

**Document Version**

Final published version

**Licence**

CC BY

**Citation (APA)**

Baki, H., & Basu, S. (2024). Estimating high-resolution profiles of wind speeds from a global reanalysis dataset using TabNet. *Environmental Data Science*, 3, Article e32. <https://doi.org/10.1017/eds.2024.41>

**Important note**

To cite this publication, please use the final published version (if applicable).  
Please check the document version above.

**Copyright**

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.  
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

**Sharing and reuse**

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

**Takedown policy**

Please contact us and provide details if you believe this document breaches copyrights.  
We will remove access to the work immediately and investigate your claim.

APPLICATION PAPER  

# Estimating high-resolution profiles of wind speeds from a global reanalysis dataset using TabNet

Harish Baki<sup>1</sup>  and Sukanta Basu<sup>2,3</sup>

<sup>1</sup>Faculty of Civil Engineering and Geosciences, TU Delft, Delft, the Netherlands

<sup>2</sup>Atmospheric Sciences Research Center, University at Albany, Albany, NY, USA

<sup>3</sup>Department of Environmental and Sustainable Engineering, University at Albany, Albany, NY, USA

**Corresponding author:** Harish Baki; Email: [h.baki@tudelft.nl](mailto:h.baki@tudelft.nl)

**Received:** 29 June 2024; **Accepted:** 28 September 2024

**Keywords:** Chebyshev polynomials; deep learning; wind resource assessment

## Abstract

The growing demand for global wind power production, driven by the critical need for sustainable energy sources, requires reliable estimation of wind speed vertical profiles for accurate wind power prediction and comprehensive wind turbine performance assessment. Traditional methods relying on empirical equations or similarity theory face challenges due to their restricted applicability beyond the surface layer. Although recent studies have utilized various machine learning techniques to vertically extrapolate wind speeds, they often focus on single levels and lack a holistic approach to predicting entire wind profiles. As an alternative, this study introduces a proof-of-concept methodology utilizing TabNet, an attention-based sequential deep learning model, to estimate wind speed vertical profiles from coarse-resolution meteorological features extracted from a reanalysis dataset. To ensure that the methodology is applicable across diverse datasets, Chebyshev polynomial approximation is employed to model the wind profiles. Trained on the meteorological features as inputs and the Chebyshev coefficients as targets, the TabNet more-or-less accurately predicts unseen wind profiles for different wind conditions, such as high shear, low shear/well-mixed, low-level jet, and high wind. Additionally, this methodology quantifies the correlation of wind profiles with prevailing atmospheric conditions through a systematic feature importance assessment.

## Impact Statement

We applied deep learning in conjunction with Chebyshev polynomials to predict unseen wind profiles from coarse-resolution meteorological features and correlate them with the prevailing atmospheric conditions. The methodology can be extended to different geographical locations and diverse wind profile datasets.

## 1. Introduction

The demand for global wind power production has experienced a significant increase, driven by the growing recognition of renewable energy sources as an essential solution to combat climate change and the pressing need for a sustainable, low-carbon future (Nagababu et al., 2023). Although offshore wind power technology is still in its initial stages, it is predicted to grow rapidly, which is primarily attributed to

  This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

offshore wind speeds being higher and more uniform as the distance from the coast increases (Guo et al., 2022). Moreover, technological advancements have facilitated the deployment of the largest wind turbines to date, such as the MySE 16–260 with a 16 MW capacity (New Atlas, 2023), having a rotor diameter of 260 m and a hub height of 152 m, making it the largest wind turbine to reach a towering height of 280 m.

For such massive wind turbines, the traditional use of hub height wind speed in estimating power output (IEC, 2005) is insufficient due to varying wind speeds across the rotor plane. To address this, the rotor equivalent wind speed approach considers a wind profile within the rotor swept area, enhancing the reliability and accuracy of power output estimates for such large wind turbines (Wagner et al., 2011; Van Sark et al., 2019). Furthermore, specific meteorological conditions lead to the formation of distinct wind profiles, such as low-level jets (LLJs) during strong stratification, well-mixed profiles during very unstable conditions, and Ekman profiles during neutral conditions (Durán et al., 2020). The atmospheric variables like wind shear and turbulence intensity associated with these diverse wind profiles significantly influence power production (Elliott and Cadogan, 1990; Wharton and Lundquist, 2012). Beyond power output, wind profiles under different stability conditions also significantly impact turbine loads (Dimitrov et al., 2015; Gutierrez et al., 2017; Park et al., 2015). These findings underscore the importance of analyzing wind profiles in both wind resource assessment and turbine design analysis.

However, characterizing wind profiles across the rotor-swept area has been hindered by the scarcity of observations at high altitudes, as the deployment of wind measurement instruments such as wind masts and lidars is generally cost-prohibitive. Numerous similarity theory-based and empirical equations exist, such as the logarithmic law of the wall, Monin–Obukhov similarity theory, and power law, among others. These have been utilized in wind resource assessment applications to extrapolate near-surface wind speeds from ground meteorological stations (Bañuelos-Ruedas et al., 2010) or satellites (Optis et al., 2021) to various vertical levels. However, these equations often require additional information that is not typically measured by ground stations. Furthermore, some of these equations are only valid within the surface layer (Basu, 2023); yet, the swept areas of contemporary turbines extend well beyond this layer, complicating their application. At present, the most reliable approach for estimating wind profiles is to use mesoscale models. There have been several such activities going on, such as, the Copernicus Regional Reanalysis for Europe (CERRA) (Schimanke et al., 2021), New European Wind Atlas (NEWA) (Hahmann et al., 2020; Dörenkämper et al., 2020), Dutch Offshore Wind Atlas (DOWA) (Wijnant et al., 2019), Winds of the North Sea in 2050 (WINS50) (Dirksen et al., 2022), Solar and Wind at Gray-Zone resolution (Baki et al., 2024a), to name a few. However, mesoscale model simulations require significant computational resources. In addition, the simulated wind speed profiles are susceptible to grid size and physical parameterizations (Baki et al., 2024b).

In recent years, numerous machine learning (ML) studies have explored extrapolating near-surface wind speeds to rotor-swept heights. Mohandes and Rehman (2018) employed deep neural networks (DNNs) to extrapolate wind speeds from lidar measurements at lower heights to 120 m height, showing superior performance over the empirical local wind shear exponent method. Optis et al. (2021) investigated methods to extrapolate near-surface wind speeds from satellite-based wind atlases to hub heights, with ML models, particularly Random Forest (RF), outperforming traditional empirical methods. They highlighted that ML models trained on a limited number of lidars could accurately extrapolate winds at various surrounding locations. Building on this, Liu et al. (2023) used three ML (RF) models to estimate wind speeds at 120 m, 160 m, and 200 m levels, incorporating large-scale weather features from ERA5 reanalysis and wind speed/direction from a remote sensing device. They concluded that, including meteorological features significantly improved ML model accuracy compared to the empirical power law method. Yu and Vautard (2022) extended this approach, constructing RF and extreme gradient boosting (XGBoost) models to estimate a gridded dataset of 100 m wind speed using meteorological variables from the ERA5 (5th generation European Centre for Medium-Range Weather Forecasts (ECMWF)) reanalysis (Hersbach et al., 2020). However, these ML studies focused on specific extrapolation levels and lack generalization for entire vertical profiles.

## 2. Problem statement

In this article, our objective is to explore a deep learning (DL) approach for wind profile estimation, leveraging coarse-resolution meteorological features from public-domain reanalysis data. Traditionally, wind resource assessments involve collecting observations for 1 year using met-masts, sodars, and lidars and extrapolating winds for other years through the measure–correlate–predict (MCP) approach. To emulate this process, our focus is on training a DL model for 1 year and predicting for a different year. As a proof-of-concept, we utilized simulated high-resolution wind profiles from the CERRA reanalysis and coarse-resolution meteorological features from the ERA5 at a specific location. The challenge lies in generalizing the methodology for easy adoption with any datasets, which could be observational data from various instruments.

To achieve this, we initially approximate the CERRA wind profiles using Chebyshev polynomials, representing them with five coefficients. Using these coefficients as targets and the ERA5 meteorological features as inputs, a DL model is trained. While the aforementioned ML models like RF and XGBoost excel in regression problems, they are designed to predict single targets. Given our objective of predicting all coefficients simultaneously to obtain collective wind profile information, we opt for state-of-the-art TabNet (Arik and Pfister, 2021), an attention-based sequential DL model. Additional details about the data used are provided in Section 3.1, the Chebyshev coefficient estimation is explained in Section 3.2, and the training procedure is outlined in Section 3.3. The results are presented in Section 4, with concluding remarks in Section 5.

## 3. Data and Methodology

### 3.1. Data

#### 3.1.1. Wind speed at different height levels

The CERRA is a state-of-the-art reanalysis developed through the collaborative efforts of the Swedish Meteorological and Hydrological Institute (SMHI), Norwegian Meteorological Institute (MET Norway), and Météo-France. CERRA provides wind speeds at 12 vertical levels: 10, 15, 30, 50, 75, 100, 150, 200, 250, 300, 400, and 500 m above sea or ground level. The dataset is available as analysis every third hour and as forecast at lead hours of 1, 2, and 3. In this study, we utilized the three-hourly analyses and corresponding forecasts at lead hours of 1 and 2, creating an hourly dataset. We analyzed data spanning 2 years, from 0000UTC on January 1, 2000, to 2300UTC on December 31, 2001. We extracted a time-series of wind profiles at the FINO1 site (54.0143N, 6.58385E), a location noted for extensive wind power meteorology research (Durán et al., 2020).

#### 3.1.2. Meteorological variables

This study utilizes 34 meteorological variables from the publicly available and globally acclaimed ERA5 reanalysis data as drivers for wind profiles, which are presented in Table 1. Of these, 25 variables are selected based on the studies of Kartal et al. (2023), adhering to the same naming convention and descriptions as mentioned in Table 1 of their work.

### 3.2. Estimating Chebyshev coefficients

Chebyshev polynomials allow one to approximate a function with smallest error as follows (Mason and Handscomb, 2002):

$$U(z) = \sum_{n=0}^{\infty} C_n T_n(z) \quad (1)$$

Here, we want to approximate the wind speed profile  $U(z)$  with the combination of Chebyshev polynomials  $T_n(z)$  multiplied by the corresponding coefficients  $C_n$ . The polynomials of the first kind can be estimated through recurrence relations as follows:

**Table 1.** Description of the meteorological variables adopted from the ERA5 reanalysis

Type	Variable	Equation	Description	Units
Derived	$\mathbf{W}_{10}$	$\sqrt{U_{10}^2 + V_{10}^2}$	Wind speed at 10 m a.g.l. computed from zonal and meridional components	$\text{m s}^{-1}$
Derived	$\mathbf{W}_{100}$	$\sqrt{U_{100}^2 + V_{100}^2}$	Wind speed at 100 m a.g.l. computed from zonal and meridional components	$\text{m s}^{-1}$
Derived	$\alpha$	$\frac{\log(\mathbf{W}_{100}/\mathbf{W}_{10})}{\log(100/10)}$	Power-law exponent of wind profile within 10–100 m above ground level (a.g.l).	–
Derived	$\mathbf{W}_{975}$	$\sqrt{U_{975}^2 + V_{975}^2}$	Wind speed at 975 hPa computed from zonal and meridional components	$\text{m s}^{-1}$
Derived	$\mathbf{W}_{950}$	$\sqrt{U_{950}^2 + V_{950}^2}$	Wind speed at 950 hPa computed from zonal and meridional components	$\text{m s}^{-1}$
Derived	$\Delta\mathbf{W}_{975-100}$	$\mathbf{W}_{975} - \mathbf{W}_{100}$	Difference in wind speed between 975 hPa level and 100 m level	$\text{m s}^{-1}$
Derived	$\Delta\mathbf{W}_{950-975}$	$\mathbf{W}_{950} - \mathbf{W}_{975}$	Difference in wind speed between 950 hPa level and 975 hPa level	$\text{m s}^{-1}$
Raw	$\mathbf{u}_*$		Friction velocity	$\text{m s}^{-1}$
Raw	$\mathbf{W}_{p10}^i$		Instantaneous wind gust at 10 m a.g.l.	$\text{m s}^{-1}$
Raw	$T_2$		Air temperature at 2 m a.g.l.	K
Raw	$T_0$		Skin temperature	K
Raw	$T_s$		Upper-level soil temperature	K
Raw	$T_{d2}$		Dew point temperature at 2 m a.g.l.	K
Raw	$P_0$		Mean sea level pressure	Pa
Raw	$\mathbf{H}$		Boundary layer height	m
Raw	$h_{cb}$		Cloud base height	m
Raw	$H_S$		Instantaneous surface sensible heat flux	$\text{W m}^{-2}$
Raw	$H_L$		Instantaneous moisture flux	$\text{kg m}^{-2} \text{s}^{-1}$
Raw	TCC		Total cloud cover	–
Raw	LCC		Low-level cloud cover	–
Raw	CAPE		Convective available potential energy	$\text{J kg}^{-1}$
Raw	CIN		Convective inhibition	$\text{J kg}^{-1}$
Raw	$\bar{\epsilon}$		Energy dissipation rate in boundary layer	$\text{J m}^{-2}$
Raw	$T_{975}$		Air temperature at 975 hPa	K
Raw	$T_{950}$		Air temperature at 950 hPa	K
Derived	$\Delta T_1$	$T_2 - T_0$	Difference in air and skin temperatures	K
Derived	$\Delta T_2$	$T_0 - T_s$	Difference in skin and soil temperatures	K
Derived	$\Delta T_3$	$T_2 - T_{d2}$	Temperature dew point spread	K
Derived	$\Delta T_4$	$T_{975} - T_2$	Difference in temperatures between 975 hPa and 2 m a.g.l.	K
Derived	$\Delta T_5$	$T_{950} - T_{975}$	Difference in temperatures between 950 hPa and 975 hPa	K
Derived	HRSin	$\sin\left(\frac{2\pi\text{Hour}}{24}\right)$	Sine encoding of hours	–
Derived	HRCos	$\cos\left(\frac{2\pi\text{Hour}}{24}\right)$	Cosine encoding of hours	–
Derived	DYSin	$\sin\left(\frac{2\pi\text{Day}}{365}\right)$	Sine encoding of Julian days	–
Derived	DYCos	$\cos\left(\frac{2\pi\text{Day}}{365}\right)$	Cosine encoding of Julian days	–

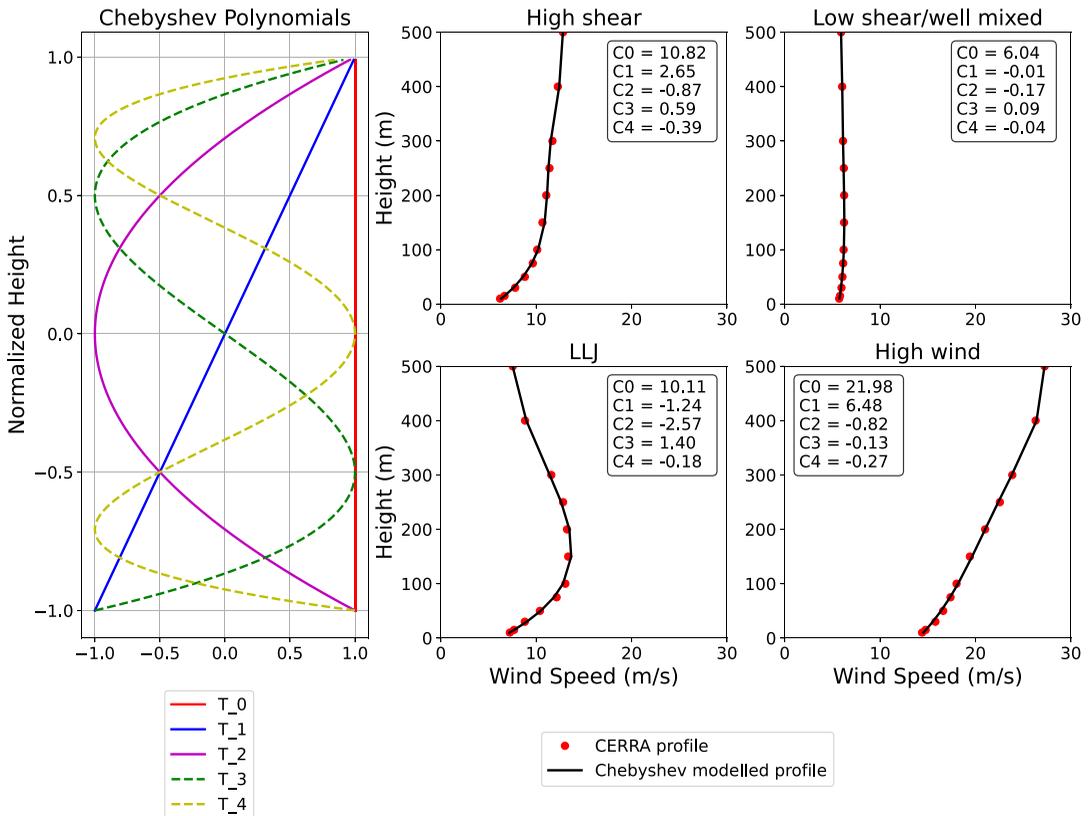
$$T_0(z) = 1 \tag{2}$$

$$T_1(z) = z \tag{3}$$

$$T_{n+1}(z) = 2zT_n(z) - T_{n-1}(z) \tag{4}$$

In this study, we employed fourth-order Chebyshev polynomials (Figure 1, 1st column). Once computed, these polynomials transform the problem into a system of linear equations, facilitating the estimation of coefficients through methods such as solving linear equations or matrix inversion. The variable  $z$  is normalized between  $-1$  and  $1$  in real data before estimating coefficients. For a wind profile with 12 vertical levels and Chebyshev polynomials of order 4 ( $T_0$  to  $T_4$ ), five Chebyshev coefficients ( $C_0$  to  $C_4$ ) are estimated, reducing the wind profile's complexity to five coefficients. We would like to point out that the approximation strategy allows us to effectively handle wind profile data from various sources. Whether the data is observed or simulated up to 500 m, and even if it does not align with the specific vertical levels of the CERRA data, the Chebyshev coefficient estimation process remains unaffected. This adaptability ensures that the proposed methodology is versatile across different wind datasets.

$T_0$  represents a constant line, with  $C_0$  approximating mean wind speed. Similarly,  $T_1$  corresponds to a diagonal line, and  $C_1$  approximates wind shear. The parabolic profile of  $T_2$  is captured by  $C_2$ , representing



**Figure 1.** Column 1: an illustration of fourth order Chebyshev polynomials plotted against the normalized height  $z = [-1, 1]$ . The remaining figures display the vertical profiles of wind speed from CERRA alongside those approximated by Chebyshev polynomials, for four well-known categories of wind regimes: high shear, low shear/well-mixed, low-level jets (LLJ), and high wind.

curvature in the wind profile. These three coefficients are expected to capture a significant portion of the wind profile, while higher-order coefficients account for small-scale variations.

To illustrate the capability of Chebyshev coefficients, we compared four wind profiles from the CERRA dataset with their Chebyshev approximations (Figure 1). These types of profiles, namely high shear, low shear/well-mixed, LLJ, and high wind, selected from Durán et al. (2020), are crucial for wind energy applications. The figures demonstrate the effective approximation of CERRA wind profiles by Chebyshev coefficients.

3.3. Experimental setup

Figure 2(a) illustrates a complete flowchart of the training procedure adopted in this study. First, the ERA5-based predictor meteorological variables as inputs and the estimated Chebyshev coefficients as targets are stacked side-by-side as a tabular dataset. Next, the entire data of year 2001 is kept aside for testing purpose. Now, among the data of year 2000, randomly selected six consecutive days of each month are used for validation purposes. The remaining data is adopted for training the models. This ensures that the training and validation data covers seasonality given 1 year sample size. The data-splitting strategy is illustrated in Figure 2(b). After splitting, there are 7056 samples in training, 1728 samples in validation, and 8760 samples in testing. After this, a min-max normalization function is constructed on the targets of training data, using which the targets of training and validation data are normalized.

The TabNet consists of several hyperparameters, in which we chose to tune  $n_d$  (width of decision prediction layer),  $n_{steps}$  (number of steps in the architecture),  $N_{independent}$  (number of independent Gated Linear Units),  $n_{shared}$  (number of shared Gated Linear Units), and  $\gamma$  (coefficient for feature reuseage). The readers are encouraged to peruse the paper by Arik and Pfister (2021) for more information about the architecture and the hyperparameters. Next, a random search is employed for tuning the model

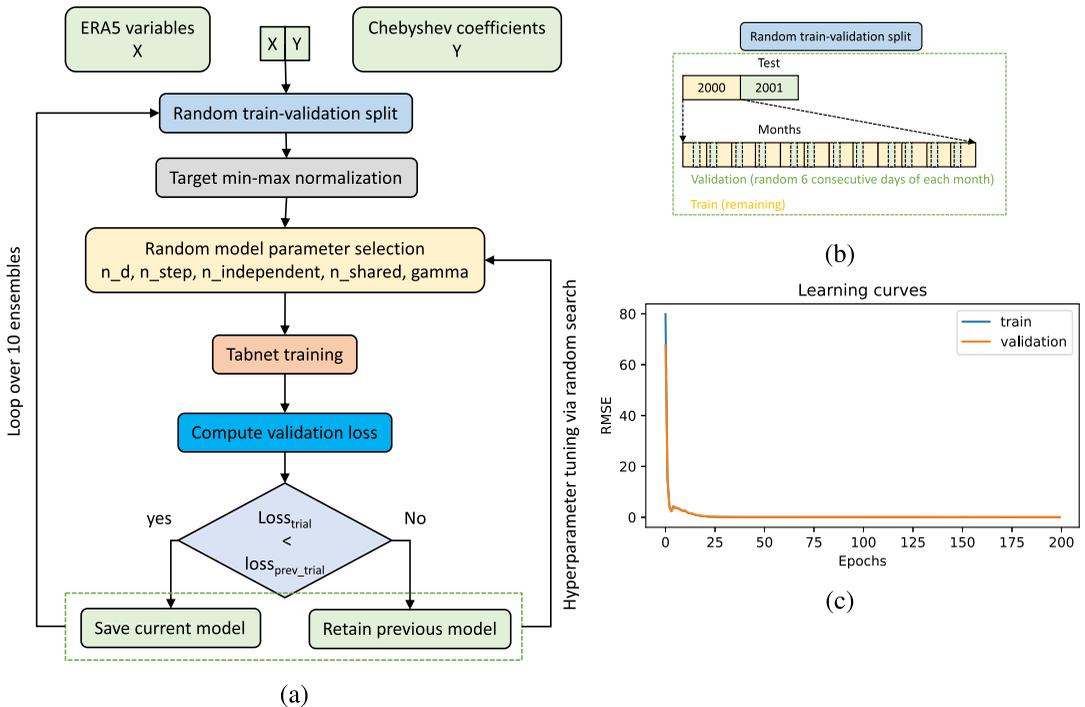


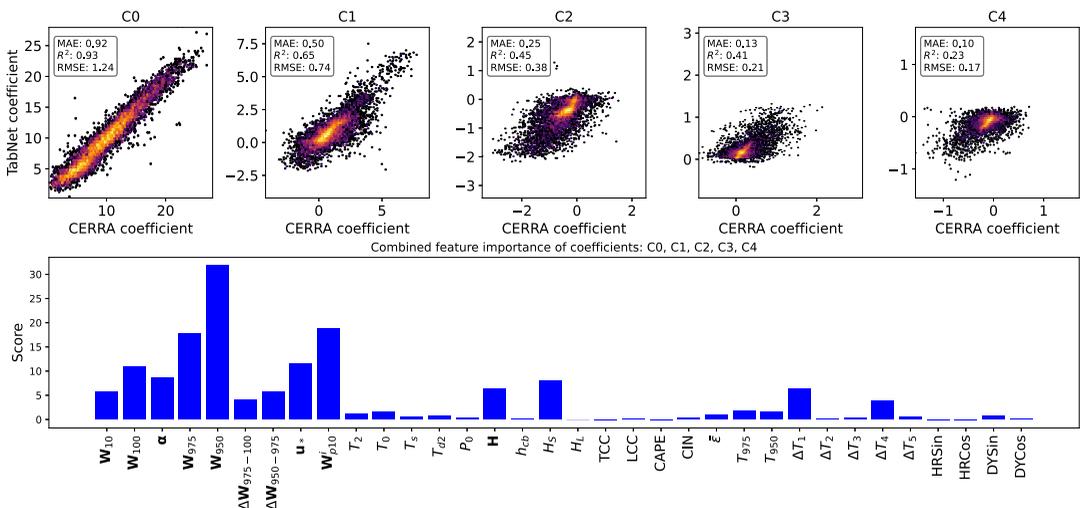
Figure 2. (a) Flowchart of the experimental setup used in this study to train the TabNet. (b) Our strategy of splitting the entire dataset into train, validation, and test. (c) Loss curves of one of the trained model, in which the train and validation RMSE values are plotted against the training epochs.

hyperparameters from parameter spaces of  $n_d:[4,8,16]$ ,  $n_{steps}:[3,4,5]$ ,  $n_{independent}:[1,2,3,4,5]$ ,  $n_{shared}:[1,2,3,4,5]$ , and  $gamma:[1.1,1.2,1.3,1.4]$ . Using these parameters, the TabNet model is trained with mean squared error (MSE) as the loss function and evaluated on the validation dataset. After training, the validation loss is calculated, and the trained model is saved as an external file. Following hyperparameter tuning, the training process is replicated across 10 ensembles, each initiated with a random train-validation split. Each ensemble saves its best-performing model. The inner loop of hyperparameter tuning enhances the model's robustness by optimizing performance across different hyperparameters on the same dataset. The outer loop, on the other hand, is crucial for generating reliable predictions through ensemble modeling. A sample learning curve from one of the saved models is depicted in Figure 2(c), demonstrating that the models are training effectively.

### 4. Results

Model predictions are generated for the test data, and performance is evaluated using key metrics such as mean absolute error (MAE), coefficient of determination ( $R^2$ ), and root mean square error (RMSE). Figure 3 (first row) illustrates a comparison of predicted coefficients from one of the ensemble models with respect to the test data using bivariate histograms. Among the coefficients,  $C_0$  and  $C_1$  predominantly align along the diagonal line, displaying narrower spreads. Notably, the model exhibits its highest predictability for  $C_0$  with an  $R^2$  of 0.93 and moderate predictability for  $C_1$  with an  $R^2$  of 0.65. Conversely, the remaining coefficients frequently register values close to zero, with less probable values demonstrating a wider spread, indicating lower predictability for these coefficients.

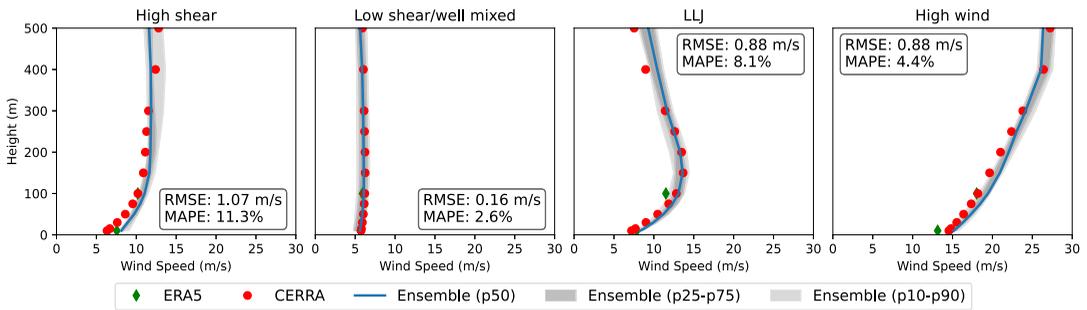
Additionally, the impact of input meteorological variables on the predicted Chebyshev coefficients is estimated by measuring (permutation) feature importance. This is done by quantifying the reduction in the score when a specific feature is absent. From the feature importance, as shown in 3 (second row), it is evident that the meteorological variables directly related to wind speed ( $W_{10}$  to  $W_{p10}^i$ ) are showing significant influence on the coefficients, which is expected. A major finding from the feature importance is that identification of atmospheric stability-related variables, such as instantaneous surface sensible heat



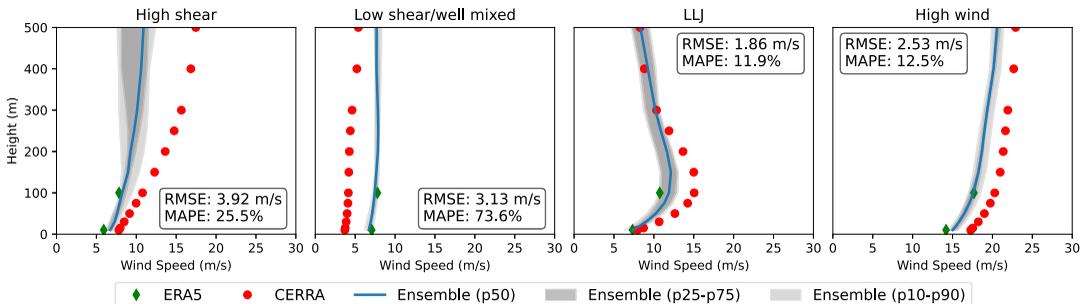
**Figure 3.** First row: a comparison of Chebyshev coefficients ( $C_0, C_1, C_2, C_3$  and  $C_4$ ) between the test data and the model predictions using bivariate histograms. The probability of occurrence is represented on a log scale with the color increasing from dark (low probability) to light (high probability). The evaluation scores, namely MAE,  $R^2$ , and RMSE for each coefficient are provided in the text boxes. Second row: the combined feature importance of input variables based on the test data.

flux ( $H_S$ ), difference in air and skin temperatures ( $\Delta T_1$ ), and the difference in 975 hPa and 2 m air temperatures ( $\Delta T_4$ ), are also exerting significant influence on wind profiles. It is evident that the boundary layer height ( $\mathbf{H}$ ) is also an important input feature for wind profile estimation. By identifying these influential variables, we can accurately characterize various wind profiles. This improved characterization aids in the better estimation of wind power and turbine loads.

Since the main objective of this study is to predict the wind profiles given coarse-resolution meteorological features from the ERA5 dataset, we reconstruct the wind profiles using the predicted Chebyshev coefficients from the test data. A sample of four profiles from the 10 ensemble model predictions is presented in Figure 4. From this figure, it is evident that the predicted profiles are in agreement with the CERRA-generated wind profiles, and the uncertainty bounds are quite narrow. The evaluation metrics RMSE and mean absolute percentage error (MAPE), which are computed between the CERRA profile and the median profile, further corroborate the strength of the TabNet model. However, not every prediction turned out to be accurate, as shown in Figure 5. For these selected profiles, the error metrics (RMSE and MAPE) are quite high. Careful analysis of Figure 5 reveals a substantial discrepancy between the ERA5 and CERRA wind speeds, which may account for the poor predictions observed in these instances. The TabNet-generated profiles lie in between the ERA5 and CERRA wind speeds. This



**Figure 4.** A comparison of vertical profiles of wind speed from CERRA and the 10 ML model predictions, on four instances of test data for the selected wind regimes. Blue line represents the 50th percentile of the ensemble, darker shade represents the ensemble between 25th and 75th percentiles, and the lighter shade represents the ensemble between 10th and 90th percentile. The wind speed from ERA5 at 10 m ( $\mathbf{W}_{10}$ ) and 100 m ( $\mathbf{W}_{100}$ ) are illustrated using green diamonds. The evaluation scores, RMSE and MAPE are computed between the CERRA and the median profile for each wind regime, are provided in the text boxes.



**Figure 5.** Same as Figure 4, but a different set of time instances of the test data for the selected wind regimes.

systematic error could be reduced by increasing the training sample size and including samples from diverse field sites.

## 5. Conclusion

In this work, we introduced a proof-of-concept methodology for estimating wind speed profiles using large-scale meteorological features from the ERA5 reanalysis using TabNet, an attention-based DL model. Chebyshev polynomials were utilized to approximate wind profiles with only five coefficients. Instead of using wind speeds at multiple heights as targets, we used these Chebyshev coefficients as targets. This approximation strategy ensures that our TabNet-based methodology can be used with different types of wind datasets, which may contain wind speeds at varying heights. Results indicated that TabNet effectively captured nonlinear dependencies between meteorological features and wind profiles across different wind regimes. Feature importance analysis highlighted the significant influence of wind speed, atmospheric stability-related variables, and boundary layer height on the Chebyshev coefficients.

Nonetheless, there is significant room for improving the accuracy of the proposed approach. Specifically, the performance of the TabNet in predicting higher-order Chebyshev coefficients is less than satisfactory. In our future work, we will incorporate additional meteorological variables from reanalysis datasets and will also explore alternative DL models. Additionally, we plan to apply this methodology to diverse wind profile datasets, collected at various geographical locations, and predict over several years in a round-robin manner.

**Open peer review.** To view the open peer review materials for this article, please visit <http://doi.org/10.1017/eds.2024.41>.

**Author contribution.** Conceptualization: H.B., S.B.; Methodology: H.B., S.B.; Data curation: H.B.; Data visualization: H.B.; Writing original draft: H.B.; Writing review and editing: H.B., S.B. All authors approved the final submitted draft.

**Competing interest.** None.

**Data availability statement.** The ERA5 and CERRA reanalysis are downloaded from ECMWF CDO, available at <https://cds.climate.copernicus.eu/cdsapp#!/search?type=dataset>. The entire workflow and the scripts used in training can be found at <https://github.com/HarishBaki/CI2024.git>.

**Funding statement.** This study was partially supported by the EU-SCORES and the Winds of the North Sea in 2050 (WINS50) projects. The EU-SCORES project is a part of the European Union's Horizon 2020 research and innovation programme under grant agreement No 101036457. The WINS50 project was sponsored by the Rijksdienst voor Ondernemend Nederland via the Top Sector Energy programme.

**Provenance statement.** This article is part of the Climate Informatics 2024 proceedings and was accepted in Environmental Data Science on the basis of the Climate Informatics peer review process.

**Ethical standard.** The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

## References

- Arik SÖ and Pfister T (2021) TabNet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 6679–6687.
- Baki, H., Basu, S., & Lavidas, G. (2024a). Estimating the offshore wind power potential of Portugal by utilizing gray-zone atmospheric modeling. *Journal of Renewable and Sustainable Energy*, 16(6). <https://doi.org/10.1063/5.0222974>
- Baki, H., Basu, S., & Lavidas, G. (2024b). Modelling Frontal Low-Level Jets and Associated Extreme Wind Power Ramps over the North Sea. *Wind Energy Science Discussion* [preprint], <https://doi.org/10.5194/wes-2024-99>.
- Bañuelos-Ruedas F, Angeles-Camacho C and Rios-Marcuello S (2010) Analysis and validation of the methodology used in the extrapolation of wind speed data at different heights. *Renewable and Sustainable Energy Reviews* 14, 2383–2391.
- Basu S (2023) Vertical wind speed profiles in atmospheric boundary layer flows. *Wind Energy Engineering: A Handbook for Onshore and Offshore Wind Turbines* (2 ed., pp. 75–85). Elsevier. <https://doi.org/10.1016/B978-0-323-99353-1.00031-1>
- Dimitrov N, Natarajan A and Kelly M (2015) Model of wind shear conditional on turbulence and its impact on wind turbine loads. *Wind Energy* 18, 1917–1931.

- Dirksen M, Wijnant I, Siebesma A, Baas P and Theeuwes N** (2022) *Validation of Wind Farm Parameterisation in Weather Forecast Model HARMONIE-AROME: Analysis of 2019*. Netherlands: Delft University of Technology.
- 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, Sile T, Trei W, Žagar M, Badger J, Gottschall J, Sanz Rodrigo J and Mann J** (2020) The making of the new European wind atlas – Part 2: Production and evaluation. *Geoscientific Model Development* 13, 5079–5102.
- Durán P, Basu S, Meißner C and Adaramola MS** (2020) Automated classification of simulated wind field patterns from multiphysics ensemble forecasts. *Wind Energy* 23, 898–914.
- Elliott DL and Cadogan JB** 1990, *Effects of Wind Shear and Turbulence on Wind Turbine Power Curves*, Technical Report, Richland, WA (USA): Pacific Northwest Lab.
- Guo Y, Wang H and Lian J** (2022) Review of integrated installation technologies for offshore wind turbines: Current progress and future development trends. *Energy Conversion and Management* 255, 115319.
- Gutierrez W, Ruiz-Columbie A, Tutkun M and Castillo L** (2017) Impacts of the low-level jet’s negative wind shear on the wind turbine. *Wind Energy Science* 2, 533–545.
- Hahmann AN, Sile T, Witha B, Davis NN, Dörenkämper M, Ezber Y, García-Bustamante E, González-Rouco JF, Navarro J, Olsen BT and Söderberg S** (2020) The making of the new European wind atlas – Part 1: Model sensitivity. *Geoscientific Model Development* 13, 5053–5078.
- Hersbach H, Bell B, Berrisford P, Hirahara S, Horányi A, Muñoz-Sabater J, Nicolas J, Peubey C, Radu R, Schepers D, et al.** (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society* 146, 1999–2049.
- IEC** (2005) 12–1: Power Performance Measurements of Electricity Producing Wind Turbines, British Standard, IEC 61400-12.
- Kartal S, Basu S and Watson SJ** (2023) A decision tree-based measure-correlate-predict approach for peak wind gust estimation from a global reanalysis dataset. *Wind Energy Science*, 8, 1533–1551, <https://doi.org/10.5194/wes-8-1533-2023>.
- Liu B, Ma X, Guo J, Li H, Jin S, Ma Y and Gong W** (2023) Estimating hub-height wind speed based on a machine learning algorithm: Implications for wind energy assessment. *Atmospheric Chemistry and Physics* 23, 3181–3193.
- Mason JC and Handscomb DC** (2002) *Chebyshev Polynomials*. CRC press.
- Mohandes MA and Rehman S** (2018) Wind speed extrapolation using machine learning methods and LiDAR measurements. *IEEE Access* 6, 77634–77642.
- Nagababu G, Srinivas BA, Kachhwaha SS, Puppala H and Kumar SVA** (2023) Can offshore wind energy help to attain carbon neutrality amid climate change? A GIS-MCDM based analysis to unravel the facts using CORDEX-SA. *Renewable Energy* 219, 119400.
- New Atlas** (July 2023) World’s largest wind turbine to feature 16 MW capacity. *New Atlas* URL <https://newatlas.com/energy/worlds-largest-wind-turbine-myse-16-260/>.
- Optis M, Bodini N, Debnath M and Doubrawa P** (2021) New methods to improve the vertical extrapolation of near-surface offshore wind speeds. *Wind Energy Science* 6, 935–948.
- Park J, Manuel L and Basu S** (2015) Toward isolation of salient features in stable boundary layer wind fields that influence loads on wind turbines. *Energies* 8, 2977–3012.
- Schimanke S, Ridal M, Le Moigne P., Berggren L, Undén P, Randriamampianina R, Andrea U, Bazile E, Bertelsen T, Brousseau P, Dahlgren P, Edvinsson L, El Said A, Glington M, Hopsch S, Isaksson L, Mladek R, Olsson E, Verrelle A, Wang Z** (2021) CERRA sub-daily regional reanalysis data for Europe on pressure levels from 1984 to present, Technical Report, Copernicus Climate Change Service (C3S) Climate Data Store (CDS). <https://doi.org/10.24381/cds.a39ff99f> (accessed 10 10 2023).
- Van Sark WG, Van der Velde HC, Coelingh JP and Bierbooms WA** (2019) Do we really need rotor equivalent wind speed? *Wind Energy* 22, 745–763.
- Wagner R, Courtney M, Gottschall J and Lindelöw-Marsden P** (2011) Accounting for the speed shear in wind turbine power performance measurement. *Wind Energy* 14, 993–1004.
- Wharton S and Lundquist JK** (2012) Assessing atmospheric stability and its impacts on rotor-disk wind characteristics at an onshore wind farm. *Wind Energy* 15, 525–546.
- Wijnant I, van Ulf B, van Stratum B, Barkmeijer J, Onvlee J, de Valk C, Knoop S, Kok S, Marseille G, Baltink HK, et al.** (2019) *The Dutch Offshore Wind Atlas (DOWA): Description of the Dataset*. De Bilt: Royal Netherlands Meteorological Institute, Ministry of Infrastructure and Water Management.
- Yu S and 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.