



UNIVERSITÀ
DI PAVIA

UNIVERSITY OF PAVIA
FACULTY OF ENGINEERING
DEPARTMENT OF ELECTRICAL, COMPUTER
AND BIOMEDICAL ENGINEERING
MASTER'S DEGREE IN ELECTRONIC ENGINEERING

MASTER THESIS

**Probabilistic forecasting of wind power production
from multivariate wind forecasts**

Predizione probabilistica della produzione di potenza eolica
a partire da previsione multivariate del vento

Candidate: Alessia Protti Melano

Supervisor: Prof. Giuseppe De Nicolao

Co-supervisor: Dr. Marco Capelletti

A. Y.: 2025 - 2026

Contents

1	Introduction	1
1.1	Context and project aim	1
1.2	Case study	2
1.3	Chapter content	4
2	Wind power forecasting	5
2.1	Chapter overview	5
2.2	Wind power forecasting methods and approaches	5
2.2.1	Scope of the problem	5
2.2.2	Point and probabilistic prediction models	6
2.2.3	Point versus probabilistic forecasting	6
2.2.4	Recalibration in wind power forecasting	7
3	Data	9
3.1	Chapter overview	9
3.2	Wind farm data	9
3.2.1	Wind speed dataset	9
3.2.2	First correlation matrix	10
3.3	Data cleaning	11
3.3.1	Reduced wind speed dataset	11
3.3.2	Second correlation matrix	12
3.4	Wind forecast	13
4	Wind farm power curve	14
4.1	Chapter overview	14
4.2	Introduction and dataset	14
4.3	Data cleaning	15
4.4	Power curve	16

5	Probabilistic recalibration	19
	of wind forecasts	
5.1	Chapter overview	19
5.2	Introduction and employed models	19
5.3	Preprocessing	20
5.4	Models	20
5.4.1	Linear regression	20
5.4.2	PCA and PCR	21
5.4.3	Lasso median regression	26
5.4.4	Other models	27
5.5	Overall findings	28
5.6	Marginal model	29
5.6.1	Pinball loss	29
5.6.2	Empirical coverage and coverage error	30
5.6.3	PIT	30
5.7	Actual recalibration	31
6	Probabilistic power forecast	35
6.1	Chapter overview	35
6.2	Predicted power	35
6.3	Comparison with a direct method	37
7	Conclusions	41
7.1	Conclusion	41
7.2	Future work	42
8	Appendix	44
8.1	Chapter overview	44
8.2	Appendix	45
8.2.1	Scattermatrices	45
8.2.2	Histograms	53
8.2.3	Windrose	56
8.2.4	Scatterplots	57
8.2.5	Spaghetti plots	60
8.2.6	PCAs	61

List of Figures

1.1	First snapshot of wind direction and speed for each turbine . . .	3
1.2	Second snapshot of wind direction and speed for each turbine . .	3
3.1	Wind speed dataset	10
3.2	Correlation matrix of wind speed dataset	11
3.3	Reduced wind speed dataset	12
3.4	Correlation matrix of reduced wind speed dataset	12
4.1	Measured wind speed and power dataset	14
4.2	Scatterplot of raw data, measured wind speed vs power	15
4.3	Scatterplot of cleaned data, measured wind speed vs power . . .	16
4.4	Linear GAM fitted and plotted on 2024 data with its confidence interval	17
4.5	Linear GAM fitted on 2024 data and plotted on 2025 data (purple) and linear GAM fitted and plotted on 2025 (orange)	18
5.1	Cumulative explained variance of the PCA per number of principal components	22
5.2	Scoreplot of data projected on the first two principal components	23
5.3	Biplot of data projected on the first two principal components . .	24
5.4	Barplots of the loadings of first and second principal components	24
5.5	mean absolute errors and root mean squared errors of the eight PCR's	25
5.6	Mean absolute errors and root mean squared errors of the lasso quantile regression varying α	26
5.7	Lasso quantile regression coefficients varying alpha	27
5.8	Quantile distributions of the measured wind speed (training: 2024, test: 2025)	29
5.9	Pinball loss of the marginal model (training: 2024, test: 2025) .	30

5.10	Empirical coverage and coverage error of the marginal model (training: 2024, test: 2025)	30
5.11	PIT of the marginal model (training: 2024, test: 2025)	31
5.12	Goodness of fit plots of the lasso quantile regression	32
5.13	Empirical coverage and coverage error of the lasso quantile regression (training: 2024, test: 2025)	33
5.14	PIT of the lasso quantile regression (training: 2024, test: 2025) .	33
5.15	Pinball loss of the lasso quantile regression (training: 2024, test: 2025)	33
5.16	Comparison between the pinball loss of the marginal model (left) and the pinball loss of the lasso quantile regression (right)	34
6.1	Scatterplot of lasso quantile regression versus TSO power with linear GAM	35
6.2	Goodness of fit plots of the predicted power	36
6.3	Pinball loss of the predicted power (training: 2024, test: 2025) .	36
6.4	Empirical coverage and coverage error of the predicted power (training: 2024, test: 2025)	37
6.5	PIT of the predicted power (training: 2024, test: 2025)	37
6.6	Direct method to predict the power with additive B-splines quantile regression fitted on 2024 and 2025 data	38
6.7	Pinball loss of the additive B-splines quantile regression (training: 2024, test: 2025)	38
6.8	Empirical coverage and coverage error of the additive B-splines quantile regression (training: 2024, test: 2025)	39
6.9	PIT of the additive B-splines quantile regression (training: 2024, test: 2025)	39
6.10	Comparison between the pinball loss of the additive B-splines quantile regression (left) and the pinball loss of the predicted power (right)	39
8.1	Scattermatrix of the complete dataset (before the subset selection of the features), with the windspeed forecasts at the four heights (10, 50, 80 and 100 metres) in the five grid points (P1-P5)	45
8.2	Scattermatrix of the reduced dataset (after the subset selection of the features), with the windspeed forecasts at the two heights (10 and 80 metres) in the five grid points (P1-P5)	46

8.3	Scattermatrix of the P5 dataset, with the wind speed forecasts at the four heights (10, 50, 80 and 100 metres) in the P5 grid point	47
8.4	Scatter matrix of the expanded dataset, with the windspeed forecasts at the two heights (10 and 80 metres) in the five grid points (P1-P5) and the Month, Day and Hour, using the Year as hue . . .	48
8.5	Scattermatrix of the expanded dataset of 2024, with the wind-speed forecasts at the two heights (10 and 80 metres) in the five grid points (P1-P5) and the Year, of 2024 data	49
8.6	Scattermatrix of the expanded dataset of 2025, with the wind-speed forecasts at the two heights (10 and 80 metres) in the five grid points (P1-P5) and the Year, of 2025 data	50
8.7	Scattermatrix of the March dataset, with the windspeed forecasts at the two heights (10 and 80 metres) in the five grid points (P1-P5) and the Year, Month, Day and Hour, of March data	51
8.8	Scattermatrix of the Noon dataset, with the windspeed forecasts at the two heights (10 and 80 metres) in the five grid points (P1-P5) and the Year, Month, Day and Hour, of Noon data	52
8.9	Histogram of the measured wind speed with the wind direction as hue	53
8.10	Histograms of the measured wind speed divided by wind direction intervals. This is an expansion of figure 8.9	53
8.11	Histogram of the measured wind speed with the forecast wind speed as hue for each windspeed forecasts at the two heights (10 and 80 metres) in the five grid points (P1-P5)	54
8.12	Histograms of the measured wind speed divided by wind speed intervals for the winspeed forecast at 80 metres in P5. This is an expansion of figure 8.11	55
8.13	Windrose of the measured wind speed versus wind direction. The former is used as hue and its frequency is shown as percentages, while the latter is depicted by using NESW coordinates	56
8.14	Scatterplot of the hour (represented with seconds from midnight) versus the wind direction with wind speed as hue	57
8.15	Scatterplots of the measured wind speed (the absolute distance from the zero) and the wind direction components of a subset of the turbines	58

8.16	Scatterplots of the wind speed forecasts (the absolute distance from the zero) and the wind direction components of a subset of the turbines	59
8.17	Spaghetti plot of the time series of the power of three turbines (T_04, T_14, T_16) to inspect their behavior	60
8.18	Spaghetti plot of the time series of the wind speed of three turbines (T_04, T_14, T_16) to inspect their behavior	60
8.19	First two principal components of the season dataset, with the windspeed forecasts at the two heights (10 and 80 metres) in the five grid points (P1-P5) and the Season, which is the hue. In particular 0 is winter, 1 is spring, 2 is summer and 3 is autumn. .	61
8.20	First two principal components of the day periods dataset, with the windspeed forecasts at the two heights (10 and 80 metres) in the five grid points (P1-P5) and the Day Period, which is the hue. In particular 0 is from 00 a.m. to 5.59 a.m., 1 is from 6 a.m. to 11.59 a.m., 2 is from 12 p.m. to 5.59 p.m. and 3 is from 6 p.m. to 11.59 p.m.	62

List of Tables

5.1	Linear regression model coefficients	21
5.2	Model coefficients of first PCR	25
5.3	Model coefficients of second PCR	25
5.4	Unscaled model coefficients of the lasso quantile regression . . .	27
5.5	Test metrics comparison among the models	28

Abstract

The aim of the project discussed in this thesis was to find a modular and interpretable method for wind power forecasting from wind forecasts, a relevant problem for environmental, management and monetary reasons. In particular, our approach, supported by the literature, strove to employ two major strategies. The first was to use probabilistic forecasting, as it is more informative and complete than its point counterpart, with all the necessary attention to the problem of how to evaluate its performance. The second was to introduce a recalibration of the wind speed forecasts and then, from its output, to predict the wind power. This two-step model aimed at providing more modularity and interpretability to the process.

Our case study was a wind farm modeled as a virtual turbine, of which we knew the wind speed and power. Focusing therefore on the former, we analyzed our dataset, which comprised the measured wind speed and twenty wind speed forecasts, and identified its most prominent issues, which we tried to mitigate by reducing the dataset to a subset of selected features that would be necessary in the recalibration phase. Then, considering the latter, we performed data cleaning and found the power curve via a linear generalized additive model using the measured wind speed versus the power from a TSO dataset.

We subsequently looked for the best possible model, in terms of performance and reduction of the dimensionality of the covariate space, for the recalibration phase. We chose, after various trials, a lasso quantile regression, which was compared to the marginal model to assess its quality.

Having obtained the recalibrated wind speed and the power curve, we predicted the wind power in two steps: first applying the lasso quantile regression to the wind speed forecasts, and then feeding the recalibrated wind speed as input to the power curve. We compared our approach to a direct method and found their performances to be equivalent, a result we considered satisfactory, as our project reached adequate performance while also being more scalable and informative. Moreover, modeling the wind farm as a single comprehensive turbine proved effective, non-negligibly reducing the complexity of the problem.

Chapter 1

Introduction

1.1 Context and project aim

As renewable energy takes on a bigger role in today's energy landscape, the need for accurate wind power forecasting has significantly increased, given its importance for the correct operation of the power system.

The intermittency and randomness of wind energy cause fluctuations in the output, which affect the management of the network (encompassing tasks such as grid scheduling and load balancing), and also raise difficulties for bidding on the energy market. This can lead to demand issues and monetary losses. It is also worth reiterating the importance of wind power in this day and age and its role in the clean energy transition, since it contributes to minimizing reliance on fossil fuels (Liu et al. (2025)).

This leads us to the aim of our work, that is, a comprehensive and interpretable approach to achieving wind power forecasting from wind speed forecasts, with the additional challenge of making it as modular as possible.

In particular, we have two main objectives. The first is to use probabilistic forecasting instead of its point counterpart. As the latter tends to be the more common strategy employed in such cases, we aim to go one step further and use the former, since it does not just predict a value, but also its distribution, thereby adding insight into the confidence of the model and the uncertainty of the forecast. Hence, since this is a more complex approach, assessing its performance becomes a challenging task, one that we tackled using specific metrics, with the added difficulty of interpreting their results (Gneiting et al. (2007)). By managing these matters, we strive to achieve a more informative and comprehensive model.

Our second objective is to use an indirect approach, that is, to first recalibrate

the wind speed from the forecasts and then, from the output of this first step, to predict the actual power by building a power curve, instead of directly predicting the power from the forecasts. This is meant to attain an approach that is, first of all, more interpretable than a black-box solution, and also more modular, as the two steps are connected yet independent and can be fitted to different wind farms without disrupting the overall method.

1.2 Case study

In this project, our case study was a specific wind farm, for which we had data on actual power, wind direction and wind speed for each turbine, along with wind speed forecasts in five grid points (P1-P5) of the farm at 10, 50, 80 and 100 metres. These forecasts have a horizon of 48 hours, and two runs were considered to ensure full coverage of the target day T, one from day T-2 for data between 00:00–17:59 and the other from day T-1 for data between 18:00–23:59.

Figures 1.1 and 1.2 show two snapshots of the wind farm at different timestamps, to better illustrate our scenario and the information at our disposal. In particular, on the axes we have the direction from which the wind blows, divided into the West-East component on the x-axis (where West is negative and East is positive) and the South-North component (where South is negative and North is positive). The absolute distance from zero is the wind speed, and each dot represents the wind direction and speed for a given turbine.

Just from these two snapshots we can see that the values can be widespread, as in figure 1.1, or more condensed, as in figure 1.2, but with some turbines still diverging from the rest. This kind of behavior may be difficult to model consistently, as it may stem from a number of possible phenomena.

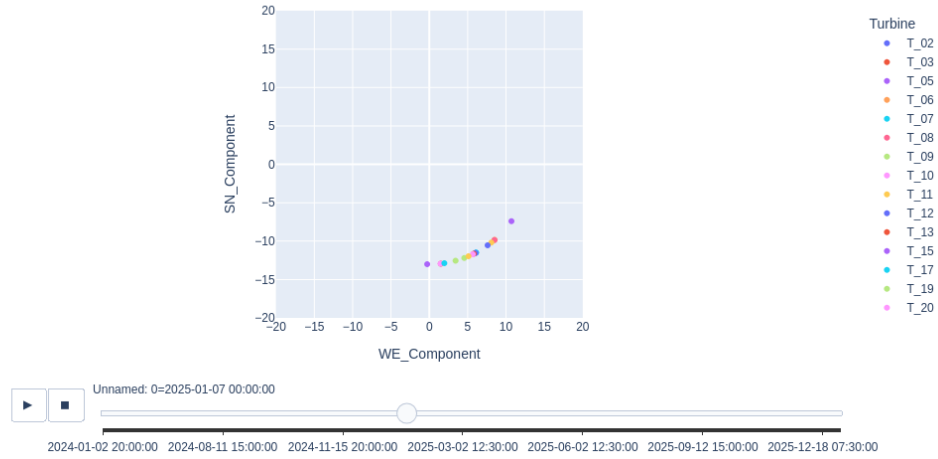


Figure 1.1: First snapshot of wind direction and speed for each turbine

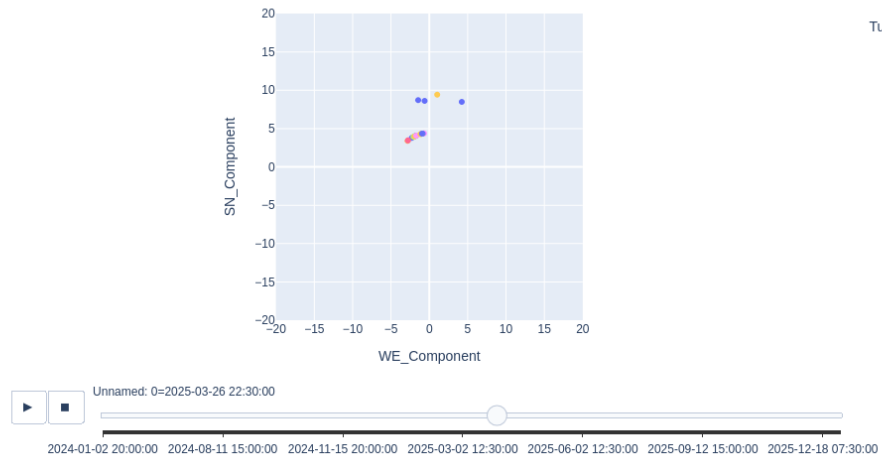


Figure 1.2: Second snapshot of wind direction and speed for each turbine

Therefore, to obtain a better representation of the problem, we also built a virtual turbine, created with the objective of representing the whole farm as a single turbine with its own wind speed, direction and power (this was obtained by computing the circular median of each feature). We focused on modeling this virtual turbine, through its wind speed and power output, to gain a global understanding of the farm, with its complexity reduced to improve the interpretability of the results.

To do so we employed the Python programming language along with various packages that will be mentioned where appropriate.

1.3 Chapter content

More specifically, this thesis is organized as follows:

- In **chapter two** we expand on the literature regarding wind power forecasting; we focus on the general problem and the viable solutions to tackle it, along with the related models. We also discuss whether point or probabilistic forecasting is preferable for predicting power, and we explore the advantages of employing a recalibration technique.
- In **chapter three** our dataset for the prediction of wind speed is thoroughly examined, with a focus on the features and the target variable. We then proceed with our data analysis, focusing on the possible issues arising from the dataset and conducting a correlation study, which leads to a new dataset comprising selected features. The latter is then the starting point for the models we employed for the recalibration.
- In **chapter four** we consider the dataset constructed to predict power from the measured wind speed. We plot our raw data and then clean it in order to obtain a better model. The power curve is therefore fitted via a linear generalized additive model, ready to be used in the second step of our approach.
- In **chapter five** we delve into the recalibration phase. In particular, we look for a viable model and, after various trials and a general performance assessment among the candidate models, we choose a lasso quantile regression. The latter is implemented and its metrics are computed for a final comparison with the marginal model.
- In **chapter six** the power is actually predicted, using the recalibrated wind speed from the lasso quantile regression and the power curve obtained through the linear generalized additive model. Its metrics are computed and compared to those attained via a direct method, which we built to evaluate our results.

Chapter 2

Wind power forecasting

2.1 Chapter overview

In this chapter we lay the foundation for how the problem of wind power forecasting is treated in the state of the art and what methods are usually employed. It also addresses the advantages and disadvantages of point versus probabilistic prediction, as well as the use of recalibration.

2.2 Wind power forecasting methods and approaches

2.2.1 Scope of the problem

In recent years the problem of wind power forecasting has been thoroughly discussed, with a variety of possible methodologies proposed to tackle it, whose merits have been considered in the literature.

Given the uncertainties and fluctuations in wind speed and the dependency on features such as geographical and environmental conditions, local terrain characteristics, temporal effects and seasonal patterns, wind power management can be challenging for distribution networks, creating issues in balancing the supply and demand of energy. Moreover, wind power forecasting is vital for wind farm maintenance, unit commitment planning and profit maximization for power traders.

It is therefore necessary to reduce the risk associated with these uncertainties and to deliver accurate predictions.

2.2.2 Point and probabilistic prediction models

Various procedures have been employed to do so, spanning a spectrum from persistence methods to neural networks. The former are usually used as a reference, as they predict that future power will equal the value measured in the present, with the added advantage of not requiring parameter evaluation or external variables (De Giorgi et al. (2011)). Greater complexity can be found in physical methods, which rely on physical characteristics and mathematical models; these are, however, computationally demanding (Zhang et al. (2019)).

With regard to statistical methods, models range from time series (Zhao et al. (2016)) to various types of neural networks, such as Adaptive Wavelet Neural Networks (Jyothi and Rao (2016)), Elman Neural Networks (Xu and Mao (2016)) and Back Propagation Neural Networks (Chang (2013)). Lastly, hybrid methods combining neural networks and logic models have been used to exploit the advantages of each technique and improve overall accuracy (Hong and Rofflorido (2019)). All of these models are usually assessed using metrics such as root mean squared error, mean absolute error and mean absolute percentage error, and rely on point forecasting to produce their results.

Concerning probabilistic forecasting, there are two main strategies: the first consists of directly modeling the distribution (by making hypotheses on its shape), while the second derives the approximate distribution from point forecasts and their associated errors (Serafin and Uniejewski (2024)); examples of these approaches are historical simulation (Maciejowska et al. (2024)) and quantile regression averaging (Nowotarski and Weron (2015)). Pinball loss and empirical coverage are usually employed for their evaluation.

2.2.3 Point versus probabilistic forecasting

In the literature, the debate between point and probabilistic prediction tends to lean towards the probabilistic approach. While the former aims to predict a single value, namely the most likely one according to the chosen metric, the latter aims to predict the expected distribution of that value, defining the interval within which the true value of the response variable is expected to fall with a given probability (Amat Rodrigo and Escobar Ortiz (2024)). Probabilistic forecasting can be seen, for specific metrics such as the mean absolute error, as a generalization of point prediction: in quantile regression, median regression is the point forecast associated with the 0.5 quantile, which is a special case of the aforementioned

probabilistic prediction. It therefore provides additional information about the confidence of the model and the uncertainty of the prediction.

Yet evaluating its performance is still something that needs to be more fully defined, and it is considerably less straightforward. On the other hand, the metrics used to evaluate point prediction, despite providing a comparable and interpretable result, yield different outcomes when optimized, leading one to wonder which metric is the right one.

Lastly, quantiles, which are often used to represent the probabilistic forecast instead of the CDF (Tibshirani (2023)), enjoy the property of equivariance to monotone transformations (Koenker (1998)), meaning that the order in which a strictly increasing transformation is applied and the quantiles of the data are computed does not matter; however, this does not hold for expected values.

With respect to our specific problem, studies have shown that probabilistic wind power forecasting is much more informative than its conventional counterpart and can be useful in cases involving volatile wind power (Zhang et al. (2014)). In particular, the quantiles of the conditional predictive distribution are considered to be related to the optimal decision for a large family of decision problems, as noted by Gneiting (2011). This provides a solid foundation for the application of probabilistic wind power forecasting.

Additionally, probabilistic forecasting is valuable for its ability to predict the probability of extreme events and, when issued on consecutive days, tends to be more consistent (Buizza (2008)). The disadvantages of deterministic solutions in our case, instead, relate to the statistical distribution of the power as its variance is usually heteroscedastic. The power distribution is also asymmetric, with a tail skewed to the right for low power and one skewed to the left for high power (Capelletti et al. (2024)).

2.2.4 Recalibration in wind power forecasting

Another point of divergence between techniques concerns whether to use a direct or an indirect method.

In a direct method, the original features of the dataset, such as wind speed forecasts, wind direction and other weather conditions, are fed directly into the model. By doing so, most of the work is left up to the algorithm, including handling the non-linearity of the power curve. Given the intricacy of modeling such information, statistical methods can prove inadequate, and as a result complex machine and deep learning models are needed (Rajaperumal and Christopher

Columbus (2025)). Moreover, these approaches tend to be resource intensive and less interpretable.

The indirect approach, conversely, consists of two main phases: first the original features are recalibrated, and second, after the power curve has been estimated, the power of the wind farm can be predicted by feeding it the recalibrated features as input. This approach can be more beneficial for wind power forecasting, as noted by Capelletti and De Nicolao (2025). Since features such as local terrain attributes, atmospheric stability and turbine wake effects interact with the validity of the wind speed forecast, recalibrating it can minimize the difference between the forecast and the actual value.

In particular, some practical benefits of recalibrating the wind speed forecast are the enhancement of operational decisions and increased general applicability. The former concerns the advantages of having a reliable forecast to reduce the uncertainty of tasks such as planning maintenance windows, ensuring grid stability and handling power purchase agreements. The latter concerns how recalibration can address issues such as partial sensor coverage or sporadic data quality.

Chapter 3

Data

3.1 Chapter overview

In this chapter we present the preliminary data analysis that precedes the recalibration phase of our model. The primary dataset is described along with its possible issues, which lead to a correlation study and the reduction of the original dataset, along with its own correlation matrix. Lastly, we draw our conclusions and consider the following steps.

3.2 Wind farm data

For this section, the main dataset we used followed the concept of a virtual turbine, that is, a representation of the entire wind farm and all of its turbines as a single comprehensive one, obtained by computing the circular median of the features. The focus of this part is not particularly on direction, but rather on the wind speed forecasts. Moreover, the power will be considered later, in accordance with the indirect approach to wind power forecasting, as this part is propaedeutic to the first phase of our model.

3.2.1 Wind speed dataset

In particular, the resulting dataset [Figure 3.1] comprises the following columns:

- the **timestamp**, that is the date and time in UTC format
- the **circular median of the measured wind speed** in m/s (the target variable of the recalibration phase)

- the **forecasts of the wind speed** in m/s in the five grid points (P1-P5) of the farm at the heights of 10, 50, 80 and 100 metres (the predictors for the recalibration phase).

Timestamp	wsa_med_clean	wsf_10M_P1	wsf_50M_P1	wsf_80M_P1	wsf_100M_P1	wsf_10M_P2	wsf_50M_P2	wsf_80M_P2	wsf_100M_P2	...
2024-01-02 18:00:00	5.019172	1.873178	3.504986	3.599350	3.657619	2.108374	4.036724	4.126384	4.177236	...
2024-01-02 18:15:00	4.955216	1.903687	3.566041	3.658929	3.716431	2.126730	4.068577	4.152473	4.200064	...
2024-01-02 18:30:00	4.919454	1.960893	3.680519	3.770638	3.826704	2.161148	4.128299	4.201391	4.242866	...
2024-01-02 18:45:00	4.942509	2.018098	3.794996	3.882347	3.936977	2.195565	4.188022	4.250308	4.285668	...
2024-01-02 19:00:00	4.695423	2.075303	3.909473	3.994057	4.047250	2.229983	4.247745	4.299225	4.328469	...

Figure 3.1: Wind speed dataset

Both the chosen regressors and the actual values suggest redundancy in the given information and a strong possibility of collinearity.

Accordingly, we conducted an exploratory analysis of the data.

3.2.2 First correlation matrix

Our first step in the exploratory study was to compute and plot the correlation coefficients between the variables [Figure 3.2]. One can easily notice the strong correlation among them, which supports the possibility of collinearity.

A closer examination reveals a checkerboard pattern: in particular, wind speed forecasts at 50, 80 and 100 metres have higher correlation coefficients with all the others compared to the forecasts at 10 metres. This could suggest that the wind speed forecasts at 10 metres hold additional information with respect to the others, and vice versa, leading one to think that weeding out some of the features would be beneficial to the problem.

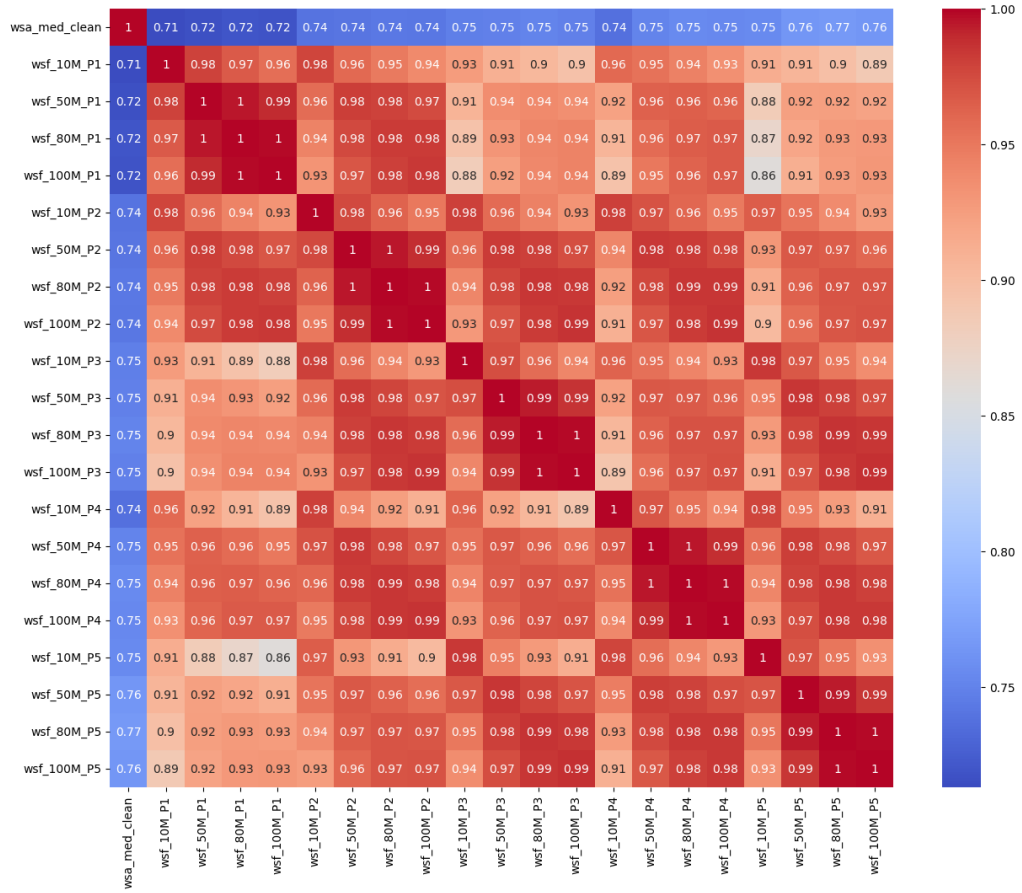


Figure 3.2: Correlation matrix of wind speed dataset

3.3 Data cleaning

3.3.1 Reduced wind speed dataset

Considering the previous findings, we chose to build a reduced dataset [Figure 3.3], comprising the timestamp, the median of the measured wind speed, and the forecasts of the wind speed in the five grid points of the farm at the heights of 10 and 80 metres from the original dataset. The first height was chosen due to its weaker correlation with the others, and the second because it was considered the most representative of 50, 80 and 100 metres. The resulting dataset is therefore more interpretable while still retaining the information needed for our problem.

Timestamp	wsa_med_clean	wsf_10M_P1	wsf_80M_P1	wsf_10M_P2	wsf_80M_P2	wsf_10M_P3	wsf_80M_P3	wsf_10M_P4	wsf_80M_P4	wsf_10M_P5	wsf_80M_P5
2024-01-02 18:00:00	5.019172	1.873178	3.599350	2.108374	4.126384	2.509670	4.687836	2.111050	4.034764	2.474903	4.651678
2024-01-02 18:15:00	4.955216	1.903687	3.658929	2.126730	4.152473	2.516352	4.682635	2.113138	4.023240	2.466867	4.621050
2024-01-02 18:30:00	4.919454	1.960893	3.770638	2.161148	4.201391	2.528882	4.672881	2.117053	4.001634	2.451798	4.563621
2024-01-02 18:45:00	4.942509	2.018098	3.882347	2.195565	4.250308	2.541411	4.663128	2.120968	3.980028	2.436730	4.506193
2024-01-02 19:00:00	4.695423	2.075303	3.994057	2.229983	4.299225	2.553941	4.653375	2.124884	3.958421	2.421662	4.448764

Figure 3.3: Reduced wind speed dataset

3.3.2 Second correlation matrix

A second correlation matrix was also plotted for this dataset. As figure 3.4 shows, we can still observe high correlation coefficients. This is most noticeable along the diagonals (not considering the main one, for obvious reasons), that is, among the forecasts of wind speed at the same height in different grid points. Regarding the target variable, the checkerboard pattern is absent, confirming the relevance not only of the wind speed forecasts at 80 metres, but also of those at 10 metres which, although less correlated with the others, remain significant to our analysis.

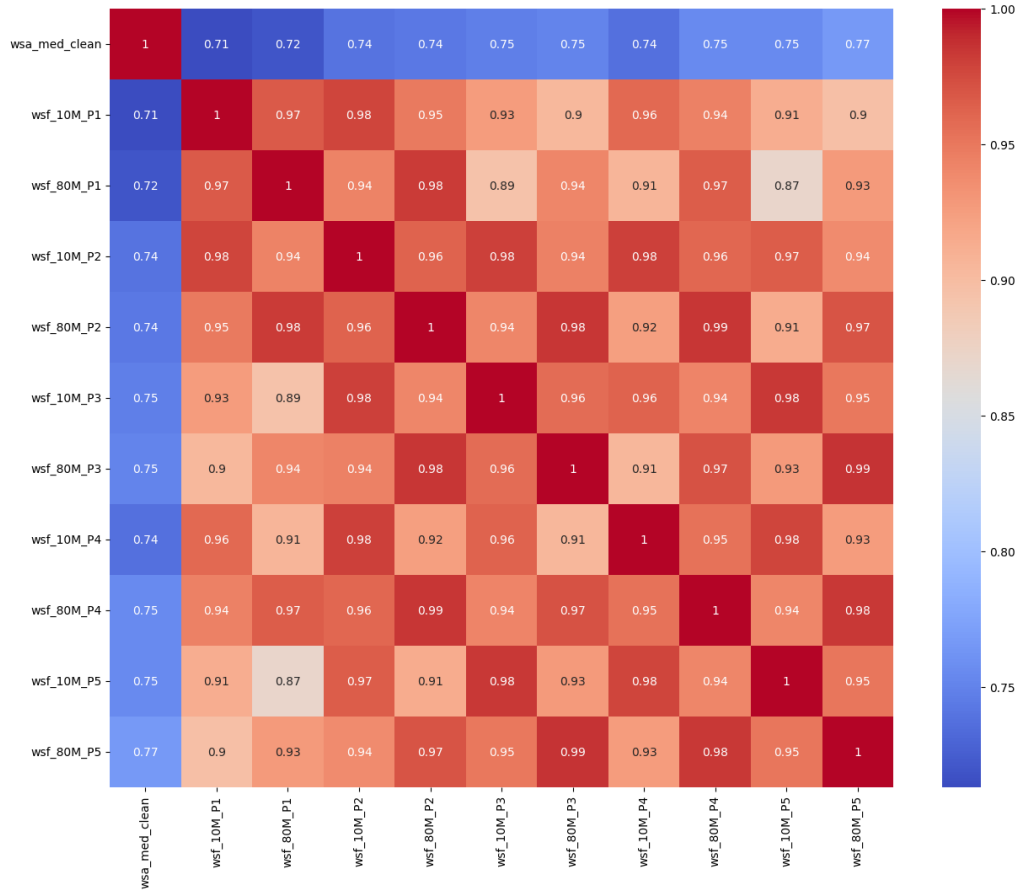


Figure 3.4: Correlation matrix of reduced wind speed dataset

3.4 Wind forecast

Given this preliminary study, which also included plotting various scatter matrices that can be viewed in the Appendix, we found that our hypothesis of possible collinearity between the regressors was validated by the correlation matrices. We also ended up with two datasets for the wind speed forecast, the original and the reduced. The latter could prove helpful for the subsequent models, as it already contained a subset of the predictors.

For the first phase of our approach, that is, the recalibration of the features to obtain a single wind speed later used for power forecasting, this analysis had to be taken into account in order to employ the models best suited to our dataset.

Chapter 4

Wind farm power curve

4.1 Chapter overview

In this chapter we describe the dataset used to plot the power curve from the measured wind speed and depict the scatter plot of the raw data. After performing data cleaning, we present the resulting scatterplot and the model built for power forecasting.

4.2 Introduction and dataset

To predict power from the recalibrated wind speed in the second phase of our approach, we needed a model for building a virtual power curve encompassing the output of all the turbines in the wind farm.

Our dataset [Figure 4.1] was the result of joining, on the timestamp, the measured wind speed with a preprocessed dataset from the TSO, from which the power of the wind farm was selected and its timestamp converted to UTC format, so that the two datasets were aligned.

timestamp_utc	wsa_med_clean	Power
2024-01-02 09:45:00	5.451536	2.4640
2024-01-02 10:00:00	5.357114	2.0944
2024-01-02 10:15:00	5.187650	1.7424
2024-01-02 10:30:00	5.224357	2.0240

Figure 4.1: Measured wind speed and power dataset

We then built the scatterplot of measured wind speed versus power. As we can see from figure 4.2, the plot is not very clean, especially at higher wind speeds, where the power is lower than it could theoretically be due to the curtailment phenomenon (that is, when turbines are deliberately limited or switched off).

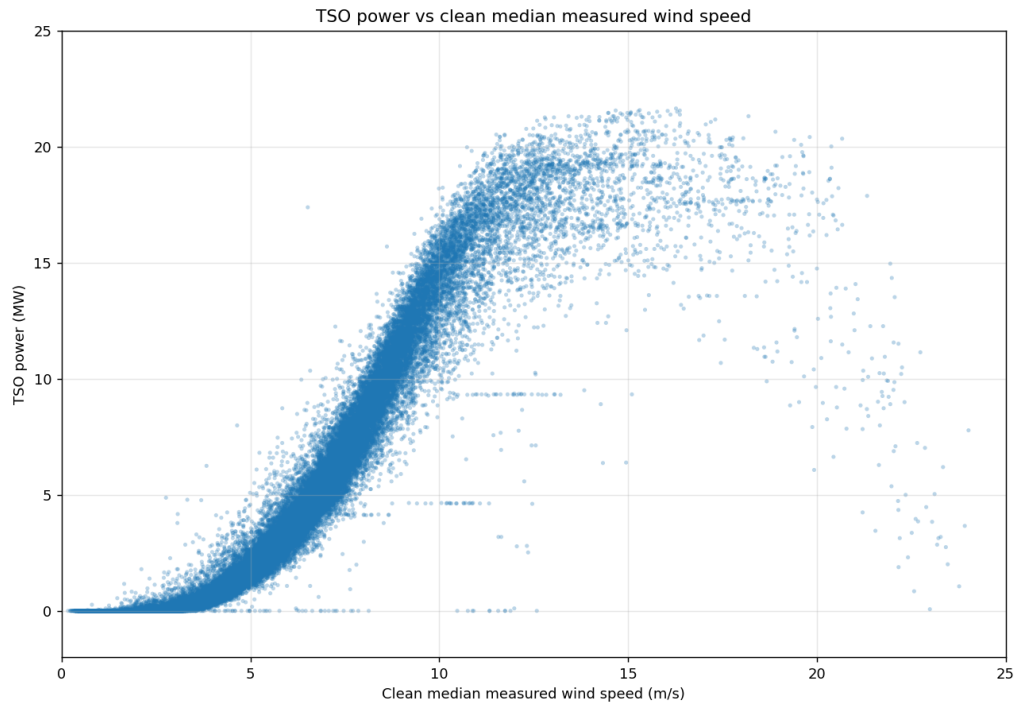


Figure 4.2: Scatterplot of raw data, measured wind speed vs power

4.3 Data cleaning

In order to obtain a cleaner result, and consequently a better power curve, we first divided the wind speed data into two hundred bins, and then applied an outlier boxplot to the corresponding power values in each bin. Figure 4.3 shows a more polished scatterplot than figure 4.2 and serves as the starting point for our model.

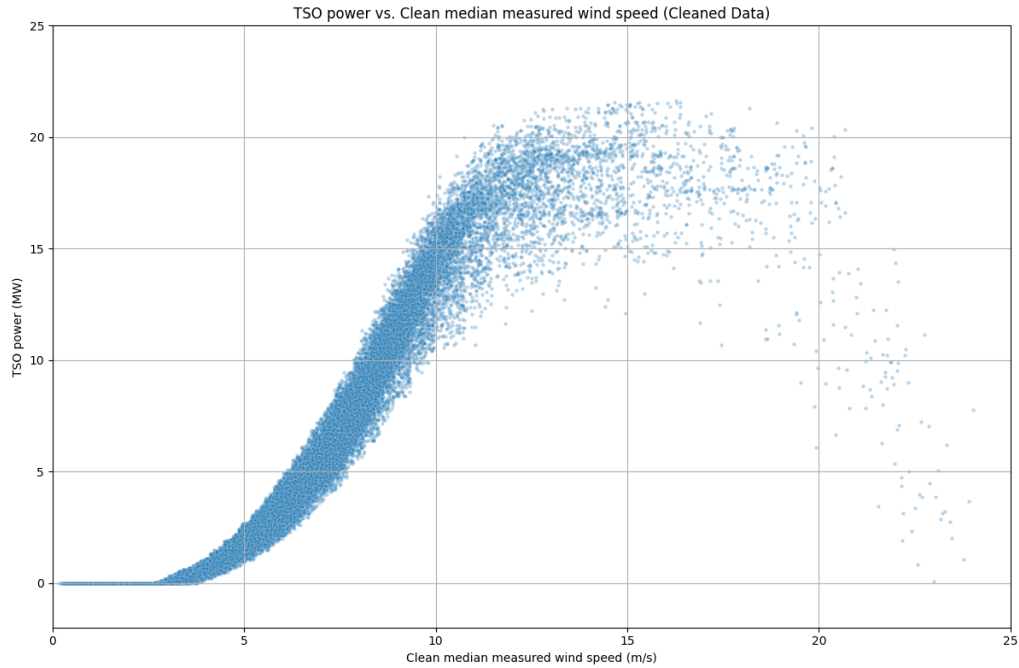


Figure 4.3: Scatterplot of cleaned data, measured wind speed vs power

This scatterplot depicts, even better than the previous one, one of the main characteristics of the distribution, namely its heteroscedasticity. As expected from a wind power distribution, its variance changes along the curve: it is very small for low wind speeds, roughly constant in the interval between 5 and 10 m/s, and finally quite large from 12 m/s onward due to curtailment. We therefore expect the following models to be less stable and to struggle to fit the distribution at higher wind speeds, but we still take into consideration this part of the data, up to 18 m/s, for a more realistic approach.

4.4 Power curve

The model we chose to build the power curve was a linear generalized additive model (GAM) with splines. It was constrained to be monotonically increasing, in accordance with the physical power distribution, while also having the advantage of being non-parametric. We used the pyGAM package (pyGAM Developers (2025)) for its modularity and interpretable output.

In particular, we employed splines because they can fit non-linear models without incurring the overfitting typical of high-degree polynomial regression, as they are piecewise low-degree polynomials with continuous derivatives at each of their knots (James et al. (2023)).

It is important to note that with this model we are making the strong assumption

that the power curve is deterministic, which is not the case in reality, but remains a useful approximation.

In order to build our model, we used data from the year 2024 and then plotted the curve on data from the year 2025, achieving a more realistic consistency with its possible real-world application. We also built a model using the 2025 data in order to compare the two and check whether a non-negligible difference existed.

Figure 4.4 shows the model fitted on the 2024 data plotted on the 2024 distribution with its confidence interval, whereas figure 4.5 depicts the previous model plotted on the 2025 data, along with a new model fitted and plotted on 2025 data. It is also worth noting that confidence intervals computed with the pyGAM package rely on large-sample results to handle non-Gaussian distributions, and treat the smoothing parameters as fixed, when in reality they are estimated from the data.

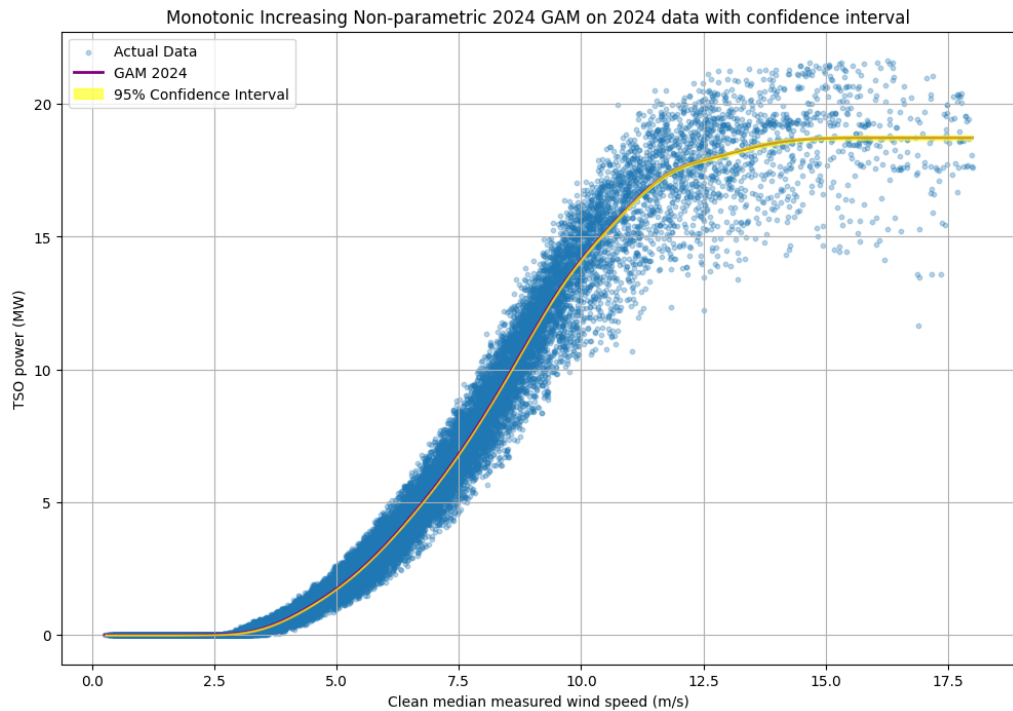


Figure 4.4: Linear GAM fitted and plotted on 2024 data with its confidence interval

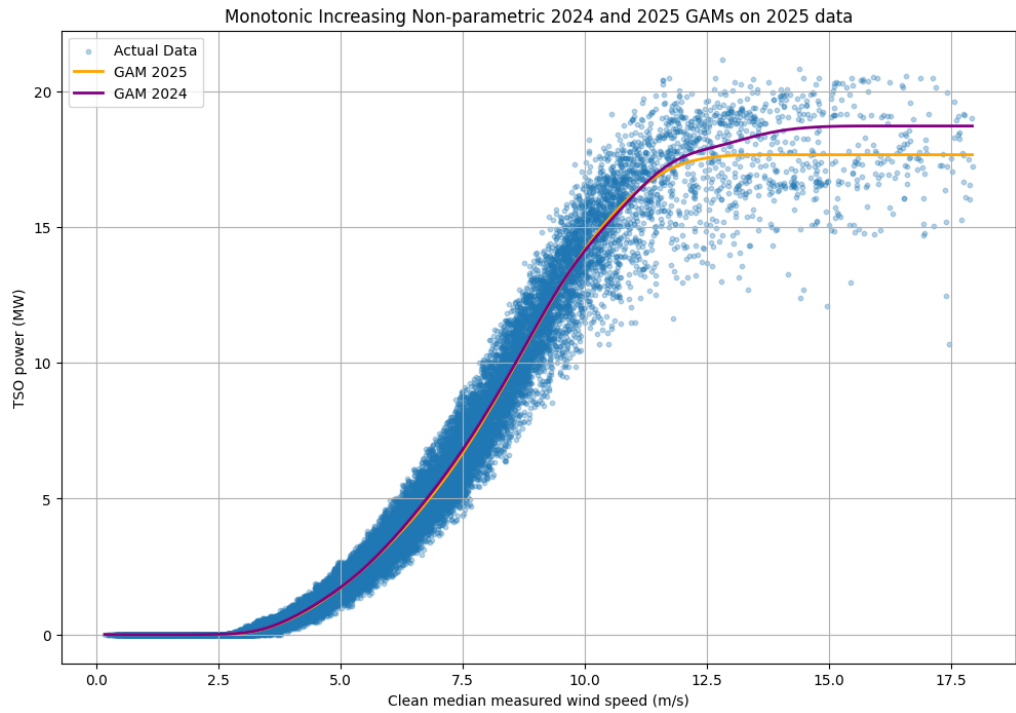


Figure 4.5: Linear GAM fitted on 2024 data and plotted on 2025 data (purple) and linear GAM fitted and plotted on 2025 (orange)

As we can see, the two GAMs, despite being fitted on different years, virtually overlap for wind speeds up to about 11 m/s, and only slightly diverge from that point on. This was to be expected given the heteroscedastic variance mentioned previously, since that is the part of the plot which is hardest to model due to the curtailment phenomenon.

Considering the 2024 GAM to be an adequate model, we then used it to predict the power from the recalibrated wind speed.

Chapter 5

Probabilistic recalibration of wind forecasts

5.1 Chapter overview

In this chapter we describe the recalibration phase built for the first part of our approach. In particular, we highlight the models that were fitted and their resulting metrics, and draw our conclusions on the model chosen for our actual recalibration, comparing it with the marginal model.

5.2 Introduction and employed models

The choice of the indirect method in wind power forecasting requires a first phase to recalibrate the wind speed. In particular, we want to use the wind speed forecasts in the five grid points to predict our target variable, the actual wind speed, thereby achieving the recalibration.

Considering this, as well as the results of the preliminary data analysis and, in particular, the collinearity we identified, we needed models that could both recalibrate our features and counteract the collinearity, either by regularizing the problem or by reducing its dimensionality. As such, we chose to implement:

- a **linear regression** acting as a baseline for the subsequent models
- a **principal component regression** derived from a principal component analysis
- a **lasso median regression**

Given that these models yield point predictions, we implemented them in order to find the best possible statistical representation for recalibrating the wind speed,

in terms of both the metrics used and the dimensionality of the solution. We carried out the following steps on both the original and reduced datasets. As expected, both datasets yield essentially equivalent outcomes, so we report only the results of the latter, for this reason and for better interpretability of the models.

5.3 Preprocessing

Our dataset was preprocessed as follows: first we randomly split it into train, validation and test sets, comprising respectively 60%, 20% and 20% of the dataset, to compute the metrics used for model evaluation. We then applied standard scaling for better overall accuracy.

5.4 Models

5.4.1 Linear regression

As our first model, we built a linear regression, intended to act as a baseline for comparison with the other models.

We obtained a linear combination of the ten features plus the intercept:

$$f_L(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon,$$

by minimizing the

$$RSS = (f_L(X) - \beta_0 - \sum_{j=1}^k \beta_j X_j)^2$$

(James et al. (2023)). We can expect coefficients with a p-value < 0.05 to be statistically significant.

Given the strong correlation between the regressors, we expected high collinearity and low interpretability, even with the reduced dataset. For the actual implementation, we used the Statsmodels package (Statsmodels Developers (2024)) for its comprehensive and interpretable output.

The resulting model comprised a weighted linear combination of the predictors plus the intercept, whose coefficients and p-values can be found in table 5.1. It is noteworthy that most of them are statistically significant. Looking at the coefficients, we see both positive and negative values, which fits with our hypothesis

of multicollinearity and low interpretability (after all, wind speed cannot be negative). Furthermore, the coefficients with the highest absolute value, besides the intercept, belong to the wind speed forecast at 80 metres in the P1 grid point, the forecast at 80 metres in P4, and lastly the forecast at 80 metres in P5.

These results are in line with the need to reduce the dimensionality of the covariate space and to regularize the problem.

Table 5.1: Linear regression model coefficients

Model coefficients		
Feature	Value	p-value
intercept	5.4861	0.000
wsf_10M_P1	-1.3386	0.000
wsf_80M_P1	2.8945	0.000
wsf_10M_P2	0.2118	0.539
wsf_80M_P2	-1.2845	0.003
wsf_10M_P3	0.5560	0.056
wsf_80M_P3	-1.0427	0.003
wsf_10M_P4	1.3829	0.000
wsf_80M_P4	-2.8729	0.000
wsf_10M_P5	-0.0773	0.759
wsf_80M_P5	3.8754	0.000

With regard to the metrics we computed, we obtained a MAE of 1.359, a MAPE of 0.374 and an RMSE of 1.811.

5.4.2 PCA and PCR

To reduce the dimensionality of the problem, we performed a Principal Component Analysis (or PCA) on our dataset using the Sklearn package (scikit-learn Developers (2026a)).

This method applies a linear transformation into a new set of coordinates, obtaining the principal components. Among them, the first one can be found as:

$$Z_1 = \phi_{11}X_1 + \dots + \phi_{p1}X_p,$$

It holds that Z_1 has the largest variance and $\sum_{j=1}^p \phi_{j1}^2 = 1$, where the ϕ coefficients are the loadings and p is the number of features. The second principal

component is, in turn, the linear combination of the features with the largest variance among all the linear combinations uncorrelated with the first one, and so on (James et al. (2023)).

As we can see from figure 5.1, which shows the cumulative explained variance with respect to the number of principal components, two or three components at most are more than enough to represent our problem. This is a valuable result, as it leads to a remarkable reduction in the size of the regressor space.

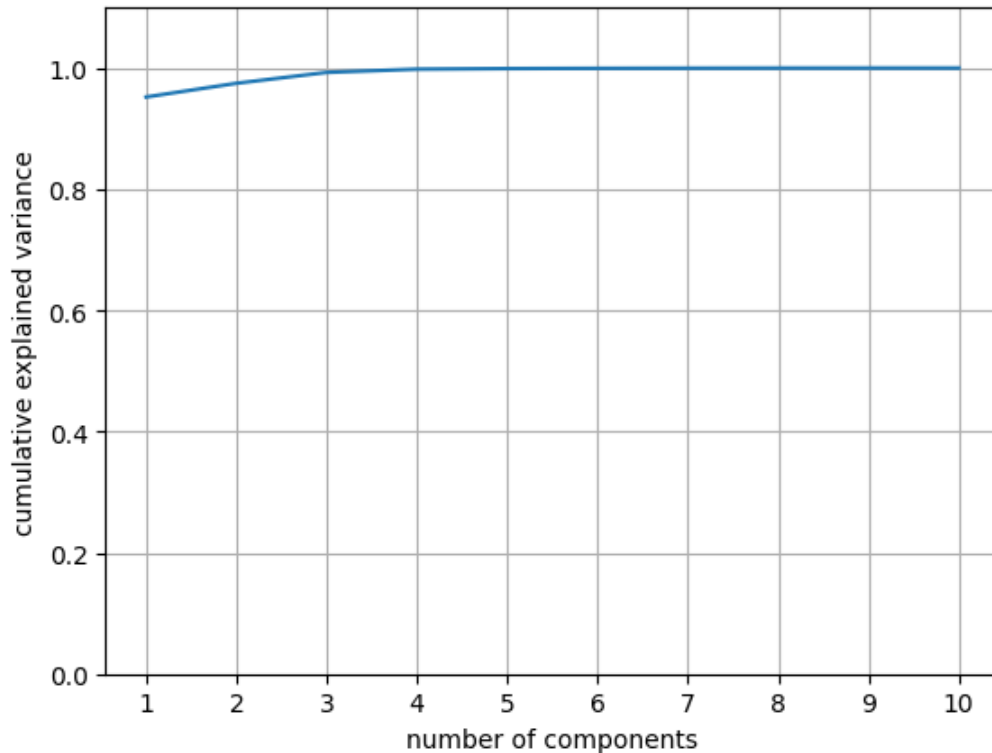


Figure 5.1: Cumulative explained variance of the PCA per number of principal components

Figure 5.2 shows a scoreplot, where the data have been projected onto the first two principal components. Level lines were added, due to the high concentration of data, to improve the readability of the graph.

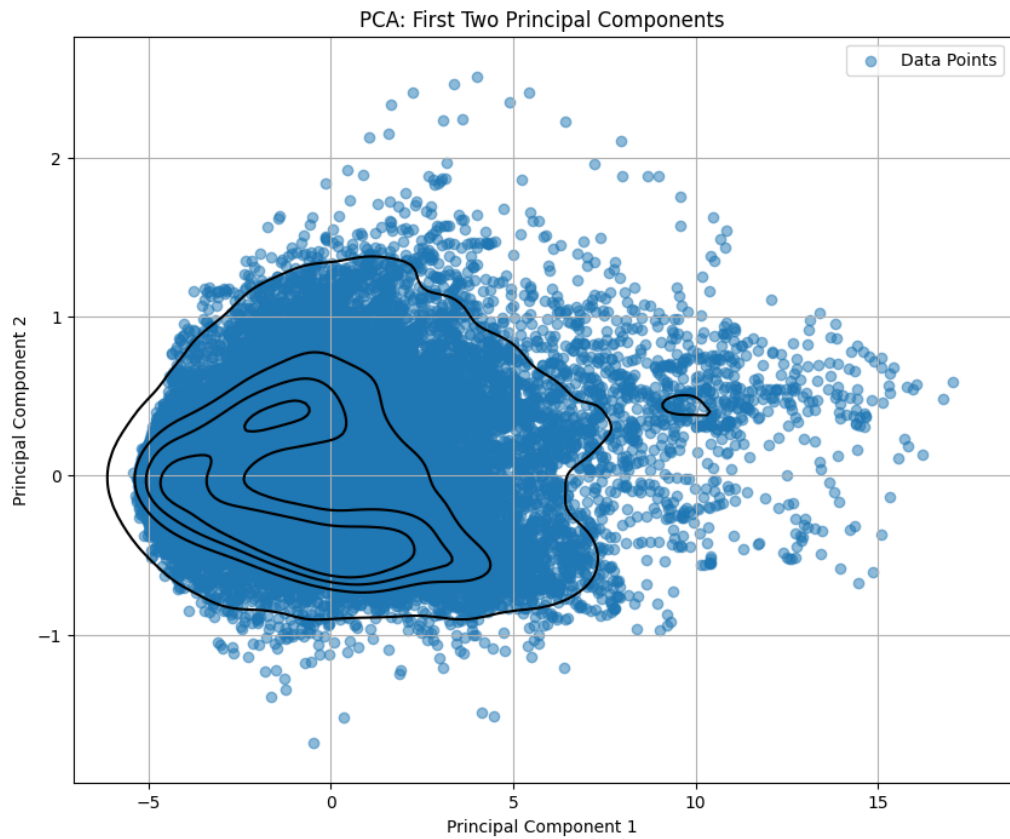


Figure 5.2: Scoreplot of data projected on the first two principal components

Another relevant graph which shed further light on our problem was the bi-plot [Figure 5.3], which combines a scoreplot [Figure 5.2] with a loadings plot, that is, a graph showing the magnitude and direction of the loadings of a PCA, specifying the relationship between the original features and the first two principal components.

In particular, it can be seen that the first component behaves in accordance with the wind speed forecasts, whereas the second is related to the differences between the forecasts, which take widely varying values along this second component.

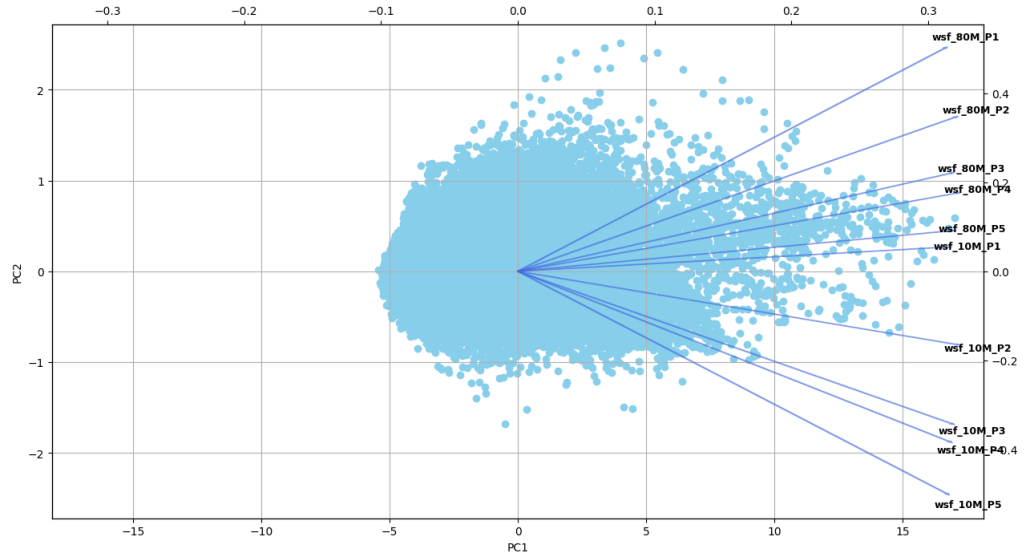


Figure 5.3: Biplot of data projected on the first two principal components

With the last set of graphs, the barplots [Figure 5.4], we delved deeper into the behavior of the regressors with respect to the first two components; on the x-axis we find the predictors, and on the y-axis the values of the corresponding loadings. The first component appears to be a weighted average of the regressors with virtually the same value, whereas the second appears to be a correction factor to which all the predictors contribute linearly, but with widely varying weights. This corroborates the behavior previously observed in the biplot.

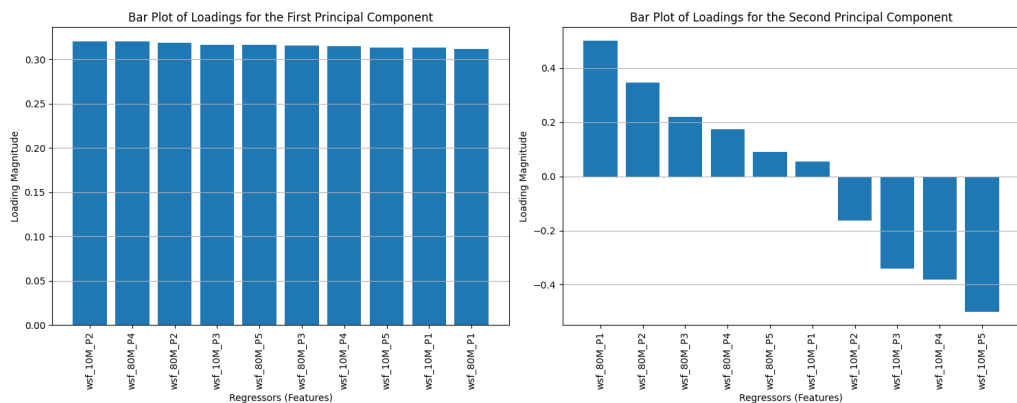


Figure 5.4: Barplots of the loadings of first and second principal components

The natural continuation of the process was to build a Principal Component Regression (or PCR) on top of the PCA, using the first n principal components found beforehand as predictors, again with the Statsmodels package, as with the linear regression. In particular, we fitted eight PCRs and plotted their corre-

sponding mean absolute errors and root mean squared errors [Figure 5.5].

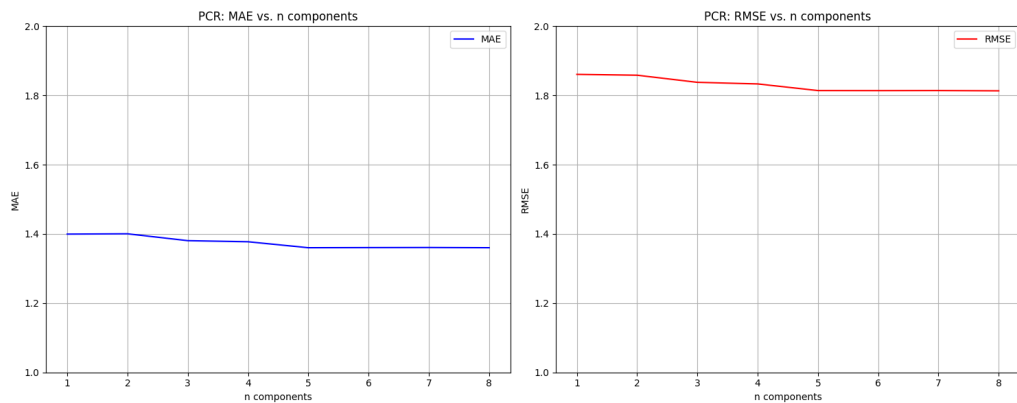


Figure 5.5: mean absolute errors and root mean squared errors of the eight PCRs

Considering these results and the cumulative explained variance, we focused on the PCRs with two and three components, and report their corresponding coefficients and p-values in table 5.2 and table 5.3. All the weights can be seen to be statistically significant. Moreover, they have, respectively, a MAE of 1.401 and 1.381, a MAPE of 0.388 and 0.381, and an RMSE of 1.858 and 1.838.

Table 5.2: Model coefficients of first PCR

Model coefficients of first PCR

Feature	Value	p-value
intercept	5.4861	0.000
PC1	0.7053	0.000
PC2	-0.1617	0.000

Table 5.3: Model coefficients of second PCR

Model coefficients of second PCR

Feature	Value	p-value
intercept	5.4861	0.000
PC1	0.7053	0.000
PC2	-0.1617	0.000
PC3	-0.6379	0.000

5.4.3 Lasso median regression

To regularize the variable weights and also tackle the multicollinearity problem, we built a lasso median regression using the Sklearn QuantileRegressor (scikit-learn Developers (2026b)).

This model was chosen because it performs feature selection by shrinking the regressors' weights, sometimes down to zero. In particular, it is obtained by computing:

$$\min_{\beta} \left[\sum_{i=1}^n \rho_{\tau}(y_i - x_i^T \beta) + \alpha \sum_{j=1}^p |\beta_j| \right],$$

where $\rho_{\tau}(u) = \begin{cases} \tau u & u \geq 0 \\ (\tau - 1)u & u < 0 \end{cases}$ is the pinball loss,

$\tau \in (0, 1)$ is the quantile (in our case 0.5) and $\alpha \sum_{j=1}^p |\beta_j|$ is the L_1 penalty weighted by the regularization parameter α . Lastly, n is the number of observations and p is the number of coefficients.

In order to tune our model, we varied α between 10^{-6} and 10^0 . Within this interval we plotted the MAEs and MSEs [Figure 5.7] and computed the values of the regressor coefficients [Figure 5.6].

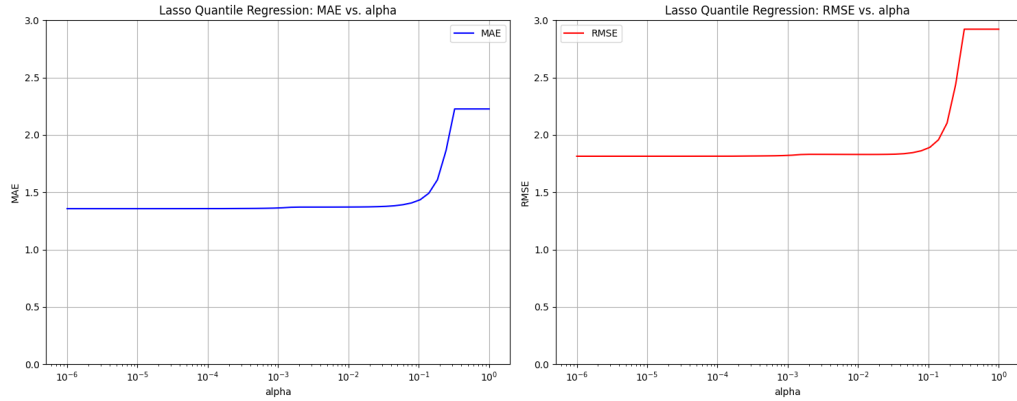


Figure 5.6: Mean absolute errors and root mean squared errors of the lasso quantile regression varying α

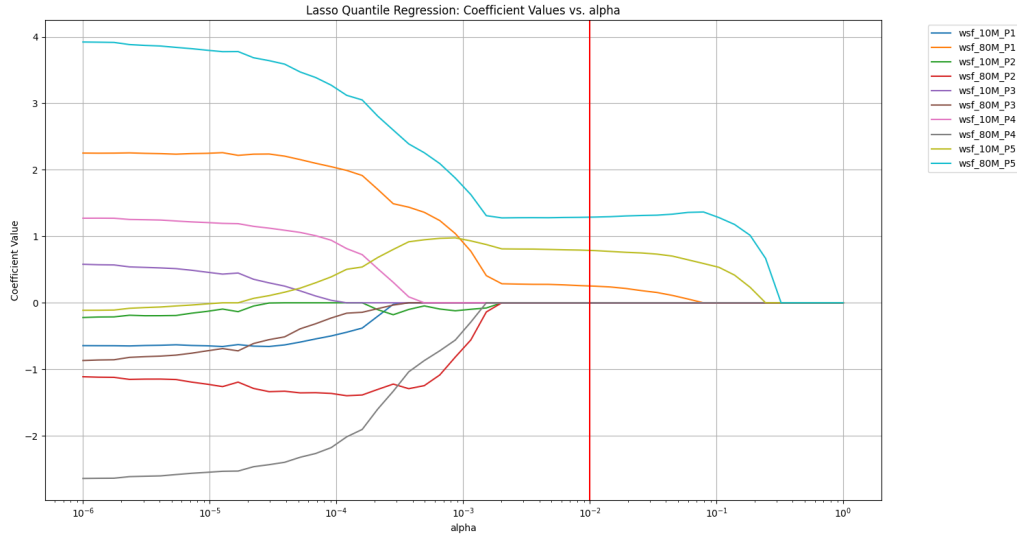


Figure 5.7: Lasso quantile regression coefficients varying alpha

As we can see from figure 5.7, a possible subset of the most significant regressors (that is, the ones that go to zero the latest) can be found between 10^{-3} and 10^{-1} . More specifically, in this range we find, from smallest to largest, the regressors for the wind speed forecasts at 80 metres in the P5 grid point, at 10 metres in P5, and at 80 metres in P1; with α equal to 10^{-2} , their unscaled coefficients, shown in table 5.4 for a more informative view of their magnitude, are as follows.

Table 5.4: Unscaled model coefficients of the lasso quantile regression

Model coefficients	
Feature	Value
wsf_80M_P5	1.28898269
wsf_10M_P5	0.7850324
wsf_80M_P1	0.24995624

Moreover, again with an α of 10^{-2} , the MAE, MAPE and RMSE are, respectively, 1.371, 0.369 and 1.829.

5.4.4 Other models

In addition to the three models just presented, we also built other solutions, still considering the need to regularize and to reduce the dimensionality of the problem. Although their ability to do so, along with their performance, was somewhat

lower than that of the previous models, they still held meaningful information. In particular, we employed forward stepwise selection and backward stepwise selection, where the selected features belonged for the most part to the grid points P1 and P5, which is in accordance with the lasso median regression. We also built a ridge regression and an elastic net, which were not considered further because they did not sufficiently reduce the dimensionality of the covariate space, as well as a lasso regression and a quantile regression which, while adequate, led us to prefer the lasso median regression as the better performing model for our problem.

5.5 Overall findings

After trialing several models, we report our results in terms of MAE, MAPE and RMSE in table 5.5.

Table 5.5: Test metrics comparison among the models

Metrics comparison			
Model	MAE	MAPE	RMSE
Linear regression	1.359	0.374	1.811
PCR with 2 PCs	1.401	0.388	1.858
PCR with 3 PCs	1.381	0.381	1.838
Lasso quantile regression	1.371	0.369	1.829

Even though the linear regression has the lowest errors, the model is not interpretable and suffers from collinearity, whereas the PCR and lasso median regression achieve comparable metrics to the former (the percentage difference between the MAE of the linear regression and that of the lasso median regression is only 0.88%) while also greatly reducing the dimensionality of the covariate space. The latter is the model we considered the best candidate, also for its interpretability. Therefore, a lasso median regression with an α coefficient of 10^{-2} is a viable model for our target variable.

We have now reduced the problem's complexity to three regressors (versus the original ten in the linear regression), which is in line with the results obtained by the PCA, suggesting that two or three principal components are enough to adequately represent our problem.

5.6 Marginal model

In order to gain a better understanding and comparison of our final model and the metrics used, we present the simple case of computing the quantiles of the distribution of our target variable [Figure 5.8], using the 2024 data for training and the 2025 data for testing.

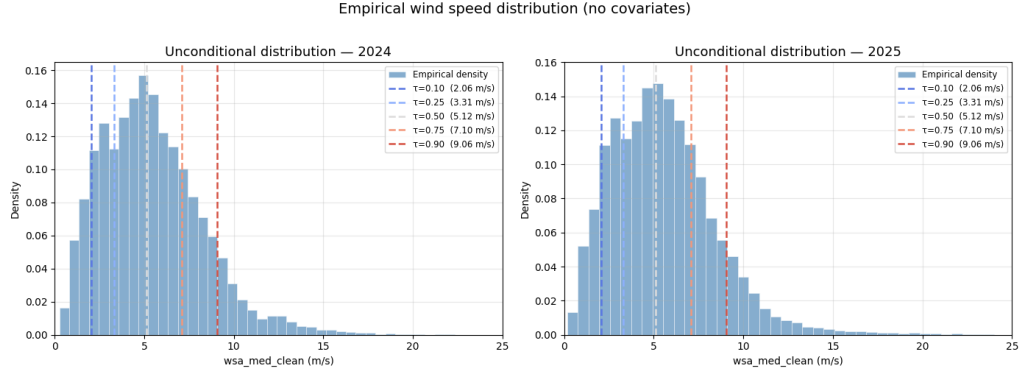


Figure 5.8: Quantile distributions of the measured wind speed (training: 2024, test: 2025)

In particular, to evaluate it we computed the pinball loss, the coverage error and the PIT.

5.6.1 Pinball loss

The optimization of the quantiles is obtained by minimizing the pinball loss, that is, the sum of absolute residuals asymmetrically weighted depending on the quantile (Koenker and Hallock (2001)). In particular, this is achieved by mini-

$$\text{mizing the sum of } \rho_{\tau}(u) = \begin{cases} \tau u & u \geq 0 \\ (\tau - 1)u & u < 0 \end{cases},$$

where u represents the absolute value of the residual, that is, the difference between the prediction and the true value, and $\tau \in (0, 1)$ is the quantile.

The typical shape of a pinball loss function across the quantiles is parabolic; its maximum is reached at the 50th quantile if the distribution is symmetric, whereas when it is not, the parabola is skewed (Narayan et al. (2024)).

Figure 5.9 shows the pinball loss of the marginal model, with its minimum at 0.22 for τ equal to 0.05 and its maximum at 1.22 for τ equal to 0.55 on the test data. As expected, the skewed curve reflects the asymmetry of our distribution.

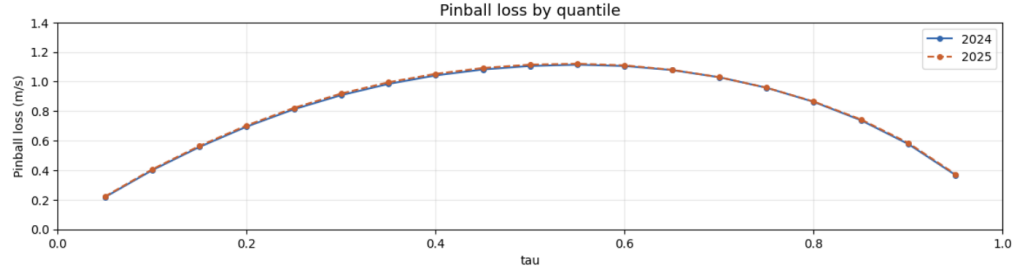


Figure 5.9: Pinball loss of the marginal model (training: 2024, test: 2025)

5.6.2 Empirical coverage and coverage error

The empirical coverage for a quantile is the percentage of predicted data that falls within it. It is to be compared with the nominal coverage, which is equal to the quantile. In the optimal case, the empirical coverage equals the nominal coverage and lies on the 45-degree line. Moreover, the coverage error is the difference between the empirical coverage and the quantile.

We plotted both in figure 5.10.

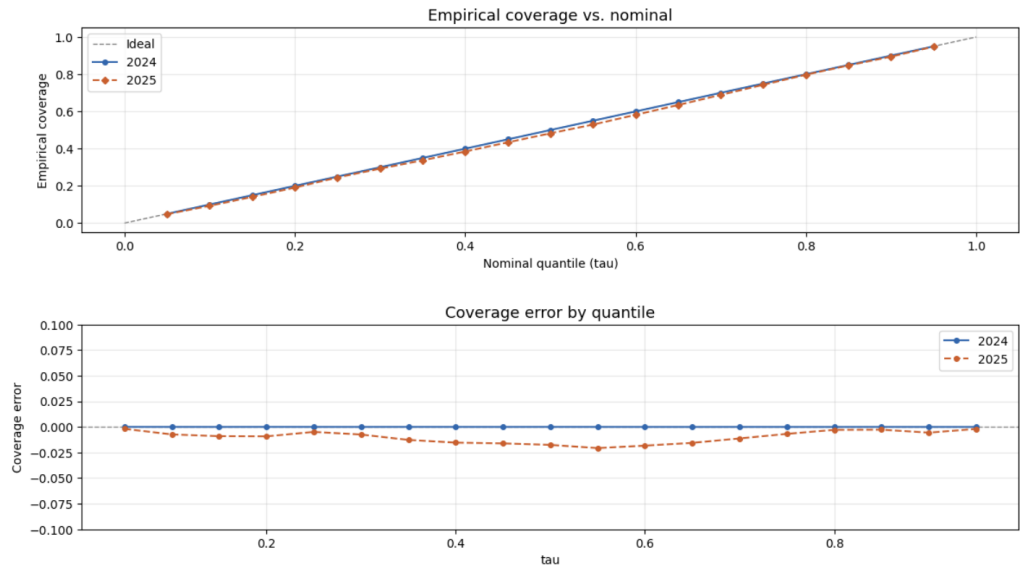


Figure 5.10: Empirical coverage and coverage error of the marginal model (training: 2024, test: 2025)

5.6.3 PIT

The Probability Integral Transform (PIT) associated with any cumulative density function F and target random variable Y is defined as

$$F^*(Y) = V \cdot F(Y) + (1 - V) \cdot F(Y^-),$$

where $V \sim \text{Unif}(0, 1)$ is independent of F and Y , and $F(y^-) = \lim_{x \rightarrow y^-} F(x)$. If F is almost surely continuous (that is, if F is a continuous function, and therefore invertible), then the PIT becomes $F^*(Y) = F(Y)$.

If F is fixed and equal to the CDF of the target variable Y , then $F^*(Y)$ (or $F(Y)$ in the continuous case) is itself a standard uniform random variable (Tibshirani (2023)). This means that we can evaluate our probabilistic forecast by how close its PIT is to the uniform distribution.

Figure 5.11 shows the results of the Probability Integral Transform for the marginal model.

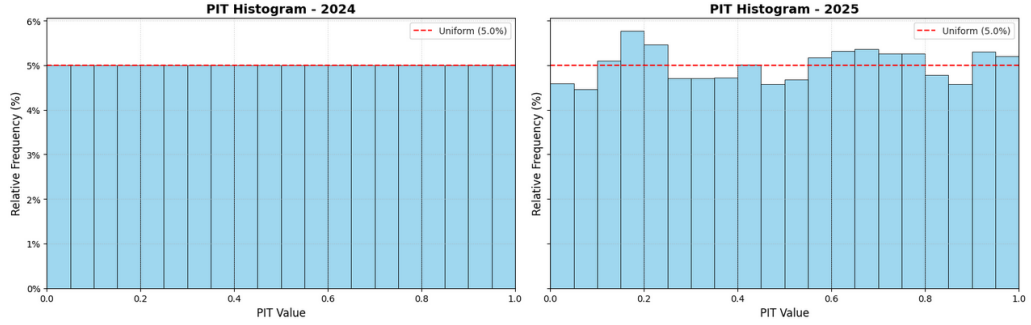


Figure 5.11: PIT of the marginal model (training: 2024, test: 2025)

5.7 Actual recalibration

This analysis was propaedeutic to the actual recalibration of the wind speed and led us to choose the best model. Having identified it, we switched from point to probabilistic forecasting by employing a lasso quantile regression with an α coefficient of 10^{-2} , considering all quantiles with a 0.05 step. In this case, as with the marginal model, we used the 2024 data as the training set and the 2025 data as the test set, as this is more consistent with how the model would be employed in practice.

Figure 5.12 shows the test goodness-of-fit plots of the model for each quantile, along with the corresponding MAE and RMSE.

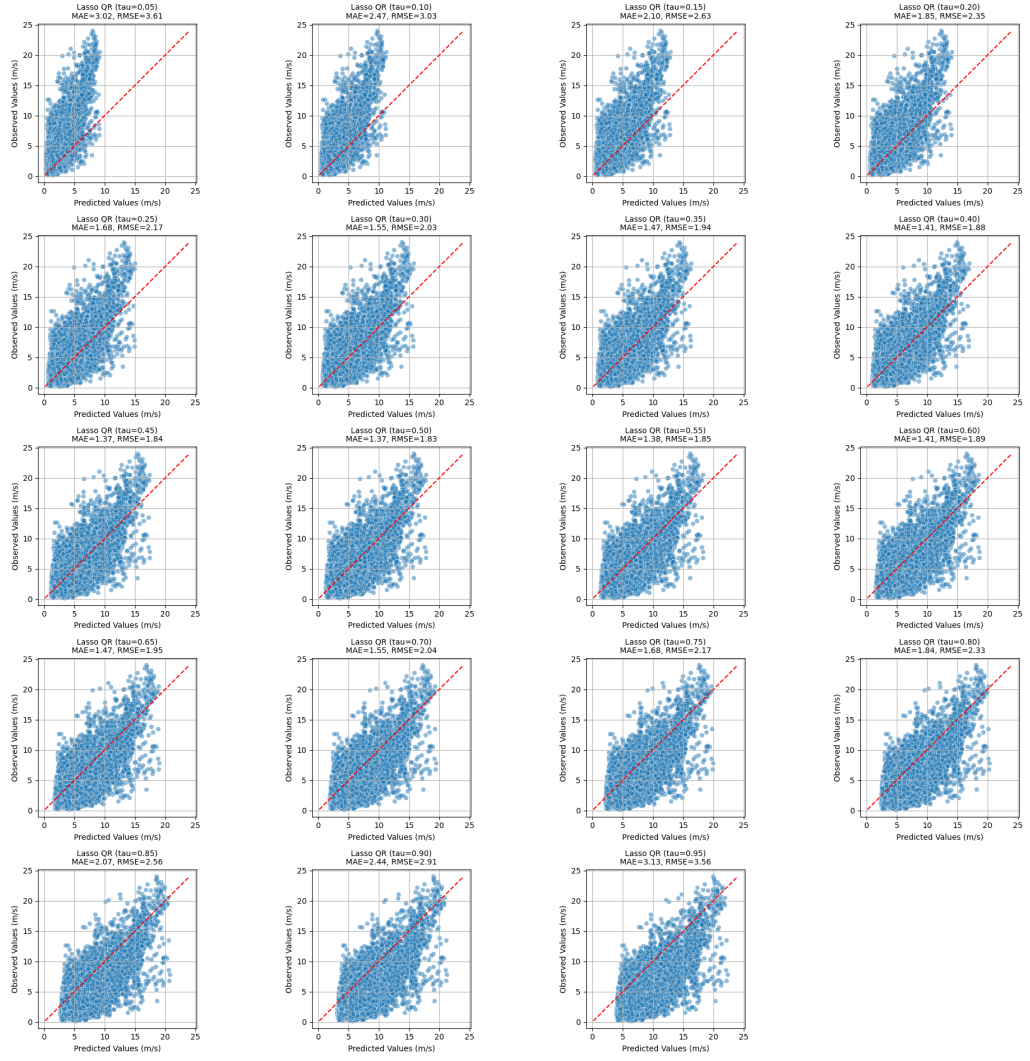


Figure 5.12: Goodness of fit plots of the lasso quantile regression

With respect to the previously mentioned metrics, figures 5.13 and 5.14 show that the empirical coverage and PIT obtained with the lasso quantile regression can be considered equivalent to those of the marginal model, as was expected.

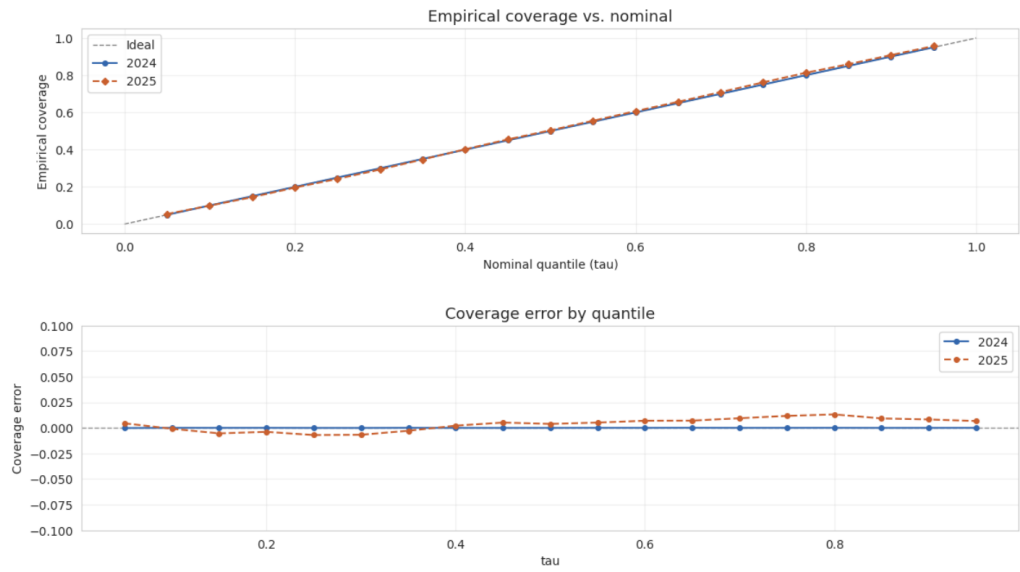


Figure 5.13: Empirical coverage and coverage error of the lasso quantile regression (training: 2024, test: 2025)

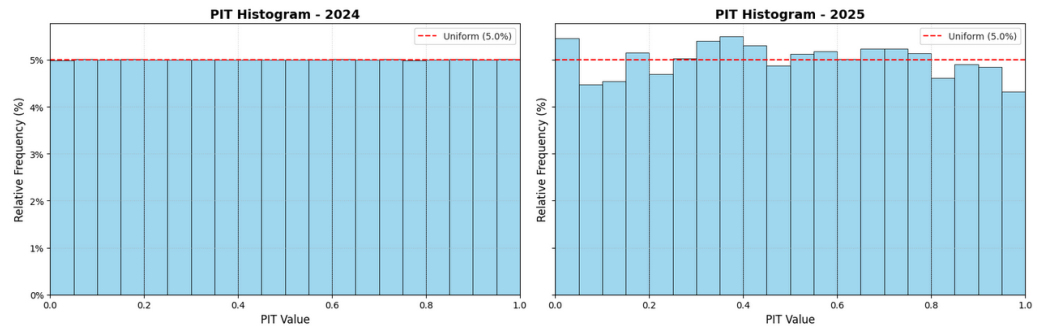


Figure 5.14: PIT of the lasso quantile regression (training: 2024, test: 2025)

On the other hand, a meaningful difference can be found in the pinball loss of this model [Figure 5.15].

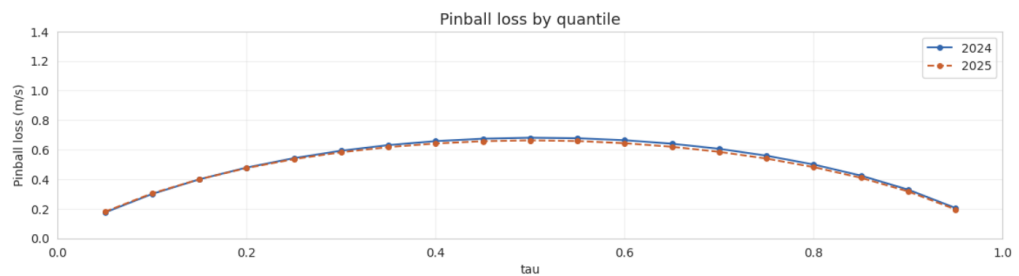


Figure 5.15: Pinball loss of the lasso quantile regression (training: 2024, test: 2025)

As we can see by comparing them [Figure 5.16], the pinball loss of the lasso quantile regression (right) is much lower than that of the marginal model (left), as the former has its minimum at 0.18 for τ equal to 0.05 and its maximum at 0.66 for τ equal to 0.5 on the test data.

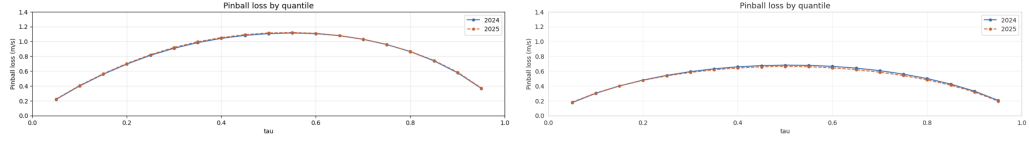


Figure 5.16: Comparison between the pinball loss of the marginal model (left) and the pinball loss of the lasso quantile regression (right)

We can conclude that decent PIT and empirical coverage results for a model are a necessary but not sufficient condition for it to be a good model, and that the lasso quantile regression outperforms the marginal model. Therefore, the output of the former, that is, the recalibrated wind speed, was used in the second phase of the indirect model as the input for predicting the power curve.

Chapter 6

Probabilistic power forecast

6.1 Chapter overview

In this chapter we describe the actual power forecasting, made possible by the recalibration of our target variable and the construction of the power model. We discuss our results using the metrics also computed for the first phase of our approach, and we make a comparison with a direct method.

6.2 Predicted power

Having obtained everything necessary for our second phase, that is, the recalibrated wind speed via the lasso quantile regression and the power curve via the linear GAM, we could finally predict the power. In particular, we fed the predicted wind speed for each quantile (again with a 0.05 step) as input to the power curve and obtained the corresponding power for each quantile [Figure 6.1].

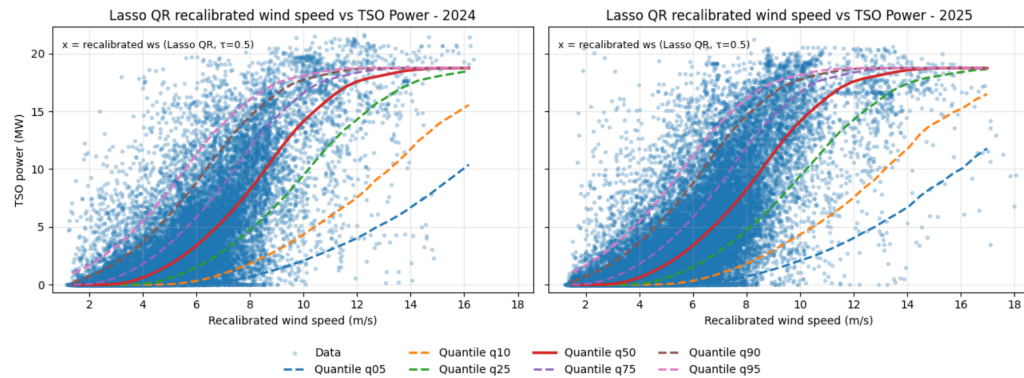


Figure 6.1: Scatterplot of lasso quantile regression versus TSO power with linear GAM

We plotted the test goodness of fit for each quantile in figure 6.2, along with

the corresponding MAE and RMSE.

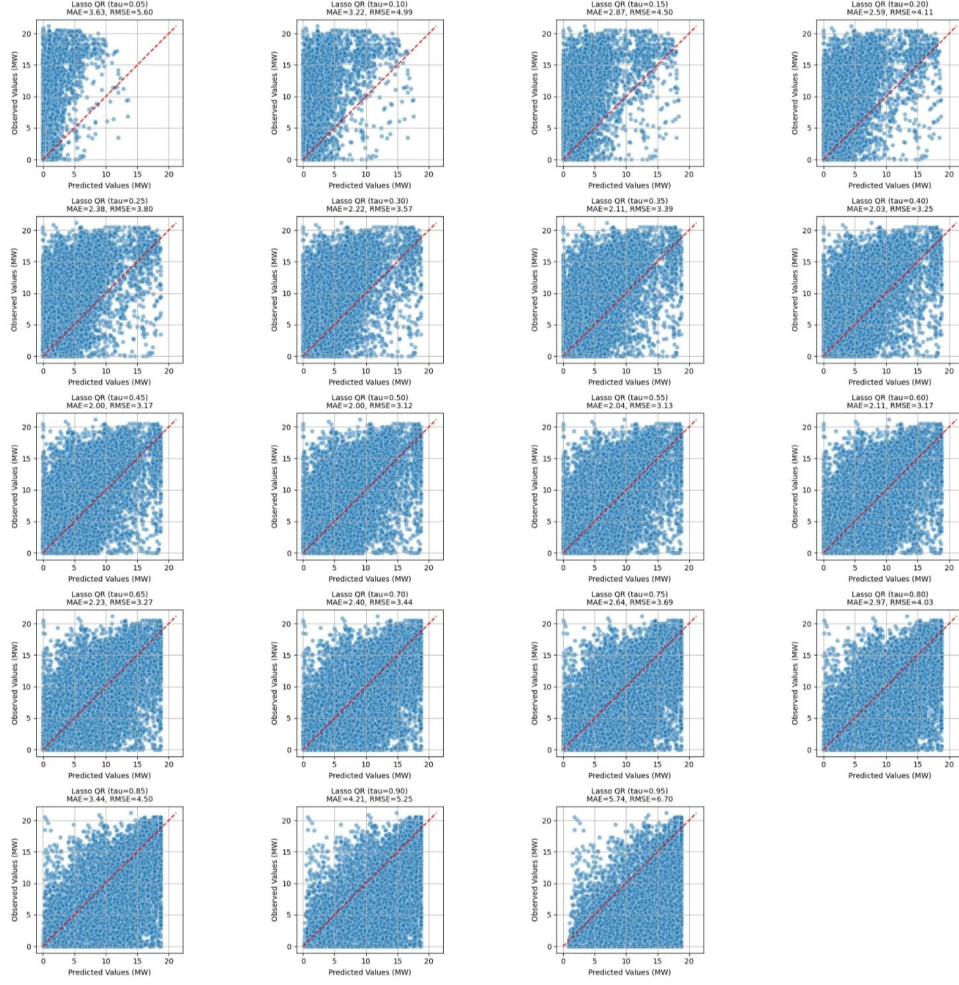


Figure 6.2: Goodness of fit plots of the predicted power

We then computed the pinball loss [Figure 6.3], the empirical coverage and coverage error [Figure 6.4], and the PIT [Figure 6.5].

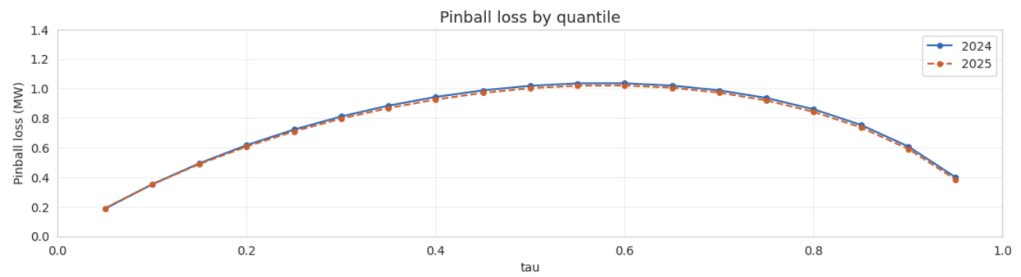


Figure 6.3: Pinball loss of the predicted power (training: 2024, test: 2025)

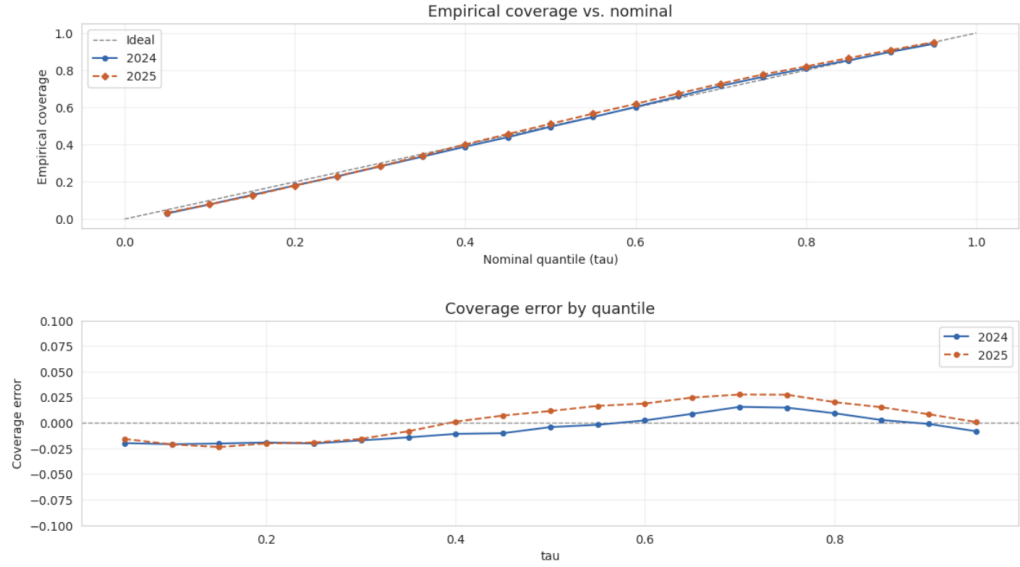


Figure 6.4: Empirical coverage and coverage error of the predicted power (training: 2024, test: 2025)

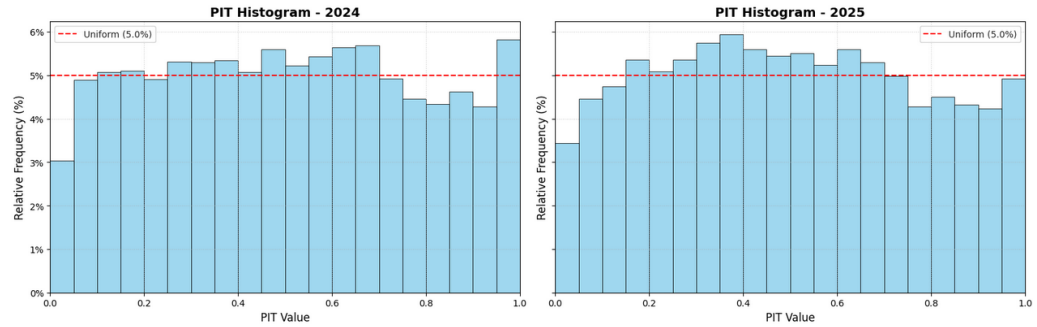


Figure 6.5: PIT of the predicted power (training: 2024, test: 2025)

As we can see, even though the goodness-of-fit plots show a widespread distribution of data, the empirical coverage is adequate, with a very low coverage error, and the pinball loss reaches its maximum at 1.02 for τ equal to 0.6, which can be considered a decent result.

6.3 Comparison with a direct method

To put our results into perspective, a simulated direct method was also used. In particular, a median regression was applied to the wind speed forecasts, that is, the covariates used for the lasso quantile regression, and the resulting predicted wind speed was taken as the input for the direct method, simulating data provided either by a second provider or by the first provider after applying data

correction techniques. The predicted wind speed was then used to forecast the power directly (the power curve was not estimated by fitting a model on the measured wind speed versus the power, as was done in the indirect method), via an additive B-splines quantile regression [Figure 6.6].

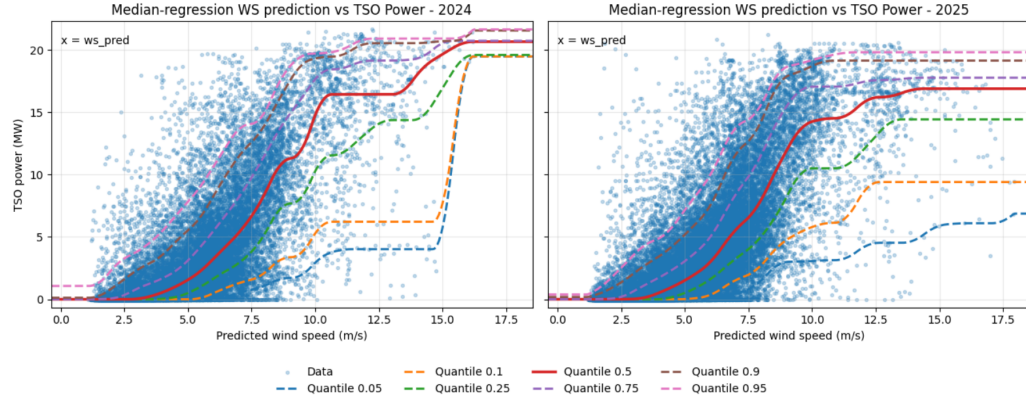


Figure 6.6: Direct method to predict the power with additive B-splines quantile regression fitted on 2024 and 2025 data

Then, with the model fitted on the 2024 data, we computed the same metrics used for the power, namely the pinball loss [Figure 6.7], the empirical coverage and coverage error [Figure 6.8], and the PIT [Figure 6.9].

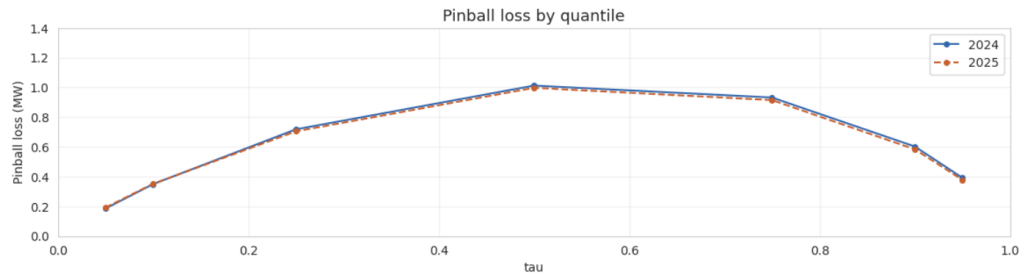


Figure 6.7: Pinball loss of the additive B-splines quantile regression (training: 2024, test: 2025)

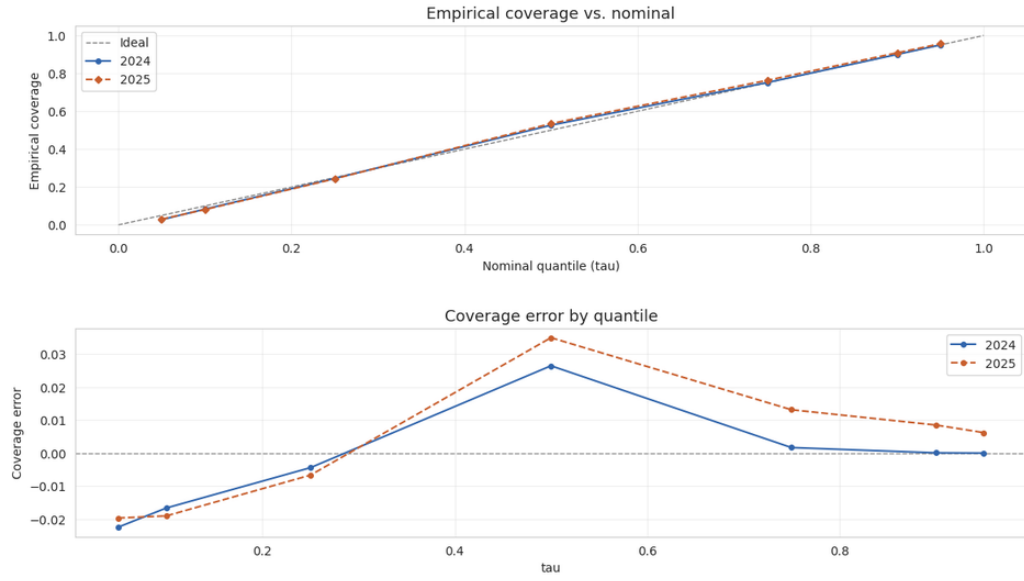


Figure 6.8: Empirical coverage and coverage error of the additive B-splines quantile regression (training: 2024, test: 2025)

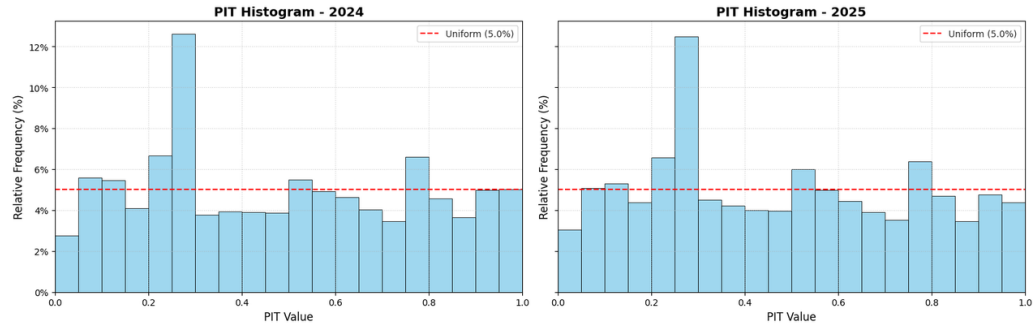


Figure 6.9: PIT of the additive B-splines quantile regression (training: 2024, test: 2025)

As we can see from figure 6.10, comparing the pinball loss of the power predicted via the direct method (left) with that of the indirect method (right), the two losses can be considered equivalent.

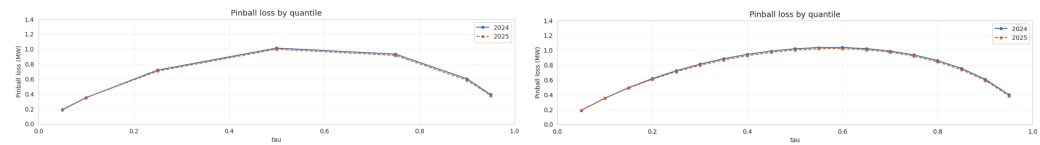


Figure 6.10: Comparison between the pinball loss of the additive B-splines quantile regression (left) and the pinball loss of the predicted power (right)

This demonstrates the strength of the indirect method, as it performs as well as the direct method while offering the additional benefits of modularity, since

it can be applied to different wind farms without requiring major changes, and greater interpretability, as its two steps convey meaningful information. As such, we considered the indirect method satisfactory for our purposes.

Conclusions

7.1 Conclusion

The aim of our project was to develop an approach to predict wind power from wind speed forecasts that would be interpretable, informative and modular. To this end, we employed probabilistic forecasting and wind speed recalibration.

In particular, we first considered the state of the art and discussed the advantages and disadvantages of probabilistic forecasting relative to point forecasting, as well as the merits of recalibrating the wind speed. We then conducted an in-depth analysis of the wind speed forecast dataset. After a correlation study, we retrieved a selected subset of the features to reduce the dimensionality of the problem, an objective we also pursued later, during the recalibration phase. We also examined the dataset comprising the measured wind speed and the wind power from the TSO, computing its scatterplot. To obtain a better model, we cleaned the data and subsequently plotted the power curve using a linear generalized additive model.

The recalibration phase entailed first trialing various models, whose performance, along with their ability to reduce the covariate space, was evaluated in order to identify the best among them. This led us to choose a lasso quantile regression for recalibrating the wind speed forecasts, whose metrics were computed and compared to those of the marginal model, with the former outperforming the latter. Lastly, we were able to predict the power by using the lasso quantile regression for recalibration and then feeding its output into the linear GAM. Our indirect approach was compared to a direct approach in order to evaluate our results.

As our metrics showed, both approaches achieved equivalent performance; this demonstrates the validity of the indirect approach, even under the assumption

of a deterministic power curve. Therefore, not only does this method achieve results comparable to the direct one, but it is also more interpretable, since its two steps both deliver clear outputs, and the characteristics of the inputs are not left for the algorithm to infer, but are thoroughly accounted for by this approach. Additionally, given its modularity, it can be easily applied to wind farms other than the one examined.

Our results also show that our second strongest assumption, after relaxing the stochasticity of the power curve, that is, modeling the entire wind farm as a single turbine, is a legitimate one. Even without considering each turbine individually, we can still model the wind farm adequately, greatly reducing the complexity of the problem and adding further modularity to our approach.

7.2 Future work

To expand our work and improve its performance, two enhancements can be considered.

The first is to actually treat the power curve as stochastic and model it accordingly. This could likely lead to better accuracy, especially in the region affected by the curtailment phenomenon, which was the least stable with the linear GAM. The second is to incorporate other covariates in addition to the wind speed forecasts. In particular, wind direction (which is further analyzed in the Appendix) could be a useful addition to the model, as we expect wind coming from a given direction to have different characteristics, including speed, than wind coming from another direction. Moreover, given that wind exhibits several kinds of periodicity, adding variables such as the hour of the day and the day of the year could improve overall performance, as different kinds of wind blow during the day and night and across the seasons of the year.

Author's Note

In the process of writing this thesis Generative AI tools (Claude 5) were utilized for revising what already written and bibliographic research. All the generated content was carefully revised and verified to ensure reliability and accuracy.

Chapter 8

Appendix

8.1 Chapter overview

In this chapter we provide additional graphs that we plotted to further study the data during our project. In particular, we focus on wind direction and the relationship between the wind speed forecasts and other possible regressors.

8.2 Appendix

8.2.1 Scattermatrices

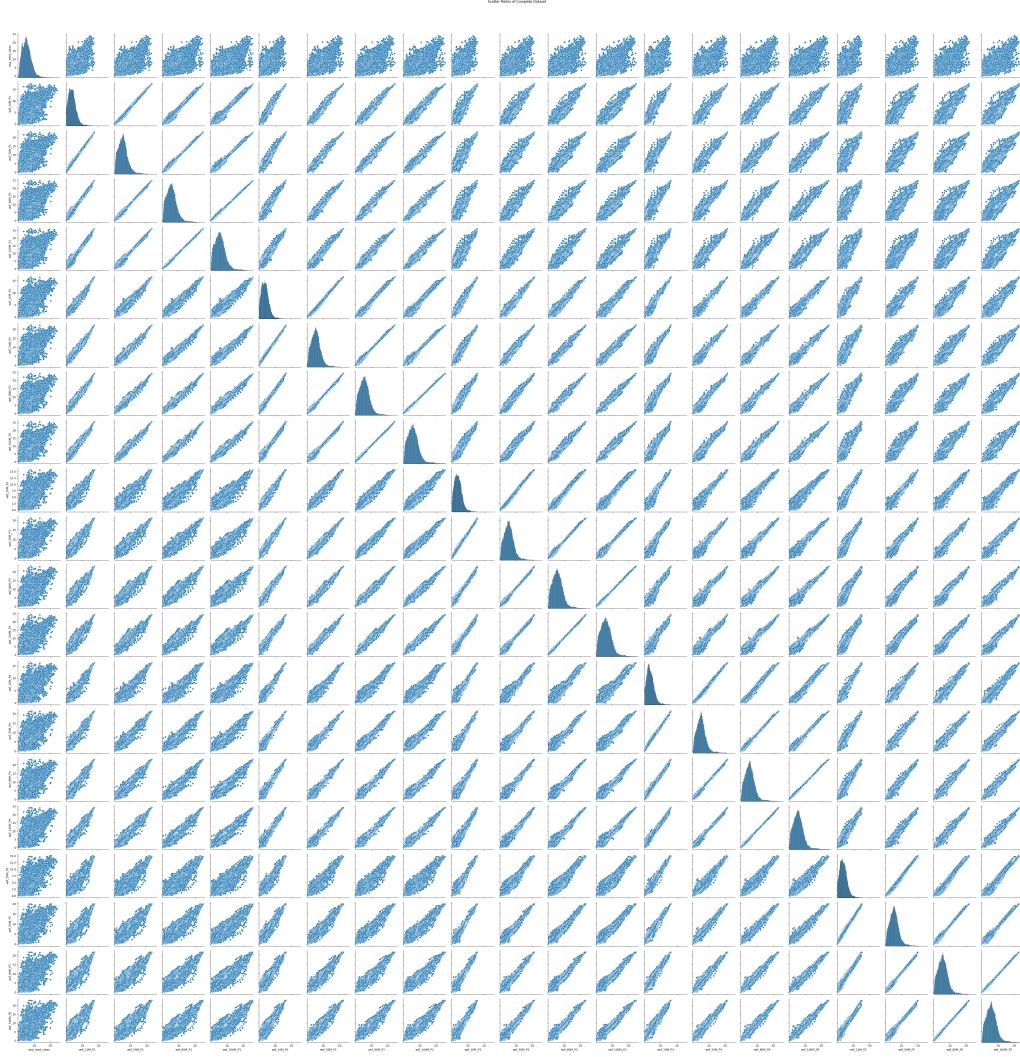


Figure 8.1: Scattermatrix of the complete dataset (before the subset selection of the features), with the windspeed forecasts at the four heights (10, 50, 80 and 100 metres) in the five grid points (P1-P5)

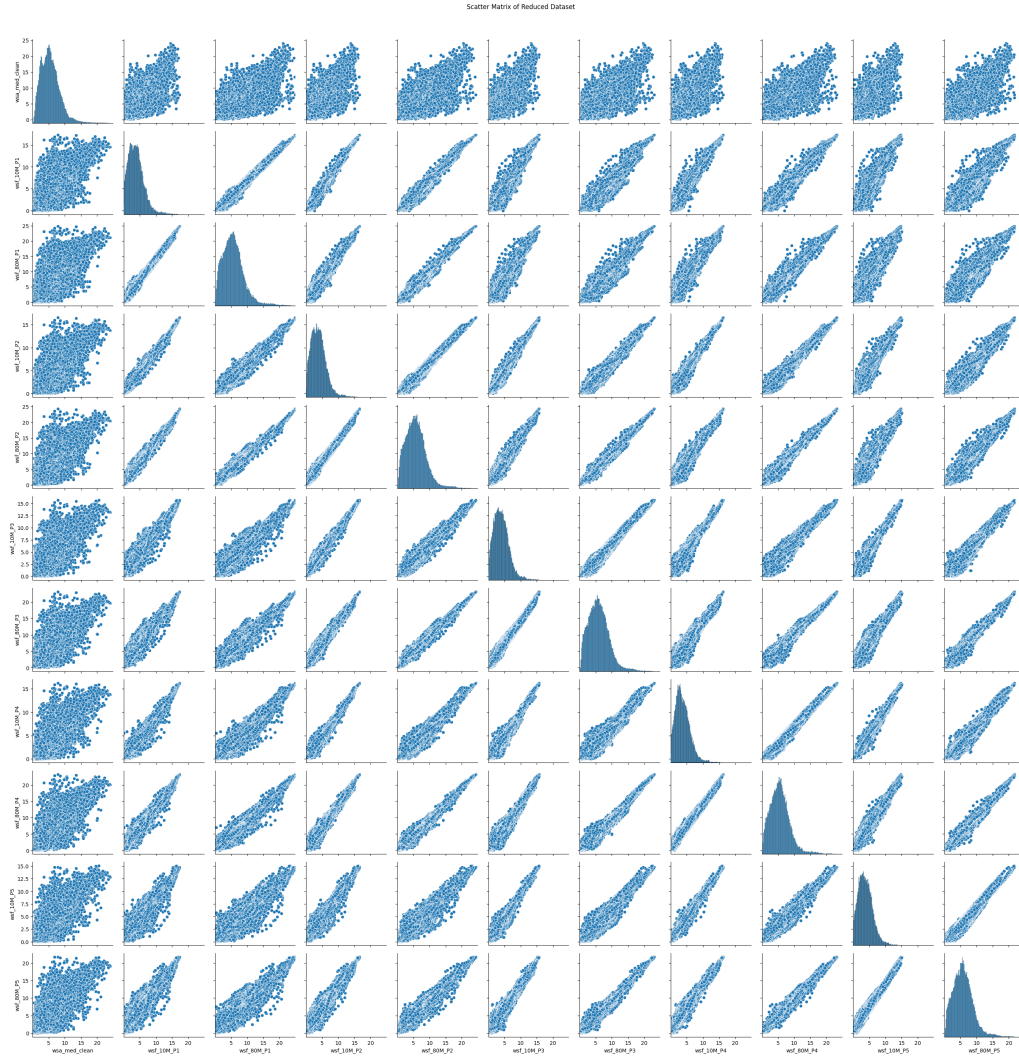


Figure 8.2: Scattermatrix of the reduced dataset (after the subset selection of the features), with the windspeed forecasts at the two heights (10 and 80 metres) in the five grid points (P1-P5)

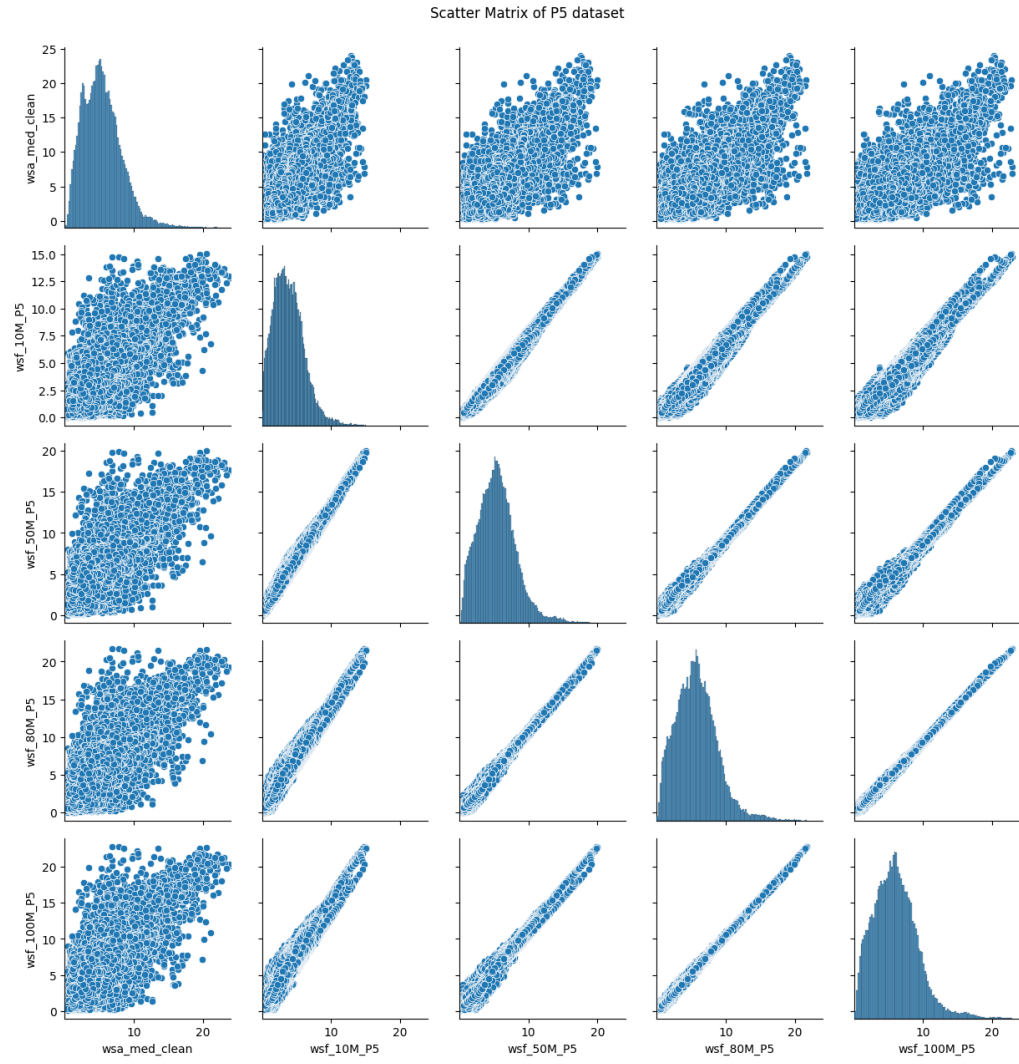


Figure 8.3: Scattermatrix of the P5 dataset, with the wind speed forecasts at the four heights (10, 50, 80 and 100 metres) in the P5 grid point

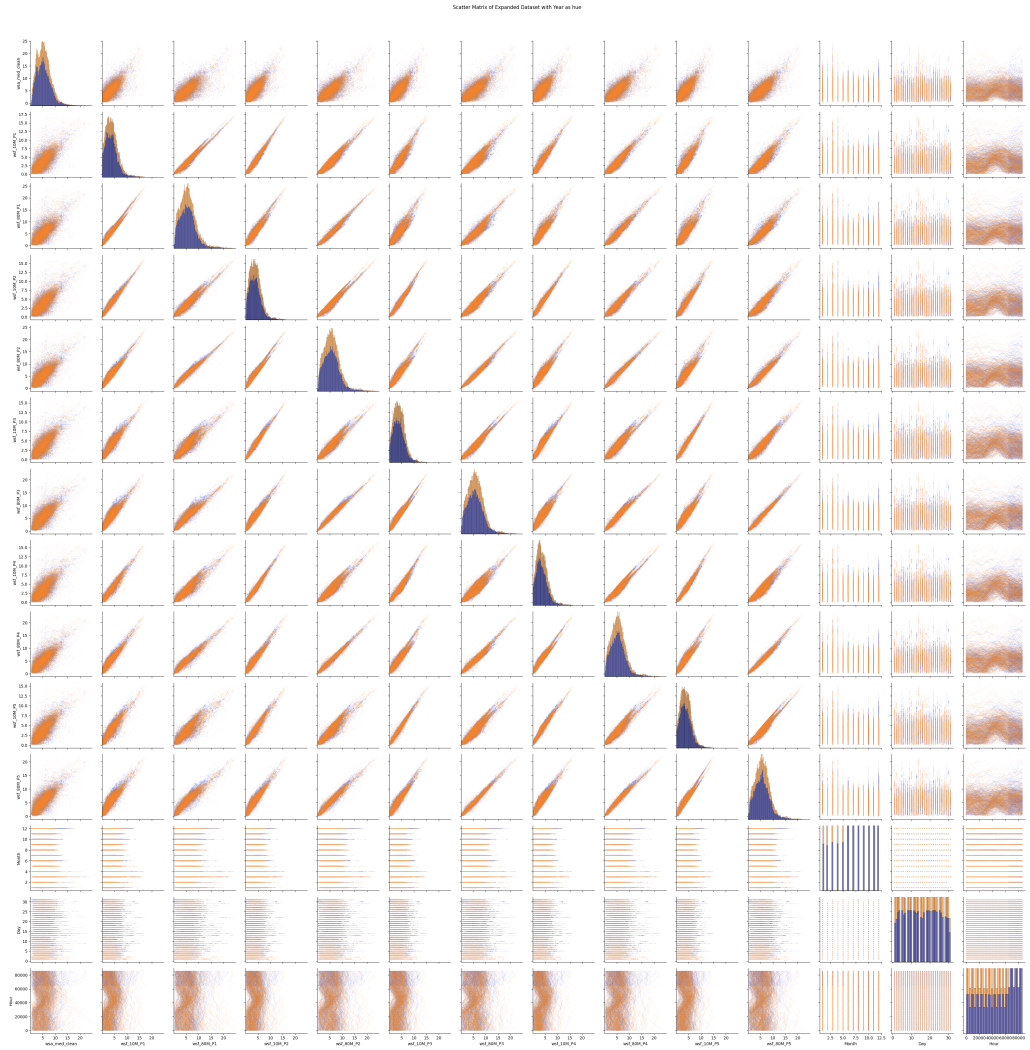


Figure 8.4: Scatter matrix of the expanded dataset, with the windspeed forecasts at the two heights (10 and 80 metres) in the five grid points (P1-P5) and the Month, Day and Hour, using the Year as hue

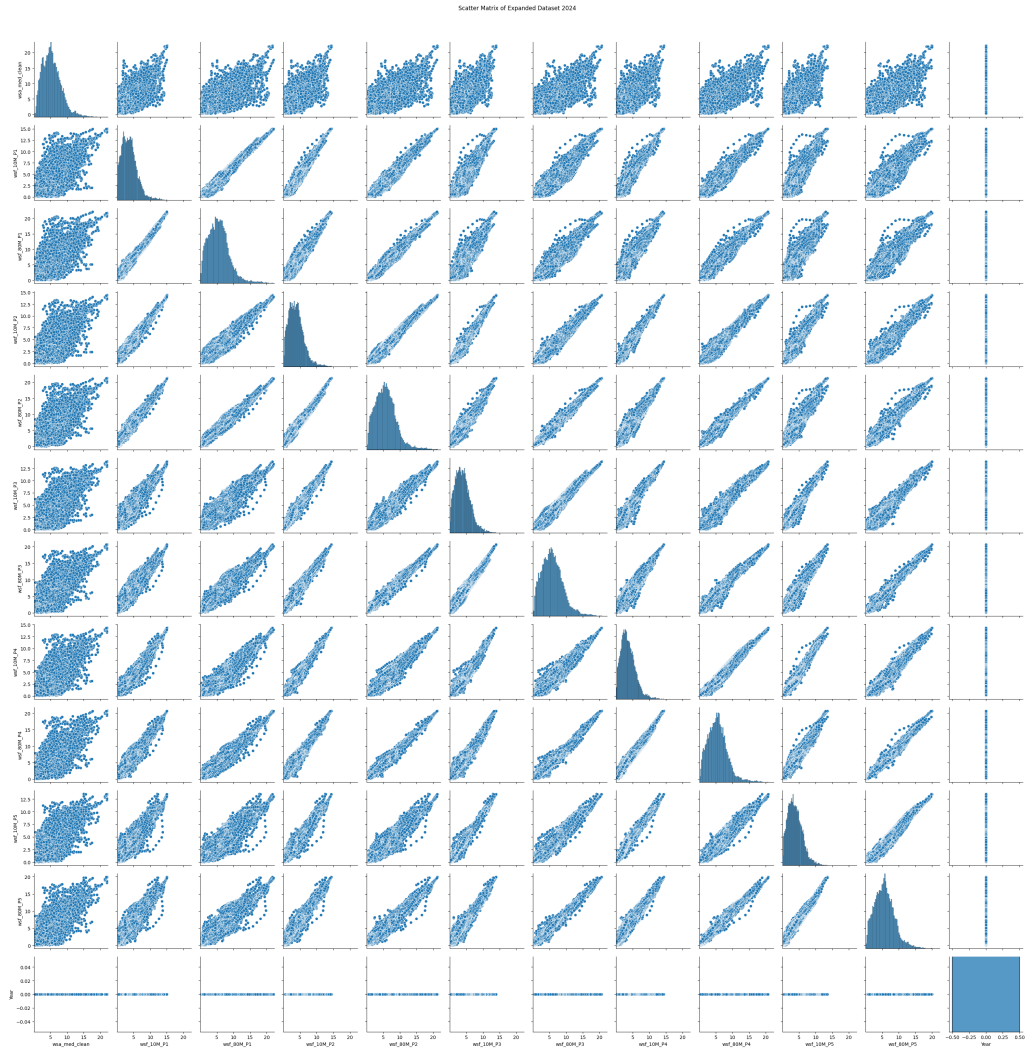


Figure 8.5: Scattermatrix of the expanded dataset of 2024, with the windspeed forecasts at the two heights (10 and 80 metres) in the five grid points (P1-P5) and the Year, of 2024 data

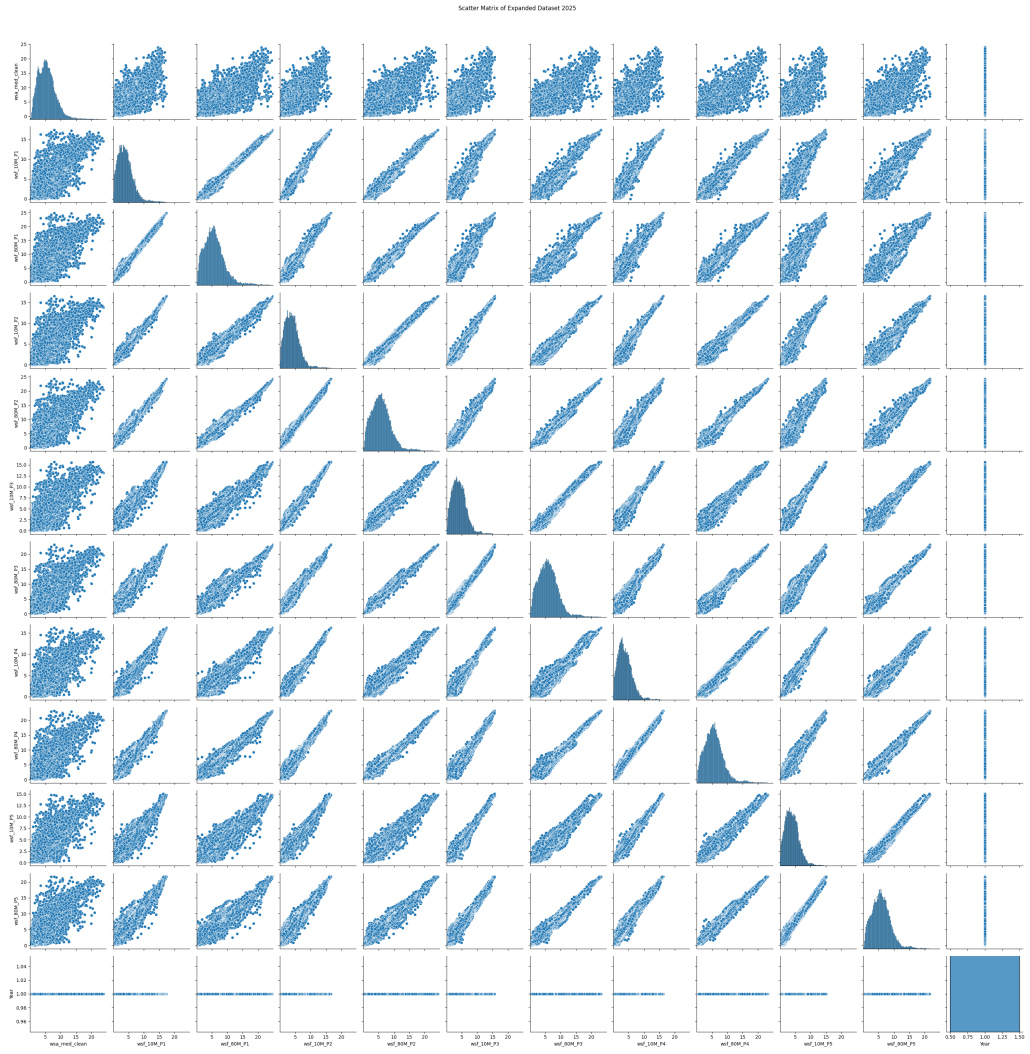


Figure 8.6: Scattermatrix of the expanded dataset of 2025, with the windspeed forecasts at the two heights (10 and 80 metres) in the five grid points (P1-P5) and the Year, of 2025 data

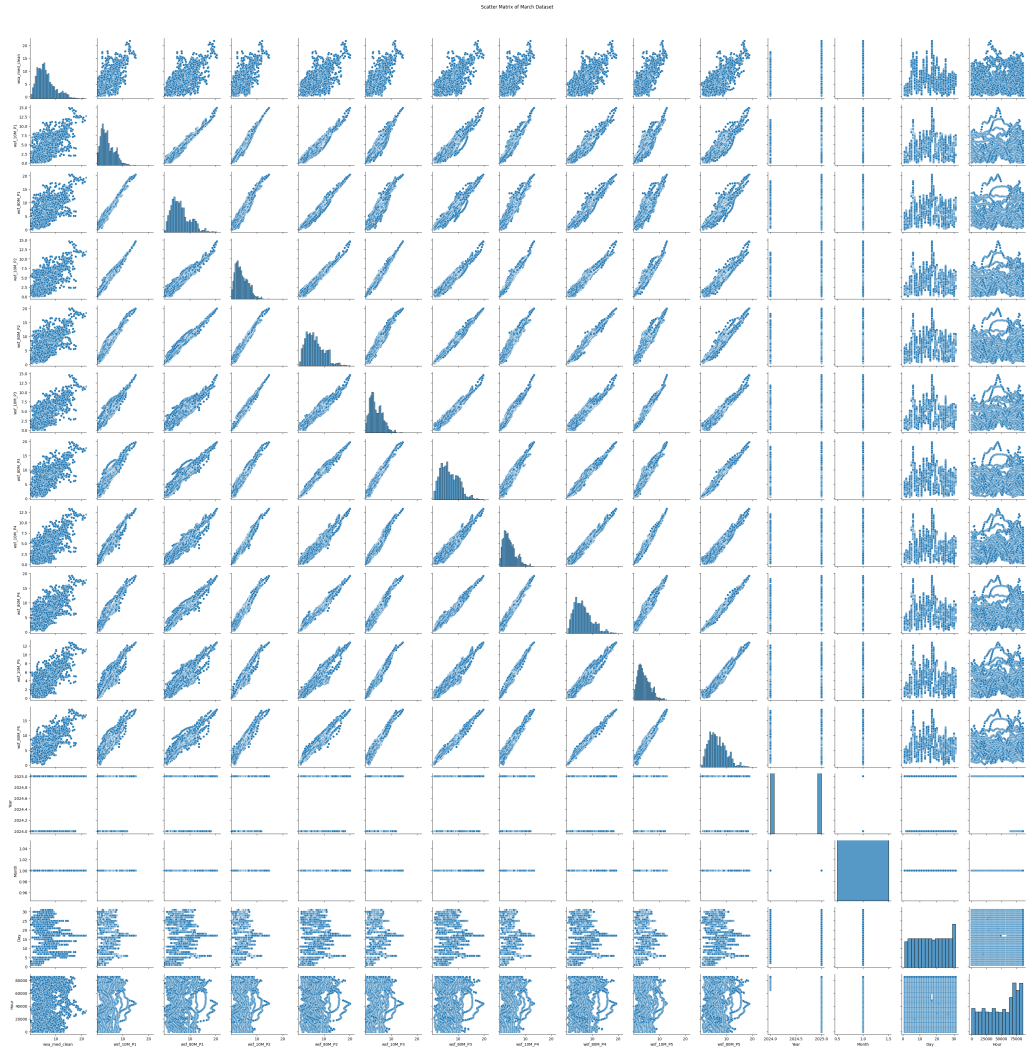


Figure 8.7: Scattermatrix of the March dataset, with the windspeed forecasts at the two heights (10 and 80 metres) in the five grid points (P1-P5) and the Year, Month, Day and Hour, of March data

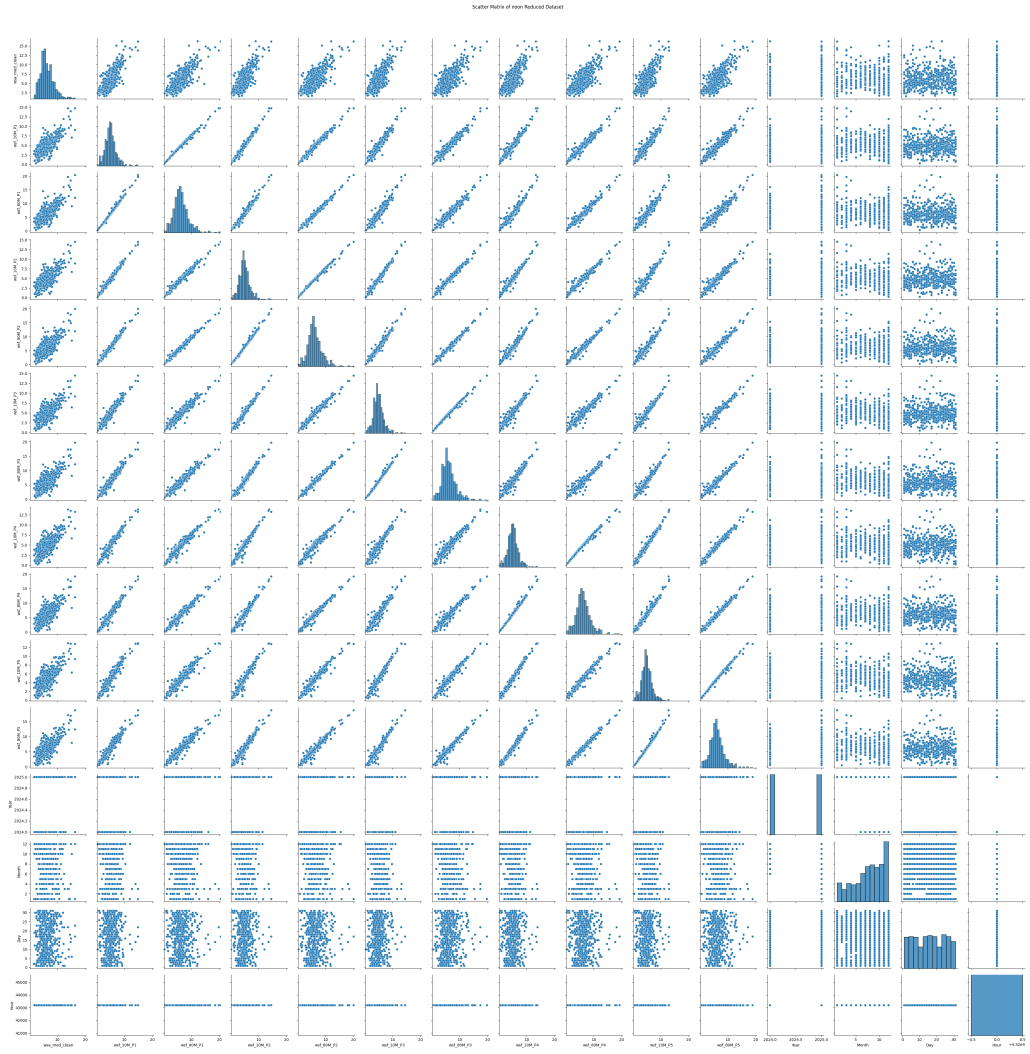


Figure 8.8: Scattermatrix of the Noon dataset, with the windspeed forecasts at the two heights (10 and 80 metres) in the five grid points (P1-P5) and the Year, Month, Day and Hour, of Noon data

8.2.2 Histograms

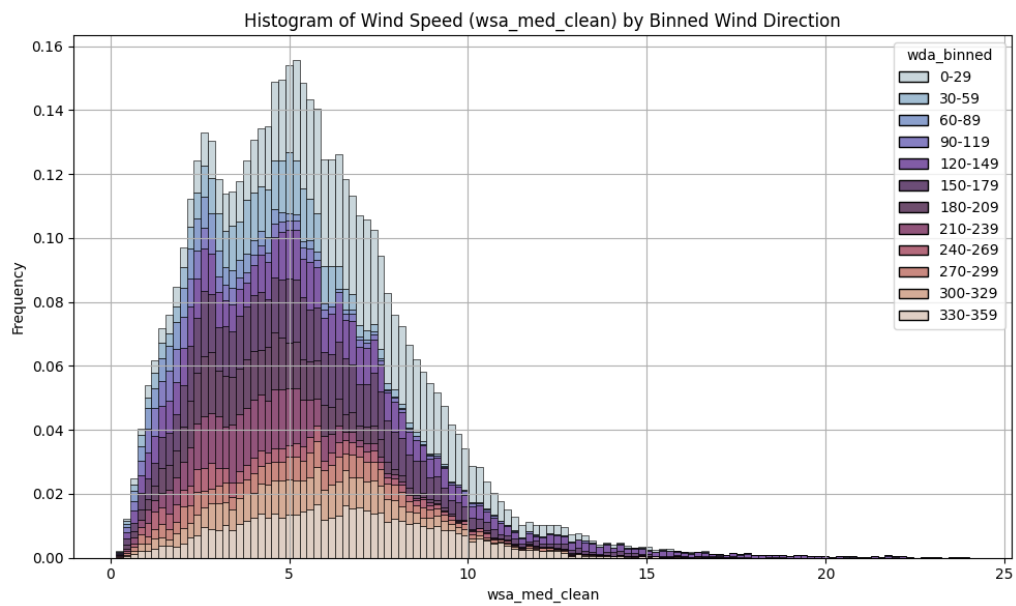


Figure 8.9: Histogram of the measured wind speed with the wind direction as hue

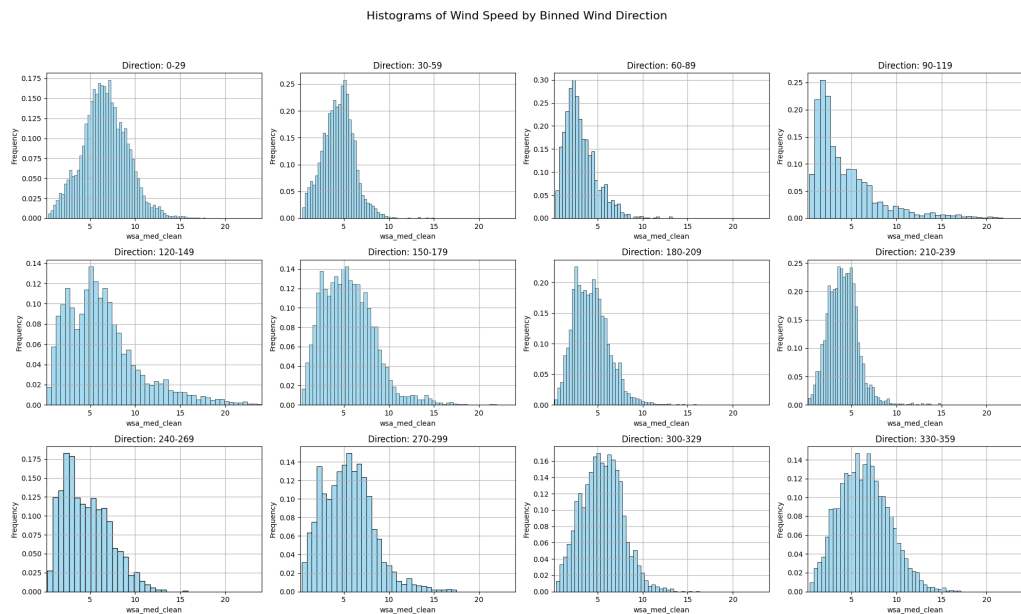


Figure 8.10: Histograms of the measured wind speed divided by wind direction intervals. This is an expansion of figure 8.9

Histograms of Wind Speed by Binned Regressor Variables

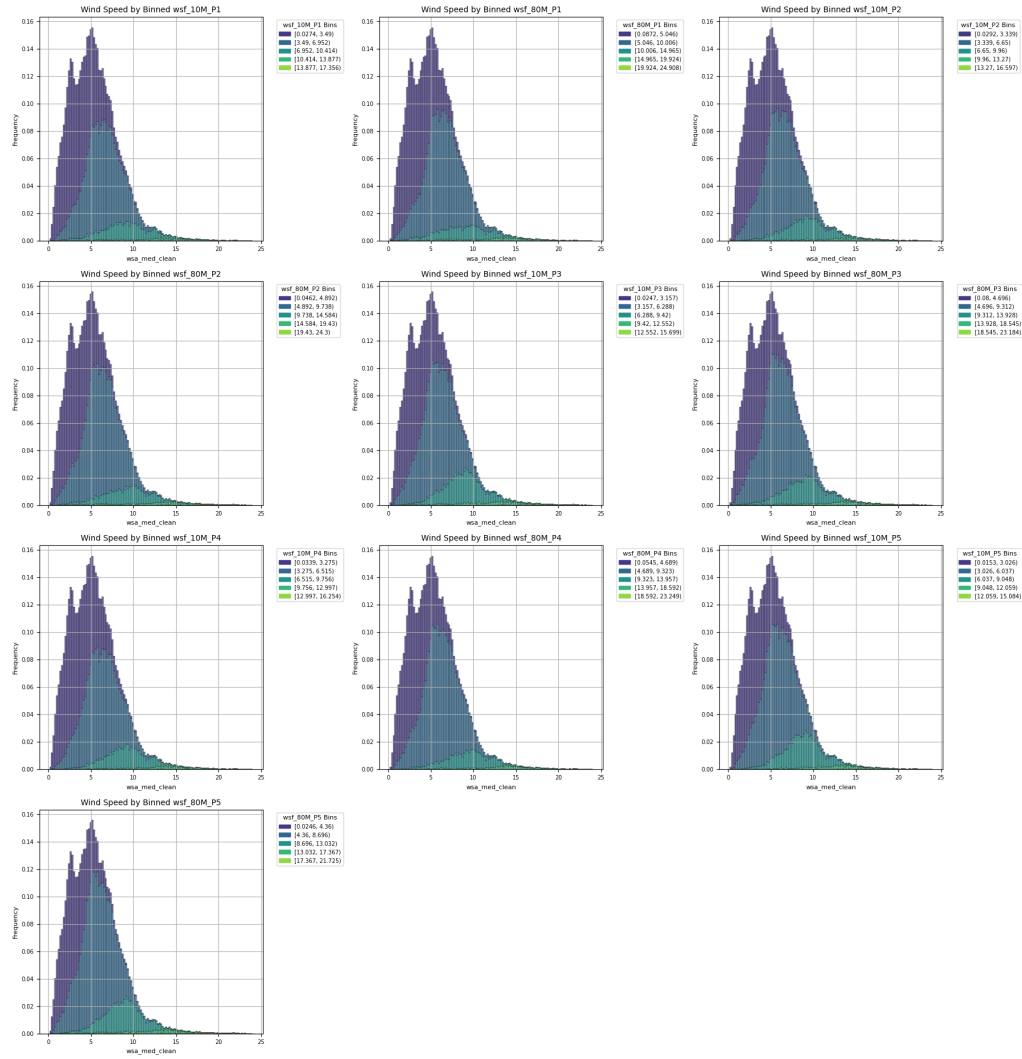


Figure 8.11: Histogram of the measured wind speed with the forecast wind speed as hue for each windspeed forecasts at the two heights (10 and 80 metres) in the five grid points (P1-P5)

Histograms of Wind Speed by Binned Wind Speed Forecast at 80 metres in P5

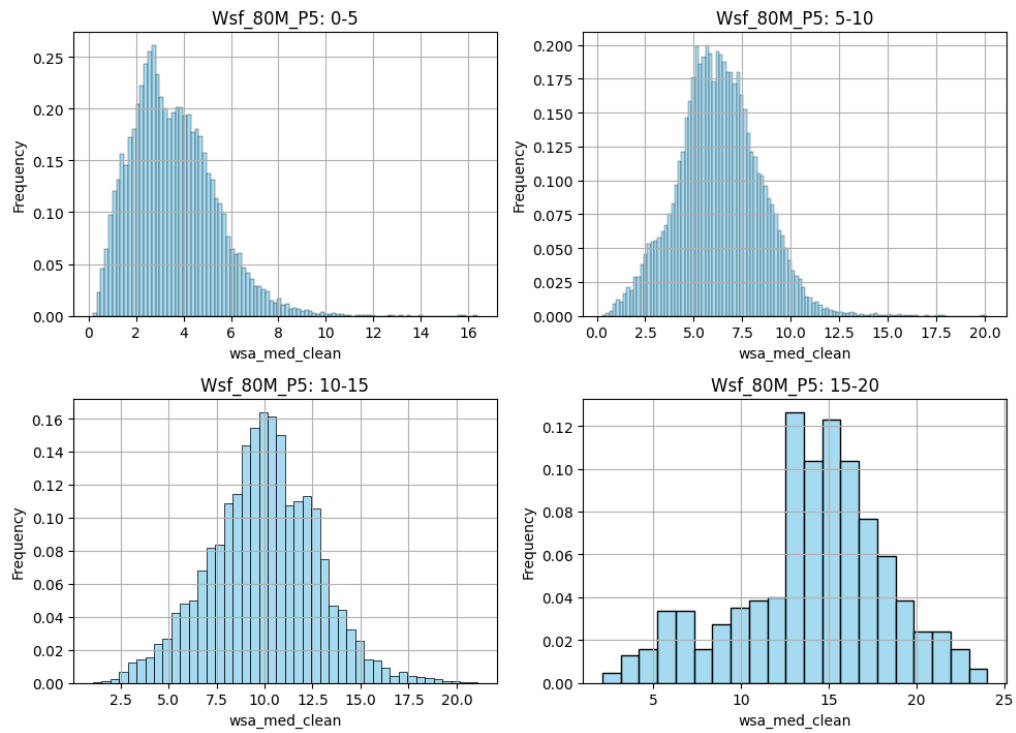


Figure 8.12: Histograms of the measured wind speed divided by wind speed intervals for the winspeed forecast at 80 metres in P5. This is an expansion of figure 8.11

8.2.3 Windrose

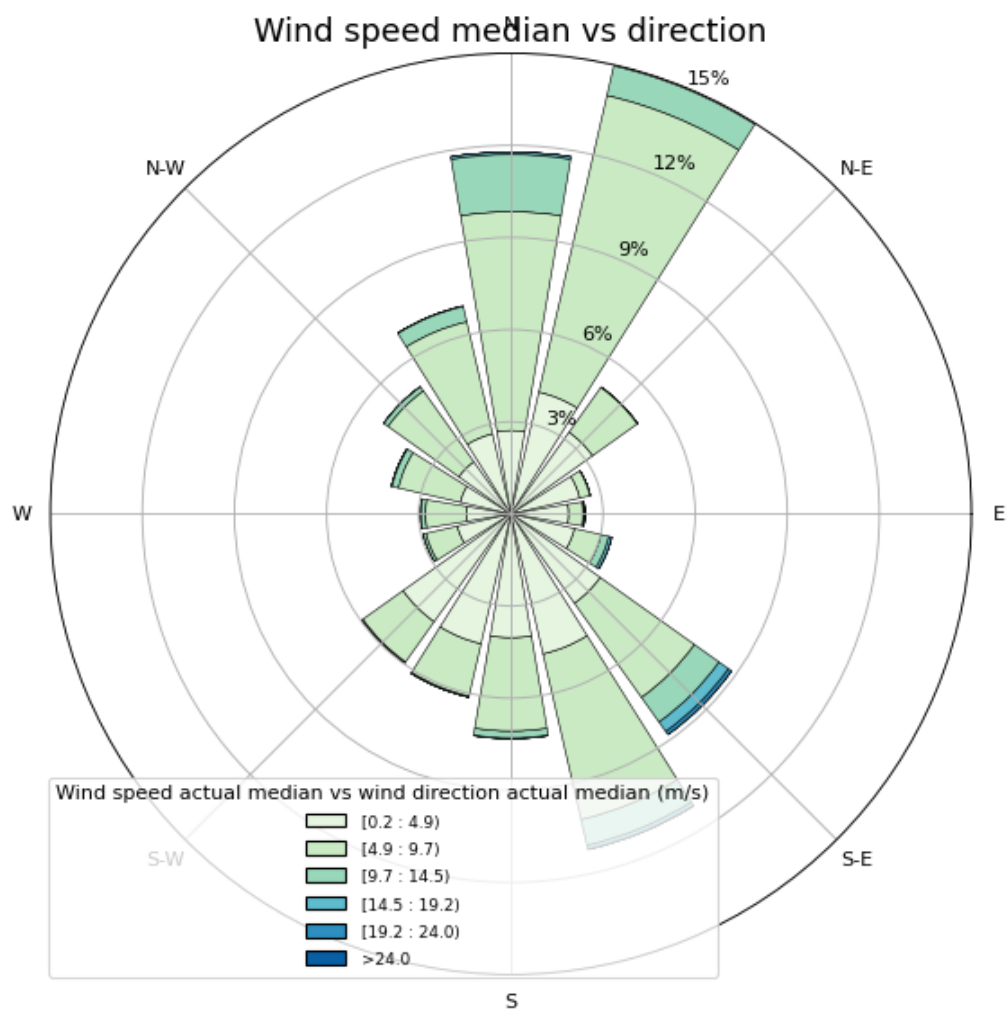


Figure 8.13: Windrose of the measured wind speed versus wind direction. The former is used as hue and its frequency is shown as percentages, while the latter is depicted by using NESW coordinates

8.2.4 Scatterplots

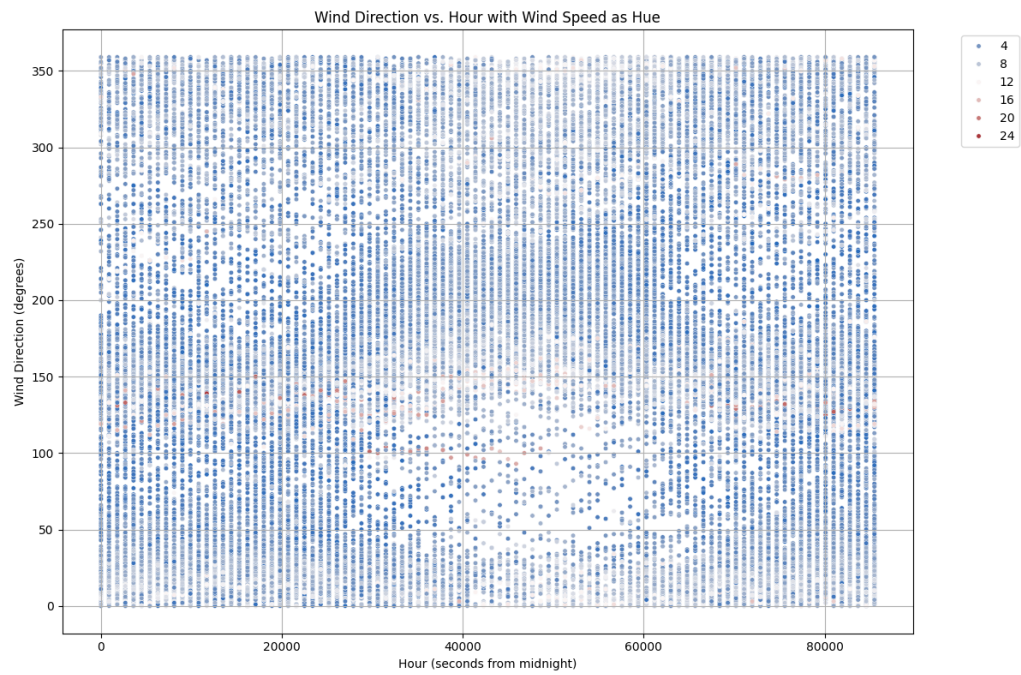


Figure 8.14: Scatterplot of the hour (represented with seconds from midnight) versus the wind direction with wind speed as hue

Scatter Plots of Wind Speed vs. Wind Direction Components (T_XX)

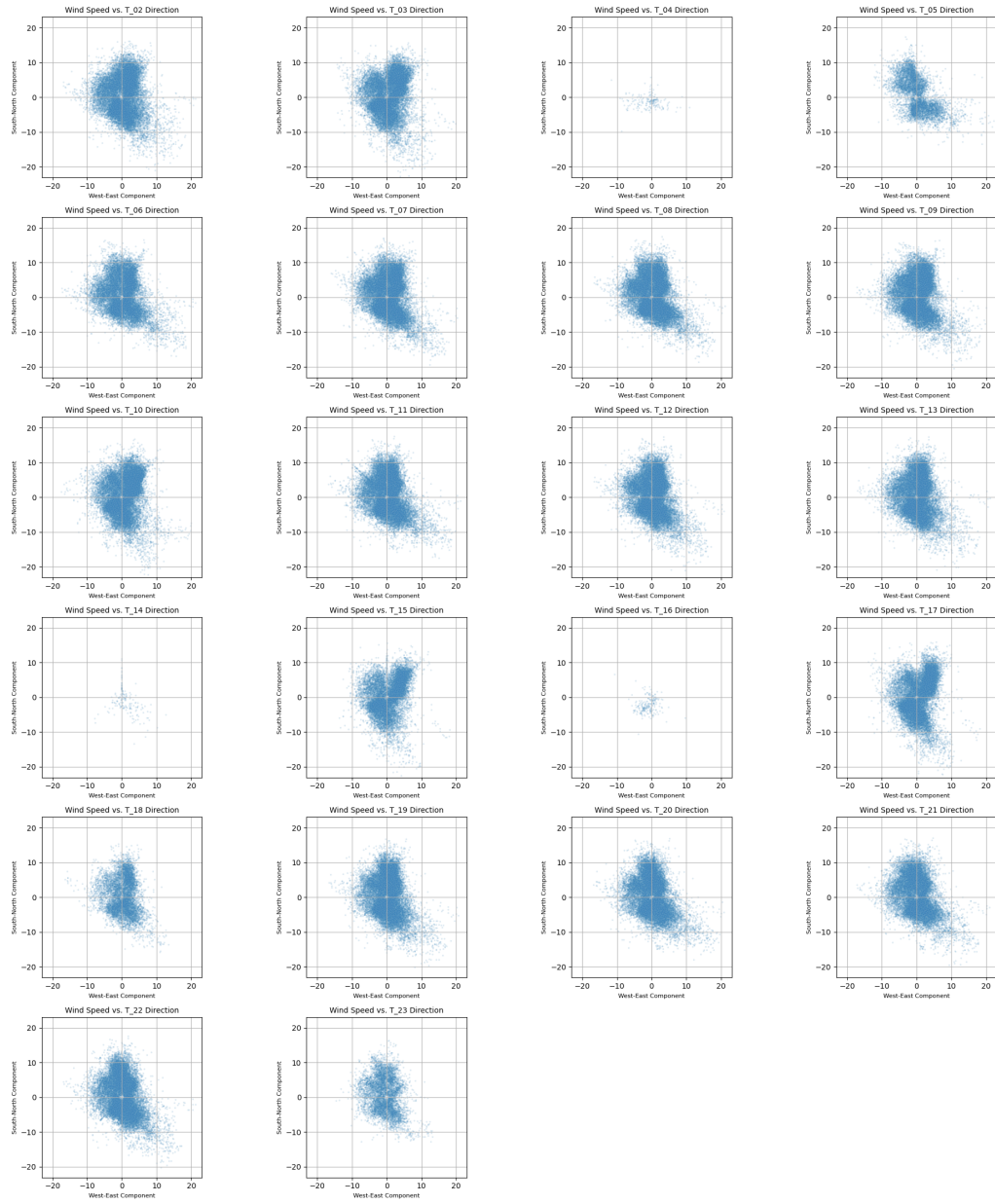


Figure 8.15: Scatterplots of the measured wind speed (the absolute distance from the zero) and the wind direction components of a subset of the turbines

Scatter Plots of Wind Speed vs. Wind Direction Components (Forecast)

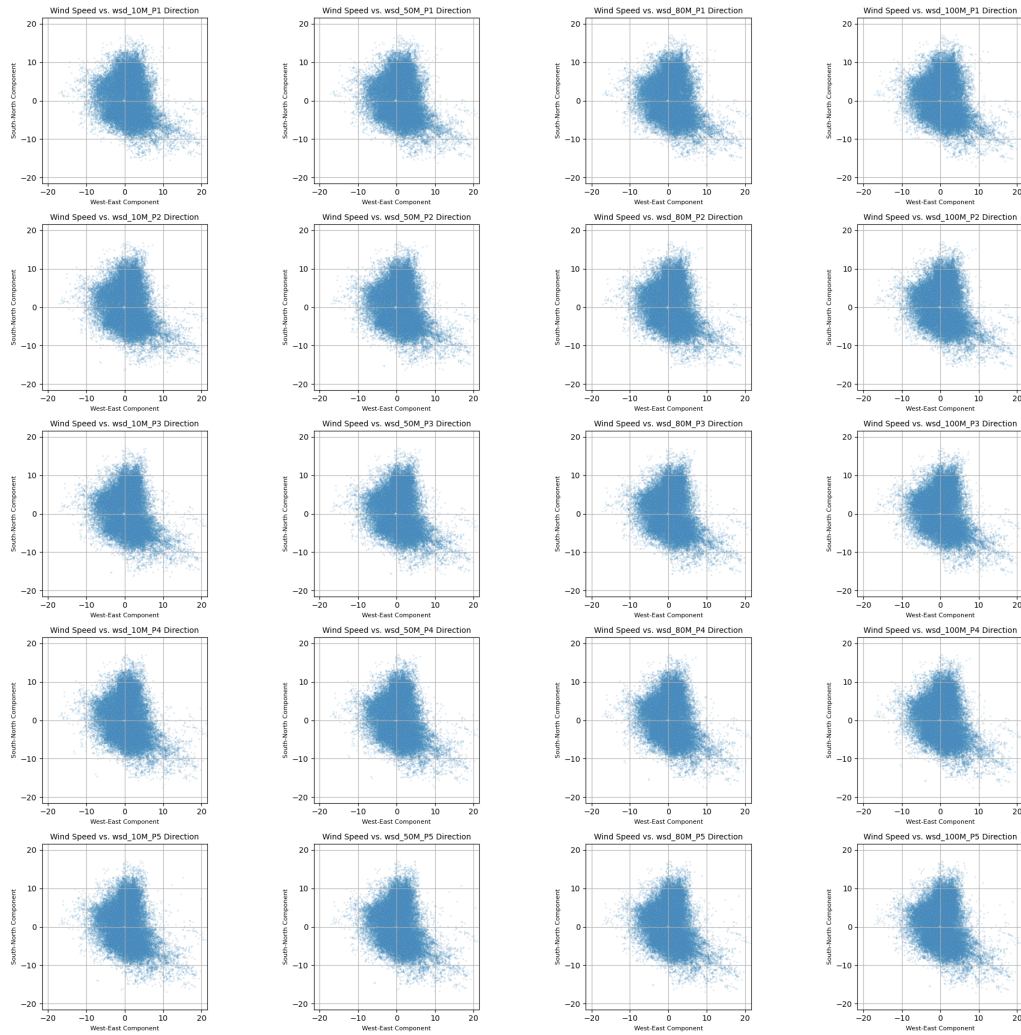


Figure 8.16: Scatterplots of the wind speed forecasts (the absolute distance from the zero) and the wind direction components of a subset of the turbines

8.2.5 Spaghetti plots

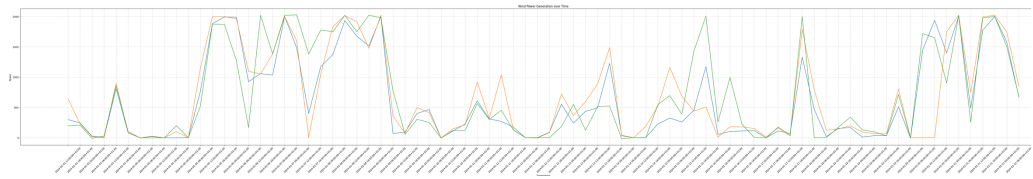


Figure 8.17: Spaghetti plot of the time series of the power of three turbines (T_04, T_14, T_16) to inspect their behavior

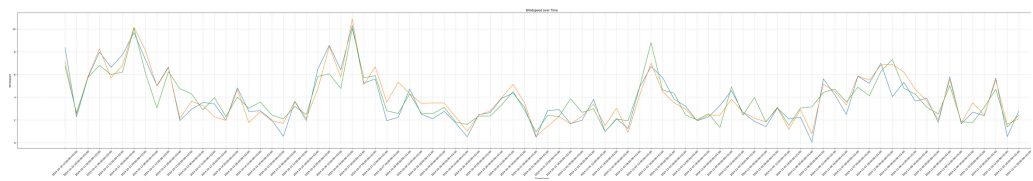


Figure 8.18: Spaghetti plot of the time series of the wind speed of three turbines (T_04, T_14, T_16) to inspect their behavior

8.2.6 PCAs

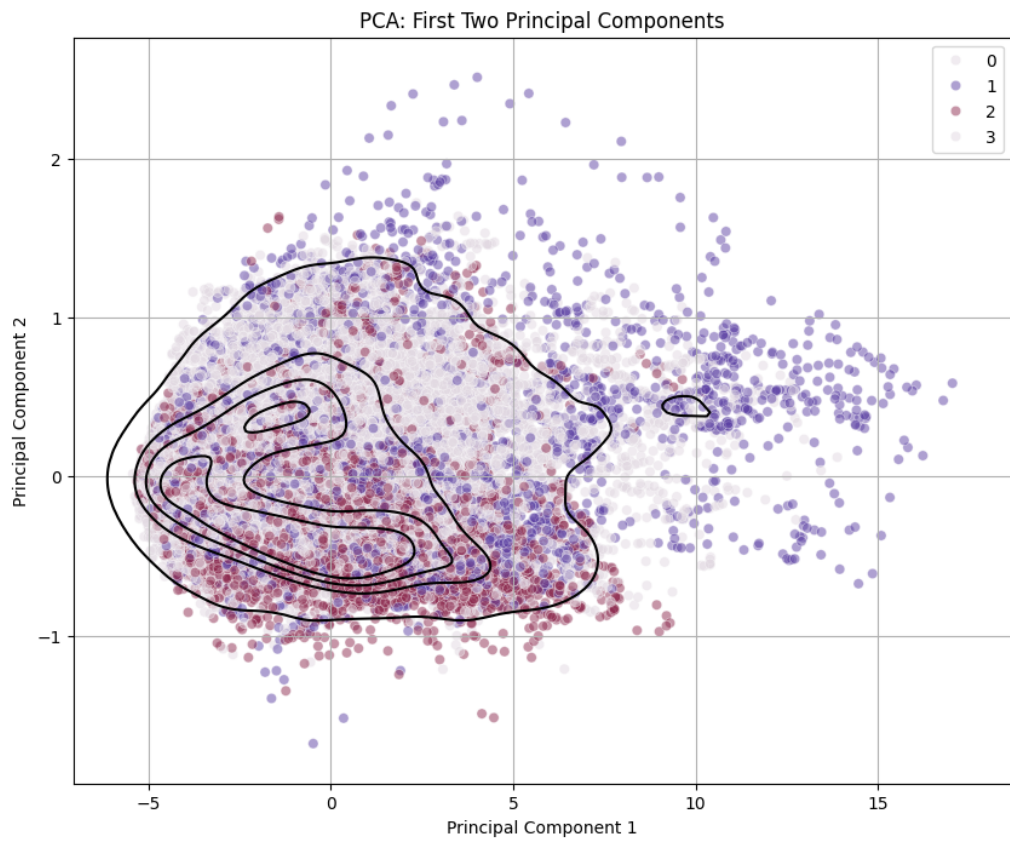


Figure 8.19: First two principal components of the season dataset, with the wind-speed forecasts at the two heights (10 and 80 metres) in the five grid points (P1-P5) and the Season, which is the hue. In particular 0 is winter, 1 is spring, 2 is summer and 3 is autumn.

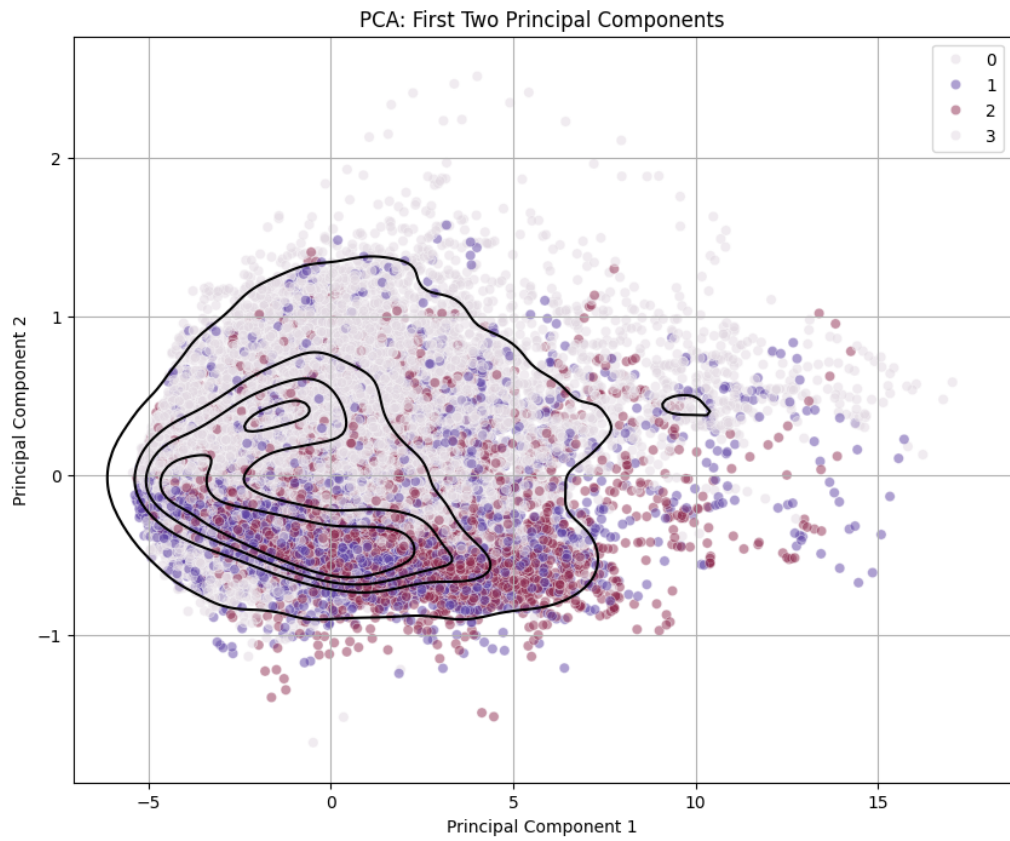


Figure 8.20: First two principal components of the day periods dataset, with the windspeed forecasts at the two heights (10 and 80 metres) in the five grid points (P1-P5) and the Day Period, which is the hue. In particular 0 is from 00 a.m. to 5.59 a.m., 1 is from 6 a.m. to 11.59 a.m., 2 is from 12 p.m. to 5.59 p.m. and 3 is from 6 p.m. to 11.59 p.m.

Bibliography

- Amat Rodrigo J. and Escobar Ortiz J. *Probabilistic forecasting: prediction intervals and prediction distribution*. Skforecast Docs, 2024. URL https://skforecast.org/0.13.0/user_guides/probabilistic-forecasting. Version 0.13.0.
- Buizza R. The value of probabilistic prediction. *Atmospheric Science Letters*, 9(2):36–42, 2008. doi: 10.1002/asl.170. URL <https://rmets.onlinelibrary.wiley.com/doi/full/10.1002/asl.170>.
- Capelletti M. and De Nicolao G. Indirect probabilistic recalibration of wind speed forecast via the inverse power curve. *IFAC-PapersOnLine*, 59(26): 289–294, 2025. ISSN 2405-8963. doi: <https://doi.org/10.1016/j.ifacol.2025.12.049>. URL <https://www.sciencedirect.com/science/article/pii/S2405896325027259>. 7th IFAC Conference on Intelligent Control and Automation Sciences ICONS 2025.
- Capelletti M., Raimondo D. M., and De Nicolao G. Wind power curve modeling: A probabilistic beta regression approach. *Renewable Energy*, 223: 119970, 2024. ISSN 0960-1481. doi: <https://doi.org/10.1016/j.renene.2024.119970>. URL <https://www.sciencedirect.com/science/article/pii/S0960148124000351>.
- Chang W.-Y. Application of back propagation neural network for wind power generation forecasting. *International Journal of Digital Content Technology and its Applications*, 7(4):502, 2013.
- De Giorgi M. G., Ficarella A., and Tarantino M. Assessment of the benefits of numerical weather predictions in wind power forecasting based on statistical methods. *Energy*, 36(7):3968–3978, 2011. ISSN 0360-5442. doi: <https://doi.org/10.1016/j.energy.2011.05.006>. URL <https://www.sciencedirect.com/science/article/pii/S0360544211003215>.

- Gneiting T. Quantiles as optimal point forecasts. *International Journal of Forecasting*, 27(2):197–207, 2011. ISSN 0169-2070. doi: <https://doi.org/10.1016/j.ijforecast.2009.12.015>. URL <https://www.sciencedirect.com/science/article/pii/S0169207010000063>.
- Gneiting T., Balabdaoui F., and Raftery A. E. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007. doi: 10.1111/j.1467-9868.2007.00587.x. URL <https://sites.stat.washington.edu/raftery/Research/PDF/Gneiting2007jrssb.pdf>.
- Hong Y.-Y. and Rioflorida C. L. P. P. A hybrid deep learning-based neural network for 24-h ahead wind power forecasting. *Applied Energy*, 250: 530–539, 2019. ISSN 0306-2619. doi: <https://doi.org/10.1016/j.apenergy.2019.05.044>. URL <https://www.sciencedirect.com/science/article/pii/S030626191930889X>.
- James G., Witten D., Hastie T., Tibshirani R., and Taylor J. *An Introduction to Statistical Learning: With Applications in Python*. Springer Texts in Statistics. Springer, Cham, 2023. doi: 10.1007/978-3-031-38747-0. URL https://hastie.su.domains/ISLP/ISLP_website.pdf.download.html.
- Jyothi M. N. and Rao P. V. R. Very-short term wind power forecasting through adaptive wavelet neural network. In *2016 Biennial International Conference on Power and Energy Systems: Towards Sustainable Energy (PESTSE)*, pages 1–6, 2016. doi: 10.1109/PESTSE.2016.7516513.
- Koenker R. Quantile regression. Course notes, Department of Economics, University of Illinois at Urbana-Champaign, 1998. URL <http://www.econ.uiuc.edu/~roger/courses/472/1998/rq.pdf>.
- Koenker R. and Hallock K. F. Quantile regression. *Journal of Economic Perspectives*, 15(4):143–156, December 2001. doi: 10.1257/jep.15.4.143. URL <https://www.aeaweb.org/articles?id=10.1257/jep.15.4.143>.
- Liu Z., Guo H., Zhang Y., and Zuo Z. A comprehensive review of wind power prediction based on machine learning: Models, applications, and challenges. *Energies*, 18(2), 2025. ISSN 1996-1073. doi: 10.3390/en18020350. URL <https://www.mdpi.com/1996-1073/18/2/350>.

- Maciejowska K., Serafin T., and Uniejewski B. Probabilistic forecasting with a hybrid factor-qra approach: Application to electricity trading. *Electric Power Systems Research*, 234:110541, 2024. ISSN 0378-7796. doi: <https://doi.org/10.1016/j.epsr.2024.110541>. URL <https://www.sciencedirect.com/science/article/pii/S0378779624004279>.
- Narayan T., Wang S., Canini K., and Gupta M. R. Expected pinball loss for quantile regression and inverse cdf estimation. *Transactions on Machine Learning Research*, 2024. URL <https://openreview.net/pdf?id=Eg8Rnb0Hdd>.
- Nowotarski J. and Weron R. Computing electricity spot price prediction intervals using quantile regression and forecast averaging. *Computational Statistics*, 30: 791–803, 2015. doi: 10.1007/s00180-014-0523-0. URL <https://link.springer.com/article/10.1007/s00180-014-0523-0>.
- pyGAM Developers . *pyGAM Documentation*. Read the Docs, 2025. URL <https://pygam.readthedocs.io/en/latest/>. Accessed 24 June 2026.
- Rajaperumal T. A. and Christopher Columbus C. Enhanced wind power forecasting using machine learning, deep learning models and ensemble integration. *Scientific Reports*, 15, 2025. doi: 10.1038/s41598-025-05250-3. URL <https://www.nature.com/articles/s41598-025-05250-3>.
- scikit-learn Developers . *sklearn.decomposition.PCA*. scikit-learn, 2026a. URL <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>. Version 1.9.0 documentation.
- scikit-learn Developers . *sklearn.linear_model.QuantileRegressor*. scikit-learn, 2026b. URL https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.QuantileRegressor.html. Version 1.9.0 documentation.
- Serafin T. and Uniejewski B. Ranking probabilistic forecasting models with different loss functions, 2024. URL <https://arxiv.org/html/2411.17743v1>. Version 1, 24 November 2024.
- Statsmodels Developers . *Linear Regression*. statsmodels, 2024. URL <https://www.statsmodels.org/v0.14.4/regression.html>. Version 0.14.4, last updated 3 October 2024.

- Tibshirani R. Forecast scoring and calibration. Lecture notes for Advanced Topics in Statistical Learning, Spring 2023, 2023. URL <https://www.stat.berkeley.edu/~ryantibs/statlearn-s23/lectures/calibration.pdf>.
- Xu L. and Mao J. Short-term wind power forecasting based on elman neural network with particle swarm optimization. In *2016 Chinese Control and Decision Conference (CCDC)*, pages 2678–2681, 2016. doi: 10.1109/CCDC.2016.7531436.
- Zhang J., Yan J., Infield D., Liu Y., and Lien F.-S. Short-term forecasting and uncertainty analysis of wind turbine power based on long short-term memory network and gaussian mixture model. *Applied Energy*, 241: 229–244, 2019. ISSN 0306-2619. doi: <https://doi.org/10.1016/j.apenergy.2019.03.044>. URL <https://www.sciencedirect.com/science/article/pii/S0306261919304532>.
- Zhang Y., Wang J., and Wang X. Review on probabilistic forecasting of wind power generation. *Renewable and Sustainable Energy Reviews*, 32: 255–270, 2014. ISSN 1364-0321. doi: <https://doi.org/10.1016/j.rser.2014.01.033>. URL <https://www.sciencedirect.com/science/article/pii/S1364032114000446>.
- Zhao Y., Ye L., Li Z., Song X., Lang Y., and Su J. A novel bidirectional mechanism based on time series model for wind power forecasting. *Applied Energy*, 177:793–803, 2016. ISSN 0306-2619. doi: <https://doi.org/10.1016/j.apenergy.2016.03.096>. URL <https://www.sciencedirect.com/science/article/pii/S0306261916304263>.