prediction of Wind Speeds based on Digital Elevation Models using Boosted Regression Trees. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XL-1/W5:197-202.

Florinsky IV (2017) An illustrated introduction to general geomorphometry. *Progress in Physical Geography: Earth and Environment* 41(6):723-752.

Foresti L, Tuia D, Kanevski M, Pozdnoukhov A (2011) Learning wind fields with multiple kernels. *Stochastic Environmental Research and Risk Assessment* 25(1):51-66.

Fouad G, Loáiciga HA (2020) Independent variable selection for regression modeling of the flow duration curve for ungauged basins in the United States. *Journal of Hydrology* 587:124975.

Frank CW, Wahl S, Keller JD, Pospichal B, Hense A, Crewell S (2018) Bias correction of a novel European reanalysis data set for solar energy applications. *Solar Energy* 164:12-24.

Frank LE, Friedman JH (1993) A statistical view of some chemometrics regression tools. *Technometrics* 35(2):109-135.

Friedman JH (2001) Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* 29(5):1189-1232.

Fuso Nerini F, Tomei J, To LS, Bisaga I, Parikh P, Black M, Borrion A, Spataru C, Castán Broto V, Anandarajah G, Milligan B, Mulugetta Y (2018) Mapping synergies and trade-offs between energy and the Sustainable Development Goals. *Nature Energy* 3(1):10-15.

Gao X, Guo W, Mei C, Sha J, Guo Y, Sun H (2023) Short-term wind power forecasting based on SSA-VMD-LSTM. *Energy Reports* 9:335-344.

Gelaro R, McCarty W, Suárez MJ, Todling R, Molod A, Takacs L, Randles CA, Darmenov A, Bosilovich MG, Reichle R, Wargan K, Coy L, Cullather R, Draper C, Akella S, Buchard V, Conaty A, da Silva AM, Gu W, Kim G-K, Koster R, Lucchesi R, Merkova D, Nielsen JE, Partyka G, Pawson S, Putman W, Rienecker M, Schubert SD, Sienkiewicz M, Zhao B (2017) The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *Journal of Climate* 30(14):5419-5454.

Genov E, Cauwer CD, Kriekinge GV, Coosemans T, Messagie M (2024) Forecasting flexibility of charging of electric vehicles: Tree and cluster-based methods. *Applied Energy* 353:121969.

Genuer R, Poggi J-M (2017) Arbres CART et Forêts aléatoires,Importance et sélection de variables.).

Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. *Machine Learning* 63(1):3-42.

Gielen D, Gorini R, Leme R, Prakash G, Wagner N, Janeiro L, Collins S, Kadir M, Asmelash E, Ferroukhi R (2021) *World Energy Transitions Outlook: 1.5° C Pathway.* International Renewable Energy Agency (IRENA), Abu Dhabip

Gleixner S, Demissie T, Diro GT (2020) Did ERA5 Improve Temperature and Precipitation Reanalysis over East Africa? *Atmosphere* 11(9).

Global Wind Energy Council (2024) *Global Wind Report 2024.* Global Wind Energy Council, Brussels, Belgiump

Goel NK, Burn DH, Pandey MD, An Y (2004) Wind quantile estimation using a pooled frequency analysis approach. *Journal of Wind Engineering and Industrial Aerodynamics* 92(6):509-528.

Gokulnath CB, Shantharajah SP (2019) An optimized feature selection based on genetic approach and support vector machine for heart disease. *Cluster Computing* 22(6):14777-14787.

González-Aparicio I, Monforti F, Volker P, Zucker A, Careri F, Huld T, Badger J (2017) Simulating European wind power generation applying statistical downscaling to reanalysis data. *Applied Energy* 199:155-168.

González-Longatt F, Medina H, Serrano González J (2015) Spatial interpolation and orographic correction to estimate wind energy resource in Venezuela. *Renewable and Sustainable Energy Reviews* 48:1-16.

Goodfellow I, Bengio Y, Courville A (2016) *Deep learning.* MIT press,

Gottschall J, Dörenkämper M (2021) Understanding and mitigating the impact of data gaps on offshore wind resource estimates. *Wind Energ. Sci.* 6(2):505-520.

Grinsztajn L, Oyallon E, Varoquaux G (2022a) Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems* 35:507-520.

Grinsztajn L, Oyallon E, Varoquaux G (2022b) Why do tree-based models still outperform deep learning on typical tabular data? *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track.*

Grohmann CH (2015) Effects of spatial resolution on slope and aspect derivation for regional-scale analysis. *Computers & Geosciences* 77:111-117.

Grohmann CH, Smith MJ, Riccomini C (2011) Multiscale Analysis of Topographic Surface Roughness in the Midland Valley, Scotland. *IEEE Transactions on Geoscience and Remote Sensing* 49(4):1200-1213.

Gruber K, Klöckl C, Regner P, Baumgartner J, Schmidt J (2019) Assessing the Global Wind Atlas and local measurements for bias correction of wind power generation simulated from MERRA-2 in Brazil. *Energy* 189:116212.

Gruber K, Regner P, Wehrle S, Zeyringer M, Schmidt J (2022) 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* 238:121520.

Gualtieri G (2019) A comprehensive review on wind resource extrapolation models applied in wind energy. *Renewable and Sustainable Energy Reviews* 102:215-233.

Gualtieri G (2021) Reliability of ERA5 Reanalysis Data for Wind Resource Assessment: A Comparison against Tall Towers. *Energies* 14(14).

Gualtieri G (2022) Analysing the uncertainties of reanalysis data used for wind resource assessment: A critical review. *Renewable and Sustainable Energy Reviews* 167:112741.

Guijarro J (2018) *Homogenization of climatic series with Climatol.* State Meteorological Agency (AEMET), Spainp

Gutiérrez C, Molina M, Ortega M, López-Franca N, Sánchez E (2024) Low-wind climatology (1979–2018) over Europe from ERA5 reanalysis. *Climate Dynamics* 62(5):4155-4170.

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. *Journal of Machine Learning Research* 3(Mar):1157-1182.

Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* 46(1):389-422.

Hammami D, Lee TS, Ouarda TBMJ, Lee J (2012) Predictor selection for downscaling GCM data with LASSO. *Journal of Geophysical Research: Atmospheres* 117(D17).

Han Q, Ma S, Wang T, Chu F (2019) Kernel density estimation model for wind speed probability distribution with applicability to wind energy assessment in China. *Renewable and Sustainable Energy Reviews* 115:109387.

Harry J (2014) *Dependence Modeling with Copulas* Chapman and Hall/CRC, New York, 1st Edition

Hassan Q, Viktor P, J. Al-Musawi T, Mahmood Ali B, Algburi S, Alzoubi HM, Khudhair Al-Jiboory A, Zuhair Sameen A, Salman HM, Jaszczur M (2024) The renewable energy role in the global energy Transformations. *Renewable Energy Focus* 48:100545.

Hastie T, Tibshirani R, Friedman JH, Friedman JH (2009) *The elements of statistical learning: data mining, inference, and prediction.* Springer, New York

Hastie TJ (2017) Generalized additive models. *Statistical models in S,* Routledge. p 249-307.

Hdidouan D, Staffell I (2017) The impact of climate change on the levelised cost of wind energy. *Renewable Energy* 101:575-592.

He Y, Li H (2018) Probability density forecasting of wind power using quantile regression neural network and kernel density estimation. *Energy Conversion and Management* 164:374-384.

He Y, Liu R, Li H, Wang S, Lu X (2017) Short-term power load probability density forecasting method using kernel-based support vector quantile regression and Copula theory. *Applied Energy* 185:254-266.

Hengl T, Heuvelink GBM, Rossiter DG (2007) About regression-kriging: From equations to case studies. *Computers & Geosciences* 33(10):1301-1315.

Hengl T, Nussbaum M, Wright MN, Heuvelink GBM, Gräler B (2018) Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ* 6:e5518.

Hennessey JP (1978) A Comparison of the Weibull and Rayleigh Distributions for Estimating Wind Power Potential. *Wind Engineering* 2(3):156-164.

Hersbach H, Bell B, Berrisford P, Hirahara S, Horányi A, Muñoz-Sabater J, Nicolas J, Peubey C, Radu R, Schepers D, Simmons A, Soci C, Abdalla S, Abellan X, Balsamo G, Bechtold P, Biavati G, Bidlot J, Bonavita M, De Chiara G, Dahlgren P, Dee D, Diamantakis M, Dragani R, Flemming J, Forbes R, Fuentes M, Geer A, Haimberger L, Healy S, Hogan RJ, Hólm E, Janisková M, Keeley S, Laloyaux P, Lopez P, Lupu C, Radnoti G, de Rosnay P, Rozum I, Vamborg F, Villaume S, Thépaut J-N (2020) The ERA5 global reanalysis. *Quarterly Journal of the royal meteorological society* 146(730):1999-2049.

Hirukawa M (2018) *Asymmetric Kernel Smoothing: Theory and Applications in Economics and Finance.* Springer Nature Singapore, Singapore. 110 p. https://books.google.ca/books?id=cjlfDwAAQBAJ

Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Computation* 9(8):1735-1780.

Hodson TO (2022) Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geosci. Model Dev.* 15(14):5481-5487.

Höglund M (2020) *Machine Learning Methods for Spatial Interpolation of Wind*. Independent thesis Basic level (degree of Bachelor) (KTH ROYAL INSTITUTE OF TECHNOLOGY, Stockholm, Sweden). http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-275743

Höhlein K, Kern M, Hewson T, Westermann R (2020) A comparative study of convolutional neural network models for wind field downscaling. *Meteorological Applications* 27(6):e1961.

Hong HP, Ye W (2014) Estimating extreme wind speed based on regional frequency analysis. *Structural Safety* 47:67-77.

Horst R (2008) *The Weibull Distribution: A Handbook.* Chapman and Hall/CRC, New York

Hosking JRM (1990) L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. *Journal of the Royal Statistical Society: Series B (Methodological)* 52(1):105-124.

Hosking JRM, Wallis JR (1997) *Regional Frequency Analysis.* Cambridge University Press, Cambridge. 240 p. https://ui.adsabs.harvard.edu/abs/1997rfa..book.....H

Houndekindo F, Ouarda TBMJ (2023a) Comparative study of feature selection methods for wind speed estimation at ungauged locations. *Energy Conversion and Management* 291:117324.

Houndekindo F, Ouarda TBMJ (2023b) A non-parametric approach for wind speed distribution mapping. *Energy Conversion and Management* 296:117672.

Houndekindo F, Ouarda TBMJ (2023c) 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*.

Houndekindo F, Ouarda TBMJ (2024) Prediction of hourly wind speed time series at unsampled locations using machine learning. *Energy* 299:131518.

Hu W, Scholz Y, Yeligeti M, Bremen Lv, Deng Y (2023) Downscaling ERA5 wind speed data: a machine learning approach considering topographic influences. *Environmental Research Letters* 18(9):094007.

Huber PJ (1964) Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics* 35(1):73-101.

Hyndman RJ, Fan Y (1996) Sample quantiles in statistical packages. *The American Statistician* 50(4):361-365.

International Renewable Energy Agency (2022) *World Energy Transitions Outlook 2022: 1.5°C Pathway.* Abu Dhabip

Jackson PL, Mayr G, Vosper S (2013) Dynamically-Driven Winds. *Mountain Weather Research and Forecasting: Recent Progress and Current Challenges*, Chow FK, De Wekker SFJ,Snyder BJ (Édit.) Springer Netherlands, Dordrecht10.1007/978-94-007-4098-3_3. p 121-218.

Jackson PS, Hunt JCR (1975) Turbulent wind flow over a low hill. *Quarterly Journal of the royal meteorological society* 101(430):929-955.

Jakob D, Reed DW, Robson AJ (1999) *Choosing a pooling-group. Flood Estimation Handbook.* Institute of Hydrology, Wallingford, UK

Jasiewicz J, Stepinski TF (2013) Geomorphons — a pattern recognition approach to classification and mapping of landforms. *Geomorphology* 182:147-156.

Jenness J (2004) Calculating Landscape Surface Area from Digital Elevation Models. *Wildlife Society Bulletin* 32:829-839.

Jeong DI, St-Hilaire A, Ouarda TBMJ, Gachon P (2012) Multisite statistical downscaling model for daily precipitation combined by multivariate multiple linear regression and stochastic weather generator. *Climatic Change* 114(3):567-591.

Jerome HF (2001) Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 29(5):1189-1232.

Jiang W, Liu B, Liang Y, Gao H, Lin P, Zhang D, Hu G (2024) Applicability analysis of transformer to wind speed forecasting by a novel deep learning framework with multiple atmospheric variables. *Applied Energy* 353:122155.

Jordan MI, Mitchell TM (2015) Machine learning: Trends, perspectives, and prospects. *Science* 349(6245):255-260.

Jourdier B (2020) Evaluation of ERA5, MERRA-2, COSMO-REA6, NEWA and AROME to simulate wind power production over France. *Adv. Sci. Res.* 17:63-77.

Jung C (2016) High Spatial Resolution Simulation of Annual Wind Energy Yield Using Near-Surface Wind Speed Time Series. *Energies* 9(5):344.

Jung C, Demant L, Meyer P, Schindler D (2022a) Highly resolved modeling of extreme wind speed in North America and Europe. *Atmospheric Science Letters* 23(5):e1082.

Jung C, Schindler D (2015) Statistical modeling of near-surface wind speed: a case study from Baden-Wuerttemberg (Southwest Germany). *Austin Journal of Earth Science* 2(1):1006.

Jung C, Schindler D (2017a) Development of a statistical bivariate wind speed-wind shear model (WSWS) to quantify the height-dependent wind resource. *Energy Conversion and Management* 149:303-317.

Jung C, Schindler D (2017b) Global comparison of the goodness-of-fit of wind speed distributions. *Energy Conversion and Management* 133:216-234.

Jung C, Schindler D (2018a) 3D statistical mapping of Germany's wind resource using WSWS. *Energy Conversion and Management* 159:96-108.

Jung C, Schindler D (2019a) Changing wind speed distributions under future global climate. *Energy Conversion and Management* 198:111841.

Jung C, Schindler D (2019b) Wind speed distribution selection – A review of recent development and progress. *Renewable and Sustainable Energy Reviews* 114:109290.

Jung C, Schindler D (2020) Integration of small-scale surface properties in a new high resolution global wind speed model. *Energy Conversion and Management* 210:112733.

Jung C, Schindler D (2021) The role of the power law exponent in wind energy assessment: A global analysis. *International Journal of Energy Research* 45(6):8484-8496.

Jung C, Schindler D (2022b) On the influence of wind speed model resolution on the global technical wind energy potential. *Renewable and Sustainable Energy Reviews* 156:112001.

Jung C, Schindler D (2022c) A review of recent studies on wind resource projections under climate change. *Renewable and Sustainable Energy Reviews* 165:112596.

Jung C, Schindler D (2023a) Comprehensive validation of 68 wind speed models highlights the benefits of ensemble approaches. *Energy Conversion and Management* 286:117012.

Jung C, Schindler D (2023b) Introducing a new wind speed complementarity model. *Energy* 265:126284.

Jung C, Schindler D, Buchholz A, Laible J (2017c) Global Gust Climate Evaluation and Its Influence on Wind Turbines. *Energies* 10(10).

Jung C, Schindler D, Laible J (2018b) National and global wind resource assessment under six wind turbine installation scenarios. *Energy Conversion and Management* 156:403-415.

Jung C, Taubert D, Schindler D (2019c) The temporal variability of global wind energy – Long-term trends and inter-annual variability. *Energy Conversion and Management* 188:462-472.

Jung S, Kwon S-D (2013) Weighted error functions in artificial neural networks for improved wind energy potential estimation. *Applied Energy* 111:778-790.

Kang D, Ko K, Huh J (2015) Determination of extreme wind values using the Gumbel distribution. *Energy* 86:51-58.

Karpatne A, Ebert-Uphoff I, Ravela S, Babaie HA, Kumar V (2019) Machine Learning for the Geosciences: Challenges and Opportunities. *IEEE Transactions on Knowledge and Data Engineering* 31(8):1544-1554.

Kartal S, Basu S, Watson SJ (2023) A decision-tree-based measure–correlate–predict approach for peak wind gust estimation from a global reanalysis dataset. *Wind Energ. Sci.* 8(10):1533-1551.

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Neural Information Processing Systems.* Guyon I LU, Bengio S, Wallach H, Fergus R,, Vishwanathan S GR (Édit.).

Keck RE, Sondell N (2020) Validation of uncertainty reduction by using multiple transfer locations for WRF–CFD coupling in numerical wind energy assessments. *Wind Energ. Sci.* 5(3):997-1005.

Keellings D (2016) Evaluation of downscaled CMIP5 model skill in simulating daily maximum temperature over the southeastern United States. *International Journal of Climatology* 36(12):4172-4180.

Keskin M, Dalgin S, Dogru A (2017) Evaluating the Topographic Effect on Estimation of Renewable Energy Sources Potential in Sakarya Basin. *Fresenius Environmental Bulletin* 26(1):64-68.

Kim B, Lee K, Ko K, Choi J (2022) Offshore wind resource assessment off the coast of Daejeong, Jeju Island using 30-year wind estimates. *Scientific Reports* 12(1):14179.

Kim J-Y, Kim K-Y (2016) Reconstructing long-term wind data at an offshore met-mast location using cyclostationary empirical orthogonal functions. *Journal of Wind Engineering and Industrial Aerodynamics* 156:146-158.

King C, Hurley B (2005) The SpeedSort, DynaSort and Scatter Wind Correlation Methods. *Wind Engineering* 29(3):217-241.

Kingma DP, Ba JL (2015) Adam: A Method for Stochastic Optimization. in *3rd International Conference for Learning Representations*San Diego), p 1–15.

Kirchmeier MC, Lorenz DJ, Vimont DJ (2014) Statistical Downscaling of Daily Wind Speed Variations. *Journal of Applied Meteorology and Climatology* 53(3):660-675.

Klink K (2002) Trends and Interannual Variability of Wind Speed Distributions in Minnesota. *Journal of Climate* 15(22):3311-3317.

Klink K (2007) Atmospheric Circulation Effects on Wind Speed Variability at Turbine Height. *Journal of Applied Meteorology and Climatology* 46(4):445-456.

Ko K, Kang D, Huh J (2015) An Estimation of Extreme Wind Speeds on Jeju Island Using Measure-Correlate-Predict Method. *Wind Engineering* 39(3):261-269.

Koenker R (2017) Quantile Regression: 40 Years On. *Annual Review of Economics* 9(1):155-176.

Koo J, Han GD, Choi HJ, Shim JH (2015) Wind-speed prediction and analysis based on geological and distance variables using an artificial neural network: A case study in South Korea. *Energy* 93:1296-1302.

Kratzert F, Klotz D, Brenner C, Schulz K, Herrnegger M (2018) Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrol. Earth Syst. Sci.* 22(11):6005-6022.

Kristianti F, Dujardin J, Gerber F, Huwald H, Hoch SW, Lehning M (2023) Combining Weather Station Data and Short-Term LiDAR Deployment to Estimate Wind Energy Potential with Machine Learning: A Case Study from the Swiss Alps. *Boundary-Layer Meteorology* 188(1):185-208.

Kruyt B, Dujardin J, Lehning M (2018) Improvement of Wind Power Assessment in Complex Terrain: The Case of COSMO-1 in the Swiss Alps. *Frontiers in Energy Research* 6.

Kumar G, Malik H (2016a) Generalized Regression Neural Network Based Wind Speed Prediction Model for Western Region of India. *Procedia Computer Science* 93:26-32.

Kumar Y, Ringenberg J, Depuru Soma S, Devabhaktuni VK, Lee Jin W, Nikolaidis E, Andersen B, Afjeh A (2016b) Wind energy: Trends and enabling technologies. *Renewable and Sustainable Energy Reviews* 53:209-224.

Laapas M, Venäläinen A (2017) Homogenization and trend analysis of monthly mean and maximum wind speed time series in Finland, 1959–2015. *International Journal of Climatology* 37(14):4803-4813.

Lafaye de Micheaux P, Ouimet F (2021) A Study of Seven Asymmetric Kernels for the Estimation of Cumulative Distribution Functions. *Mathematics* 9(20).

Laib M, Kanevski M (2016) Spatial Modelling of Extreme Wind Speed Distributions in Switzerland. *Energy Procedia* 97:100-107.

Lamb WF, Wiedmann T, Pongratz J, Andrew R, Crippa M, Olivier JGJ, Wiedenhofer D, Mattioli G, Khourdajie AA, House J, Pachauri S, Figueroa M, Saheb Y, Slade R, Hubacek K, Sun L, Ribeiro SK, Khennas S, de la Rue du Can S, Chapungu L, Davis SJ, Bashmakov I, Dai H, Dhakal S, Tan X, Geng Y, Gu B, Minx J (2021) A review of trends and drivers of greenhouse gas emissions by sector from 1990 to 2018. *Environmental Research Letters* 16(7):073005.

Lambert T, Grue A (2012) The matrix time series method for MCP. *Proceedings of the AWEA Windpower Conference and Exhibition, Atlanta, GA, USA.* p 3-6.

Landberg L, Mortensen N (1993) A comparison of physical and statistical methods for estimating the wind resource at a site. *Proc. BWEA* 15.

Langer J, Zaaijer M, Quist J, Blok K (2023) Introducing site selection flexibility to technical and economic onshore wind potential assessments: New method with application to Indonesia. *Renewable Energy* 202:320-335.

Lapen DR, Martz LW (1993) The measurement of two simple topographic indices of wind sheltering-exposure from raster digital elevation models. *Computers & Geosciences* 19(6):769-779.

Largeron Y, Guichard F, Bouniol D, Couvreux F, Kergoat L, Marticorena B (2015) Can we use surface wind fields from meteorological reanalyses for Sahelian dust emission simulations? *Geophysical Research Letters* 42(7):2490-2499.

Latifovic R, Pouliot D, Olthof I (2017) Circa 2010 Land Cover of Canada: Local Optimization Methodology and Product Development. *Remote Sensing* 9(11):1098.

Lawan SM, Abidin WAWZ, Lawan S, Lawan AM (2016) An Artificial Intelligence Strategy for the Prediction of Wind Speed and Direction in Sarawak for Wind Energy Mapping. in *Recent Advances in Mathematical Sciences*, Édit Kılıçman A, Srivastava, H., Mursaleen, M., Abdul Majid, Z (Springer Singapore, Singapore), p 71-82.

Leardi R, Boggia R, Terrile M (1992) Genetic algorithms as a strategy for feature selection. *Journal of Chemometrics* 6(5):267-281.

Leblanc M, Schoborg D, Cox S, Haché A, Tindal A (2009) Is a Non-linear MCP method a useful tool for North American wind regimes. *Proceedings of the AWEA 2009 Windpower Conference and Exhibition, Chicago, IL, USA.*

Lee C (2022) Long-term wind speed interpolation using anisotropic regression kriging with regional heterogeneous terrain and solar insolation in the United States. *Energy Reports* 8:12-23.

Lee J, Wang W, Harrou F, Sun Y (2020) Reliable solar irradiance prediction using ensemble learning-based models: A comparative study. *Energy Conversion and Management* 208:112582.

Lee JCY, Fields MJ (2021) An overview of wind-energy-production prediction bias, losses, and uncertainties. *Wind Energ. Sci.* 6(2):311-365.

Lee K, Kim E, Ko K (2019) Feasibility study of energy self-sufficiency on Chuja island, Korea using wind energy. *International Journal of Renewable Energy Research* 9(3):1224-1231.

Li D, Feng J, Dosio A, Qi J, Xu Z, Yin B (2020) Historical Evaluation and Future Projections of 100-m Wind Energy Potentials Over CORDEX-East Asia. *Journal of Geophysical Research: Atmospheres* 125(15):e2020JD032874.

Li D, Feng J, Xu Z, Yin B, Shi H, Qi J (2019) Statistical Bias Correction for Simulated Wind Speeds Over CORDEX-East Asia. *Earth and Space Science* 6(2):200-211.

Li J, Wang X, Yu X (2018) Use of spatio-temporal calibrated wind shear model to improve accuracy of wind resource assessment. *Applied Energy* 213:469-485.

Li T, Zheng X, Dai Y, Yang C, Chen Z, Zhang S, Wu G, Wang Z, Huang C, Shen Y, Liao R (2014) Mapping near-surface air temperature, pressure, relative humidity and wind speed over Mainland China with high spatiotemporal resolution. *Advances in Atmospheric Sciences* 31(5):1127-1135.

Liemohn MW, Shane AD, Azari AR, Petersen AK, Swiger BM, Mukhopadhyay A (2021) RMSE is not enough: Guidelines to robust data-model comparisons for magnetospheric physics. *Journal of Atmospheric and Solar-Terrestrial Physics* 218:105624.

Lim B, Arık SÖ, Loeff N, Pfister T (2021a) Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting* 37(4):1748-1764.

Lim B, Zohren S (2021b) Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379(2194):20200209.

Lin H, Tang J, Wang S, Wang S, Dong G (2023) Deep learning downscaled high-resolution daily near surface meteorological datasets over East Asia. *Scientific Data* 10(1):890.

Lindsay JB (2014) The Whitebox Geospatial Analysis Tools project and open-access GIS. *GIS Research UK 22nd Annual Conference.* (University of Glasgow, 16-18 April) The University of Glasgow.

Lins DR, Guedes KS, Pitombeira-Neto AR, Rocha PAC, de Andrade CF (2023) Comparison of the performance of different wind speed distribution models applied to onshore and offshore wind speed data in the Northeast Brazil. *Energy* 278:127787.

Liu C, Zhang X, Mei S, Zhen Z, Jia M, Li Z, Tang H (2022) Numerical weather prediction enhanced wind power forecasting: Rank ensemble and probabilistic fluctuation awareness. *Applied Energy* 313:118769.

Liu Y, Wu Y (2011) Simultaneous multiple non-crossing quantile regression estimation using kernel constraints. *Journal of Nonparametric Statistics* 23(2):415-437.

Lopez A, Mai T, Lantz E, Harrison-Atlas D, Williams T, Maclaurin G (2021) Land use and turbine technology influences on wind potential in the United States. *Energy* 223:120044.

Lorenz T, Barstad I (2016) A dynamical downscaling of ERA-Interim in the North Sea using WRF with a 3 km grid—for wind resource applications. *Wind Energy* 19(10):1945-1959.

Lowe DG (2004) Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60(2):91-110.

Luo W, Taylor MC, Parker SR (2008) 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(7):947-959.

Luzia G, Koivisto MJ, Hahmann AN (2023) Validating EURO-CORDEX climate simulations for modelling European wind power generation. *Renewable Energy* 217:118989.

Manuca R, Savit R (1996) Stationarity and nonstationarity in time series analysis. *Physica D: Nonlinear Phenomena* 99(2):134-161.

Manwell JF, McGowan JG, Rogers AL (2010) *Wind energy explained: theory, design and application.* John Wiley & Sons,

Martinez A, Iglesias G (2024) Global wind energy resources decline under climate change. *Energy* 288:129765.

Maxwell AE, Shobe CM (2022) Land-surface parameters for spatial predictive mapping and modeling. *Earth-Science Reviews* 226:103944.

McKenna R, Ostman v.d. Leye P, Fichtner W (2016) Key challenges and prospects for large wind turbines. *Renewable and Sustainable Energy Reviews* 53:1212-1221.

McKenna R, Pfenninger S, Heinrichs H, Schmidt J, Staffell I, Bauer C, Gruber K, Hahmann AN, Jansen M, Klingler M, Landwehr N, Larsén XG, Lilliestam J, Pickering B, Robinius M, Tröndle T, Turkovska O, Wehrle S, Weinand JM, Wohland J (2022) High-resolution large-scale onshore wind energy assessments: A review of potential definitions, methodologies and future research needs. *Renewable Energy* 182:659-684.

McVicar TR, Van Niel TG, Li LT, Roderick ML, Rayner DP, Ricciardulli L, Donohue RJ (2008) Wind speed climatology and trends for Australia, 1975–2006: Capturing the stilling phenomenon and comparison with near-surface reanalysis output. *Geophysical Research Letters* 35(20).

Mehta D, van Zuijlen AH, Koren B, Holierhoek JG, Bijl H (2014) Large Eddy Simulation of wind farm aerodynamics: A review. *Journal of Wind Engineering and Industrial Aerodynamics* 133:1-17.

Meinshausen N, Ridgeway G (2006) Quantile regression forests. *Journal of Machine Learning Research* 7(6).

Mestre O, Domonkos P, Picard F, Auer I, Robin S, Lebarbier E, Boehm R, Aguilar E, Guijarro J, Vertachnik G, Klančar M, Dubuisson B, Stepanek P (2013) HOMER: A homogenization software - methods and applications. *Idojaras* 117(1):47-67.

Miao H, Dong D, Huang G, Hu K, Tian Q, Gong Y (2020) Evaluation of Northern Hemisphere surface wind speed and wind power density in multiple reanalysis datasets. *Energy* 200:117382.

Michelangeli PA, Vrac M, Loukos H (2009) Probabilistic downscaling approaches: Application to wind cumulative distribution functions. *Geophysical Research Letters* 36(11).

Mifsud MD, Sant T, Farrugia RN (2018) A comparison of Measure-Correlate-Predict Methodologies using LiDAR as a candidate site measurement device for the Mediterranean Island of Malta. *Renewable Energy* 127:947-959.

Mifsud MD, Sant T, Farrugia RN (2020) Analysing uncertainties in offshore wind farm power output using measure–correlate–predict methodologies. *Wind Energ. Sci.* 5(2):601-621.

Miguel JVP, Fadigas EA, Sauer IL (2019) The Influence of the Wind Measurement Campaign Duration on a Measure-Correlate-Predict (MCP)-Based Wind Resource Assessment. *Energies* 12(19):3606.

Millstein D, Solomon-Culp J, Wang M, Ullrich P, Collier C (2019) Wind energy variability and links to regional and synoptic scale weather. *Climate Dynamics* 52(7):4891-4906.

Minola L, Azorin-Molina C, Chen D (2016) Homogenization and Assessment of Observed Near-Surface Wind Speed Trends across Sweden, 1956–2013. *Journal of Climate* 29(20):7397-7415.

Modarres R (2008) Regional maximum wind speed frequency analysis for the arid and semi-arid regions of Iran. *Journal of Arid Environments* 72(7):1329-1342.

Moemken J, Reyers M, Feldmann H, Pinto JG (2018) Future Changes of Wind Speed and Wind Energy Potentials in EURO-CORDEX Ensemble Simulations. *Journal of Geophysical Research: Atmospheres* 123(12):6373-6389.

Molina MO, Gutiérrez C, Sánchez E (2021) Comparison of ERA5 surface wind speed climatologies over Europe with observations from the HadISD dataset. *International Journal of Climatology* 41(10):4864-4878.

Molod A, Takacs L, Suarez M, Bacmeister J (2015) Development of the GEOS-5 atmospheric general circulation model: evolution from MERRA to MERRA2. *Geosci. Model Dev.* 8(5):1339-1356.

Mombeni HA, Mansouri B, Akhoond M (2021) Asymmetric kernels for boundary modification in distribution function estimation. *REVSTAT-Statistical Journal* 19(4):463–484-463–484.

Mona Ahmed Yassen, Mohamed Gamal Abdel-Fattah, Islam Ismail, EL-Sayed M. El Kenawy, Moustafa HE-D (2024) An AI-Based System for Predicting Renewable Energy Power Output Using Advanced Optimization Algorithms. *Journal of Artificial Intelligence and Metaheuristics* 8(1):01-08, 2024.

Moradian S, Olbert AI, Gharbia S, Iglesias G (2023) Copula-based projections of wind power: Ireland as a case study. *Renewable and Sustainable Energy Reviews* 175:113147.

Morales-Ruvalcaba CF, Rodríguez-Hernández O, Martínez-Alvarado O, Drew DR, Ramos E (2020) Estimating wind speed and capacity factors in Mexico using reanalysis data. *Energy for Sustainable Development* 58:158-166.

Morgan EC, Lackner M, Vogel RM, Baise LG (2011) Probability distributions for offshore wind speeds. *Energy Conversion and Management* 52(1):15-26.

Muhammad SL, Abidin WAWZ, Chai WY, Baharun A, Masri T (2014) Development of Wind Mapping Based on Artificial Neural Network (ANN) for Energy Exploration in Sarawak. *International Journal of Renewable Energy Research* 4:618-627.

Murcia JP, Koivisto MJ, Luzia G, Olsen BT, Hahmann AN, Sørensen PE, Als M (2022) Validation of European-scale simulated wind speed and wind generation time series. *Applied Energy* 305:117794.

Nabipour N, Mosavi A, Hajnal E, Nadai L, Shamshirband S, Chau K-W (2020) Modeling climate change impact on wind power resources using adaptive neuro-fuzzy inference system. *Engineering Applications of Computational Fluid Mechanics* 14(1):491-506.

Naizghi MS, Ouarda TBMJ (2017) Teleconnections and analysis of long-term wind speed variability in the UAE. *International Journal of Climatology* 37(1):230-248.

Nascimento EGS, de Melo TAC, Moreira DM (2023) A transformer-based deep neural network with wavelet transform for forecasting wind speed and wind energy. *Energy* 278:127678.

Nasri B, Bouezmarni T, St-Hilaire A, Ouarda TBMJ (2017) Non-stationary hydrologic frequency analysis using B-spline quantile regression. *Journal of Hydrology* 554:532-544.

Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics* Volume 7 - 2013.

Nefabas KL, Söder L, Mamo M, Olauson J (2021) Modeling of Ethiopian Wind Power Production Using ERA5 Reanalysis Data. *Energies* 14(9).

Nelson V, Starcher Kre (2018) *Wind Energy: Renewable Energy and the Environment.* CRC Press, Boca raton, Floride USA

Németh CP, Bartholy J, Pongrácz R (2014) Homogenization of Hungarian daily wind speed data series. *Idojaras* 118:119-132.

Newman DR, Cockburn JMH, Drăguţ L, Lindsay JB (2022) Evaluating Scaling Frameworks for Multiscale Geomorphometric Analysis. *Geomatics* 2(1):36-51.

Niermann D, Borsche M, Kaiser-Weiss Andrea K, Kaspar F (2019) Evaluating renewable-energy-relevant parameters of COSMO-REA6 by comparison with satellite data, station observations and other reanalyses. *Meteorologische Zeitschrift* 28(4):347-360.

Olauson J (2018) ERA5: The new champion of wind power modelling? *Renewable Energy* 126:322-331.

Olivier CP, Sebastian E (2024) Neural networks for extreme quantile regression with an application to forecasting of flood risk. *The Annals of Applied Statistics* 18(4):2818-2839.

Ouali D, Chebana F, Ouarda TBMJ (2016) Quantile Regression in Regional Frequency Analysis: A Better Exploitation of the Available Information. *Journal of Hydrometeorology* 17.

Ouarda TBMJ, Charron C (2018) On the mixture of wind speed distribution in a Nordic region. *Energy Conversion and Management* 174:33-44.

Ouarda TBMJ, Charron C (2021) Non-stationary statistical modelling of wind speed: A case study in eastern Canada. *Energy Conversion and Management* 236:114028.

Ouarda TBMJ, Charron C, Chebana F (2016) 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* 124:247-265.

Ouarda TBMJ, Charron C, Shin JY, Marpu PR, Al-Mandoos AH, Al-Tamimi MH, Ghedira H, Al Hosary TN (2015) Probability distributions of wind speed in the UAE. *Energy Conversion and Management* 93:414-434.

Ouarda TBMJ, Charron C, St-Hilaire A (2022) Regional estimation of river water temperature at ungauged locations. *Journal of Hydrology X* 10.1016/j.hydroa.2022.100133.

Öztopal A (2006) Artificial neural network approach to spatial estimation of wind velocity data. *Energy Conversion and Management* 47(4):395-406.

Palomino I, Martín F (1995) A Simple Method for Spatial Interpolation of the Wind in Complex Terrain. *Journal of Applied Meteorology and Climatology* 34(7):1678-1693.

Palutikof JP, Brabson BB, Lister DH, Adcock ST (1999) A review of methods to calculate extreme wind speeds. *Meteorological Applications* 6(2):119-132.

Park J, Moon J, Jung S, Hwang E (2020) Multistep-Ahead Solar Radiation Forecasting Scheme Based on the Light Gradient Boosting Machine: A Case Study of Jeju Island. *Remote Sensing* 12(14):2271.

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (2019) PyTorch: an imperative style, high-performance deep learning library. *Proceedings of the 33rd International Conference on Neural Information Processing Systems,* Curran Associates Inc. p Article 721.

Paul HCE, Marx BD (1996) Flexible Smoothing with B-splines and Penalties. *Statistical Science* 11(2):89-102.

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

Pedruzzi R, Silva AR, Soares dos Santos T, Araujo AC, Cotta Weyll AL, Lago Kitagawa YK, Nunes da Silva Ramos D, Milani de Souza F, Almeida Narciso MV, Saraiva Araujo ML, Medrado RC, Camilo Júnior WO, Neto AT, de Carvalho M, Pires Bezerra WR, Costa TT, Bione de Melo Filho J, Bandeira Santos AÁ, Moreira DM (2023) Review of mapping analysis and complementarity between solar and wind energy sources. *Energy* 283:129045.

Pelser T, Weinand JM, Kuckertz P, McKenna R, Linssen J, Stolten D (2024) Reviewing accuracy & reproducibility of large-scale wind resource assessments. *Advances in Applied Energy* 13:100158.

Pennock DJ, Zebarth BJ, De Jong E (1987) Landform classification and soil distribution in hummocky terrain, Saskatchewan, Canada. *Geoderma* 40(3):297-315.

Perkins SE, Pitman AJ, Holbrook NJ, McAneney J (2007) Evaluation of the AR4 Climate Models' Simulated Daily Maximum Temperature, Minimum Temperature, and Precipitation over Australia Using Probability Density Functions. *Journal of Climate* 20(17):4356-4376.

Petersen EL, Mortensen NG, Landberg L, Højstrup J, Frank HP (1998) Wind power meteorology. Part II: siting and models. *Wind Energy* 1(2):55-72.

Petersen EL, Troen I, Frandsen S, Hedegaard K (1981) *Danish Windatlas: A Rational Method of Wind Energy Siting.* Riso National Laboratory. https://books.google.ca/books?id=ru2XoAEACAAJ

Pfenninger S, Hawkes A, Keirstead J (2014) Energy systems modeling for twenty-first century energy challenges. *Renewable and Sustainable Energy Reviews* 33:74-86.

Philippopoulos K, Deligiorgi D (2012) Application of artificial neural networks for the spatial estimation of wind speed in a coastal region with complex topography. *Renewable Energy* 38(1):75-82.

Potisomporn P, Adcock TAA, Vogel CR (2023) Evaluating ERA5 reanalysis predictions of low wind speed events around the UK. *Energy Reports* 10:4781-4790.

Powers JG, Klemp JB, Skamarock WC, Davis CA, Dudhia J, Gill DO, Coen JL, Gochis DJ, Ahmadov R, Peckham SE, Grell GA, Michalakes J, Trahan S, Benjamin SG, Alexander CR, Dimego GJ, Wang W, Schwartz CS, Romine GS, Liu Z, Snyder C, Chen F, Barlage MJ, Yu W, Duda MG (2017) The Weather Research and Forecasting Model: Overview, System Efforts, and Future Directions. *Bulletin of the American Meteorological Society* 98(8):1717-1737.

Pryor SC, Barthelmie RJ (2010) Climate change impacts on wind energy: A review. *Renewable and Sustainable Energy Reviews* 14(1):430-437.

Pryor SC, Barthelmie RJ (2011) Assessing climate change impacts on the near-term stability of the wind energy resource over the United States. *Proceedings of the National Academy of Sciences* 108(20):8167-8171.

Pryor SC, Barthelmie RJ (2021) A global assessment of extreme wind speeds for wind energy applications. *Nature Energy* 6(3):268-276.

Pryor SC, Barthelmie RJ, Bukovsky MS, Leung LR, Sakaguchi K (2020) Climate change impacts on wind power generation. *Nature Reviews Earth & Environment* 1(12):627-643.

Pryor SC, Barthelmie RJ, Schoof JT (2006a) Inter-annual variability of wind indices across Europe. *Wind Energy* 9(1-2):27-38.

Pryor SC, Barthelmie RJ, Schoof JT (2012) Past and future wind climates over the contiguous USA based on the North American Regional Climate Change Assessment Program model suite. *Journal of Geophysical Research: Atmospheres* 117(D19).

Pryor SC, Barthelmie RJ, Young DT, Takle ES, Arritt RW, Flory D, Gutowski Jr. WJ, Nunes A, Roads J (2009) Wind speed trends over the contiguous United States. *Journal of Geophysical Research: Atmospheres* 114(D14).

Pryor SC, Schoof JT, Barthelmie RJ (2006b) Winds of change?: Projections of near-surface winds under climate change scenarios. *Geophysical Research Letters* 33(11).

Pryor SC, Shepherd TJ, Barthelmie RJ (2018) Interannual variability of wind climates and wind turbine annual energy production. *Wind Energ. Sci.* 3(2):651-665.

Pya N, Wood SN (2015) Shape constrained additive models. *Statistics and Computing* 25(3):543-559.

Qin Z, Li W, Xiong X (2011) Estimating wind speed probability distribution using kernel density method. *Electric Power Systems Research* 81(12):2139-2146.

Qiu R, Liu C, Cui N, Gao Y, Li L, Wu Z, Jiang S, Hu M (2022) Generalized Extreme Gradient Boosting model for predicting daily global solar radiation for locations without historical data. *Energy Conversion and Management* 258:115488.

Rabbani R, Zeeshan M (2020) Exploring the suitability of MERRA-2 reanalysis data for wind energy estimation, analysis of wind characteristics and energy potential assessment for selected sites in Pakistan. *Renewable Energy* 154:1240-1251.

Rajczak J, Kotlarski S, Salzmann N, Schär C (2016) Robust climate scenarios for sites with sparse observations: a two-step bias correction approach. *International Journal of Climatology* 36(3):1226-1243.

Ramon J, Lledó L, Pérez-Zanón N, Soret A, Doblas-Reyes FJ (2020) The Tall Tower Dataset: a unique initiative to boost wind energy research. *Earth Syst. Sci. Data* 12(1):429-439.

Ramon J, Lledó L, Torralba V, Soret A, Doblas-Reyes FJ (2019) What global reanalysis best represents near-surface winds? *Quarterly Journal of the royal meteorological society* 145(724):3236-3251.

Raupach MR, Finnigan JJ (1997) The influence of topography on meteorogical variables and surface-atmosphere interactions. *Journal of Hydrology* 190(3):182-213.

Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N, Prabhat (2019) Deep learning and process understanding for data-driven Earth system science. *Nature* 566(7743):195-204.

Reinhardt K, Samimi C (2018) Comparison of different wind data interpolation methods for a region with complex terrain in Central Asia. *Climate Dynamics* 51(9):3635-3652.

Reiss RD (1989) *Approximate Distributions of Order Statistics: With Applications to Nonparametric Statistics.* Springer New York. XII, 355 p. https://books.google.ca/books?id=st7TBwAAQBAJ

Ren G, Liu J, Wan J, Guo Y, Yu D (2017) Overview of wind power intermittency: Impacts, measurements, and mitigation solutions. *Applied Energy* 204:47-65.

Ren K, Fang W, Qu J, Zhang X, Shi X (2020) Comparison of eight filter-based feature selection methods for monthly streamflow forecasting – Three case studies on CAMELS data sets. *Journal of Hydrology* 586:124897.

Requena AI, Ouarda TBMJ, Chebana F (2017) Flood Frequency Analysis at Ungauged Sites Based on Regionally Estimated Streamflows. *Journal of Hydrometeorology* 18(9):2521-2539.

Riley SJ, DeGloria SD, Elliot R (1999) Index that quantifies topographic heterogeneity. *intermountain Journal of sciences* 5(1-4):23-27.

Robert S, Foresti L, Kanevski M (2013) Spatial prediction of monthly wind speeds in complex terrain with adaptive general regression neural networks. *International Journal of Climatology* 33(7):1793-1804.

Roderick ML, Rotstayn LD, Farquhar GD, Hobbins MT (2007) On the attribution of changing pan evaporation. *Geophysical Research Letters* 34(17).

Rodriguez-Galiano VF, Luque-Espinar JA, Chica-Olmo M, Mendes MP (2018) Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods. *Science of The Total Environment* 624:661-672.

Rogers AL, Rogers JW, Manwell JF (2005a) Comparison of the performance of four measure–correlate–predict algorithms. *Journal of Wind Engineering and Industrial Aerodynamics* 93(3):243-264.

Rogers AL, Rogers JW, Manwell JF (2005b) Uncertainties in results of measure-correlate-predict analyses. *American Wind Energy Association*.

Ryberg DS, Caglayan DG, Schmitt S, Linßen J, Stolten D, Robinius M (2019) The future of European onshore wind energy potential: Detailed distribution and simulation of advanced turbine designs. *Energy* 182:1222-1238.

Saarnak E, Bergström H, Söderberg S (2014) Uncertainties Connected to Long-term Correction of Wind Observations. *Wind Engineering* 38(3):233-248.

Safari B (2011) Modeling wind speed and wind power distributions in Rwanda. *Renewable and Sustainable Energy Reviews* 15(2):925-935.

Şahin AD, Dincer I, Rosen MA (2006) Thermodynamic analysis of wind energy. *International Journal of Energy Research* 30(8):553-566.

Sailor DJ, Smith M, Hart M (2008) Climate change implications for wind power resources in the Northwest United States. *Renewable Energy* 33(11):2393-2406.

Salehi Borujeni M, Dideban A, Akbari Foroud A (2021) Reconstructing long-term wind speed data based on measure correlate predict method for micro-grid planning. *Journal of Ambient Intelligence and Humanized Computing* 12(11):10183-10195.

Salinas D, Flunkert V, Gasthaus J, Januschowski T (2020) DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting* 36(3):1181-1191.

Samal RK (2021) Assessment of wind energy potential using reanalysis data: A comparison with mast measurements. *Journal of Cleaner Production* 313:127933.

Sapuan MS, Razali AM, Ibrahim K (2011) Forecasting and Mapping of Extreme Wind Speed for 5 to 100-years Return Period in Peninsula Malaysia. *Australian journal of basic and applied sciences* 5:1204-1212.

Schicker I, Ganglbauer J, Dabernig M, Nacht T (2023) Wind power estimation on local scale—A case study of representativeness of reanalysis data and data-driven analysis. *Frontiers in Climate* 5.

Schoof JT (2013) Statistical Downscaling in Climatology. *Geography Compass* 7(4):249-265.

Schultz MG, Betancourt C, Gong B, Kleinert F, Langguth M, Leufen LH, Mozaffari A, Stadtler S (2021) Can deep learning beat numerical weather prediction? *Philosophical Transactions* of the Royal Society A: Mathematical, Physical and Engineering Sciences 379(2194):20200097.

Schwegmann S, Faulhaber J, Pfaffel S, Yu Z, Dörenkämper M, Kersting K, Gottschall J (2023) Enabling Virtual Met Masts for wind energy applications through machine learning-methods. *Energy and AI* 11:100209.

Seidou O, Asselin JJ, Ouarda TBMJ (2007a) Bayesian multivariate linear regression with application to change point models in hydrometeorological variables. *Water Resources Research* 43(8).

Seidou O, Ouarda TBMJ (2007b) Recursion-based multiple changepoint detection in multiple linear regression and application to river streamflows. *Water Resources Research* 43(7).

Sekulić A, Kilibarda M, Heuvelink GBM, Nikolić M, Bajat B (2020) Random Forest Spatial Interpolation. *Remote Sensing* 12(10).

Shang Y, Sang S, Tiwari AK, Khan S, Zhao X (2024) Impacts of renewable energy on climate risk: A global perspective for energy transition in a climate adaptation framework. *Applied Energy* 362:122994.

Shen C, Zha J, Li Z, Azorin-Molina C, Deng K, Minola L, Chen D (2022) Evaluation of global terrestrial near-surface wind speed simulated by CMIP6 models and their future projections. *Annals of the New York Academy of Sciences* 1518(1):249-263.

Shi H, Dong Z, Xiao N, Huang Q (2021) Wind Speed Distributions Used in Wind Energy Assessment: A Review. *Frontiers in Energy Research* 9.

Shin J-Y, Ouarda TBMJ, Lee T (2016) Heterogeneous mixture distributions for modeling wind speed, application to the UAE. *Renewable Energy* 91:40-52.

Shu C, Ouarda TBMJ (2012) Improved methods for daily streamflow estimates at ungauged sites. *Water Resources Research* 48(2).

Shwartz-Ziv R, Armon A (2022) Tabular data: Deep learning is not all you need. *Information Fusion* 81:84-90.

Si P, Luo C, Liang D (2018) Homogenization of Tianjin monthly near-surface wind speed using RHtestsV4 for 1951–2014. *Theoretical and Applied Climatology* 132(3):1303-1320.

Simões T, Estanqueiro A (2016) A new methodology for urban wind resource assessment. *Renewable Energy* 89:598-605.

Sliz-Szkliniarz B, Vogt J (2011) GIS-based approach for the evaluation of wind energy potential: A case study for the Kujawsko-Pomorskie Voivodeship. *Renewable and Sustainable Energy Reviews* 15:1696-1707.

Smith G (2018) Step away from stepwise. *Journal of Big Data* 5(1):32.

Staffell I, Pfenninger S (2016) Using bias-corrected reanalysis to simulate current and future wind power output. *Energy* 114:1224-1239.

Štěpánek P, Zahradníček P, Skalák P (2009) Data quality control and homogenization of air temperature and precipitation series in the area of the Czech Republic in the period 1961–2007. *Adv. Sci. Res.* 3(1):23-26.

Sun S, Liu Y, Li Q, Wang T, Chu F (2023a) Short-term multi-step wind power forecasting based on spatio-temporal correlations and transformer neural networks. *Energy Conversion and Management* 283:116916.

Sun Y, Zhu D, Li Y, Wang R, Ma R (2023b) 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* 289:117198.

Sweeney C, Bessa RJ, Browell J, Pinson P (2020) The future of forecasting for renewable energy. *WIREs Energy and Environment* 9(2):e365.

Szentimrey T (1999) Multiple Analysis of Series for Homogenization (MASH). in *Second Seminar for Homogenization of Surface Climatological Data*, Édit WmoGeneva, switzerland), p 27-46.

Tadono T, Ishida H, Oda F, Naito S, Minakawa K, Iwamoto H (2014) Precise Global DEM Generation by ALOS PRISM. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* II4:71-76.

Tang X-Y, Zhao S, Fan B, Peinke J, Stoevesandt B (2019) Micro-scale wind resource assessment in complex terrain based on CFD coupled measurement from multiple masts. *Applied Energy* 238:806-815.

Thé J, Yu H (2017) A critical review on the simulations of wind turbine aerodynamics focusing on hybrid RANS-LES methods. *Energy* 138:257-289.

Thøgersen ML, Motta M, Sørensen T, Nielsen P (2007) Measure-correlate-predict methods: case studies and software implementation. *European Wind Energy Conference & Exhibition.* p 10.

Thomas SR, Nicolau S, Martínez-Alvarado O, Drew DJ, Bloomfield HC (2021) How well do atmospheric reanalyses reproduce observed winds in coastal regions of Mexico? *Meteorological Applications* 28(5):e2023.

Tibshirani R (1996) Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1):267-288.

Tobin I, Jerez S, Vautard R, Thais F, van Meijgaard E, Prein A, Déqué M, Kotlarski S, Maule CF, Nikulin G, Noël T, Teichmann C (2016) Climate change impacts on the power generation potential of a European mid-century wind farms scenario. *Environmental Research Letters* 11(3):034013.

Torralba V, Doblas-Reyes FJ, Gonzalez-Reviriego N (2017) Uncertainty in recent near-surface wind speed trends: a global reanalysis intercomparison. *Environmental Research Letters* 12(11):114019.

Tsvetkova O, Ouarda TBMJ (2023) Use of the Halphen distribution family for mean wind speed estimation with application to Eastern Canada. *Energy Conversion and Management* 276:116502.

Tuller SE, Brett AC (1984) The Characteristics of Wind Velocity that Favor the Fitting of a Weibull Distribution in Wind Speed Analysis. *Journal of Applied Meteorology and Climatology* 23(1):124-134.

Tuller SE, Brett AC (1985) The goodness of fit of the weibull and rayleigh distributions to the distributions of observed wind speeds in a topographically diverse area. *Journal of Climatology* 5(1):79-94.

Turner R, Eriksson D, McCourt M, Kiili J, Laaksonen E, Xu Z, Guyon I (2021) Bayesian Optimization is Superior to Random Search for Machine Learning Hyperparameter Tuning: Analysis of the Black-Box Optimization Challenge 2020. in *Proceedings of the*

NeurIPS 2020 Competition and Demonstration Track, Édit Hugo Jair E, Katja H (PMLR, Proceedings of Machine Learning Research), p 3--26.

Tyralis H, Papacharalampous G (2021) Boosting algorithms in energy research: a systematic review. *Neural Computing and Applications* 33(21):14101-14117.

UN General Assembly (2015) *Transforming our world: the 2030 agenda for sustainable development* https://www.refworld.org/docid/57b6e3e44.html, Accessed 26 October 2022

Urbanowicz RJ, Meeker M, La Cava W, Olson RS, Moore JH (2018) Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics* 85:189-203.

Valmassoi A, Keller JD, Kleist DT, English S, Ahrens B, Ďurán IB, Bauernschubert E, Bosilovich MG, Fujiwara M, Hersbach H, Lei L, Löhnert U, Mamnun N, Martin CR, Moore A, Niermann D, Ruiz JJ, Scheck L (2023) Current Challenges and Future Directions in Data Assimilation and Reanalysis. *Bulletin of the American Meteorological Society* 104(4):E756-E767.

Valsaraj P, Thumba DA, Asokan K, Kumar KS (2020) Symbolic regression-based improved method for wind speed extrapolation from lower to higher altitudes for wind energy applications. *Applied Energy* 260:114270.

Van Ackere S, Van Eetvelde G, Schillebeeckx D, Papa E, Van Wyngene K, Vandevelde L (2015) Wind Resource Mapping Using Landscape Roughness and Spatial Interpolation Methods. *Energies* 8(8).

Van Lieshout P (2010) Improvements in AEP calculations using IEC 61400. *Windtech International* 6(3).

Vapnik VN (2000) Methods of Function Estimation. *The Nature of Statistical Learning Theory*, Vapnik VN (Édit.) Springer New York, New York, NY10.1007/978-1-4757-3264-1_7. p 181-224.

Vassallo D, Krishnamurthy R, Fernando HJS (2020) Decreasing wind speed extrapolation error via domain-specific feature extraction and selection. *Wind Energ. Sci.* 5(3):959-975.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. *Advances in Neural Information Processing Systems* 30.

Vautard R, Cattiaux J, Yiou P, Thepaut J-N, Ciais P (2010a) Northern Hemisphere atmospheric stilling partly due to an increase in surface roughness. *Nat. Geosci.* 3:756-761.

Vautard R, Cattiaux J, Yiou P, Thépaut J-N, Ciais P (2010b) Northern Hemisphere atmospheric stilling partly attributed to an increase in surface roughness. *Nature Geoscience* 3(11):756-761.

Veers P, Dykes K, Lantz E, Barth S, Bottasso CL, Carlson O, Clifton A, Green J, Green P, Holttinen H, Laird D, Lehtomäki V, Lundquist JK, Manwell J, Marquis M, Meneveau C, Moriarty P, Munduate X, Muskulus M, Naughton J, Pao L, Paquette J, Peinke J, Robertson A, Sanz Rodrigo J, Sempreviva AM, Smith JC, Tuohy A, Wiser R (2019) Grand challenges in the science of wind energy. *Science* 366(6464):eaau2027.

Veronesi F, Grassi S, Raubal M (2016) Statistical learning approach for wind resource assessment. *Renewable and Sustainable Energy Reviews* 56:836-850.

Veronesi F, Grassi S, Raubal M, Hurni L (2015) Statistical Learning Approach for Wind Speed Distribution Mapping: The UK as a Case Study. *AGILE 2015: Geographic Information Science as an Enabler of Smarter Cities and Communities*, Bacao F, Santos MY,Painho

M (Édit.) Springer International Publishing, Cham10.1007/978-3-319-16787-9_10. p 165-180.

Vu Dinh Q, Doan Q-V, Ngo-Duc T, Nguyen Dinh V, Dinh Duc N (2022) Offshore wind resource in the context of global climate change over a tropical area. *Applied Energy* 308:118369.

Wais P (2017) A review of Weibull functions in wind sector. *Renewable and Sustainable Energy Reviews* 70:1099-1107.

Waldmann E (2018) Quantile regression: A short story on how and why. *Statistical Modelling* 18(3-4):203-218.

Wan H, Wang XL, Swail VR (2010) Homogenization and Trend Analysis of Canadian Near-Surface Wind Speeds. *Journal of Climate* 23(5):1209-1225.

Wang Q, Li B, Xiao T, Zhu J, Li C, Wong DF, Chao LS (2019) Learning Deep Transformer Models for Machine Translation. (Florence, Italy, July) Association for Computational Linguistics, p 1810-1822.

Wang X (2008) Penalized Maximal F Test for Detecting Undocumented Mean Shift without Trend Change. *Journal of Atmospheric and Oceanic Technology - J ATMOS OCEAN TECHNOL* 25.

Wang XL, Wen QH, Wu Y (2007) Penalized Maximal t Test for Detecting Undocumented Mean Change in Climate Data Series. *Journal of Applied Meteorology and Climatology* 46(6):916-931.

Wang Y, Zou R, Liu F, Zhang L, Liu Q (2021) A review of wind speed and wind power forecasting with deep neural networks. *Applied Energy* 304:117766.

Wang Z, Kong F, Feng S, Wang M, Yang X, Zhao H, Wang D, Zhang Y (2025) Is Mamba effective for time series forecasting? *Neurocomputing* 619:129178.

Watson S (2019) Quantifying the Variability of Wind Energy. *Advances in Energy Systems,* https://doi.org/10.1002/9781119508311.ch21. p 355-368.

Weekes SM, Tomlin AS (2014a) Comparison between the bivariate Weibull probability approach and linear regression for assessment of the long-term wind energy resource using MCP. *Renewable Energy* 68:529-539.

Weekes SM, Tomlin AS (2014b) Data efficient measure-correlate-predict approaches to wind resource assessment for small-scale wind energy. *Renewable Energy* 63:162-171.

Weekes SM, Tomlin AS (2014c) Low-cost wind resource assessment for small-scale turbine installations using site pre-screening and short-term wind measurements. *IET Renewable Power Generation* 8(4):349-358.

Weekes SM, Tomlin AS, Vosper SB, Skea AK, Gallani ML, Standen JJ (2015) Long-term wind resource assessment for small and medium-scale turbines using operational forecast data and measure–correlate–predict. *Renewable Energy* 81:760-769.

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

Wen Q, Zhou T, Zhang C, Chen W, Ma Z, Yan J, Sun L (2023) Transformers in time series: a survey. in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*Macao, P.R.China), p Article 759.

Whiteman CD (2000) *Mountain Meteorology: Fundamentals and Applications.* Oxford University Press, https://doi.org/10.1093/oso/9780195132717.001.0001

Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP (2006) Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology* 75(5):1182-1189.

Wiernga J (1993) Representative roughness parameters for homogeneous terrain. *Boundary-Layer Meteorology* 63(4):323-363.

Wilk MB, Gnanadesikan R (1968) Probability plotting methods for the analysis for the analysis of data. *Biometrika* 55(1):1-17.

Wilks DS (2011) *Statistical methods in the atmospheric sciences.* Academic press,

Wilson JP (2018) *Environmental applications of digital terrain modeling.* John Wiley & Sons,

Wilson T, Tan P-N, Luo L (2022) DeepGPD: A Deep Learning Approach for Modeling Geospatio-Temporal Extreme Events. *Proceedings of the AAAI Conference on Artificial Intelligence* 36(4):4245-4253.

Winstral A, Elder K, Davis RE (2002) Spatial Snow Modeling of Wind-Redistributed Snow Using Terrain-Based Parameters. *Journal of Hydrometeorology* 3(5):524-538.

Winstral A, Jonas T, Helbig N (2017) Statistical Downscaling of Gridded Wind Speed Data Using Local Topography. *Journal of Hydrometeorology* 18(2):335-348.

Wiser R, Rand J, Seel J, Beiter P, Baker E, Lantz E, Gilman P (2021) Expert elicitation survey predicts 37% to 49% declines in wind energy costs by 2050. *Nature Energy* 6(5):555-565.

Woldesellasse H, Marpu PR, Ouarda TBMJ (2020) Long-term forecasting of wind speed in the UAE using nonlinear canonical correlation analysis (NLCCA). *Arabian Journal of Geosciences* 13(18):962.

Wood AW, Maurer EP, Kumar A, Lettenmaier DP (2002) Long-range experimental hydrologic forecasting for the eastern United States. *Journal of Geophysical Research: Atmospheres* 107(D20):ACL 6-1-ACL 6-15.

World Meteorological Organization (2020) *Guidance on homogenization.* Geneva, Switzerland, (WMO-No. 1245). 63 p

Wu H, Meng K, Fan D, Zhang Z, Liu Q (2022) Multistep short-term wind speed forecasting using transformer. *Energy* 261:125231.

Wu J, Chen X-Y, Zhang H, Xiong L-D, Lei H, Deng S-H (2019) Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimizationb. *Journal of Electronic Science and Technology* 17(1):26-40.

Xie Z-X, Hu Q-H, Yu D-R (2006) Improved Feature Selection Algorithm Based on SVM and Correlation. *Advances in Neural Networks - ISNN 2006.* (Berlin, Heidelberg, 2006//), Wang J, Yi Z, Zurada JM, Lu B-L,Yin H (Édit.) Springer Berlin Heidelberg, p 1373-1380.

Xiong L, Jiang C, Xu C-Y, Yu K-x, Guo S (2015) A framework of change-point detection for multivariate hydrological series. *Water Resources Research* 51(10):8198-8217.

Yang X, Jiang X, Liang S, Qin Y, Ye F, Ye B, Xu J, He X, Wu J, Dong T, Cai X, Xu R, Zeng Z (2024) Spatiotemporal variation of power law exponent on the use of wind energy. *Applied Energy* 356:122441.

Ye L, Dai B, Li Z, Pei M, Zhao Y, Lu P (2022) An ensemble method for short-term wind power prediction considering error correction strategy. *Applied Energy* 322:119475.

Ye W, Hong HP, Wang JF (2015) Comparison of Spatial Interpolation Methods for Extreme Wind Speeds over Canada. *Journal of Computing in Civil Engineering* 29(6):04014095.

Yu S, Vautard R (2022) A transfer method to estimate hub-height wind speed from 10 meters wind speed based on machine learning. *Renewable and Sustainable Energy Reviews* 169:112897.

Yu Y, Si X, Hu C, Zhang J (2019) A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Computation* 31(7):1235-1270.

Yue C-D, Liu C-C, Tu C-C, Lin T-H (2019) Prediction of Power Generation by Offshore Wind Farms Using Multiple Data Sources. *Energies* 12(4):700.

Zahradníček P, Brázdil R, Štěpánek P, Řezníčková L (2019) Differences in wind speeds according to measured and homogenized series in the Czech Republic, 1961–2015. *International Journal of Climatology* 39(1):235-250.

Zajaczkowski FJ, Haupt SE, Schmehl KJ (2011) A preliminary study of assimilating numerical weather prediction data into computational fluid dynamics models for wind prediction. *Journal of Wind Engineering and Industrial Aerodynamics* 99(4):320-329.

Zardi D, Whiteman CD (2013) Diurnal Mountain Wind Systems. *Mountain Weather Research and Forecasting: Recent Progress and Current Challenges*, Chow FK, De Wekker SFJ,Snyder BJ (Édit.) Springer Netherlands, Dordrecht10.1007/978-94-007-4098-3_2. p 35-119.

Zeng Z, Ziegler AD, Searchinger T, Yang L, Chen A, Ju K, Piao S, Li LZX, Ciais P, Chen D, Liu J, Azorin-Molina C, Chappell A, Medvigy D, Wood EF (2019) A reversal in global terrestrial stilling and its implications for wind energy production. *Nature Climate Change* 9(12):979-985.

Zhang G, Azorin-Molina C, Wang X, Chen D, McVicar TR, Guijarro JA, Chappell A, Deng K, Minola L, Kong F, Wang S, Shi P (2022a) Rapid urbanization induced daily maximum wind speed decline in metropolitan areas: A case study in the Yangtze River Delta (China). *Urban Climate* 43:101147.

Zhang H, Chapman S, Trancoso R, Toombs N, Syktus J (2024) Assessing the impact of bias correction approaches on climate extremes and the climate change signal. *Meteorological Applications* 31(3):e2204.

Zhang H, Jeffrey S, Carter J (2022b) Improved quality gridded surface wind speed datasets for Australia. *Meteorology and Atmospheric Physics* 134(5):85.

Zhang J, Chowdhury S, Messac A, Hodge B-M (2014) A hybrid measure-correlate-predict method for long-term wind condition assessment. *Energy Conversion and Management* 87:697-710.

Zhang J, Draxl C, Hopson T, Monache LD, Vanvyve E, Hodge B-M (2015) Comparison of numerical weather prediction based deterministic and probabilistic wind resource assessment methods. *Applied Energy* 156:528-541.

Zhang NR, Siegmund DO (2007) A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* 63(1):22-32.

Zhang S, Li X (2021) Future projections of offshore wind energy resources in China using CMIP6 simulations and a deep learning-based downscaling method. *Energy* 217:119321.

Zhang Z, Qin H, Liu Y, Yao L, Yu X, Lu J, Jiang Z, Feng Z (2019) Wind speed forecasting based on Quantile Regression Minimal Gated Memory Network and Kernel Density Estimation. *Energy Conversion and Management* 196:1395-1409.

Zhao W, Zhong Y, Li Q, Li M, Liu J, Tang L (2022) Comparison and correction of IDW based wind speed interpolation methods in urbanized Shenzhen. *Frontiers of Earth Science* 16(3):798-808.

Zhao Z, Anand R, Wang M (2019) Maximum Relevance and Minimum Redundancy Feature Selection Methods for a Marketing Machine Learning Platform. *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA).*5-8 Oct. 2019), p 442-452.

Zhong X, Du F, Chen L, Wang Z, Li H (2024) Investigating transformer-based models for spatial downscaling and correcting biases of near-surface temperature and wind-speed forecasts. *Quarterly Journal of the royal meteorological society* 150(758):275-289.

Zhou F, Zhao Z, Azorin-Molina C, Jia X, Zhang G, Chen D, Liu J, Guijarro JA, Zhang F, Fang K (2022) Teleconnections between large-scale oceanic-atmospheric patterns and interannual surface wind speed variability across China: Regional and seasonal patterns. *Science of The Total Environment* 838:156023.

Zhou H, Zhang S, Peng J, Zhang S, Li J, Xiong H, Zhang W (2021a) Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence.* p 11106-11115.

Zhou J, Erdem E, Li G, Shi J (2010) Comprehensive evaluation of wind speed distribution models: A case study for North Dakota sites. *Energy Conversion and Management* 51(7):1449-1458.

Zhou Y, Li Y, Wang D, Liu Y (2023) A multi-step ahead global solar radiation prediction method using an attention-based transformer model with an interpretable mechanism. *International Journal of Hydrogen Energy* 48(40):15317-15330.

Zhou Y, Liu Y, Wang D, Liu X, Wang Y (2021b) A review on global solar radiation prediction with machine learning models in a comprehensive perspective. *Energy Conversion and Management* 235:113960.

Zhou Y, Luckow P, Smith SJ, Clarke L (2012) Evaluation of Global Onshore Wind Energy Potential and Generation Costs. *Environmental science & technology* 46(14):7857-7864.

Zhu Z (2017) Change detection using landsat time series: A review of frequencies, preprocessing, algorithms, and applications. *ISPRS Journal of Photogrammetry and Remote Sensing* 130:370-384.

Zilitinkevich SS, Deardorff JW (1974) Similarity Theory for the Planetary Boundary Layer of Time-Dependent Height. *Journal of Atmospheric Sciences* 31(5):1449-1452.

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2):301-320.