ENVIRONMENTAL RESEARCH
LETTERS

LETTER • OPEN ACCESS

Downscaling ERA5 wind speed data: a machine learning approach considering topographic influences

To cite this article: Wenxuan Hu *et al* 2023 *Environ. Res. Lett.* **18** 094007

ENVIRONMENTAL RESEARCH
LETTERS

LETTER

Downscaling ERA5 wind speed data: a machine learning approach considering topographic influences

Wenxuan Hu[1,*], **Yvonne Scholz**[1], **Madhura Yeligeti**[1], **Lueder von Bremen**[2] and **Ying Deng**[1]

[1] German Aerospace Center (DLR), Institute of Networked Energy Systems, Curiestr. 4, 70563 Stuttgart, Germany
[2] German Aerospace Center (DLR), Institute of Networked Energy Systems, Carl-von-Ossietzky-Str. 15, 26129 Oldenburg, Germany
* Author to whom any correspondence should be addressed.

E-mail: wenxuan.hu@student.kit.edu

Keywords: downscaling, wind speed, time series, machine learning, reanalysis, ERA5

OPEN ACCESS

RECEIVED
10 February 2023

REVISED
4 July 2023

ACCEPTED FOR PUBLICATION
27 July 2023

PUBLISHED
10 August 2023

Abstract

Energy system modeling and analysis can provide comprehensive guidelines to integrate renewable energy sources into the energy system. Modeling renewable energy potential, such as wind energy, typically involves the use of wind speed time series in the modeling process. One of the most widely utilized datasets in this regard is ERA5, which provides global meteorological information. Despite its broad coverage, the coarse spatial resolution of ERA5 data presents challenges in examining local-scale effects on energy systems, such as battery storage for small-scale wind farms or community energy systems. In this study, we introduce a robust statistical downscaling approach that utilizes a machine learning approach to improve the resolution of ERA5 wind speed data from around 31 km × 31 km to 1 km × 1 km. To ensure optimal results, a comprehensive preprocessing step is performed to classify regions into three classes based on the quality of ERA5 wind speed estimates. Subsequently, a regression method is applied to each class to downscale the ERA5 wind speed time series by considering the relationship between ERA5 data, observations from weather stations, and topographic metrics. Our results indicate that this approach significantly improves the performance of ERA5 wind speed data in complex terrain. To ensure the effectiveness and robustness of our approach, we also perform thorough evaluations by comparing our results with the reference dataset COSMO-REA6 and validating with independent datasets.

1. Introduction

Renewable energy technologies play a crucial role in mitigating climate change impacts. As the world shifts towards a more sustainable future, many nations have already started the transition toward renewable energy sources [1, 2]. In energy system models, values such as installable capacity and power generation potentials can be calculated from meteorological data, such as wind speed [3, 4]. The most frequently used meteorological dataset is reanalysis data [5, 6]. Reanalysis data can provide meteorological data in a spatially and temporally consistent way [7]. However, its low spatial resolution impedes its ability to capture small-scale details, which are crucial for the accurate simulation of local-scale energy systems [8]. Over the years, in-depth studies of local effects and the development of local storage models have drawn more and more attention in the energy system modeling community, which cannot be achieved through the use of coarse-resolution reanalysis data. Therefore, to capture local effects in energy system models, a dataset with high spatial resolution is necessary.

Many efforts have contributed to the improvement of the spatial resolution of reanalysis data, with downscaling being the most widely adopted technique. Downscaling is a method for obtaining high-resolution climate or climate change information from global climate models [9]. Downscaling can be categorized into dynamic downscaling and statistical downscaling. Dynamic downscaling refers to the use of high-resolution regional climate models to dynamically extrapolate the impacts of large-scale climate processes to a regional or local scale of interest [10]. It provides individual variables that are physically consistent in time and space and is internally consistent across different variables [11]. However, the complexity of performing large-scale calculations of

various physical and chemical equations can be limiting and challenging due to the large amount of computational resources required.

On the other hand, statistical downscaling offers a computationally efficient solution by analyzing statistical relationships without considering complex physical and chemical processes. For example, Curry *et al* [12] investigated the statistical correlations between climate forecast variables and reanalysis data to derive monthly Weibull distribution parameters. Kirchmeier *et al* [13] and González-Aparicio *et al* [14] used vector generalized linear model to predict a daily-varying probability density function of local wind speeds conditioned on large-scale daily wind speed predictors. Other notable works in this field include [15–17]. These studies are focused on downscaling the probability distribution parameters of wind speed, which proves beneficial in quantifying the range of local-scale wind speeds and calculating energy output. However, for the purpose of energy system modeling, time series data are often required and cannot be derived from wind speed probability distributions.

Meanwhile, other studies developed statistical downscaling methods in different directions. For instance, Monahan [18] downscaled monthly wind speed time series at the buoy locations by using multiple linear regression. Jung *et al* [19] employed a least square boosting approach to downscale the monthly extreme wind speed value for North America and Europe. Winstral *et al* [20] developed an optimization scheme to downscale the wind speed time series in Switzerland taking the local terrain structure into consideration. These and other studies such as [21–23] are explicitly focused on the acquisition of high spatial resolution time series data, rather than the probability distribution parameters. However, many studies only developed site-specific corrections or aimed at obtaining daily or even monthly wind speed time series, limiting spatial and temporal capabilities when applied to an energy system model.

In recent years, the emergence of wind atlas platforms has significantly enhanced the acquisition of high spatial resolution wind speed data. Alongside various national and regional wind atlases in Europe [24], two prominent and widely recognized atlases, namely Global Wind Atlas [25] and the New European Wind Atlas [26], have gained recognition for their ability to provide high spatial resolution and open-access estimations of wind characteristics at specific locations. While these wind atlases offer high-resolution maps of the wind climate, such as long-term averaged wind speed and variability, they do not provide wind speed time series data for specific time spans. To address this limitation, researchers have explored the use of wind atlases to bias-adjust reanalysis data. One common approach involves scaling the reanalysis time series data to align with the long-term averaged wind speed provided by the wind atlas [27–29]. However, this method has a notable drawback: the resulting time series tends to be excessively smooth when compared to point measurements. As a consequence, it fails to accurately capture the significant fluctuations in wind speed commonly observed in measurements.

To overcome these limitations, we propose a machine learning-based statistical downscaling method to improve the spatial resolution of ERA5 wind speed time series data to around 1 km $\times$ 1 km. This method offers robust and computationally convenient solutions with simple input requirements by considering the importance of topographic conditions in local wind speed estimates. By investigating the relationship between large-scale wind speed, local-scale wind speed, and topographical metrics, our approach can improve the quality of reanalysis data for specific regions, thereby providing wind speed time series data at high spatial resolution. This enables us the capability to perform a close examination of local-scale energy systems, such as battery storage systems for small-scale wind farms or decentralized community energy systems.

2. Data

2.1. Local-scale observation data

In this study, we use two observation datasets: MeteoSwiss observations for Switzerland and German Meteorological Service (DWD) observations for Germany. MeteoSwiss observations are used for training, testing, and cross-dataset validation of our machine learning model, while the DWD observations representing an independent dataset are exclusively used for cross-dataset validation.

2.1.1. MeteoSwiss observation data

The MeteoSwiss observations are collected from the website of the Federal Office of Meteorology and Climatology of Switzerland [30]. This dataset comprises 10-meter hourly wind speed measurements gathered from weather stations distributed throughout Switzerland. In total, there are more than 150 measurement stations. However, stations with missing values accounting for over 10% of the data are excluded. The final dataset contains measurements from 116 weather stations for the years 2017, 2018, and 2019. Of these years, only observations from 2018 are used to develop our model, while measurements from 2017 and 2019 are used for cross-dataset validation. The distribution of the MeteoSwiss weather stations used in this study is presented in figure 1(a).

2.1.2. DWD observation data

The DWD (German Meteorological Service) observations, accessible from the DWD Open Data Server, also provide 10-meter hourly wind speed observations from over 500 weather stations [31]. As the

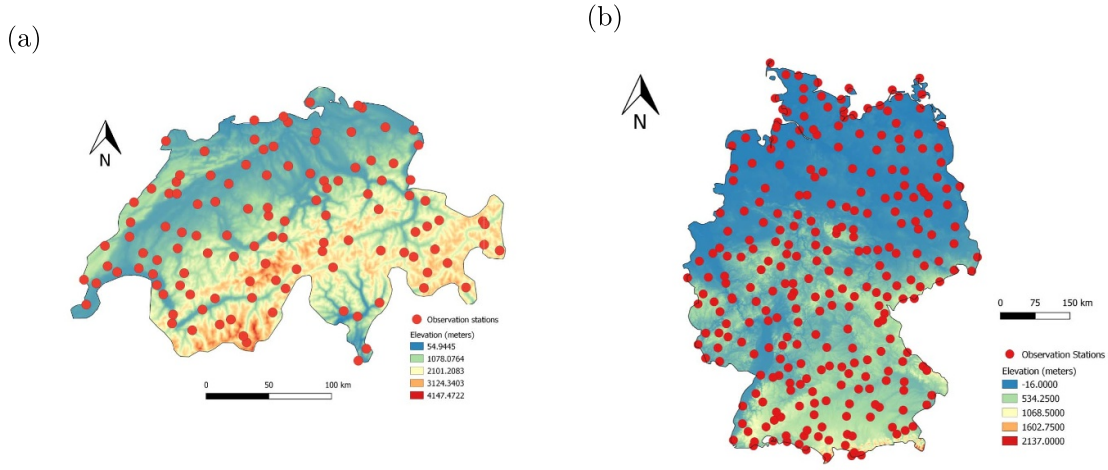

Figure 1. The study incorporates data from observation stations, including (a) the distribution of 116 MeteoSwiss stations across Switzerland and (b) the distribution of 272 DWD (German Meteorological Service) stations across Germany.

focus of this study is on topographic influence, offshore weather stations are excluded. In addition, stations with missing values exceeding 10% of the data are also excluded. The final dataset contains measurements from 272 stations in 2018. Figure 1(b) shows the distribution of these DWD weather stations.

2.2. Large-scale reanalysis data

Reanalysis data are gridded datasets that represent the atmosphere state, incorporating observations and outputs of numerical weather prediction models from past to present-day [32]. ERA5 is a widely used reanalysis dataset due to its extensive temporal and spatial coverage [8, 33–35]. Moreover, it also provides more than 200 other variables, some of which are topography-related and have significant impacts on wind speeds. Leveraging the advantages of ERA5, we employ ERA5 as the source datasets for our downscaling model development, where the wind speed and wind direction time series are calculated from the '10 meters U-component of wind' and '10 meters V-component of wind' in ERA5. In addition to ERA5, we also utilize COSMO-REA6 [36], another reanalysis dataset, as the reference dataset for result comparison. Compared to ERA5, COSMO-REA6 has a higher spatial resolution, approximately 6 km × 6 km, offering valuable insights when assessing the performance of our results. We then employ nearest-neighbor interpolation to identify the nearest grid point based on the latitude and longitude of each weather station to obtain continuous time series data.

2.3. Topographic metrics

To thoroughly examine the effect of topography on local-scale wind speed, we calculate and analyze six crucial topographic metrics: elevation, slope, aspect, small and large-scale topographic position index (TPI), and terrain diversity index (TDI). These metrics offer insight into the complexities of the terrain and provide a comprehensive picture of its impact on wind dynamics [20, 34, 37].

These topographic metrics can all be derived from a Digital Elevation Model (DEM). The DEM used in this study is Global Land One-Kilometer Base Elevation (GLOBE), which has a spatial resolution of 0.0083 degrees (around 1 km × 1 km) [38]. Among these topographic metrics, elevation data can be directly retrieved from DEM. Slope can be calculated by dividing the vertical change in elevation by the horizontal distance [39]. Aspect is the orientation of slope, measured clockwise in degrees from 0 to 360 [40]. TPI, first proposed by Weiss in 2001 [41], provides a broad description of the elevation of a particular location relative to its surroundings. The TPI value of a point is a measure of its relative elevation, determined by subtracting the average elevation of all pixels within a specified radius from the elevation of the point. In this study, a radius of 75 km and 5 km TPI are calculated respectively to better capture both major landscape units and smaller local features.

In addition to these widely recognized and established topographic metrics, we introduce an index called TDI. It quantifies the topographic variety of an area by computing the ratio of the range of elevations to the mean elevation as indicated in equation (1), thus reflecting the diversity of the terrain. In our study, we employ an 11 km radius window for TDI calculation. The value of TDI serves as an indicator of the degree of topographic diversity in a given region. A higher TDI value signifies a greater range of elevations, thereby implying a more intricate topography. Conversely, a lower TDI implies a more uniform landscape. The maps of these topographic metrics for Switzerland are provided in figure 2.

$$\mathrm{TDI} = \frac{H_{\max} - H_{\min}}{H_{\mathrm{mean}}} \tag{1}$$

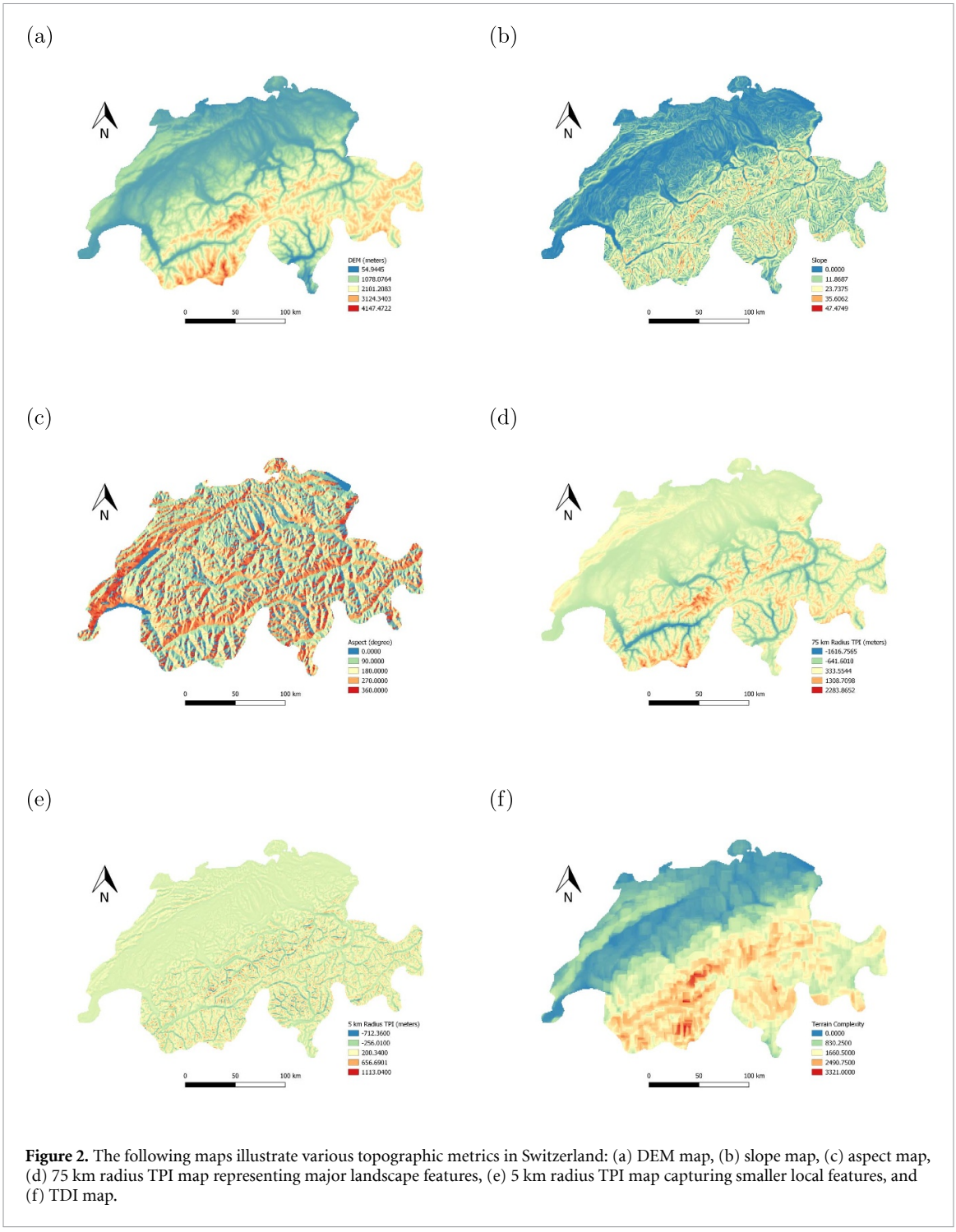

Figure 2. The following maps illustrate various topographic metrics in Switzerland: (a) DEM map, (b) slope map, (c) aspect map, (d) 75 km radius TPI map representing major landscape features, (e) 5 km radius TPI map capturing smaller local features, and (f) TDI map.

where:

$$\begin{aligned} \text{TDI} &= \text{Terrain diversity index} \\ H_{\max} &= \text{Maximum elevation of an area} \\ H_{\min} &= \text{Minimum elevation of an area} \\ H_{\text{mean}} &= \text{Mean elevation of an area} \end{aligned}$$

3. Methods

Our statistical downscaling approach involves a regression analysis aimed at establishing the relationship between ERA5, observed data, and topographic metrics. Before this process, we conduct a data preprocessing step that classifies the study region into various categories based on the aforementioned topographic metrics, thereby ensuring a thorough consideration of the terrain impact.

3.1. Data preprocessing

When comparing ERA5 with observations, we discover that the biases of ERA5 vary based on the site location. More specifically, in mountainous areas or valleys, ERA5 tends to have large biases, whereas, in plain areas, the biases are usually small. This is

Table 1. The classification rule employed in this study is based on the RMSE value between the station observations and ERA5. This scheme is used to determine the observation station classes according to their level of discrepancy with the observed values. Class1 represents a good ERA5 quality with an RMSE less than 1.5 m s^{-1}, Class2 represents a moderate ERA5 quality with an RMSE bigger than 1.5 m s^{-1} but less than 3 m s^{-1}, while Class 3 indicates a poor ERA5 quality with an RMSE greater than 3 m s^{-1}.

Class	Number of stations	RMSE range
Class1	276	$\mathrm{RMSE} \leqslant 1.5\,\mathrm{m\,s^{-1}}$
Class2	88	$1.5\,\mathrm{m\,s^{-1}} < \mathrm{RMSE} \leqslant 3\,\mathrm{m\,s^{-1}}$
Class3	24	$\mathrm{RMSE} > 3\,\mathrm{m\,s^{-1}}$

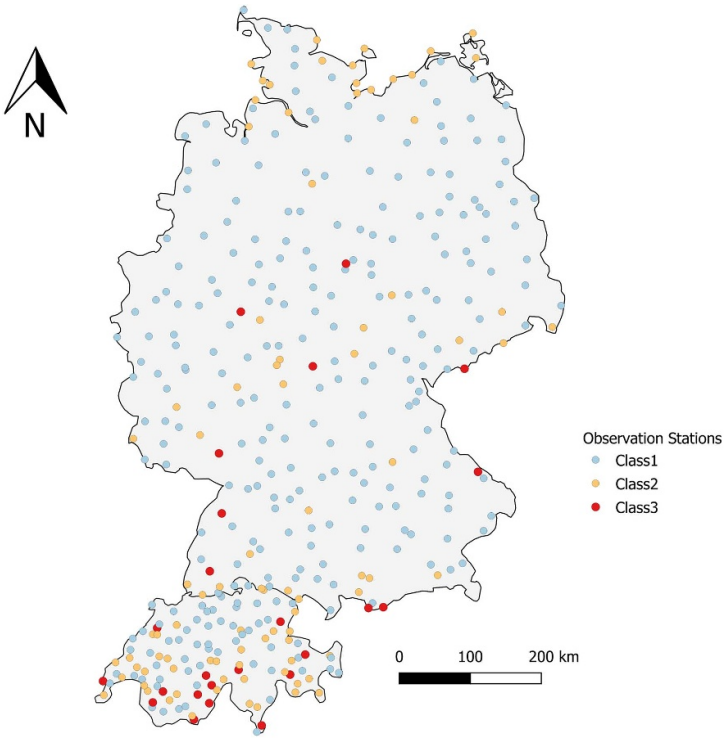

Figure 3. Distribution of measurement stations classified into three classes based on ERA5 quality. Class1 (276 stations) represents stations with good ERA5 quality, Class2 (88 stations) indicates moderate ERA5 quality, and Class3 (24 stations) corresponds to stations with poor ERA5 quality.

consistent with other studies [8, 20, 42]. To examine the influence of topography on the quality of ERA5, we first employ the method proposed by Winstral *et al* [20] who utilized TPI to assess the performance of COSMO-REA6. They observed that COSMO-REA6 overestimates wind speeds in regions with low TPI values and underestimates wind speeds in regions with high TPI values. However, our findings differ in the case of ERA5, where we observe that ERA5 often underestimates wind speed even for regions with low TPI values. This discrepancy could be due to the coarser spatial resolution of ERA5 compared to COSMO-REA6 and the insufficiency of relying on a single topographic metric such as TPI.

To investigate the impact of various topographic metrics on the accuracy of ERA5 and predict the potential quality of ERA5 for any given region solely based on the topographic conditions, we propose a preprocessing step. This involves integrating multiple topographic metrics, as outlined in section 2.3, into a random forest classification model implemented using the scikit-learn python package [43].

Initially, all weather stations are classified into three classes based on the root mean square error (RMSE) value between observations and ERA5, as indicated in table 1. These classes are used as the target variables in the classification process. Additionally, we calculate elevation, slope, aspect, TPI (with 5 km and 75 km radii), and TDI for each station and include them as input features for the classification process. It is worth noting that due to the limited sample size of MeteoSwiss stations, the classification process is performed for both MeteoSwiss and DWD weather stations. The distribution of the stations across different classes is illustrated in figure 3. To assess the model performance, a data split of 80% training data (310 stations) and 20% testing data (78 stations) is applied.

3.2. Regression of wind speed time series

After the preprocessing step, a regression analysis is performed for each class to investigate the

correlations between input features and the target variable. In our case, input features include both large-scale ERA5 data and local-scale topographic metrics. These large-scale ERA5 data are time-dependent and provided as time series, including ERA5 wind speed, ERA5 wind direction, and gravity wave dissipation (GWD). The local-scale topographic metrics in our study are time-independent and provided as constants, including 5 km and 75 km Radius TPI. And the target variable is observed wind speed, which is also a time-dependent variable and provided as time series. Including GWD as an input feature in our analysis serves a specific purpose. GWD is the cumulative conversion of kinetic energy into thermal energy in the mean flow over the entire atmospheric column per unit area, which is due to the effects of stress associated with low-level, orographic blocking and orographic gravity waves [7]. Incorporating GWD can provide additional insights into the impact of unresolved valleys, hills, and mountains at a scale between 5 km and the ERA5 grid. This added dimension provides a better understanding of the topography influences on local-scale wind speed.

Before being introduced to the machine learning model, all predictors are normalized to set the features on a common scale. The regression process is then implemented using the eXtreme gradient boosting (XGBoost) algorithm, a scalable and optimized tree-boosting framework [44]. XGBoost offers great accuracy, scalability, and efficient handling of missing data, while its regularization capabilities can prevent overfitting [44]. Data for both input features and target variables are collected for all MeteoSwiss stations in 2018. Similar to the preprocessing step, all time series data are split into 70% for training and 30% for testing. To estimate the performance of the regression model, we use four crucial statistical metrics: RMSE, Pearson Correlation Coefficient (PCC), R^2 score, and Kolmogorov-Smirnov D statistic (KSD). KSD can quantify the degree of matching between two distributions, with $D = 0$ indicating a perfect match.

4. Results and cross-dataset validation

4.1. Results

4.1.1. Preprocessing results

In the preprocessing phase, our random forest classification model yields a model accuracy of 0.76 for the entire testing data set. However, when specifically considering Class3 stations, the model accuracy increased to 0.90. Figure 4 compares the relative significance of each input feature. A comprehensive map of the region class predictions in Europe is presented in figure 5. This map is generated by applying our random forest classification model to regions where no observation data is available.

4.1.2. Regression results

To determine if a machine learning model is prone to overfitting, it is important to compare its performance on training and testing datasets, as a comparable error in both datasets suggests that the model is generalized. Therefore, we compare the statistical metrics for both training and testing datasets as summarized in table 2. Meanwhile, table 3 showcases the improvements resulting from the regression process. To visually demonstrate these improvements, figure 6 presents scatter plots and histograms for a randomly selected station in each class. Meanwhile, the extent to which various input features impact the target variables is demonstrated in figure 7.

4.2. Cross-dataset validation

To verify the robustness of our regression model, we conduct a cross-dataset validation scheme in two ways. Firstly, we apply the regression model to all MeteoSwiss weather stations but using data from different years. Secondly, we test the model for DWD observations.

4.2.1. Cross-dataset validation across different years

To acquire the downscaled wind speed time series for MeteoSwiss observation stations across multiple years, the time-dependent input features are first adjusted for the current year prior to being fed into the regression model. The results of this cross-validation are summarized in table 4. To provide a visual representation of these results, scatter plots and histograms for three representative stations are shown in figure 8, and time series plots are displayed in figure 9. Due to limited data availability in COSMO-REA6 after August 2019, the comparison for the year 2019 is focused on the first 8 months.

4.2.2. Cross-dataset validation across different locations

To perform cross-validation with DWD observation stations, the input features are first computed at DWD station locations before being fed into the regression model. The results are presented in table 5. To delve deeper into the improvement observed in Class3 stations, table 6 presents a comparison of statistic metrics for all Class3 DWD stations. For a more visual representation of the results, figures 10 and 11 present scatter plots, histograms, and time series plots for selected Class 3 stations.

5. Discussion

In the preprocessing step, we investigated the impact of various topographic metrics on the accuracy of ERA5. Our findings reveal that all of the topographic metrics considered in this study play a crucial role in determining the performance of ERA5.

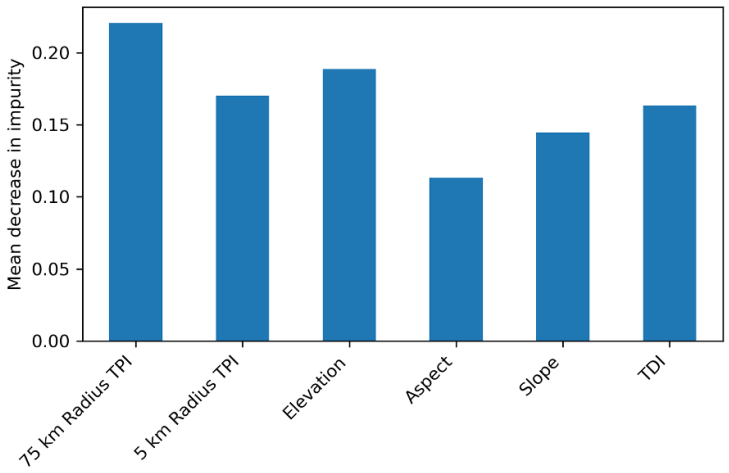

Figure 4. Feature importance plot for the Random Forest classification trained on all the available observation stations. The plot showcases the relative importance of different features in predicting the station classes. The analysis reveals that 75 km Radius TPI has the most pronounced impact on the local-scale wind speed, with Elevation, 5 km Radius TPI, TDI, Slope, and Aspect following in that order.

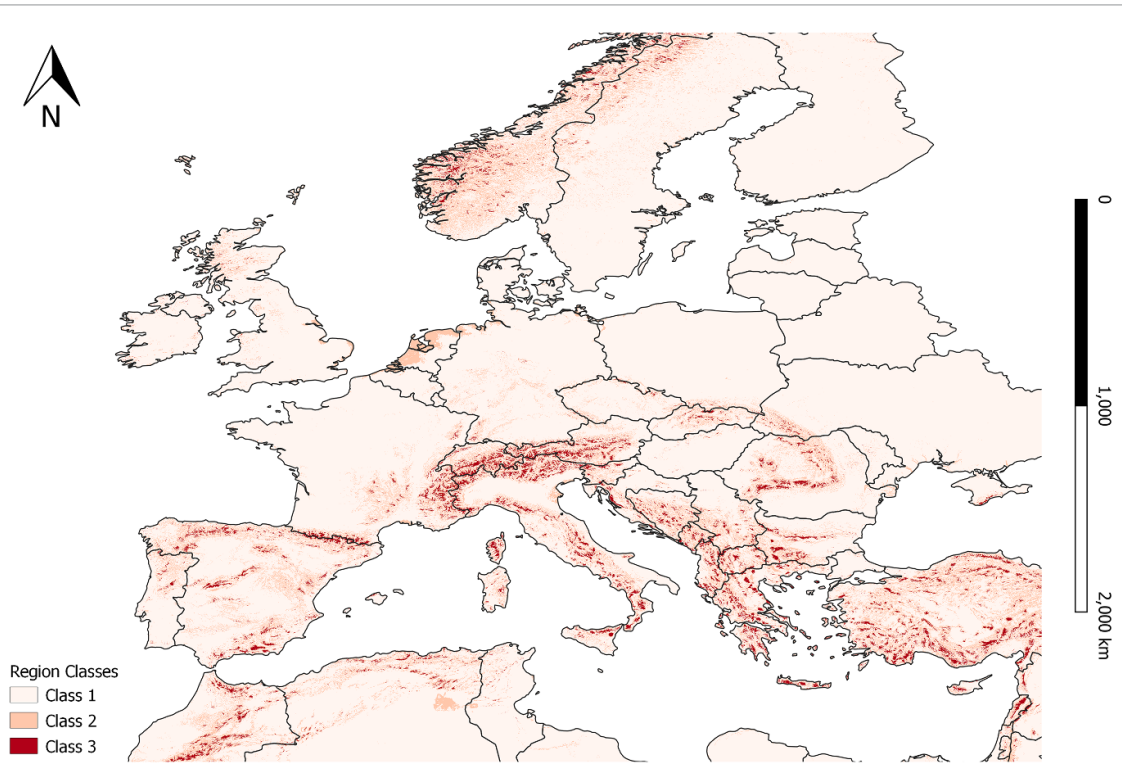

Figure 5. Map displaying the prediction of region classes after implementing the Random Forest classification method across Europe. It showcases the division of areas into Class 1 (small biases of ERA5), Class 2 (moderate biases of ERA5), and Class 3 (significant biases of ERA5) regions. The spatial resolution of the map is 1 km × 1 km.

Table 2. Comparison of statistical metrics (RMSE, PCC, R^2, KSD) for training and testing data for different classes in the machine learning model. The comparison reveals variations in the model's performance for each class between the training and testing phases. A comparable error in both datasets suggests a robust generalization of the model.

	RMSE ($m\ s^{-1}$)		PCC		R^2		KSD	
Class	training	testing	training	testing	training	testing	training	testing
Class1	0.85	0.85	0.76	0.76	0.58	0.58	0.23	0.23
Class2	1.33	1.35	0.72	0.72	0.52	0.52	0.30	0.30
Class3	2.08	2.12	0.82	0.81	0.67	0.66	0.19	0.19

Table 3. Comparison of statistical metrics (RMSE, PCC, R^2, KSD) across the stations for each class between observations and three datasets: the original ERA5, COSMO-REA6, and the corrected wind speed time series obtained from the machine learning model. The comparison highlights the effectiveness of the machine learning model, as evidenced by improved statistical indicators. The results reveal a decrease in RMSE and KSD and an increase in PCC and R^2 for all classes, with the greatest improvement observed in Class3 stations, where RMSE decreases from 4.22 m s^{-1} to 2.05 m s^{-1}, PCC increases from 0.47 to 0.73, R^2 increases from 0.22 to 0.53, and KSD decreases from 0.59 to 0.24.

Class	Statistic metrics	ERA5	COSMO-REA6	Corrected
Class1	RMSE (m s^{-1})	1.03	1.39	0.84
	PCC	0.54	0.59	0.67
	R^2	0.29	0.35	0.49
	KSD	0.20	0.18	0.35
Class2	RMSE (m s^{-1})	1.70	1.58	1.32
	PCC	0.52	0.62	0.65
	R^2	0.27	0.38	0.42
	KSD	0.25	0.09	0.36
Class3	RMSE (m s^{-1})	4.22	4.03	2.05
	PCC	0.47	0.51	0.73
	R^2	0.22	0.26	0.53
	KSD	0.59	0.40	0.24

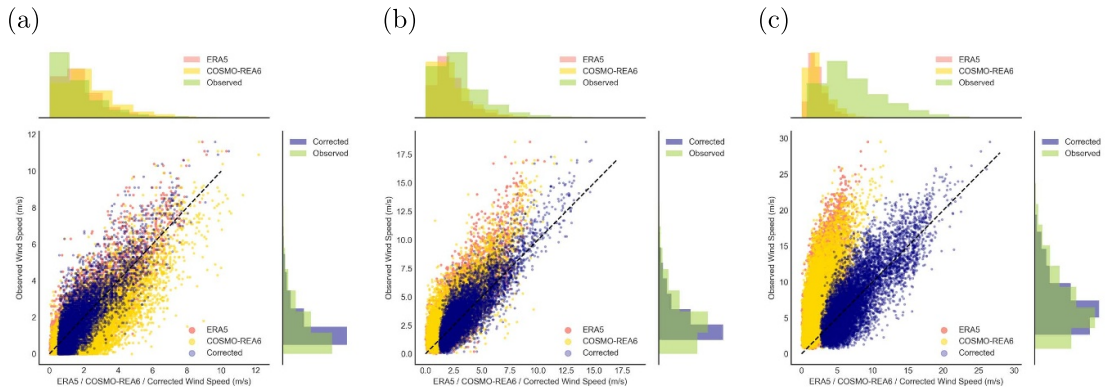

Figure 6. Scatter plots and histograms comparing observations with three datasets (original ERA5, COSMO-REA6, and corrected wind speed) are shown for three MeteoSwiss stations: (a) Station REH in Zürich Affoltern (Class1), (b) Station HAI in Salen-Reutenen (Class2), and (c) Station DOL at the mountain peak La Dôle (Class3). Notably, for Class3 station DOL, the scatter plot of the corrected wind speed aligns much closer with the diagonal line, and the histogram distribution more closely resembles the observations compared to ERA5 and COSMO-REA6 data. In contrast, the scatter plots and histograms of the original ERA5 and COSMO-REA6 indicate an underestimation of the wind speed. These comparisons suggest a significant reduction in biases and a better agreement with the observations for the corrected wind speed, particularly for Class3 station DOL.

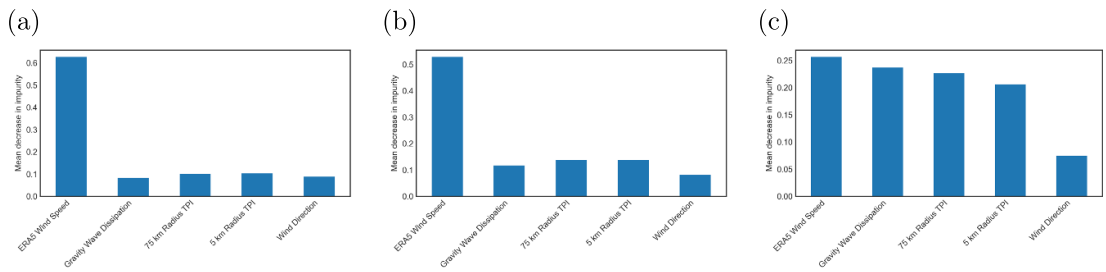

Figure 7. Feature importance plot for the XGBoost regression process for (a) Class1, (b) Class2, and (c) Class3. The plot showcases the relative importance of different features in predicting the wind speed time series at the local scale. The analysis reveals that ERA5 wind speed is the most dominant feature among all classes, and its importance significantly surpasses all other features in Class1 and Class2 regions. However, this is not observed in Class3 regions, where the influences of all features are more evenly distributed. For Class3 regions, GWD is the second most crucial predictor, following ERA5 wind speed, and is followed by 75 km Radius TPI, 5 km Radius TPI, and wind direction.

Table 4. Comparison of statistical metrics (RMSE, PCC, R^2, KSD) across the stations for each class in 2017 and 2019 between observations and three datasets: the original ERA5, COSMO-REA6, and the corrected wind speed time series. The regression model demonstrates notable improvements across all classes, particularly in Class3 stations. In 2017, the RMSE decreases to 2.48 m s^{-1}, the PCC increases to 0.77, the R^2 increases to 0.59, and the KSD decreases to 0.18. Similar improvements are observed in 2019. These results showcase the model's robustness when applied to independent datasets.

		ERA5		COSMO-REA6		Corrected	
Class	Statistic metrics	2017	2019	2017	2019	2017	2019
Class1	RMSE (m s^{-1})	1.03	1.10	1.41	1.45	0.92	0.95
	PCC	0.51	0.61	0.57	0.57	0.61	0.72
	R^2	0.26	0.37	0.32	0.32	0.37	0.52
	KSD	0.21	0.11	0.22	0.18	0.37	0.23
Class2	RMSE (m s^{-1})	1.64	1.77	1.52	1.63	1.41	1.40
	PCC	0.50	0.57	0.61	0.60	0.60	0.70
	R^2	0.25	0.32	0.37	0.36	0.36	0.49
	KSD	0.24	0.20	0.07	0.08	0.41	0.29
Class3	RMSE (m s^{-1})	5.03	5.18	4.56	4.64	2.48	2.45
	PCC	0.53	0.55	0.51	0.54	0.77	0.78
	R^2	0.28	0.30	0.26	0.29	0.59	0.61
	KSD	0.62	0.62	0.48	0.48	0.18	0.15

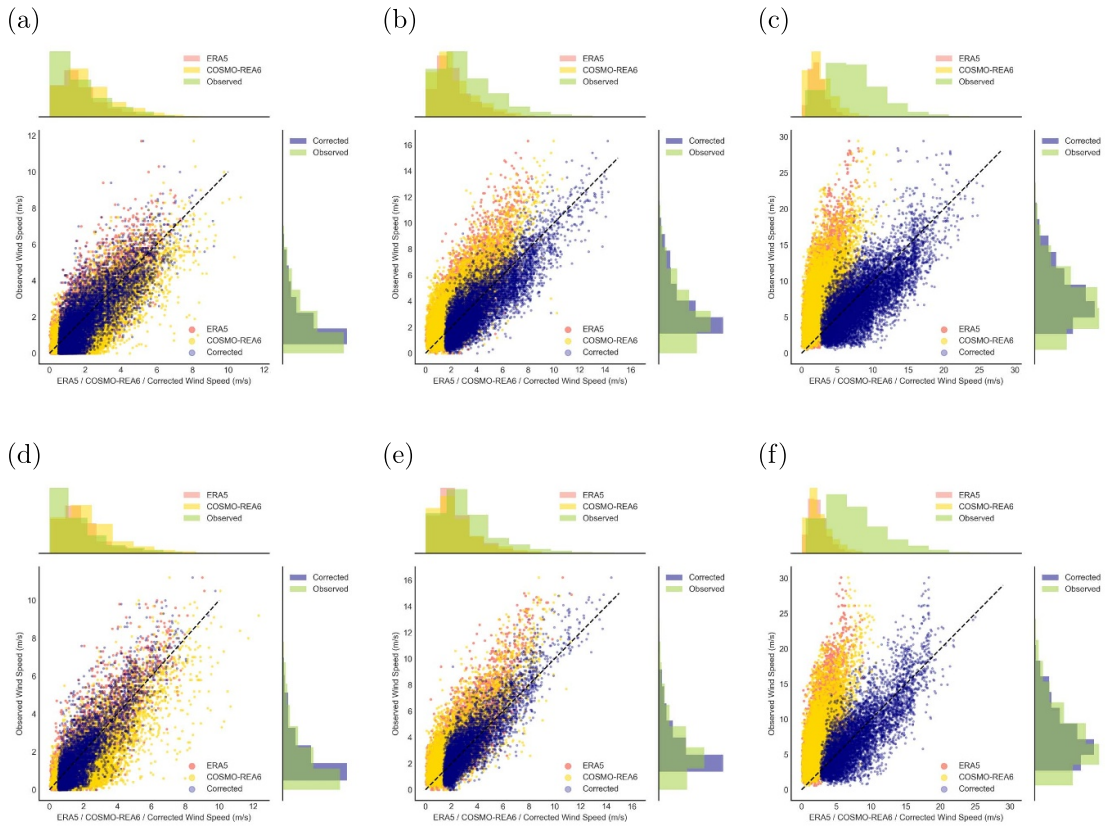

Figure 8. The scatter plots and histograms compare observations with three datasets (original ERA5, COSMO-REA6, and corrected wind speed) in 2017 and 2019. (a) Station REH (Class1) in 2017, (b) Station HAI (Class2) in 2017, and (c) Station DOL (Class3) in 2017. Similarly, (d)–(f) depict the same stations in 2019. For Class3 station DOL in both 2017 and 2019, the scatter plot of the corrected wind speed more closely aligns with the diagonal line, and the histogram distribution more closely resembles the observations compared to ERA5 and COSMO-REA6 data. This suggests a significant reduction in biases and a better agreement with the observations, suggesting the model's robustness when applied to independent datasets.

The model accuracy of the random forest classification indicates a strong performance overall, especially in Class3 region. Additionally, we present a map of the region classes in Europe, which reveals that most of Europe falls under Class1 regions with relatively high-quality ERA5 data. Conversely, Class3 regions, which indicate complex topographic conditions, occupy a small fraction of the total area, primarily surrounding

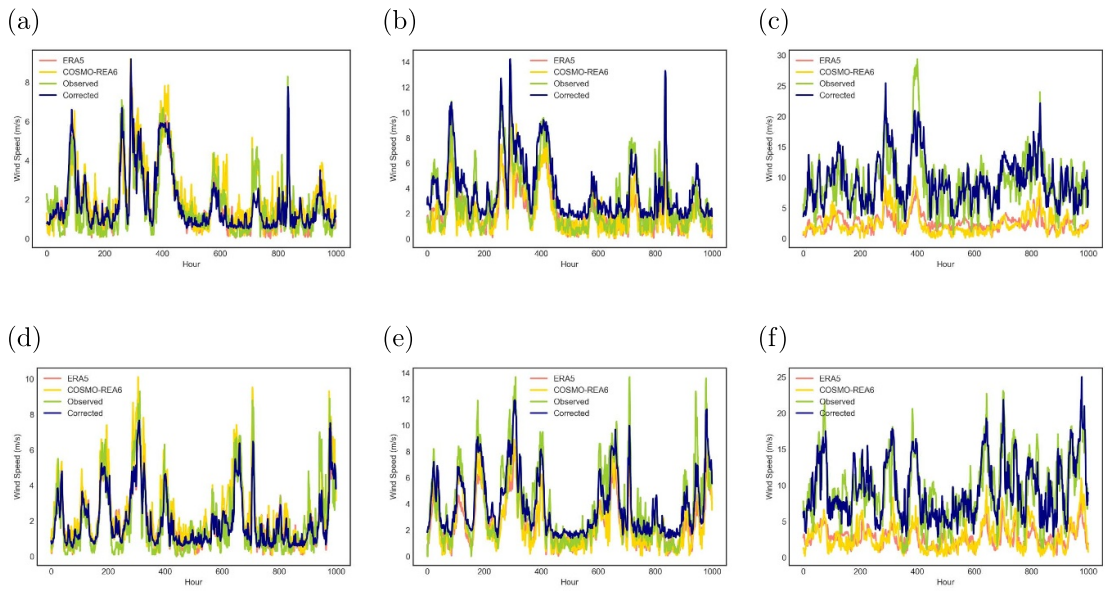

Figure 9. Time series plots in 2017 and 2019 for the observations, original ERA5, COSMO-REA6, and the corrected wind speed. For clarity, only the first 1000 hours are displayed. (a) Station REH (Class1) in 2017, (b) Station HAI (Class2) in 2017, and (c) Station DOL (Class3) in 2017. Similarly, (d)–(f) depict the same stations in 2019. For Station DOL, the stronger overlap between the corrected wind speed and observation data also suggests a significant reduction in biases, suggesting the model's robustness when applied to independent datasets.

Table 5. Comparison of statistical metrics (RMSE, PCC, R^2, KSD) across the stations for each class between DWD observations and three datasets: the original ERA5, COSMO-REA6, and the corrected wind speed time series. This table shows marked improvement in Class3 stations, with RMSE reducing to 2.77 m s^{-1}, PCC increasing to 0.76, R^2 increasing to 0.58, and KSD decreasing to 0.21. However, this improvement is not reflected in Class2 and Class1 stations, where a slight increase in RMSE and KSD and a corresponding decrease in PCC and R^2 have been observed. Overall, while significant improvement was observed in Class3 stations upon expanding the regression model to Germany, Class1 and Class2 stations showed no improvement.

Class	Statistic metrics	ERA5	COSMO-REA6	Corrected
Class1	RMSE (m s^{-1})	1.40	1.24	1.90
	PCC	0.79	0.80	0.78
	R^2	0.62	0.64	0.61
	KSD	0.14	0.05	0.34
Class2	RMSE (m s^{-1})	1.96	1.71	2.08
	PCC	0.73	0.79	0.74
	R^2	0.53	0.62	0.55
	KSD	0.10	0.06	0.19
Class3	RMSE (m s^{-1})	5.07	4.84	2.77
	PCC	0.69	0.71	0.76
	R^2	0.48	0.50	0.58
	KSD	0.62	0.47	0.21

mountainous regions. This suggests the fact that for a large geographic extent, it is not necessary to apply downscaling techniques to all regions. Focusing on areas with complex topographic conditions can save a significant amount of computational effort.

In the regression process, our results indicate that the regression model is well-fitted, as evidenced by the small differences in statistic metrics between training and testing datasets. The regression results reveal that our model can significantly improve the quality of ERA5, particularly in Class3 regions. Furthermore, the improvement in KSD suggests that this downscaling process not only decreases statistical error but also maintains a high degree of wind speed variability. The feature importance analysis indicates that TPI and GWD play a crucial role in estimating local-scale wind speed in Class3 regions. These topographic metrics serve as important indicators of the topographic conditions. Conversely, Class2 and Class1 regions, which are predominantly composed of flat terrain, are less affected by topographic conditions, making the topographic metrics less significant.

Table 6. Comparison of statistical metrics (RMSE, PCC, R^2, KSD) for all class3 DWD stations between observations and three datasets: the original ERA5, COSMO-REA6, and the corrected wind speed time series. The results highlight that there are noteworthy improvements for all Class 3 stations, with some showing significant improvements. For instance, station Brocken exemplifies such progress, as RMSE decreases from 8.23 m s^{-1} to 2.96 m s^{-1}, PCC increases from 0.79 to 0.88, R^2 increases from 0.62 to 0.77, and KSD decreases from 0.70 to 0.17.

Station	Statistic metrics	ERA5	COSMO-REA6	Corrected
Brocken	RMSE (m s^{-1})	8.23	7.94	2.96
	PCC	0.79	0.84	0.88
	R^2	0.62	0.71	0.77
	KSD	0.70	0.68	0.17
Kahler	RMSE (m s^{-1})	3.03	2.83	2.11
	PCC	0.86	0.84	0.86
	R^2	0.73	0.71	0.73
	KSD	0.49	0.46	0.32
Feldberg	RMSE (m s^{-1})	6.72	6.14	3.49
	PCC	0.71	0.80	0.78
	R^2	0.50	0.64	0.61
	KSD	0.69	0.60	0.17
Fichtelberg	RMSE (m s^{-1})	6.08	5.12	4.80
	PCC	0.77	0.81	0.84
	R^2	0.59	0.66	0.71
	KSD	0.64	0.49	0.51
Großer	RMSE (m s^{-1})	4.89	4.06	2.29
	PCC	0.67	0.75	0.78
	R^2	0.45	0.56	0.61
	KSD	0.63	0.45	0.11
Hornisgrinde	RMSE (m s^{-1})	4.89	4.40	2.34
	PCC	0.76	0.76	0.81
	R^2	0.58	0.58	0.66
	KSD	0.59	0.49	0.10
Wasserkuppe	RMSE (m s^{-1})	3.96	3.03	2.00
	PCC	0.71	0.80	0.77
	R^2	0.50	0.64	0.59
	KSD	0.53	0.34	0.08
Weinbiet	RMSE (m s^{-1})	3.64	3.94	1.69
	PCC	0.81	0.73	0.86
	R^2	0.66	0.53	0.74
	KSD	0.49	0.52	0.11
Zugspitze	RMSE (m s^{-1})	6.22	5.68	3.57
	PCC	0.35	0.51	0.61
	R^2	0.12	0.26	0.37
	KSD	0.82	0.72	0.31
Mittenwald	RMSE (m s^{-1})	3.03	2.57	2.40
	PCC	0.48	0.58	0.36
	R^2	0.23	0.34	0.13
	KSD	0.60	0.41	0.26

From the cross-dataset validation with MeteoSwiss observations across various years, we can observe a large degree of improvement, particularly in Class3 stations. This demonstration of generalizability suggests that our model can effectively apply to unseen datasets and is applicable across different years. However, when extending our model to Germany, the results reveal a marked improvement only for Class3 stations. The corrected wind speeds in Class1 and Class2 stations show a larger bias compared to ERA5 and COSMO-REA6. One possible explanation is that the topographic influences in these regions are so minimal that considering topographic-related features in a machine-learning regression model would result in inaccurate predictions. Therefore, we strongly recommend limiting the use of our regression process to Class3 regions alone. This will prevent potential errors in Class1 or

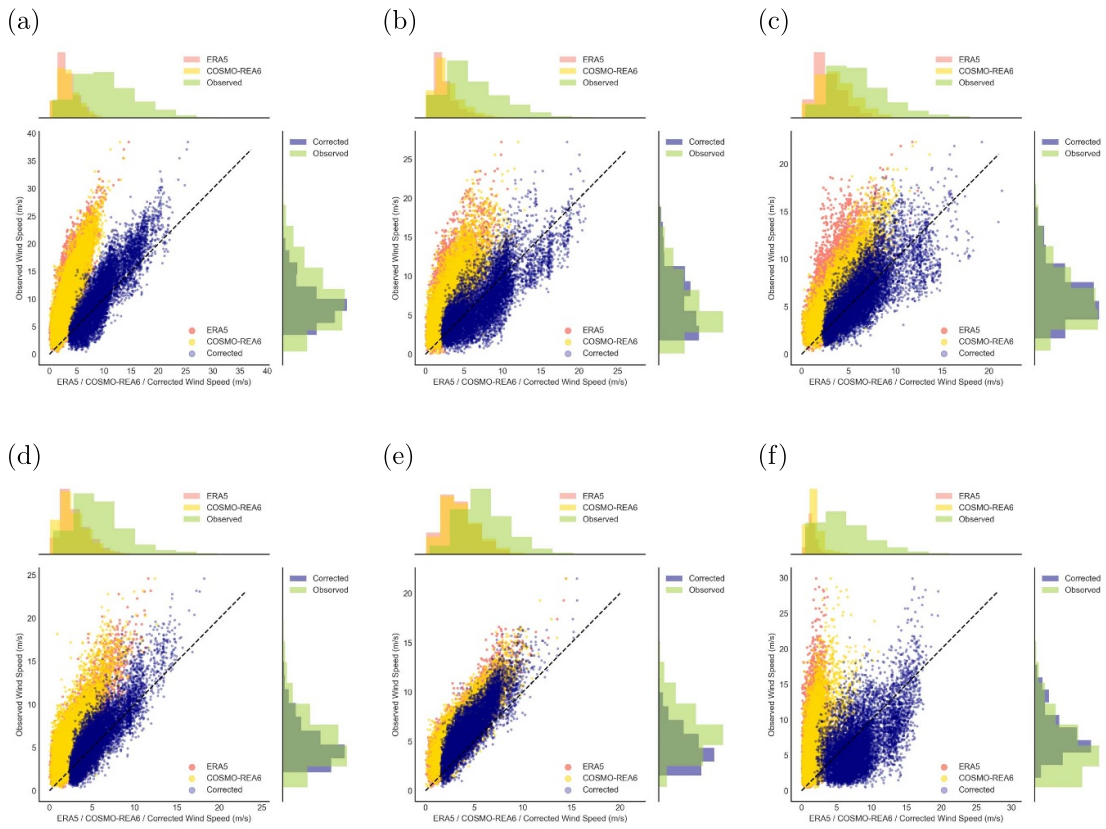

Figure 10. Scatter plots and histogram comparisons for 6 representative DWD stations between observations and three datasets: the original ERA5, COSMO-REA6, and the corrected wind speed time series. (a) Station Brocken, (b) Station Hornisgrinde, (c) Station Wasserkuppe, (d) Station Weinbiet, (e) Station Kahler and (f) Station Zugspitze. For all these stations, the scatter plot of the corrected wind speed aligns closer with the diagonal line, and the distribution of the histogram also exhibits a closer resemblance to the distribution of the observations compared to ERA5 and COSMO-REA6, suggesting a significant reduction in biases and a better agreement with the observations for the corrected wind speed, highlighting the robustness of our model when applied to independent datasets.

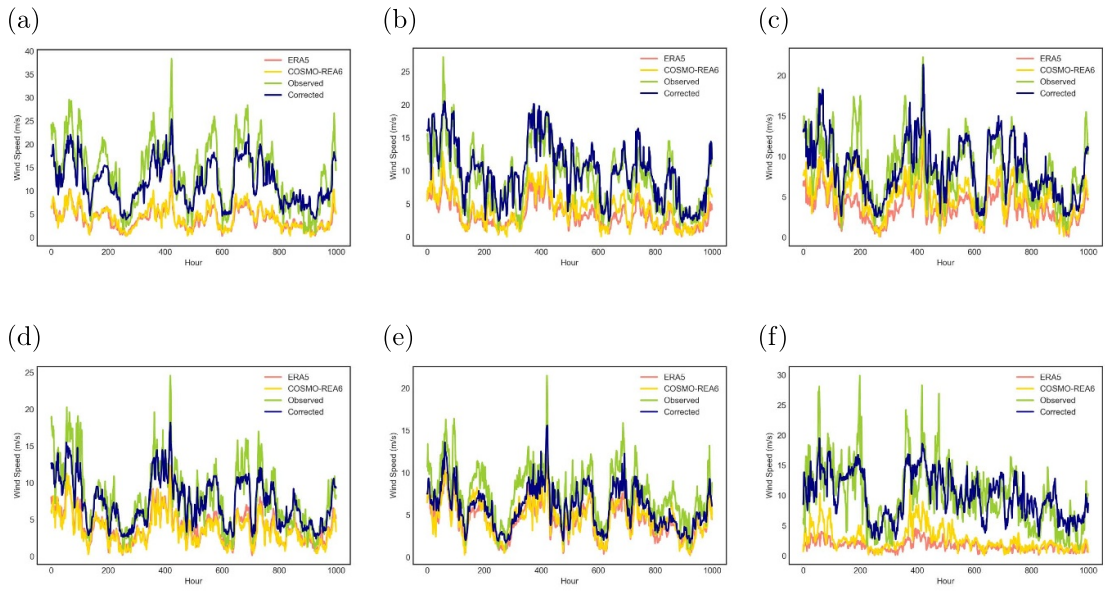

Figure 11. Time series plots for 6 representative DWD stations between observations and three datasets: original ERA5, COSMO-REA6, and the corrected wind speed. For clarity, only the first 1000 hours are displayed. (a) Station Brocken, (b) Station Hornisgrinde, (c) Station Wasserkuppe, (d) Station Weinbiet, (e) Station Kahler and (f) Station Zugspitze. For all these stations, the stronger overlap between the corrected wind speed and observation data also suggests a significant reduction in biases and a better agreement with the observations for the corrected wind speed, highlighting the robustness of our model when applied to independent datasets.

Class2 regions while significantly reducing computational efforts. However, to better correct wind speeds in Class1 and Class2 regions, one possible approach is to increase the amount of training data for the regression model. Given that our regression model is solely trained for MeteoSwiss data, incorporating more observations from Class1 and Class2 stations in diverse regions could lead to improved outcomes. Another possibility is to incorporate topography-independent features into the model, such as weather regimes and air pressure. These features, as previously noted in studies such as [45, 46], play a more significant role in affecting the quality of reanalysis data in Class1 and Class2 regions, and should be considered accordingly.

Furthermore, it is important to note that our study focuses on correcting wind speeds at 10 meters height, as the availability of wind speed observations at higher heights is limited. However, the same model can be applied to downscale wind speeds at higher heights once a sufficient amount of observational data becomes available.

6. Conclusion

In summary, this study highlights the crucial role played by topographic conditions in determining the spatially disparate quality of ERA5 wind speed data. Complex terrain regions, as characterized as Class3 regions through preprocessing step, show the highest degree of inaccuracies in the ERA5 wind speed data. To downscale ERA5 to higher spatial resolution, we apply a machine learning-based regression model to interpret the relationship between the large-scale ERA5 data, local-scale observations, and topographic metrics. The robustness of the regression model is evaluated by comparing its output against measurement data across different years and locations. The results demonstrate that while the method results in considerable improvement in ERA5 data quality in Class3 regions, the improvement is not as pronounced in Class1 and Class2 regions, which already have better ERA5 quality. By utilizing this method, ERA5 wind speed data can be downscaled from a spatial resolution of 31 km × 31 km to as fine as 1 km × 1 km, depending on the resolution of DEM used.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://10.5281/zenodo.8100208.

Acknowledgments

This study would not have been possible without the VERMEER project (support code: 03EI1010A) funded by the Federal Ministry for Economic Affairs and Climate Action (BMWK) of Germany. Thanks also to Ronald Stegen for his valuable programming suggestions, Marion Schroedter-Homscheidt for her suggestions and insights on the meteorological data, and to Hans Christian Gils and Patrick Jochem for their meticulous comments on the paper. Additional thanks to European Centre for Medium-Range Weather Forecasts (ECWMF), DWD, MeteoSwiss, and National Oceanic and Atmospheric Administration (NOAA) for providing the essential data that greatly contributed to this study.

ORCID iDs

Wenxuan Hu https://orcid.org/0000-0002-6704-9904
Yvonne Scholz https://orcid.org/0000-0002-1633-3825
Madhura Yeligeti https://orcid.org/0000-0002-9643-465X
Lueder von Bremen https://orcid.org/0000-0002-7072-0738
Ying Deng https://orcid.org/0000-0001-5063-3015

References

[1] Gielen D, Boshell F, Saygin D, Bazilian M D, Wagner N and Gorini R 2019 The role of renewable energy in the global energy transformation *Energy Strategy Rev.* **24** 38–50
[2] Thellufsen J Z and Lund H 2016 Roles of local and national energy systems in the integration of renewable energy *Appl. Energy* **183** 419–29
[3] Scholz Y 2012 Renewable energy based electricity supply at low costs: development of the REMix model and application for Europe *PhD Dissertation* University of Stuttgart
[4] Stetter D 2014 Enhancement of the REMix energy system model: global renewable energy potentials, optimized power plant siting and scenario validation Universtität Stuttgart
[5] Staffell I and Pfenninger S 2016 Using bias-corrected reanalysis to simulate current and future wind power output *Energy* **114** 1224–39
[6] Ritter M, Shen Z, Cabrera B L, Odening M and Deckert L 2015 A new approach to assess wind energy potential *Energy Proc.* **75** 671–6
[7] Hersbach H *et al* 2020 The ERA5 global reanalysis *Q. J. R. Meteorol. Soc.* **146** 1999–2049
[8] 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
[9] Pielke Sr R A and Wilby R L 2012 Regional climate downscaling: what's the point? *EOS Trans. Am. Geophys. Union* **93** 52–53
[10] Castro C L, Pielke Sr R A and Leoncini G 2005 Dynamical downscaling: assessment of value retained and added using the Regional Atmospheric Modeling System (RAMS) *J. Geophys. Res. Atmos.* **110** D5
[11] Giorgi F and Gutowski W J 2015 Regional dynamical downscaling and the CORDEX initiative *Annu. Rev. Environ. Resour.* **40** 467–90
[12] Curry C L, van der Kamp D and Monahan A H 2012 Statistical downscaling of historical monthly mean winds over a coastal region of complex terrain. I. Predicting wind speed *Clim. Dyn.* **38** 1281–99
[13] Kirchmeier M C, Lorenz D J and Vimont D J 2014 Statistical downscaling of daily wind speed variations *J. Appl. Meteorol. Climatol.* **53** 660–75

[14] Gonzalez-Aparicio I, Monforti F, Volker P, Zucker A, Careri F, Huld T and Badger J 2017 Simulating European wind power generation applying statistical downscaling to reanalysis data *Appl. Energy* **199** 155–68
[15] Davy R J, Woods M J, Russell C J and Coppin P A 2010 Statistical downscaling of wind variability from meteorological fields *Bound.-Layer Meteorol.* **135** 161–75
[16] Oh M, Lee J, Kim J and Kim H 2022 Machine learning-based statistical downscaling of wind resource maps using multi-resolution topographical data *Wind Energy* **25** 1121–41
[17] Alizadeh M J, Kavianpour M R, Kamranzad B and Etemad-Shahidi A 2019 A weibull distribution based technique for downscaling of climatic wind field *Asia Pac. J. Atmos. Sci.* **55** 685–700
[18] Monahan A H 2012 Can we see the wind? statistical downscaling of historical sea surface winds in the subarctic northeast Pacific *J. Clim.* **25** 1511–28
[19] Jung C, Demant L, Meyer P and Schindler D 2022 Highly resolved modeling of extreme wind speed in North America and Europe *Atmos. Sci. Lett.* **23** e1082
[20] Winstral A, Jonas T and Helbig N 2017 Statistical downscaling of gridded wind speed data using local topography *J. Hydrometeorol.* **18** 335–48
[21] Tang B H and Bassill N P 2018 Point downscaling of surface wind speed for forecast applications *J. Appl. Meteorol. Climatol.* **57** 659–74
[22] Goubanova K, Echevin V, Dewitte B, Codron F, Takahashi K, Terray P and Vrac M 2011 Statistical downscaling of sea-surface wind over the Peru–Chile upwelling region: diagnosing the impact of climate change from the IPSL-CM4 model *Clim. Dyn.* **36** 1365–78
[23] van der Kamp D, Curry C L and Monahan A H 2012 Statistical downscaling of historical monthly mean winds over a coastal region of complex terrain. II. Predicting wind components *Clim. Dyn.* **38** 1301–11
[24] Badger J et al 2019 Report on link to global wind atlas and national wind atlases (Deliverable D4. 7). Tech Rep.
[25] Technical University of Denmark (DTU) 2023 Global wind atlas (available at: https://globalwindatlas.info/en)
[26] Dörenkämper M *et al* 2020 The making of the new european wind atlas–part 2: production and evaluation *Geosci. Model Dev.* **13** 5079–102
[27] Murcia J P, Koivisto M J, Luzia G, Olsen B T, Hahmann A N, Sørensen P E and Als M 2022 Validation of European-scale simulated wind speed and wind generation time series *Appl. Energy* **305** 117794
[28] Gruber K and Schmidt J 2019 Bias-correcting simulated wind power in Austria and in Brazil from the ERA-5 reanalysis data set with the DTU wind atlas *11th Internationale Energiewirtschaftstagung an der TU Wien*
[29] Gruber K, Klöckl C, Regner P, Baumgartner J and 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
[30] MeteoSwiss 2013 The data portal of meteoswiss for research and teaching
[31] DWD Climate Data Center (CDC) 2013 Historical hourly station observations of wind speed and wind direction for Germany
[32] Gibson J 1997 ERA description *ECMWF re-analysis project report series* 1
[33] Olauson J 2018 ERA5: the new champion of wind power modelling? *Renew. Energy* **126** 322–31
[34] Molina M O, Gutiérrez C and Sánchez E 2021 Comparison of ERA5 surface wind speed climatologies over Europe with observations from the HadISD dataset *Int. J. Climatol.* **41** 4864–78
[35] Doddy Clarke E, Griffin S, McDermott F, Monteiro Correia J and Sweeney C 2021 Which reanalysis dataset should we use for renewable energy analysis in Ireland? *Atmosphere* **12** 624
[36] Bollmeyer C *et al* 2015 Towards a high-resolution regional reanalysis for the European CORDEX domain *Q. J. R. Meteorol. Soc.* **141** 1–15
[37] Solbakken K, Birkelund Y and Samuelsen E M 2021 Evaluation of surface wind using WRF in complex terrain: atmospheric input data and grid spacing *Environ. Model. Softw.* **145** 105182
[38] GLOBE Task Team The global land one-kilometer base elevation (GLOBE) digital elevation model, version 1.0 *National Oceanic and Atmospheric Administration, National Geophysical Data Center* 325 (available at: http://www.ngdc.noaa.gov/mgg/topo/globe.html)
[39] ESRI (Environmental Systems Research Institute) Slope ESRI ArcGIS resource center (available at: https://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/slope.htm) (Accessed 27 June 2023)
[40] ESRI (Environmental Systems Research Institute) Aspect ESRI ArcGIS resource center (available at: https://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/aspect.htm) (Accessed 27 June 2023)
[41] Weiss A 2001 Topographic position and landforms analysis *Poster Presentation, ESRI User Conf.* vol 200
[42] Vanella D *et al* 2022 Comparing the use of ERA5 reanalysis dataset and ground-based agrometeorological data under different climates and topography in Italy *J. Hydrol. Reg. Stud.* **42** 101182
[43] Pedregosa F *et al* 2011 Scikit-learn: machine learning in python *J. Mach. Learn. Res.* **12** 2825–30
[44] Chen T and Guestrin C 2016 Xgboost: a scalable tree boosting system *Proc. 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining* pp 785–94
[45] Brayshaw D J, Troccoli A, Fordham R and Methven J 2011 The impact of large scale atmospheric circulation patterns on wind power generation and its potential predictability: a case study over the UK *Renew. Energy* **36** 2087–96
[46] Garrido-Perez J M, Ordónez C, Barriopedro D, García-Herrera R and Paredes D 2020 Impact of weather regimes on wind power variability in Western Europe *Appl. Energy* **264** 114731