

Enhancing Weather Data Reconstruction through Hybrid Methods with Dimensionality Reduction

MURILO MONTANINI BREVE

Work carried out under the guidance of
Prof. Dr. Carlos Balsa; Prof. Dr. José Rufino
Prof. Dr. Fabricio Martins

Master's degree in Industrial Engineering
2022-2023

Enhancing Weather Data Reconstruction through Hybrid Methods with Dimensionality Reduction

MURILO MONTANINI BREVE

*This thesis is done under the scope of the double degree agreement between Instituto
Politécnico de Bragança and Universidade Tecnológica Federal do Paraná*

Master's degree in Industrial Engineering

2022-2023

The School of Technology and Management is not responsible for the opinions expressed in this report.

I declare that the work described in this report is of my own authorship
and it is my wish that it be submitted for evaluation.

Murilo Montanini Breve

Acknowledgment

I would like to express my gratitude to my family for their support throughout this thesis journey. My mother, Lara Montanini, has been a constant source of encouragement, while my brother, Matheus Montanini, has provided motivation and belief in my abilities.

I am deeply grateful to my esteemed professors, Balsa, Rufino and Fabricio, for their invaluable help and guidance. Their dedication to my academic progress and mentorship have been crucial to the completion of this thesis.

I also want to thank my friends, Lucas Viveiros, Alexandre Jarozs, and Felipe Gimenez, for their support and friendship. Their companionship has been a great source of encouragement, and I am truly grateful for our shared experiences.

To all those who have contributed to this achievement, thank you for your unwavering support and dedication.

Abstract

Accurate weather analysis and forecasting rely on complete historical data. However, missing weather data often occurs due to sensor failures, data transmission issues, or limited monitoring capabilities. Reconstructing this missing data is crucial for reliable weather analysis. The Analog Ensemble (AnEn) method leverages past weather events and information from nearby stations to reconstruct and forecast data. However, incorporating nearby stations significantly increases computational costs, making the reconstruction process time consuming. To address this challenge, this dissertation integrates AnEn with dimension reduction techniques: Principal Component Analysis (PCA) and Partial Least Squares (PLS). Four hybrid methods—PCAnEn, PLSAnEn, PCClustAnEn, and PLSClustAnEn—are developed to enhance computational performance while maintaining or improving accuracy.

Through four studies using three datasets, this research focuses on reconstructing six variables: wind-related variables, temperature, pressure, and humidity. The hybrid methods improved accuracy compared to the original AnEn. Notably, PLSAnEn achieves the highest reconstruction accuracy, while PLSR exhibits the fastest processing times. Additionally, PLSClustAnEn also proves to be a alternative for data reconstruction. The findings of this research contribute to the portfolio of strategies for addressing missing weather data.

Keywords: Analog Ensemble; Principal Component Analysis; Partial Least Squares; Weather Data Reconstruction; Hindcasting.

Resumo

A análise e a previsão climática beneficiam de dados históricos completos. No entanto, é comum faltarem dados meteorológicos devido a falhas nos sensores, problemas na transmissão de dados ou limitações nas capacidades de monitoramento. A reconstrução desses dados ausentes é crucial para uma análise climática confiável. O método Analog Ensemble (AnEn) utiliza eventos meteorológicos passados e informações de estações próximas para reconstruir e prever dados. No entanto, a incorporação de estações próximas aumenta significativamente os custos computacionais, tornando o processo de reconstrução bastante demorado. Para enfrentar esse desafio, esta dissertação integra o AnEn com técnicas de redução de dimensionalidade: Análise de Componentes Principais (PCA) e Mínimos Quadrados Parciais (PLS). Quatro métodos híbridos - PCAnEn, PLSAnEn, PCClustAnEn e PLSClustAnEn - são desenvolvidos para melhorar o desempenho computacional, mantendo ou aumentando a precisão.

Por meio de quatro estudos utilizando três conjuntos de dados, esta pesquisa concentra-se na reconstrução de variáveis meteorológicas. Os métodos híbridos aprimoraram a precisão em comparação com o AnEn original. Notavelmente, o PLSAnEn alcança a maior precisão de reconstrução, enquanto o PLSR é mais eficiente em termos computacionais. Além disso, o PLSClustAnEn também se mostra uma alternativa eficiente para a reconstrução de dados. Os resultados desta pesquisa contribuem para um portfólio de estratégias de reconstrução de dados meteorológicos.

Palavras-chave: Analog Ensemble; Análise de Componentes Principais; Mínimos Quadrados Parciais; Reconstrução de Dados Meteorológicos; Hindcasting.

Contents

1	Introduction	1
1.1	Objectives	3
1.2	Contributions	4
1.3	Document Structure	4
2	Background	5
2.1	Weather Prediction and Decision-making	5
2.2	The Evolution of Weather Prediction	8
2.2.1	Early Weather Forecasting	8
2.3	AnEn: Analog Ensemble	10
2.4	Literature Review	11
2.4.1	Analog Ensemble and Hybrid Methods	11
2.4.2	Missing Weather Data: A Challenge in Forecasting	12
2.4.3	Addressing Weather Data Challenges	14
3	Combining Principal Component Analysis with Analog Ensemble	19
3.1	Meteorological Dataset	20
3.2	Data Correlation	22
3.2.1	Principal Components Analysis	23
3.2.2	Analogues Ensemble Method	27
3.3	Evaluation Metrics for Reconstruction Accuracy	30
3.4	Reconstruction with AnEn Method on PCs	31

3.5	Conclusions and Future Directions	33
4	PCAnEn: Consolidating the combination of PCA and AnEn	35
4.1	Meteorological Dataset	36
4.1.1	Data Correlation	38
4.2	Dataset Decomposition Using Principal Components	39
4.3	Experiments with the PCAnEn Method	41
4.3.1	Comparing Accuracy	42
4.3.2	Comparing Performance	44
4.4	Key Findings	46
5	PCClustAnEn: Enhancing PCAnEn with K-means Clustering	49
5.1	The ClustAnEn Method	50
5.2	Experimental Evaluation	51
5.3	Main Takeaways	55
6	PLS AnEn-based Methods and Regression Techniques for Meteorological Data Re- construction	57
6.1	Meteorological Data Sets	58
6.2	Reconstruction Methods	61
6.2.1	Principal Component Regression	61
6.2.2	Partial Least Squares Regression	64
6.3	Selecting the Principal Components	66
6.4	Selecting the Latent Variables	69
6.5	Results	73
6.5.1	Reconstruction of Meteorological Variable in one Station	74
6.5.2	Reconstruction of Meteorological Variable in all Stations	77
6.6	Computational Performance	79
6.7	Final Remarks	81
7	Conclusion	83

List of Tables

3.1	Meteorological datasets characterization.	21
3.2	Variable loadings in each PC and variance proportion of each PC.	25
3.3	Variable loadings in each PC and variance proportion of each PC, without WDIR variable.	26
3.4	Valongo variables predicted by Soutelo PCs and by the classical AnEn.	31
3.5	Valongo variables predicted by Edroso PCs and by the classical AnEn.	32
3.6	Valongo variables predicted by Soutelo PCs and by the original AnEn method without the WDIR variable.	32
3.7	Valongo variables predicted by Edroso PCs and by the original AnEn method without the WDIR variable.	33
4.1	Meteorological dataset characterization.	37
4.2	Standard deviation of the <i>PCs</i> generated from the variables WSPD and GST coming from different stations together.	40
4.3	Standard deviation of the <i>PCs</i> generated from the variables PRES and ATMP coming from different stations together.	41
4.4	Accuracy comparison between the PCAnEn and AnEn methods.	43
5.1	RMSE of the reconstruction with different methods.	52
6.1	Caracterization of the Dataset. Available data (Avail.) is represented as a percentage.	59
6.2	Errors of the prediction of PPX station for different principal components. . . .	68

6.3 Errors of the prediction of the PPX station for a different number of latent variables (LVs). 73

6.4 Number of PCs or LVs used by each method. 74

6.5 Mean execution times in seconds across methods, variables, and steps. 79

List of Figures

3.1	Geolocation of the meteorological stations.	20
3.2	Correlation between variables.	22
3.3	Correlation between stations.	23
3.4	Reconstruction of missing meteorological records using the AnEn method with a dependent analog search.	28
3.5	Reconstruction of missing meteorological records using the AnEn method with an independent analog search.	29
4.1	Geolocation of the meteorological stations.	36
4.2	Correlation between variables at each station.	38
4.3	Correlation between stations for each variable.	39
4.4	Comparison of the RMSE for different variables and number of stations.	42
4.5	RMSE of PCAnEn method used in a dependent and independent way.	44
4.6	Processing time for different number of stations and different methods.	45
4.7	Processing time for different number of CPU cores used by PCAnEn.	46
5.1	Reconstruction of missing meteorological records with the ClustAnEn method.	51
5.2	Reconstruction time of WSPD with 16 cores (2 to 6 stations).	52
5.3	Reconstruction time of WSPD with 1 to 16 cores (6 stations).	54
6.1	Geolocation of the selected NDBC meteorological stations.	58
6.2	Correlation between stations for the WSPD variable.	60
6.3	Correlation between stations for the GST variable.	60

6.4	Correlation between stations for the ATMP variable.	61
6.5	Correlation between stations for the PRES variable.	61
6.6	Standard deviation of the first PCs for different predictor variables.	68
6.7	Normalized RMSEP values for different latent variables on a logarithmic scale. The same variable was used in both X and y . The normalization factor used was the difference between the maximum and minimum values [99].	70
6.8	Normalized RMSEP values for different latent variables on a logarithmic scale. All variables were included in X	71
6.9	Q^2 metric for a different number of latent variables when using the same vari- able as predictor and predicted.	72
6.10	Q^2 metric for a different number of latent variables when using all variables as predictor.	72
6.11	Comparison between reconstructed and observed values of the meteorological a) WSPD, b) ATMP and c) PRES variables, from the PPX station, at January 9, 2021, from 10 am to 4pm.	75
6.12	Power spectral densities of the reconstructed/observed time series from the PPX station.	76
6.13	RMSE for the reconstruction of WSPD variable across all available stations. . .	77
6.14	RMSE for the reconstruction of ATMP variable across all available stations. . .	78
6.15	RMSE for the reconstruction of PRES variable across all available stations. . .	78
6.16	CPU time for the reconstruction of WSPD in station PPX in function of the number of cores (excluding loading time).	81

Abbreviations

AF Analog Forecasting. 9

AN ANalog-space. 10

AnEn Analog Ensemble. 2

ANKF ANalog-space Kalman Filter. 10

ANN Artificial Neural Network. 12

ATMP air temperature. 21

CeDRI The Research Centre in Digitalization and Intelligent Robotics. 41

CNN Convolutional Neural Networks. 17

CRPS Continuous Ranked Probability Score Minimisation. 12

CV Cross-validation. 69

ECMWF European Centre for Medium-Range Weather Forecasts. 9

GFS Global Forecast System. 9

GST peak gust speed. 21

HRH High Relative Humidity. 21

HWRF The Hurricane Weather Research and Forecasting. 12

KF Kalman Filter. 10

MLR Multiple Linear Regression. 16

MSEP Mean-Squared Error Predicted. 69

NA Not Available. 21

NDBC The National Data Buoy Center. 36

NN Neural Network. 16

NWP Numerical Weather Predictions. 2

PCA Principal Components Analysis. 3

PCAnEn PCA-based AnEn. 4

PCClustAnEn Cluster-based PCAnEn. 4

PCR Principal Components Regression. 3

PLS Partial Least Squares. 3

PLSAnEn PLS-based AnEn. 4

PLSClustAnEn Cluster-based PLSAnEn. 4

PLSR Partial Least Squares Regression. 3

RI-AnEn Rapid Intensification Analog Ensemble. 12

RMSE Root Mean Square Error. 30

RMSEP Root-Mean-Squared Error Predicted. 70

VAEs Variational Autoencoders. 12

WDIR wind direction. 21

WSPD wind speed. 21

Chapter 1

Introduction

Filling gaps in observed time series is a critical issue in numerous areas of applied sciences that rely on data analysis. Without addressing these gaps, making predictions of any kind becomes challenging or even impossible. This problem is especially prominent in fields where the volume of stored information is rapidly growing, such as weather forecasting. In this context, big data analytics can play a pivotal role in enhancing predictions by uncovering patterns and correlations within the data, as well as reconstructing missing information in areas where limited data is available.

Handling varying degrees of missing data (unrecorded observations) is essential in weather analysis, as different levels of data incompleteness can lead to distinct consequences. For example, when less than 1% of the total data is missing, the impact on analysis is minimal. If the missing data rate lies between 1% and 5%, the dataset is still manageable for analysis. However, when over 5% of the data is missing, it becomes necessary to implement appropriate solutions to utilize the data effectively. Moreover, when missing data rates exceed 15%, prediction models can be significantly and negatively impacted [1].

Even more impactful is when historical data for an area is absent for extended time periods, which can last weeks, months, or even years. This lack of data can be attributed to various factors, such as the unavailability of sensors in the area, inadequate monitoring, technical issues, or even natural disasters [2]. This issue is particularly prevalent in remote, underdeveloped, or geopolitically unstable regions where data collection is often limited or

unreliable.

Places without historical data create obstacles in estimating or analyzing their future, making it especially difficult to identify their renewable energy potential. Without access to this data, it is impractical to evaluate the suitability of these locations for energy generation projects, such as solar or wind farms. As a result, there is a growing demand for methods capable of reconstructing data from limited inputs and locations to perform simulations targeting these areas. Developing and implementing such techniques would not only significantly enhance the understanding of potential advancements in these regions but also promote the establishment of clean energy solutions, contributing to global environmental goals and the well-being of local communities.

Reconstructing missing or absent historical weather data may be addressed through techniques known as *hindcasting*. Hindcasting enables the reconstruction of missing historical data through the use of a generic forecast model to recreate past weather conditions. One of the key techniques employed in meteorological data reconstruction is the Analog Ensemble (AnEn) method [3], [4]. The AnEn method has its origins in the postprocessing of Numerical Weather Predictions (NWP), but it has been applied in several areas, such as forecasting of wind and solar variables [5], [6]. The method involves the identification of analogs in past observations to generate a probability density function for the prediction of future weather conditions.

The rapid increase in stored information, resulting from recent advancements in data collection, creates possibilities for more accurate data reconstruction and improved weather analysis. However, to effectively process this massive amount of data, hybrid AnEn-based methods capable of handling Big Data are essential. This situation involves processing extensive data from multiple sources and diverse variables. In terms of weather analysis, it entails integrating data from various weather stations, each with different variables and data availability, to accurately predict conditions at nearby stations. To accomplish this, enhancing AnEn's computational performance is necessary, as well as refining variable selection, based on their correlation to the target, for greater reconstruction precision.

The AnEn method offers a promising solution for weather data reconstruction, tackling

the issues of missing or absent data in weather forecasting. With the ever-growing volume of data in weather forecasting, it is vital to create hybrid AnEn-based techniques to manage Big Data. Continued research is crucial to boost the computational performance and precision of AnEn-based methods in weather data reconstruction. Exploring the potential of statistical methods, such as Principal Components Analysis (PCA) and Partial Least Squares (PLS), could provide valuable insights and enhancements to this field. Moreover, clustering techniques can be employed to group similar data records, reducing the number of operations required to reconstruct data [7], [8].

1.1 Objectives

This study addresses limitations in the AnEn method by proposing hybrid methods that leverage a large number of predictor variables while also reducing dimensionality without loss of critical information. This approach involves exploring dimension reduction through two techniques: PCA and PLS, with the aim of improving the quality of reconstructions while optimizing computational efficiency. Specifically, PCA reduces the dataset's dimensions by identifying the principal components that capture the maximum variance, while PLS, as a supervised method, extracts latent variables that possess the greatest predictive power.

By merging PCA and PLS with the AnEn approach, our goal is to exploit their strengths for superior data reconstruction in terms of both numerical accuracy and computational efficiency. To better comprehend their advantages in various data scenarios, this study conducts a comparative analysis evaluating the performance of these hybrid AnEn methods against traditional Principal Components Regression (PCR) and Partial Least Squares Regression (PLSR) techniques within the context of a hindcasting problem. The task entails reconstructing missing data at a meteorological station by using information from a collection of predictor stations with diverse geographical locations.

1.2 Contributions

This thesis contributes to the progress of reconstructing missing or absent weather data by combining various techniques. In particular, this study utilizes PCA and PLS for dimensionality reduction of the predictor dataset, alongside regression and analog-based approaches. This research represents the first exploration of applying AnEn methods with PLS dimensionality reduction, giving rise to the PLS-based AnEn (PLSAnEn) method. Furthermore, this study pioneers the integration of PCA, PLS, and AnEn with clustering techniques, introducing three new methods for data reconstruction: Cluster-based PCAnEn (PCClustAnEn), Cluster-based PLSAnEn (PLSclustAnEn) and PCA-based AnEn (PCAnEn).

1.3 Document Structure

This thesis begins with an exploration of weather prediction and its significance in decision-making processes (Chapter 1), followed by a discussion on the evolution of weather forecasting, from early methods to current technologies (Chapter 2). Furthermore, a Literature Review is presented in Section 2.4, which delves into analog ensemble and hybrid methods in weather forecasting, addressing the challenges posed by missing or incomplete weather data, aiming to overcome the limitations and improve the accuracy of predictions.

The core of the thesis is a sequence of studies presented in chronological order, each building on the findings of the previous study. Topics covered include the application of Analog Ensemble and Principal Components Analysis for reconstructing meteorological variables (Chapter 3), the consolidation of PCA and AnEn in a new hybrid method called PCAnEn (Chapter 4), the enhancement of the PCAnEn method with K-means clustering (Chapter 5), and a comparative study that combines AnEn with PCA and Partial Least Squares for meteorological data reconstruction (Chapter 6). The final section of the thesis brings together the main findings and contributions of the research, summarizing the key advancements made throughout the studies and addressing whether the results have tackled the challenges in weather forecasting related to missing or absent data (Chapter 7).

Chapter 2

Background

This chapter provides an examination of the significance of weather predictions in decision-making across various fields. In Section 2.1, the importance of weather predictions in multiple fields is outlined, along with an overview of how these predictions are used in these fields and their influence on decision-making processes. In Section 2.2, a historical and developmental analysis of weather prediction methods and technologies is presented, including examination of early and current methods and technologies used for forecasting.

2.1 Weather Prediction and Decision-making

Weather forecasting is a crucial tool for decision making, providing individuals and organizations with vital information on the potential impacts of weather on their activities. Thanks to the use of sophisticated forecasting models, meteorologists are able to generate precise predictions on the future state of the atmosphere, allowing people to prepare in advance and reduce any risks related to extreme weather events. For instance, farmers can use weather forecasts to plan planting and harvesting times, decide when to fertilise, and even know when to move livestock [9]. Companies may also utilise forecasts to plan routes, schedule flights and ships, and avoid hazardous weather conditions. In addition, weather can influence consumer spending [10] and stock market volatility [11]. In the renewable energy sector, weather predictions are used to calculate a site's energy generation potential prior

to investing large amounts of funding [12]. By providing accurate predictions about future weather conditions, forecasts enable decision-makers to make more informed choices and reduce potential risks.

Wind energy and weather forecasting have become increasingly intertwined as wind energy has grown globally. Wind energy has gained attention due to its potential to reduce dependence on fossil fuels and mitigate greenhouse gas emissions [13]. However, the inherent intermittency and unpredictability of wind variables pose significant challenges in maintaining the safe and stable functioning of power systems [14]. From this, a wind's accurate prediction is crucial for the effective integration of wind energy into power systems. According to the Global Wind Energy Council [15], wind energy has the potential to significantly reduce CO₂ emissions, with an annual reduction of over 1.2 billion tons, equivalent to the consumption of the entire South American continent. This is important because efficient energy production is a key factor in determining a country's economic and social development [16]; thus, access to energy is a crucial task for governments and regions [17]. In light of these considerations, the development and improvement of the prediction of wind-related variables is important for the functioning of wind power systems, as well as for supporting the transition to a sustainable energy future.

Weather predictions also contribute to the planning of agricultural activities. The use of weather forecasts in agriculture is extensive, affecting many aspects of decision-making. In sub-Saharan Africa, the use of seasonal climate forecasting improves knowledge about the atmosphere and oceans; and provides probabilistic forecasts about future climatic conditions [18], which is especially important in underdeveloped regions. In fact, efforts are being made to improve limited access to seasonal weather forecasts in many parts of the world [19]. Furthermore, seasonal climate forecasts can help farmers reduce investment risks by adjusting their agricultural management strategies. For example, forecasting the La Niña or El Niño phases can benefit farmers by allowing them to take advantage of potential favorable conditions [20]. The close relationship between weather and agriculture highlights the importance of the ability to translate meteorological information into agricultural terms, which can help farms around the world to be more effective.

Weather prediction has a wide range of uses beyond just wind energy and agriculture. The use of weather forecasts can significantly improve decision making processes across society, as well as increase economic efficiency, as Thompson pointed out in 1973 [21]. Recent studies have supported this statement by demonstrating the significant impact of forecasts from Numerical Weather Prediction models on various sectors of the economy, such as the better integration of wind energy resources into the electric grid, increased worker output due to more accurate precipitation forecasts, and improved decisions by agricultural producers in preparation for freezing conditions [22]. In Europe, there is a growing need for better access to climate information across economic sectors in order to inform planning and decision-making [23]. To provide more detail, predictions on a 10-year time scale have potential applications in multiple sectors in Europe [23]. As a consequence, efforts to improve access to this information in underdeveloped areas are ongoing and it is crucial to continue to research and develop ways to effectively predict weather to improve decision-making processes and increase economic efficiency.

Weather predictions indeed play a crucial role in decision-making across fields such as agriculture, transportation, consumer spending, stock market volatility and renewable energy. The use of weather forecasts can improve the efficiency of these industries by providing the necessary information to make informed decisions. The development and improvement of weather predictions can also support the integration of renewable energy sources, such as wind energy, into power systems and contribute to a sustainable energy future. Additionally, weather predictions can help farmers in underdeveloped regions make informed decisions and reduce investment risks. Overall, better weather prediction improves economic efficiency and can bring benefits to all sorts of fields. The next section explores the evolution of weather prediction through the advances of numerical weather prediction - which has resulted from a steady accumulation of scientific knowledge and technological advances over many years – and is considered to be one of the greatest areas of progress in physical science [24].

2.2 The Evolution of Weather Prediction

Human weather forecasting has a long history, with evidence of its use dating back to the early days of civilization [25]. In particular, farmers and sailors relied on their observations of the sky and the natural world to make decisions about planting crops and embarking on voyages [26]. Although fishermen were likely to be the earliest practitioners of weather forecasting, their methods were primitive and often inaccurate [26]. Nevertheless, their work laid the foundation for the development of modern, sophisticated weather forecasting techniques.

The following sections presents an overview of the methods for weather prediction that have been developed over time, beginning by discussing NWP, a method of predicting the weather using mathematical models without relying on subjective human observations. NWP has a long history, with early efforts dating back to the 1920s. Next, the Analog-based methods are described, which involve comparing current weather conditions to those of the past, with the belief that similar weather patterns will recur. These Analog-based methods were developed first to make a NWP prediction probabilistic, and more recently to reconstruct data of a station by means of neighboring stations. It is also presented an overview of other available methods of weather prediction in the current state-of-the-art. The study of early predictions' attempts may guide future solutions in weather prediction.

2.2.1 Early Weather Forecasting

In 1922, Richardson proposed a mathematical method for forecasting the weather called NWP (*Numerical Weather Prediction*) [27]. Richardson used the time integration of basic equations of fluid mechanics to simulate and predict atmospheric circulation. His method used data from observations of the current state of the atmosphere, including temperature, humidity, wind speed and direction, and other variables, to forecast future weather conditions. Richardson's first attempt to predict using the numerical method for prediction did not go well, though. The results were incorrect, with a surface pressure change that was about 100 times larger than the actual change, leading to his method being initially called

“Richardson’s dream”.

Richardson’s idea was not successful at his time because he had neither enough meteorological data for initial condition nor a high-speed computer for numerical calculation [28]. Richardson attributed the failure partly to inadequate upper wind data, but later recognised that the inability to use observed winds disturbed the calculation of pressure changes, a principle identified by Margules [29]. Because of its failures, NWP had no influence on practical weather forecasting until 1950.

The success of the numerical weather prediction came later, after 1950, when the first electronic computer was developed, NWP forecasted a 24-hour pressure change at the 500 hPa level (in the middle of the troposphere). The later success of NWP, due to both the development of electronic computers and advances in dynamic meteorology, rapidly improved the prediction of weather variables. Computers have continued to support the advancement of numerical weather prediction, and today, computer simulations of the global atmospheric circulation are used to make forecasts for the next two days, week, month, and even 50 years. Examples of current NWP models include the Global Forecast System (GFS) [30], developed by the National Weather Service in the United States; also the European Centre for Medium-Range Weather Forecasts (ECMWF) [31] model, and the UK Met Office’s Unified Model [32]. Richardson’s dream of accurate numerical weather prediction has finally been realized [24].

In contrast to NWP, Analog-based methods rely on the principle that past atmospheric states can predict future ones, as stated by Lorenz in 1969 [33]. His method, called Analog Forecasting (AF), introduced the concept of predicting weather by finding a past weather event similar to the current situation as a guide to forecast. But at the time, after analyzing the results of his efforts based on this assumption, Lorenz dismissed this hypothesis, believing it was impossible to find analogous states of the atmosphere, since it would take many years before two similar states of the atmosphere could be found.

The use of analogues for weather forecasting was largely discarded until the H. M. Van den Dool’s study in 1989 [34]. He showed that, in a limited area, the AF model can effectively make 12-hour forecasts at target points by searching for analogues within a 900 km radius related to the target (the predicted one). The idea is that by finding several analogs in a relative

small area, these analogs can be used to predict accurately, since they are likely to correlated with each other. As well, different historical analogues can be used at different target points based on their proximity and reliability. Hence, Van den Dool’s study demonstrated that Analog Forecasting is possible if the area of possible analogues is reduced.

After Analog Forecasting (AF) was proved to be effective, the development of new techniques, inspired by Lorenz’s original concept, has been a significant advancement in the field of weather prediction. These methods often aim to improve the efficiency and precision of AF by handling large data sets [35]. Despite the computational demands of working with large data sets, AF remains a valuable tool for weather forecasting, particularly for short-term forecasting scenarios. It is able to extract useful information from historical data and apply it to predictions, making it a data-driven method [36], which relies on large and reliable datasets. As such, many improvements made to AF were related to handling a large amount of data.

Weather forecasting has evolved over time, from ancient humans using observations of the sky and the natural world to make predictions. Numerical Weather Prediction was not successful at first but with the advancements in technology and meteorology, NWP improved and is now widely used for weather forecasting. Analog-based methods, which involve comparing current weather conditions to past patterns, were also developed and proved to be effective in weather forecasting. Today, NWP and AF are widely used for weather prediction and both have played an important role in improving weather forecasts and understanding the atmosphere.

2.3 AnEn: Analog Ensemble

Building on the Analog methods proposed by Lorenz and Van den Dool [37], [38], Monache introduced two analog ensemble methods in 2011: ANalog-space Kalman Filter (ANKF) and ANalog-space (AN) [3]. These methods use similar past events to create probabilistic forecasts of NWP predictions and reduce systematic and random errors. ANKF combines the Kalman Filter (KF) method with an analog approach to address the limitations of the KF

method in cases of sudden changes in weather conditions [39]. AN, on the other hand, is based on taking weighted averages of observations when verifying past predictions for current forecast features. To assess the effectiveness of ANKF and AN at reducing systematic and random errors compared to existing correction techniques such as KF or 7-day running mean corrections, both methods were tested using wind speed prediction data from 400 surface stations over the western United States for a 6-month period. The results indicated that AN was consistently the best, showing significant improvements compared to the other methods.

The Analog (AN) method proposed by Monache in his 2013 study [4] was so successful that he introduced a new term for it, Analog Ensemble (AnEn). He demonstrated the effectiveness of the AnEn method in using past observations as analogs to estimate the probability distribution of the atmosphere's future states. The AnEn method was shown to have equal or superior accuracy compared to other ensemble systems and also had better computational performance. Another study [40] extended the prediction to six days ahead and independently at each location on a two-dimensional grid, using only 1/6th of the computational resources of traditional methods. The AnEn method has also been used successfully to generate probabilistic wind speed forecasts at 80 m height for large-scale projects [41], and to provide high-quality long-term estimation of wind speed time series and frequency distribution at target sites, as well as reliable uncertainties based on physical processes [42]. Overall, the AnEn method has been proven to be an effective and efficient approach for generating probabilistic forecasts from a deterministic forecast.

2.4 Literature Review

2.4.1 Analog Ensemble and Hybrid Methods

The Analog Ensemble (AnEn) method has become a widely used technique in weather forecasting to extract useful information from historical data to make predictions. Studies have shown that AnEn is effective in predicting rare and high-risk wind situations [43], and when

combined with other methods such as multimodel intensity consensus forecast, it can improve accuracy [44]. The Hurricane Weather Research and Forecasting (HWRF) model with a variant of AnEn called the Rapid Intensification Analog Ensemble (RI-AnEn) was found to be more accurate than other models in predicting rapid intensification, which is a critical factor in forecasting hurricanes [44]. Additionally, AnEn has been found to be effective in predicting meteorological variables such as wind and solar power [45], [46]; also precipitation [47] and air quality [48].

Recently, efforts have been made to improve the AnEn method by combining it with other techniques. For example, Variational Autoencoders (VAEs) have been used in combination with AnEn to improve performance [49]. Additionally, weighting predictor combinations with Continuous Ranked Probability Score Minimisation (CRPS) and PCA improved performance by up to 20% while also providing reliable estimation of forecast uncertainty [50]. Furthermore, a combination of Artificial Neural Network (ANN) and AnEn was proposed to generate 72-hour deterministic and probabilistic forecasts of power generated by photovoltaic plants [51], and the results showed that the combination of AnEn and ANN yielded the best results, indicating its potential for large-scale computation tasks. The Hybrid Ensemble approach, which combines the strengths of the AnEn and NWP ensembles, has also been found to have potential in forecasting tasks [52]. The AnEn method has been demonstrated to be an effective tool for weather forecasting, and its potential applications have been demonstrated across different combinations of methods.

2.4.2 Missing Weather Data: A Challenge in Forecasting

Despite the use of advanced models, weather forecasting is still not perfect and errors can occur [53]. Studies have found that errors in weather predictions can be significant, and weather forecasting is a challenging task due to its unpredictable nature [54]. Weather data are currently collected through various means, such as satellites, radar, and ground stations, and are then analysed through statistical, machine learning, or numerical techniques to make predictions about future weather conditions. Due to general problems, the gathered

data is often incomplete or lacking parts, making accurate predictions even more difficult. Researchers are continuously working to improve the accuracy of weather forecasts, but it is still an ongoing challenge.

The missing values are inevitable in the weather databases. They are caused by the temporary absence of observers, equipment failure, infrequent calibration of sensors, among others. In fact, the estimation of missing values is the first step in climatological and environmental studies [55]. The literature on weather forecasting has highlighted the negative impact of missing data on forecast precision [56], [57], which is particularly concerning as it has been shown that the estimation error increases with the number of consecutive missing values. On top of that, historical datasets may be affected by other problems such as systematic data quality issues, further compounding the negative impact of missing data. To mitigate the effects of missing data, solutions to reconstruct and fill gaps in data are essential part of weather forecasting.

Weather prediction is increasingly dependent on technology to process large amounts of data from various sources. The amount of stored information is growing four times faster than the world economy, while the processing power of computers is growing nine times faster [58]. In view of this, big data analytics can help improve predictions by uncovering patterns and correlations in the data [59]. This can also help to reconstruct missing data in areas where there is limited information. Conversely, this growth in data also means that the amount of missing data is increasing, making accurate reconstruction a crucial task. To handle this challenge, forecasting methods must be able to handle large amounts of data, multiple sources of data – and a wide variety of meteorological variables. This requires advanced methodologies that can adapt to the unique characteristics of big data in weather forecasting.

Despite the abundance of weather data available, there are still many areas without historical record. These locations, which may be remote or under-developed, have the potential to be significant generators of renewable energy; but, without historical weather data, it is difficult to accurately predict the potential for energy generation in these areas. Therefore, there is a growing need for methods that can generate weather data from limited inputs

and location. This is particularly important for locations without detailed historical weather records. The ability to generate weather data from limited inputs would allow simulations of environmentally driven systems to be run at these locations. This would greatly improve the understanding of the potential for renewable energy generation and would facilitate the development of sustainable energy systems in these regions [60].

Therefore, the field of weather prediction is currently facing two major challenges:

- Missing or absent weather data
- Handling large volumes of data

2.4.3 Addressing Weather Data Challenges

Having a complete and large data set is critical to making accurate weather predictions; yet, weather data are often incomplete or missing, which presents a significant challenge for forecasting. This section explores various methods for addressing this challenge, including techniques such as AnEn, Multiple Imputation, and Neural Network methods, as well as Machine Learning techniques. AnEn, in particular, has been shown to be effective in both weather reconstruction and forecasting. Despite its effectiveness, there are still limitations in the AnEn-based literature that need to be addressed, such as a limited focus on reconstructing more recent weather data, a scarcity of techniques that involve more than two neighboring stations, and a lack of exploration on combining different techniques such as dimension reduction to improve computational performance. By understanding the process of reconstructing weather data, it is possible to identify problems and solutions based on existing methods and ideas, and ultimately improve the accuracy of weather predictions.

2.4.3.1 Reconstructing Incomplete Weather Data

The first challenge of handling missing or absent data should be addressed through weather data reconstruction techniques. Also known as Hindcasting, Weather Data Reconstruction is

a technique used to recreate past weather conditions by applying a forecast model on a historical starting point. This approach is adopted primarily for the validation of forecast models by comparing their output with past observations. Hindcasting also serves as a means of reconstructing missing historical data, known as non-recorded observations, through the use of a generic forecast model. Besides rebuilding data, research in the field of hindcasting also aims to improve various aspects of meteorology, such as downscaling and forecasting methods. One of the key techniques employed in meteorological data reconstruction is the AnEn method [3], [4], which was presented in Section 2.3. Although the original AnEn was first proposed to make a NWP prediction probabilistic, other Analog-based methods have also aimed to reconstruct data using another way to take advantage of AnEn [7], [61]–[63]. In other words, instead of giving a probabilistic prediction, AnEn can recreate data by utilising the analogs' found values from neighbouring stations or a related variable from the same station.

To gain a comprehensive understanding of the literature on Hindcasting, it was conducted a search for papers on data reconstruction with a focus on AnEn-related methods. The goal was to identify gaps and issues in current literature and find potential solutions by examining existing approaches that have been combined (or not) to improve reconstruction. It is important to note that, in contrast to the abundance of forecasting models available, techniques that aim to reconstruct weather data are relatively scarce. Many of these techniques focus on rebuilding historical data, particularly data from previous centuries. Although understanding weather patterns of the past can be valuable for a variety of applications, such as climate change research, it may not be as useful for forecasting current weather patterns. To ensure accuracy in forecasting, it is essential to have accurate data from the recent past. Therefore, studies that focus on reconstructing weather data from more recent periods are of particular significance.

The literature on methods for weather data reconstruction is extensive, featuring a wide range of approaches such as neural networks, multiple imputation, data assimilation, evolutionary algorithms, multiple linear regression, and more. Among these methods, AnEn stands out due to its adaptability and potential applicability in various contexts [7]. In this

study, AnEn was successfully adapted to accurately reconstruct pressure, temperature, wind speed, and wind direction variables, highlighting its relevance in renewable energy management.

In comparison, Schneider [64] showcased the effectiveness of the regularized EM algorithm for estimating missing values in climate datasets, while Barrios [65] found that Neural Network (NN), Multiple Linear Regression (MLR), and the modified inverse distance weighting performed best for monthly precipitation records reconstruction. Carro-Calvo [66] utilized an evolutionary algorithm for wind speed reconstruction. Chen [67] employed stochastic models for reconstructing precipitation, radiation, and temperature, and Torade [68] used data assimilation for cloud cover, air pressure, precipitation, wind speed, and humidity reconstruction. Rao [69] and Mahjoobi [70] both applied NNs for storm wave and wave and wind variables reconstruction, respectively, while Malekmohamadi [71] combined numerical wave model with artificial neural networks for the same purpose. AnEn has also been effectively employed for weather data reconstruction in [72], with possibilities for further refinement.

Lastly, other methodologies, including spatial regression [73], multilayer perceptron [74], and additional techniques, may be successful in this area. However, it is worth noting that AnEn's adaptability and successful application in various studies underscore its potential as a leading technique in weather data reconstruction.

2.4.3.2 Tackling Large Data Sets in Reconstruction

The second challenge is about effectively managing large data sets. One approach is to use data reduction techniques, as discussed in [75]. Another strategy is to employ clustering techniques, as recommended by [76]. These techniques are useful for grouping similar data points and reducing the computational burden of data analysis. Specific methods that have been exploited include the k-means [7] and the kd-tree algorithm [8]. These methods aim to group data points with similar characteristics, making data processing and analysis more efficient, such as searching for analogous events. Similarly, dimension reduction techniques also aim to improve computational performance and enable more information to be added

without compromising method performance. This is achieved by identifying the part of the data that can best represent it. These methods, such as principal component analysis, are particularly useful for data-driven techniques. In summary, there are several potential approaches to address the challenges of large data sets and computational performance.

The AnEn approach has been successfully combined with k-means clustering technique to improve algorithm efficiency, as demonstrated in studies such as [7], [62], [63]. One study in special [62] compared the proposed k-means metric to search for analogs with metrics based on fuzzy C-means clustering, cosine, and normalization, and found that the k-means clustering metric performed better in terms of prediction accuracy. The k-means approach groups similar weather data points, reducing the need to compare entire time series and reducing the computational burden. The studies reconstructed data using a database of predictor stations close to the station with missing data, and also exploited the use of multiple predictor stations or variables, showing that the use of two predictor stations improved hindcasting performance, leading up to 16% lower error, depending on the correlation between the predictor stations. Nevertheless, the reconstruction with more nearby stations was not pursued, which is a gap in the research that could be further exploited. One potential explanation for this is that utilizing more time series increases the amount of data to be processed, slowing down processing time. To address this, dimension reduction techniques could be employed to improve the computational performance of AnEn as well as its combination with the k-means clustering.

In view of the factors discussed above, it is clear that the Analogue Ensemble (AnEn) approach is a highly promising method for addressing the challenge of data reconstruction. To provide further evidence, a number of studies have also been conducted to compare the effectiveness of the AnEn approach with other methods, such as Convolutional Neural Networks (CNN). For example, studies such as [77] and [78] have found that the AnEn approach can improve prediction accuracy equally or even outperform CNNs when used as post-processing methods for a Weather Research and Forecasting model. To supplement, the implementation of the AnEn approach is relatively straightforward compared to other machine learning methods [79]. These findings provide a strong basis for the continued use

of AnEn and k-means clustering in future research and weather prediction and reconstruction applications.

While AnEn-based methods have been widely researched, there are still some unexplored areas in the literature that need to be addressed, such as:

- Using more than two neighbouring stations for reconstructing data from a single station.
- Combining AnEn with dimension reduction techniques, to improve the computational performance of AnEn in data reconstruction.

To sum it up, weather data reconstruction is a vital process for tackling the challenges of missing or incomplete data in weather prediction. Various techniques such as AnEn, multiple imputation, and neural networks have been shown to be successful methods for reconstructing weather data. Notably, the AnEn technique has been found to be a highly promising method for data reconstruction, with studies demonstrating that it can either improve equally or perform better than other methods like convolutional neural networks. However, there are still areas of research that need further exploration in the field of AnEn methods, such as reconstructing weather data from recent periods, utilizing techniques that involve more than two neighboring stations, and investigating the combination of different techniques to enhance the computational performance of AnEn. These research gaps highlight potential opportunities for further studies in the field of weather data reconstruction.

Chapter 3

Combining Principal Component Analysis with Analog Ensemble

Addressing the research gaps identified in Section 2.4.2, this chapter presents a study that combines a modified version of the Analog Ensemble (AnEn) method with Principal Component Analysis (PCA) for data reconstruction and dimensionality reduction of meteorological datasets. This approach was chosen to enable the integration of data from multiple weather stations in close proximity while maintaining computational performance. This is particularly relevant because while AnEn benefits from large training datasets, the data processing required to identify analogues can sometimes make the computational cost prohibitive [80].

In this chapter, the hindcasting technique is applied to reconstruct five meteorological variables at a specific station using data from the same variables recorded at neighboring stations. By combining AnEn and PCA, the aim is to decrease the dimensionality of the meteorological variables and retain the most relevant information, followed by AnEn to identify analogs in fewer dimensions and generate reconstructed values for an incomplete dataset. The performance of this method is evaluated using a dataset of meteorological variables collected at three villages in the Northern region of Portugal over a seven-year period.

Section 3.1 presents the dataset used in the chapter. In Section 3.2 the correlations between the different meteorological variables and stations are studied. Section 3.2.1 introduces the PCA technique and Section 3.2.2 describes the AnEn method. Section 3.4 is dedicated to the application of AnEn method to the principal components, in the reconstruction of meteorological variables of a single station. Some final considerations are presented in section 3.5.

3.1 Meteorological Dataset

The data used in this chapter was obtained from meteorological stations belonging to the the Polytechnic Institute of Bragança (IPB). The oldest records started in 1999. The stations are located in the northeast region of Portugal, near the villages of Edroso (*latitude* : 41.912778; *longitude* : -7.152833), Soutelo (*latitude* : 41.92116; *longitude* : -6.808528) and Valongo (*latitude* : 41.923056; *longitude* : -6.950833). Notably, these locations are near the border with Spain (refer to Figure 3.1 for the map of the station locations).

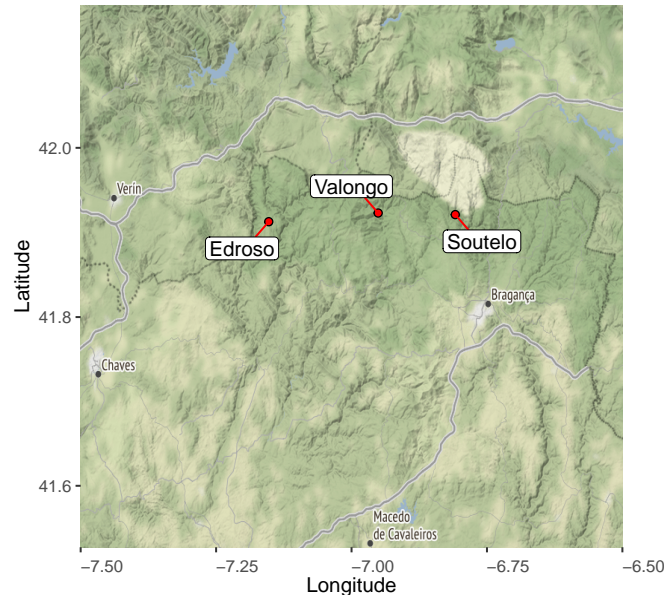


Figure 3.1: Geolocation of the meteorological stations.

The meteorological variables available in this dataset, measured at each station, are:

High Relative Humidity (HRH) [%]; air temperature (ATMP) [°C]; wind speed (WSPD) [m/s] averaged over a 30 minutes period; peak gust speed (GST) [m/s] during the same 30 minutes period; wind direction (WDIR) [°] from the North in a clockwise direction.

All three stations have data available from 2000 to 2007, with a sampling frequency of 30 minutes. However, one often finds time windows where there is no record, and also different time series intervals. To overcome these problems, an interpolation (nearest-neighbour or linear) was performed on the data, in order to standardize the data intervals to every 30 minutes. This interpolation process is limited to 4 missing values, since larger intervals could distort the data in an excessive manner.

Despite the interpolation process, a considerable amount of values Not Available (NA) persisted. Table 3.1 displays the number of missing values, the minimum and maximum values for each variable, and the percentage of data availability. The Valongo station has most data, while Soutelo data is the sparsest.

Table 3.1: Meteorological datasets characterization.

Station	Variable	Min	Max	#NA	Availability (%)
Soutelo	WSPD	0	17	63915	47.93
	GST	0	30.3	49114	59.98
	WDIR	0,00	337.5	66628	45.71
	ATMP	-10.3	33.6	63915	47.93
	HRH	10.0	100	65353	46.75
Edroso	WSPD	0	113.5	27330	77.73
	GST	0	114.0	27330	77.73
	WDIR	0	337.5	56898	77.73
	ATMP	-32.10	33.2	28411	76.85
	HRH	9	100	28410	76.86
Valongo	WSPD	0	10.3	18279	85.11
	GST	0	19.2	18279	85.11
	WDIR	0	337.5	20253	83.50
	ATMP	-9.60	40.1	18307	85.08
	HRH	-11.0	100	18307	85.08

For the reconstruction experiments, the meteorological stations of Soutelo and Edroso are used as predictor stations, while Valongo is the predicted station. The data was separated into two datasets. The first has data for a training period, from the beginning of 2000

to the end of 2006. The second contains data from the year of 2007, for a prediction (reconstruction) period.

3.2 Data Correlation

Consider the original multivariable historical dataset represented by the matrix \mathbf{X}

$$\mathbf{X} = [\mathbf{x}_1 \mathbf{x}_2 \cdots \mathbf{x}_q] \in \mathbb{R}^{n \times q}, \quad (3.1)$$

Then, the matrix \mathbf{X} is used to obtain the correlation matrix, given by

$$\mathbf{C} = \frac{1}{n} \mathbf{X}^T \mathbf{X} \quad (3.2)$$

where each (i, j) -entry of the matrix C is the correlation between the meteorological variables \mathbf{x}_i and \mathbf{x}_j .

Figure 3.2 shows the correlation between the five variables, for each station. It can be observed that: WSPD and GST were the most correlated variables, HRH and ATMP showed inverse correlation almost equally in all stations, at Valongo station, WSPD and GST presented a low correlation with WDIR and ATMP; also, HRH was somewhat inversely correlated with WSPD and GST.

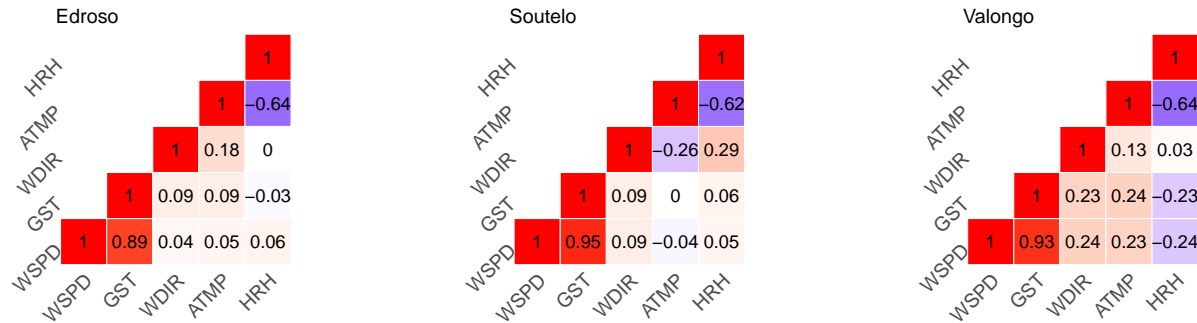


Figure 3.2: Correlation between variables.

Similarly, Figure 3.3 presents the correlations between stations, for all the five variables,

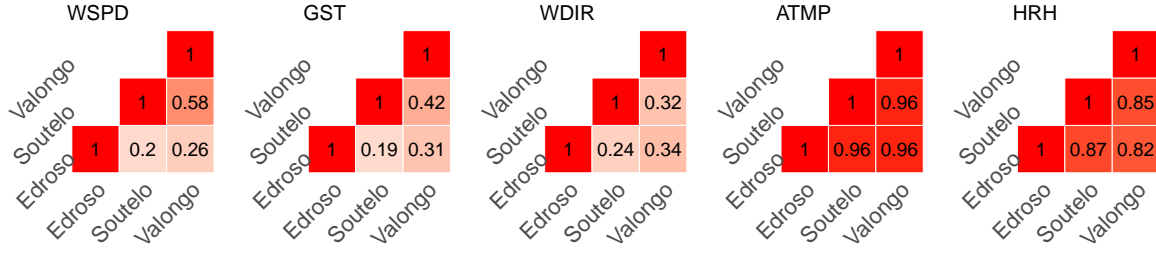


Figure 3.3: Correlation between stations.

the three stations showed high correlation in the ATMP and HRH. Between Valongo and Soutelo, WSPD and GST showed a moderate correlation and a slightly lesser correlation from Edroso and Valongo. It is worth noting that all the variables presented some degree of correlation.

3.2.1 Principal Components Analysis

The Principal Components Analysis (PCA) technique enables to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the information of variation present in the data set. This is achieved by transforming to a new set of uncorrelated variables, called the principal components (PCs), which are ordered so that the first few contain most of the variation information present in original variables data set [81]. The application of PCA to the dimension reduction of the predictor variables is next briefly described.

The original data set of predictor variables may be represented by the matrix

$$\mathbf{X} = [\mathbf{x}_1 \mathbf{x}_2 \cdots \mathbf{x}_q] \in \mathbb{R}^{n \times q}, \quad (3.3)$$

where predictor variables are represented by the q column vectors \mathbf{x}_j , with $j = 1, \dots, q$, each one with n records of the value of a given meteorological variable.

To identify the dimensions along which the data are most dispersed, i.e., the dimensions that best differentiate the predictor data set, it is necessary to compute the principal components (PCs) vectors. Such can be achieved by the thin singular value decomposition of the

predictor matrix \mathbf{X} , given by

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T, \quad (3.4)$$

where the columns of the matrix $\mathbf{U} \in \mathbb{R}^{n \times q}$ contains the left singular vectors, the diagonal matrix $\Sigma \in \mathbb{R}^{q \times q}$ contains the singular values σ_i , with $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_q \geq 0$, and the matrix $\mathbf{V} \in \mathbb{R}^{q \times q}$ contains the right singular vectors \mathbf{v}_j , with $j = 1, \dots, q$, that are the *principal components directions* of \mathbf{X} (for details see [82]). The matrices of the left and right singular vectors are orthonormal, i.e., $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$, where \mathbf{I} is the identity matrix.

The vectors

$$\mathbf{z}_j = \mathbf{X}\mathbf{v}_j, \quad \text{with } j = 1, \dots, q, \quad (3.5)$$

are the principal components (PCs) of the original data set and define new variables that will be used instead of the original predictor variables. The first principal component, \mathbf{z}_1 , has the largest sample variance, equal to σ_1^2/n , amongst all normalized linear combinations of the columns of \mathbf{X} [82]. The second principal component, given by $\mathbf{z}_2 = \mathbf{X}\mathbf{v}_2$ is the new variable with the second largest variance (σ_2^2/n). Likewise, the remaining principal components define new variables with decreasing variances.

The new variables \mathbf{z}_j are linear combinations of the columns of \mathbf{X} , i.e., the original predictor variables $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_q$, being given by

$$\mathbf{z}_j = v_{1j}\mathbf{x}_1 + v_{2j}\mathbf{x}_2 + \dots + v_{qj}\mathbf{x}_q, \quad \text{with } j = 1, \dots, q, \quad (3.6)$$

where the coefficients v_{ij} , with $i = 1, 2, \dots, q$, designated as *loadings*, are the elements of the vector \mathbf{v}_j . The magnitude of a coefficient is related to the relative importance of the corresponding original variable in the principal component.

The substitution criterion of the original predictor variables, $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_q$, by p PCs, $\mathbf{z}_1, \dots, \mathbf{z}_p$, with $p < q$, in the AnEn or ClustAnEn methods, must take into account the influence of the new variables in the original data set. This influence is directly proportional to the respective variances which are given by σ_i^2/n , with $i = 1, 2, \dots, q$. It is expected that the first few principal components, corresponding to the largest singular values, account for a

large proportion of the total variance, being all that is needed to describe the original data set [83]. Therefore, one of the possible criteria that can be used to choose how many PCs should be used, is the magnitude of the respective singular values. If the original variables are previously scaled, by dividing each variable by the respective standard deviation, each of them will have a standard deviation equal to one. If a PC has a standard deviation greater than one, it means that it contains more information than any of the original variables and, as such, should be chosen to represent the original data set.

The substitution criterion of the original variables by a few of the new variables must take into account the influence of the new variable on the variance of the original data. This influence is directly related to singular values. Usually, the first few principal components, corresponding to the largest singular values, account for a large proportion of the total variance, so they are all that is needed for future analyses [83].

A decomposition into principal components (*PCs*) of the five original meteorological variables was performed. Tables 3.2 and 3.3 show the loadings of each *PC* in the case of the stations of Soutelo and Edroso. The Valongo station did not participate in this analysis because it was used only as the predicted station. Additionally, it is also included the proportion between the variance of each *PC* and the sum of the variances over all the *PCs*.

Table 3.2: Variable loadings in each PC and variance proportion of each PC.

Station	Variable	PC_1	PC_2	PC_3	PC_4	PC_5
Soutelo	WSPD	0.616	-0.341	0.058	-0.067	0.704
	GST	0.611	-0.352	0.047	0.034	-0.706
	WDIR	0.267	0.338	-0.901	-0.037	0.001
	ATMP	-0.278	-0.574	-0.326	0.694	0.058
	HRH	0.313	0.561	0.274	0.714	0.043
	Variance Proportion	0.404	0.350	0.161	0.075	0.010
Edroso	WSPD	-0.668	0.221	-0.089	-0.061	-0.701
	GST	-0.684	0.166	-0.057	0.120	0.699
	WDIR	-0.137	-0.144	0.962	0.178	-0.053
	ATMP	-0.218	-0.670	0.002	-0.707	0.058
	HRH	0.132	0.675	0.251	-0.671	0.114
	Variance Proportion	0.385	0.330	0.199	0.067	0.020

Table 3.2 presents the coefficients that multiply each variable into the principal components. It shows that for both stations, PC_1 mainly contains the effects of the variables WSPD, GST, and accounted for 40.4% of all the variance. These are wind-related variables, and therefore more correlated, in according with Figure 3.2. The PC_2 mainly reflects the effect of the ATMP and HRH, the second highest correlation between variables. Note also that the PC_3 is essentially dominated by WDIR. Regarding the variance proportion, the PC_2 and PC_3 represent 35.0% and 16.1% of the total variance, respectively.

It should be mentioned that for Soutelo, PC_1 , PC_2 and PC_3 accounted for 91.5% of all variances and, for Edroso, the three PCs accounted for 91.4%. This shows that in addition to presenting the same PC decomposition pattern, both stations also showed almost the same proportion of variance in the first three components.

Table 3.3: Variable loadings in each PC and variance proportion of each PC, without WDIR variable.

Station	Variable	PC_1	PC_2	PC_3	PC_4
Soutelo	WSPD	0.677	-0.201	0.060	-0.705
	GST	0.674	-0.213	-0.034	0.706
	ATMP	-0.195	-0.681	-0.703	-0.053
	HRH	0.219	0.671	-0.706	-0.042
	Variance Proportion	0.496	0.400	0.092	0.012
Edroso	WSPD	-0.562	0.439	-0.069	-0.698
	GST	-0.582	0.391	0.023	0.712
	ATMP	-0.401	-0.591	-0.699	0.020
	HRH	0.429	0.551	-0.711	0.072
	Variance Proportion	0.568	0.331	0.088	0.013

Table 3.3, similarly to Table 3.2, shows the contribution of each variable in each principal component, but without WDIR in the analysis. It can be observed that, for Edroso, the variable loadings are more evenly distributed for PC_1 and PC_2 . However, at both stations, the same pattern observed with WDIR repeats in the first two components. That is, the predominance of wind-related variables in PC_1 and temperature-related (ATMP and HRH) in PC_2 .

As expected, since there are only 4 variables in the analysis, the first three components represent 98.8% and 98.7% of the proportion of variance, respectively for Soutelo and Edroso.

This shows that by decreasing the number of variables in the analysis, it is possible to keep more information in the first principal components and, at the same time, have less information overall.

3.2.2 Analogues Ensemble Method

Besides providing probabilistic forecasts, the AnEn method offers the additional capability of reconstructing missing meteorological records within a time series. This data reconstruction process relies on one or more predictor time-series that exhibit a certain degree of correlation with the incomplete series to be reconstructed or predicted.

A practical application scenario consists of reconstructing data from a meteorological station using data from neighbouring stations. In this context, several time series are used as predictors, being represented by the column vectors

$$\mathbf{x}_j \in \mathbb{R}^n, \quad \text{with } j = 1, 2, \dots, q, \quad (3.7)$$

each one containing n records of the values of certain meteorological variables. For simplicity, these vectors are often referred to as predictor variables.

The predictor variables can be used in a *dependent* or *independent* way. In the dependent variant, the analogs selected in different predictor variables must be concomitant (overlapping) in time. In the independent version such is not mandatory.

In Figure 3.4 the dependent version of the AnEn method is illustrated with q predictor variables. The historical data is complete in the predictor variables and incomplete in the reconstructed/predicted one ($\mathbf{y} \in \mathbb{R}^m$). The period of missing records is denoted as the *reconstruction period*, but, often, it is also designated as *prediction period*. This designation originates from the application of the AnEn method to the post-processing of meteorological forecasts, in which the predictor series contains the history of forecasts. In this study, the reconstruction period corresponds to the part of the time-series in which the records are reconstructed (or, by analogy, predicted). The period for which all series contain full data is

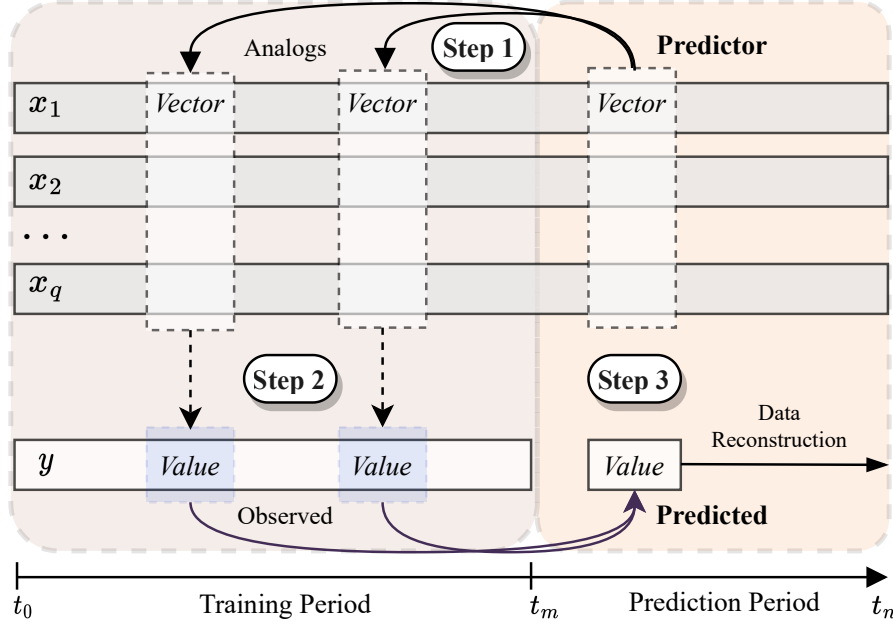


Figure 3.4: Reconstruction of missing meteorological records using the AnEn method with a dependent analog search.

known as the training period. The longer the training period (compared to the prediction period), the better the AnEn method is expected to perform (the more comparison data exists, the more likely it will be to find meteorological conditions similar to those sought).

As depicted in Figure 3.4, firstly (step 1), a certain number of analogs are selected in the training period of the predictor variables, due to being the past observations most similar to the predictor record at instant $t \in \{t_m, \dots, t_n\}$. All compared records are vectors of $2k + 1$ elements, where each element is the value of a predictor at successive $2k + 1$ instants of the same time window, and $k > 0$ is an integer that represents the width of each half-window (into the past, and into the future) around the central instant of the time window. In this work, $k = 1$ and the resulting time window corresponds to one hour, as the time series had a sampling period of 30 minutes. Additionally, the training period ranges from 2000 to the end of 2006 and the prediction period is 2007.

At the end of step 1, a predefined number of analogs was selected - an analog ensemble. The choice of these analogs is made according to the resulting value of a metric that enables to compare the vectors of records (see [62]). Note that comparing vectors, instead of single

values, accounts for the evolutionary trend of the meteorological variable around the central instant of the time window, allowing for the selection of analogs to take into consideration weather patterns (instead of single isolated values).

In step 2, the analogs are mapped onto observations of the predicted station. This mapping is done only for the central time of each analog time window, i.e., for each analog vector a single observed value is selected in the training period.

Finally, in step 3, the observed values selected are used to predict (reconstruct) the missing values in the predicted variable y , through its average (weighted or not). When this target value is actually available as real observational data (as it happens in this study), it becomes possible to compute the error of the reconstruction/prediction and, consequentially, to validate the method.

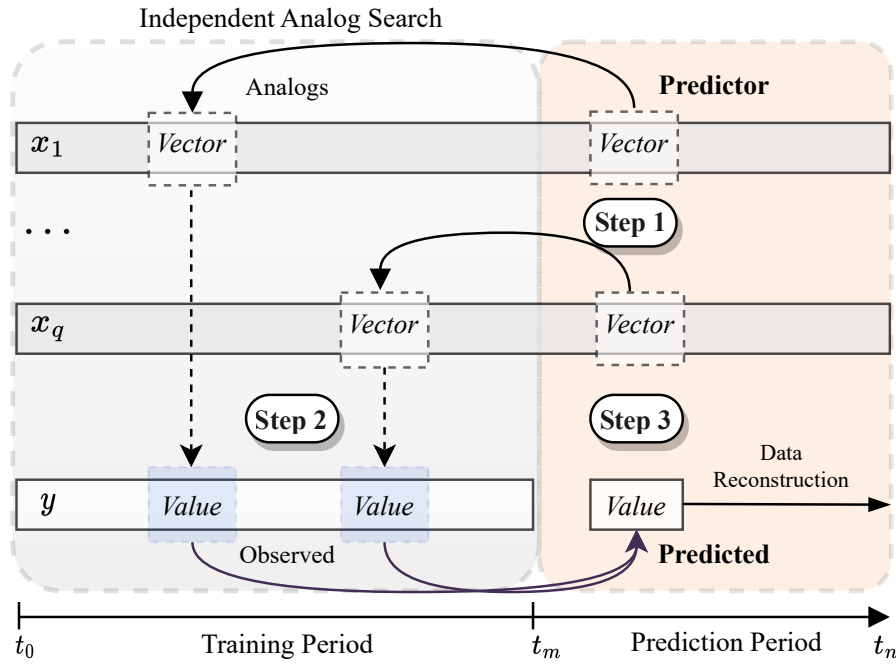


Figure 3.5: Reconstruction of missing meteorological records using the AnEn method with an independent analog search.

To provide clarity on the independent AnEn method, Figure 3.5 illustrates the independent method of analogs definition during the back search process. The independent AnEn method searches for analogs individually across predictors (x_n). This results in an increase in

the calculations required for searching all potential analogs, proportional to the number of predictors n . The remaining steps in the process are consistent with those in the dependent AnEn version.

3.3 Evaluation Metrics for Reconstruction Accuracy

Root Mean Square Error (RMSE) [84] is a widely used metric to assess the accuracy of a prediction. It represents the square root of the average squared differences between the predicted and observed values. The smaller the RMSE value, the better the forecast model's accuracy. RMSE is given by

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (f_i - o_i)^2} \quad (3.8)$$

Bias is another measure used to evaluate accuracy [84]. It quantifies the average deviation between the predicted and observed values. A positive Bias value indicates that the model overestimates the observations, while a negative Bias value signifies underestimation. The ideal Bias value is zero, which indicates a perfectly unbiased model. Bias is expressed as

$$\text{Bias} = \frac{1}{N} \sum_{i=1}^N (f_i - o_i) \quad (3.9)$$

Furthermore, the Standard Deviation Error (SDE) is another measure for evaluating accuracy. However, SDE is derived from RMSE, and therefore, RMSE can also be described as a function of Bias and SDE:

$$\text{RMSE} = \sqrt{\text{SDE}^2 + \text{Bias}^2} \quad (3.10)$$

In this dissertation, some experiments will use SDE as an evaluation metric; however, most of the experiments will employ one or two metrics to display the accuracy: (1) RMSE or (2) Bias in combination with RMSE. As SDE can be described as a function of RMSE and Bias, it is unnecessary to display SDE with the others.

3.4 Reconstruction with AnEn Method on PCs

In this section the AnEn method is applied to a hindcasting problem with the dataset presented in Section 3.1. Several experiments were conducted with the aim of evaluating the effects of using principal components instead of the original historical variables. The Bias error (Bias) and the root mean square error (RMSE) were used to assess the accuracy of the results.

The tests were divided into three combinations of PCs, according to the magnitude of the variance proportion. Additionally, the results obtained with the classical AnEn method applied to the original variables are also presented, allowing for a comparison with the results obtained with the PCs. For each combination of PCs, one test with dependency and one without is presented. It is important to note that when using only PC_1 , or using the classical AnEn method, there cannot exist dependency, since there is only one time-series (see [85]).

Table 3.4 shows that for Soutelo predicting Valongo, in general, the use of two or three PCs results in smaller errors in WSPD, GST and WDIR. Meanwhile, AnEn was superior in comparison to all PC combination in the HRH and ATMP variables, with the exception of the Bias measure in the HRH.

Table 3.4: Valongo variables predicted by Soutelo PCs and by the classical AnEn.

Soutelo predicting Valongo									
Variable	Depend	1 PC		2 PC		3 PC		AnEn	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
WSPD	Yes	-0.13	0.76	-0.05	0.65	-0.08	0.65	-0.08	0.71
	No	—	—	-0.04	0.73	-0.06	0.76	—	—
GST	Yes	-0.29	1.78	-0.09	1.50	-0.15	1.52	-0.24	1.64
	No	—	—	-0.06	1.72	-0.09	1.82	—	—
WDIR	Yes	1.25	1.75	1.23	1.74	1.26	1.78	1.28	1.84
	No	—	—	1.23	1.73	1.23	1.73	—	—
ATMP	Yes	-1.18	8.55	0.48	4.33	-0.96	4.22	-0.08	2.31
	No	—	—	-0.42	6.17	-1.41	6.60	—	—
HRH	Yes	-4.18	19.89	-8.76	18.70	-5.14	15.74	-5.05	13.89
	No	—	—	-6.16	16.74	-4.05	16.28	—	—

In turn, Table 3.5 shows the results of Valongo variables predicted by Edroso PC s. It can

Table 3.5: Valongo variables predicted by Edroso PCs and by the classical AnEn.

Edroso predicting Valongo									
Variable	Depend	1 PC		2 PC		3 PC		AnEn	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
WSPD	Yes	-0.07	0.74	-0.06	0.69	-0.06	0.69	-0.03	0.85
	No	—	—	-0.05	0.76	-0.04	0.77	—	—
GST	Yes	-0.09	1.72	-0.09	1.62	-0.07	1.62	-0.04	1.87
	No	—	—	-0.05	1.82	-0.03	1.85	—	—
WDIR	Yes	0.63	1.76	0.64	1.77	0.63	1.75	0.64	1.78
	No	—	—	0.64	1.76	0.64	1.77	—	—
PRES	Yes	0.32	5.36	0.42	5.20	0.43	4.36	-0.83	2.01
	No	—	—	0.21	5.36	0.16	5.12	—	—
ATMP	Yes	-0.49	8.619	-0.50	5.11	-0.45	3.44	-0.14	2.69
	No	—	—	-0.28	6.61	-0.21	7.34	—	—

also be observed that, for most variables, the smallest errors are obtained with two or three PCs. The AnEn method achieves good results in the case of the RMSE measure in the ATMP and HRH variables.

Tables 3.6 and 3.7 concern to the same type of experiments presented in Tables 3.4 and 3.5, but now without the wind direction (WDIR) variable. Thus, in these tests the WDIR variable was not used to obtain the PCs.

Table 3.6: Valongo variables predicted by Soutelo PCs and by the original AnEn method without the WDIR variable.

Soutelo predicting Valongo									
Variable	Depend	1 PC		2 PC		3 PC		AnEn	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
WSPD	Yes	-0.06	0.72	-0.03	0.65	-0.02	0.65	-0.08	0.71
	No	—	—	-0.04	0.72	-0.04	0.76	—	—
GST	Yes	0.11	1.68	0.04	1.50	-0.01	1.47	-0.24	1.64
	No	—	—	-0.06	1.72	-0.05	1.82	—	—
ATMP	Yes	-1.54	8.83	-0.53	4.15	-0.17	2.10	0.08	2.31
	No	—	—	-1.05	6.02	-1.14	6.19	—	—
HRH	Yes	3.12	19.91	-5.61	16.37	-4.95	13.97	5.02	13.89
	No	—	—	-4.50	15.45	-3.59	15.24	—	—

In Table 3.6, relative to Valongo predicted by Soutelo, the best scores of RMSE (lowest)

Table 3.7: Valongo variables predicted by Edroso PCs and by the original AnEn method without the WDIR variable.

Edroso predicting Valongo									
Variable	Depend	1 PC		2 PC		3 PC		AnEn	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
WSPD	Yes	-0.04	0.81	-0.06	0.79	-0.08	0.81	-0.04	0.86
	No	—	—	-0.04	0.83	-0.04	0.84	—	—
GST	Yes	-0.06	2.0	-0.12	1.9	-0.16	2.0	-0.05	1.9
	No	—	—	-0.07	2.0	-0.08	2.0	—	—
ATMP	Yes	0.06	6.1	0.43	4.0	-0.30	2.7	-0.19	2.8
	No	—	—	0.44	5.5	0.00	5.5	—	—
HRH	Yes	-3.2	15.4	-4.3	15.8	-5.7	15.0	-8.2	16.6
	No	—	—	-4.2	15.1	-4.6	15.3	—	—

are generally obtained with three PCs, with the exception of the HRH variable, where the smallest RMSE is obtained with the AnEn method.

In the case of Valongo predicted by Edroso (Table 3.7), the lowest errors obtained were more equally distributed among the different combinations of components. But PC_3 still remained superior in two (ATMP and HRH) of the four variables analyzed.

Without WDIR in the analysis, the method was able to predict the ATMP and HRH variables more successfully at both stations. For example, in comparison to the analysis with all the 5 variables, ATMP had an RMSE decrease of 50.27% at Soutelo, which allowed the 3 PC combination to outperform the classic method (AnEn).

The predictions of the wind-related variables were weakened by the absence of wind direction at Edroso, since the WSPD and GST errors increased. This was not observed at the Soutelo station, indicating that only at Edroso the WDIR was important towards predicting the wind-related variables.

3.5 Conclusions and Future Directions

This first study exploited the possibility of applying the AnEn method to principal components instead of original variables. The results showed that this technique is effective in reconstructing wind-dependent variables, allowing to reduce the number of original variables

to be processed in the historical dataset. Such reduction happens because the variables are correlated with each other. Consequently, the five original variables can be replaced by only two or three principal components.

With the WDIR in the analysis, the effectiveness of the reconstruction is not as good in the non-wind dependent variables, such as relative humidity and temperature. The presence of wind direction can slightly improve the prediction of wind-related variables at Edroso, but at the cost of significantly worsening the ATMP and HRH reconstruction at both stations. Therefore, the non-presence of WDIR promoted more balanced predictions among the variables, enabling the combination of 3 *PC* to have lower errors in 6 out of 8 RMSE measures.

This study was very conditioned by the dataset used, where there are many missing records. On the other hand, most of the variables, with the exception of the wind-dependent variables, are not correlated with each other and, therefore, it is difficult to reduce their dimension.

In the next chapter, the same methodology (PCA and AnEn) is applied to a better-quality dataset. Such will also allow to better assess the virtues of this new approach when using more predictors that are correlated.

Chapter 4

PCAnEn: Consolidating the combination of PCA and AnEn

The exploratory study of Chapter 3 laid the foundation for incorporating correlated and uncorrelated variables into the prediction process. It was observed that when correlated variables are included in the PCA analysis, the initial new dimensions capture a higher proportion of the variance, thereby retaining more information from the original dataset. On the contrary, the inclusion of uncorrelated variables was found to reduce prediction accuracy due to the decreased information content in the transformed components and the increased difficulty of compressing information into fewer components.

In light of these findings, this chapter seeks to combine PCA and AnEn once more, but with an enhanced dataset comprising a larger number of weather stations with more available data. By applying PCA to reduce the dimensionality of the dataset, when variables are correlated, these new dimensions can be used in predicting variables for nearby weather stations based on the data from other stations included in the PCA. The overarching objective is to merge PCA and AnEn methodologies to effectively reduce dataset dimensions, ultimately resulting in the development of the PCAnEn method. This approach facilitates data reconstruction for a specific weather station using neighboring station data, while allowing for the incorporation of additional stations and information without compromising processing

time.

The remainder of this chapter is structured as follows: Section 4.1 introduces the dataset employed, along with an analysis of the correlations between meteorological variables and stations, Section 4.2 details the process of decomposing the new dataset into principal components using PCA, Section 4.3 focuses on the application of the innovative PCAnEn method to the principal components for reconstructing meteorological variables at a single station, along with the presentation of accuracy and computational efficiency test results. Lastly, Section 4.4 offers concluding remarks and research directions used in the next chapter.

4.1 Meteorological Dataset

The National Data Buoy Center (NDBC), located in the southern Mississippi, in the United States, operates and maintains a network of data collection buoys and coastal stations, with collected data being publicly available [86]. The buoy network is spread worldwide, with the largest numbers located in North America.

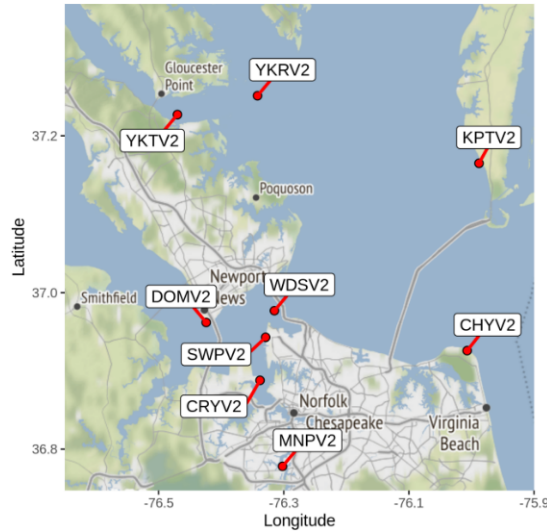


Figure 4.1: Geolocation of the meteorological stations.

Figure 4.1 shows weather stations maintained by NDBC in the region near Hampton and Newport News. In this study, the WDSV2 station is the predicted station. The predictors stations are within a radius of approximately 30 km. For the experiments, the stations were

ordered based on their proximity to WDSV2. The closest are SWPV2, CRYV2 and MNPV2, and so were used first in the test setups.

Table 4.1: Meteorological dataset characterization.

Station	WSPD				GST			
	Min	Mean	Max	Avail(%)	Min	Mean	Max	Avail(%)
WDSV2	0.0	5.7	26.7	97.5	0.0	6.6	32.2	97.5
YKRV2	0.0	5.9	27.6	98.0	0.0	6.9	39.6	98.0
YKTV2	0.0	4.3	23.8	97.7	0.0	5.4	32.8	97.7
MNPV2	0.0	2.6	18.6	96.4	0.0	4.1	30.7	96.5
CHYV2	0.0	5.4	29.7	95.5	0.0	6.9	34.9	95.5
DOMV2	0.0	3.9	24.3	97.5	0.0	5.3	32.1	97.5
KPTV2	0.0	4.7	29.6	97.4	0.0	6.0	35.6	97.5
SWPV2	NA	NA	NA	0	NA	NA	NA	0
CRYV2	0.0	4.1	22.2	82.5	0.0	15.6	30.5	80.5

Station	PRES				ATMP			
	Min	Mean	Max	Avail(%)	Min	Mean	Max	Avail(%)
WDSV2	970.1	1017.4	1044.9	93.6	-12.7	16.5	44.4	87.9
YKRV2	972.6	1017.4	1043.9	98.6	-12.8	15.9	36.3	98.5
YKTV2	974.7	1017.3	1044.3	98.4	-13.5	16.0	37.8	98.2
MNPV2	968.5	1017.5	1044.1	97.9	-13.8	16.8	37.3	97.7
CHYV2	985.2	1017.0	1042.7	31.1	-12.2	16.1	36.5	97.0
DOMV2	972.8	1017.8	1044.5	98.3	-12.6	16.1	37.2	98.2
KPTV2	NA	NA	NA	0	NA	NA	NA	0
SWPV2	972.0	1017.7	1044.1	96.1	NA	NA	NA	0
CRYV2	970.3	1017.6	1044.3	82.8	-10.5	16.5	36.3	34.3

The meteorological variables available in this dataset, measured at each station, are: air pressure (PRES) [bar]; air temperature (ATMP) [°C]; wind speed (WSPD) [m/s], peak gust speed (GST) [m/s]. WSPD and GST can vary significantly over short time intervals, and sporadic data gathering incompletely describes their real behavior. NDBC solved this problem by sampling the wind speed (WSPD) every 6/8 seconds and averaging the readings across 6 minutes; also, it considers the maximum wind speed on the same interval as the peak gust speed (GST). In turn, the collection of ATMP and PRES is straightforward: the instantaneous value every 6 minutes is the one recorded in the NDBC database.

The characterization of these variables is shown in Table 4.1. Variables with more than

85% data availability were selected for the analysis. However, variables have different availability at different stations. Thus, for each variable, different stations combinations were chosen to maximize data availability.

4.1.1 Data Correlation

A correlation study between variables and stations was performed to define the test setups. Variables and stations that are sufficiently correlated with each other can be used together in the PCA technique, as more correlation allows more information to be kept in fewer dimensions.

Figure 4.2 presents the correlations between different meteorological variables within the same station. In stations KPTV2 and CHYV2 there are only records of two (WSPD and GST) and three (WSPD, GST and ATMP) variables, respectively. In the other stations, there are records of four variables: WSPD, GST, ATMP and PRES.

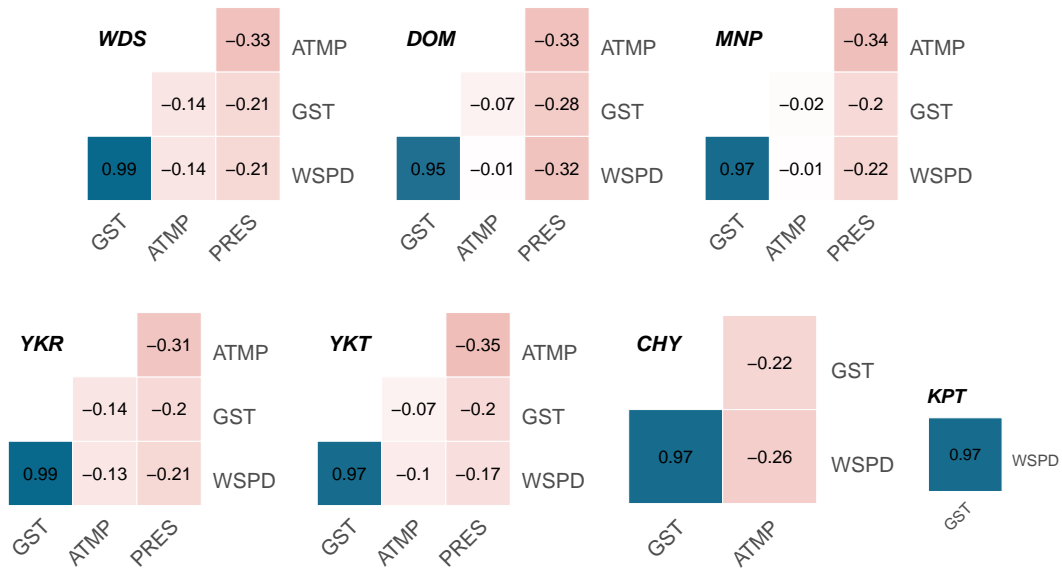


Figure 4.2: Correlation between variables at each station.

In all stations there is a high correlation between WSPD and GST. A mild inverse correlation is observable between ATMP and PRES. The other variable interactions showed low and inconsistent correlation among the stations.

Furthermore, Figure 4.3 shows the correlations for the same variable between different stations. It can be observed that, with minor exceptions, correlations are high, with variables PRES and ATMP showing the highest correlations between different stations.

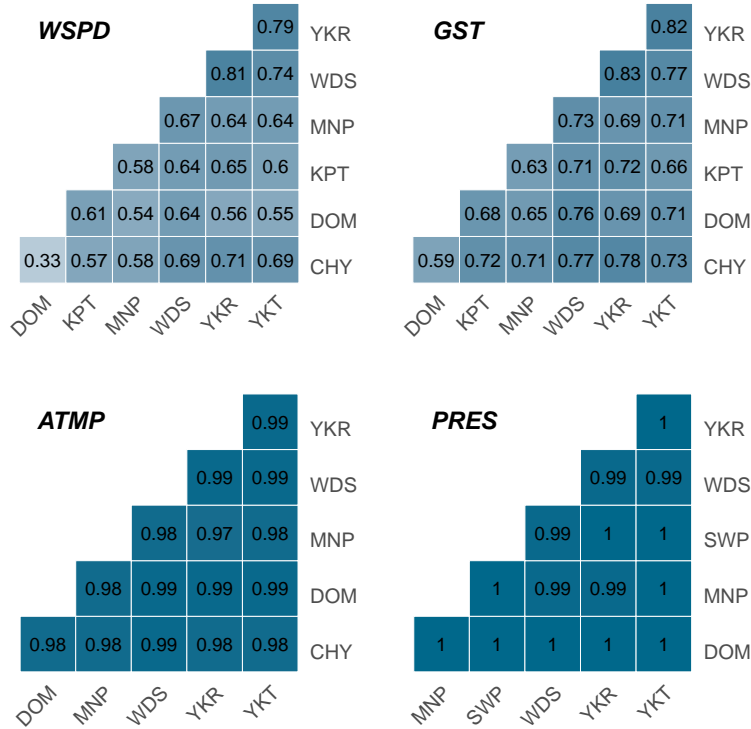


Figure 4.3: Correlation between stations for each variable.

4.2 Dataset Decomposition Using Principal Components

As seen in 3.2.1 the PCA technique identifies the dimensions along which the data are most dispersed (the dimensions that best differentiate the data set under analysis, that is, its principal components).

A decomposition into principal components (*PCs*) of the original meteorological variables, coming from different stations, was performed. Tables 4.2 and 4.3 show the standard deviations of each *PC* for a different amount of input stations. In table 4.2, *PCs* were calculated from the variables WSPD and GST coming from a number of neighboring stations

between 2 and 6. In Table 4.3, PC s were calculated from the variables PRES and ATMP from a number of neighboring stations ranging from 2 to 5. The variables from the WDSV2 station were not included in the original variables because WDSV2 was used only as the predicted station.

In Table 4.2 the PCA is performed from the data matrix that includes two meteorological variables, GST and WSPD, coming from different stations. This is because GST and WSPD are highly correlated (recall section 4.1.1) and so it is possible to use them together in the PCA. In this table, standard deviations above 1 are highlighted. When this occurs, the corresponding PC has a higher variance than the original scaled variables and, consequently, more information.

Table 4.2: Standard deviation of the PC s generated from the variables WSPD and GST coming from different stations together.

Stations	Standard Deviation					
	WSPD and GST					
	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6
2	1.771	0.881	0.234	0.169	—	—
3	2.059	1.027	0.782	0.207	0.172	0.148
4	2.386	1.030	0.807	0.688	0.207	0.178
5	2.689	1.038	0.848	0.691	0.609	0.207
6	2.913	1.047	0.885	0.834	0.652	0.607

It can also be seen in Table 4.2 that, for most cases, the standard deviation is greater than 1 for PC_1 and PC_2 , meaning that these two new variables concentrate the information contained in all the original variables (included in the data matrix H). As in [12], the PC s with standard deviations higher than 1 were chosen to represent the original dataset. It can be seen that for WSPD and GST, PC_1 and PC_2 showed values higher than 1, except for the 2-station configuration. As expected, by increasing the amount of input stations, more components are needed to represent the original dataset.

Table 4.3 shows the standard deviations of the PC s computed from the ATMP and PRES variables. It is important to note that, unlike Table 4.2, these variables were analyzed separately, because they do not correlate sufficiently with each other. The same pattern of values

was observed for both variables, that is, PC_1 was sufficient to represent the data in all configurations.

Table 4.3: Standard deviation of the PC s generated from the variables PRES and ATMP coming from different stations together.

Stations	Standard Deviation							
	PRES				ATMP			
	PC_1	PC_2	PC_3	PC_4	PC_1	PC_2	PC_3	PC_4
2	1.414	0.039	—	—	1.408	0.137	—	—
3	1.731	0.040	0.039	—	1.721	0.154	0.126	—
4	1.998	0.063	0.039	0.028	1.987	0.154	0.129	0.102
5	2.233	0.086	0.050	0.037	2.220	0.185	0.139	0.102

4.3 Experiments with the PCAnEn Method

In this section, the PCAnEn method is applied to a hindcasting problem with the dataset presented in Section 6.1. Several experiments were conducted in order to evaluate the effects of using principal components instead of the original historical data. The accuracy of the reconstructed values is assessed by comparison to the exact values recorded at the WDSV2 station during the prediction period.

The available historical data ranges from 2011 to the last hour of 2018, and the reconstruction period is 2019. Because of the high resolution (6 minutes) and large amount of data, it was decided to make predictions only between 10 am and 4 pm, every 6m. For the classical AnEn experiments, the original data is used instead of the PC s.

All tests were performed in duplicate, using two different implementations of the methods, one in R [87] and another in MATLAB [88]. This provided confidence on the numerical results obtained (which were expected, and verified, to be identical) and also allowed to compare the respective computational performance. The computer system used in the experiments was a virtual machine hosted on the The Research Centre in Digitalization and Intelligent Robotics (CeDRI) virtualization cluster, running Ubuntu 20.04.4 LTS. The resources associated with the virtual machine were 16 virtual cores of a Intel Xeon W-2195 CPU, 64GB

of RAM and 256GB of secondary storage (SSD-based).

The tests were divided between different amounts of predictor stations. In addition, the results obtained with the classical AnEn method applied to the original variables are also presented, allowing a comparison with the results obtained with the PCAnEn methodology. Note that for AnEn, the same stations as the PCAnEn (2-station configuration) were chosen for the prediction; the variable is predicted from the same variable located in the two closest stations, in order to ensure the most favourable configuration of the AnEn method.

The Subsection 4.3.1 presents and discusses the accuracies obtained from the experiments. The Subsection 4.3.2 shows a comparison of performance between the AnEn and PCAnEn methods, and between R and MATLAB implementations.

4.3.1 Comparing Accuracy

Figure 4.4 allows to compare the accuracy of the AnEn and PCAnEn methods, with different amounts of stations, for the four meteorological variables considered in this study. For each combination, the number of *PCs* used is 1 or 2, as indicated in Tables 4.2 and 4.3.



Figure 4.4: Comparison of the RMSE for different variables and number of stations.

The chart is based on the accuracies provided by the R implementation; however, they

Table 4.4: Accuracy comparison between the PCAnEn and AnEn methods.

Method	#St	Errors	R				MATLAB			
			WSPD	GST	ATMP	PRES	WSPD	GST	ATMP	PRES
PCAnEn	2	Bias	-0.09	-0.35	0.20	0.44	-0.09	-0.35	0.20	0.44
		RMSE	1.67	1.97	1.02	0.51	1.68	1.97	1.02	0.51
	3	Bias	-0.06	-0.33	0.18	0.37	-0.06	0.33	0.18	0.37
		RMSE	1.35	1.55	0.85	0.47	1.35	1.55	0.85	0.47
	4	Bias	-0.04	-0.29	0.14	0.36	-0.04	-0.29	0.14	0.36
		RMSE	1.29	1.48	0.79	0.44	1.29	1.48	0.79	0.44
	5	Bias	-0.06	-0.30	0.12	0.51	-0.06	-0.30	0.12	0.51
		RMSE	1.21	1.39	0.73	0.60	1.21	1.39	0.73	0.60
	6	Bias	-0.07	-0.31	—	—	-0.07	-0.31	—	—
		RMSE	1.26	1.44	—	—	1.26	1.44	—	—
AnEn	2	Bias	-0.15	-0.35	0.16	0.41	-0.15	-0.35	0.16	0.41
		RMSE	1.73	1.77	0.88	0.52	1.73	1.77	0.88	0.52

are very similar to those of the MATLAB implementation, as may be seen in Table 4.4; this table provides the full accuracy results, including also the Bias in addition to the RMSE.

The smallest errors were obtained by the PCAnEn method in the configurations with 4 or 5 stations. For the variables ATMP, GST and WSPD, 5 stations showed better RMSE. In contrast, for PRES, the 4-station configuration provided the most accurate prediction.

For instance, as shown in Table 4.4, the predictions of WSPD and GST with the PCAnEn method generated 30% and 21.8% lower RMSE errors compared to the classical AnEn method, respectively. To a lesser effect, the reconstructions of the PRES and ATMP variables showed a reduction of 13.6% and 16.7% with the PCAnEn method. Moreover, the lowest BIAS measurements were obtained with 4 stations in all variables.

As depicted in Figure 4.4, the non-PRES variables predicted by PCAnEn consistently exhibited superior prediction performance across all configurations when compared to AnEn, except for the configuration involving 2 stations. This outcome is unsurprising since these stations were identical to the ones utilized in the AnEn tests. Although PCA can effectively condense data into fewer components, it inevitably results in some loss of information.

Regarding the issue of *dependency*, Figure 4.5 shows the values of the RMSE obtained in

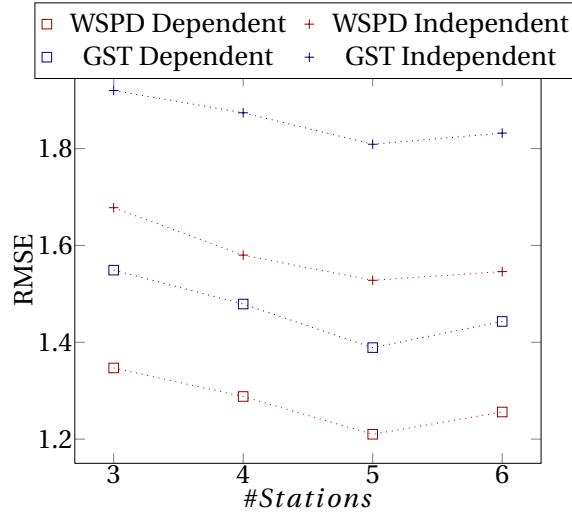


Figure 4.5: RMSE of PCAnEn method used in a dependent and independent way.

the prediction of WSPD and GST by the PCAnEn method used in a *dependent* and *independent* way, with 3 or more stations. The results show clearly that an *independent* PCAnEn did not improve the results in any configuration or station, in comparison with the *dependent* version (as the results in Chapter 3). It is also observed that increasing the number of stations up to more than 5 leads to a reduction in the RMSE error, but the increase to 5 stations no longer brings advantages, since the RMSE increases.

4.3.2 Comparing Performance

Figure 4.6 shows the processing times obtained by the MATLAB and R implementations of the PCAnEn and AnEn methods, with different station quantities. The processing times were measured when using 14 CPU cores (above that number, the decrease in overall execution time was negligible – see Figure 4.7). As previously mentioned, for the AnEn method only 2 stations were used, and so Figure 4.6 only provides two execution times (one for each AnEn implementation).

The PCAnEn method significantly reduces the total processing time compared with the classical AnEn method, for both of the implementation used. This is evident in the 2 stations scenario: using MATLAB, the PCAnEn method consumes 38% (30.4/79.3) of the time spent

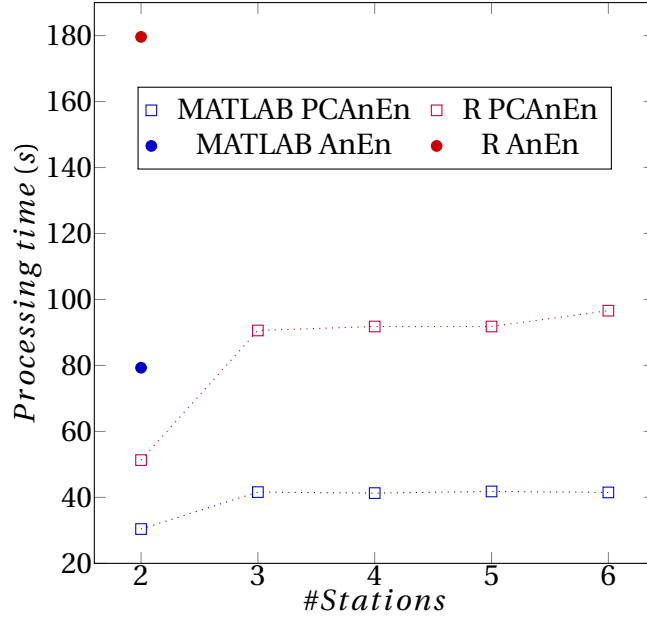


Figure 4.6: Processing time for different number of stations and different methods.

by the AnEn method (a speedup of 2.6x); in turn, using the R implementations, the PCAnEn method runs in 28% (51.3/179.6) of the time needed by the AnEn method (a speedup of 3.5x).

Focusing only on the PCAnEn method, the processing times varies little with the number of stations, in both implementations (the exception is the 2 stations scenario, where the processing time is visibly smaller than with more stations).

For any number of stations used, the MATLAB implementation was always found to be faster than the R implementation. For instance, with 6 stations, PCAnEn in MATLAB was 2.3x (96.6/41.5) faster than in R, though with 2 stations the speedup was only 1.7x (51.3/30.4). In turn, also with 2 stations, AnEn in MATLAB was 2.3x (179.6/79.3) faster than its implementation in R.

Finally, the processing times in function of the number of CPU cores used was also evaluated. Figure 4.7 shows the processing times for the PCAnEn method with 6 stations (using *PCs* generated from the variables WSPD and GST), when varying the number of CPU cores from 1 up to 14. It can be observed that both implementations scale reasonably well, though with diminishing returns past 8 CPU cores. Again, the MATLAB implementation offers superior performance and slightly better scalability. It should be noted, however, that MATLAB

is known to be particularly optimized to take advantage of Intel CPUs (as the one where this evaluation was performed), once it relies on the Intel MKL library by default.

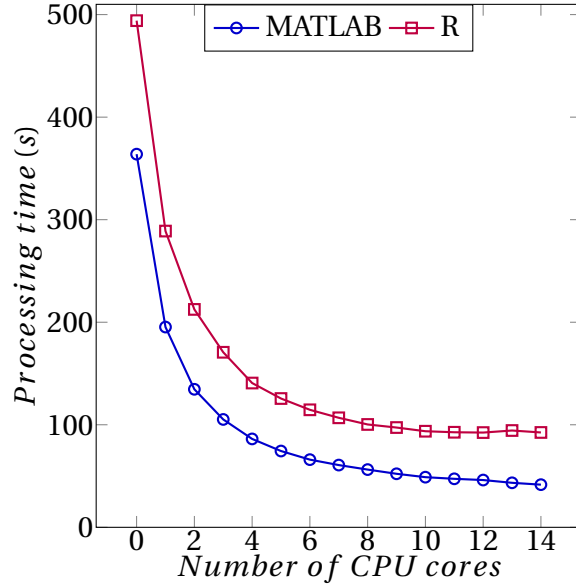


Figure 4.7: Processing time for different number of CPU cores used by PCAnEn.

4.4 Key Findings

This chapter presented a study on the combination of Principal Component Analysis (PCA) and Analog Ensemble (AnEn) in the form of the novel PCAnEn method. By utilizing an expanded dataset, which includes a greater number of weather stations and more reliable data, it was successfully demonstrated the potential of PCAnEn for weather data reconstruction.

The combination of the PCA technique with the AnEn method offers a better hindcasting accuracy than the classical AnEn method. In the present study, the data reconstruction of the WDSV2 station by means of the 5 nearest station seems to be optimal. However, the choice of predictor stations must take into account the proximity and correlation between them (which needs to be assessed prior to the determination of the *PCs*).

In terms of computational performance, the PCAnEn method allows to reduce the processing time considerably, compared to the classical AnEn method. It was also verified that the implementation in MATLAB is faster (and by which magnitude) than the implementation

in R. This information may then be considered in the choice between a proprietary non-free platform and a an open-source free one, to solve the same kind of hindcasting problems.

Chapter 5

PCClustAnEn: Enhancing PCAnEn with K-means Clustering

ClustAnEn (Cluster-based AnEn) is a variant of the AnEn method, which uses K-means clustering to group feasible analogues, reducing computational costs without sacrificing the accuracy of the reconstructions [7], [85]. Another approach to enhance the AnEn method is to combine it with PCA, as seen in Chapters 3 and 4. Therefore, this chapter builds upon previous investigations of the PCAnEn method (as discussed in Section 4) and expands it by applying the same approach of integrating the PCA technique to the ClustAnEn method, resulting in a new variant called PCClustAnEn.

Combining PCA with ClustAnEn further reduces processing times while maintaining high accuracy in the reconstruction of meteorological data. The aim is to develop a novel approach that outperforms existing methods in terms of accuracy and computational speed. A comparative study of the performance of all these methods in a hindcasting problem is also presented, corresponding to the reconstruction of missing data in a given meteorological station by means of data coming from a set of predictor stations with different geographical locations.

The reconstruction experiments utilize the same data source used in Chapter 4 from the

US government’s National Data Buoy Center (NDBC)[86]. The configuration of stations remains also the same, with WDSV2 serving as the predicted station and the others functioning as predictors. As a result, the correlations between stations and variables are identical to those in Chapter 4. Consequently, the experiments are conducted using the same configuration of combined variables: WSPD and GST are integrated for predictions, while the remaining variables are used individually.

5.1 The ClustAnEn Method

The extension of the training period has an influence in the performance of the AnEn method. The longer the training period, the more accurate the predictions/reconstructions are expected to be. On the other hand, longer training periods imply greater computational effort to identify the analogs in each reconstruction. To alleviate this problem, an alternative version of the AnEn method was developed in which all possible analogs are previously classified into a predefined number of clusters [7], [85], with the number of clusters set to the square root of the total number of possible analogs. This heuristic is based in the empirical results previously obtained [89].

As presented in Figure 5.1, the ClustAnEn method starts by clustering (with k-means) the data of the predictor \mathbf{x}_1 (training period). Then, the predictor vector in the prediction period is compared with each cluster centroid to identify the analog cluster, which contains a certain number of vectors.

ClustAnEn significantly speeds up the process of identifying the analog ensemble compared to the classic AnEn method. The reason behind this acceleration is the reduced number of clusters compared to the total number of possible analogs. Consequently, the number of comparison calculations required to find the centroid most similar to the value is greatly reduced, enhancing computational efficiency. It is important to note that steps 2 and 3 of the ClustAnEn method remain the same as the original AnEn method (see Section 3.4).

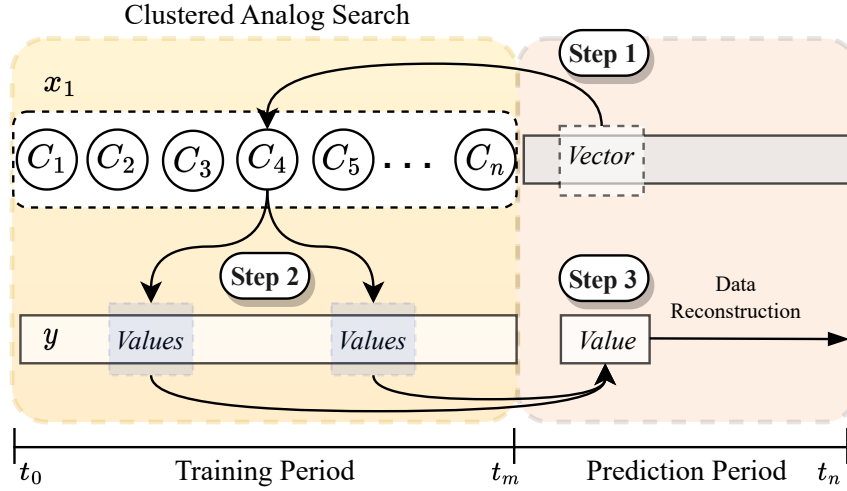


Figure 5.1: Reconstruction of missing meteorological records with the ClustAnEn method.

5.2 Experimental Evaluation

As the previous chapter, the reconstruction was performed by two separate implementations of the methods, one in R and another in MATLAB. The computing system used is also the same as Section 4.3.

Besides testing the PCAnEn and PClustAnEn methods, the corresponding non-PCA variants (AnEn and ClustAnEn) applied to the original datasets were also tested. This way, the specific impact of the PCA technique may also be assessed. To make the comparison fair, AnEn and ClustAnEn were tested using as predictors the same two stations used to test the PCAnEn and PClustAnEn with a 2-station configuration ($\#Stations = 2$). This means that the variable is predicted from the same variable located in the two closest stations, thus ensuring the most favourable configuration to the AnEn and ClustAnEn methods.

Table 5.1 presents the RMSE values for all tests performed. For each test, the number of PCs used was 1 or 2, after the values of the respective standard deviation, as explained in Section 3.3. Between PCAnEn and PClustAnEn, there were no noteworthy changes in accuracy. The 5-station setup demonstrated a lower RMSE than the non-PCA approaches for the majority of variables. The higher errors are obtained with the 2-station configurations, in which case there's no sensible advantage in using the PCA variants over the non-PCA ones.

Table 5.1: RMSE of the reconstruction with different methods.

Method	#St	MATLAB				R			
		WSPD	GST	ATMP	PRES	WSPD	GST	ATMP	PRES
PCAnEn	2	1.65	1.95	1.01	0.51	1.65	1.97	1.01	0.51
	3	1.32	1.52	0.84	0.48	1.32	1.52	0.84	0.48
	4	1.27	1.46	0.78	0.45	1.27	1.45	0.78	0.45
	5	1.19	1.36	0.71	0.61	1.19	1.36	0.71	0.61
	6	1.24	1.42	—	—	1.24	1.44	—	—
AnEn	2	1.68	1.77	0.86	0.59	1.67	1.76	0.86	0.59
PCClust	2	1.65	1.95	1.01	0.52	1.65	1.95	1.01	0.52
	3	1.32	1.50	0.84	0.48	1.32	1.51	0.84	0.48
	4	1.27	1.45	0.78	0.45	1.28	1.45	0.78	0.45
	5	1.20	1.35	0.72	0.61	1.20	1.36	0.72	0.60
	6	1.27	1.40	—	—	1.25	1.42	—	—
ClustAnEn	2	1.69	1.73	0.88	0.60	1.69	1.74	0.87	0.56

The reductions in error rates from the PCA implementations ranged from $\approx 18\%$ to $\approx 30\%$, for the best setting of each variable, compared to the non-PCA methods. These considerations apply to both implementations (R and MATLAB).

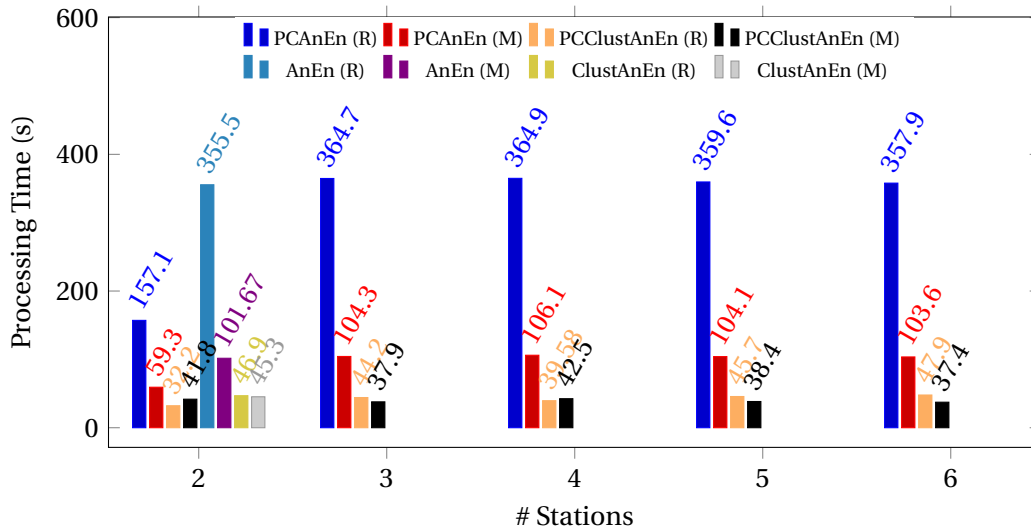


Figure 5.2: Reconstruction time of WSPD with 16 cores (2 to 6 stations).

Regarding the computational performance, Figure 5.2 shows the processing times of the

MATLAB (M) and R (R) codes, with different amounts of stations, for the reconstruction of the WSPD variable, using all the CPU cores (16) available in the test bed computational system. The WSPD variable was chosen for the performance evaluation because a) it is available for more stations (recall Table 4.3), and b) it requires 2 *PCs* to represent the original variables when using 3 or more stations. The same is also valid for the GST variable, whose processing times are either the same (PCA-based approaches) or similar (other approaches).

When using clustering (ClustAnEn and PCClustAnEn), the reconstruction times are the lowest, and the variations are small for different numbers of stations, whether using PCA or not; also, for the only scenario where it makes sense to use the non-PCA approaches (2 stations), using clustering alone (ClustAnEn) is slower than combining it with PCA (PCClustAnEn).

Without clustering (AnEn and PCAnEn), the processing times are noticeably higher. When applying PCA (PCAnEn), the highest times are obtained with 3 or more stations (using 2 *PCs*), and they are similar; these times roughly double the time with 2 stations (using 1 *PC*); thus, without clustering, the number of *PCs* used has a noticeable influence (direct proportionality) on the processing times. For the 2-stations scenario, not using PCA (AnEn) doubles the processing times compared to using PCA (PCAnEn), being equivalent to using PCA with more than 2 stations, once it is also using two time series.

Lowering the processing times is important, but it shouldn't be at the expense of higher reconstruction errors. Ideally, the reconstruction should be faster and also more accurate. The smallest RMSE errors for the WSPD variable are obtained with PCA-based methods using 5 stations, whether clustering is used (PCClustAnEn) or not (PCAnEn) – recall Table 5.1. However, clustering ensures much lower processing times, with a speedup between $\gtrsim 2.7$ (MATLAB code) and $\gtrsim 7.8$ (R code). Comparing the processing times of PCClustAnEn with 5 stations, with the ones of ClustAnEn with 2 stations (the best provided by not using PCA) yields almost none speedup ($46.9/45.7=1.03$ and $45.3/38.4=1.18$); however, the RMSE error of PCClustAnEn with 5 stations is only $1.2/1.69 \approx 70\%$ of the error of ClustAnEn with 2 stations, thus favouring the first approach.

The impact on performance of using or not the PCA method is perceivable in the 2-stations scenario. Here, using PCA provides speedups ranging from 2.26 to 1.08, for comparable methods (AnEn vs PCAnEn, and ClustAnEn vs PCClustAnEn).

Another advantage of adding PCA emerges when two variables, like WPSD and GST, are used together in the analysis. Once they share the same time series, PCA-based methods can predict both variables in a single run, unlike the non-PCA approaches, which would require two runs of the reconstruction code.

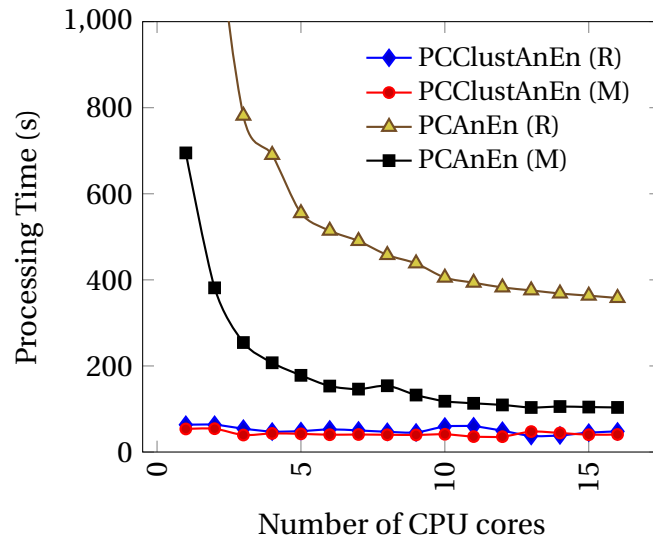


Figure 5.3: Reconstruction time of WSPD with 1 to 16 cores (6 stations).

The MATLAB code was found consistently faster than the R code. This is visible in Figure 5.2 for 16 cores, and can also be seen in Figure 5.3 for a variable number of cores. However, under PCClustAnEn the differences were minor, meaning both implementations are equally efficient when applying the K-means clustering. More important, PCClustAnEn required much less processing times, in all configurations, compared to PCAnEn. Also, PCClustAnEn mostly doesn't benefit from the extra cores, in opposition to the PCAnEn method, where the search for analogues is the biggest code hotspot and is easily parallelizable.

It should also be stressed that both R and MATLAB were used with default configurations, without any extra performance tuning to optimize their behavior.

5.3 Main Takeaways

Findings focus on the combination of the ClustAnEn method, which employs K-means clustering to group feasible analogues, thereby reducing computational costs without sacrificing reconstruction accuracy, as previously demonstrated in the literature [7], [85]. The investigation built upon earlier studies of the PCAnEn method, described in Chapters 3 and 4, and expands upon it by incorporating PCA into the ClustAnEn method, resulting in a new variant called PCClustAnEn.

By utilizing PCA, it is possible to reduce the data from several stations into a smaller number of time series, corresponding to the Principal Components, which are then employed to reconstruct missing data in the records of a meteorological site. As demonstrated in the experiments, the PCA technique improves prediction assertiveness without compromising computational performance, allowing for an increased number of stations without raising the number of input time series. However, the efficacy of PCA is heavily influenced by the correlation between the time series of several predictors, with higher correlation leading to a greater proportion of information/variance in the first components.

Moreover, this chapter also exploited two different implementations of the methods, one in MATLAB and another in R, which enable to double-check numerical results and assess the potential performance impact of choosing either implementation. It was also investigated the scalability of both codes within a medium-scale multicore system, revealing the superiority of AnEn methods that combine PCA with clustering.

Chapter 6

PLS AnEn-based Methods and Regression Techniques for Meteorological Data Reconstruction

In the previous chapter, it was shown that the combination of Principal Component Analysis (PCA) and clustering significantly improves prediction accuracy without compromising computational performance. This chapter moves further by exploiting the Partial Least Squares (PLS) technique with the AnEn method and its clustering variant, originating the new PLSAnEn and PLSClustAnEn methods (supervised methodologies, in contrast to PCA). Typically, PCA and PLS are combined with multivariate linear regression, such as the Principal Components Regression (PCR) and Partial Least Square Regression (PLSR) methods. Then, a comparative analysis of the performance of all these methods in a hindcasting problem is also conducted.

Additionally, previous studies were conducted on datasets containing three to nine stations, with each station often having data for fewer than four variables. In this chapter, the experiments employ a dataset with a larger volume of available data. Moreover, in these experiments, each station serves as both a predictor and a predicted variable at different instances. This approach enhances the robustness of the experiments by eliminating the

possibility of a single station yielding better results with a particular method due to chance.

The remaining of the chapter is organized as follows: Section 6.1 introduces the meteorological data sets used, Section 6.2 presents the various reconstruction methods employed in this study; Sections 6.3 and 6.4 focus on the selection of principal components and latent variables, respectively. In Section 6.5, the numerical results of the tests performed with the various reconstruction methods are presented. Section 6.6 provides an analysis of the computational performance of the same methods. Finally, Section 6.7 concludes, summarizing the main findings and their implications for the solving of hindcasting and forecasting problems with a high number of predictors.

6.1 Meteorological Data Sets

Similar to the studies in Chapters 4 and 5, this research makes use of the NDBC as the data source for the experiments. The NDBC database covers regions across almost the entire US coast, as well as various other locations worldwide. For this study, it was sought a region with the highest possible station density. The selected region, centered north of the San Francisco Bay (California, USA), features records from 16 meteorological stations in close proximity. Figure 6.1 illustrates the chosen region and its 16 NDBC stations.

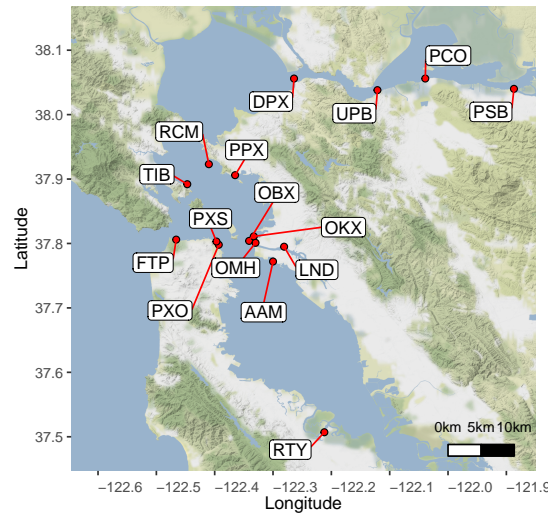


Figure 6.1: Geolocation of the selected NDBC meteorological stations.

Records of various meteorological variables are available for each station, with a sampling period of 6 minutes. The variables used in this study are: atmospheric pressure (PRES) [mbar], air temperature (ATMP) [°C], wind speed (WSPD) [m/s] and peak gust speed (GST) [m/s].

Records start in January 1st, 2016, and end in December 31st, 2021. Because of sensor failures and maintenance operations in the stations, variables often present time series with incomplete data. Table 6.1 presents the mean, standard deviation (SD) and availability (in percentage), for the considered variables (WSPD, GST, ATMP and PRES), in the 16 stations.

Table 6.1: Characterization of the Dataset. Available data (Avail.) is represented as a percentage.

Station	WSPD		GST		ATMP		PRES	
	Mean	Avail.	Mean	Avail.	Mean	Avail.	Mean	Avail.
AAM	2.2	93.9	3.1	93.7	12.7	98.8	1016.9	98.9
DPX	3.6	97.6	4.9	97.6	12.9	98.7	1016.1	98.7
FTP	2.5	97.8	4.2	97.8	14.3	98.9	1016.8	99
LND	2.1	92.5	3	92.5	12.8	93.6	1016.6	93.6
OBX	NA	NA	NA	NA	12.8	95.7	NA	NA
OKX	2.6	81.7	3.6	81.7	NA	NA	NA	NA
OMH	3.0	85.5	4.0	85.5	NA	NA	NA	NA
PCO	4.3	88.5	5.6	88.5	12.3	93.3	1016.2	93.4
PPX	3.8	95.2	5.1	95.2	13.1	96.6	1016.8	96.6
PSB	3.9	95.1	5.4	95.1	13.7	96.2	1015.8	96.2
PXO	2.2	84.7	3.4	84.7	12.6	87.8	1015.6	76.1
PXS	NA	NA	NA	NA	13.2	96.9	NA	NA
RCM	2.6	93.9	3.9	93.9	12.8	98	1016.4	98
RTY	1.7	96.1	2.5	95.7	12.5	97.6	1017.0	97.7
TIB	1.7	3.3	2.7	3.3	NA	NA	1016.3	3.3
UPB	4.2	41.2	5.6	41.2	NA	NA	NA	NA

To maximize the amount of training data, only variables with more than 85% of availability (in bold in Table 6.1) were chosen for this study. The presence of NA in Table 6.1 means that the corresponding variable is not available at the corresponding station. Consequently, the working dataset consists of a total of 41 time series: 10 series with records of the WSPD variable, 10 of the GST variable, 12 of the ATMP variable and 9 of the PRES variable.

Figures 6.2 to 6.5 show the correlations between the stations with more records for each

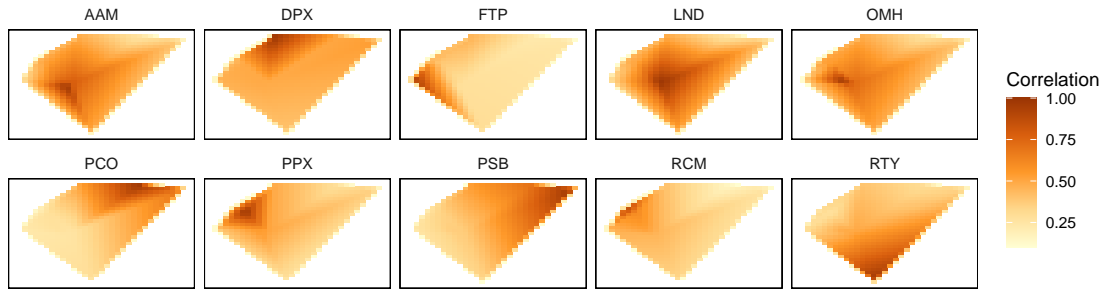


Figure 6.2: Correlation between stations for the WSPD variable.

variable (10 stations for the WSPD, GST and ATPM variables, and 9 stations for the PRES variable). In each figure, a heat-map is shown for each station; the polygonal format of this map matches (approximately) a 4-side polygon that would include the 16 stations, preserving their geographical positions and distances, considering the layout of Figure 6.1. In each heat-map, only 10 (9) points represent station correlations - the points corresponding to their locations; the correlations for the other points were produced by interpolation.

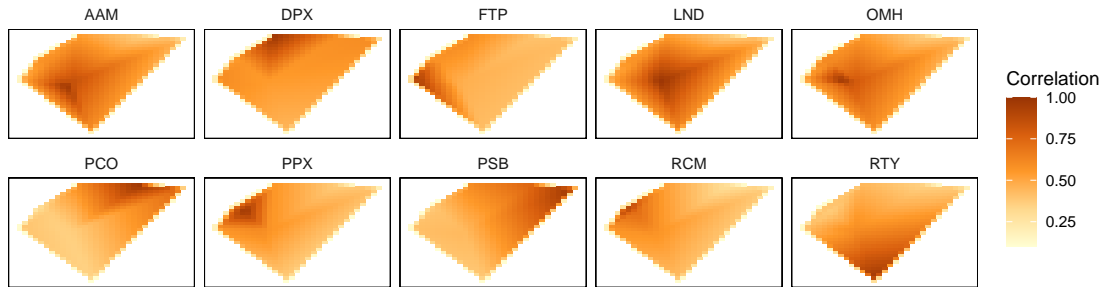


Figure 6.3: Correlation between stations for the GST variable.

As may be observed in Figures 6.2 and 6.3, for the WSPD and GST variables only the closest stations present a strong correlation. Also, the GST variable presents correlations between stations slightly higher than those observed for the WSPD variable. These observations confirm that the wind-dependent variables have a local (and not regional) variation, depending a lot on the morphology of the terrain where the station is implemented.

Figures 6.4 and 6.5 show that the ATPM and PRES variables have a different behavior from that of the WSPD and GST variables (Figures 6.2 to 6.3), as ATPM and PRES present a very high correlation between stations, even among the most distant ones. This shows that

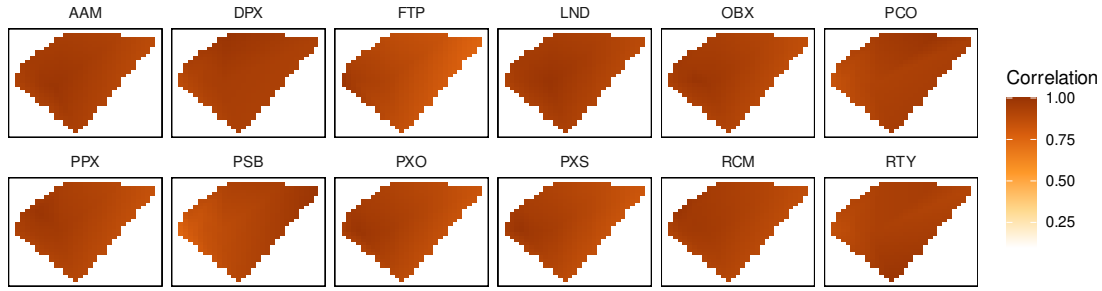


Figure 6.4: Correlation between stations for the ATMP variable.

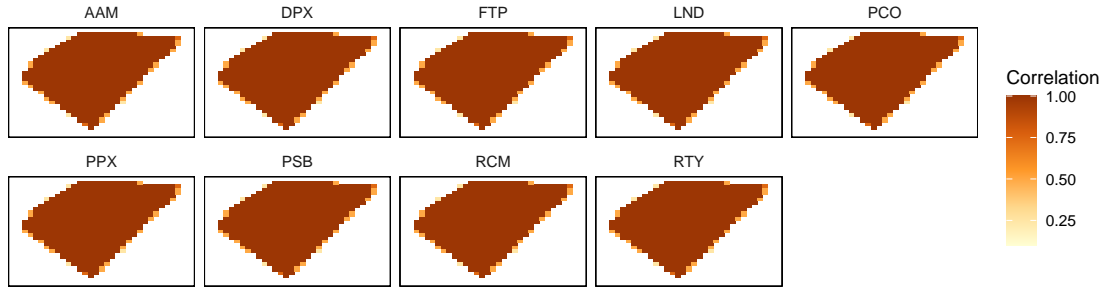


Figure 6.5: Correlation between stations for the PRES variable.

the ATMP and PRES variables have a regional character, varying little or nothing locally.

6.2 Reconstruction Methods

This section introduces two additional reconstruction methods that were not previously discussed in detail: Principal Components Regression (PCR) and Partial Least Squares Regression (PLSR).

6.2.1 Principal Component Regression

As an alternative to the reduction of the size of the predictor dataset, the reconstruction of missing data can be accomplished using *multivariate linear regression*. This method, unlike the AnEn method, allows the direct use of all the original predictor variables.

The goal of the multivariate regression is to predict \mathbf{y} from \mathbf{X} , where $\mathbf{X} \in \mathbb{R}^{m \times q}$ contains in the columns the values of all the predictor variables recorded during the training period,

and $\mathbf{y} \in \mathbb{R}^m$ the corresponding values of the predicted variable. This problem involves the determination of the vector $\mathbf{b} \in \mathbb{R}^q$ that is the approximated solution of the linear system of equations

$$\mathbf{X}\mathbf{b} \approx \mathbf{y}. \quad (6.1)$$

Such is equivalent to solve the linear least squares problem

$$\min_{\mathbf{b}} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|, \quad (6.2)$$

where $\|\cdot\|$ is the usual 2-norm (see [82] for details). If \mathbf{X} is a full rank column matrix, then the solution of the problem (6.2) is given by

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \quad (6.3)$$

The expectation is that the solution vector \mathbf{b} can be used to predict values in the reconstruction period based on the predictor variables for that same period, that is:

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\mathbf{b}, \quad (6.4)$$

where $\tilde{\mathbf{y}} \in \mathbb{R}^{(n-m)}$ represents the reconstructed/predicted variable during the reconstruction/prediction period and $\tilde{\mathbf{X}} \in \mathbb{R}^{(n-m) \times q}$ contain the values of the predictor variables along the same period.

The multivariate regression model given by Eq. (6.4) can be implemented only if the matrix \mathbf{X} has full column rank (its column vectors are linearly independent). The near-collinearity of columns can occur if there are highly correlated predictor variables. In this case, the least squares problem (6.2) becomes ill-conditioned and difficult to solve.

The *principal component regression* (PCR) [90] method circumvents the rank deficiency by replacing the original predictor variables \mathbf{X} by its principal components (PCs) in the regression model. Once the principal components, $\mathbf{Z} = \mathbf{X}\mathbf{V}$, are obtained from matrix \mathbf{X} in the same way as described in section 3.2.1, a few of them (p) are used in the regression model to

estimate \mathbf{y} .

Therefore, the PCR method consists in regressing \mathbf{y} not on \mathbf{X} itself but on the principal components matrix $\mathbf{Z} \in \mathbb{R}^{m \times p}$, assuming that p PCs have been previously selected. This implies to solve by linear least squares the system

$$\mathbf{Z}\mathbf{c} \approx \mathbf{y}, \quad (6.5)$$

whose solution is the parameter vector

$$\mathbf{c} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{y}. \quad (6.6)$$

The PCR regression model on \mathbf{Z} is then given by

$$\tilde{\mathbf{y}} = \tilde{\mathbf{Z}}\mathbf{c}, \quad (6.7)$$

where $\tilde{\mathbf{y}} \in \mathbb{R}^{(n-m)}$ is, as before, the vector of the reconstructed/predicted values of \mathbf{y} during the reconstruction/prediction period, $\tilde{\mathbf{Z}} = \tilde{\mathbf{X}}\mathbf{V} \in \mathbb{R}^{(n-m) \times p}$ contains the values of the p selected PCs along the reconstruction/prediction period, and $\mathbf{c} \in \mathbb{R}^p$ is the parameter vector of the PCR.

The regression model (6.7) can be expressed in function of $\tilde{\mathbf{X}}$ instead of $\tilde{\mathbf{Z}}$ by replacing $\tilde{\mathbf{Z}}$ by $\tilde{\mathbf{X}}\mathbf{V}$, thus originating

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\mathbf{V}\mathbf{c}. \quad (6.8)$$

PCR is an alternative to the AnEn based methods that combines the size reduction provided by PCA with linear regression. This combination prevents collinearity problems between vectors of predictor variables. Another advantage of PCR is the reduction of the number q of original predictor variables to a lower number p of principal components, but which contain most of the original information. Thus, the good performance of this method depends strongly on the choice of the PCs.

It is expected that a few of the PCs, which have a higher variance, are enough to describe

the evolution of the original predictor data set. However, these components were chosen to explain the evolution of the original predictor variables, contained in the matrix \mathbf{X} and, as such, there is no guarantee that these PCs will be relevant for the prediction of \mathbf{y} .

6.2.2 Partial Least Squares Regression

In contrast with the PCR method, the Partial Least Squares Regression (PLSR) method uses the components from \mathbf{X} that best predict \mathbf{y} . These components, also called the latent variables (because they are not directly observed or measured), are coming from the joint decomposition of \mathbf{X} and \mathbf{y} , taking into account the obligation of the components to explain the covariance between \mathbf{X} and \mathbf{y} as best as possible [91], [92].

PLSR computes the latent variables that model \mathbf{X} and \mathbf{y} and best predict \mathbf{y} , resulting in the variable decompositions

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \quad \text{and} \quad \mathbf{y} = \mathbf{R}\mathbf{q}^T + \mathbf{f}, \quad (6.9)$$

where: $\mathbf{T} \in \mathbb{R}^{m \times p}$ and $\mathbf{R} \in \mathbb{R}^{m \times p}$ are the matrix with p latent vectors (also known as scores) extracted from \mathbf{X} and \mathbf{y} , respectively; $\mathbf{P} \in \mathbb{R}^{q \times p}$ and $\mathbf{q} \in \mathbb{R}^p$ represent the loadings vectors; the matrix $\mathbf{E} \in \mathbb{R}^{m \times q}$ and vector $\mathbf{f} \in \mathbb{R}^m$ represent the residuals, whose norms are minimized. Additionally, the scores matrix \mathbf{T} is orthogonal, that is, $\mathbf{T}^T\mathbf{T} = \mathbf{T}\mathbf{T}^T = \mathbf{I}$. The decompositions (6.9) can be achieved by different procedures, such as the *nonlinear iterative partial least squares* (NIPALS) algorithm [93] or the SIMPLS algorithm [94].

The decompositions (6.9) are performed in order to minimize the norm of the residual matrices, \mathbf{E} and \mathbf{f} , and to maximize the covariance between the latent vectors, columns of \mathbf{T} and \mathbf{R} . Consequentially, there is a linear relation between \mathbf{T} and \mathbf{R} , expressed as

$$\mathbf{R} = \mathbf{T}\mathbf{D} + \mathbf{H}, \quad (6.10)$$

where $\mathbf{D} \in \mathbb{R}^{p \times p}$ is a diagonal matrix with the regression weights and \mathbf{H} denotes the matrix of

the residuals. Combining (6.10) with the decomposition of \mathbf{y} , given by (6.9), leads to

$$\mathbf{y} = \mathbf{T}\mathbf{D}\mathbf{q}^T + (\mathbf{H}\mathbf{q}^T + \mathbf{f}), \quad (6.11)$$

or simply

$$\mathbf{y} = \mathbf{T}\mathbf{c}^T + \mathbf{f}^*, \quad (6.12)$$

where $\mathbf{c}^T = \mathbf{D}\mathbf{q}^T \in \mathbb{R}^p$ denotes the regression vector and $\mathbf{f}^* = \mathbf{H}\mathbf{q}^T + \mathbf{f}$ is the residual vector, so that \mathbf{y} can be estimated as

$$\hat{\mathbf{y}} = \mathbf{T}\mathbf{c}^T. \quad (6.13)$$

The regression model (6.13) enables to estimate \mathbf{y} based on the latent variables \mathbf{T} , but it is useful regressing \mathbf{y} on the original predictor variables \mathbf{X} . To accomplish this, the matrix $\mathbf{W} = \mathbf{X}^T\mathbf{U}$, of the PLS weights, computed such that $\mathbf{E}\mathbf{W} = \mathbf{0}$, is post-multiplied by the decomposition of \mathbf{X} in (6.9):

$$\begin{aligned} \mathbf{X}\mathbf{W} &= \mathbf{T}\mathbf{P}^T\mathbf{W} + \mathbf{E}\mathbf{W} \\ \Leftrightarrow \mathbf{T} &= \mathbf{X}\mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}. \end{aligned} \quad (6.14)$$

Replacing (6.14) on (6.13) enables to obtain the PLS regression for training data:

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X}\mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{c}^T \\ &= \mathbf{X}\mathbf{X}^T\mathbf{R}(\mathbf{T}^T\mathbf{X}\mathbf{X}^T\mathbf{R})^{-1}\mathbf{c}^T \\ &= \mathbf{X}\mathbf{d} \end{aligned} \quad (6.15)$$

where

$$\mathbf{d} = \mathbf{X}^T\mathbf{R}(\mathbf{T}^T\mathbf{X}\mathbf{X}^T\mathbf{R})^{-1}\mathbf{c}^T \quad (6.16)$$

is the parameter vector of the PLSR regression model. Since the solution of (6.12) by linear least squares, with orthogonal latent predictors \mathbf{T} , leads to $\mathbf{c} = \mathbf{y}^T \mathbf{T}$, the parameter vector (6.16) can be written as

$$\mathbf{d} = \mathbf{X}^T \mathbf{R} (\mathbf{T}^T \mathbf{X} \mathbf{X}^T \mathbf{R})^{-1} \mathbf{T}^T \mathbf{y} = \mathbf{T} \mathbf{T}^T \mathbf{y}. \quad (6.17)$$

For the reconstruction/prediction period, the PLS regression model will be given by

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}} \mathbf{d} = \tilde{\mathbf{T}} \mathbf{c}^T = \tilde{\mathbf{T}} \mathbf{T}^T \mathbf{y}, \quad (6.18)$$

where $\tilde{\mathbf{X}} \in \mathbb{R}^{(n-m) \times q}$ is the matrix of the predictor variables along the reconstructions/prediction period and

$$\tilde{\mathbf{T}} = \tilde{\mathbf{X}} \mathbf{X}^T \mathbf{R} (\mathbf{T}^T \mathbf{X} \mathbf{X}^T \mathbf{R})^{-1} \in \mathbb{R}^{(n-m) \times p} \quad (6.19)$$

represents the matrix of latent variables in the same period. Other formulations of the PLSR model can be derived based on the properties and identities between the vectors resulting from the algorithm used to obtain the decompositions (6.9) (see, for instance, [92], [95], [96]).

As Equation (6.19) enables to extend the latent variables along the prediction period ($\tilde{\mathbf{T}} = [\tilde{\mathbf{t}}_1 \ \tilde{\mathbf{t}}_2 \ \dots \ \tilde{\mathbf{t}}_p]$), it is possible to use them as predictors in the AnEn-based method. In this work, it is also exploited the combination of the AnEn and ClustAnEn methods with the PLS decomposition, in order to use the latent variables as predictors instead of the original variables. The resulting methods are denoted PLSAnEn and PLSClustAnEn, respectively.

The PLS regression method is also, by itself, an alternative method to PCR and AnEn based methods for the reconstruction/prediction of the missing records, by estimating them via the regression model (6.18) and, therefore, is also included in the present study.

6.3 Selecting the Principal Components

This section shows how the selection of the principal components (PCs) that are used in the PCAnEn, PCClustAnEn and PCR methods is done. The number of PCs to be included

in these methods is of great importance. An insufficient number of PCs translates into the loss of information necessary for data reconstruction, whilst a high number translates into redundant information and increased computational costs.

The identification of the dimensions with most data dispersion enables to identify the principal components \mathbf{z}_j , with $j = 1, \dots, p$, that best distinguish the dataset under study. The dataset corresponding to the multiple predictor time series, from the various meteorological stations described in the Section 6.1, is represented by the data matrix $\mathbf{X} = [\mathbf{x}_1 \mathbf{x}_2 \dots \mathbf{x}_q]$, where each column vector \mathbf{x}_j , with $j = 1, \dots, q$, includes the centred and scaled records of a single variable. Then, the thin singular value decomposition of \mathbf{X} enables to obtain the principal components \mathbf{z}_j , $j = 1, \dots, q$, each one corresponding to a singular value σ_j , $j = 1, \dots, q$, where $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_q$ (see Section 3.2.1). The PC vector \mathbf{z}_1 has the largest sample variance (σ_1^2/m), \mathbf{z}_2 has the second largest variance (σ_2^2/m) and so on. On the other hand, as the original predictor time series are previously scaled by the respective standard deviation, if a PC has a standard deviation greater than 1 it means that this PC defines a dimension with more dispersion, i.e., it contains more information than the original variables. This will be the criteria used to select the PCs, as employed in [12].

Figure 6.6 shows the standard deviations of the first 10 PCs obtained from 37 predictor variables of the data set used. These predictor variables do not include the time series from the PPX station because these series are used only as predicted/reconstructed variables.

Based on the same figure, the principal component analysis is performed from different predictor matrices \mathbf{X} : in the first case (blue bars) the PCs are obtained from all predictor variables ($q = 37$); in the second case (orange bars) the PCs are obtained by means of the predictor variables WSPD and GST ($q = 18$); in the third case (green bars), only the ATMP time series were used ($q = 11$); finally, in the fourth case (grey bars), only the PRES series were considered as predictors ($q = 8$). The idea behind this analysis is to verify if there is any advantage in using predictor variables different from the predicted ones. The variables WSPD and GST were merged as they are highly correlated with each other, hence it makes no sense to use them separately.

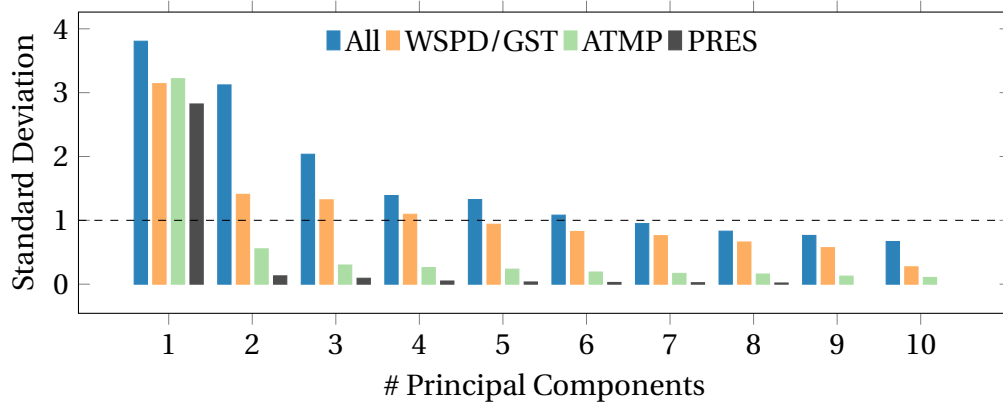


Figure 6.6: Standard deviation of the first PCs for different predictor variables.

Table 6.2: Errors of the prediction of PPX station for different principal components.

Predicted	Predictor	# PCs (p)	PCR			PCClustAnEn			PCAnEn		
			BIAS	RMSE	SDE	BIAS	RMSE	SDE	BIAS	RMSE	SDE
WSPD	All	6	0.47	1.99	1.93	0.55	1.91	1.83	0.55	1.8	1.71
	WSPD/GST	4	0.39	1.85	1.81	0.38	1.79	1.75	0.42	1.75	1.7
ATMP	All	6	-0.01	0.54	0.54	-0.05	1.00	1.00	-0.07	0.8	0.8
	ATMP	1	-0.04	0.64	0.64	-0.03	0.66	0.66	-0.03	0.66	0.66
PRES	All	6	0.04	0.29	0.29	0.08	1.59	1.58	0.05	0.97	0.97
	PRES	1	0.04	0.20	0.19	0.04	0.32	0.31	0.04	0.32	0.32

As can be seen in Figure 6.6, where the PC choice threshold is indicated by a dashed horizontal line, if all variables are used as predictors then the first six PCs ($p = 6$) must be selected to represent the predictor data set. In the case of wind-related predictor variables (WSPD and GST), the first four PCs ($p = 4$) must be used. Finally, for a single predictor variable (ATMP or PRES), only the first PC ($p = 1$) should be chosen.

Table 6.2 shows the errors for the prediction/reconstruction of the variables WSPD, ATMP and PRES, of the PPX station, using the PCR, PCClustAnEn and PCAnEn methods, with the number of PCs previously defined (6, 4 or 1). The reconstruction period was the year of 2021 (the last one of the data set) and the training period spanned from 2016 to 2020. Due to computing resources constraints, the daily records were reconstructed only from 10 am to 4 pm; also, the analogs were searched only in the same period.

As may be observed in Table 6.2, with the exception of the ATMP variable, there seems to be no advantage in using all available predictor variables and, in most cases, it is preferable

to use as predictor the same variable to be predicted/reconstructed. Moreover, comparing the results obtained by the three methods, they result in errors very close to each other. At most, errors are in the order of tenths of a unit. Noteworthy, the consistent spatial correlation of ATMP and PRES, as seen in Figures 6.4 and 6.5, gives PCR a slight advantage in their reconstruction, which suggests that a regression model uses this correlation more effectively. In contrast, WSPD has lower spatial correlation and more frequent temporal variations, making PCAnEn the best method for its reconstruction. Finally, PCClusAnEn yields comparable or nearly identical results to PCAnEn, for all the three variables.

6.4 Selecting the Latent Variables

Choosing the number of components (latent variables) is an important step in the application of the PLSR method or variants. As a latent variable is relevant only if it improves the prediction of \mathbf{y} , it is firstly necessary to solve the problem of which and how many latent variables should be kept in the PLSR model to achieve optimal predictions.

This section proposes two approaches that can be used to determine the number of latent variables (p). To achieve this, the variables of the PPX station were predicted/reconstructed by means of the variables of the neighbouring stations. Thus, there are $q = 37$ original predictor variables that can be used to obtain latent variables.

The performance of a PLSR model can be evaluated with computer-based re-sampling techniques such as cross-validation (c.f. [97]). In this technique, the data of the training period (see Figure 3.4) are split into a *learning* set (used to build a PLSR model) and *testing* set (used to test the model). In particular, in the Cross-validation (CV) approach, the initial training data set is partitioned into exactly k subsets (k -Fold). In turn, each subset is then used to test the PLSR model built by means of the data included in the $k - 1$ learning subsets (for details refer to [92] and [98]). The predicted observations for each testing set are stored in the vector $\hat{\mathbf{y}}^{[p]}$, which is used to determine the overall quality of the PLSR model using p latent variables. The quality of the PLSR model is evaluated by measuring the similarity between \mathbf{y} and $\hat{\mathbf{y}}^{[p]}$. This similarity can be provided by the Mean-Squared Error Predicted

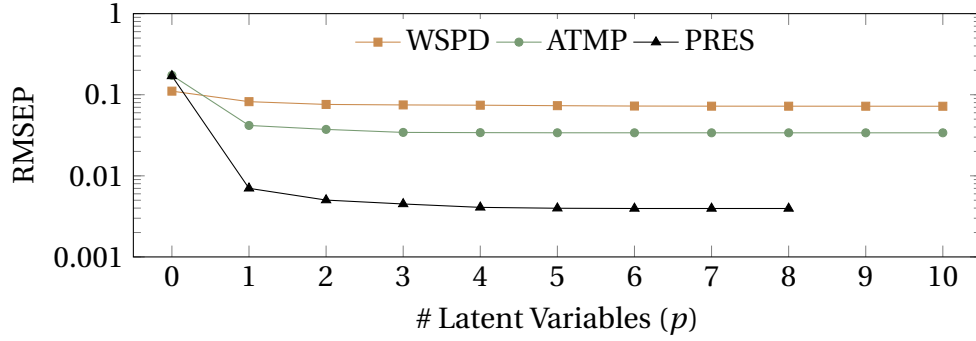


Figure 6.7: Normalized RMSEP values for different latent variables on a logarithmic scale. The same variable was used in both \mathbf{X} and \mathbf{y} . The normalization factor used was the difference between the maximum and minimum values [99].

(MSEP), as given by

$$\text{MSEP} = \frac{1}{m} \left\| \mathbf{y} - \hat{\mathbf{y}}^{[p]} \right\|^2, \quad (6.20)$$

or by the Root-Mean-Squared Error Predicted (RMSEP), such that $\text{RMSEP} = \sqrt{\text{MSEP}}$.

Figure 6.7 shows the values of the normalized RMSEP in function of the number of latent variables (p), originated by the predictions of the four meteorological variables (WSPD, GST, ATMP and PRES), from the PPX stations, by means of the 10-Fold CV technique. Each variable was predicted only by the time series corresponding to the same meteorological variable (\mathbf{X} and \mathbf{y} contain data from the same meteorological variable), except for WSPD and GST, that are used together to predict WSPD, based on the analysis in Section 6.1. It can be observed, in the case of WSPD, that the first four components are responsible for the highest decrease in RMSEP. For a number of latent variables greater than four, the decrease in the RMSEP is not significant. For the ATMP and PRES meteorological variables, the largest reduction in RMSEP happens for the first three components.

Figure 6.8 shows the values of the RMSEP generated in the same conditions of Figure 6.7, with the exception that each variable was predicted by all meteorological variables in the neighbouring stations. In the case of WSPD, the first two or three latent variables are responsible for the highest decrease in RMSEP. For the case of ATMP and PRES, the largest reduction in RMSEP occurs for the first three or four components.

An alternative approach to determine the optimal number of latent variables, is based on

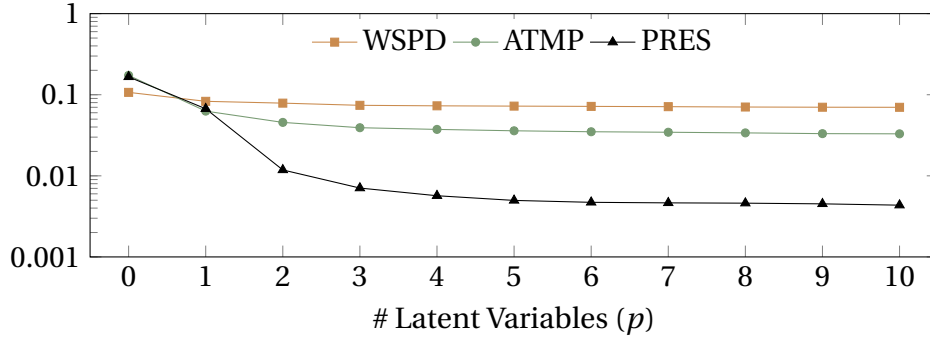


Figure 6.8: Normalized RMSEP values for different latent variables on a logarithmic scale. All variables were included in \mathbf{X} .

the metric

$$Q_p^2 = 1 - \frac{\text{PRESS}_p}{\text{RESS}_{p-1}}, \quad (6.21)$$

where PRESS_p is the predicted residual sum of squares originated by the CV technique, with the p latent variable, being computed through

$$\text{PRESS}_p = \left\| \mathbf{y} - \hat{\mathbf{y}}^{[p]} \right\|^2, \quad (6.22)$$

and

$$\text{RESS}_{p-1} = \left\| \mathbf{y} - \hat{\mathbf{y}}^{[p-1]} \right\|^2 \quad (6.23)$$

is the residual sum of squares originated by the PLSR model, obtained with the $p - 1$ latent variable and built with all data from the training period. The idea of this criteria, proposed in [100], is that a latent variable is kept if the value of the metric (6.21) is larger than a certain threshold (ϵ) generally set to $\epsilon = 0.0975$, i.e.

$$Q_p^2 \geq 0.0975. \quad (6.24)$$

Figure 6.9 shows the values of Q^2 in function of the number of latent variables, in the case of predictions with the same meteorological variables. It can be observed that the number of latent variables that verify the criterion (6.24) is two for the WSPD meteorological variable, three for the ATMP variable, and four for the PRESS variable.

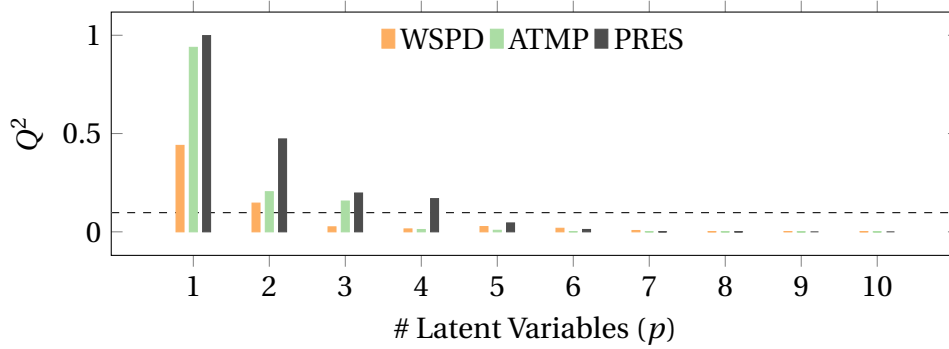


Figure 6.9: Q^2 metric for a different number of latent variables when using the same variable as predictor and predicted.

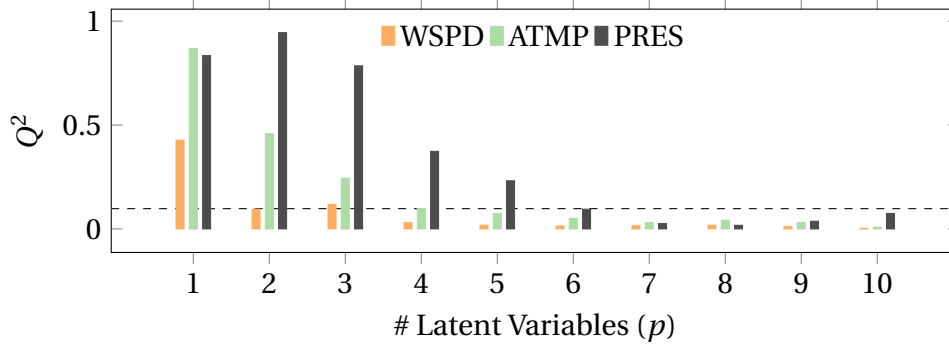


Figure 6.10: Q^2 metric for a different number of latent variables when using all variables as predictor.

Similarly, Figure 6.10 represents the values of Q^2 in function of the number of latent variables, in the case of predictions with all the meteorological variables. It can be observed that the number of latent variables that verify the criterion (6.24) is now three for the variable WSPD, three or four for the ATMP variable, and six for the PRESS variable.

Table 6.3 shows the errors obtained in the prediction of the variables WSPD, ATMP and PRES, of the PPX station, with the PLSAnEn, PLSClustAnEn and PLSR methods, for the cases where a different number of latent variables (LVs) are used as predictors, chosen according to the previously discussed criteria. As in Section 6.3, the prediction/reconstruction period is the year of 2021 (restricted, everyday, to the same period - 10 am to 4 pm) and the remaining years, from 2016 to 2020, make up the training period. For each number (p) of predictor LVs, and each method, the smallest errors are highlighted in bold.

Table 6.3: Errors of the prediction of the PPX station for a different number of latent variables (LVs).

Predicted	Predictor	# LVs	PLSR			PLSClustAnEn			PLSAnEn		
			BIAS	RMSE	SDE	BIAS	RMSE	SDE	BIAS	RMSE	SDE
WSPD	All	2	0.41	1.91	1.86	0.46	1.95	1.89	0.39	1.90	1.85
		3	0.47	1.81	1.75	0.50	1.84	1.77	0.50	1.80	1.73
	WSPD/GST	2	0.44	1.86	1.81	0.43	1.85	1.80	0.35	1.84	1.81
		3	0.41	1.80	1.75	0.41	1.81	1.76	0.37	1.76	1.72
ATMP	All	3	-0.02	0.60	0.60	-0.02	0.82	0.82	-0.03	0.71	0.7
		4	0.02	0.58	0.58	0	0.85	0.85	-0.01	0.7	0.7
	ATMP	3	-0.01	0.53	0.53	-0.04	0.57	0.57	-0.04	0.56	0.56
		4	-0.01	0.53	0.53	-0.03	0.59	0.59	-0.03	0.57	0.57
PRES	All	3	0.03	0.20	0.20	-0.03	1.04	1.04	-0.01	0.75	0.75
		6	0.01	0.13	0.13	-0.01	1.27	1.27	-0.02	0.85	0.85
	PRES	3	0.04	0.12	0.12	0.03	0.33	0.33	0.03	0.31	0.31
		4	0.01	0.12	0.12	0.03	0.31	0.31	0.02	0.31	0.31

In agreement with what was also observed in Section 6.3, the use of all variables as predictors did not bring any advantages, and the errors obtained were smaller for predictor variables corresponding to the same meteorological variables being predicted. It may be also concluded that the increase in the number of LVs does not always translate into a smaller error in the reconstructed/predicted values; for instance, the errors obtained for ATMP were higher with 4 LVs, rather than with 3. The PLSAnEn method showed the best results in the reconstruction/prediction of WSPD. The PLSR method showed smallest errors in the reconstruction/prediction of PRES and ATMP, with results not far to those produced by the PLSAnEn method (for WSPD and ATMP).

6.5 Results

This section examines the errors resulting from the data reconstruction using data from neighboring stations. The results are divided into two subsections: Subsection 6.5.1, featuring the prediction of the station PPX, and Subsection 6.5.2, which focuses on the reconstruction of each station contained in the dataset. As in previous sections, the reconstruction period is the year of 2021 (only the daily period from 10 am to 4 pm) and the remaining

years, from the begin of 2016 to the end of 2020, constitute the training period. For each meteorological variable all reconstruction/prediction methods where applied with the optimal number of PCs or LVs determined in the two previous sections.

Table 6.4: Number of PCs or LVs used by each method.

Predicted Predictor	WSPD	ATMP	PRES
WSPD/GST	All/ATMP	PRES	
# Original Preditors	18	37/11	8
PCR	4	6	1
PCClustAnEn	4	1	1
PCAnEn	4	1	1
PLSR	3	3	4
PLSClustAnEn	3	3	4
PLSAnEn	3	3	4

Table 6.4 resumes the number of PCs or LVs used by each method. As shown in Table 6.2, ATMP reconstruction using the PCR method obtains better results when all meteorological variables are used as predictors. For this reason, this is the only case where the PCs ($p = 6$) are obtained from all $q = 37$ original predictor variables. In the remaining cases, both PCs and LVs are obtained from predictor variables corresponding to the same meteorological variable that is predicted, i.e., PRES is predicted from PRES and WSPD is predicted from WSPD and GST, as shown in Table 6.4.

6.5.1 Reconstruction of Meteorological Variable in one Station

Figure 6.11 allows to compare the real/observed values with the reconstructed/predicted ones, for the three meteorological variables, at the PPX station, in January 9, 2021, from 10 am to 4 pm. Only the values obtained by the methods with the smallest errors in Table 6.2 and Table 6.3 are presented. It can be seen that the WSPD variable varies much more in time than ATMP , and especially in comparison to the PRES variable. Visually, WSPD shows more variance at higher frequencies, i.e., more fluctuations between consecutive records.

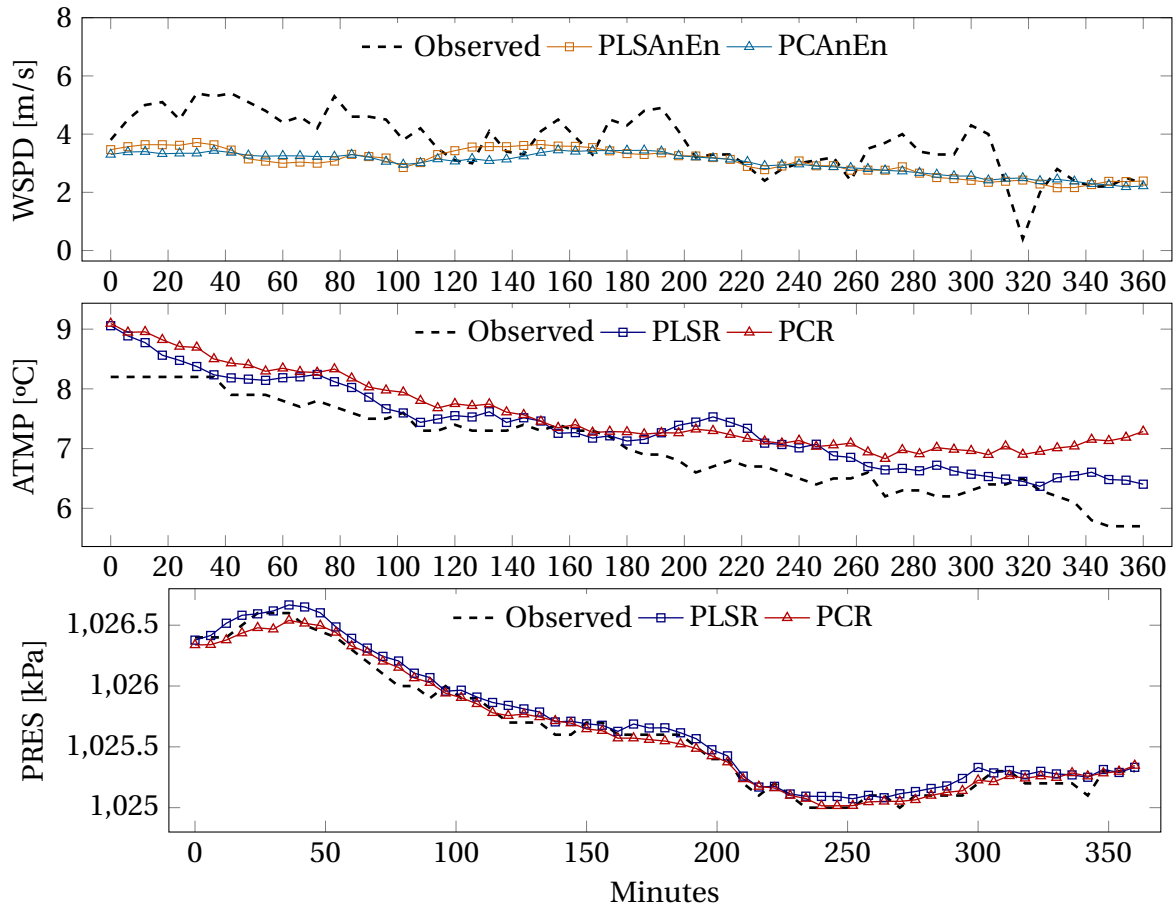


Figure 6.11: Comparison between reconstructed and observed values of the meteorological a) WSPD, b) ATMP and c) PRES variables, from the PPX station, at January 9, 2021, from 10 am to 4pm.

For the WSPD variable, the reconstructed values are not able to reproduce specific variations. However, they reproduce the general trend of variation of the variable. It should be stressed that this meteorological variable undergoes permanent changes over time. These changes are greatly influenced by factors inherent to the location, such as the orientation and topography of the site. It can also be seen that there are no major differences between the values reconstructed by the PCAnEn and PLSAnEn methods.

For the ATMP variable, the values obtained by the PLSR method are closer to the observed values than those produced by the PCR method. Both sets of values show some distance in relation to the observed values. This highlights the difficulty in reconstructing, in a very exact way, the temperatures in a certain place, from the temperatures in other places with different

characteristics, such as sun exposure.

The values reconstructed by the PLSR and PCR methods for PRES are very close to the observed values. This happens because this meteorological variable does not have oscillations due to location, and because it often contains less high-frequency fluctuations. Its variation is carried out on a regional scale without local influence, since the stations are all located at the same altitude (near sea level).

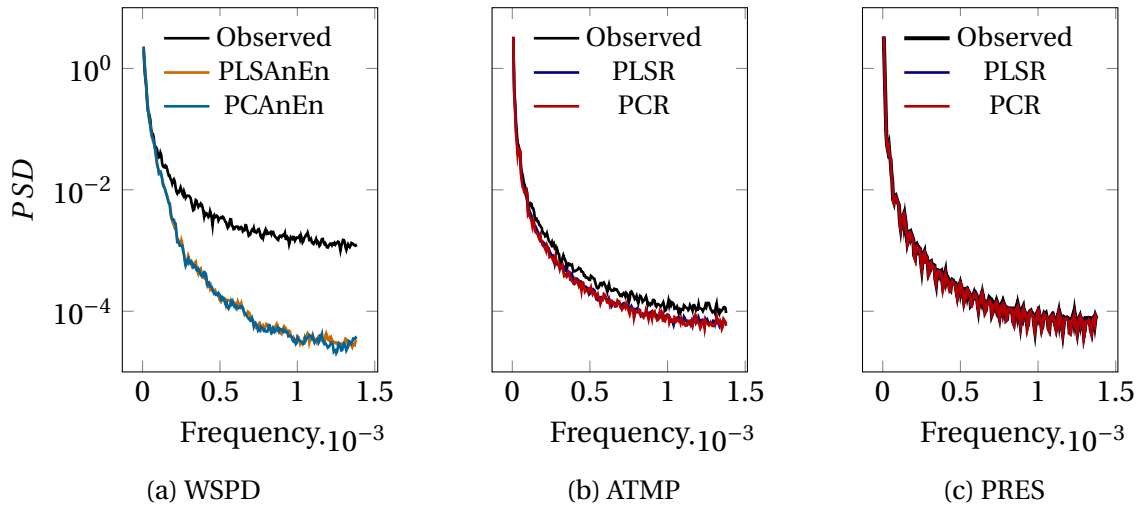


Figure 6.12: Power spectral densities of the reconstructed/observed time series from the PPX station.

Although Figure 6.11 presents a one-day comparison of predicted and observed values, it is insufficient to conclude that these patterns remain constant throughout the entire time series (1 year). Thus, an analysis is necessary to verify the consistency of these patterns in the complete dataset. To evaluate the prediction methods' ability to capture high-frequency patterns, the normalized Power Spectrum Densities (PSD) [101] of both reconstructed and observed series are compared in Figure 6.12. Upon initial observation, the WSPD data display greater high-frequency density than that of ATMP and PRES. Generally, at low frequencies, the methods' densities align with the observed data across all variables. However, for WSPD, the predictions fail to accurately capture the observed high-frequency variance. In contrast, for ATMP and PRES, the reconstructions exhibit similar high-frequency variance to the observed data. These results corroborate the previous one-day analysis.

6.5.2 Reconstruction of Meteorological Variable in all Stations

Figure 6.13 shows the values of the RMSE error per each station, resulting from the reconstruction of the WSPD meteorological variable by the different methods. For all methods, it is observed, in general, that the lowest errors are obtained for the stations that occupy central positions in relation to the others. In general, it is also observed that PLS-based methods obtain lower errors than PC-based methods, although the differences are small (on the order of a tenth or a hundredth of a unit). The PLSAnEn method presents the best results in the reconstruction of the WSPD in all meteorological stations.

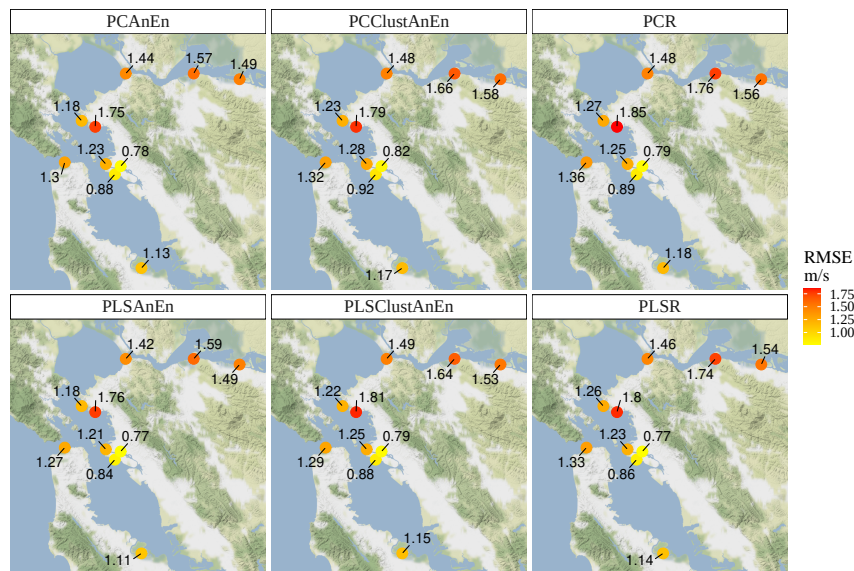


Figure 6.13: RMSE for the reconstruction of WSPD variable across all available stations.

In Figure 6.14 it is possible to observe the values of the same error for the reconstruction of the ATMP variable. In general, the errors obtained are smaller than in the case of WSPD. Here, also, the highest errors are obtained in the most peripheral stations, that have less correlation with the remaining. For this meteorological variable, the supremacy of methods based on the PLS is also verified. The best results at all stations are obtained by the PLSR method, followed closely by the results obtained by the PLSAnEn method.

Finally, Figure 6.15 show also the RMSE errors, this time during the reconstruction of the

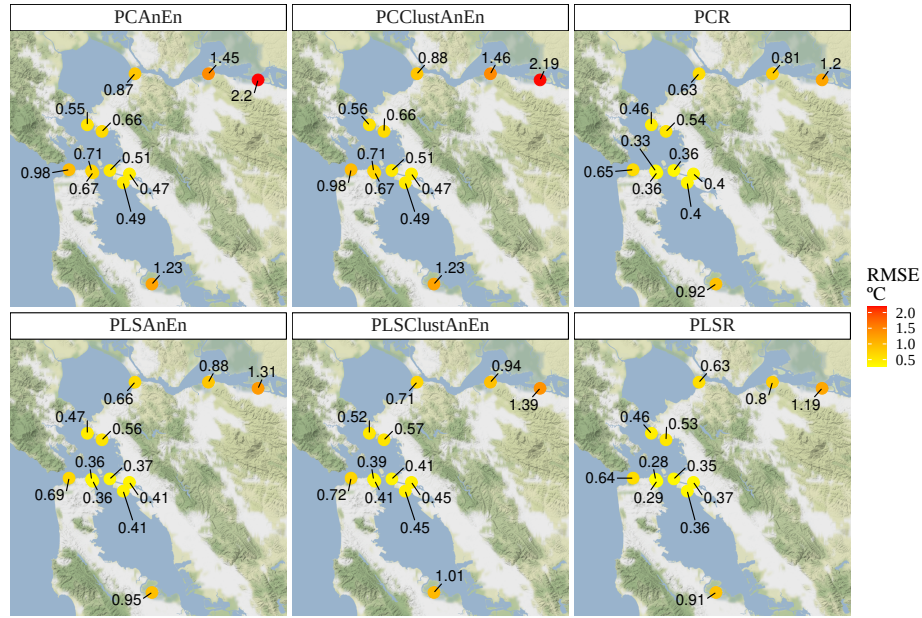


Figure 6.14: RMSE for the reconstruction of ATMP variable across all available stations.

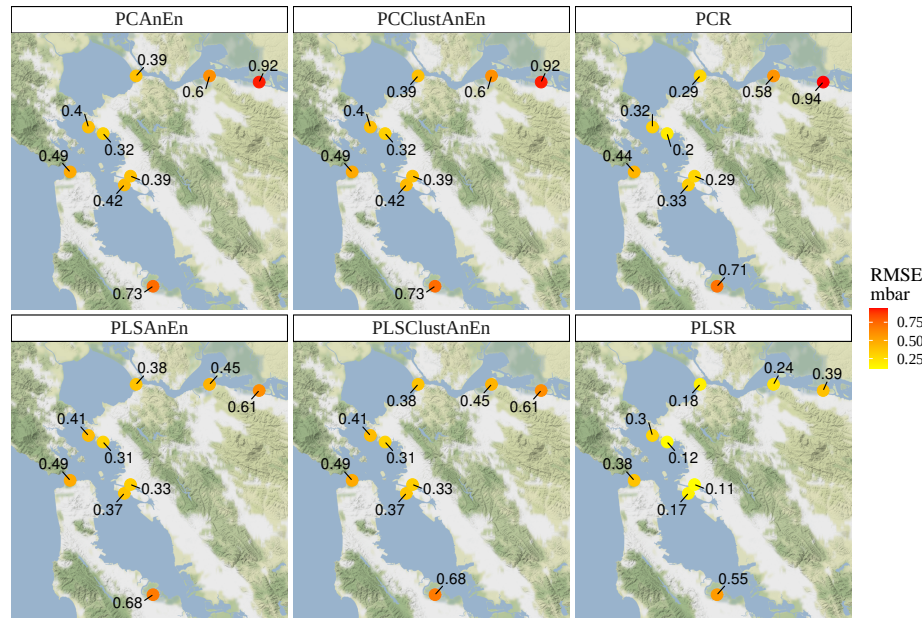


Figure 6.15: RMSE for the reconstruction of PRES variable across all available stations.

meteorological variable PRES. The low error values show that PRES is clearly the easiest variable to reconstruct. In this case, PLS-based methods do not always get the best score, which is achieved by the PLSR method. Following PLSR, the method that obtains the best results

is PCR, especially in the most central stations. The PLSAnEn and PLSClustAnEn methods obtain the same results at all stations.

6.6 Computational Performance

This section analyses the computational performance of the reconstruction methods considered in this study. The computational system used for the evaluation was the same as described in Section 4.3.

Table 6.5 presents the mean execution times, in seconds, needed by each method in the reconstruction of the meteorological variables for all the different stations analyzed in the previous section (each execution time presented in the table corresponds to the average of the times of all stations). These execution times concern the execution in a parallel regime (by instructing the R platform to exploit, whenever possible, all the available CPU cores).

Table 6.5: Mean execution times in seconds across methods, variables, and steps.

Variable	Steps	PCAnEn	PLSAnEn	PCClustAnEn	PLSClustAnEn	PCR	PLSR
WSPD	Loading	37.2	37.2	37.2	37.2	37.2	37.2
	PCA/PLS	0.5	0.7	0.5	0.7	3.7	3.4
	Prediction	341.4	289.9	13.2	11.8	0	0
ATMP	Loading	24.1	24.1	24.1	24.1	61.2	24.1
	PCA/PLS	0.3	0.4	0.3	0.4	6.8	2.1
	Prediction	72.1	287.1	4.2	6.0	0	0
PRES	Loading	24.6	24.6	24.6	24.6	24.6	24.6
	PCA/PLS	0.3	0.3	0.3	0.3	2.0	1.7
	Prediction	72.1	340.3	5.5	7.1	0	0

The execution times are broken down into three consecutive stages: Loading, Decomposition (PCA or PLS), and Prediction/Reconstruction. The Loading stage corresponds to the reading of the data set from CSV files and interpolation of missing values (when they are not more than four consecutive values). In the case of the reconstruction of WSPD, the Loading step involves the reading of both the WSPD and GST dat set files, which takes ≈ 37.2 seconds; even longer, the loading of the ATMP data set, when using the PCR method, takes ≈ 61.2 seconds because this method uses all 37 predictor variables (recall Table 6.4).

The PCA or PLS decomposition step corresponds to the calculation of PCs or LVs, respectively. This step is done by internal R functions that are already highly optimized. As such, the execution time of this step is very fast (although, this step can include also the choice of the number of components through the CV method, which may have some computational costs).

The final step consists in the Reconstruction/Prediction of the missing values. In the PCR and PLSR methods, this stage executes very fast, through linear regression with previously determined PCs or LVs. But for the PCAnEn and PLSAnEn, this stage is very demanding because, for each value to be reconstructed/predicted, the training period is swept in search for analogs. In turn, for the PCClustAnEn and PLSClustAnEn methods, this step is not as demanding, once all possible analogs are previously clustered, and the sweeps are reduced to a single operation in which the predictor value is compared with the cluster centroid. Overall, the slowest methods are thus the PCAnEn method (WSPD) or the PLSAnEn method (ATMP and PRES) and the faster (for all variables) is the PLSR method.

The execution times presented in Table 6.5 were obtained using all CPU cores available. The impact of using a varying number of CPU cores may be apprehended by inspecting Figure 6.16. This figure represents the execution time for the reconstruction of WSPD in station PPX in function of the number of CPU cores, without including the loading time (more IO sensitive). In these experiments PLSR and PCR cross-validation were parallelized to accelerate the 10-fold cross-validation process. It can be seen that the PCR and PLSR methods perform similarly for any number of CPUs. In turn, the PLSClustAnEn and PCClustAnEn methods benefit from increasing the number of CPU cores, especially up to 6/8; however, as these methods depend heavily on the clustering phase of possible analogs, which is not always performed in the same number of iterations, their performance does not always improve with the increased calculation capacity.

Regarding the PCAnEn and PLSAnEn methods, it may be observed that they are more sensitive to the increase of the CPU cores employed, with their computational efficiency improving when using up to about ≈ 10 cores. It turns out that these methods are highly parallelizable: many searches for analogs may be carried out simultaneously once they are

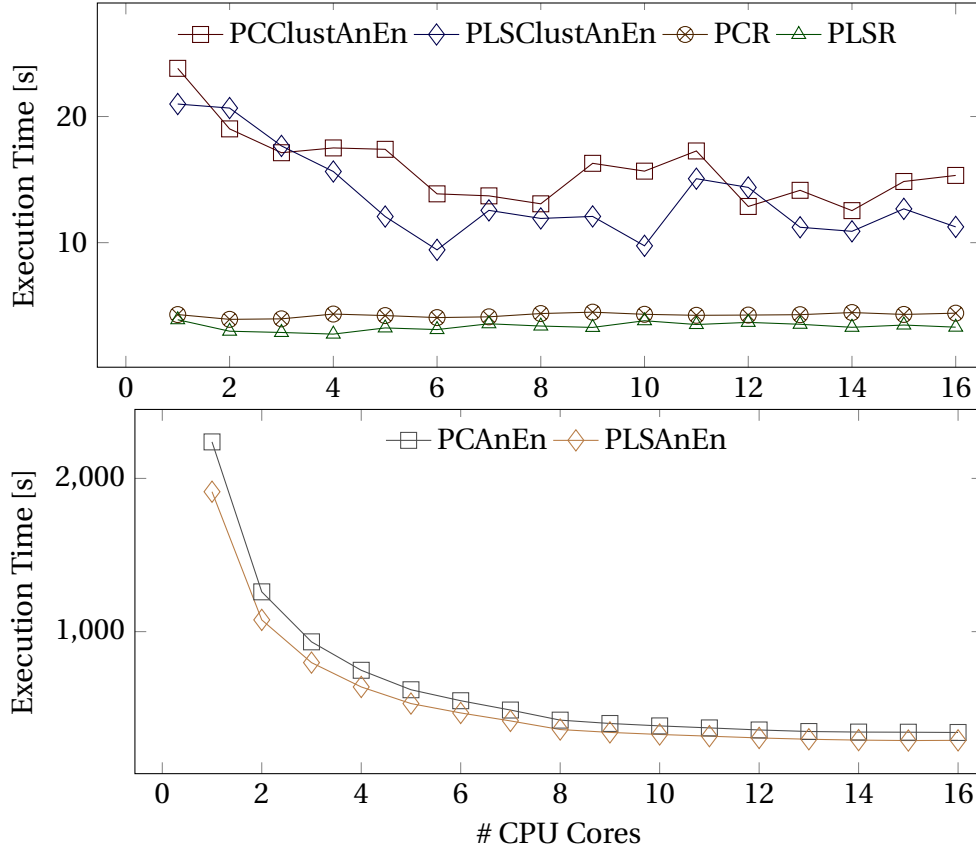


Figure 6.16: CPU time for the reconstruction of WSPD in station PPX in function of the number of cores (excluding loading time).

inherently independent from each other. However, despite the performance gains brought by the parallel execution, the PCAnEn and PLSAnEn methods are still considerably slower than the others.

6.7 Final Remarks

This last study presents methods that address hindcasting and forecasting problems with a high number of predictors. Solving these problems using the classical Analog Ensembles methodology can be computationally inefficient without dimensionality reduction techniques, due to the computational load involved.

The methods presented here combine the robustness of the AnEn method (with or without Clustering) and the PCA and PLS techniques for dimension reduction of the predictor data set. Using these techniques, the predictor variables are reduced to a small number of new variables that mostly retain (PCA) and may even enhance (PLS) the meteorological information used by the AnEn method to reconstruct or forecast the records sought.

The results produced by the PLS-based techniques were found to be slightly more accurate than those obtained with the PCA-based ones, especially in the reconstruction or forecast of meteorological variables with a lot of oscillation, such as wind speed (WSPD). This happens because PLS builds the latent variables in such a way that they simultaneously explain the variation of the predictor variables and the predicted variable, while the main components, obtained by PCA, only explain the variation of the predictor variables.

The combination of the AnEn method with PLS results in a hybrid method, PLSAnEn, that is very accurate in the reconstruction or prediction of wind speed. It is therefore a highly suitable forecasting method for meteorological time series with potential applications in wind-resource assessment and wind energy. At the same time, PLSAnEn is very demanding from the computational point of view, that benefits from a parallel implementation.

PLSclustAnEn, which combines the AnEn methods with the prior clustering of analogs, could be an alternative to the PLSAnEn method, as it is much more computationally efficient (also seen in Chapter 5). It is, however, a method that depends on many parameters that have to be properly chosen in order to improve the accuracy of the results.

Simultaneously with the AnEn-based methods, regression methods were also tested on the new variables determined by PCA or PLS. The resulting methods, PCR and PLSR, are very fast and allow for very accurate reconstructions. In particular, the PLSR method is the most suitable for reconstructing or forecasting highly correlated predictor variables.

Chapter 7

Conclusion

In this dissertation, the challenges identified in Section 2.4.2 were addressed, specifically the issues of missing or absent weather data and handling large volumes of data in the context of weather prediction.

The integration of PCA and PLS techniques was investigated with the AnEn method, as well as its clustered version, for reconstructing missing meteorological data. By employing three datasets and conducting comparative analyses with multivariate regression techniques, such as PLSR and PCR, the findings underline the potential of these hybrid methods for solving hindcasting and forecasting problems with a high number of predictors.

Initially, in Chapter 3, it was exploited the possibility of applying the AnEn method to principal components instead of original variables. This approach proved effective in reconstructing wind-dependent variables and reducing the number of variables requiring processing. Building upon this foundation, Chapter 4 incorporated a more reliable dataset to test the PCAnEn method, leading to improved hindcasting accuracy and reduced processing time. This emphasized the importance of predictor station correlation and the potential performance impact of using different software implementations.

Continuing the research, Chapter 5 examined the PCAnEn method in conjunction with clustering, resulting in the development of PCClustAnEn. This new method maintained the same numerical accuracy as previous approaches but offered significantly faster computational performance.

Finally, in Chapter 6, it was exploited the potential of integrating dimension reduction techniques, PCA and PLS, with the AnEn method and its clustered version for reconstructing missing meteorological data in scenarios with a high number of predictors. Our comparative analyses with multivariate regression techniques, such as PLSR and PCR, showcased their efficiency and accuracy in handling large datasets. It was also observed that AnEn-based methods performed better for wind-related variables, while regression-based methods excelled in atmospheric temperature (ATMP) and pressure (PRES) variables, which are often highly correlated and exhibit fewer high-frequency signals. These methods can be further fine-tuned and adapted to various meteorological applications, offering more reliable and computationally efficient solutions to address hindcasting challenges.

This dissertation has contributed to a deeper understanding of the benefits of combining PCA and PLS techniques with AnEn methods and their clustered variants for reconstructing missing meteorological data. These achievements may address the issue of missing or absent historical weather data, providing valuable insights for various sectors of society. This research has demonstrated the potential of these integrations for solving hindcasting involving a large number of predictors. Future research could focus on further optimizing these methods for more efficient computation and exploring their applications in domains beyond meteorology.

Bibliography

- [1] E. Acuña and C. Rodriguez, “The treatment of missing values and its effect on classifier accuracy,” D. Banks, F. R. McMorris, P. Arabie, and W. Gaul, Eds., pp. 639–647, 2004.
- [2] M. Hasanpour Kashani and Y. Dinpashoh, “Evaluation of efficiency of different estimation methods for missing climatological data,” *Stochastic environmental research and risk assessment*, vol. 26, no. 1, pp. 59–71, 2012.
- [3] L. D. Monache, T. Nipen, Y. Liu, G. Roux, and R. Stull, “Kalman filter and analog schemes to postprocess numerical weather predictions,” *Monthly Weather Review*, vol. 139, no. 11, pp. 3554–3570, 2011. DOI: 10.1175/2011mwr3653.1.
- [4] L. D. Monache, F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight, “Probabilistic Weather Prediction with an Analog Ensemble,” *Monthly Weather Review*, vol. 141, no. 10, pp. 3498–3516, 2013. DOI: 10.1175/mwr-d-12-00281.1.
- [5] S. Alessandrini, L. D. Monache, S. Sperati, and J. N. Nissen, “A novel application of an analog ensemble for short-term wind power forecasting,” *Renewable Energy*, vol. 76, pp. 768–781, 2015. DOI: 10.1016/j.renene.2014.11.061.
- [6] G. Cervone, L. Clemente-Harding, S. Alessandrini, and L. Delle Monache, “Short-term photovoltaic power forecasting using artificial neural networks and an analog ensemble,” *Renewable Energy*, vol. 108, pp. 274–286, 2017, ISSN: 0960-1481. DOI: <https://doi.org/10.1016/j.renene.2017.02.052>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0960148117301386>.

- [7] C. Balsa, C. V. Rodrigues, L. Araújo, and J. Rufino, "Hindcasting with cluster-based analogues," in *Communications in Computer and Information Science*, Springer International Publishing, 2021, pp. 346–360. DOI: 10.1007/978-3-030-90241-4_27.
- [8] D. Yang, "Ultra-fast analog ensemble using kd-tree," *Journal of Renewable and Sustainable Energy*, vol. 11, no. 5, p. 053 703, Sep. 2019. DOI: 10.1063/1.5124711.
- [9] T. Klemm and R. A. McPherson, "The development of seasonal climate forecasting for agricultural producers," *Agricultural and Forest Meteorology*, vol. 232, pp. 384–399, 2017, ISSN: 0168-1923. DOI: <https://doi.org/10.1016/j.agrformet.2016.09.005>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0168192316303847>.
- [10] K. B. Murray, F. Di Muro, A. Finn, and P. Popkowski Leszczyc, "The effect of weather on consumer spending," *Journal of Retailing and Consumer Services*, vol. 17, no. 6, pp. 512–520, 2010, ISSN: 0969-6989. DOI: <https://doi.org/10.1016/j.jretconser.2010.08.006>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0969698910000822>.
- [11] L. Symeonidis, G. Daskalakis, and R. N. Markellos, "Does the weather affect stock market volatility?" *Finance Research Letters*, vol. 7, no. 4, pp. 214–223, 2010, ISSN: 1544-6123. DOI: <https://doi.org/10.1016/j.frl.2010.05.004>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1544612310000371>.
- [12] F. Davò, S. Alessandrini, S. Sperati, L. D. Monache, D. Airolidi, and M. T. Vespucci, "Post-processing techniques and principal component analysis for regional wind power and solar irradiance forecasting," *Solar Energy*, vol. 134, pp. 327–338, Sep. 2016. DOI: 10.1016/j.solener.2016.04.049.
- [13] B. Yang, T. Yu, H. Shu, *et al.*, "Passivity-based sliding-mode control design for optimal power extraction of a pmsg based variable speed wind turbine," *Renewable Energy*, vol. 119, pp. 577–589, 2018, ISSN: 0960-1481. DOI: <https://doi.org/10.1016/j.renene.2017.12.047>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S096014811731251X>.

- [14] B. Yang, L. Zhong, J. Wang, *et al.*, “State-of-the-art one-stop handbook on wind forecasting technologies: An overview of classifications, methodologies, and analysis,” *Journal of Cleaner Production*, vol. 283, p. 124 628, 2021, ISSN: 0959-6526. DOI: <https://doi.org/10.1016/j.jclepro.2020.124628>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0959652620346722>.
- [15] A. M. Update, “Global wind report,” *Global Wind Energy Council*, 2017.
- [16] D. Bogdanov, M. Ram, A. Aghahosseini, *et al.*, “Low-cost renewable electricity as the key driver of the global energy transition towards sustainability,” *Energy*, vol. 227, p. 120 467, 2021, ISSN: 0360-5442. DOI: <https://doi.org/10.1016/j.energy.2021.120467>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0360544221007167>.
- [17] Y. L. Zhukovskiy, D. E. Batueva, A. D. Buldysko, B. Gil, and V. V. Starshaia, “Fossil energy in the framework of sustainable development: Analysis of prospects and development of forecast scenarios,” *Energies*, vol. 14, no. 17, 2021, ISSN: 1996-1073. DOI: [10.3390/en14175268](https://doi.org/10.3390/en14175268). [Online]. Available: <https://www.mdpi.com/1996-1073/14/17/5268>.
- [18] “Review of seasonal climate forecasting for agriculture in sub- saharan africa,” *Experimental Agriculture*, vol. 47, no. 02, pp. 205–240, 2011. DOI: [10.1017/S0014479710000876](https://doi.org/10.1017/S0014479710000876).
- [19] P. Calanca, “Weather forecasting applications in agriculture,” in *Encyclopedia of Agriculture and Food Systems*, N. K. Van Alfen, Ed., Oxford: Academic Press, 2014, pp. 437–449, ISBN: 978-0-08-093139-5. DOI: <https://doi.org/10.1016/B978-0-444-52512-3.00234-5>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/B9780444525123002345>.
- [20] “The value of climate information when farm programs matter,” *Agricultural Systems*, vol. 93, no. 1, pp. 25–42, 2007. DOI: [10.1016/J.AGSY.2006.04.005](https://doi.org/10.1016/J.AGSY.2006.04.005).
- [21] J. M. C. Thompson, “The potential economic benefits of improvements in weather forecasting,” 1972.

- [22] D. D. Turner, H. Cutler, M. Shields, *et al.*, “Evaluating the economic impacts of improvements to the high-resolution rapid refresh (hrrr) numerical weather prediction model,” *Bulletin of the American Meteorological Society*, vol. 103, no. 2, E198–E211, 2022. DOI: 10.1175/BAMS-D-20-0099.1. [Online]. Available: <https://journals.ametsoc.org/view/journals/bams/103/2/BAMS-D-20-0099.1.xml>.
- [23] “Sectoral use of climate information in europe: A synoptic overview,” *Climate Services*, vol. 9, pp. 5–20, 2018. DOI: 10.1016/J.CLISER.2017.06.001.
- [24] P. Bauer, A. Thorpe, and G. Brunet, “The quiet revolution of numerical weather prediction,” *Nature*, vol. 525, no. 7567, pp. 47–55, Sep. 2015, ISSN: 1476-4687. DOI: 10.1038/nature14956. [Online]. Available: <https://doi.org/10.1038/nature14956>.
- [25] R. Kimura, “Numerical weather prediction,” *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 90, no. 12, pp. 1403–1414, 2002, Fifth Asia-Pacific Conference on Wind Engineering, ISSN: 0167-6105. DOI: [https://doi.org/10.1016/S0167-6105\(02\)00261-1](https://doi.org/10.1016/S0167-6105(02)00261-1). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167610502002611>.
- [26] C. A. Doswell, “Weather forecasting by humans—heuristics and decision making,” *Weather and Forecasting*, vol. 19, no. 6, pp. 1115–1126, 2004. DOI: 10.1175/WAF-821.1. [Online]. Available: https://journals.ametsoc.org/view/journals/wefo/19/6/waf-821_1.xml.
- [27] L. F. Richardson, *Weather prediction by numerical process*. University Press, 1922.
- [28] F. G. Shuman, “History of numerical weather prediction at the national meteorological center,” *Weather and Forecasting*, vol. 4, no. 3, pp. 286–296, 1989. DOI: 10.1175/1520-0434(1989)004<0286:HONWPA>2.0.CO;2. [Online]. Available: https://journals.ametsoc.org/view/journals/wefo/4/3/1520-0434_1989_004_0286_honwpa_2_0_co_2.xml.
- [29] G. W. Platzman, “A retrospective view of richardson’s book on weather prediction,” *Bulletin of the American Meteorological Society*, vol. 48, no. 8, pp. 514–551, 1967. DOI:

- 10.1175/1520-0477-48.8.514. [Online]. Available: https://journals.ametsoc.org/view/journals/bams/48/8/1520-0477-48_8_514.xml.
- [30] V. R. Durai and S. K. Roy Bhowmik, "Prediction of indian summer monsoon in short to medium range time scale with high resolution global forecast system (gfs) t574 and t382," *Climate Dynamics*, vol. 42, no. 5, pp. 1527–1551, Mar. 2014, ISSN: 1432-0894. DOI: 10.1007/s00382-013-1895-5. [Online]. Available: <https://doi.org/10.1007/s00382-013-1895-5>.
- [31] M. Leutbecher, S.-J. Lock, P. Ollinaho, *et al.*, "Stochastic representations of model uncertainties at ecmwf: State of the art and future vision," *Quarterly Journal of the Royal Meteorological Society*, vol. 143, no. 707, pp. 2315–2339, 2017. DOI: <https://doi.org/10.1002/qj.3094>. eprint: <https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3094>. [Online]. Available: <https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3094>.
- [32] D. Walters, A. J. Baran, I. Boutle, *et al.*, "The met office unified model global atmosphere 7.0/7.1 and jules global land 7.0 configurations," *Geoscientific Model Development*, vol. 12, no. 5, pp. 1909–1963, 2019. DOI: 10.5194/gmd-12-1909-2019. [Online]. Available: <https://gmd.copernicus.org/articles/12/1909/2019/>.
- [33] E. N. Lorenz, "Atmospheric predictability as revealed by naturally occurring analogues," *Journal of Atmospheric Sciences*, vol. 26, no. 4, pp. 636–646, 1969. DOI: 10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2. [Online]. Available: https://journals.ametsoc.org/view/journals/atsc/26/4/1520-0469_1969_26_636_aparbn_2_0_co_2.xml.
- [34] "A new look at weather forecasting through analogues," *Monthly Weather Review*, vol. 117, no. 10, pp. 2230–2247, 1989. DOI: 10.1175/1520-0493(1989)117<2230:ANLAWF>2.0.CO;2.
- [35] D. Burov, D. Giannakis, K. Manohar, and A. Stuart, "Kernel analog forecasting: Multi-scale test problems," *Multiscale Modeling & Simulation*, vol. 19, no. 2, pp. 1011–1040,

2021. DOI: 10.1137/20M1338289. eprint: <https://doi.org/10.1137/20M1338289>.
[Online]. Available: <https://doi.org/10.1137/20M1338289>.
- [36] H. M. van den Dool, "A new look at weather forecasting through analogues," *Monthly Weather Review*, vol. 117, no. 10, pp. 2230–2247, 1989. DOI: 10.1175/1520-0493(1989)117<2230:ANLAWF>2.0.CO;2. [Online]. Available: https://journals.ametsoc.org/view/journals/mwre/117/10/1520-0493_1989_117_2230_anlawf_2_0_co_2.xml.
- [37] E. N. Lorenz, "Atmospheric Predictability as Revealed by Naturally Occurring Analogues," *Journal of the Atmospheric Sciences*, vol. 26, no. 4, pp. 636–646, 1969. DOI: 10.1175/1520-0469(1969)26<636:aparbn>2.0.co;2.
- [38] H. M. V. D. Dool, "A New Look at Weather Forecasting through Analogues," *Monthly Weather Review*, vol. 117, no. 10, pp. 2230–2247, 1989. DOI: 10.1175/1520-0493(1989)117<2230:anlawf>2.0.co;2.
- [39] L. D. Monache, T. Nipen, X. Deng, Y. Zhou, and R. B. Stull, "Ozone ensemble forecasts: 2. a kalman-filter predictor bias correction," *J. Geophys. Res.*, vol. 111, p. D05308, 2006.
- [40] "Gridded probabilistic weather forecasts with an analog ensemble," *Quarterly Journal of the Royal Meteorological Society*, vol. 143, no. 708, pp. 2874–2885, 2017. DOI: 10.1002/QJ.3137.
- [41] "Using the analog ensemble method as a proxy measurement for wind power predictability," *Renewable Energy*, vol. 146, pp. 789–801, 2020. DOI: 10.1016/J.RENENE.2019.06.132.
- [42] "Wind resource estimates with an analog ensemble approach," *Renewable Energy*, vol. 74, pp. 761–773, 2015. DOI: 10.1016/J.RENENE.2014.08.060.
- [43] "Probabilistic prediction of tropical cyclone intensity with an analog ensemble," *Monthly Weather Review*, vol. 146, no. 6, pp. 1723–1744, 2018. DOI: 10.1175/MWR-D-17-0314.1.

- [44] “Performance of the hwrp rapid intensification analog ensemble (hwrp ri-anen) during the 2017 and 2018 hrip real-time demonstrations,” *Weather and Forecasting*, vol. 35, no. 3, pp. 841–856, 2020. DOI: 10.1175/WAF-D-19-0037.1.
- [45] S. Alessandrini, S. Sperati, and L. D. Monache, “Improving the analog ensemble wind speed forecasts for rare events,” *Monthly Weather Review*, vol. 147, no. 7, pp. 2677–2692, 2019. DOI: 10.1175/MWR-D-19-0006.1. [Online]. Available: <https://journals.ametsoc.org/view/journals/mwre/147/7/mwr-d-19-0006.1.xml>.
- [46] “Observation-based analog ensemble solar forecast in coastal california,” 2019. DOI: 10.1109/PVSC40753.2019.8980546.
- [47] “Statistical downscaling of a high-resolution precipitation reanalysis using the analog ensemble method,” *Journal of Applied Meteorology and Climatology*, vol. 56, no. 7, pp. 2081–2095, 2017. DOI: 10.1175/JAMC-D-16-0380.1.
- [48] “Improving air quality predictions over the united states with an analog ensemble,” *Weather and Forecasting*, vol. 35, no. 5, pp. 2145–2162, 2020. DOI: 10.1175/WAF-D-19-0148.1.
- [49] “Analog ensemble data assimilation and a method for constructing analogs with variational autoencoders,” *Quarterly Journal of the Royal Meteorological Society*, vol. 147, no. 734, pp. 139–149, 2021. DOI: 10.1002/QJ.3910.
- [50] “Predictor-weighting strategies for probabilistic wind power forecasting with an analog ensemble,” *Meteorologische Zeitschrift*, vol. 24, no. 4, pp. 361–379, 2015. DOI: 10.1127/METZ/2015/0659.
- [51] “Short-term photovoltaic power forecasting using artificial neural networks and an analog ensemble,” *Renewable Energy*, vol. 108, pp. 274–286, 2017. DOI: 10.1016/J.RENENE.2017.02.052.
- [52] “A hybrid nwp–analog ensemble,” *Monthly Weather Review*, vol. 144, no. 3, pp. 897–911, 2016. DOI: 10.1175/MWR-D-15-0096.1.

- [53] “A comprehensive review of numerical weather prediction models,” *International Journal of Computer Applications*, vol. 74, no. 18, pp. 44–48, 2013. DOI: 10 . 5120 / 12989-0246.
- [54] “Intelligent methods used for obtaining weather derivatives: A review,” vol. 4, no. 6, pp. 144–, 2019. DOI: 10 . 11648/J.EAS . 20190406 . 12.
- [55] N. Kanda, H. Negi, M. S. Rishi, and M. Shekhar, “Performance of various techniques in estimating missing climatological data over snowbound mountainous areas of karako-ram himalaya,” *Meteorological Applications*, vol. 25, no. 3, pp. 337–349, 2018.
- [56] R. Tawn, J. Browell, and I. Dinwoodie, “Missing data in wind farm time series: Properties and effect on forecasts,” *Electric Power Systems Research*, vol. 189, p. 106 640, 2020, ISSN: 0378-7796. DOI: <https://doi.org/10.1016/j.epsr.2020.106640>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0378779620304430>.
- [57] L. Massetti, “Analysis and estimation of the effects of missing values on the calculation of monthly temperature indices,” *Theoretical and Applied Climatology*, vol. 117, no. 3, pp. 511–519, Aug. 2014, ISSN: 1434-4483. DOI: 10 . 1007/s00704-013-1024-8. [Online]. Available: <https://doi.org/10.1007/s00704-013-1024-8>.
- [58] *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. 2013.
- [59] H. Jain and R. Jain, “Big data in weather forecasting: Applications and challenges,” in *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*, 2017, pp. 138–142. DOI: 10 . 1109/ICBDACI . 2017 . 8070824.
- [60] “A methodology for the synthesis of hourly weather data,” *Solar Energy*, vol. 46, no. 2, pp. 109–120, 1991. DOI: 10 . 1016/0038-092X(91)90023-P.
- [61] A. Chesneau, C. Balsa, C. V. Rodrigues, and I. M. Lopes, “Hindcasting with multistations using analog ensembles,” in *CEUR Workshop Proceedings*, CEUR-WS, vol. 2486, 2019, pp. 215–229.

- [62] C. Balsa, C. V. Rodrigues, I. Lopes, and J. Rufino, "Using analog ensembles with alternative metrics for hindcasting with multistations," *ParadigmPlus*, vol. 1, no. 2, pp. 1–17, Jun. 2020. [Online]. Available: <https://journals.ititd.org/index.php/paradigmplus/article/view/11>.
- [63] "Parametric study of the analog ensembles algorithm with clustering methods for hindcasting with multistations," pp. 544–559, 2021. DOI: 10.1007/978-3-030-72651-5_52.
- [64] T. Schneider, "Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values," *Journal of Climate*, vol. 14, no. 5, pp. 853–871, 2001. DOI: 10.1175/1520-0442(2001)014<0853:A0ICDE>2.0.CO;2. [Online]. Available: https://journals.ametsoc.org/view/journals/clim/14/5/1520-0442_2001_014_0853_a0icde_2.0.co_2.xml.
- [65] A. Barrios, G. Trincado, and R. Garreaud, "Alternative approaches for estimating missing climate data: Application to monthly precipitation records in south-central chile," *Forest Ecosystems*, vol. 5, no. 1, pp. 1–10, 2018.
- [66] "Wind speed reconstruction from synoptic pressure patterns using an evolutionary algorithm," *Applied Energy*, vol. 89, no. 1, pp. 347–354, 2012. DOI: 10.1016/J.APENERGY.2011.07.044.
- [67] L. Chen, J. Xu, G. Wang, and Z. Shen, "Comparison of the multiple imputation approaches for imputing rainfall data series and their applications to watershed models," *Journal of Hydrology*, vol. 572, pp. 449–460, 2019, ISSN: 0022-1694. DOI: <https://doi.org/10.1016/j.jhydrol.2019.03.025>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0022169419302653>.
- [68] K. Toride, P. Neluwala, H. Kim, and K. Yoshimura, "Feasibility study of the reconstruction of historical weather with data assimilation," *Monthly Weather Review*, vol. 145, no. 9, pp. 3563–3580, 2017. DOI: 10.1175/MWR-D-16-0288.1. [Online]. Available: <https://journals.ametsoc.org/view/journals/mwre/145/9/mwr-d-16-0288.1.xml>.

- [69] S. Rao and S. Mandal, "Hindcasting of storm waves using neural networks," *Ocean Engineering*, vol. 32, no. 5, pp. 667–684, 2005, ISSN: 0029-8018. DOI: <https://doi.org/10.1016/j.oceaneng.2004.09.003>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0029801804002057>.
- [70] J. Mahjoobi, A. Etemad-Shahidi, and M. Kazeminezhad, "Hindcasting of wave parameters using different soft computing methods," *Applied Ocean Research*, vol. 30, no. 1, pp. 28–36, 2008, ISSN: 0141-1187. DOI: <https://doi.org/10.1016/j.apor.2008.03.002>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S014111870800014X>.
- [71] I. Malekmohamadi, R. Ghiassi, and M. Yazdanpanah, "Wave hindcasting by coupling numerical model and artificial neural networks," *Ocean Engineering*, vol. 35, no. 3, pp. 417–425, 2008, ISSN: 0029-8018. DOI: <https://doi.org/10.1016/j.oceaneng.2007.09.003>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0029801807002107>.
- [72] L. O. Guth, "Estudo paramétrico do método dos conjuntos análogos aplicado a dados meteorológicos," M.S. thesis, IPB, 2021.
- [73] J. Estévez, P. Gavilán, and A. Garcíea-Marién, "Spatial regression test for ensuring temperature data quality in southern spain," *Theoretical and applied climatology*, vol. 131, no. 1, pp. 309–318, 2018.
- [74] J. A. Bellido-Jiménez, J. E. Gualda, and A. P. Garcíea-Marién, "Assessing machine learning models for gap filling daily rainfall series in a semiarid region of spain," *Atmosphere*, vol. 12, no. 9, p. 1158, 2021.
- [75] E. Namey, G. Guest, L. N. Thairu, and L. Johnson, "Data reduction techniques for large qualitative data sets," 2007.
- [76] D. Mercer, "Clustering large datasets," 2003.

- [77] S. Meech, S. Alessandrini, W. Chapman, and L. Delle Monache, "Post-processing rain-fall in a high-resolution simulation of the 1994 piedmont flood," *Bulletin of Atmospheric Science and Technology*, vol. 1, no. 3, pp. 373–385, Dec. 2020, ISSN: 2662-1509. DOI: 10.1007/s42865-020-00028-z. [Online]. Available: <https://doi.org/10.1007/s42865-020-00028-z>.
- [78] C. M. Rozoff and S. Alessandrini, "A comparison between analog ensemble and convolutional neural network empirical-statistical downscaling techniques for reconstructing high-resolution near-surface wind," *Energies*, vol. 15, no. 5, 2022, ISSN: 1996-1073. DOI: 10.3390/en15051718. [Online]. Available: <https://www.mdpi.com/1996-1073/15/5/1718>.
- [79] S. Alessandrini, "Predicting rare events of solar power production with the analog ensemble," *Solar Energy*, vol. 231, pp. 72–77, 2022, ISSN: 0038-092X. DOI: <https://doi.org/10.1016/j.solener.2021.11.033>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0038092X21009920>.
- [80] S. Vannitsem, J. B. Bremnes, J. Demaeyer, *et al.*, "Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world," *Bulletin of the American Meteorological Society*, vol. 102, no. 3, E681–E699, Mar. 2021. DOI: 10.1175/bams-d-19-0308.1.
- [81] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. Springer-Verlag, 2002. DOI: 10.1007/springerreference_205537.
- [82] L. Eldén, *Matrix methods in data mining and pattern recognition*. Philadelphia, PA, USA: SIAM, 2007.
- [83] L. Spence, A. Insel, and S. Friedberg, *Elementary Linear Algebra: A matrix Approach*. Pearson Education Limited, Jul. 2013, 632 pp., ISBN: 1292025034.
- [84] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, 2014. DOI: 10.5194/gmd-7-1247-2014.

- [85] C. Balsa, C. V. Rodrigues, L. Araújo, and J. Rufino, “Cluster-based analogue ensembles for hindcasting with multistations,” *Computation*, vol. 10, no. 6, p. 91, Jun. 2022. DOI: 10.3390/computation10060091.
- [86] National Oceanic and Atmospheric Administration, National Weather Service, *National Data Buoy Center (accessed: 28.07.2022)*. [Online]. Available: <https://www.ndbc.noaa.gov>, (accessed: 28.07.2022).
- [87] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: <https://www.R-project.org/>.
- [88] MATLAB, *version 9.10.0.1602886 (R2021a)*. Natick, Massachusetts: The MathWorks Inc., 2021.
- [89] L. Araújo, C. Balsa, C. V. Rodrigues, and J. Rufino, “Parametric Study of the Analog Ensembles Algorithm with Clustering Methods for Hindcasting with Multistations,” in *Advances in Intelligent Systems and Computing*, Springer International Publishing, 2021, pp. 544–559. DOI: 10.1007/978-3-030-72651-5_52.
- [90] W. F. Massy, “Principal components regression in exploratory statistical research,” *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 234–256, Mar. 1965. DOI: 10.1080/01621459.1965.10480787.
- [91] H. Abdi, “Partial least squares regression and projection on latent structure regression (PLS regression),” *WIREs Computational Statistics*, vol. 2, no. 1, pp. 97–106, Jan. 2010. DOI: 10.1002/wics.51.
- [92] B. Mevik and R. Wehrens, “The pls package: Principal component and partial least squares regression in r,” *Journal of Statistical Software*, vol. 18, no. 2, 2007. DOI: 10.18637/jss.v018.i02.
- [93] H. Wold, “Nonlinear iterative partial least squares (NIPALS) modelling: Some current developments,” in *Multivariate Analysis-III*, Elsevier, 1973, pp. 383–407. DOI: 10.1016/b978-0-12-426653-7.50032-6.

- [94] S. Jong, "SIMPLS: An alternative approach to partial least squares regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 18, no. 3, pp. 251–263, Mar. 1993. DOI: 10.1016/0169-7439(93)85002-x.
- [95] X.-D. Zhang, *A Matrix Algebra Approach to Artificial Intelligence*. Springer Singapore, 2020. DOI: 10.1007/978-981-15-2770-8.
- [96] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Subspace, Latent Structure and Feature Selection*, Springer Berlin Heidelberg, 2006, pp. 34–51. DOI: 10.1007/11752790_2.
- [97] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 36, no. 2, pp. 111–133, Jan. 1974. DOI: 10.1111/j.2517-6161.1974.tb00994.x.
- [98] P. A. Lachenbruch and M. R. Mickey, "Estimation of error rates in discriminant analysis," *Technometrics*, vol. 10, no. 1, pp. 1–11, 1968. DOI: 10.1080/00401706.1968.10490530.
- [99] M. V. Shcherbakov, A. Brebels, N. L. Shcherbakova, A. P. Tyukov, T. A. Janovsky, V. A. Kamaev, *et al.*, "A survey of forecast error measures," *World applied sciences journal*, vol. 24, no. 24, pp. 171–176, 2013.
- [100] M. Tenenhaus, *La Regression PLS - Théorie et Pratique*. Paris: Editions Technip, 1998, ISBN: 2-7108-0735-1.
- [101] A. J. Barbour and R. L. Parker, "Psd: Adaptive, sine multitaper power spectral density estimation for r," *Computers and Geosciences*, vol. 63, pp. 1–8, 2014, ISSN: 0098-3004. DOI: <https://doi.org/10.1016/j.cageo.2013.09.015>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0098300413002550>.

Appendix A

Publications

This thesis is based on a collection of articles/abstracts that were published during the master's degree. The articles/abstracts included in this appendix are:

1. C. Balsa, M. M. Breve, C. V. Rodrigues, L. S. Costa, and J. Rufino, "An Exploratory Study on Hindcasting with Analogue Ensembles of Principal Components," *Communications in Computer and Information Science*, vol. 1378, Springer Nature Switzerland, pp. 488–499, 2022. DOI: 10.1007/978-3-031-20319-0_36.
2. C. Balsa, M. M. Breve, B. André, C. V. Rodrigues, and J. Rufino, "PCAnEn - Hindcasting with Analogue Ensembles of Principal Components," *CSEI: International Conference on Computer Science, Electronics and Industrial Engineering (CSEI)*, Springer Nature Switzerland, pp. 169–183, 2023. DOI: 10.1007/978-3-031-30592-4_13.
3. C. Balsa, M. M. Breve, C. V. Rodrigues, and J. Rufino, "Reconstruction of Meteorological Records by Methods Based on Dimension Reduction of the Predictor Dataset," *Computation*, vol. 11, no. 5, MDPI AG, p. 98, May 12, 2023. DOI: 10.3390/computation11050098.
4. M. M. Breve, C. Balsa, and J. Rufino, "Reconstruction of Meteorological Records with PCA-based Analog Ensembles Methods," *WorldCist'23 Conference*, (Accepted for publication).

5. Abstract published in *Symposium of Applied Science for Young Researchers (SASYR)*.

Note that the first page of the articles listed above are presented bellow in the appendix in the same order.



An Exploratory Study on Hindcasting with Analogue Ensembles of Principal Components

Carlos Balsa¹(✉) , Murilo M. Breve¹ , Carlos V. Rodrigues² ,
Luís S. Costa³ , and José Rufino¹

¹ Research Centre in Digitalization and Intelligent Robotics (CeDRI),
Instituto Politécnico de Bragança, Bragança, Portugal
{balsa,murilo.breve,rufino}@ipb.pt

² Vestas Wind Systems A/S, Design Centre Porto, Porto, Portugal
calvr@vestas.com

³ Mountain Research Center (CIMO), Instituto Politécnico de Bragança,
Bragança, Portugal
lcosta@ipb.pt

Abstract. The aim of this study is the reconstruction of meteorological data that are missing in a given station by means of the data from neighbouring stations. To achieve this, the Analogue Ensemble (AnEn) method was applied to the Principal Components (PCs) of the time series dataset, computed via Principal Component Analysis. This combination allows exploring the possibility of reducing the number of meteorological variables used in the reconstruction. The proposed technique is greatly influenced by the choice of the number of PCs used in the data reconstruction. The number of favorable PC varies according to the predicted variable and weather station. This choice is directly linked to the variables correlation. The application of AnEn using PCs leads to improvements of 8% to 21% in the RMSE of wind speed.

Keywords: Hindcasting · Analogue ensembles · Principal Component Analysis · Time series

1 Introduction

Classical weather *hindcasting* is the recreation of past weather conditions by applying a forecast model on a past starting point (reanalysis). This is done to validate the forecast model if comparable past observations are available. It may also be used to derive absent past data (non-recorded past observations) from the forecast model (reconstruction).

Hindcasting is also a field of research aiming to improve methods in other fields of meteorology such as *downscaling* or *forecasting*. Meteorological data reconstruction techniques are essentially based on the Analogue Ensembles (AnEn) method [8,9]. Hindcasting with the AnEn method allows to reconstruct



PCAnEn - Hindcasting with Analogue Ensembles of Principal Components

Carlos Balsa^{1,2}, Murilo M. Breve^{1,2}, Baptiste André³,
Carlos V. Rodrigues⁴, and José Rufino^{1,2}

¹ Research Centre in Digitalization and Intelligent Robotics (CeDRI), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal
`balsa@ipb.pt`, `murilo.breve@ipb.pt`, `rufino@ipb.pt`

² Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal

³ Université de Toulouse - INP - ENSEIHT, 31071 Toulouse Cedex 7, France
`baptiste.andre@etu.inp-n7.fr`

⁴ Vestas Wind Systems A/S, Design Centre Porto, Centro Empresarial Lionesa, R. Lionesa Edifício B, 4465-671 Leça do Balio, Portugal
`calvr@vestas.com`

Abstract. The focus of this study is the reconstruction of missing meteorological data at a station based on data from neighboring stations. To that end, the Principal Components Analysis (PCA) method was applied to the Analogue Ensemble (AnEn) method to reduce the data dimensionality. The proposed technique is greatly influenced by the choice of stations according to proximity and correlation to the predicted one. PCA associated with AnEn decreased the errors in the prediction of some meteorological variables by 30% and, at the same time, decreased the prediction time by 48%. It was also verified that our implementation of this methodology in MATLAB is around two times faster than in R.

Keywords: Hindcasting · Analogue ensembles · Principal component analysis · Time series · R · MATLAB

1 Introduction

The meteorological field is used daily for many purposes in our society and has a great impact on many decision-making processes. For instance, renewable energy management often requires information about weather conditions in places without available historical data or weather forecasts.

Weather conditions can be recreated by applying a forecast model to a past starting point, a process known as *hindcasting*. Its main function is to validate the forecast model when comparable past observations are available. It can also be used for *reconstruction* purposes, whereby it derives missing past data (past observations not recorded) from the forecast model.

Article

Reconstruction of Meteorological Records by Methods Based on Dimension Reduction of the Predictor Dataset

Carlos Balsa ^{1,*} , Murilo M. Breve ¹ , Carlos V. Rodrigues ²  and José Rufino ¹ 

¹ Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal; murilo.breve@ipb.pt (M.M.B.); rufino@ipb.pt (J.R.)

² Vestas Wind Systems A/S, Design Centre Porto, 4465-671 Leça do Balio, Portugal; calvr@vestas.com

* Correspondence: balsa@ipb.pt

Abstract: The reconstruction or prediction of meteorological records through the Analog Ensemble (AnEn) method is very efficient when the number of predictor time series is small. Thus, in order to take advantage of the richness and diversity of information contained in a large number of predictors, it is necessary to reduce their dimensions. This study presents methods to accomplish such reduction, allowing the use of a high number of predictor variables. In particular, the techniques of Principal Component Analysis (PCA) and Partial Least Squares (PLS) are used to reduce the dimension of the predictor dataset without loss of essential information. The combination of the AnEn and PLS techniques results in a very efficient hybrid method (PLSAnEn) for reconstructing or forecasting unstable meteorological variables, such as wind speed. This hybrid method is computationally demanding but its performance can be improved via parallelization or the introduction of variants in which all possible analogs are previously clustered. The multivariate linear regression methods used on the new variables resulting from the PCA or PLS techniques also proved to be efficient, especially for the prediction of meteorological variables without local oscillations, such as the pressure.

Keywords: hindcasting; forecasting; analog ensemble; principal component analysis; partial least square; multivariate regression



Citation: Balsa, C.; Breve, M.M.; Rodrigues, C.V.; Rufino, J. Reconstruction of Meteorological Records by Methods Based on Dimension Reduction of the Predictor Dataset. *Computation* **2023**, *11*, 98. <https://doi.org/10.3390/computation11050098>

Academic Editor: Shengkun Xie

Received: 29 March 2023

Revised: 7 May 2023

Accepted: 9 May 2023

Published: 12 May 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Filling gaps in observed time series is an important problem in many areas of applied sciences that depend on data analysis. Without this filling, data reconstruction is difficult or even impossible. This assumption is particularly true in weather forecasting, where the amount of stored information is growing four times faster than the world economy [1]. In view of this, big data analytics can help to improve predictions by uncovering patterns and correlations in the data [2] and reconstructing missing data in areas where there is limited information. Conversely, this growth in data also means that the amount of missing data is increasing, which makes accurate reconstruction a crucial task. To handle this challenge, forecasting methods must be able to handle large amounts of data, multiple data sources and a wide variety of meteorological variables. This requires advanced methodologies that can adapt to the particular characteristics of big data in weather forecasting.

Despite the general abundance of weather data available, there are still many regions without historical data records. These locations, which may be remote or under-developed, have the potential to be significant generators of renewable energy. However, without historical weather data, it is difficult to accurately predict the potential for energy generation in such places. Therefore, there is a growing need for methods that can generate weather data from limited inputs and locations, with the purpose of running simulations of environmentally driven systems that target these locations. This may greatly enhance our understanding of the potential for renewable energy generation, and may facilitate the development of sustainable energy systems in such regions [3].

Reconstruction of meteorological records with PCA-based Analog Ensembles methods

Murilo M. Breve^{1,2}, Carlos Balsa^{1,2} and José Rufino^{1,2}

¹ Research Centre in Digitalization and Intelligent Robotics (CeDRI),
Instituto Politécnico de Bragança, Campus de Santa Apolónia,
5300-253 Bragança, Portugal

² Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha
(SusTEC), Instituto Politécnico de Bragança, Campus de Santa Apolónia,
5300-253 Bragança, Portugal
`{murilo.breve,balsa,rufino}@ipb.pt`

Abstract. The Analogue Ensembles (AnEn) method has been used to reconstruct missing data in time series with base on other correlated time series with full data. As the AnEn method benefits from the use of large volumes of data, there is a great interest in improving its efficiency. In this paper, the Principal Component Analysis (PCA) technique is combined with the classical AnEn method and a K-means cluster-based variant, within the context of reconstructing missing meteorological data at a particular station using information from neighboring stations. This combination allows to reduce the dimension of the number of predictor time series, while ensuring better accuracy and higher computational performance than the AnEn methods: it reduces prediction errors by up to 30% and achieves a computational speedup of up to 2x.





Keywords: Meteorological data reconstruction, Analogue ensembles, K-means clustering, Principal component analysis, MATLAB, R

1 Introduction

Information about past weather states is crucial to many scientific domains and practical applications. In the renewable energy field, for instance, it is vital to know the historical weather data and meteorological patterns, in order to estimate the productive potential of a given site, before making substantial financial investments [9]. However, full meteorological data may not always be available or may be absent altogether. In this scenario, data reconstruction techniques come into play. These should be numerically accurate and computationally efficient.

A well-known approach for meteorological data reconstruction is the Analogue Ensembles (AnEn) method. Initially, it was used as a post-processing technique, to improve the accuracy of deterministic numerical forecast models [13]: past observations that are similar to the forecast are used to enhance the accuracy of the forecast. The AnEn method can also be used directly for weather forecasting [18, 6]. More recently [5], AnEn was used to reconstruct data of a

Applications of the Analog Ensembles Method to Meteorological Data Reconstruction in the Northeast of Portugal

Murilo Montanini Breve¹ , José Rufino¹ , Carlos Balsa² , and Luís de Sousa Costa² 

¹ Research Centre in Digitalization and Intelligent Robotics (CeDRI), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal

{murilo.breve,rufino,balsa}@ipb.pt

² Mountain Research Center (CIMO), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal

lcosta@ipb.pt

Abstract

The observation of weather states has always been a human need. Our most distant ancestors already tried to understand and predict the weather, but did not have reliable methods. In the 19th century, modern meteorology took its first steps: the French government, motivated by the sinking of ships near the coast of Crimea, because of a heavy rainstorm, created a network of 24 stations spread across Europe, which began to observe the weather. In recent years, due to computational advances, different methods of predicting weather states have begun to emerge, increasing the forecast extent and its accuracy.

The Analog Ensembles method (AnEn), introduced by Luca Delle Monache in 2011 [1], is a post-processing tool that has shown good results to improve whether predictions or perform hindcasting (reconstruction of missing meteorological data). The goal of this study is to use the AnEn method to perform hindcasting, in order to reconstruct past weather conditions in a specific area of the northeast of Portugal and verify its similarity with the actual forecast.

The AnEn method uses a two different time series: one with historical data (from a predictor station) and another with observed data (concerning a predicted station). The historical data is complete, while the observed data is missing or sparse in the prediction period. In Figure 1, which illustrates the methodology, a number of analogs are selected from the historical data set, according to their similarity to a predictor value (see step 1). At the same time instant, but at the predicted station, the corresponding observed data is selected and is used to produce a predicted value (step 3). This process is performed successively until the end of the prediction period data, and thereby it is possible to reconstruct the full predicted data from station 2. The AnEn method also allows using more than one predictor station (or more than one variable from the same station); in this scenario, the data from the predictor stations (or variable) can be used either dependently or independently (i.e., with the analogs selected in different predictor series having to overlap in time, or not).

The data for this research comes from weather stations managed by IPB and located in the northeast region of Portugal, near the villages of Edroso (*latitude* : 41.912778;