Statistical modelling of wind energy using Principal Component Analysis

Christina Skittides

A dissertation submitted for the degree of Doctor of Philosophy

Heriot-Watt University

Institute of Mechanical, Process and Energy Engineering

January 2015

Abstract

The statistical method of Principal Component Analysis (PCA) is developed here from a time-series analysis method used in nonlinear dynamical systems to a forecasting tool and a Measure-Correlate-Predict (MCP) and then applied to wind speed data from a set of Met.Office stations from Scotland. PCA for time-series analysis is a method to separate coherent information from noise of measurements arising from some underlying dynamics and can then be used to describe the underlying dynamics. In the first step, this thesis shows that wind speed measurements from one or more weather stations can be interpreted as measurements originating from some coherent underlying dynamics, amenable to PCA time series analysis. In a second step, the PCA method was used to capture the underlying time-invariant short-term dynamics from an anemometer. These were then used to predict or forecast the wind speeds from some hours ahead to a day ahead. Benchmarking the PCA prediction against persistence, it could be shown that PCA outperforms persistence consistently for forecasting horizons longer than around 8 hours ahead. In the third stage, the PCA method was extended to the MCP problem (PCA-MCP) by which a short set of concurrent data from two sites is used to build a transfer function for the wind speed and direction from one (reference) site to the other (target) site, and then apply that transfer function for a longer period of data from the reference site to predict the expected wind speed and direction at the target site. Different to currently used MCP methods which treat the target site wind speed as the independent variable and the reference site wind speed as the dependent variable, the PCA-MCP does not impose that link but treats the two sites as joint observables from the same underlying coherent dynamics plus some independent variability for each site. PCA then extracts the joint coherent dynamics. A key development step was then to extend the identification of the joint dynamics description into a transfer function in which the expected values at the target site could be inferred from the available measurements at the reference site using the joint dynamics. This extended PCA-MCP was applied to a set of Met.Office data from Scotland and benchmarked a standard linear regression MCP method. For the majority of cases, the error of the resource prediction in terms of wind speed and wind direction distributions at the target site was found to be between 10% and 50% of that made using the standard linear regression.

The target mean absolute error was also found to be only the 29% of the linear regression one.

Dedication

I dedicate this thesis first and foremost to my father Dr. Phil Skittides and my supervisor Dr. Wolf Gerrit Früh. Without them this thesis would not have been possible. I would also like to dedicate this thesis to my mother and friends who went to great lengths to support me throughout.

Acknowledgements

I would like to thank ETP (Energy Technology Partnership) and SgurrEnergy Ltd. for their financial support throughout this PhD research. I am very grateful to my supervisor Dr. Wolf Gerrit Früh, to Darran Gardner from ETP, Shona Quinn and Richard Boddington from SgurrEnergy Ltd and last but not least to Babis for his valuable help.

ACADEMIC REGISTRY
Research Thesis Submission

Name:	Christina Skittides		
School/PGI:	EPS/IMPEE		
Version: *(i.e. First, Resubmission, Final)*	FINAL	Degree Sought (Award **and** Subject area)	PhD in Mechanical Engineering

Declaration

In accordance with the appropriate regulations I hereby submit my thesis and I declare that:

1) the thesis embodies the results of my own work and has been composed by myself
2) where appropriate, I have made acknowledgement of the work of others and have made reference to work carried out in collaboration with other persons
3) the thesis is the correct version of the thesis for submission and is the same version as any electronic versions submitted*.
4) my thesis for the award referred to, deposited in the Heriot-Watt University Library, should be made available for loan or photocopying and be available via the Institutional Repository, subject to such conditions as the Librarian may require
5) I understand that as a student of the University I am required to abide by the Regulations of the University and to conform to its discipline.

* *Please note that it is the responsibility of the candidate to ensure that the correct version of the thesis is submitted.*

Signature of Candidate:		Date:	5/10/2015

Submission

Submitted By *(name in capitals):*	
Signature of Individual Submitting:	
Date Submitted:	

For Completion in the Student Service Centre (SSC)

Received in the SSC by *(name in capitals):*			
Method of Submission *(Handed in to SSC; posted through internal/external mail):*			
E-thesis Submitted ***(mandatory for final theses)***			
Signature:		Date:	

Please note this form should bound into the submitted thesis.

Updated February 2008, November 2008, February 2009, January 2011

Table of Contents

List of Tables

List of Figures

Glossary

Chapter 1

c : Weibull distribution scale factor

CP: capacity factor

GCM: Global Circulation Model

k : Weibull distribution shape factor

LIDAR: Light Detection And Ranging

MCP: Measure-Correlate-Predict

PCA: Principal Component Analysis

SODAR: Sonic Detection And Ranging

η : wind turbine efficiency

Chapter 2

AEP: Annual Energy Production

ANN: Artificial Neural Network

b : constant in linear regression relationship

c : subscript denoting the concurrent data

DAMS: Detailed Aspect Method of Scoring

ECMWF: European Center for Medium-range Weather Forecasting

m : gradient in linear regression relationship

NCAR: National Centre for Atmospheric Research

NOABL: Numerical Objective Analysis Boundary Layer

$pred,hist$: subscripts denoting the predicted and historical data

tar,ref : subscripts denoting the target and reference sites

v : wind speed

VMM: Virtual Met Mast

VR: Variance Ratio method

WAsP: Wind Atlas Analysis and Application

θ : wind direction

Chapter 3

c : subscript denoting calibrated data in MCP methodology

d_i : Euclidean distance of a single point

D_j : distance to nearest neighbours

EOF: Empirical Orthogonal Function

h : subscript denoting reference only data of concurrent data set used for calibration

i_0 : time-delay row index

j : time-delay column index

j_0 : observable time-delay index

k : constant factor characteristic of the spring

k : entry of neighbour to latest measurements

k' : entry of training principal components

M : number of columns in time-delay matrix

M_w : number of lags (window length) in time-delay matrix

N : number of rows in time-delay matrix; equation (11)

n : subscript denoting new time series projection in forecasting methodology

N_0 : number of channels in the time-delay matrix; N_c in schematic of Figure 8

n_n : nearest neighbours

N_t : length of observations ; equation (11)

n_x : orbit length

P : principal components matrix

p : subscript denoting prediction in MCP methodology

pc: principal component

$P_f^{\,j}(T)$: ensemble prediction based on nearest neighbours

r : subscript denoting historical reference data

S : singular vectors matrix

SSA: Singular Systems Analysis

SVD: Singular Value Decomposition

svec: singular vector

T : leading time

t : subscript denoting truncation in forecasting and MCP methodology

$u_p^j(t)$: predicted wind speed in forecasting methodology

v : phase space variable of a dynamical system usually associated with the velocity

x : phase space variable of a dynamical system usually associated with the position

x : signal 1 of pendulum system

Y : time-delay matrix

y : signal 2 of pendulum system

$y_{j0}(t)$: delay vector, time-delay matrix entries

$\bar{y}_{j0}(t)$: mean value of forecasting time-delay matrix entries

ε : error term of pendulum system

$\theta_p^j(t)$: predicted wind direction in forecasting methodology

Λ: singular values diagonal matrix

$\mu = \bar{y}_{j0}$: mean value of original time-delay matrix

μ_c: mean value of calibrated time-delay matrix

μ_p: mean value of predicted time-delay matrix

σ: standard deviation of original time-delay matrix

σ_c: standard deviation of calibrated time-delay matrix

σ_p: standard deviation of predicted time-delay matrix

$\sigma_p(t)$: predicted standard deviation in forecasting methodology

τ: lag in time-delay matrix

Chapter 4

ANFIS: Adaptive Neuro-Fuzzy Inference Systems

ARIMA: Autoregressive Integrated Moving Average

ARMA: Auto Regressive Moving Average

$BIAS(T)$: bias in forecasting methodology

e_t: error in forecasting methodology

$\mathrm{Im}\, p_{ref,BIAS}(T)$: improvement measure of forecasting methodology

$MAE(T)$: mean absolute error of forecasting methodology

M_t: truncation value in forecasting and MCP methodology

M_w: window length value in forecasting and MCP methodology

NWPs: Numerical Weather Prediction systems

PI : performance index of forecasting methodology

$RMSE(T)$: root mean square error of forecasting methodology

SVM: Support Vector Machine

$u_{f,i}$: filter applied to prediction in forecasting methodology

u_x : horizontal velocity component of wind speed

v_x : horizontal velocity component of wind direction

Chapter 5

A : minimum value of $u_{1,p}$

A_1, A_2 : local wind magnitude of noisy pendulum system

B : minimum value of $u_{2,p}$

BFH: Blackford Hill Met. Office station, target site

E_R : benchmark error of GGB M_w : 7 days, M_t : 6

$E_{u,1}, E_{u,2}$: error of GGB and BFH respectively

e_x, e_y : errors of signals x, y of noisy pendulum system

$\bar{e}_x, \bar{e}_y$: mean of errors of signals x, y of noisy pendulum system

E_1, E_2 : relative error of GGB and BFH respectively

$|E_1|, |E_2|$: absolute relative error of GGB and BFH respectively

f_1, f_2 : dynamics of modulating wind of noisy pendulum system

GGB: Gogarbank Met. Office station, reference site

$half$: subscript denoting half information time-delay matrix of noisy pendulum system

$MAE_{1,ws}, MAE_{2,ws}$: mean absolute error of GGB and BFH respectively

p: subscript denoting predicted time-delay matrix of noisy pendulum system

$r_{\sigma_x}, r_{\sigma_y}$: time-shifted standard deviation ratios of signals x, y of noisy pendulum system

S: time shift at which maximum correlation occurs of noisy pendulum system

$sd(e_x), sd(e_y)$: standard deviation of errors of signals x, y of noisy pendulum system

$test$: subscript denoting testing time-delay matrix of noisy pendulum system

u_1, u_2: wind speeds of GGB and BFH respectively

$u_{1,p}, u_{2,p}$: predicted wind speeds of GGB and BFH respectively

$u^*_{1,a}, u^*_{2,a}$: rescaled predicted wind speeds of GGB and BFH respectively

$u^*_{1,p}, u^*_{2,p}$: normalised predicted wind speeds of GGB and BFH respectively

x_0: equation representative of UK weather characteristics of noisy pendulum system

$\delta\phi$: time shift, time of flight between two sites of noisy pendulum system

ε: error term of noisy pendulum system i.e. turbulence

μ_1, μ_2: mean of training GGB and BFH data respectively

σ_1, σ_2: standard deviation of training GGB and BFH data respectively

Chapter 6

$Bias$: bias of the reference, target site and linear regression model of PCA-MCP methodology based on the $Bias$ of the forecasting methodology equation (33)

E_{lr}: linear regression model error

e_{lr} : error based on difference of linear regression prediction pdf and actual reference data pdf

e_{ref}, e_{tar}: error of the reference and target site of PCA-MCP methodology

MAE_j : calculated MAE for each distribution

$MAE_{ref}, MAE_{tar}, MAE_{lr}$: mean absolute errors of the reference, target site and linear regression model of PCA-MCP methodology

N : number of wind speed bins of width $\Delta u = 1 m/s$

pdf : probability density function

PI : performance index of PCA-MCP methodology

U_1, U_2 : original wind speed of the reference and target site respectively

$u_{1,past}, v_{1,past}$,: linear regression model wind speed and wind direction combination of reference historical data

$u_{2,past}, v_{2,past}$: linear regression model wind speed and wind direction combination of target historical data

u_1, v_1 : vector combination of wind speed and wind direction of the reference site

$U_{2,pred}$: linear regression model target wind speed prediction

u_2, v_2 : vector combination of wind speed and wind direction of the target site

β_0 : linear regression model intercept

$\hat{\beta}_0$: linear regression model estimated intercept

β_1 : linear regression model slope

$\hat{\beta}_1$: linear regression model reference wind speed estimate

ε : linear regression model error term

θ : wind direction of the reference and target site (in degrees)

Chapter 7

N_G : Gap length used for the data linear interpolation

Chapter 1 Introduction to wind resource

This chapter will give a brief introduction on the current status of wind energy. Wind resource assessment which is the aspect of wind energy of interest for this research will then be discussed.

1.1 Wind energy industry

Wind energy is one of the most established renewable energy forms. It has been one of the fastest growing renewable industries for the past two decades. Its growth has appeared at the beginning of the 90's and ever since it has become a more mature, clean energy generating technology. As facts indicate, wind industry is expected to continue existing with lower costs as energy security threats and the immediate need to meet the CO_2 reduction standards so as to prevent climate change [1].

In more details, wind energy's key role as a renewable energy form can be verified by various statistics. In Europe, there is currently 128.8 GW of installed wind capacity where 8GW come from offshore and 120.6 GW from onshore installations [2]. Wind power installations have increased annually over the past 14 years from 3.2GW in 2000 to 11.8GW in 2014. Germany followed by Spain, UK and France are the leading EU countries in the wind installed capacity [2]. In a normal wind year, the installed wind capacity by the end of 2014 could produce 284TW of electricity enough to cover 10.2% of the EU's electricity consumption needs where 9.1% originates from onshore and 1.1% from offshore wind [2]. The EU wind farms investment ranges between €13.1bn and €18.7bn with onshore wind farms in particular having investments from €8.9bn to €12.8bn. Furthermore, wind energy technology installations had the highest installation rate in 2014 with 43.7% of all new installations and 11.8GW [2]. In Europe, 79.1% of the newly installed capacity came from renewable energy sources with the installation of 21.3GW renewable power capacity.

In 2014, UK installed 1,736.4MW of wind power with 813.4MW originating from offshore wind [2]. Renewable energy currently provides 19% of the UK's electricity

needs and wind energy covers half of it. UK power needs were met by 13GW with wind energy [3]. More specifically, onshore wind is providing 5% of the electricity in the UK and this percentage is expected to rise to 10% by 2020. Another attractive factor of wind as a form of energy is that UK has an excellent wind resource potential and its cost as a form of energy is also small compared to the benefits of decarbonisation achieved [3]. The economic benefit from wind energy to local communities is very important also considering small scale wind installations; by the end of 2014, 27,819 small and medium turbines have been deployed across UK saving 168,257 tonnes of CO_2 [3].

Scotland in particular, has shown a steady growth in renewable electricity capacity currently at 7.3GW in the end of 2014 compared with 2.7 MW in 2007 [4]. Onshore wind, accounts for over 69% of the installed capacity in Scotland followed by hydro, offshore and bioenergy. Furthermore, Scotland's renewable electricity output has raised to 19,067GWh in 2014 from 8,215GWh in 2007 and the electricity generation originating from renewables was around 49.8% in 2014 [4]. Onshore wind investment was 4,513MW in the end of 2013 and 5,015MW in the end of 2014 with a total of £701.8m whereas offshore was 190MW in the end of 2003 and 197MW in the end of 2014 with a £22.8m [4].

The constant technological development alongside with the increasing energy needs and the necessity of turning to a more sustainable future will undoubtedly establish even further wind energy in the near future.

1.2 Importance of wind resource assessment

Wind energy is a strongly intermittent form of energy with a large variability [5]. Given that modern wind farms have installed capacities of several hundred megawatts or more, even a small overestimate or an uncertainty in the predicted resource can result in a shortfall of income of several million pounds annually per wind farm. For this reason, wind resource assessment is an important part of siting and developing wind farms.

Since wind is highly dependent on regional factors affecting the larger weather systems, wind energy production is significantly challenging in terms of prediction [6-8]. In order to investigate the variability in wind temporally and spatially, wind speed and direction are monitored in different locations (reference sites) close to a potential site (target site) so that the characteristics of the wind resource can be established [9].

Different types of data sets from different wind masts in wind farms and meteorological office stations from other nearby sites are being used by wind farm developers. For example, meteorological stations report hourly wind speed and direction data from instruments typically 10m above ground, but dedicated wind farm masts are typically taller and would sample measurements at a second by second or minute by minute frequency. When these datasets are analysed by well-established MCP methods, they can significantly aid in the wind uncertainty reduction and prediction [10].

A typical problem concerning wind data collection is the large variability between areas and sources containing wind data [11]. Part of this is because the local wind resource is strongly affected by its immediate environment and part is due to the instrumentation used. A further problem that wind data analysis often faces is that poor quality and inadequate data which in both cases can lead to poor predictions [12]. Sometimes, other global datasets can be used instead such as reanalysis data [13]. Reanalysis data are produced by combining a range of different meteorological datasets such as: remote sensing observations, satellite data, surface observations coming from land. These data are used as an input to a Global Circulation Model (GCM), in other words a numerical weather prediction model, so at to result in a range of values of meteorological variables at discrete time intervals [14]. Their accuracy however, could be questioned since they are not always representative of the area and the data readings might not be recorded frequently enough. Satellite data specifically, look into for example temperature measurements and cloud coverage.

Instruments such as LIDAR, SODAR have been used since they can provide with high resolution wind data as an alternative [8]. As mentioned, some of the main factors that influence the wind resource assessment quality and reliability are the location of the

wind farm itself [15], the wind statistics of measurements before the turbine installation and during their operation that are expected but also all the technical equipment used in the farm.

Furthermore, as far as the location of the wind farm is concerned, its topography is very important. In more detail, the effects of the terrain on the wind flow, the roughness and the small/large scale weather of the surrounding area are some of the very important factors that should be taken into consideration when planning a wind farm. In Velasquez research the islands orography was used as part of his analysis and it was found to be an important factor for the results since it can affect the sites correlation [16]. Furthermore, wind speed and direction which are some of the most important factors that determine the wind statistics can vary at all-time scales.

A semi-empirical methodology was developed by the UK Met Office to estimate small-scale wind energy potential which consists of the application of corrections to wind data locally and regionally based, according to average surface roughness parameters [17]. Weekes and Tomlin in their research tried to evaluate the aforementioned method [9] for 38 UK sites in order to examine the errors between the actual and predicted wind speeds and wind power density, the surface roughness due to the difference of terrains and the morphology of the UK sites used. They concluded that the method has some limitations; however, even with simple modifications of these aerodynamic parameters improvements could be made. They also noted that semi-empirical modes can be applied easily and with a small cost for wind resource assessment however they include uncertainties that could pose a problem when assessing large wind investments. Therefore, in these cases they should be used in addition to onsite measurements.

Due to its large variability interannually it is important to obtain a sufficient amount of wind data measurements [11, 18-24]. Different authors suggest different periods such as 3 years, 10 years or even 20 to 30 years to be able to characterize sufficiently a sites wind resource [11, 16, 18, 21, 25-27]. Different cycles from daily to seasonal and interannual ones [11, 19, 26, 27] can be observed but also wind turbulence and gusts are quite common too. Turbulence represents rapid fluctuations in wind speed and direction

at all time scales, including those shorter than is usually available from resource assessment measurements which can impact the wind turbine performance. Wind gusts which are a sudden increase in the wind speed which last long enough to affect turbine performance [23]. In addition, the turbines have to adjust to the wind fluctuations at all time but that is not always the case since they often have a delay and a lack of immediate response. The anemometers also can be of low precision and response depending on their quality so data measurements can be of poor quality and quantity.

Albers et al. highlight the extreme importance of good quality of wind speed measurements as they think that it is the only way to limit the financial risk of wind farm projects especially for complex terrain sites [28]. As Angelis-Dimakis et al. note in their review for wind energy, one of wind resource's greatest challenge is to come up with flow and numerical models which can identify the wind flow features while being at a complex terrain and at the same time keep the calculation cost at a low level [8] Gerdes et al. also mentioned in their study, one has to be careful with the wind speed measurements used since if the measurements are undertaken in a 'good' energy production year, it could lead to an overestimation of the forecasted energy production of a wind farm. Thus he highlights that the long-term effects of measurements must be taken into account [20]. Hence, the wind farm equipment quality and the methods of analysis and prediction of wind behaviour are indeed of extreme importance for a good resource assessment.

1.3 From wind to electricity

Figure 1 illustrates the power output for a given wind turbine. The maximum value of the y axis power coefficient is 1 since it is scaled to rated power. The typical wind turbine efficiency η is described as

$$\eta = \frac{P}{\frac{1}{2}\rho U^{3} A} \tag{1}$$

where P is the power output from the turbine and the denominator is the power carried by the wind with ρ the air density, U the wind speed and $A = \frac{\pi D^2}{4}$ the swept area of the blades.

At cut-in winds when the turbine starts to operate, the efficiency and output increase rapidly until the rated power is reached. At that point the typical turbine efficiency is about 40%-50% and thus it is the typical best efficiency for a wind turbine. After the rated power output reaches 1 at the rated wind, it becomes flat which means that even if wind speed is increased, the efficiency decreases. It can also be seen that, usually, above u =25 m/s the turbine is turned off.

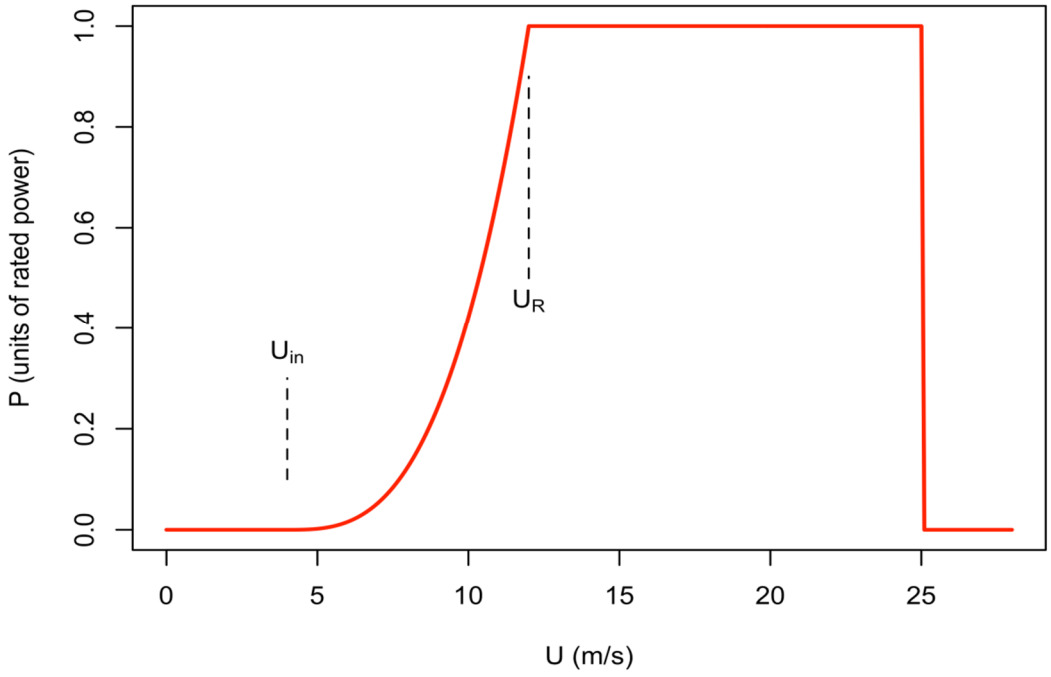


Figure 1. Performance curve of a typical wind turbine.

The Weibull distribution is often found to best describe wind speed distribution, at least in Europe. The Weibull distribution equation is of the following form

$$P(U) = \left(\frac{k}{c}\right)\left(\frac{U}{c}\right)^{k-1} \exp\left[-\left(\frac{U}{c}\right)^{k}\right] \qquad (2)$$

where U is the wind speed k is the shape factor and c the scale factor measured in m/s.

Figure 2 indicates the probability density function of the Rayleigh distribution, a special case of Weibull distribution with k =2, frequently used for well- behaved sites with wind speeds above 4.5 m/s. The Rayleigh distribution equation is of the form

$$P_R(U) = \frac{2U}{c^2}\exp\left[-\left(\frac{U}{c}\right)^2\right] \tag{3}$$

The capacity factor (CP) for wind turbines/ farms is the ratio of actual output of the wind turbine/ farm for some time, over their full potential. The typical estimate of CP for wind farms is 30% [29, 30].

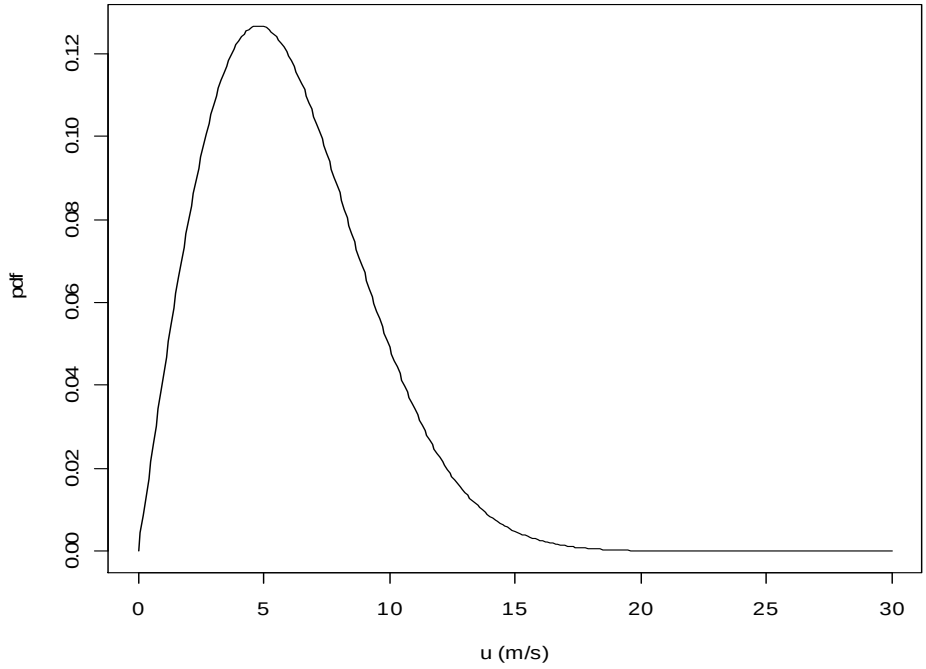


Figure 2. Probability density function (pdf) of Rayleigh distribution.

To examine the sensitivity of wind turbine performance so as to verify the rationale behind the importance of a good wind resource assessment an example will be illustrated in Figure 3. Taking a typical wind turbine 45m above the ground nearby Edinburgh airport with average annual wind speed of u=6 m/s, the expected CP of the

Rayleigh distribution with scale factor c =6.77 m/s is expected to be CP=16.4%. Changing the scale factor randomly by ± 0.4 % resulted in a change of the CP by ± 1.3%, as it can be seen in Figure 3.

Früh in his research for 43-year-long wind data in Scotland concluded that the electricity output is subject to sensitivity when small changes in mean wind exist, especially for poor wind resource sites. He also highlighted the variation in the expected output from year-to-year ranging from 10% to 15% in Scotland [30].

The conclusions that can be derived from this example are that the turbine electricity output is very sensitive to changes in the wind statistics and hence the electricity distribution and the sites chosen are of great importance. Finding the best possible estimate of the sites distribution results in choosing a good site. Furthermore, it should be noted that wind direction apart from the speed measurements is essential to be correct since together they can affect the electricity outputs.

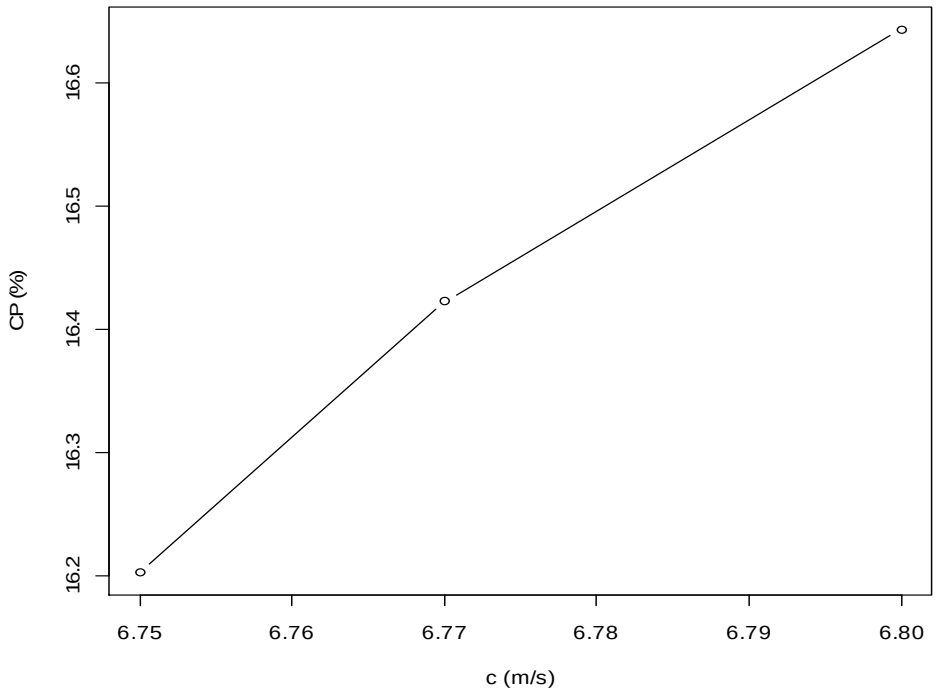


Figure 3. Change of CP in response of changing c.

1.4 Aims and objectives

In order to address the issues mentioned in sections 1.2 and 1.3 a novel Measure-Correlate-Predict (MCP) technique was developed in this research which is based on the statistical methodology of Principal Component Analysis (PCA) and applied through the statistical package R [31].

1.5 Thesis outline

This thesis structure will be as following: Chapter 2 will explain the fundamental principles of the MCP methods and give their overview. Chapter 3 will explain the theory behind the PCA methodology, Chapter 4 will illustrate the results of PCA being used as a forecasting method for wind purposes. Chapter 5 will contain the initial results of PCA used as an MCP method followed by Chapter 6 which will exploit the main results of this research i.e. using PCA and an MCP method for wind speed and direction. Finally, Chapter 7 will include discussion and conclusions of this research followed by the Appendices which contain the R [31] scripts which were used throughout this research.

Chapter 2 Measure-Correlate-Predict (MCP) methods

This chapter will discuss the principles of MCP methods as well as describe the already established MCP methods in research and in the wind energy industry.

2.1 Fundamental principles of MCP methods

The principle behind the MCP methodology is to correlate short-term wind data of a target site, usually the wind farm of interest, with long-term wind data of a reference site, usually a meteorological office site nearby, so that a relationship between them is established [8, 9, 18, 32, 33]. A typical (concurrent) data measurement period used is a year or more [24]. Current commercial practice in companies is around 18 months; this period is enough to capture the annual wind cycle and not too long so as to be longer and more costly than necessary. The goal of MCP in this application is to characterise the wind speed distribution as a function of wind direction and other invariants so as to estimate the annual energy capture of a wind farm [24, 34]. As mentioned in section 1.2, the electricity sales are directly proportional to the annual energy production and hence a major factor in the economic analysis of a potential site and, for that reason, a reliable wind resource estimate is a key factor for investors and developers for their planning and decision-making.

In general, the following function mathematically describes MCP for the wind application:

$$V_{tar,c} = f(V_{ref,c}, \theta_{ref,c}) \tag{4}$$

where V is the wind speed, θ the wind direction, c denotes the concurrent data and subscripts tar and ref are the target and reference sites, respectively. The function $f(.)$ has a fixed form which is determined from the concurrent data and then applied to the historical record. Hence equation (4) indicates that the wind speed of the target site is a function of wind speed and direction of the reference site. Using the concurrent set of measurements, $(V_{ref,c}, \theta_{ref,c})$ and $(V_{tar,c}, \theta_{tar,c})$ it is possible to determine the function f ,

which can then be used to estimate historical data for the target site by $(V_{tar,h}, \theta_{tar,h}) = f(V_{ref,h}, \theta_{ref,h})$. A simple example of this function estimation is the linear regression MCP, where a line-fit for V_{tar} (on the *y* axis) is regressed against the V_{ref} (on the *x* axis) for the concurrent measurements. This gives a best-fit line with a form of the function *f* only depending on the wind speed, V_{ref} , but not the direction, θ_{ref}. This can be expressed as

$$V_{tar} = f\left(V_{ref,}\theta_{ref}\right) = b + mV_{ref} \tag{5}$$

where b is a constant and m is the gradient. Thus, using the historically available reference wind speed record, $V_{ref,h}$ in equation (5) gives the prediction for the corresponding wind speed at the target site, $V_{tar,h}$. Figure 4 is depicting a straight forward case of the target site and reference site line fitting where the solid red line shows the line of best fit found from linear regression, resulting in the equation given at the top left.

MCP methods can be divided into analytical and empirical models depending on the function that they use. Analytical models make a clear assumption regarding the form of the function they use [18] where the parameters of the function are determined by the regression analysis results. The most common is the linear relationship used in equation (5) to illustrate the principle, but non-linear MCP regressional methods also exist. Historical data are referred as the data originating from previous years readings taken from the reference site datasets.

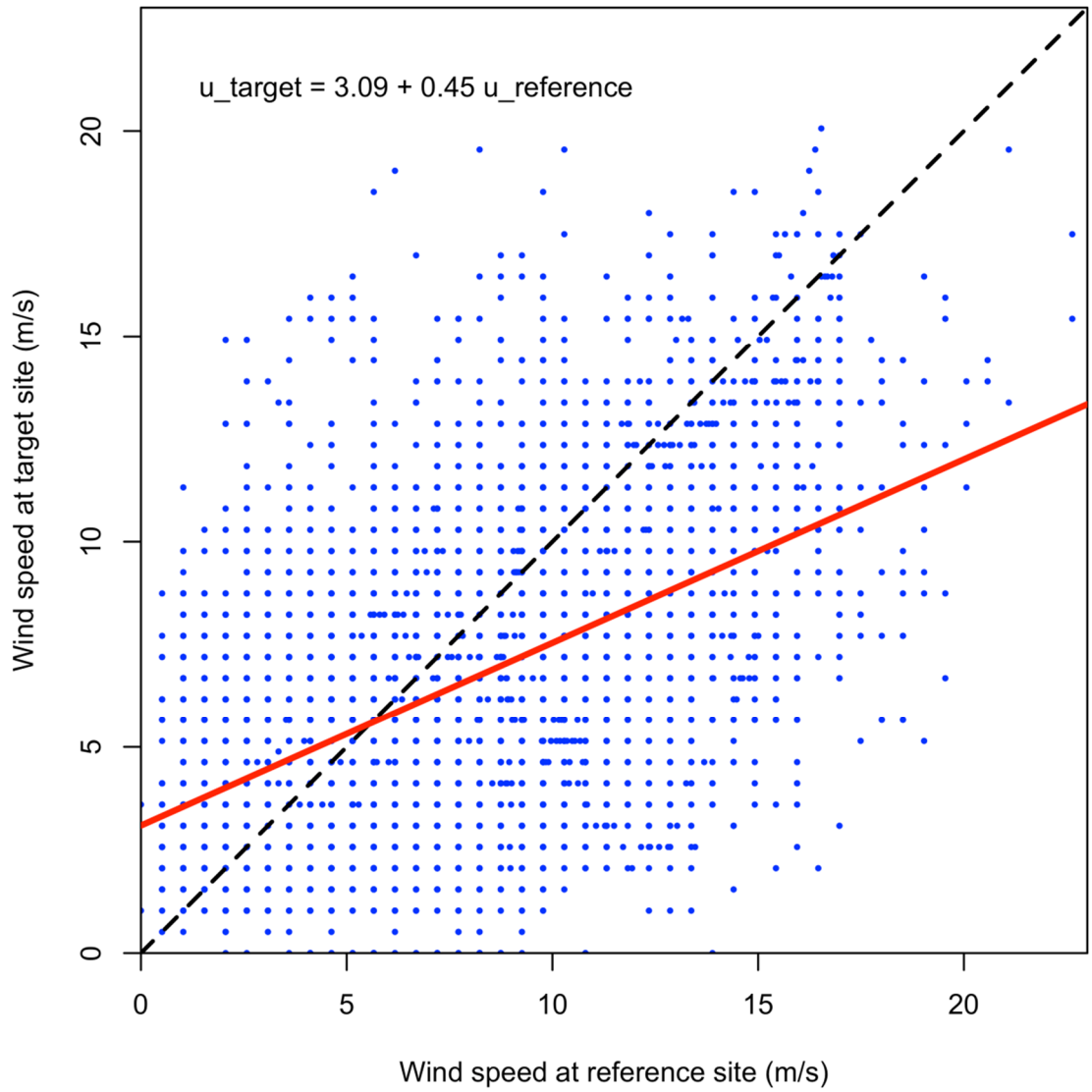


Figure 4. A straight forward case of line fitting where the solid red line shows the line of best fit found from linear regression.

Almost all regression-based analysis methods assume that wind at the target site at a point in time is directly linked to the wind at the reference point from the same time period. Occasionally, and especially in complex terrain, this assumption is not valid and other, empirical, MCP methods have been developed to overcome this. Rather than linking the wind speeds at certain times to each other, the wind distribution or wind rose at the target site is compared to that for the reference site. Empirical MCP methods often use a matrix to link a wind situation at the target site with a corresponding situation at the reference site using a matrix form or lookup table for a 'case-by-case'

correspondence where no underlying analytical assumption is applied to all cases [35]. Like the analytical case, the matrix or distribution methods include concurrent data to extract the relationship [36] between the two distributions or wind roses for the period of the concurrent measurements, and subsequently apply that relationship to the historical wind rose from the reference site to predict the corresponding wind rose for the target site.

2.2 MCP methods in literature

2.2.1 Overview of established MCP methods in literature

Essentially, the fundamental approaches can be classified into (a) analytical, of which most use a form of regression, (b) empirical, which mostly determine links between wind roses or wind speed distributions (also known as frequency tables), and (c) non-linear modelling, of which the Artificial Neural Network (ANN) has been used most widely. Table 1 is attempting to provide an overview over the range of the fundamental MCP approaches in literature. The first column indicates a classification according to the main approach, and the second column lists common variations on that basic assumption. For example, a linear regression can be applied to all data indiscriminately (‘Simple’) or it can be applied individually to subsets of the measurements where each subset only considers the measurements when the wind direction was in a chosen range (‘Binned’). The third column presents the equation of the method where applicable, followed by the last column the explanation of the equation’s variables.

Class	Variant	Assumptions / Limitations	Equation	Explanation
Analytical				
Linear Regression	Simple	Single fixed linear relationship of wind speed at each point in time between sites.	$\hat{y} = mx + b$	$\hat{y}$ is predicted target wind speed, x observed reference wind speed and m, b slope and offset [37].
Linear Regression	Binned	A set of fixed linear relationships, one for each wind direction sector.	$u_j = \frac{m_j \sum_{i=1}^{12} Z_{i,j} u_i + c_j}{100}$	u_j is the predicted mean wind speed, u_i the mean wind speed of Met. Site sectors, m, c the regression gradient and intercept and $Z_{i,j}$ are the percentage weights [35].
Linear Regression	Gaussian scatter	Represents a variety of processes which are not accounted for in the simple linear model and result in scatter in the individual data points about the mean prediction.	$\sigma_{res} = \sqrt{\frac{1}{N-2} \sum_{i=1}^{N} \left(u_{tar,i} - \hat{u}_{tar,i}\right)^2}$	σ_{res} is the average wind speed for a particular 30^o angular sector, N is the number of observations, $u_{tar,i}$ is the wind speed at target site and $\hat{u}_{tar,i}$ the mean target prediction [33].
Variance Ratio	Linear relationship but the coefficients (slope and intercept) are determined by ratio of variance	This approach differs from the linear regression one because no direct attempt is made to model the error term in order to reconstruct the residual scatter.	$\hat{y} = \left(\mu_y - \left(\frac{\sigma_y}{\sigma_x}\right)\mu_x\right) + \left(\frac{\sigma_y}{\sigma_x}\right)x$	$\hat{y}$ is predicted target wind speed, $\mu_x, \mu_y, \sigma_x, \sigma_y$ is the mean and standard deviation of x the reference and y the target

	from both sites and not through regression.			site respectively of the concurrent datasets [36].
Multiple Principal Least Squares (MPLS)	Based on regression but combining information from different sources.	Does not require the existence of concurrent data. It benefits from data of multiple sites. Reference site data do not need to be in same length with each other or with the target site data.	$\hat{T} = u_T \sum_x V_T^*$	$\hat{T}$ is the predicted target site data, U_T, V_T^* orthogonal matrices and $\sum_x$ diagonal matrix based on singular value decomposition of the reference data [32].
Empirical				
Distribution Methods	Probabilistic	Limited dependence between the reference and candidate site data and limited representation of the long-term wind characteristics for the data period used for training purpose.	$P_c^{LT}(v_i,d_j) = \frac{1}{NF}\sum_{k=1}^{N_w}\sum_{z=1}^{N_d} P_{c-r}^{ST}(v_i,d_j,v_k,d_z) P_r^{LT}(v_k,d_z)$	$P_r^{LT}(v_k,d_z)$ is the probability mass function of long-term wind speed and direction of reference site, NF is the Normalization factor and $P_{c-r}^{ST}(v_c,d_c,v_r,d_r)$ is the joint probability mass function of the short-term candidate and reference site wind speed and direction. N_w, N_d are the wind speed and direction bins of the joint probability mass function [38].
Distribution Methods	Simple linear regression pdf (SLRpdf)	Models the underlying distribution of target site wind speeds rather than the historical time-series.	$f_{long}(u_t) = \int \frac{f(u_r,u_t)}{f_{short}(u_r)} f_{long}(u_r) du_r$	u_t, u_r are the wind speed observations of target and reference sites, $f_{long}(u_t), f_{long}(u_r)$ is the long-

		Also rather than the restriction that a specific reference site wind speed corresponds to a specific target site wind speed, it predicts a distribution of target site wind speeds for every reference site wind speed in the form of a conditional probability distribution.		term marginal probability density function of wind speed at target and reference site and $f_{short}(u_r)$ short-term training marginal pdf of the reference site [39].
Distribution Methods	Weibull pdf (Wpdf), Nonlinear regression with bivariate cumulative Weibull pdf's (WR)	As above but assumes that both reference and target sites are described by Weibull distributions**.**	$F(x,y)=1-\exp\left\{-\left[\left(\frac{x}{\lambda_x}\right)^{\frac{k_x}{\delta}}+\left(\frac{y}{\lambda_y}\right)^{\frac{k_y}{\delta}}\right]^{\delta}\right\}$	k is the shape and λ the scale factors, δ the parameter controlling the degree of association between x, y the reference and target sites [40].
Nonlinear Modelling				
Artificial Neural Networks (ANNs)	Multilayer Perceptron Topologies (MLPs) [41] Group Method of Data Handling method (GMDH) [41] Extreme Learning Machine (ELM) [41]	No initial assumption of relationship between sites. Link is established through 'learning' of hidden layers. Interpretation and error estimation is not trivial [15, 16, 41].	$V_c^{LT}=\left[\overline{V}_c^{ST}-\left(\frac{S_c^{ST}}{S_r^{ST}}\right)\overline{V}_r^{ST}\right]+\left(\frac{S_c^{ST}}{S_r^{ST}}\right)V_r^{LT}$	LT, ST are the long-term and short-term data, S is the standard deviation and V_r, V_c is the reference and candidate wind speed [16] based on the Variance Ratio method [36].

Hybrid MCP		Correlates the wind data at the targeted site with that at multiple reference stations. Accounts for the local climate and the topography information. The weight of each reference station is determined based on: (i) the distance and (ii) the elevation differences between the target wind site and each reference station.	$w_i = \frac{1}{2(n_{ref}-1)}\left(\frac{\sum_{j=1,\, j\neq 1}^{n_{ref}} \Delta d_j}{\sum_{j=1}^{n_{ref}} \Delta d_j} + \frac{\sum_{j=1,\, j\neq 1}^{n_{ref}} \Delta h_j}{\sum_{j=1}^{n_{ref}} \Delta h_j}\right)$	w_i is the weight of each reference station, n_{ref} the number of reference stations, $\Delta d_j, \Delta h_j$ are the distance and elevation difference between target and j^{th} reference station [42] .

Table 1. MCP methods in literature.

All the linear regression, variance ratio and MPLS methods for MCP attempt to find a prescribed fixed function to describe the link between the wind at the reference and target sites at each point in time. In the case of the simple linear regression, the a single function $v_{target} = mv_{reference} + b$ (as in equation (5) above) is determined by the linear regression between the full contemporary records of the reference and target site, and then that single function is applied to the entire historical wind speed record from the reference site to calculate the corresponding predictions of the wind speed at the target site. In the case of refinements, the binned regression for example, repeated linear regressions are applied, each only to the selection of contemporary wind speeds which belong to a wind direction bin. With that, different values of m and b are calculated for each wind direction sector. Then again, only the selection of wind speeds from the historical records with their respective m and b are used to predict the corresponding target wind speeds. To illustrate this with just using two bins: one for easterly winds (direction between 0° and 180°) and one for westerly winds (direction between 180° and 360°): m_1 and b_1 are found by a line fit between only those concurrent v_{target} and $v_{reference}$ for which the reference wind speed has a direction between 0° and 180°; likewise a different line fit is applied to the wind speed data from the other wind direction bin to get values m_2 and b_2.

Having identified the two best-line fits, the historical target wind speeds are predicted by again splitting the historical reference data into those where the wind direction is between 0° and 180° (say, $v_{hist,1}$) and those where it is between 180° and 360° ($v_{hist,2}$). Then the historical, climatological wind speed at the target site, v_{pred}, is predicted as the set made up of

$$V_{pred,1} = b_1 + m_1 v_{hist,1} \tag{6}$$

and

$$V_{pred,2} = b_2 + m_2 v_{hist,2} \tag{7}$$

This principle can be extended to more wind directional bins which can be arbitrarily chosen, or could be extended to incorporate a time-delay between reference and target site to allow for wind speed changes to be transported by the wind from the reference site to the target site. For example, for a target site 40 km downwind of the reference site at a wind speed of 5 m/s, a sensible delay would be 40000/5 = 4000s or a little over an hour. However, choosing the best time-delay is difficult because physically time-delay incorporates wind speed and direction and it is correct for one particular wind speed and wind direction only.

Distribution methods for MCP are fundamentally different in that they do not attempt to find a link between a wind speed measurement at a particular time at the reference site and a corresponding speed at the target site, but the link is made between the wind statistics (distribution) over the period of the availability of the concurrent data. So, while regression links two wind speed measurements, the distribution methods link how often the wind was at a particular strength at the reference site with how often the wind was at that strength at the target site. In that sense it looks at dependencies.

Artificial Neural Networks attempt to find a link between the sites by training the ANN on the concurrent set, and then using that trained ANN to predict the target site wind speeds using the historical reference data. While such a method can be very powerful, as the link between the sites is not prescribed in a simple formula but explored in the training process, it is notoriously difficult to predict how good an ANN model has performed, and the training and prediction process is a fairly complex and non-transparent process.

The potential of ANNs to perform very well suggests that using approaches which can 'learn' not only parameters in a specified function f but the form of the function itself potentially can provide a much better prediction without having to resort to expensive and time-consuming flow modeling. One such method, but through a much more transparent process, uses empirical functions based on maximizing the covariance between two measurement series.

2.2.2 Applications of MCP methods

The establishment of the relationship between the two sites, reference and target can be complicated and is based on several stochastic variables. Some of these invariants are: wind speed and direction over time, which are used as the inputs for the MCP algorithms in most cases, the distance between the reference and the target site, for example the time of flight delays, the effects of the terrain on the flow, e.g. local obstructions such as forests, hills and the large and small scale weather patterns, e.g. atmospheric stability [36].

Due to the existence of these invariants, wind speed observations can be binned in accordance with their wind direction values. It can be found that wind direction is binned in different sectors at the target and reference site and hence the binning might not always coincide. Thus a decision must be made in order to select which site, reference or target binning will be used [40]. An alternative binning methodology was proposed by Woods and Watson where the predicted wind speeds related with a specific wind direction sector are obtained from the weighted average of the linear regression relations among all sectors when correlated with that specific bin [35]. Probst et al. treated the wind speed data corresponding to each sector as being analysed individually so as to find their correlation [43]. A general binning of the wind direction sectors happens normally at 8, 12, 30 or 45 [16].

The reference data can be derived from weather monitoring stations close to an airport or national weather services. As Probst et al. suggest in their work, the reference and target data used for the MCP analysis should be coming from similar heights. They continue by saying and that a difference in the heights could cause a reduction in the correlation coefficient between the two sites [43]. In a research conducted by Carta et al. the cross correlation of hourly mean wind speeds for 2010 at the Gran Canaria airport and some other wind installations on average 13km apart were examined. The reference station data were measured at one height (10m agl) but the target ones at different heights (10, 20, 40 and 60m agl). They found that the correlation coefficient between 10m agl for the reference and 20m agl. for the target was the highest one [18]. The data readings used are usually hourly wind speed averages for the long-term assessment but

there exist other possible data readings intervals such as a 10-minute interval, a half-an-hour one but also 1- or 2-minute intervals [18].

2.2.3 Comparison of MCP methods

Rogers compared four MCP methodologies [36]. These methods included the linear regression model, a model using ratio distributions of wind speeds at two sites, a vector regression method and a method based on the ratio of standard deviations of two datasets. The most popular of the analytical methods is linear regression developed by Derrick [37]. A refinement of this method was found by Woods and Watson [35] which uses again linear regression for modelling the wind speed but treats the wind direction in matrix bins.

Some other methods include the 'Variance Ratio' (VR) method [36] where its approach lies in the relationship of variances from the reference and target site. VR's advantage lies in the fact that it preserves the data variance whereas other MCP methods don't. Mortimer also developed a similar method where the wind speed is binned according to the direction sector and speed at reference site. The standard deviation of the ratios in each bin in a matrix form was taken into account and another matrix was created with the ratios averages [44]. According to Mortimer, this method could predict better extreme winds in comparison with linear regression.

Artificial Neural Networks (ANNs) were used for short term wind measurements in order to estimate the annual wind energy potential [15]. More specifically, one-year measurements were used from three different sites in Ireland to examine the annual wind regime using a training period of one and two months. The authors concluded that the ANNs method performed well in predicting the annual energy yield for both training periods thus using short term data could be a successful representative of such an analysis. They compared the results also with WAsP (Wind Atlas Analysis and Application Program) [45] and found that they were similar.

Carta et al. proposed another MCP method to estimate long-term wind speed characteristics at 6 wind energy sites located in the Canary Islands, Spain. The method

was based on the probability density function of the wind speed of the target site conditioned with respect to the wind speed at the reference site [38]. Then they compared their method with the VR method [36], the Weibull Scale method [12] and the joint probabilistic approach [46]. The results indicated a better estimation of the wind speeds in most of the cases however it was underlined by the authors that the degree of correlation between reference and target site is of great importance and hence it can pose a limitation in the quality of the results [47] as well as the effect of climate change [48]. Wind resource covers the lifetime of a wind farm (around 10-25 years) so if historical wind speeds are included i.e. from 40 years before then climate change is included and hence bias is introduced [49] [30]. Thus, the reliability of the resource assessment would be questioned. They also underlined the importance of examining wind regime and wind speed correlations in the regions of interest so that an appropriate MCP methodology would be used. They also noted that observed wind speed data of the target site may actually give better estimates rather than long-term estimations made by MCP techniques [38].

Three new MCP methods were evaluated and compared in another study with simple LR and the VR method [36] on concurrent synthetic wind speed data sets from two sites. Perea et al. [40] developed three new models, two based on conditional probability density functions (pdf's) with termed kernel methods named: Weibull pdf (Wpdf) and the simple linear regression pdf (SLRpdf) and one based on nonlinear regression with bivariate cumulative Weibull pdf's (WR). They investigated 5 metrics for all the different MCP methods: mean, standard deviation, Weibull scale factor, shape factor and energy density. The results indicated that the combination of the modelling approach and the parameter estimation both are reliable criteria for the choice of the most appropriate MCP method. The Wpdf seemed to outperform all the other methods and give the most accurate prediction for all the metrics and input data combinations but also portray in the best way the natural distribution of the data. Finally, they concluded that the Wpdf method performs with more accuracy than the VR method [36] even though the VR method still can predict very well. However, the drawback of Wpdf over the VR method is that it entails more programming cost.

In another MCP study, the authors [16] compared the linear MCP method based on the VR method [36] and ANNs comprised of Multilayer Perceptron Topologies (MLPs) [16]. They used 6 weather stations with mean hourly wind speed data spanning from a 10-year period on the Canary Archipelago in Spain, estimated their long term wind speeds and based on this knowledge, their energy costs. The uniqueness of their research lied in the fact that for the first time a linear MCP algorithm and ANNs where compared in the cost estimation per kWh of a wind turbine at a target site. Furthermore, the errors calculated were based on short-term and long-term data from the target site and were compared with each other. In general, the ANNs cost per kWh was lower than the linear MCP. The errors and hence the cost tended to be higher when using the short-term data as a representation of the target site.

Two particular neural network models, which have an efficient training algorithm and therefore are not tine consuming to reconstruct and predict time series, were applied to wind series reconstruction and predictions of a real wind farm in Guadalajara in Spain [41] named Group Method of Data Handling method (GMDH) and Extreme Learning Machine (ELM). The methods performed accurately and fast when compared with other well-known methodologies such as multi-layer perceptron (MLP) [16] and support vector regression algorithms. A software based usage of GMDH and ELM was also undertaken and indicated fast wind speed reconstruction and prediction from reference sites.

Weekes and Tomlin based their MCP research on short period wind speed data, only three months, for 22 UK stations [33]. They examined 3 different MCP approaches: simple LR, the VR method [36] and linear regression with Gaussian scatter (LR2). They concluded that using such small short-term data can introduce challenges such as the effect of seasonality but nevertheless, they can lead to successful predictions. For the seasonality specifically, they found that in the UK the lowest errors were observed when using autumn or spring data as their training period whereas the highest errors occurred for winter and summer. Since these MCP approaches performed very well in this case, i.e. using such a short period of data this subsequently can be quite beneficial to small-scale wind farm developers.

In another study conducted by the same authors, [39] an MCP approach based on the bivariate Weibull (BW) probability distribution of wind speeds pairs of correlation sites and a variation of the BW method, (BW2) were compared with simple LR and the VR method [36] for 11-year-long wind observations of 22 UK sites. In addition, they created 22 artificial wind data based on ideal BW distributions. Regarding the artificial data, the BW method performed better than the linear MCP methods but the contrary was the case for the actual wind data for short training periods. For training periods of 12 months, all methods performed in a similar way. Hence, they came to the conclusion that whereas BW performs better for artificial data, when used for actual ones the method might not work as well since data might not follow exactly the idealised BW distributions assumed.

Dinler [32] in his study tested a new MCP method, Multiple Principal Least Squares (MPLS) on hourly wind data from 4 different regions. The main advantage of MPLS lies in the fact that it can be applied even when there is a low correlation between the target and reference site and thus when poor quality of or non-concurrent data exist as well. MPLS was proven to be as good as the VR method [36] for 95% of the cases when concurrent data exist. It also performed well in the lack of concurrent data and for different lengths of data with 84% better predictions than the actual data. The method had a 40% improvement when using one-year or six-month data. According to the author, Principal Component Analysis (PCA) could identify discrepancies in wind speed data and be useful in order to extract signal from noise but however it may not give reliable predictions [32]. In this research the opposite will be proven for PCA standing as an MCP method with more details to follow in the next chapters.

Zhang et al. used an advanced hybrid MCP methodology with wind speed and direction as input variables for 6 reference stations at North Dakota USA using the years of 2008-2010 [42]. They examined two cases, in the first one each reference station used one of the established MCP algorithms 1) LR, 2) VR [36], 3) ANNs and 4) support vector regression and the best hybrid strategy of the MCP methods and station was assessed. It was found that the hybrid algorithm’s accuracy was influenced by the use of individual MCP algorithms and stations and that the best scenario was achieved

when considering the length of the correlation period. In the second case, both wind speed and direction were taken into account and the best correlation period was found to be from approximately 8 months to a year. Lastly, they found that the power generation was over predicted by ANNs, hybrid neural networks, LR, hybrid LR and hybrid VR methods. However, the opposite occurred for the support vector regression, hybrid support vector regression and for the VR method.

2.3 Alternative resource assessment approaches

The fundamental aim of MCP is to evaluate the long-term, climatological resource at the chosen location. MCP does this through linking a local measurement campaign to a climatological record from a specific nearby locations. One alternative approach is to use climatological information synthesised from many data sources and compiled in an atlas format. Another method is to use the results from climate models.

2.3.1 Atlas methods

An initial wind resource assessment can be made by using a database, such as the now NOABL free database [50] or the UKCP09 [51] for the UK, or the WAsP Wind Atlas [45] used virtually world-wide, where long-term wind data are used to report an average wind speed for a fairly large area. Such an approach only gives a rough indication as to whether a region may have, on average, a good resource or not. However, the accuracy of the prediction is well below that expected from developers, or the institutions providing the financing of the development. As a next step to improving this accuracy prediction, the regional average is refined by using computer model to simulate the flow through the proposed development site. They may use a Wind Atlas as an input and with some correction for the terrain and obstacles. Then, it creates a map where the location's wind regime can be obtained.

One common refinement tool is WAsP [45] which is widely used in the wind industry. These packages tend to work well for sites with fairly simple topography, but they are reputed to struggle to produce satisfactory results for areas with a complex

topography. Other commercial packages performing essentially the same service include WindSim [52], ZephyTools [53], Windie [54] WindFarmer [55], WindPro [56], and AWS openWind [57] which are explained in more detail in section 2.4 of this chapter.

Another resource assessment method that has been applied to wind resource prediction involves the utilisation of global atmospheric databases, with the most commonly used being from the NCAR (National Centre for Atmospheric Research) the NCEP/ NCAR reanalysis data [13] the ECMWF (European Centre for Medium-range Weather Forecasting) which includes datasets such as HRES, ENS [58] and from NASA the MERRA (Modern-Era Retrospective Analysis for Research and Applications) [59]. The Virtual Met Mast (VMM) [60] (developed by the Met Office) is a type of British Wind Atlas and atmospheric model.

2.3.2 Climate model methods

Other methods developed include statistical downscaling from climate models and dynamic downscaling from climate models. The climate models use mesoscale atmospheric results with high resolution to convert it to local atmospheric models and thus a series of nested downscaled models are created [61, 62]. Mesoscale modelling consists of a dynamical statistical approach to express global (large-scale) climatology into regional wind climatology. Typical mesoscale models are KAMM [63], MM5 [11, 64].

The usual problem with this type of methods is that they are very expensive in terms of computing resources and time. An advantage of MCP methods is that they give specific results for a location whereas many of the alternative methods do not. For example, WAsP is a fixed terrain flow model which produces an Atlas to give a wind resource estimate for a region rather than a site. In more detail, WAsP initially uses the observed wind at a mast to derive the wind resource at a terrain absence i.e. the wind atlas and then using the reverse procedure, it uses the background wind as an input for the wind profile prediction of other points [23]. If the local topography is simple it can give good results for a site, but it is less useful for more complex sites [45]. VMM is a

more sophisticated method than WAsP and uses topographic information combined with computational modelling but not local measurements.

Bowen et al. performed a study investigating the WAsP limits and found that the errors in predictions could be of importance if the terrain or climate are outside the standard conditions- for example, having a non-flat terrain and with extreme weather conditions [65]. Another study was conducted in order to compare the offshore wind resource for the German Bight between the mesoscale model and WAsP by Jimenez et al. It was found that WAsP can depend on the reference measurement stations to a large extend and that the wind profile used by WAsP for the North Sea is in good accordance. MM5 seemed to yield good results with a roughly 4% offshore deviation. Its main advantage is that there is no need to use measurement data for it [66].

Suarez et al. compared the wind speed predictions of three different methods, WAsP, MS-Micro/3 [67] and DAMS (Detailed Aspect Method of Scoring) [68] in a complex forested terrain in the Cowal Peninsula, Scotland. They concluded that all three methods yielded similar predictions and in general outperformed their expectations based on previous case studies. The possible reason behind this is that they considered wind direction which could compensate for overpredictions on a hill in relation with underpredictions of a wind coming from a different direction. However the prediction variability seemed to be larger for WAsP followed by MS-Micro/3 and lastly by DAMS. Thus they concluded that for WAsP, the variability of predictions tends to become high over small distances [69].

2.4 Industrial MCP tools

A lot of well-established industrial software packages performing MCP analysis have been developed to fulfil the wind energy industry's needs for good resource assessment. A very well-known software package that performs sectorwise linear regression is WindFarm [70]. It accepts inputs in the form of time series used for the concurrent data or of a long-term frequency table used for the historic data. Since the least squares method only takes into account uncertainty in the vertical axis, the best-

fitted straight line in WindFarm is found through orthogonal linear regression, also known as the York method [71]. Since the least squares method which is explained below cannot be applied, orthogonal regression becomes more appropriate. The orthogonal regression estimators can be found by the minimisation of the perpendicular distance between observations and fitted line. The output given is presented as sectorwise wind distribution but also as an overall mean wind speed. Furthermore, it enables the predicted mean wind speed value to be directly compared to the measured one.

Another powerful MCP package is WindPro [56] which uses among other techniques least squares regression and bases its predictions on a Weibull distribution fit. The least squares regression is attempting to minimise the vertical distance between observations and fitted line. However, since the predictions made are based on a Weibull fit, it is likely that they could include error and thus uncertainty. The input data are derived this time not from times series but from the Weibull distribution and the output mean wind speed prediction is now not compared with the actual measured one but with a calculated one originating from frequency tables.

From the empirical methods, the Inhouse Matrix Tool developed by the renewable energy consultancy company SgurrEnergy Ltd. builds a correlation matrix between reference and target site for each sector of the wind direction which is then applied to historical data for prediction. WindPro [56] also includes a matrix approach, in which the concurrent data are grouped in bins in order to define a matrix based analysis of the behaviour between the reference and target site. Thus, in order to smoothen the pattern of the behaviour obtained, normal distributions are fitted to the data. The final stage includes the use of the smoothed surface to transfer long term data from the reference to the target site. This is undertaken by the Weibull- Monte Carlo fit, to the data which are computer based [71].

A more comprehensive review was carried out by SgurrEnergy Ltd [71] which compares six MCP techniques. The MCP techniques used for this analysis were the WindFarm and WindPro linear regression, the WindPro residuals and WindPro matrix and the Inhouse matrix and Inhouse WindFarm. The data originated from wind masts

situated at potential wind farm sites but meteorological station data were included too. At first it was essential to determine an overall site correlation for each pair of sites for all the datasets included in the analysis. That was achieved with the use of least squares regression for all the concurrent data of each pair and r, the correlation coefficient was determined so as to quantify this relationship. However, the overall site correlation found was different from the correlation obtained by the MCP analysis results.

A brief explanation on the procedure of the comparison of each MCP technique used for several pairs of sites is given next. Firstly, a pair of two year reference and target site concurrent data was chosen. Then, for each pair the first year of concurrent data was used in order to explore the sectorwise MCP relationships between the two sites. Furthermore, the relationships obtained were then applied to the second year of the reference data to enable a prediction for the second year of the target sites wind resource. A way to verify the results obtained by the MCP relationships was with the comparison of the predicted mean wind speed with the targets measured mean wind speed from the second year. In addition, the application of the measured and predicted wind speed against a generic power curve yielded mean energy outputs which were then compared and indicated the accuracy of the predicted against measured wind distribution.

The results of the undertaken analysis indicated that the WindFarm MCP tool performed well consistently under most types of analysis and yielded accurate results. The Inhouse Matrix tool also had a good performance but indicated sensitivity to the site correlation. Furthermore, the Inhouse WindFarm technique seemed to result in a better long-term analysis than the Inhouse Matrix. WindPro Matrix also performed well but its drawbacks were that it could give different answers if the same datasets were to be reanalysed. Finally, out of all the analysed techniques, the WindPro linear ones were the ones that performed poorly and therefore are not recommended.

2.5 Main challenges for resource prediction

2.5.1 Wind uncertainty

Apart from the MCP techniques comparison results, some other significant observations were made regarding the MCP methods. In some cases it was found that the underlying MCP assumptions did not hold; more specifically the relationships established between the two sites were varying depending on the concurrent year which could result in the existence of uncertainty in the MCP methods used. Hence, the uncertainty of a good prediction between two sites for future wind resource exists since the analysis indicated no ideal method to overcome this but at the same time the uncertainty that has been accounted for during the MCP analysis was not exceeded [71].

Frandsen and Christensen [72] analysed the theory and evaluated the uncertainties based on several parameters for the annual power curve output of a turbine. They also indicated how to combine different types of uncertainty. It was found that the uncertainty in many cases could be ranging from 10 % to 15% on power curve determination and above 20% for wind resources. As for production, considering various Danish wind turbines the production estimation was well fitted on average but the standard deviation of actual and predicted production was high: 20%-30%. The authors emphasized on the fact that the economic viability of wind power will highly depend on the future relation of interest rate and the prices of fossil fuels.

An uncertainty analysis was conducted by Lackner et al. in terms of wind resource assessment and energy production estimation [24]. The authors examined three major aspects related to uncertainty: wind resource, wind turbine power output and losses and finally the AEP (Annual Energy Production) uncertainty which was accounted for with a new method based on the Weibull distribution. Since they used sensitivity factors to combine different uncertainty causes, it was found that the sensitivity factors related to wind speed measurement and Weibull factors uncertainty can be accurately accounted for when this method was used. Thus the advantage which arises is that the site assessment uncertainty can be derived more accurately.

Jung [73] performed an analysis for two sites in the Korean peninsula to evaluate the uncertainty based on the variability which characterizes the nature of wind energy. He firstly proposed probability distribution models which included wind characteristics such as mean wind speed, Weibull parameters, air density etc. Then he created an empirical probability model based on a Monte-Carlo simulation for the power curve performance. It was found that the aforementioned method performed well in quantifying the annual energy production and the uncertainty can successfully be assessed when considering the site characteristics which can be obtained from the short-term measurements of the target site in the form of probability parameters.

In another study performed in Austria, the authors used statistical simulation methods to come up with profitability calculations based on wind speeds and uncertainty [74]. They used the VR method [36] to obtain wind speed estimates and then the Conditional Value at Risk (CVaR) as a risk measure for profit returns. The originality of the method was based on the fact that measured wind speed distributions, uncertainty and the CVaR measures were used successfully to assess the risks of wind profitability for the first time.

Messac et al. [10] presented a new method which characterizes uncertainties in the annual wind distribution predictions and models these uncertainties with respect to overall wind farm performance and local wind power density (WPD). They used a 10 year period for two sites; one onshore and one offshore and developed two uncertainty models; a parametric and a nonparametric one. They investigated the period of payback, annual energy production (AEP) and cost of energy (COE) regarding the wind farm performance and found that the WPD was 30% for the offshore site and 11 % for the onshore site. It was found that wind speed and direction uncertainties are not proportional to annual predictions of the same conditions. Therefore they remarked that it should be taken into a great account how these conditions occur from year to year but also in the long term when it comes to designing a wind farm.

2.5.2 The example of the year 2010

In order to understand further the significance of the good quality resource assessment as mentioned in Chapter 1, there is a representative example of wind speed measurements in 2010. The wind statistics that year indicated significantly low wind speeds in comparison with the ones in previous years which caused a lot of insecurity in general regarding the wind industry [1]. Investors, wind farm developers and others directly and indirectly affected in the wind industry have been concerned about the 2010 issue and have been seeking answers regarding what will follow in the future.

Regarding 2010 as a significant low wind year, several questions have arisen and been of concern to many companies and investors in the wind industry. These questions are for example: should 2010 be included as a reference site historical data period in order to predict wind resource? If yes, is there a trend created from including 2010 in the analysis and what kind of trend could occur? (over/under prediction) Is there bias? What is the effect of using low or high wind speed years in a concurrent dataset? In a similar situation but with the opposite wind affects i.e. extremely high wind which was observed in 1993 further analysis that have been carried out indicated that including 1993 in the datasets did in fact tend to over predict wind resource.

Früh investigated this specific year and found that it did not fit the overall trend of Scottish climate. However, he emphasized on the need for continuous research regarding the climate indicators and wind observations connection [30]. In order to start answering these significant questions some analysis has been carried out with the use of two different MCP methods, the WindFarm [70] linear regression and the matrix method. The datasets were including 2010 as a historical year from a total of a ten-year historical period (2000-2010) and as a concurrent year alongside other concurrent year models such as 2008 and 2009.

Furthermore, the same analysis has been carried out excluding 2010 from the concurrent and historical data period so as to observe any significant results [75]. Figure 5 indicates a summary of this analysis, which shows the ratio of the predicted mean

wind speed over the actual wind speed for the two analysis methods against the prediction period.

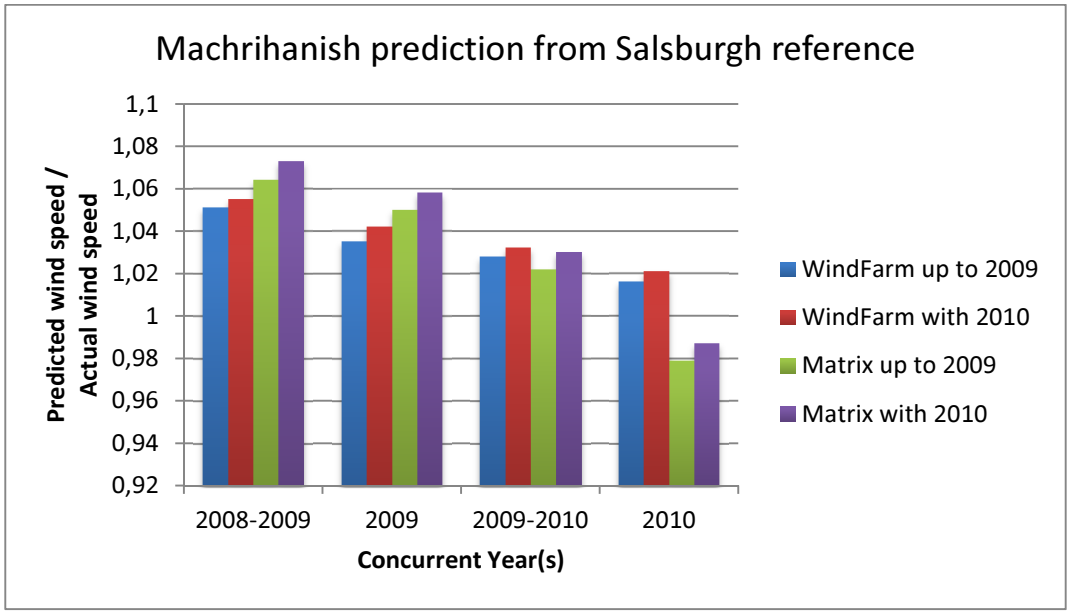


Figure 5. MCP Results using WindFarm and Matrix method for Machrihanish target Salsburgh reference for the historical period (2000-2010).

As it can be seen from Figure 5, when the ratio of wind speed exceeds 1 the prediction periods which fall fully or partially within it can be interpreted as over predicting the wind speed. On the other hand, for example for the 2010 concurrent dataset the matrix method for both long term models seem to under predict wind speed since the wind speed ratio is below 1. The highest over predictions can be observed for 2008-2009 and 2009 concurrent models with minor variations within all methods and historical datasets. Hence from this dataset analysis the question of whether the inclusion or not of 2010 in the historical and concurrent data could under predict wind speed cannot be clearly verified since in most concurrent datasets the inclusion of 2010 in historical data but also in the 2009-2010 concurrent model does not seem to under predict wind speed. However, that seems to be the case for the model including only 2010 as a concurrent dataset but not for both MCP methods. The general conclusions that can be made from this example are that for the 2010 concurrent model the linear method seems to over predict but the matrix method seems to under predict irrespective of the length of measurements. Thus, regression seems to be less sensitive in terms of

the inclusion or exclusion of 2010 for correlation however that is not the case for the method itself.

2.6 Opportunity for improvement

The following remarks can be drawn regarding MCP methods. An ‘easy’ site is usually predicted by all methods quite well. With ‘difficult’ sites, one can sometimes conclude that they are difficult to predict as different methods return different prediction, and it is not clear which is the best prediction. With other ‘difficult’ sites, different models have given similar predictions but those predictions were wrong. One of the fundamental assumptions and therefore potential limitations of MCP methods is that they all assume the existence of a relationship between the two sites, reference and target, and that this relationship does not change over time.

One of the key limitations of analytical MCP methods is that they also specify the relationship rigidly, hence it is restricted. On the other hand, matrix or distribution methods predict only the distribution curve (which is also often referred to as the frequency table of wind speeds) [9]. By doing that, the amount of information used to specify the relationship is reduced, i.e., not all possible information is utilised in the prediction. Furthermore, some distribution methods assume explicitly a specific shape of both distributions, such as the Weibull distribution [12], which can also be a limitation.

The main conclusion which can be made is that MCP are methods which can be used for wind resource assessment by wind farm developers without being highly complex expert systems. Furthermore, they could be further improved so as to become more accurate, minimise bias and estimate the reliability or precision of the prediction. Factors such as the trends of the constant climate change which includes the climate variability and oscillations on a multiple year time scale are essential for the MCP analysis to be identified [30]. The wind related data could be treated as dynamical systems so that cycles and random unusual behaviours that often characterise them can be identified, explained and understood. Thus, there would be a lot of benefit in having

a common tool that is capable of identifying trends, climate cycles and true outliers. Current industrial experience with available MCP methods is that there is no consistently best prediction, and that a convergence of the prediction from different methods to a common answer is not necessarily a conversion to the correct answer [1] .

2.7 Reason for improvement

A study was conducted via a questionnaire named WAUDIT (Wind Resource Assessment Questionnaire) [76] for the first half of 2010 which was addressed to Europe based wind analysts in academia and industry with a response from 72 people from 48 different organisations. The conclusions drawn were that there is a need for developing more remote sensing instruments for wind measurements purposes but also examine the used models for more complex terrains and for offshore wind purposes. Furthermore, turbulence models should be taken into account since they are related with turbine wakes. Finally, the need of validation and further development of the wind resource assessment techniques is of vital importance [76].

A novel MCP technique was developed which is based on the statistical methodology of Principal Component Analysis (PCA). This new PCA-MCP method is designed to capture the relationship between the target and reference site empirically without enforcing common assumption such as linearity in that relationship. This method is also in the general framework of MCP in that it assumes a fixed relationship between the sites but does not specify the shape of the relationship. Instead, it allows for a selection of empirical relationships to be combined for the prediction and selects the best predictor for different weather types, where both, the predictor and the appropriate weather type, are identified through the PCA algorithm. PCA explores the interrelationship of the reference and target sites and is used rather than 'assuming or using' a fixed relationship of time-delays; it trains the optimum relationship between the two sites. This new method is based on the theory of Dynamical Systems and extracting

[1] S.Quinn, SgurrEnergy Ltd., pers.comm., 2014

the optimum signal using PCA. Chapter 3 will illustrate all the theoretical background behind this new MCP method.

Chapter 3 Development of Principal Component Analysis as a forecasting and MCP method

In this chapter the theory of Principal Component Analysis for optimising the time series analysis of a dynamical system will be explained, and the extension of the method for the purposes of this research, namely forecasting and MCP, will be developed.

3.1 Dynamical systems

A *dynamical system* is used to model physical phenomena whose state (or instantaneous description) changes over time [77]. The system is described by fixed and deterministic rules and, in order to describe those rules, the space where the system evolves geometrically has to be defined. Their applications range, among others, from financial and economic forecasting, environmental modelling to medical diagnosis. Their applications can be divided into three main categories: predictive, in which the future states of the system are being predicted with the use of past observations and the system's present states, diagnostic, where the aim is to investigate what possible past states (or observations) of the system might have led to its present state and finally applications where the aim is neither predict the future nor explain the past but actually explore the theory of a physical phenomenon or the underlying dynamics.

3.1.1 Phase space

The dynamical systems involve differential equations that depend on position and momentum. A simple example of a dynamical system is the linear pendulum which can be derived from the equation (Hooke's law applied to Newtion's second law)

$$\begin{aligned} F &= -kx \\ F &= ma = m\ddot{x} \\ &= -kx = m\ddot{x} \end{aligned} \tag{8}$$

and can be re-arranged to the second differential equation of the form

$$\ddot{x} + \frac{k}{m} x = 0 \tag{9}$$

that leads to the dynamical system of two coupled first-order differential equations

$$\begin{aligned} v &= \dot{x} \\ \dot{v} &+ \frac{k}{m} x = 0 \end{aligned} \tag{10}$$

where v, x, α are the velocity, position and acceleration respectively and k is the constant factor characteristic of the spring. As it can be seen from the relationships in (5) their form is quite simple and they indicate the change of v, x given their current condition. Thus the system can be described as deterministic since no random equations exist i.e. the future changes of v, x can be predetermined [78]. However taking into consideration specific cases of dynamical systems, it is often possible to observe irregular behaviour in some of them, for example in the Lorenz [79] attractor case. Other important definitions regarding dynamical systems are: the *phase space* which describes the system's variables, the *attractor* which defines the actual solution of the system and finally the *orbit* which is the path that the system follows during its evolution.

The principle in terms of a dynamical system is that the dynamic evolution of the system takes place on a time-invariant object, called 'attractor', after initial transients have decayed. This attractor is a geometric object in the phase space defined by the dynamic variables of the dynamical system.

3.1.2 Time-delay method

A method is needed so as to define equivalent variables to the phase space ones which is the *time-delay* method [80]. It is a practical implementation of the dynamical systems since it aids in reconstructing the phase space of a dynamical system from an

observed deterministic time series. The reconstruction of a phase space is indeed significant since it can extract useful information about the time series that characterise the system. Using previous measurements is equivalent but not practical with data containing noise or turbulence [81]. In complex systems, where the phase space is not fully accessible from measurements, one can use Takens' method of delays [80]. The phase-space equivalent variables can then be constructed using Takens' Method of Delays [80], which postulates that the dynamic variables not directly measured have influenced the evolution of the measured variables and are therefore somehow represented by the previous measurements. Thereby, a sufficient representation of the state complete phase space at time t is given by the delay vector $y_1(t), y_2(t-\tau), y_3(t-2\tau),..., y_j(t-M_w\tau)$, where M_w is the number of time lags, τ, used, and the same can be done for further variables measured, e.g., $y_2(t)$.

With a time series of N_0 variables of length N_t, the delay matrix will have $N = N_t - M_w\tau$ rows and $M = N_0 M_w$ columns with

$$Y^{i,j,+(j_0-1)M_w} = y_{j_0}(j+(i-1)\tau) \tag{11}$$

with the row index $i = 1,..., N$, the column index $j = 1,..., M$, and the observable index, $j_0 = 1,...,N_0$ [81]. In this matrix, a row m is equivalent to a complete phase-space description of the system at time t_m as long as M is sufficiently large. Taken's method of delays is therefore able to create a space equivalent to the phase space but this phase space reconstruction cannot separate the important dynamics from measurement noise or turbulence.

3.2 Principal Component Analysis (PCA)

PCA is a non-parametric statistical method which can optimize phase space reconstruction [82]. By non-parametric it is assumed that it is a method not limited to be of a certain distribution or linear relationship. It can identify the number of needed time-delays and give a picture of their shape. It is also known as Empirical Orthogonal

Function (EOF) Analysis in the Meteorological and Oceanographic community to identify the main circulation patterns in the atmosphere and oceans, e.g. [83, 84]. This technique is now widely used for time series analysis of nonlinear dynamical systems in general, e.g.[80, 85] as the analysis is very powerful to separate coherent dynamics from noise. PCA uses samples of data whereas EOF uses sets of spatial images and Singular Systems Analysis (SSA) time series with a size of 1 observation which later extended to more observations. All of them use the principle of the Singular Values Decomposition (SVD). SVD is a mathematical matrix operation technique which is described by the same theory as PCA but from a linear algebra point of view whereas PCA comes from a statistical point of view.

PCA's goal is to explain important variability of the time series data and to extract useful information (i.e. hidden structures of the data) from its more relevant components in a reduced number of dimensions. Applying PCA to the set of delay time series is a method to redefine the phase space to concentrate the coherent information in a few directions (or dimensions) of the phase space, which then allows to 'delete' the weaker and uncorrelated dynamics from the description of the system.

PCA's advantage lies in the fact that it can separate noise from useful information applied to time-delay series [80]**.** More specifically, PCA was devised to separate coherent dynamical information from noisy experimental data, known also as SSA [86] [87] .

The mathematical procedure to carry out a PCA is through SVD of the delay matrix. In terms of the linear algebra of the SVD, it is a transformation of the basis vectors of the phase space which finds orthonormal basis vector to maximise the variance described by as few basis vectors as possible. The three SVD/PCA outputs are the *singular vectors* which are the basis vector for each dimension (they are also the eigenvectors of the covariance matrix of Y), the *singular values* which measure the time-averaged contribution of each dimension to the total variance**,** and the *principal components* (pc's) which form an attractor and describe the system's time series. In matrix notation, the Singular Value Decomposition is written as

$$Y = P\Lambda S \tag{12}$$

where $Y(n,m)$ is the *time-delay* matrix with $n = 1,..,N$ the time point within time series and $m = 1,..,M$ the index of the dimension. $P(n,m)$ is the principal component matrix, $\Lambda(n,m)$ is the diagonal matrix of singular values, and $S(m,m)$ contains the singular vectors.

The singular values represent a measure of the variance, more specifically the square root of the variance of the time series in the corresponding dimensions and they can pick out the important variability of the data. The singular values represent the square root of the eigenvalues of the covariance matrix C, of Y in equation (12), $C = Y^T Y = S^T \Lambda^2 S$. If the training data set consists of N_0 variables, $y_{j0}(t)$, for example wind speed and wind direction with $N_0 = 2$, covering N_t time steps, the first step is to rescale them in such a manner that they both contribute equally to the analysis. This is achieved by rescaling them both to time series of zero mean and unit variance, i.e., subtracting the mean from each variable in turn and then dividing by the variance. The singular vectors have the property of being orthonormal, i.e. orthogonal and of unit length and they span the dimensions of the phase space. They represent a measure of those dimensions that define a dynamical system, for instance they can replace position and momentum, two variables which can form a dynamical system. The singular vectors, S are also the eigenvectors of the covariance matrix of Y in equation (12). The principal components are the time series of the system in the coordinate system defined by the singular vectors. This means that plotting the principal components against each other draws the orbit of the measurements and thereby provides an estimate of the underlying attractor. They represent a measure of those dimensions that define a dynamical system, for instance in the aforementioned example of section 3.1.1 they can replace position and momentum as variables.

When PCA is applied to the time-delay matrix, PC's are the time series of the coordinates of that trajectory in respect of these dimensions. Using the example of section 3.3.1 again, PC's can replace the values of the position and momentum at any time. In more detail, this dynamical system's position of the reconstructed phase space

can be given at any time precisely by position and momentum however when PCA is applied the PC's take over this role. Since an eigenvalue matrix exists in PCA analysis, it should be noted that both eigenvectors and PC's are normalised i.e. scaled to the amplitude of the dimensions used by PCA.

3.3 Application of PCA for time series analysis

3.3.1 Single variable series

One of the first attempts to apply PCA for phase space reconstruction to a clean system was undertaken by Broomhead et al. [86]. In order to achieve that, the application of SSA to time series generated by passing sinusoidal signals through a cubic non-linearity i.e. coming from nonlinear dynamical systems was used. The number of degrees of freedom resulting from SSA enabled the identification of the dimension of the reduced subspace of the series. After this in the same subspace, a more extensive analysis was used in order to discover underlying patterns that contributed to the motion. Thus, the use of this methodology for this specific time series was successful since it enabled useful information for the system to be extracted.

3.3.2 Multivariate variable series

As Broomhead et al. concluded [86] SSA offers a high potential yet to be explored. Thus, the next step in PCA/SSA analysis was developed by Read [81] and was to apply SSA in multivariate data series (M-SSA). He firstly applied single variable SSA in data from one probe obtained from the full set of experimental time series of temperature measurements with sixteen probes in total which were part of a rotating thermal convection experiment. Furthermore, the SSA results were used to characterise the different types of flow and their dimension correlation. Then, he applied M-SSA to multiple probes and found that it was able to improve the reconstructions of the signal, in terms of noise ratio and uniformity of the attractor's structure in comparison with previous methods. Thus, he concluded that the M-SSA superiority lies in the fact that it

can combine information simultaneously from all spatial cross-correlation functions of complex spatio-temporal structure signals.

The method of M-PCA has been used in other cases, too. Früh [88] used spatially extended times series for temperature measures to identify the most important vortex patterns in a rotating fluid experiment. He achieved this by separating steady drifting fluid patterns as the key mechanisms in the onset of variability. It succeeded in extracting specific non-linear wave interactions leading to chaos and disordered flow, e.g. velocity images.

3.3.3 PCA between two signals

Allen and Smith [89] used some simple stochastic sinusoidal system and applied singular PCA. They used PCA on test series and performed analysis on noise data so they created differed time series and correlated noise. Moreover, from the extraction of noise they managed to obtain a confidence interval (CI) which had the form of a decaying spectrum containing eigenvalues. They found that the eigenvalues that lay outside the CI contained useful information.

Furthermore, in the second part of their experiment, they applied the same method to real temperature climate data and to the eigenvalues outside the CI. They attempted to refine the sign of difference and to look into it in relation with noise. Finally, they separated the random fluctuations originating from the signal and performed PCA again so as to compare maximum signal above the fluctuations.

3.3.4 PCA for combined system

The aforementioned attempts that used PCA justified its importance as a method. Additionally, it has had a wide range of applications that involve current big issues of general significance. These applications include the analysis of temperature time series as used by Allen et al. [90] and the detection of global climatic changes as used by Allen et al. [91]. Hardy and Walton [92] used PCA for a one-year record of mean wind velocities from 10 different locations. They found that the method can be used

successfully for large datasets of wind velocity data and can extract the useful spatial and temporal properties of the data.

Benestad [93] used common EOF's for statistical downscaling purposes of future climate scenarios. According to his research, the method has the advantage of minimizing the errors related to the downscaling procedure hence this method is recommended for downscaling purposes. In another research, Guillou and Dreverton [94] generated daily time series for weather-derivatives market purposes. PCA was used in order to analyse daily average temperatures of each year and it was found that the interannual variance of the climate was captured correctly. In another research, Martinez et al. [95] used EOF for generation of large-scale atmospheric component patterns using an NCEP-NCAR dataset for the years 1958-2004 for the region of Gaspe in Quebec, Canada. They concluded that the method is able when using large-scale data to generate time series at a regional level and accurate numerical atlases to an extent. The success of the method according to the authors lies in the fact that it can summarise statistical information to a few most dominant patterns according to the variance explained. Also the fact that it can construct time series regionally is very important for the industries that base their research on daily time series of for example wind or temperature.

Moreover, PCA has potential of further extension of its applications since not only can it be used to analyse one or multiple variables for one experiment but it can also be applied to analyse different but coupled systems. These coupled systems could be treated as one large system containing two sub systems that include the variables from the first and the second system. With the use of PCA for a combined system some of the main challenges of the MCP methods mentioned in section 2.2 will be hopefully overcome. More specifically, it will contribute to the reduction of uncertainty due to poor quality data but also to the reduction of bias since years such as 2010 might not affect the analysis so as to result in under or over prediction of wind resource.

The following section depicts the rationale behind a coupled system with the use of a simple dynamical system which is in fact the initial implementation of the method and

continues with analysis of real wind resource data. In fact, it gives a first impression of how well PCA can stand as an MCP tool.

3.4 A first illustration of PCA on a dynamical system

In the example of a harmonic oscillator, the phase space is defined by the position and momentum of the oscillating object, and the motion of it takes place on a limited cycle. This cycle is the attractor, and the trace drawn by the oscillation, or its 'orbit', would draw repeating copies of that cycle over and over. An illustration is shown below using a simple oscillator.

3.4.1 Case A, fully quasi-periodic system (noiseless pendulum)

The first implementation of PCA was attempted to a two signal pendulum divided into two cases, with or without noise. The pendulum was chosen since it is a case of a simple dynamical coupled system and thus illustrative of the dynamical systems theory. This example is an idealised case of the pendulum consisting of two variables (signals) x, y representing signal 1 and signal 2 respectively which could represent the reference and target site respectively. The two different cases were selected in order to investigate how the PCA was influenced by the existence or not of noise. The phase space in the pendulum's case is defined by angular displacement and velocity i.e. the two dimensions of the dynamical system. Performing PCA for this example can characterise x, y together as a linked system. More specifically, with the application of PCA to x and going back to the description of the combined system found, it can be identified which y has the best fit to the linked system.

In more detail, the time interval of the pendulum's movement was set to be from 0.1 to 1000 seconds and several lags and window sizes were used as inputs so as to examine the PCA results in both cases. The range of values that were examined for both signals x, y for the lags were from 1 to 25 and for the window length from 10 to 100. Different values were also used for the noise term, 0.6ε in equation (14) of case B

ranging from 0.1 to 1. In case A when excluding noise, the pendulum inputs were of the form:

$$x = 3\sin\left(\frac{t}{0.7}\right)$$
$$y = x + 0.4\sin\left(\frac{t}{\pi}\right) \tag{13}$$

As it can be seen from Figure 6, the top right graph shows the singular values that result from the PCA analysis. Two large singular values followed by 2 more can be observed here which means that they dominate in terms of their contribution as dimensions to the total variance. It can also be seen that the first two largest singular values pick out the highest frequencies i.e. the most important movement of the pendulum and the next two the lower frequencies i.e. the more hidden patterns. The first graph of the first row indicates the periodic movement of y overtime and the two different period movement which originates from the two signals. Oscillations can also be observed in the top and bottom ends of it. The middle graph of the first row is a picture of the time-delay phase portrait and its circular shape represents the frequencies of the pendulum with the thickness of the orbit line indicating their modulation.

The first two graphs of the second row show the singular vectors and both contain the two signals at the same time. The sinusoidal pattern that characterizes them is due to the existence of the sine function in the input equations for both signals x, y and the singular vector 1 plot indicates a perfect sine curve. The first graph of the third row depicts the PC's P_1 versus P_2 and it is of torus shape. It summarises the whole pendulum's periodic movement and is an amplified version of the time delay matrix graph described above. The last two graphs of the third row which also depict the rest of the PC's versus each other are representing the same periodic shape as the P_1 versus P_2 graph but illustrated from a different angle.

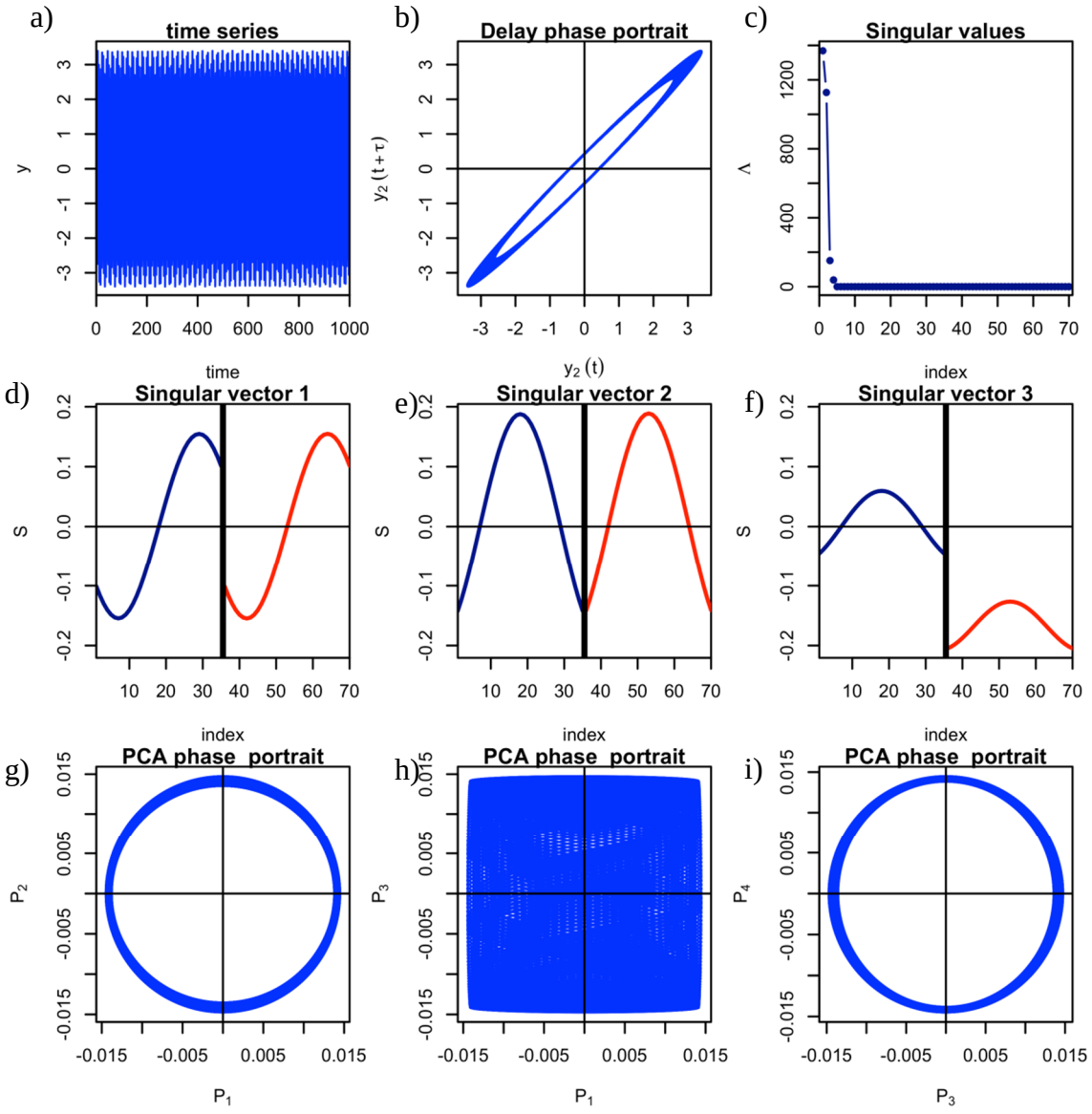


Figure 6. MCP results for noiseless pendulum (case A) with lag 1, window 35.

3.4.2 Case B, noisy oscillations

The next case in the pendulum example, case B, included the term 0.6ε as the white noise and the inputs were of the following form:

$$\begin{aligned} x &= 3\sin\left(\frac{t}{0.7}\right) \\ y &= x + 0.4\sin\left(\frac{t}{\pi}\right) + 0.6\varepsilon \end{aligned} \tag{14}$$

Figure 7 indicates the results from case B and it can be observed that there are some differences in comparison with the case A results. As it can be seen, if the two singular values plots are compared (top right graphs), in case B there is a little gap in the measurements in the first part of the flat line and this is due to the existence of noise in the system. However, the two large singular values lie further up from the rest in both cases. This means that the highest frequencies of the pendulum are being picked out with or without the existence of noise. As it can be seen in the top left graph of the time series the movement of y over time in case B appears to be less periodic than in case A which is due to the existence of noise. The time-delay phase portrait in the middle of the top row of case B due to noise again does not retain the clear circular shape like the one of case A. On the other hand, the singular vector graphs of the second row are of the same shape in both cases A and B. However, the PC's plots as shown in the third row do not have an identical shape as in case A. Some broadening in the frequencies is observed due to noise existence but the general periodic movement can still be observed.

A few initial conclusions could be drawn from this first implementation of PCA. It was found to be robust and useful method for time series of multiple inputs since both cases managed to extract the most significant outputs. It can be observed that even in the presence of noise; PCA can pick out the most important oscillation patterns by looking at the second row of Figure 6 and Figure 7 which are identical. Moreover, the main oscillations are indicated by the 2 dominating singular values of the top right row graphs of Figure 6 and Figure 8 and can reliably reconstruct even in the presence of noise the PC's for example looking at P_1 versus P_2 of the bottom left row graphs which are identical for both cases A and B. Thus, it can be concluded that noisy or 'clean' data do not play a significant role in PCA. Since for several trials of different time-delay and gap of entries in the data for the pendulum example the results were kept unchanged, it can also be concluded that the choice of the time-delay length and gap of entries in the matrix did not seem to play a significant role in the PCA results for this specific system.

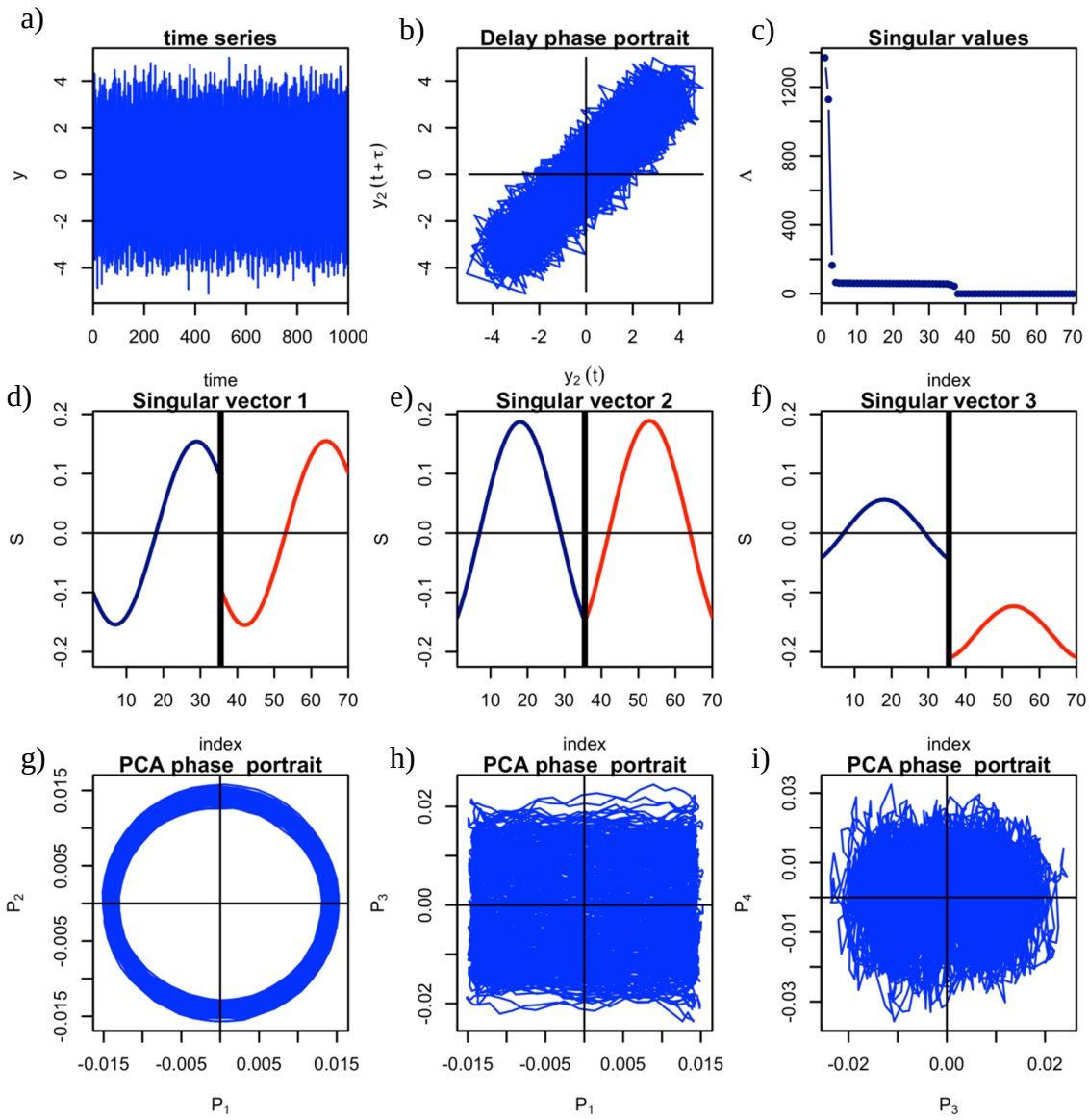


Figure 7. MCP results for noisy pendulum (case B) with lag 1, window 35.

3.5 PCA used for forecasting wind energy resource

Underlying all statistical and empirical approaches is the need to separate the predictable component from the turbulent component in an effective and efficient manner. For example, for mean daily or hourly wind speed forecasts, i.e., short-term horizons, the underlying atmospheric dynamics become of great importance [96]. The wind related data could be treated as dynamical systems so that cycles and random unusual behaviours that often characterise them can be identified, explained and understood. Based on this understanding, PCA was proposed to be used as a time series

analysis technique based on the dynamical systems theory for wind forecasting purposes.

The challenge that arises from the previous chapters is that if we have measurements from only one site, can we use similar analysis concepts to identify the state of a combined system on the phase space? If so, can we then predict for the second site, for which we have half the information we need? More precisely, the question that arises is by taking the defined points of the combined two sites system and adding the new measurements can we project them to the existing attractor and predict from the nearby points?

The creation of this system based on a training set of wind data defines the model for the forecasting. New measurements can then be mapped onto the cleaned-up attractor to find previous measurements which are, in dynamical terms, similar to the current measurements. Finding one or more ‘similar’ previous measurements, then allows us to the evolution of those measurements as equivalent to predicting the current measurements. In addition to a prediction, however, this method predicts a number of similar events and following how their distances change over the lead time of the prediction also provides a measure of how sensitive the system is to uncertainties in measurements or out-of-system perturbations. Hence, it provides a measure of the uncertainty of the prediction at the same time.

This section contains background information regarding phase space reconstruction as well as PCA and explains in detail how they will be used for the forecasting purposes. The extended results of this section will follow in Chapter 4. The stages for the training of the predictor are preparation of the phase space using the training set of data (e.g., wind speed and direction), PCA of the phase space to optimise the phase space and truncation of the phase space to the relevant components only to define the predictor.

The application of the predictor goes through the preparation of the test data to the same specifications as the training set, mapping the test data onto the truncated phase space, finding an ensemble of nearest neighbours on the attractor as defined by the test

data, tracing the evolution of that ensemble for the lead period of the prediction, and finally re-transforming the ensemble of predictions into the original variables (e.g., wind speed and direction). A summary of the PCA forecasting algorithm is presented in Table 2, and the remainder of this section will describe each of these steps in turn.

Training:	
1) Normalise wind speed measurements	$y_{j0} = \frac{\left(y^*_{j0} - \mu_{j0}\right)}{\sigma_{j0}}$
2) Create time-delay matrix; equation (11)	$Y^{i,j+(j_0-1)M_w} = y_{j0}\left(j+(i-1)\tau\right)$
3) Perform PCA to optimise; equation (12)	$Y = P\Lambda S$
4) Truncate to the relevant components to define predictor; equation (15)	$Y_t = P_t \Lambda_t S_t$
Forecasting:	
5) Normalise new measurements using Training normalisation	$y_{jn} = \frac{\left(y^*_{jn} - \mu_{j0}\right)}{\sigma_{j0}}$
6) Create time-delay matrix using same parameters as for training	$y_{j0} = \frac{\left(y^*_{j0} - \mu_{j0}\right)}{\sigma_{j0}}$
7) Map time-delay matrix onto attractor coordinates; equation (17)	$P_n = Y_n S_t^T \Lambda_t^{-1}$
8) Find number of similar events in training period and follow evolution of past events i.e. nearest neighbours; equation (18)	$d_i = \frac{1}{n_x}\sum_j \left\lvert P_n^j - P_t^{i+j-1}\right\rvert$
9) Find distance vector due to n. neighbours; equation (19)	$D_j = P_t^{kj} - P_n^{nx}$
10) Use ensemble prediction based on n. neighbours; equation (20)	$P_f^j(T) = P_t^{kj+T} + D_j$
11) Map back to delay matrix and return predicted wind speed; equation (21)	$Y_f^j = P_f \Lambda_t S_t$
12) Re-scale back to proper units	$y^*_{jf} = y_{jf}\sigma_{j0} + \mu_{j0}$

Table 2. The PCA forecasting algorithm.

3.5.1 The Forecasting Model

The singular values are a key measure on which the determination of the best predictor is based, since our initial assumption was that the wind conditions several hours ahead is better predicted by the slower atmospheric dynamics than the short-time fluctuations. The PCA has separated the coherent (slower) dynamics from the temporally uncorrelated short-term fluctuations, such that uncorrelated fluctuations are visible as a noise floor in the singular value spectrum. Persistent variance from the atmospheric dynamics is concentrated in the leading singular values of much higher magnitude. For that reason, the phase space can now be truncated to a much smaller dimension than the original delay matrix.

By creating a reduced set of M_t principal components, P_t^{N,M_r} singular values, $\Lambda_t^{M_r,M_r}$ and singular vectors, $S_t^{M_r,M}$, one can produce a filtered time series of the original data by

$$Y_t = P_t \Lambda_t S_t \tag{15}$$

There, the filtered time series of the first observable, y_1, is contained in the first column of Y_t, the filtered time series of the second observable in column $M_w + 1$, and so on. However, due to the method of delays, those columns only cover the time steps 1 to N_T $N_T - M_w$ and one has to append the bottom row to the end of that variable, i.e., time step $N_T - M_w + 1$ of the first variable is at the end of column 2 in $Y_r(N,2)$ and the last time step in $Y_r(N, M_w)$:

$$\mathbf{y_{j0}(t_1, t_2, \dots, t_{Nt}) = \{Y_t(1 \dots N, 1 + (j_o - 1)M_w), Y_t(N, \{2 \dots M_w\} + (j_o - 1)M_w)\}} \tag{16}$$

The forecasting model therefore consists of the truncated dynamical system P_t, Λ_t, S_t and the principle is to interpolate the current measurements to 'close' examples of the filtered training data, where 'close' is in terms of dynamic behaviour rather than time.

3.5.2 Preparing new data for the forecasting model

It is possible to project a new time series onto this reduced set of singular vectors by creating a delay matrix following the same procedure as for the training set, including using the mean and standard deviation from the training data set to rescale the new data. This projection will then give principal components, $P_{n,}$ to place the new data in this phase space as

$$P_n = Y_n S_t^T \Lambda_t^{-1} \tag{17}$$

To generate a single point in this phase space, the new time series must contain $M_w \tau$ measurements. Conversely, if the new time series contains $M_w \tau + n_x - 1$ points, its time delay matrix contains n_x columns for that observable and its projection onto the singular vectors results in a section of orbit containing n_x points.

3.5.3 Finding nearest neighbours

Ensemble forecasting in dynamical forecasting makes several forecasts, each initialised with a slightly different initial condition but within the measurement accuracy of the initial point to predict a large sample of possible future outcomes. The results are then evaluated by examining the distribution across all ensemble members of the forecast variables. A useful feature of ensemble forecasting is that it also provides an estimation of the reliability of the forecast. The idea is that when the different ensemble members differ widely, the actual event we try to forecast could shadow any of the modelled ensemble members. This then means that the forecast is affected by a large uncertainty; when there is a closer agreement between the ensemble member forecasts, the uncertainty in the prediction is lower [96]. This principle can also be applied to PCA forecasting where the attractor represents the model. Now, current measurements can be mapped onto the attractor and previously observed wind states close to the current measurements can be found. They can then be taken as an ensemble of initial conditions close to the current state and thus, be used for prediction.

The two key stages in the forecasting part of the method are, firstly, to find a number of 'similar' events in the training period, which is done by finding a chosen number of nearest neighbours in the attractor and, secondly, to follow the evolution of those past similar events. From that evolution one can calculate an expected mean evolution which is the prediction, and one can also calculate by how much the evolution of the ensemble of similar past events either stayed close (giving confidence in the mean forecast) or diverged over the forecasting horizon (indicating that the currently measured wind comes from a part of the attractor which is unstable and not well predictable).

The nearest neighbours are found by calculating the Euclidean distance between the new point, or the mean distance of each point of the section of orbit, to all other points or sections of the training attractor; for a single point: $d_i = \left| P_n - P_t^{i+j-1} \right|$ or for a section of orbit with n_x points

$$d_i = \frac{1}{n_x} \sum_j \left| P_n^j - P_t^{i+j-1} \right| \tag{18}$$

From this complete set of distances to all points of the training attractor, a specified number of nearest neighbours is selected, subject to a constraint that they do not come from adjacent points on the training orbit but from different passes of the orbit through the neighbourhood. This can either be done by sorting all distances and rejecting those which come from adjacent points of the training time series, or by stepping through all distances, and skipping a set number of time points after having identified a local minimum of the distances. If entry k' of the training principal components has been identified as one of the nearest neighbours, then the entry $k = k' + n_x - 1$ is the neighbour to the latest measurement.

The number of nearest neighbours, n_n, to use for the forecasting depends on the dimension of the reduced system and how densely the phase space is covered by the training attractor. If too few neighbours are chosen, the ensemble prediction might not capture the divergence or convergence of the attractor and hence may not give a good

estimate of the forecasting error. If too many neighbours are chosen, the nearest neighbours may not be that near and no longer be a good representation of the local dynamics, hence introducing errors into the forecasting.

3.5.4 Predicting using nearest neighbours

Once the nearest neighbours have been identified, each can be moved forward in time by the lead time or forecasting horizon while sampling all intervening time steps. A key assumption in the implicit forecasting here is that the current point will evolve alongside the identified nearest neighbours from the training data. This means that the relative position of the point from the training attractor at time $k = k' + n_x - 1 + T$ will have a similar position relative to that of the current measurement predicted a lead time T ahead. If the current distance vector to nearest neighbour j is

$$D_j = P_t^{kj} - P_n^{nx} \tag{19}$$

then the prediction based on this nearest neighbour is

$$P_f^j(T) = P_t^{kj+T} + D_j \tag{20}$$

The ensemble of $P_f^j(T),\ j = 1...n_n$ is then the ensemble prediction, each member of the ensemble is mapped back onto the delay matrix space by using

$$Y_f^j = P_f \Lambda_t S_t \tag{21}$$

Each of the Y_f^j returns the predicted wind speeds for the next T time steps as the entries $u_p^j(+1...T) = Y_f^j(N - T + 1...N, M_w)$. This ensemble of predicted wind speeds can then be used to calculate the expected velocity as their average, and an estimate of the uncertainty based on the standard deviation

$$\sigma_p(t) = \left\langle u_p^j(t) \right\rangle_j \tag{22}$$

Likewise, if wind direction is used as a second observable, this can be reconstructed by

$$\boldsymbol{\theta_p^j(+1 \dots T) = Y_f^j(N - T\ + 1 \dots N, \dots, 2M_w)} \tag{23}$$

3.6 PCA used as a Measure-Correlate-Predict methodology

In this section, the background of the usage of PCA as a MCP methodology, which is the main focus of this research will be explained. Figure 8 is depicting a first illustration of the PCA-MCP method as a structure of matrix formalism.

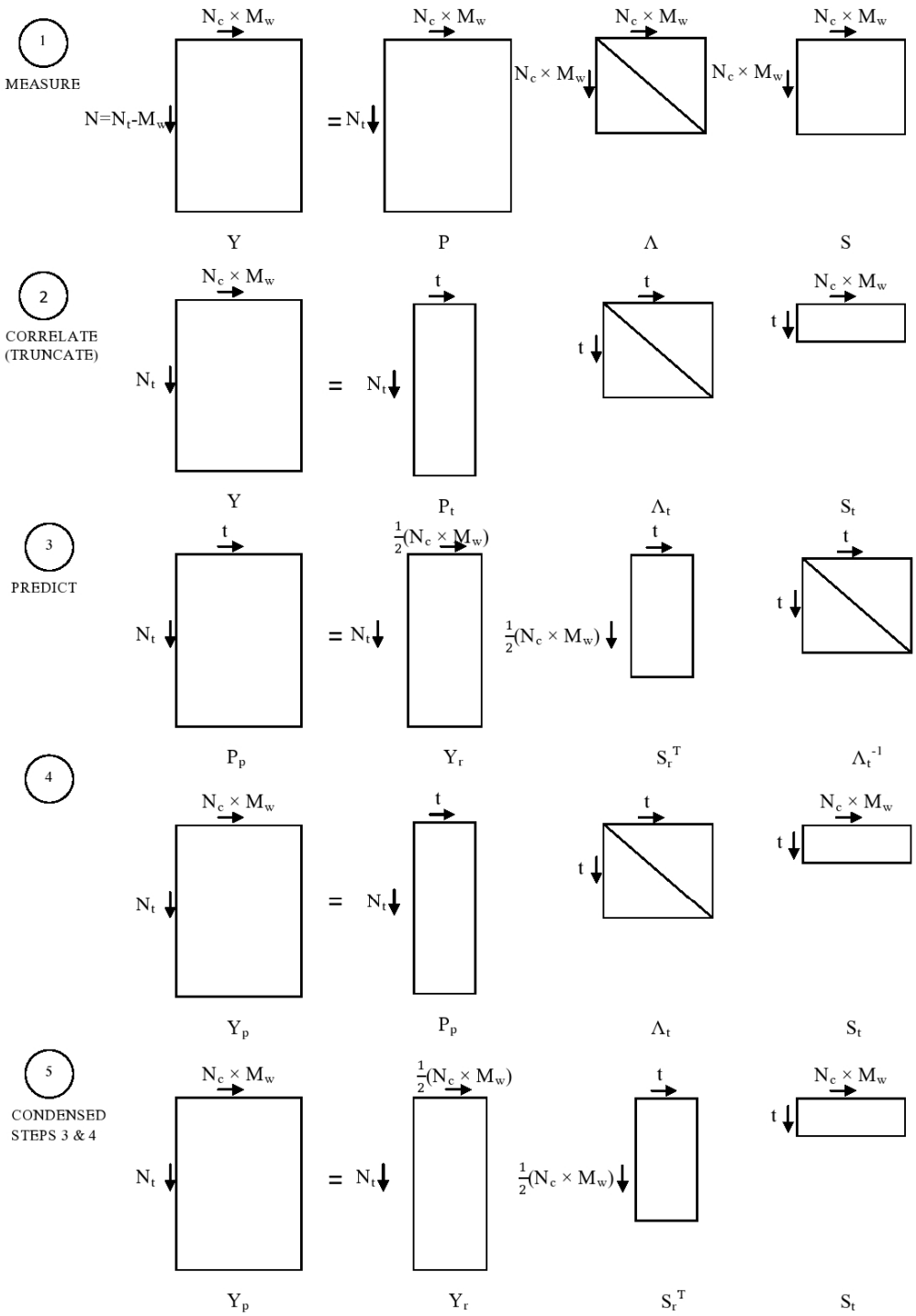


Figure 8. The PCA-MCP schematic.

3.6.1 Mapping from part knowledge onto full attractor for prediction of MCP: the underlying idea.

The challenge that arises from Figure 8 stage 3 is that if we have wind measurements from only one site, can we use similar analysis concepts to identify the state of the combined system on the phase space? If so, can we then predict for the second site for which we have only half information available? More precisely, the question that arises is by taking the defined points of the combined two sites system and adding the new measurements, can we project them to the existing attractor and predict from the nearby points? The experience from previous results all shown in Chapter 4 actually tells us that points under a specific state of the phase space have moved to the next state. Thus, in practice, with the use of data the reference site and the knowledge of their current and next position in the phase space, the challenge is whether we can do the same for the target site.

In practice, following equation (12) in order to achieve the MCP principle, we perform PCA to the original wind data and then truncate to the relevant components with the singular values and vectors becoming: Λ_t, S_t where t here is denoting the truncation and will be used throughout, which is the adjustment of r used as the truncation notation in the PCA forecasting principle of the previous section, section 3.5. The notation of section 3.5 was followed directly from the published paper. Then, the prediction of the wind speeds using that truncation would be of the form:

$$Y_p = P_p \Lambda_t S_t \tag{24}$$

where Y_p, P_p are the predicted time-delay matrix and principal components respectively However, as a continuance of the principle described in the previous paragraph, we use the reference site data only to predict from, since they are the known information and thus equation (12) becomes

$$Y_r = P_p \Lambda_t S_r \tag{25}$$

where Y_r, S_r are the reference site time-delay matrix and singular vectors respectively i.e. of half size since they do not contain the target site data. As it can be seen from equation (25) Λ_t which is diagonal is used instead of Λ_r since as long as there are more time-delays than truncations i.e. $r > t$ where $r = \frac{1}{2} * M_w * j_0$, then Λ_t will be used. Solving for the PC's to get the predicted ones, P_p from equation (25) we obtain

$$P_p = Y_r S_r^T \Lambda_t^{-1} \tag{26}$$

but the predicted wind speeds are of interest and hence we need to go back again to equation (24) and transform it by substituting P_p from equation (26) into the following relationship

$$\begin{aligned} Y_p &= P_p \Lambda_t S_t \\ Y_p &= Y_r S_r^T S_t \end{aligned} \tag{27}$$

which can then be normalised back to actual wind speeds.

3.6.2 Predictor calibration

This is the underlying MCP idea, to predict the wind speeds of the reference and target sites using only the reference site data and truncated singular values and singular vectors by performing PCA once. While testing this idea during development of the technique, the issue which occurred was that applying the operations to the historical reference data only consistently resulted in prediction with far too small a variance compared to that expected. The reason for this can be found in the fact that the correct variance is maintained if each complete principal component is multiplied with its corresponding singular value and singular vector. This works for each principal component – singular vector pair individually as the singular vectors form an orthonormal basis. However, only using partial principal components and singular vectors, as in equation (26), do not preserve the variance. Hence, even though the method is applied to reference historical data it needs to be calibrated with respect to the

known data which is the training period i.e. the concurrent reference data. Several calibration methods were explored, all based on empirically matching the predictions of applying equation (27) to the reference data only from the training period to the actual reference and target data for the training period. They are presented in section 5.4 but only the final method, proposed as the most reliable method found so far, is introduced here.

Going back to equation (26) P_p is now transformed into P_c which are the PC's used for the calibration and it becomes of the form:

$$P_c = Y_h S_r^T \Lambda_t^{-1} \tag{28}$$

where Y_h is the half from the original concurrent time-delay matrix and S_r^T is the reference only singular vectors. Since S_r is the same size as Y_h (i.e. half) this is the rationale behind its use in equation (28). Thus, equation (27) after substituting P_c from equation (28) will become

$$\begin{aligned} Y_c &= P_c \Lambda_t S_t \\ Y_c &= Y_h S_r^T S_t \end{aligned} \tag{29}$$

i.e. the PCA predictors of the concurrent wind speeds Y_c.

Next step was to normalise the calibrated results back to actual wind speeds and calculate the calibrated mean and standard deviation μ_c, σ_c of the Y_c matrix. Hence, the resulting prediction has μ_c, σ_c but we know that it should have μ, σ. The rescaling thus method was of the form:

$$\mu_p = \mu\left(\frac{\mu}{\mu_c}\right) = \frac{\mu^2}{\mu_c} \tag{30}$$

and

$$\sigma_p = \sigma\left(\frac{\sigma}{\sigma_c}\right) = \frac{\sigma^2}{\sigma_c} \tag{31}$$

where μ_p, σ_p are the mean and standard deviation of the predicted data, calculated as the ratio of the original mean and standard deviation μ, σ of Y over the calibrated mean and standard deviation μ_c, σ_c of Y_c respectively. This rescaling method was selected for the PCA-MCP procedure because it gives all variables i.e. wind speed and wind direction from both reference and target sites equal rating so that the skewness towards one variable can be avoided.

Finally, going back to equation (27) $Y_p = Y_r S_r^T S_t$ we can normalise the results back to actual wind speeds using μ_p, σ_p as found from equations (30) and (31). Thus Y_p now contains the predicted reference and target data and Y_r the reference actual data.

3.6.3 The PCA- MCP algorithm

Table 3 presents the steps taken in the development of the PCA-MCP algorithm as described also in Figure 8.

Measure:	
1) Normalisation of measurements of wind speed and direction so that all measurements are given equal weight (no bias of any instrument dominating the signals). Here the number of channels, j_0, are 2 or 4 depending if only wind speed or wind speed and wind direction are used.	$y_{j0} = \frac{\left(y^*_{j0} - \mu_{j0}\right)}{\sigma_{j0}}$
Correlate:	
2) Creation of time-delay matrix. This is the fundamental step to move from direct measurements to the phase-space description	$Y^{i,j+(j_0-1)M_w} = y_{j0}\left(j+\left(i-1\right)\tau\right)$
3) Perform PCA to optimise the attractor (= predictor). This finds the best combination of measurements into patterns with the highest contribution to the signal and least noise. The result of the optimisation is then the phase space description of the relationship between the two sites; equation (12)	$Y = P\Lambda S$
4) Choice of appropriate truncation. This determines how many patterns are thought to contain the important signal. A truncation too small ignores useful information, while a truncation too high includes too much noise; equation (24)	use truncation choice t
5) Truncation of the PCA output to the relevant components to define predictor. This with the use of the appropriate PCA results determined in the steps 3 and 4 above helps in finding the predictions.	Λ_t, S_t
6) Using the reference only data and truncation equation (12) becomes; equation (25)	$Y_r = P_p \Lambda_t S_r$
Calibrate:	
7) The reference historical data need to be calibrated with respect to the known data which is the training period i.e. the concurrent reference data; equation (28)	$P_c = Y_h S_r^T \Lambda_t^{-1}$
8) The calibrated time-delay matrix is of the following form; equation (29)	$Y_c = P_c \Lambda_t S_t$ $Y_c = Y_h S_r^T S_t$

9) Rescaling of the predictor. Since the truncation deletes (unwanted) information, variance is lost from the system. This requires the original mean and standard deviation $\mu = \overline{y}_{j0}, \sigma$ and the calibrated ones μ_c, σ_c ; equation (30) and equation (31)	$\mu_p = \frac{\mu^2}{\mu_c}, \sigma_p = \frac{\sigma^2}{\sigma_c}$
Predict:	
10) Normalisation of the historical data set from the reference site. This must use the same offset and scaling applied in steps 4 and 5 above to ensure that the historical data set is compatible with the PCA attractor created from the concurrent data.	$y_{r0} = \frac{\left(y^*_{r0} - \mu_{j0}\right)}{\sigma_{j0}}$
11) Creation of time-delay matrix using same parameters as for training, this creates a time-delay matrix using the new normalised measurements which will become the new wind speed predictions.	$Y_r^{i,j+(j_0-1)M_w} = y_{r0}\left(j+(i-1)\tau\right)$
12) Projection of the time-delay matrix onto the predictor gives prediction in phase space; equation (26) and equation (27)	$P_p = Y_r S_r^T \Lambda_t^{-1}$ $Y_p = P_p \Lambda_t S_t$ $Y_p = Y_r S_r^T S_t$
13) Mapping of the prediction in phase space back to delay matrix in physical space returns predicted wind speed	$y^*_p = y_{jp}\sigma_{j0,p} + \mu_{j0,p}$

Table 3. The PCA- MCP algorithm.

The way PCA is used as an MCP method is quite similar to the forecasting methodology described in the previous chapter. However, there are some differences in some steps of the procedure since for the MCP case; we do not need to use from past events the nearest neighbours in order to predict the wind resource for a day ahead. In this case, we use PCA for our reference site to train and truncate to the relevant components and then by having only half the information matrix, we predict for both reference and target sites. We calibrate around the mean and standard deviation ratios in order to recover all the lost variance caused by the PCA analysis and finally rescale back to actual wind speeds. Furthermore, this tool also enables us to examine the performance of our predictions when comparing with the actual wind data from both

sites. Hence, the PCA- MCP tool not only gives us a prediction of the wind resource of a site but also measures the reliability of this prediction.

Chapter 4 PCA as a wind forecasting method

This chapter is describing the attempt of using PCA as a forecasting wind speed method. It was undertaken as a preliminary step for this research's purposes aside from the PCA-MCP principle which is investigated in the chapters to follow. The methodology for this PCA application is described in detail in section 3.5. This application of PCA was also published in the journal of Renewable Energy, Elsevier [97].

4.1 Literature review in forecasting methods

The wind variability can be characterised by slow cycles (daily and longer), fast (unpredictable) turbulence, and synoptic weather changes which tend to changes only slowly, the forecasting horizon can be divided into the three following categories: 1: immediate-short-term (up to 8 hours ahead), 2: short-term (8 to 24 hours ahead), and 3: long-term (multiple-days-ahead) forecasting [98-100]. It is more common to use hourly forecasts of winds for dispatching decisions and for scheduling the loads strategy it is common to use daily forecasts of hourly winds. For maintenance purposes, weekly forecast of day-to-day winds are more commonly used [101].

Several forecast models have been created which can be categorised into physical, such as the Numerical Weather Prediction systems (NWPs) [98], statistical, including linear methods such as Auto Regressive Moving Average models (ARMA) or methods coming from artificial intelligence and machine learning fields such as Artificial Neural Networks (ANNs) or even by hybrid approach methods which are a combination of statistical and physical methods with a use of weather forecasts and analysis of time series [99]. Erdem and Shi [102] used four ARMA approaches in order to obtain wind speed and direction forecasts and found that the ARMA model based on the decomposition of wind speed into lateral and longitudinal components was better in predicting direction in comparison to the traditional ARMA model. However, that was the opposite case for wind speed. De Giorgi et al. [103] used ARMA models in combination with different types of ANNs and Adaptive Neuro-Fuzzy Inference

Systems (ANFIS) for several testing period models but also time horizons. For all the attempts it was found that the forecast was worse as the prediction length was increasing.

An integration of ANNs with NWPs for forecasting purposes was undertaken again by De Giorgi et al. [104]**.** The neural network was initially based on the statistic model of wind power time series and was later integrated with NWPs which indicated a significant improvement on the performance. Specifically, pressure and temperature as NWP parameters seemed to improve the forecasting model. Früh [105] explored a simple a linear predictor and based on the observed mean daily cycle model with wind speed or power output data as inputs and noted that increased sophistication in the forecasting methods surprisingly seemed to deteriorate the predictive ability.

Hybrid approaches typically employ an ARIMA (Autoregressive Integrated Moving Average) model for the linear characteristics and an ANN or SVM (Support Vector Machine) model for the nonlinear characteristics. Wang et al. [100] found that depending on forecasting horizon, hybrid methods or ARIMA method perform better in forecasting than the ANN and SVM methods. They also concluded that hybrid methods add significantly in the short-term forecasting modelling for wind speed and power generation, but in general, they do not outperform the other methods [106].

4.2 Data and methodology

4.2.1 Dataset

The data used for this analysis originated from the Gogarbank surface station in Edinburgh provided through the UK Met. Office – MIDAS Land Surface Station record [75]. The site used an anemometer 10m high above ground and the data records used spanned from 1998-2010 with hourly mean wind readings with the wind speed stored to the nearest knot (1 kn=0.5144 m/s) and the wind direction in degree to the nearest 10°. Details of the dataset used are shown in Table 10 and Figure 31 of section 6.3. For this analysis purposes wind speed and wind direction data were used with the wind speed converted to m/s. An illustration of the data, i.e. the wind speed is shown in Figure 9 for

the 2-year period covering 2008 and 2009. The data not used as the training data set were then used for testing the method. A section from the test period was used to apply the prediction model, and the predictions for the 24 hours following that section were then compared against the actual data for the 24h period following that section.

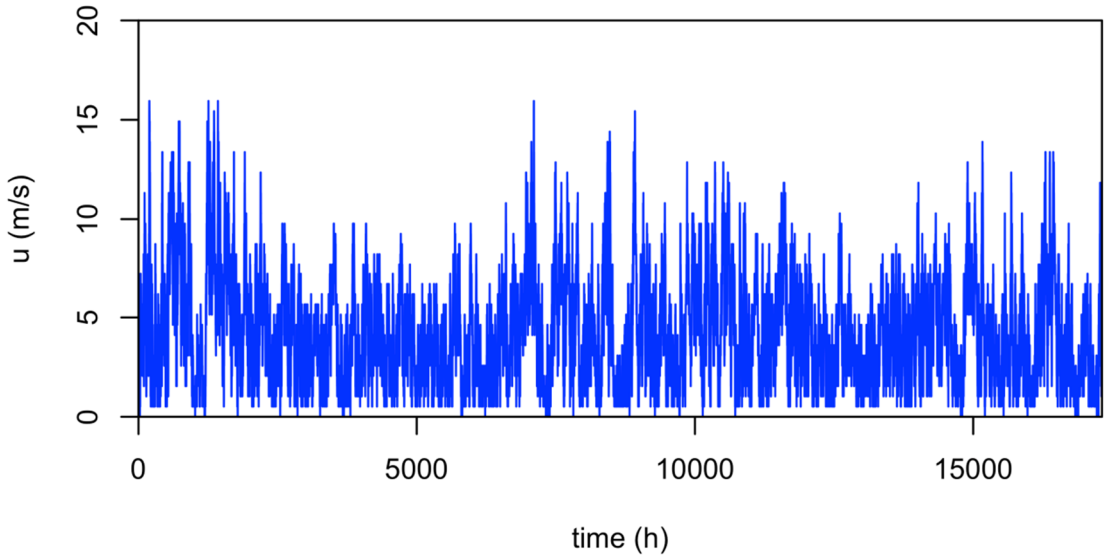


Figure 9. Wind speed time series for Gogarbank 2008 and 2009.

4.2.2 Analysis setup

From the available records, two 2-year records were chosen as the training period, either the years 2008 – 2009 or 2000 – 2001. For all examples discussed in section 4.3, the time lag chosen to create the delay matrix was equal to the sampling period of the data, $\tau = 1\text{h}$, but a range of delay window lengths, M_w, ranging from 1 day (i.e., 24 readings) to 2 weeks (336 readings). The reference case for the discussion in the results section is the window length of 1 day for the training period 2008–2009 but two days for the training period 2000–2001, as indicated in Table 2 which also summarises the other parameter chosen for testing the method. For the case of a 2-year training period (17520 hours), a 2-week window (336 hours) of wind speed and direction, the delay matrix will have 672 columns and 16848 rows, leading to a principal component matrix of the same dimension, 672 singular values, and 672 singular vectors of length 672 each.

Because wind direction is a circular variable, one either has to be aware that there is an apparent discontinuity between 360° and 0° or transform the wind speed and wind direction variables into a pair of horizontal velocity components, $u_x = u \sin\theta$ and $v_y = u\cos\theta$. In the present case, we used the direction as a direct input. As there were virtually no cases of the direction jumping across the 0°/360° boundary, it was decided that no error was introduced. However, for locations with a wider spread of wind directions, it is recommended that the data should be transformed to the velocity components.

Of the singular values (lambda), of which the first 90 are shown in Figure 10, only a few have high values which drop off rapidly and then settle to a plateau from the 20th on. From this figure it is clear that at least the leading four dimensions must be retained in the model but that including more than 20 would add increasingly noise to the predictions. For that reason, a truncation of M_t = 5 to 35 was explored.

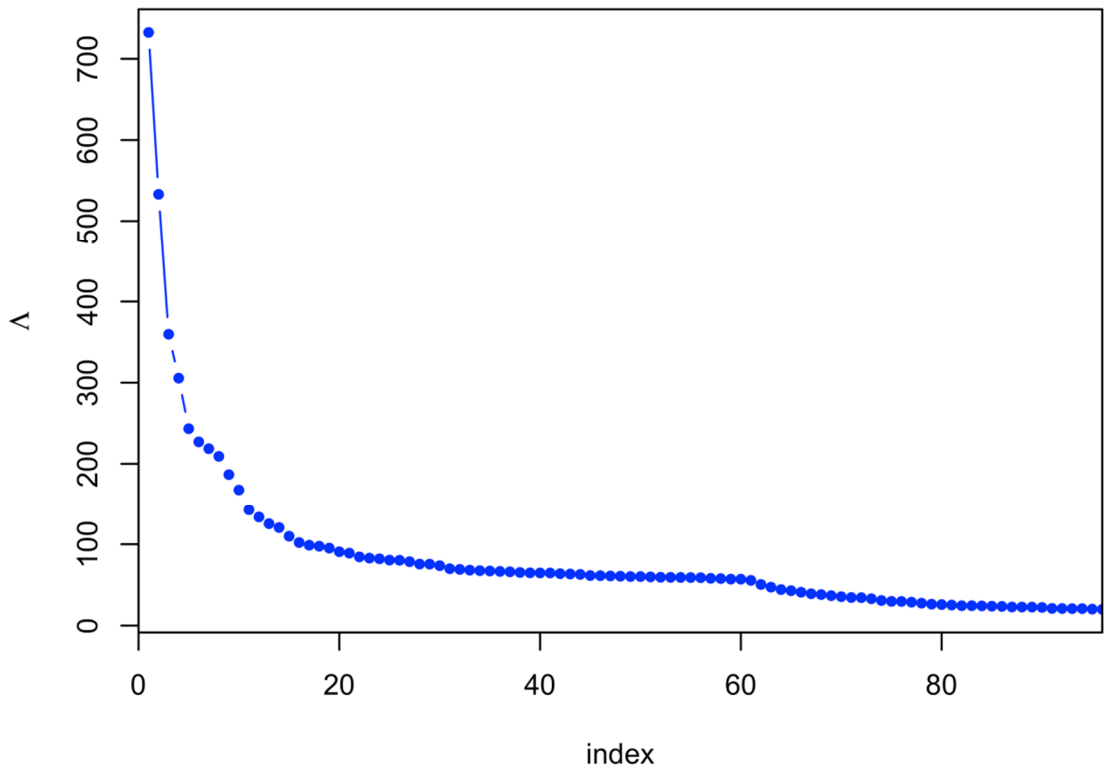


Figure 10. The first 90 singular values for the PCA of 2008-2009 training set with window length (M_w) of 2 weeks.

The first three singular vectors (svec[,1],svec[,2] and svec[,3] respectively) for the PCA applied to the 2008–2009 data using M_w =48h are shown in Figure 11. Since the input data are the wind speed and the wind direction, each singular vector contains two distinct sections, where the first 48 entries correspond to the temporal evolution of the wind speed attributed to that singular vector and the entries 49 to 96 correspond to the wind direction. Figure 11a and Figure 11b show that the first two singular vectors are associated with a slow modulation of the weather, while the third singular vector in Figure 11c and the fourth singular vector (not shown) correspond to a daily cycle. The phase space diagram drawn by the first two principal components (P_1 and P_2), shown in Figure 12, shows an attractor with a clear structure associated with the prevailing weather conditions in Scotland, and the transition between them.

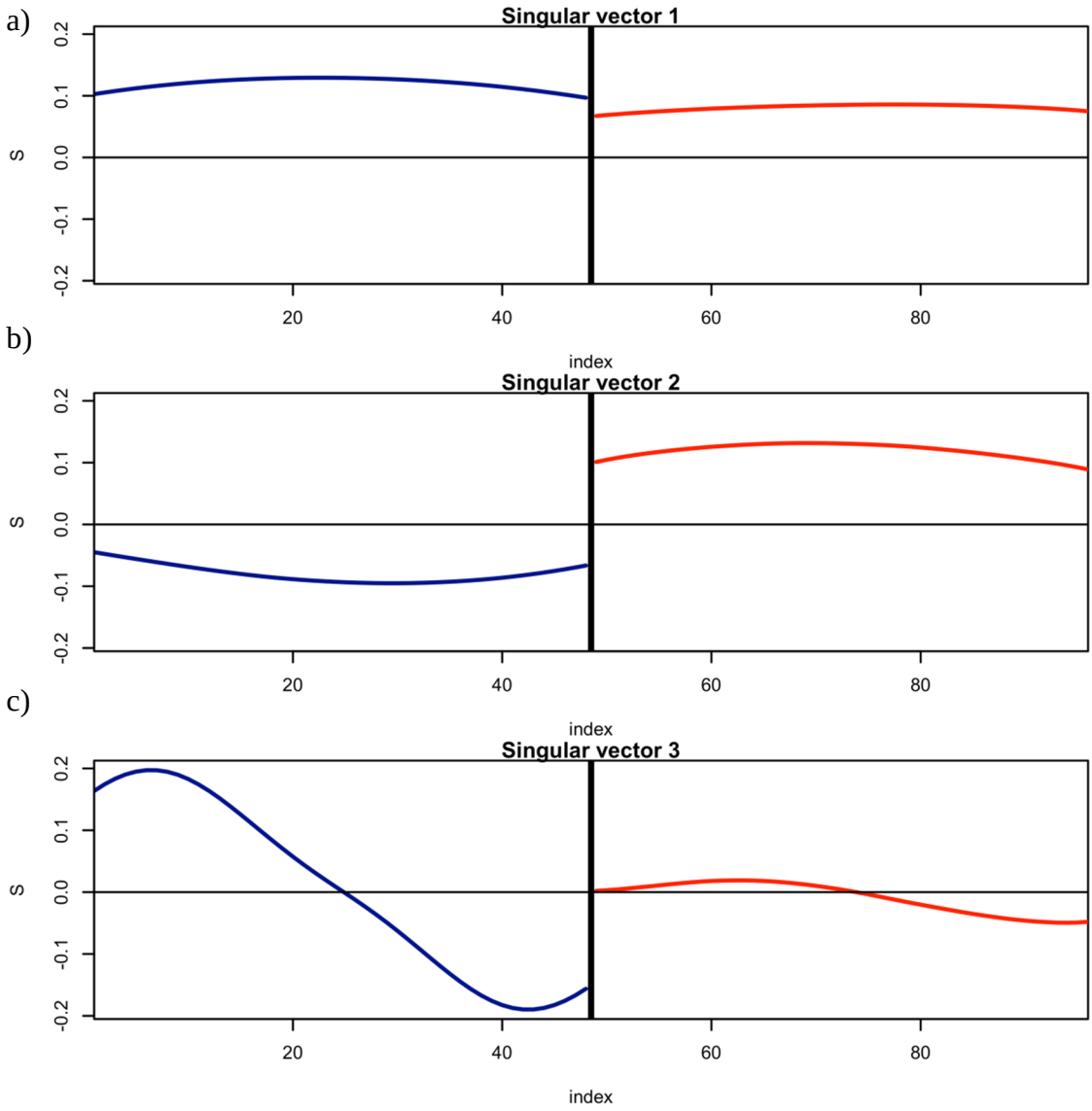


Figure 11. First three singular vectors of the 2008-2009 M_w =48h model in Fig. 10(a), 10(b), 10(c). The line between index 48 and 49 separates wind speed on left from the wind direction on the right.

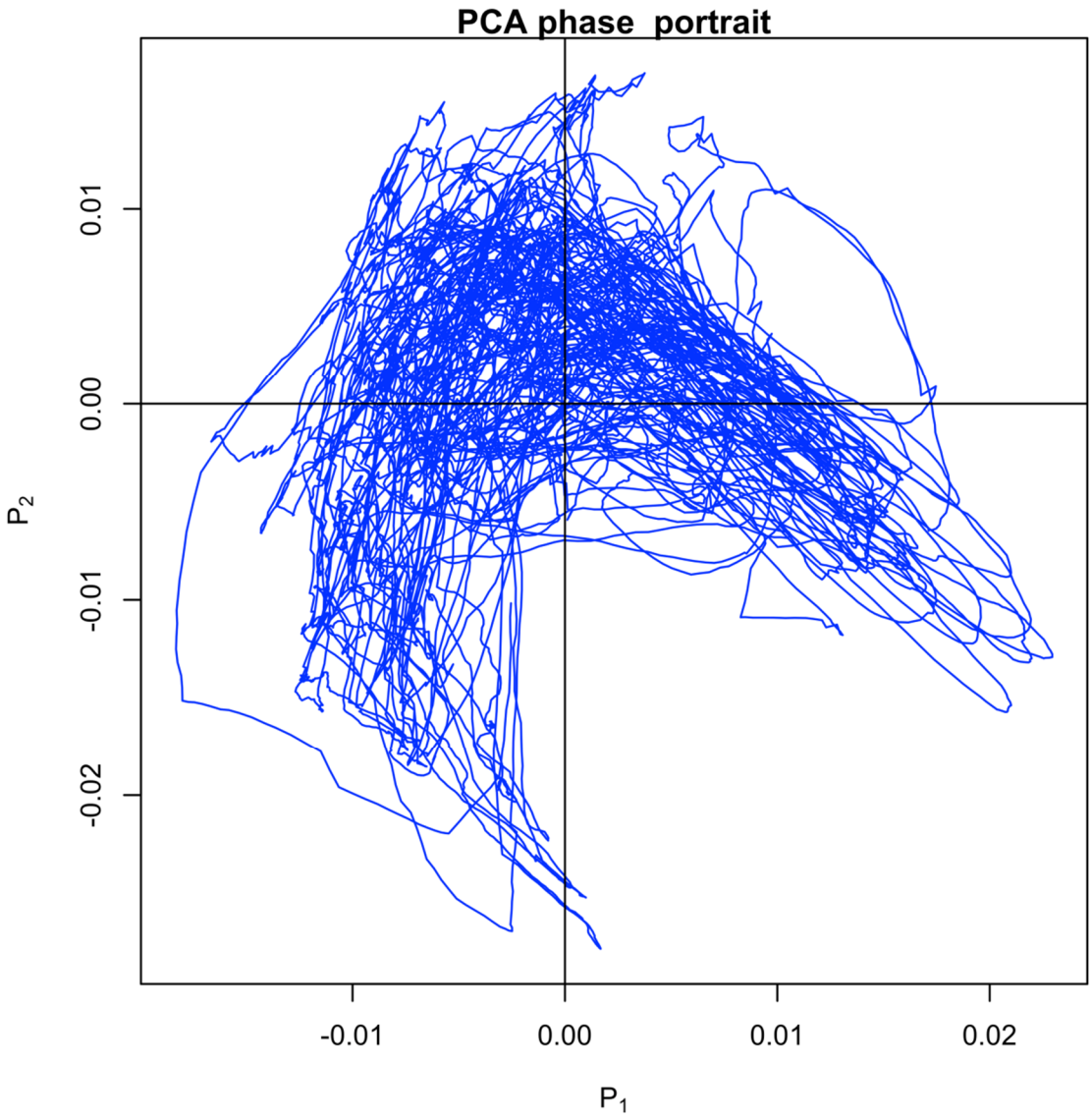


Figure 12. Phase portrait constructed from the first two principal components P_1, P_2 for the 2008-2009 M_w =48h model.

Finally, the parameters for the forecasting component were the length of the orbit section to be projected onto the attractor and the number of nearest neighbours which had to be chosen. For the orbit length a range of 1 to 3 was chosen which means that, for a window length of, for example 48 hours, a section of 48h, 49h, or 50h, respectively was chosen from the test data to create a delay matrix consisting of 1, 2 or 3 rows, correspondingly. The number of nearest neighbours explored in the analysis ranged from 2 to 10, as summarised in Table 4. The reference case was fixed based on the optimal results as shown in Figure 19 to Figure 21.

	M_w	M_t	n_x	n_n	**Forecast**	**Forecasting horizon**
Reference case values	1 day	16	1	5	2010	24h
Range	1 day-2 weeks	5-35	1-3	2-10	1999-2007	1-24h

Table 4. Summary of data used for training and forecasting, with parameter settings used for 2008-2009.

With the model defined by the M_t singular vectors and the past data describing the observed dynamics through the M_t principal components, the new measurements for the forecasting were transformed using the same parameters and then projected onto the observed dynamics. This is illustrated in Figure 13 where the attractor from the training data is the grey object. The blue circle is a single point in the phase space created by a time series section of the window length M_w. In this example, n_n = 5 i.e. the five nearest neighbours on the orbit of the training data are, in order of proximity, identified by the red numbers in Figure 13. These five nearest neighbours can then be traced forward in time over the forecasting horizon, which is shown by the red curves evolving from the numbered positions. Each of these can then be re-transformed to wind speed and direction to produce the ensemble forecast. The final result is then a forecast of the predicted mean wind speed and the uncertainty in that prediction for all lead times from one hour ahead to the specified forecasting horizon, 24 hours in our analysis.

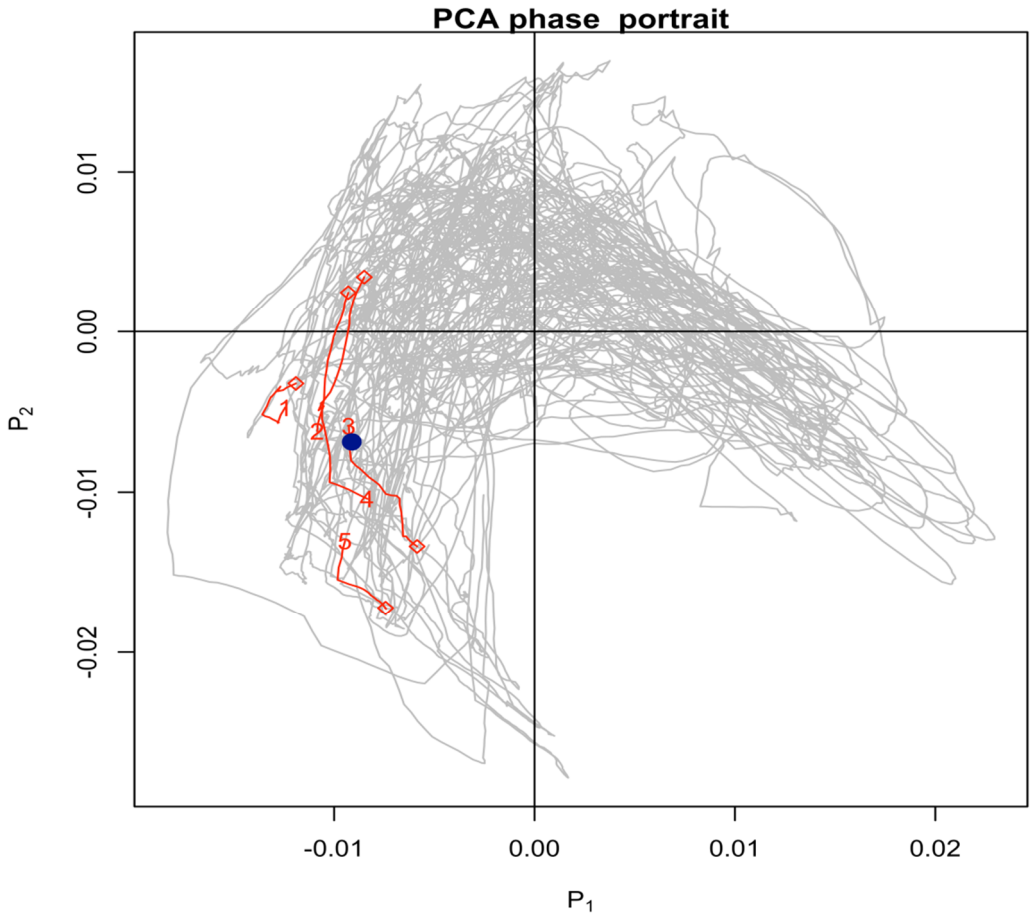


Figure 13. New data mapped onto training set for the 2008-2009 M_w =48h model. The blue circle is the new 'current' observation, and the five red numbers are the nearest neighbours which were then found to evolve for the specified forecasting horizon as shown by the red lines.

4.3 Performance evaluation

To evaluate the performance of the predictions, the predictions are compared against the actual values from the test data, using the three main measures recommended by Madsen et al. [107] albeit for wind speed rather than power output. They are all based on the prediction calculated as the difference between actual observation, u, at time $t+T$ from the test set and the wind speed predicted for that time based on the observation at time t $\hat{u}$,as

$$\mathbf{e(t+T\,|t) = u(t+T) - \hat{u}(t+T\,|t)} \tag{32}$$

These three measures are the bias

$$\mathbf{BIAS(T)} = \hat{\boldsymbol{\mu}}_{\mathbf{e}}\mathbf{(T)} = \overline{\mathbf{e(T)}} = \frac{\mathbf{1}}{\mathbf{N}}\sum_{\mathbf{t=1}}^{\mathbf{N}} \mathbf{e(t+T \mid t)} \tag{33}$$

the mean absolute error (MAE), frequently used in the literature, e.g. [104]

$$\mathbf{MAE(T)} = \overline{\mathbf{|e(T)|}} = \frac{\mathbf{1}}{\mathbf{N}}\sum_{\mathbf{t=1}}^{\mathbf{N}} \mathbf{|e(t+T \mid t)|} \tag{34}$$

and the root mean squared error (RMSE)

$$\mathbf{RMSE(T)} = \sqrt{\frac{\mathbf{1}}{\mathbf{N}}\sum_{\mathbf{t=1}}^{\mathbf{N}} \mathbf{(e(t+T \mid t)\,)^2}} \tag{35}$$

These errors for the predictions using the PCA forecasting were then benchmarked against the frequently used persistence, $\hat{u}_{ref}\ (t+T \mid t) = u(t)$. This benchmarking is quantified by an improvement measure as defined [107], e.g., for the BIAS (and likewise for MAE and RMSE) as

$$\mathbf{Imp_{ref,BIAS}(T)} = \frac{\mathbf{BIAS_{ref}(T) - BIAS(T)}}{\mathbf{BIAS_{ref}(T)}} \tag{36}$$

Since the PCA forecasting intrinsically returns all predicted time steps at the sampling interval until the prediction horizon or lead time T, we also use average of $\mathrm{Im}\, p(T)$ over $T = 1,....,T_{\max}$. The sensitivity of the PCA forecasting method to different choices of the parameters is here described in terms of the overall improvement of the MAE over persistence:

$$\mathbf{PI} = \frac{\mathbf{1}}{\mathbf{T_{max}}}\sum_{\mathbf{T=1}}^{\mathbf{T_{max}}} \mathbf{Imp_{p,MAE}(T)} \tag{37}$$

where the maximum lead time in our case is 24 hours.

4.4 Results

4.4.1 Forecasts of wind speed and uncertainty

The uncertainty of the actual and forecasted wind speed was examined in Figure 14.

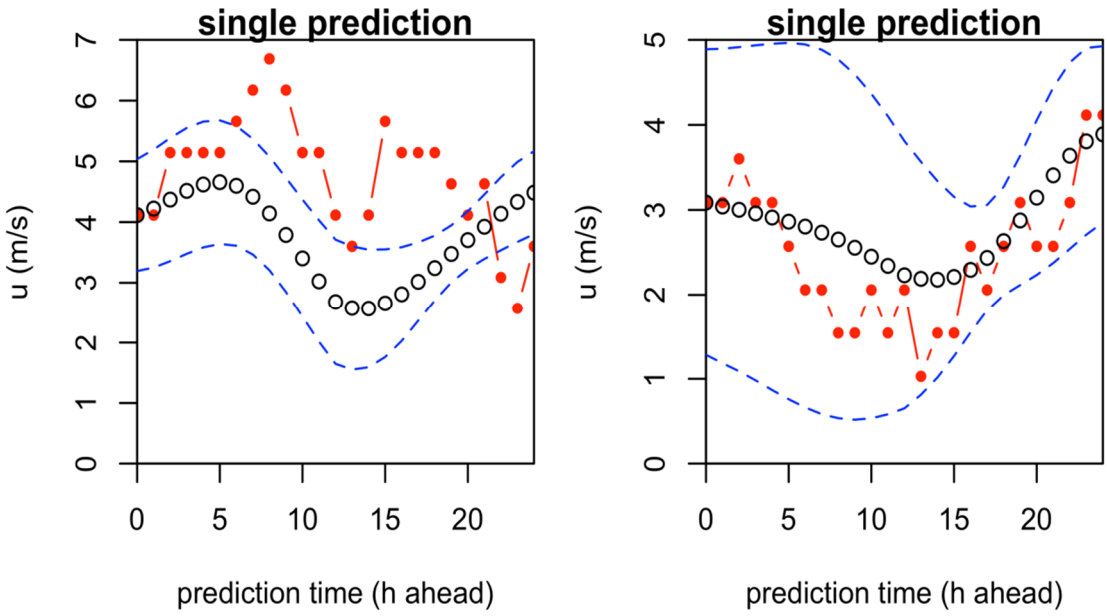


Figure 14. Comparison of actual wind speed (red line), forecasted wind speed (open black circles) and uncertainty of wind speed (dashed blue lines). Fig. 13 (a) is a 'bad' prediction example whereas Fig.13 (b) is a 'good'example.

Figure 14 illustrates a comparison of the ensemble forecast representing all 24 hours of lead time for two of the 100 predictions made for this analysis. As outlined in section 4.2.2, the predictions made in the phase space were re-transformed to real wind speed and direction. From the ensemble of $n_n = 5$ forecasts, the prediction was calculated from the mean of the ensemble (open black circles) and the prediction uncertainty was also found with the use of the standard deviation (dashed blue lines). Hence the comparison to the actual wind events was made (red lines).

As both examples in Figure 14 show, the predicted wind speeds form a strongly smoothed curve compared to the actual winds, as the PCA has successfully separated the slow atmospheric dynamics from the unpredictable local turbulence. For a very good prediction at all lead times from one hour ahead to the forecasting horizon, the

actual with the predicted wind speeds are closely aligned, while for an acceptable prediction, the actual wind speed should lie within the band specified by the uncertainty of the prediction. Conversely, wind speeds outside the band would have been poorly predicted.

Figure 14(a) is an example where the forecast is relatively poor at times due to very large hourly variations in the wind speed. The prediction does not capture the substantial increase in the first 8 hours of the forecast to a degree where the actual wind speed is well above the predicted uncertainty band. A consequence of this is that the actual wind speed is outside the expected range indicated by the dashed blue lines. Finally, the prediction toward the end of the horizon is for the wind to increase while the actual wind speed decreases. Figure 14(b) is a case where the prediction is good: the decrease of the wind speed over the first 14 hours is predicted as is the increase beyond. Furthermore, the model predicts a higher uncertainty for lead times between 10 and 20 hours after which the predicted uncertainty suggests a return of predictability for the day-ahead forecast. This is exactly borne out by the actual observations which follow the predicted mean very well but shows a persistent error within the 10h to 18h lead time.

4.4.2 Forecasting quality

To quantify the performance of this model we used as the first measure the mean absolute error, MAE, as defined in equation (34) by averaging the absolute forecasting error at a lead time T ahead for a large sample ($N = 200$) of forecasts. The reason for concentrating on this measure is that it gives a direct comparison of the error with the predicted uncertainty. If the MAE is less than the uncertainty, the prediction is as good as it can be (and is known to be) but if the MAE is much larger than the predicted uncertainty, the model does not work for that data set.

Figure 15 shows the $MAE(T)$ as the solid red line against the lead time for the reference case of Table 4, i.e. the case of a 2-week training window M_w = 336h, a model predictor dimension of M_t =16 matching a point on the attractor n_x = 1, and using

n_n = 5 nearest neighbours where the predictor was applied to 200 samples from the year 2010. The open black circles are the average of the uncertainties predicted for that lead time and the dotted line is the standard deviation of these predicted uncertainties. As the figure shows, the actual MAE is very close to the predicted uncertainty at short lead time, at lead times approaching the 24h prediction horizon. The model performs slightly worse than predicted from its own internal dynamics at lead times between 8 and 20 hours but still within the range of calculated predictions. The model-internal performance check is also compared against persistence. The mean absolute error for persistence, $MAE_p(T)$ is shown as the green dash-dotted line. The key features of the error of persistence compared to that of the PCA model is that persistence is much better than PCA at short lead times up to 6 hours but that PCA outperforms persistence at longer lead times. The fact that persistence is often the best predictor for short lead times was also supported by Madsen et al. [107] and can be explained through the short-term fluctuations affecting the local wind at these times more than any slow synoptic weather changes. Based on this, we propose a refinement of the PCA-predictor by merging it with a persistence-based correction at short lead times.

4.4.3 Combining persistence and PCA

After performing this comparison and applying several inputs for the different parameters used by PCA, it was concluded that the respective strengths of persistence and PCA could be exploited in a combined forecast by applying a filter to the PCA prediction [99]. This filter constructs a weighted average of the persistence prediction and the PCA prediction for a filter length long enough to cover the range where persistence outperforms PCA prediction. Over that filter length, the weights of the averaged change linearly from 1 for persistence and 0 for PCA at the 'current' time (lead time = 0h) to the other extreme of 0 for persistence and 1 for PCA at the end of the filter length. The filter is of the form:

$$\boldsymbol{u_{f,i}} = \begin{cases} \left(\mathbf{1} - \dfrac{\boldsymbol{i}}{\boldsymbol{N_f}}\right)\boldsymbol{u_0} + \dfrac{\boldsymbol{i}}{\boldsymbol{N_f}}\,\boldsymbol{u_{PCA,i}} & \boldsymbol{for\ i = 0 \ldots N_f} \\ \boldsymbol{u_{PCA,i}} & \boldsymbol{for\ i > N_f} \end{cases} \tag{38}$$

where i is the lead time, N_f the filter length, $u_{PCA,i}$ the ensemble forecast and u_0 the current wind speed. By trial and error, a good filter length was found to be between 10h and 15h, with little change of the results in that range.

The effect of applying such a correction on the performance of the predictor is shown in Figure 16, where it is clear that the very short term prediction, up to a lead time of 6h is now as good as for persistence and that the prediction for longer lead times is dominated by the ability of PCA to extract the slower atmospheric dynamics. The reason behind the dip at a lead time of 5h as seen in Figure 16 for both, PCA and persistence, is unclear. However, a speculation could be made that this reflects the gap in the typical wind speed power spectrum at the period of a few hours [108].

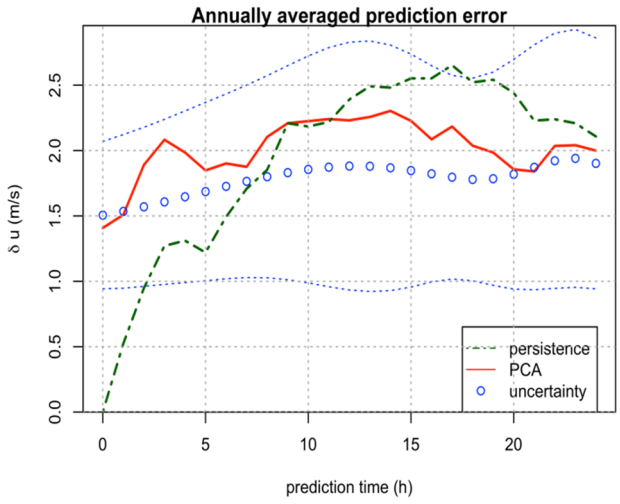


Figure 15. Comparison of annual mean forecasting error and uncertainty (unfiltered data) for the reference case.

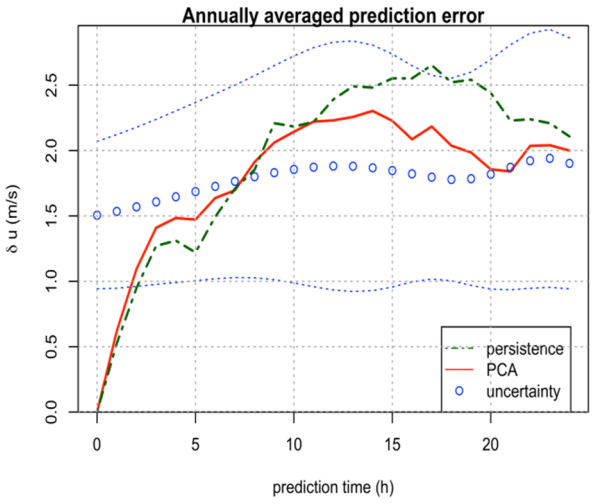


Figure 16. Comparison of annual mean forecasting error and uncertainty (filtered data) for the reference case.

4.4.4 4.4.4. Other error measures

Following the recommendations of Madsen et al. [107] the alternative error measures of Bias (33) and RMSE (35) were calculated and are shown in Figure 17 and Figure 18. They both indicate that PCA outperformed the persistence method and specifically for the bias error measure, PCA performed substantially better than persistence.

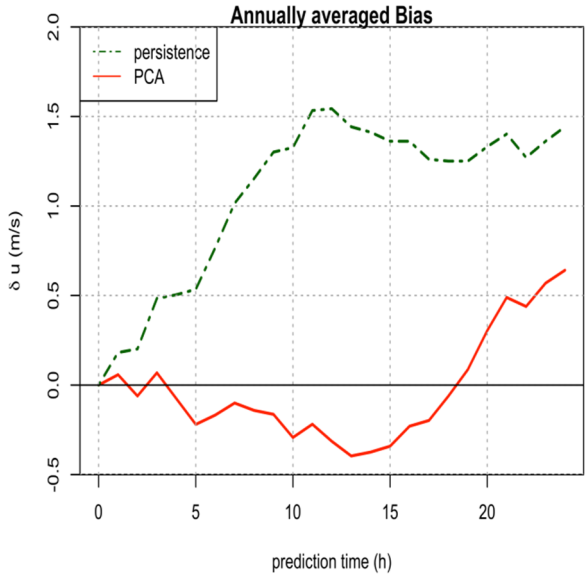


Figure 17. Comparison of bias between PCA and persistence method.

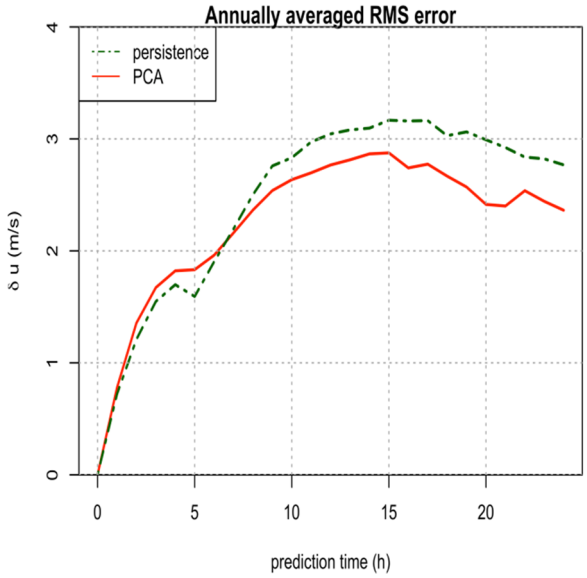


Figure 18. Comparison of RMSE between PCA and persistence method.

4.5 Sensitivity analysis of parameters

Figure 19, Figure 20 and Figure 21 show the performance index of the results for the different choices of the length of orbit, n_x, to use for finding the nearest neighbours on the attractor, the number of nearest neighbours, n_n and the embedding dimension M_t, respectively.

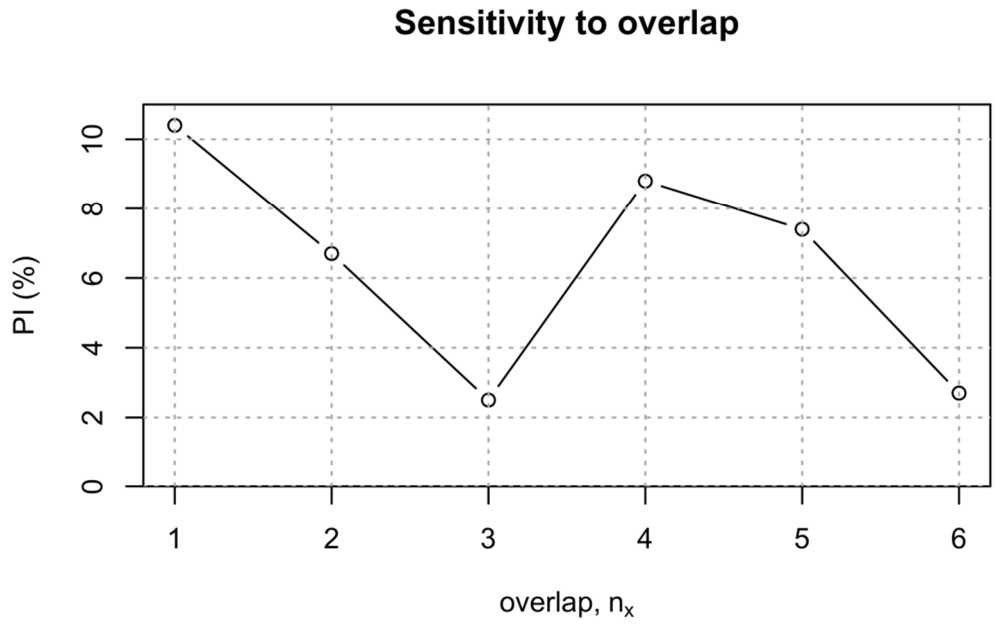


Figure 19. Performance Index of PCA results in % for different overlap values.

Figure 19 indicates that using a single point ($n_x = 1$) rather than fitting a short time series of point ($n_x > 1$) overlap seems to yield the best improvement (around 11.2%) of the results. This means that the PCA results are 11.2% closer to the actual results in comparison with the persistence method. Using $n_x > 1$ did not work for our data set as there were not enough nearest neighbours. After determining that $n_x = 1$ seems to be the best, this was used for analysing the sensitivity to the number of nearest neighbours, n_n, Figure 20 shows that the overall improvement initially rises substantially from below 8% for only two neighbours to above 11% for five nearest neighbours but then drops again to around 9%. Using too few or too many neighbours might not be appropriate since with too few (i.e. less than 5) the information we use for the analysis might be too little whereas on the contrary, using too many (i.e. more than 5) might initially show that we can obtain more information; however, these neighbours might actually lie very far apart from each other in the phase space. There is clearly a distinct optimum which needs to be determined but it is not clear whether it is at or around five nearest

neighbours for any data set or whether this must be determined from optimising the parameters through experience at each site individually.

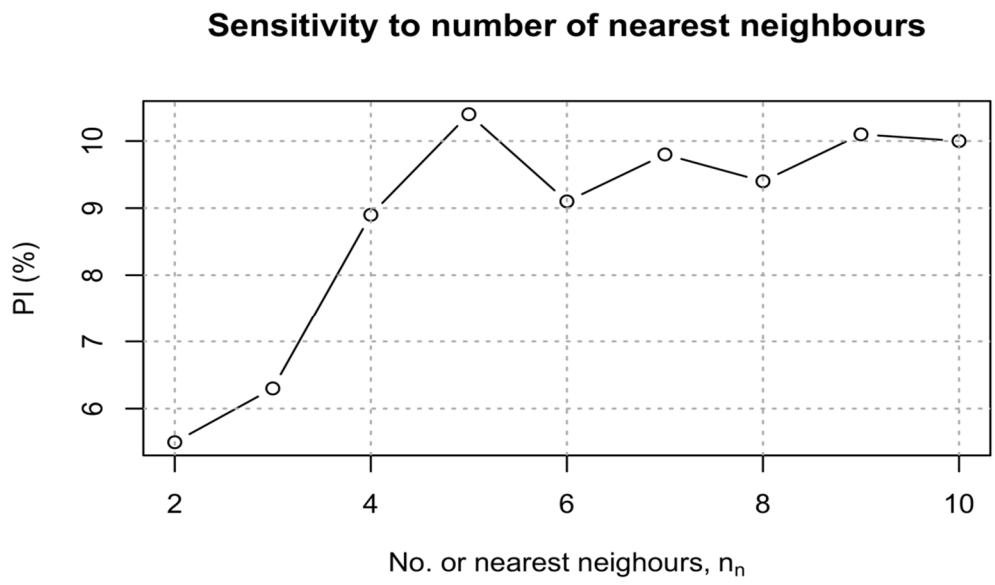


Figure 20. Performance Index of PCA results in % for different nearest neighbours values.

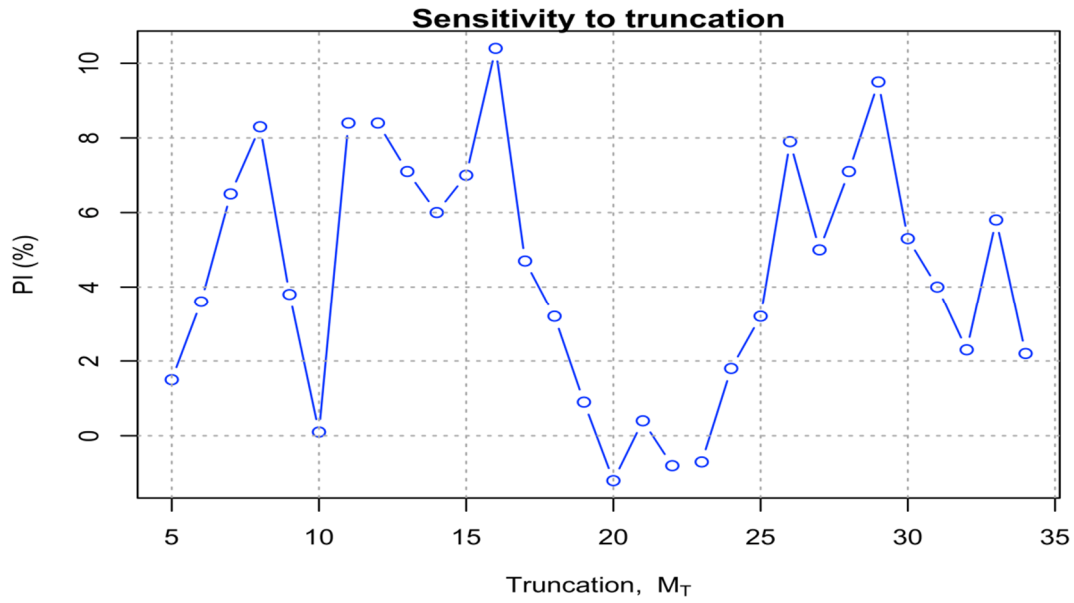


Figure 21. Performance Index of PCA results in % for different truncation values.

Finally, Figure 21 shows the sensitivity of the model to the choice of the model truncation. Here, it can be seen that different choice of truncation results in a big variation of the percentage of improvement. Hence, a careful choice of the number of truncations is important. The number of truncations for which the improvement seems to be more consistently high for (5.6%) is around 16. Truncations up to 16 represent variations at time scales from that of the window length down to 4 cycles per day so the left side of Figure 21 could be interpreted in daily cycles caused by the sun. It should be noted that adding more truncations results in adding more information but whether this information is useful or not is another issue which should be of further investigation and of course depends on the site and wind dynamics used for the analysis.

It can be concluded that applying PCA for wind forecasting purposes demonstrated that the method is a reliable forecasting method for forecasting wind speeds hours ahead to day ahead. By combining the PCA prediction with persistence prediction at very short time scales, it was possible to eliminate the weakness of applying PCA to a coarsely sampled wind record. One of the most useful aspects of PCA over some other forecasting techniques is that it is based on an ensemble forecast using ensembles of similar past events. This allows an estimation of the forecast accuracy at the time when the forecast is made. The analysis showed that this estimated forecast uncertainty is a reliable predictor of the actual forecasting error.

Chapter 5 PCA as an MCP method initial applications and results

This chapter is an intermediate step in the application of PCA used as an MCP method starting from the simple concept of a pendulum which was initially used in Chapter 3 and moving on to real wind data. Initially, wind speed was only used as an input variable and furthermore, where the main part of this research focuses on, wind speed and direction were used as input variables for the analysis described in more detail in Chapter 6.

5.1 First application of PCA as an MCP method in noisy pendulum

The initial step in developing PCA as an MCP method was to test it on the first application of PCA which was presented in section 3.4 and more specifically for the noisy oscillator. The initial noisy pendulum equations as described in equation (14) were developed as

$$\begin{aligned} x_0 &= 3\sin\left(\frac{t}{1.7}\right) \\ x &= x_0\left(1+A_1\cos\left(\frac{t}{f_1}\right)\right)+0.1\varepsilon(x_0) \\ y &= x_0\left(1+A_2\cos\left(\frac{t}{f_2}+\delta\phi\right)\right)+0.1\varepsilon(x_0) \end{aligned} \tag{39}$$

The rationale behind reconstructing the pendulum equations of section 3.4 was to construct simple dynamical systems representing some key characteristic of a potential wind site. Extending this to PCA would need to have two signals representing a target site and a reference site, respectively, which somehow are linked by a common underlying signal representing the synoptic weather pattern. In this case, x_0 would represent, for example, the UK weather characteristics and is a common factor in both equations and x, y i.e. are equivalent to reference and target sites have very similar equations since in real life the wind speed of both sites should not have significant differences in their overall behaviour. Each of the two sites contains the common signal

x_0 but modulated by local effects, represented by modulations with different amplitudes, frequencies and phases, in addition to a noise term representing local turbulence completely without any correlation between target and reference site. The systematic change of them with time should be examined by investigating changes in the following variables: $A_1, A_{2,} f_1, f_2, \delta\phi$. More specifically, A_1, A_2 represent the local magnitude of the wind of x_0, f_1, f_2 indicate the more coherent dynamics of A_1, A_2 modulating the wind and $\delta\phi = \frac{\pi}{(0,2\pi)}$ is the time shift (time of flight) between the two sites and finally, ε the turbulence (noise).

The use of PCA in the noisy pendulum case was structured as it is described in the following steps. First, a time series for the pendulum was set and then the equations as described in equation (36) were introduced. Then, a time-delay matrix Y containing the equations x, y was created. Furthermore, a new reduced time series was created containing only half the information of the system i.e. just the channel of x. The reason behind this was to treat the system like the reference and target site data in an MCP analysis case. Thus, another time-delay matrix Y_{half} using this half information was created containing only x_{half} with y_{half} being the unknown information of our interest. PCA was then performed on the initial time-delay matrix Y which contained both x, y and the singular values Λ, singular vectors S and principal components P were harvested. The singular values matrix was inverted Λ^{-1} and the singular vectors matrix was transposed S^T according to the equations (24) and (25) of section 3.6.1.

A new principal component matrix P_p was then created where instead of the old time-delay matrix Y the new half information Y_{half} one was used alongside with the singular values Λ^{-1} and vectors S^T in the aforementioned form. Equation (40) describes it

$$P_p = Y_{half} S_r{}^T \Lambda^{-1} \tag{40}$$

Another time-delay matrix Y_{test} and a principal component matrix $P_{p,test}$ were then created in order to check the results of combining the initial time-delay matrix Y with the half information one Y_{half} thus containing x_{half} and obtaining the new information y_{half}. Finally, we projected back from the attractor in the phase space to the delay space using the new time series matrix Y_p like described in equation (29) so that we can extract the predicted signals x_p, y_p like it would happen in the case of MCP for the reference and target site.

5.2 PCA-MCP noisy pendulum results

The range of values used for the variables A_1, A_2, f_1, f_2 of equation (39) were 0.1, 0.2, 0.3, 0.5, 0.71, 0.9, 1.11, 1, 2, 3.14, 10 and all these values were examined for the different time shifts $\delta\phi$: 0, π/2, π/4, π/9, π/12,1, $\sqrt{5\pi}$, π, 1.11π, $\sqrt{2\pi}$, 1.5π. These values were chosen randomly, nevertheless attempting to cover different scenarios for the system of equations. A more detailed sensitivity analysis for the aforementioned range of values is found in the graphs of the Appendix A.

5.2.1 Qualitative Results

The PCA results of different principal components plotted against each other (P_1,P_2,P_3,P_4) when examining an indicative range of values based on a reference case of the system for the variables: $A_1, A_2, f_1, f_2, \delta\phi$ are presented below and summarized in Table 5. In some cases it was observed that PCA was not performing very well since some of the PC's of the prediction time series (middle row) were not of similar shape when compared with the actual full time series PC's (last row). Looking at A_1, A_2 for different A_1 values the PC predicted graphs have differences with the actual ones (middle and last row) which indicates that PCA did not perform with accuracy for all PC's. As it can be seen in Figure 22 in the middle and last row all PC's are not similar especially for P_3 and P_4 thus PCA failed to extract similar patterns when the half information time series were used. However this was not the case for A_2 since it can be

clearly seen from Figure 23 that all PC graphs have similar shapes when compared to the real time series which indicates that PCA performed well. Hence, since PCA seemed to be performing better for different A_2 values than for A_1. Further investigation of the singular vectors should be conducted.

Figures	$\delta\phi$	A_1	A_2	f_1	f_2
Reference Case	π/9	4	0.3	0.5	0.3
Figure 22	π/9	0.1	0.3	0.5	0.3
Figure 23	π/9	4	2	0.5	0.3
Figure 24	π/4	4	0.3	0.5	0.3
Figure 25	π/9	4	0.3	10	0.3

Table 5. Range of values for A_1, A_2, f_1, f_2, δφ used for Figure 22 to Figure 25.

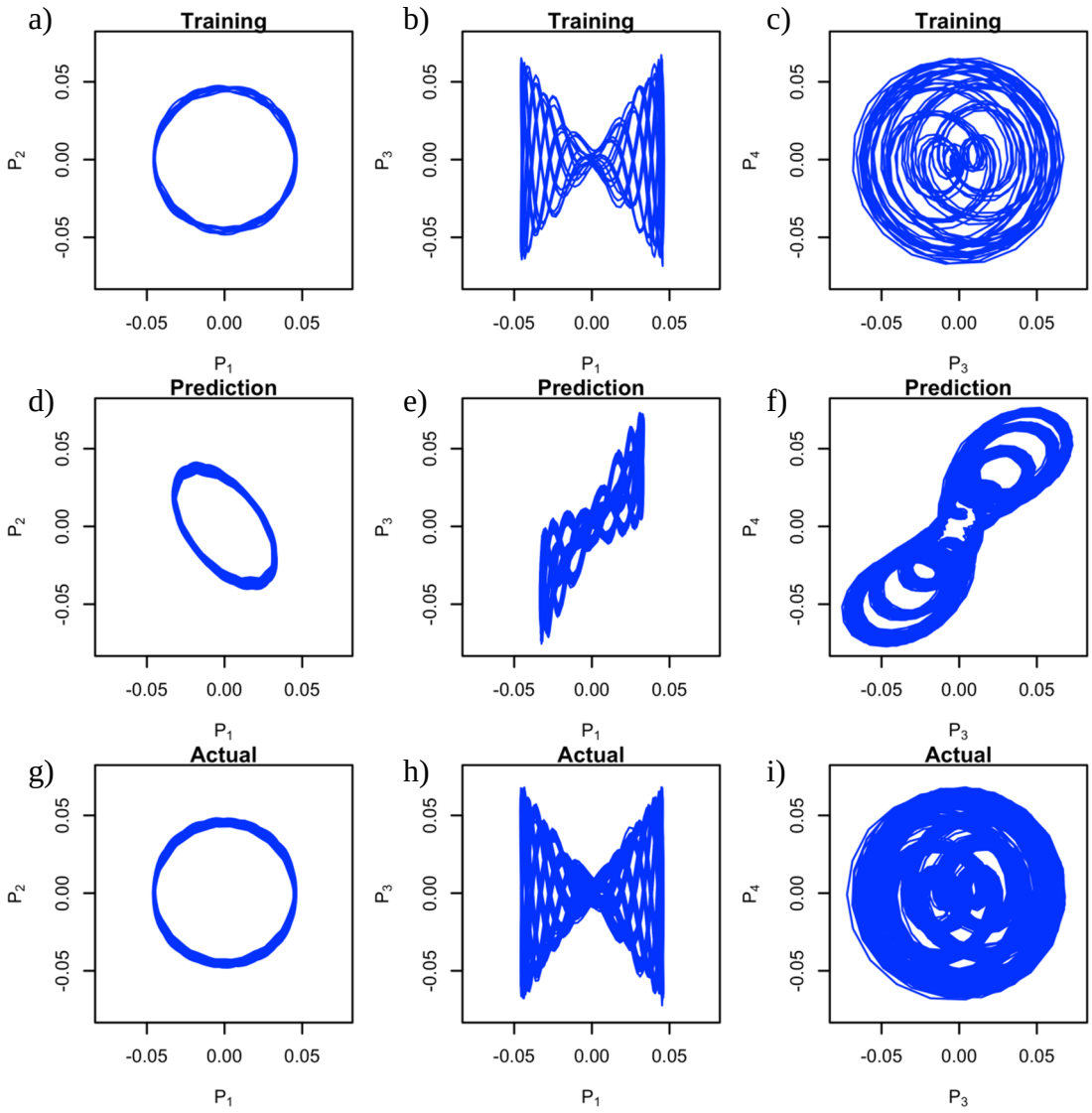


Figure 22. Principal components results for A_1=0.1 and rest of settings originating from the reference case i.e. A_2=0.3, f_1=0.5, f_2=0.3, $\delta\phi = \pi/9$.

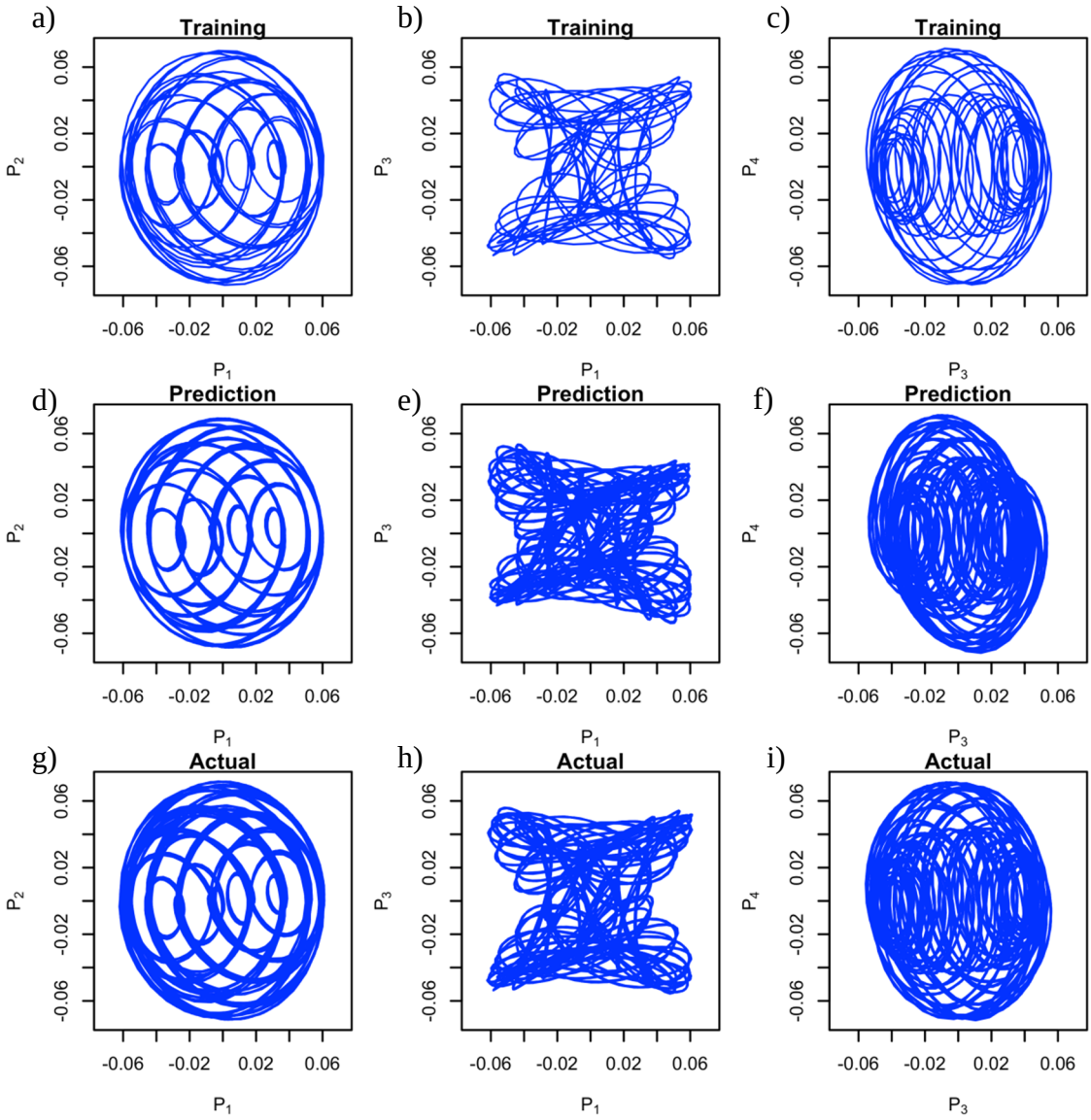


Figure 23. Principal components results for A_2=2 and rest of settings originating from the reference case i.e. A_1=4, f_1=0.5, f_2=0.3, $\delta\phi = \pi/9$.

Regarding the range of $\delta\phi$ we can see in Figure 24 that all PCs have similar shape when compared to each other for the actual and predicted time series. However, for f_1 as seen in Figure 25, PCA results indicate different shapes of the predicted and actual time series for P_3 and the P_4 graphs in the last two columns.

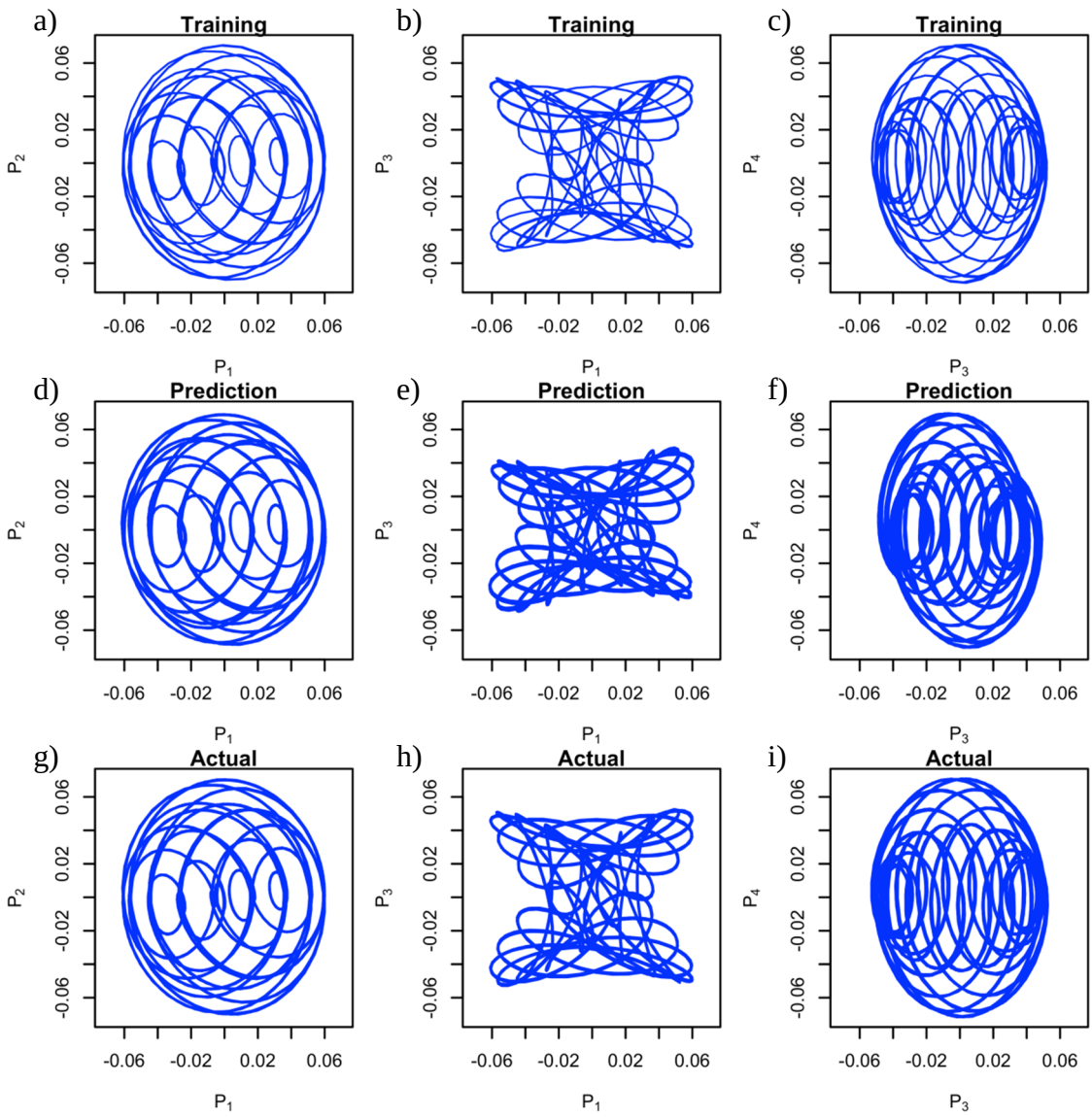


Figure 24. Principal components results for $\delta\phi = \pi/4$ and rest of settings originating from the reference case i.e. A_1=4, A_2=0.3, f_1=0.5, f_2=0.3.

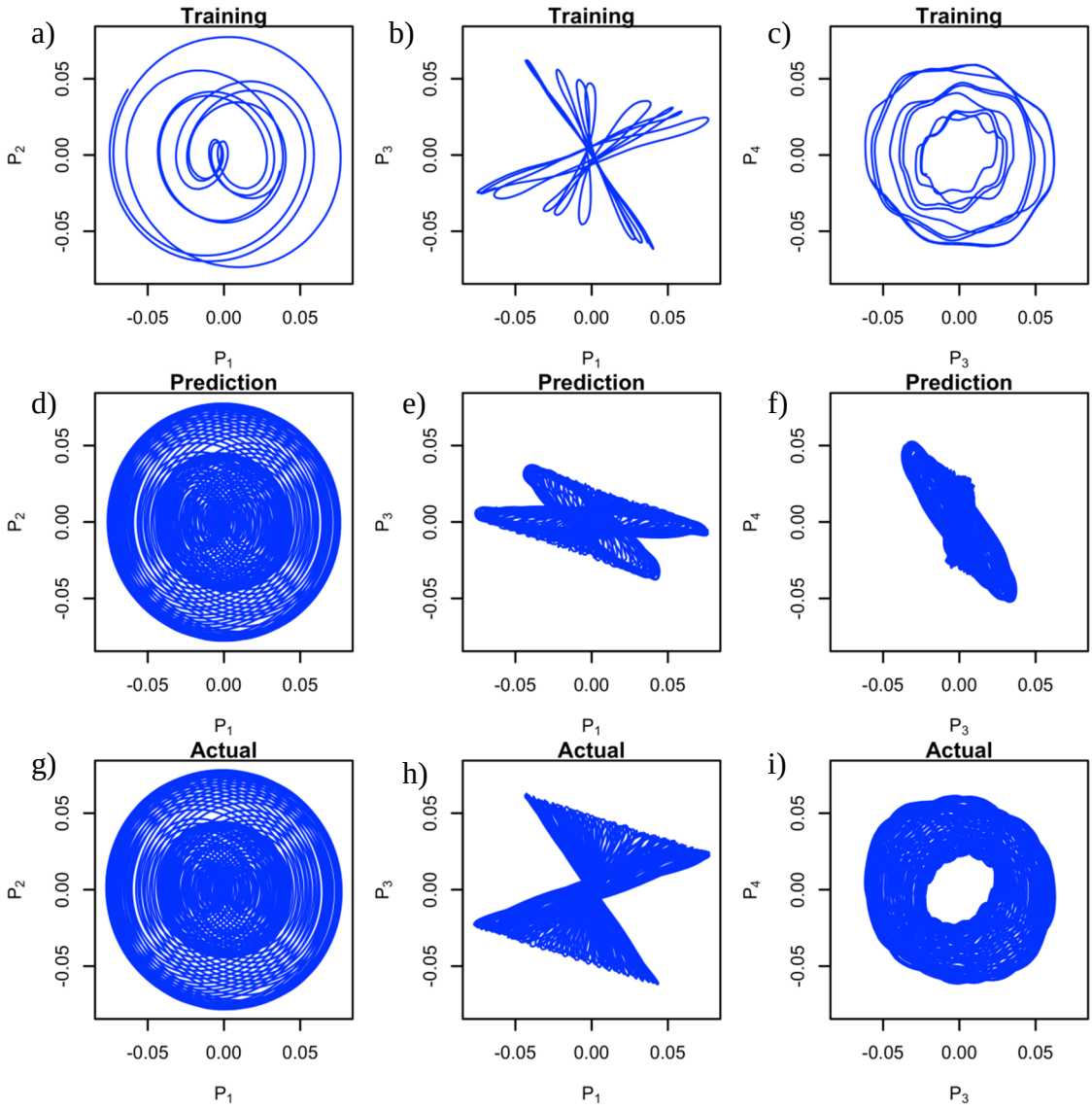


Figure 25. Principal components results for f_1=10 and rest of settings originating from the reference case i.e. A_1=4, A_2=0.3, f_2=0.3, $\delta\phi = \pi/9$.

From the application of different values to the variables: $A_1, A_2, f_1, f_2, \delta\phi$ in order to investigate how PCA performs qualitatively when it is used for MCP purposes, it can be concluded that A_2 and $\delta\phi$ when choosing different values did not in general seem to affect the PC graphs which indicates that PCA performed well and yielded accurate results. On the other hand, when f_1 and A_1 were examined PCA did not perform so well since the predicted PC graphs were of different shape than the actual ones i.e. the predicted PC's did not reproduce secondary oscillations.

5.2.2 Survey of parameter sensitivity

After the examination of the qualitative results in the noisy pendulum example, a survey of the sensitivity analysis was then conducted. A selection of a range of values for the variables: $A_1, A_2, f_1, f_2, \delta\phi$ was chosen. This range was more limited than the ones used for the analysis in 5.2.1 and it was for A_1, A_2, f_1, f_2: 0.3, 0.71, 1, 2, 10 and for $\delta\phi$: π/2, π/4, π/12. For the aforementioned range of values for the variables, the following quantitative criteria were investigated.

The time shift s_1, s_2 at which the maximum correlation occurs of x_{half}, x_p and y_{half}, y_p. Then, the time-shifted standard deviation ratio $r_{\sigma_x}, r_{\sigma_y}$ is described by the equation

$$\begin{aligned} r_{\sigma_x} &= \frac{sd\,(x_{half})}{sd\,(x_p)} \\ r_{\sigma_y} &= \frac{sd\,(y_{half})}{sd\,(y_p)} \end{aligned} \tag{41}$$

and finally the mean $\bar{e}_x, \bar{e}_y$ and standard deviation $sd(e_x), sd(e_y)$ of the errors e_x, e_y which were derived from the following equation

$$\begin{aligned} e_x &= r_{\sigma_x} x_p - x_{half} \\ e_y &= r_{\sigma_y} y_p - y_{half} \end{aligned} \tag{42}$$

Ideally, we want the time shift s_1, s_2 between the two signals x, y to be the same, the $r_{\sigma_x}, r_{\sigma_y}$ to also be the same as well and small values for $\bar{e}_x, \bar{e}_y$ and for $sd\,(e_x), sd\,(e_y)$. Similar time shift s_1, s_2 and standard deviation ratio $r_{\sigma_x}, r_{\sigma_y}$ values indicate that the predictions were of good quality and representative of the two signals x, y. Using the same rationale, small values for $\bar{e}_x, \bar{e}_y$ and for $sd\,(e_x), sd\,(e_y)$ are indicators of a good prediction as well.

The results indicated that using the A_2 range of values seems to be in good accordance with the mentioned desired criteria i.e. very similar s_1, s_2 and $r_{\sigma_x}, r_{\sigma_y}$ with small $\bar{e}_x, \bar{e}_y$ values. Similarly, the same held for the f_1 range of values whereas for the A_1, f_2 values, they seem to result in different s_1, s_2 and $r_{\sigma_x}, r_{\sigma_y}$ and very different $sd(e_x), sd(e_y)$ in the case of f_2. Table 6 indicates the results for each one of the A_1, A_2, f_1, f_2 variables for the range of values used.

	Time shift where max correlation occurs s_1, s_2	**Standard deviation ratio** $r_{\sigma_x}, r_{\sigma_y}$	**Mean of errors** $\bar{e}_x, \bar{e}_y$	**Standard deviation of errors** $sd(e_x), sd(e_y)$
f_1=0.7 ,δφ : π/2	s_1 , s_2 =14	r_{σ_x} =1, r_{σ_y} =2.86	$\bar{e}_x$ =-0.001, $\bar{e}_y$ = -0.0003	$sd(e_x)$ =2.28, $sd(e_y)$ =2.19
f_2=0.3, δφ= π/4	s_1 =8, s_2 =26	r_{σ_x} =1.18, r_{σ_y} =2.87	$\bar{e}_x$ = -0.001, $\bar{e}_y$ =0.44	$sd(e_x)$ =0.78, $sd(e_y)$ =4.24
A_1=0.3 , δφ= π/12	s_1 =3, s_2 =17	r_{σ_x} =1.13, r_{σ_y} =4.76	$\bar{e}_x$ = 0.005, $\bar{e}_y$ = 0.010	$sd(e_x)$ =2.80, $sd(e_y)$ =3.06
A_2=1, δφ =π/12	s_1, s_2 =15	r_{σ_x} =2.94, r_{σ_y} =2.87	$\bar{e}_x$ = -0.004, $\bar{e}_y$ =-0.006	$sd(e_x)$ =2.62, $sd(e_y)$ =3.18

Table 6. Semi-quasi quantitative results for range of values A_1, A_2, f_1, f_2: 0.3, 0.7, 1, 2, 10 and δφ : π/2, π/4, π/12.

5.3 PCA as an MCP method on real wind data

The next attempt for PCA to be used as an MCP method was made on real wind speed data taken from Gogarbank (GGB) and Blackford Hill (BFH) meteorological stations in Scotland, UK [75]. The data specifications were explained in detail in

Chapter 4 and in Table 9 and Figure 31 of Chapter 6. In our case, Gogarbank was treated as the reference site, where the historical data are being used from, and Blackford Hill as the target site, the site for which our aim is to predict for.

At this point it is important to mention that the gaps in the data used throughout all the real wind data analysis in this chapter but also in Chapter 6 were identified and then interpolated. The implications of this and possible solutions will be discussed in the final chapter, Chapter 7.

Variables such as temperature and pressure were initially included in the PCA-MCP analysis but were emitted later on. Pressure specifically did not seem to play an important role in the analysis and temperature was excluded since it was found that it introduces seasonality in the data which would result in more biased results.

5.3.1 PCA-MCP for wind speed

The MCP methodology steps described in section 3.6.3 were followed for this analysis. The number of channels in this case are $N_0 = 2$ since only wind speed is used i.e. $y_{j_0}(t) = u_{1,} u_2$ where u_1 is the wind speed of Gogarbank and u_2 the wind speed of Blackford Hill.

5.3.2 PCA-MCP wind speed parameter analysis setup

The parameters which were examined were: the window length, M_w i.e. the number of days used for the columns of the time-delay matrix and the truncations, M_t i.e. the number of principal components used for the PCA analysis which were originally obtained from the singular values spectrum (Λ from equation (12)) graph. Window length M_w from 1 to 21 days was used. Table 7 indicates the values of these examined parameters. For small window length, the results were poor for large truncations. For window lengths of more than 14 days, the physical limitations of the computer prevented the use of truncations larger than the very shortest of 3 and 6. For

this reason here, the intermediate cases are presented which will then be used to explore the full sensitivity analysis in Chapter 6.5.

Training periods (concurrent data)	1999, …, 2010	
Window length for training (M_w)	1 and 3 days	7 and 14 days
Number of principal components retained for prediction (truncations) (M_t)	3,4,6	6,12,18
Prediction period	1999 - 2010	

Table 7. Parameter settings used for the PCA-MCP wind speed analysis.

5.4 PCA-MCP wind speed calibration and results

The calibration of the predicted wind speed results and comparison with the original data was essential to be undertaken since the PCA-MCP predictions were shorter than the length of the historical data by the size of the window length M_w. Hence, the aim was to match the predicted 'shorter' reference data with the historical data. The calibration method presented here was an intermediate step to achieve the final calibration method used as described in section 3.6.2, Table 3 and used in Chapter 6. Various possibilities were explored regarding the calibration, all guided by the aim to calibrate the predictions so as to have the same variance and mean values as the actual data. The earlier attempts explained here, base this calibration on matching the mean and variation of the prediction to those of the training period. The rationale was that, if the calibration of the target site was similar to that of the reference site, then it would be possible to determine the calibration for the reference site and transfer that calibration to the target site. The two calibration methods used are two variants of linear regression, described in equations (43) and (44), respectively.

In contrast to this approach, the final calibration, which was introduced in the formal development of the model in section 3.6.2, does not require the calibration between reference and target site to be similar, but that the loss of variance by truncating the singular vectors (both in dimensions and number of input channels) will

be the same irrespective of whether the method is applied to the reference data from the training period or from the historical period.

The mean average error MAE was calculated and therefore shifted across the data to examine how well it matches the predicted values to different sections of the historical data. In more detail, initially the data were normalised as following:

$$\begin{aligned} u^{*}_{1,p} &= u_{1,p} * \sigma_1 + \mu_1 \\ u^{*}_{2,p} &= u_{2,p} * \sigma_2 + \mu_2 \end{aligned} \tag{43}$$

where μ_1, μ_2 are the mean and σ_1, σ_2 standard deviation coming from the original training data, $u_{1,p}, u_{2,p}$ and $u^{*}_{1,p}, u^{*}_{2,p}$ are the normalised predicted wind speeds in m/s for GGB and BFH respectively. Furthermore, after several rescaling attempts so that the original data 'match' with the predicted ones, the best rescaling method found by trial and error was of the form:

$$\begin{aligned} A &= \min(u_{1,p}), B = \min(u_{2,p}) \\ u^{*}_{1,a} &= u_{1,p} * (\frac{-\mu_1}{A}) + \mu_1 \\ u^{*}_{2,a} &= u_{2,p} * (\frac{-\mu_2}{B}) + \mu_2 \end{aligned} \tag{44}$$

With the use of the minimum values and the mean of the training signals $u_{1,p}, u_{2,p}$ for both GGB and BFH as correction factors and by taking also into consideration that wind speeds are bigger than zero the rescaled signals for GGB and BFH were found to be $u^{*}_{1,a}, u^{*}_{2,a}$.

All data were then examined in order to find where the best matches between the predicted and original data existed. Then, the histograms of $u_1, u_2, u^{*}_{1,a}, u^{*}_{2,a}$ were produced in order to investigate how well the predicted calibrated PCA results performed when compared to the actual wind speed data for both sites. Some indicative results can be shown in Figure 26 and Figure 27.

Figure 26 presents in the top left plot (a) the singular values plot from the PCA analysis results. This graph was used for the determination of the truncation values used in the parameter settings models. It shows three sections, an initial set of six large but rapidly reducing singular values, followed by three singular values of similar magnitude, and final a long tail of gradually decreasing values. From this, one there are two choices of truncation suggested, either a truncation of $M_t = 6$ to include only the first set or a truncation of $M_t = 9$ to include the three singular values forming the central set. In Figure 26b) the mean absolute error (MAE) for GGB historical and predicted data (black line) and with the red line the MAE for BFH historical and predicted data is depicted against time shift, s for matching the actual wind speed against prediction which is shorter than the actual by the window length used to create delay matrix. The MAE formula is given in equation (45):

$$\begin{aligned} MAE_{1,ws} &= \sum \left| u_1 - u_{1,s}^* \right| \\ MAE_{2,ws} &= \sum \left| u_2 - u_{2,s}^* \right| \end{aligned} \tag{45}$$

As it can be seen for these specific training and historical periods and parameter choices, there is a clear minimum in the *MAE* for both sites at the same time shift of around 10. The PCA results seem to be relatively good since the MAE of BFH is relatively close to that for the GGB, they are of the same shape and the amount of error is around 1.5-2 m/s, i.e. relatively small compared to the much larger values at larger time shifts.

The rest of the graphs in Figure 26, depict the probability density function (pdf) histograms of the actual and predicted data for both sites. In Figure 26c) shows the distribution of the wind speeds at the reference site (Gogarbank) for the prediction period while in Figure 26d) shows the prediction of the wind speeds at the reference site as the histogram, with the actual data from panel c) reproduced as the red line for direct comparison. Likewise, Figure 26e) and f) show the actual wind speed at the target site for the prediction period and its prediction. It can be shown from graphs d) and f) that PCA seems to over predict for small wind speed values of less than 6 m/s but above this

value, the actual and predicted measurements seem to be in good accordance for both sites.

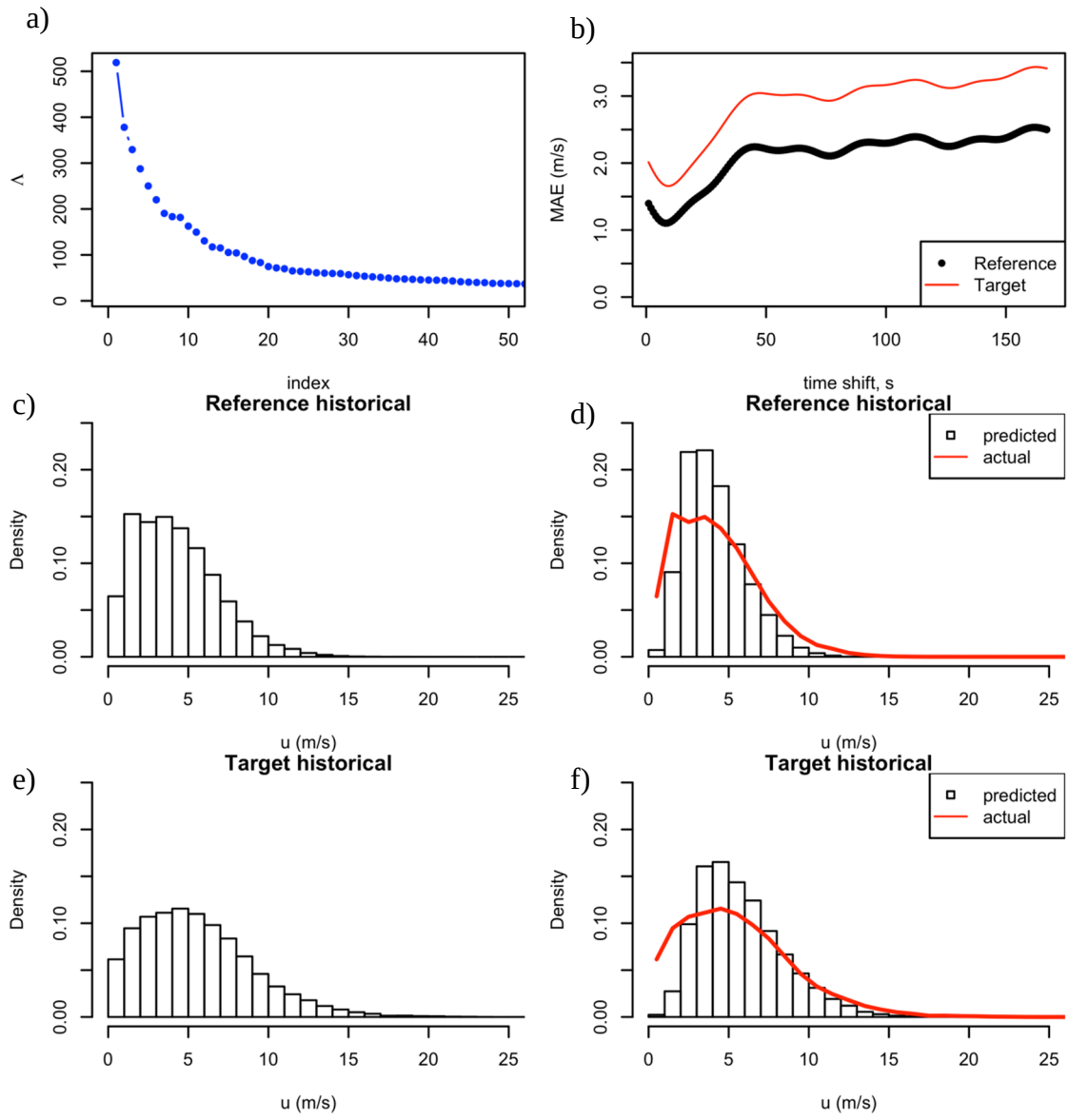


Figure 26. PCA-MCP results for historical data 1999-2008, training year 2009-2010, truncation M_t =6, window length M_w =7 days.

On the other hand, a 'bad' result of PCA is depicted in Figure 27. It can be seen that the reference and target site errors indicate an oscillatory movement which could possibly suggest a daily cycle in the data. However, it should be noted that the errors seem to be of similar shape and a minimum value can be identified for both GGB and BFH similar

to the one of Figure 26. From the differences in the two Figure 26, Figure 27 which contain different window lengths and truncations, it can be concluded that the choice of the parameters seems to be of great importance for the quality of the PCA-MCP results. This leads to the initial observation that prior to undertaking PCA, a careful evaluation and choice of the parameters which will be used for the analysis purposes should be conducted.

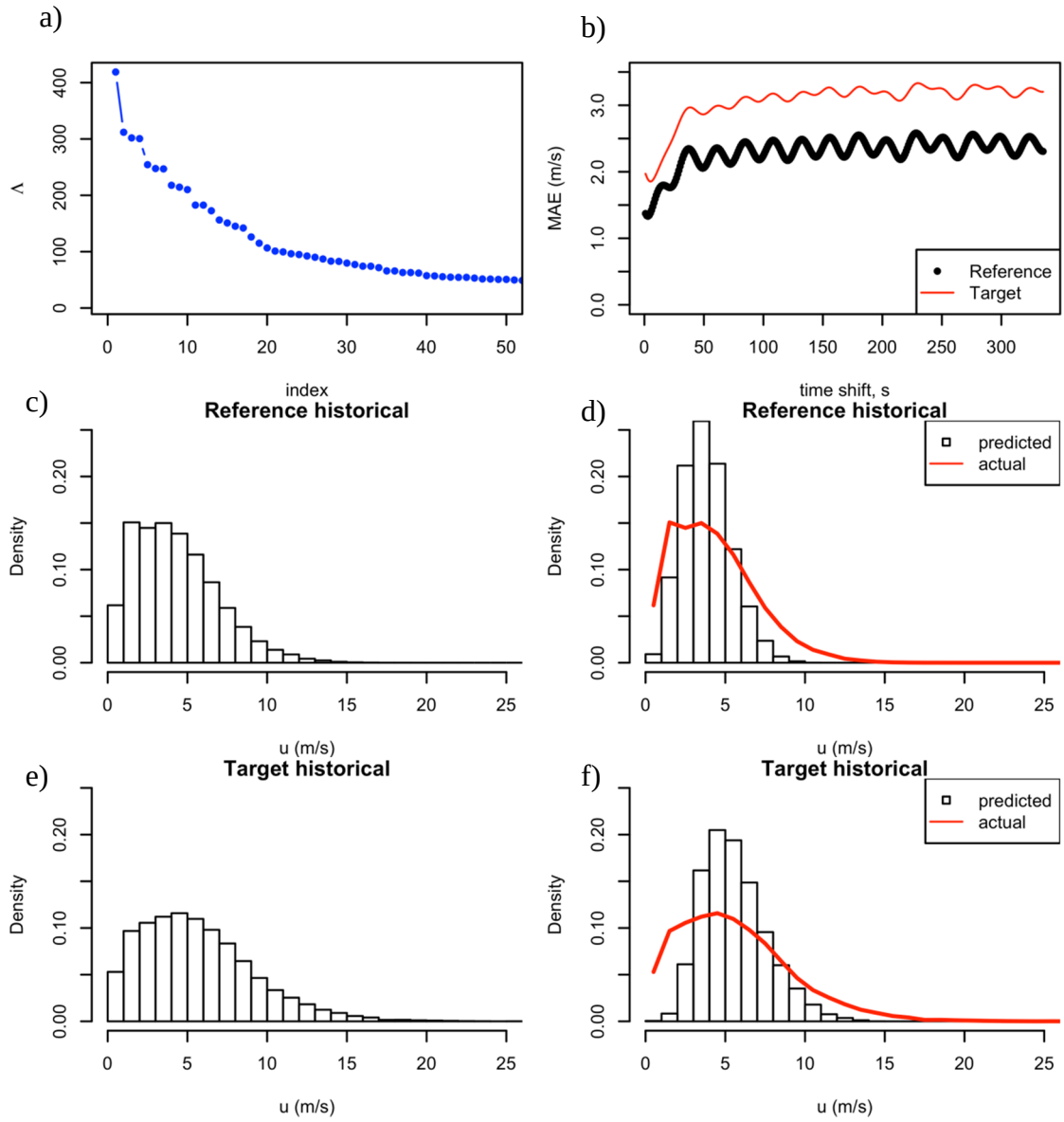


Figure 27. PCA-MCP results for historical data 1999-2009, training year 2010, truncation M_t =18, window length M_w =14 days.

In general, the histograms as it can be seen in Figure 27 are quite similar to the ones of Figure 26. They are very good illustrations of the performance found at all parameter settings and hence it can be concluded that there exists a persistent tendency for over prediction for low wind speeds for both GGB and BFH but more accurate predictions for larger wind speeds for the majority of the models. In addition, the knowledge of the errors for the reference site enables 1) a calibration for the predicted results 2) an error estimation for the target site also and most importantly 3) the evaluation of the quality of these predictions. Overall, PCA seemed to perform well for GGB and BFH regarding wind speed measurements considering that a sheltered (reference) site was used to predict for an exposed (target) site. In the following section, the performance of all models will be compared by using the error in the wind speed prediction to explore how the performance varies as the model parameters are varied.

5.5 PCA-MCP wind speed sensitivity analysis

The next step was to find the relative error for the different settings of window length (M_w) and truncation (M_t) as described in Table 7. The models used for the error analysis are shown in Table 8:

Years	
Historical	**Truncation**
1999-2009	2010
1999-2008	2009-2010
2000-2009	2010
2000-2008	2009-2010
2001-2009	2010
2001-2008	2009-2010
2002-2009	2010
2002-2008	2009-2010
2003-2009	2010
2003-2008	2009-2010
2004-2009	2010
2004-2008	2009-2010

Table 8. Datasets used for the relative error analysis.

Initially, the errors $E_{u,1}, E_{u,2}$ of GGB and BFH respectively were calculated for all the models and the aforementioned different window and dimension settings as following:

$$E_{u,1} = \left(\frac{u_1 - u_{1,a}^*}{u_1} \right) * 100$$
$$E_{u,2} = \left(\frac{u_2 - u_{2,a}^*}{u_2} \right) * 100 \qquad (46)$$

Then, the relative error E_1, E_2 of GGB and BFH which is the measure of sensitivity of $E_{u,1}, E_{u,2}$ was calculated. That is the error $E_{u,1}, E_{u,2}$ as calculated in equation (46) but rescaled and normalised with respect to the set up value of window length M_w=7 and truncation: 6 for the GGB dataset, named E_R for each model of Table 8 respectively. The value of E_R was specifically chosen since it was a benchmark setting of 'middle' values for the PCA-MCP parameter settings of Table 7.

$$E_1 = \left(\frac{E_{u,1}}{E_R} \right)$$
$$E_2 = \left(\frac{E_{u,2}}{E_R} \right) \qquad (47)$$

Figure 28 indicates the boxplots of the relative error E_1, E_2 in the top (a) and absolute relative error $|E_1|, |E_2|$ in the bottom graph (b) for GGB (blue) and BFH (red) for the different window lengths used as mentioned above. Boxplots show the range of values observed divided into quartiles, with the central horizontal line showing the median of the error, the boxes either side the 2nd and 3rd quartile, and the range from the box to the whiskers the 1st and 4th quartile, respectively. Outliers, defined as 2.5 times the extend of the 2nd or 3rd quartile from the median, are shown as individual circles. As it can be seen, more window length results in less E_1 for GGB. For example when comparing the error for window 1 and 14 with windows 3 and 7 we can see a significant

reduction in E_1. However, E_2 tends to become more for BFH for the same window length choices. The closest pair of E values for both sites seems to be the window length of 3 days.

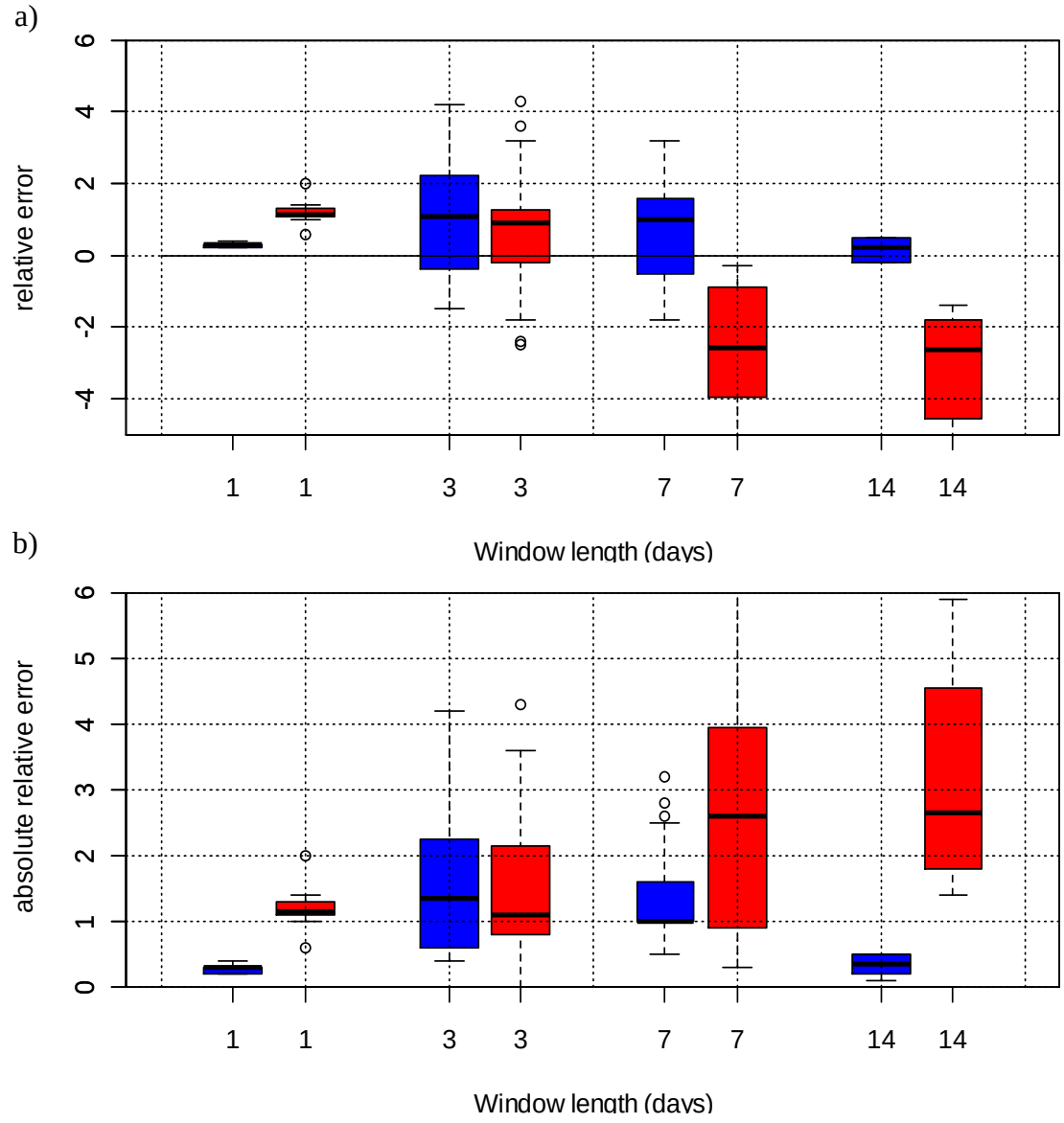


Figure 28. Relative and absolute relative error for window length M_w: 1, 3, 7, 14 days where blue represents GGB (reference) and red BFH (target).

Next, Figure 29 illustrates E_1, E_2 and $|E_1|, |E_2|$ for the different truncation values. As it can be seen for GGB, the 3 different truncation values do not seem to result in a significant change of E_1. On the other hand, regarding BFH for the highest

number of truncations, i.e. 12 dimensions in this case, E_2 seems to be the least and also closest to the error of GGB when compared with truncations 3 and 6.

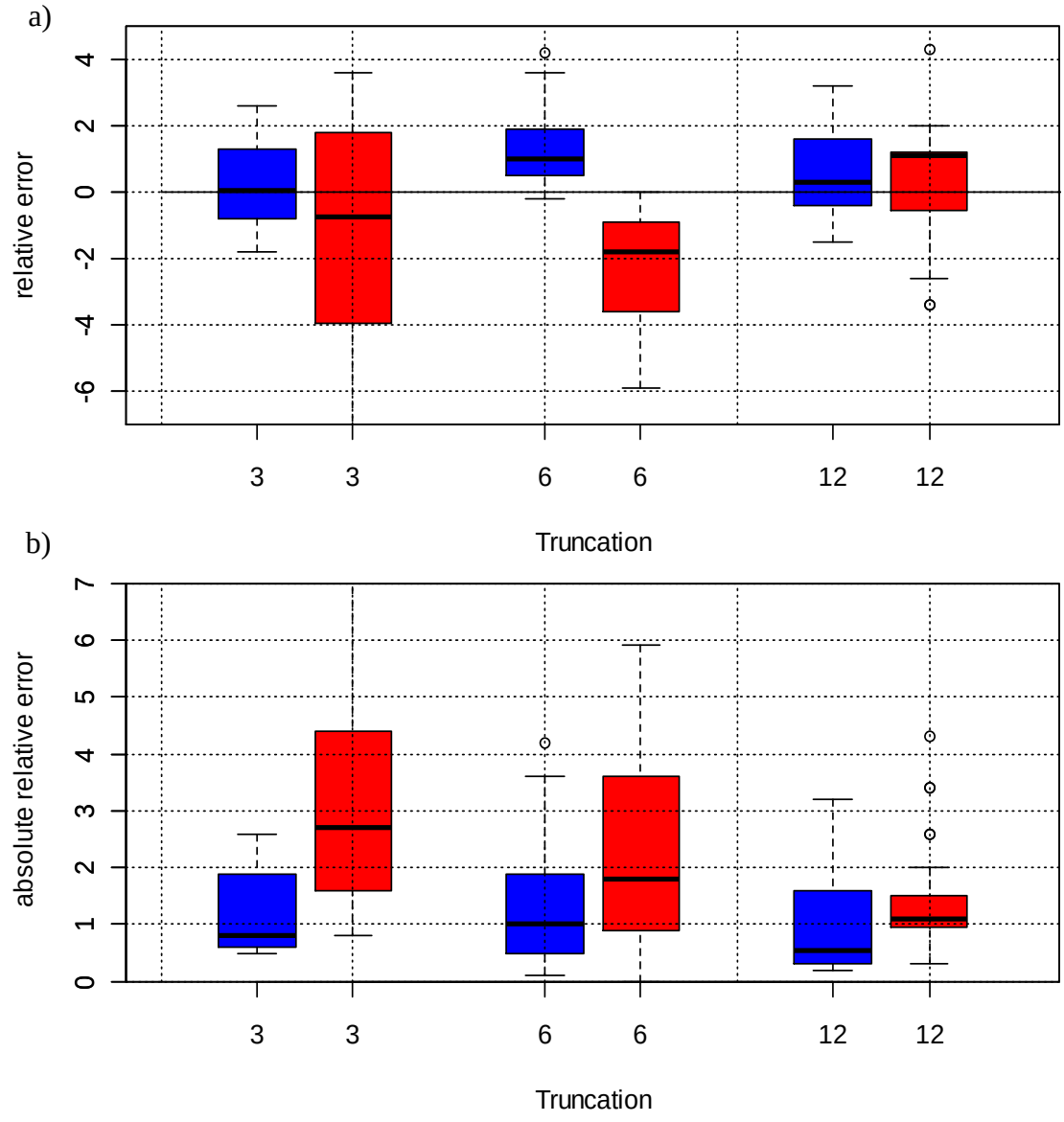


Figure 29. Relative and absolute relative error for truncation M_t: 3, 6, 12 where blue represents GGB (reference) and red BFH (target).

Since it was observed from Figure 28 and Figure 29 that E seems consistent between the two sites for the parameter choices of window length 3 and truncation 12, more graphs were created to investigate further the error behaviour with respect to these

interesting parameter choices. In the top graph of Figure 30 for truncation 12 and 1 day window, we can observe a very low error for GGB, the reference site meaning that the wind regime of the local site is captured and no irrelevant information from other days is added. It can also be observed that the median error of BFH for all window lengths is relatively constant but its spread is increasing as the window increases too. Thus, it can be seen that in order to get the best parameter match and thus lowest error at both sites we need to use the shortest window. For the different truncations with respect to window 3 in the bottom graph of Figure 30 it can be clearly seen that the lowest errors for both target and reference sites are for truncation 12. Some outliers can also be observed which indicates that some years models did not yield 'good' PCA results. Therefore from Figure 30 it can be concluded that the best combination which results in the lowest E for both sites is observed for truncation 12 and window length 3.

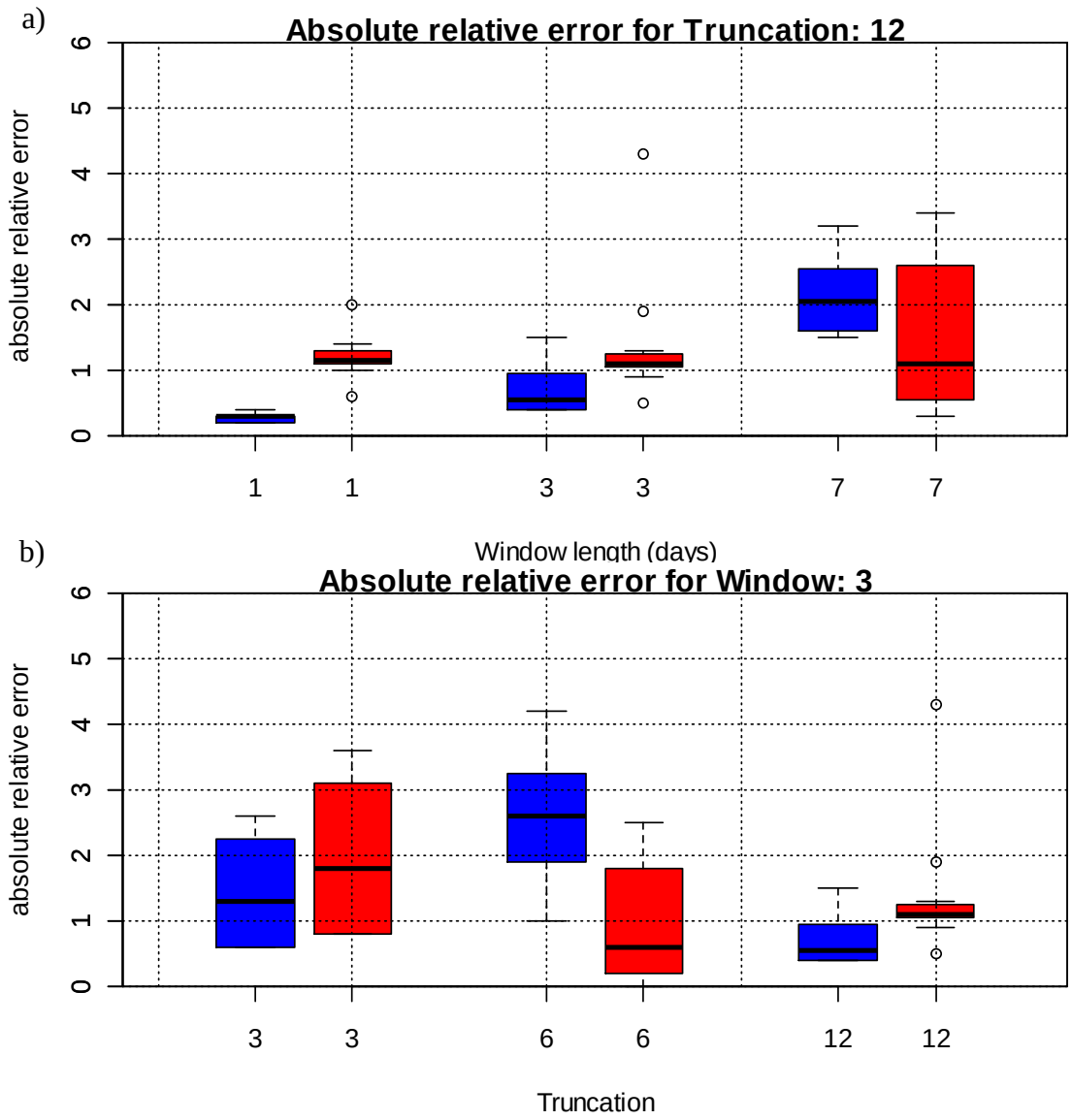


Figure 30. Absolute relative error for truncation M_t =12 and window length M_w =3 where blue represents GGB (reference) and red BFH (target).

Hence, the main initial observations from applying PCA as an MCP method on wind speed data for two datasets using several years and different parameter choices can be drawn. Adding more time series information seemed to be better for local predictions (reference site) but worse for the target site (Figure 28). On the contrary, more global information dynamics, i.e. truncations added, yielded better results for the target site but worse for the local site (Figure 29). To sum up, the choice and length of historical and training periods for GGB and BFH did not seem to affect significantly the results, however the choice of PC's firstly and window length secondly seem to be of big

importance for the quality of the results. However, a key weakness of the method developed so far is the persistent error in the low wind speed range, where the predicted frequency of very low wind speeds (0 to 3 m/s) is too high at the expense of moderate wind speeds (4 – 6 m/s). This problem was traced back to the calibration approach used at this stage of the research, leading to the final calibration approach outlined in section 3.6.2, which was used throughout the final testing and validation work presented in the following chapter. Chapter 6 will use similar analysis principles as Chapter 5; however, this will be done in more depth and by introducing in the analysis one more very important MCP variable, wind direction.

Chapter 6 PCA-MCP method final applications and results

This chapter will present and evaluate the final stage of the PCA-MCP algorithm development. Initially, the inclusion of another variable, wind direction, in the PCA-MCP analysis will be introduced, and then a presentation of examples from a poor and good performance of the method will be explored. Finally, error and bias measures criteria as well as a comparison with simple linear regression as an alternative MCP method will be presented in order to assess the performance of the PCA-MCP technique and draw the final conclusions of this research.

6.1 Wind direction as an PCA-MCP invariant

Next step in the MCP analysis is to include wind direction as a variable. Wind direction is usually included in the MCP analysis alongside with the wind speed since these too variables provide more integrated wind information about both reference and target sites and hence contribute in a more robust MCP analysis [18]. The method to introduce wind direction as a variable which will be used for this MCP analysis is to create a vector combination consisting of wind speed and wind direction with the components u, v .

The selection of using these components was also based on the fact that, for example, north winds jump from 0° to 360°. Thus, there exists artificial discontinuity in the data whereas using the u, v components this is avoided. In general, wind direction prediction alone is weak. Therefore, in our case the actual wind direction was not looked at but the link between the reference and target wind direction was examined as an invariant. Positive u is coming from west to east direction and negative u from east to west, positive v from north to south and negative v from south to north. The u, v linear combination is of the following form:

$$u_1 = -U_1 \sin\left(\frac{2\pi}{360}\theta_1\right) \tag{48}$$

and

$$v_1 = -U_1 \cos\left(\frac{2\pi}{360}\theta_1\right) \tag{49}$$

where U_1 is the wind speed and θ_1 is the wind direction (in degrees) for the reference site. Equivalently, the same relationships hold for U_2 and θ_2 of the target site i.e.:

$$u_2 = -U_2 \sin\left(\frac{2\pi}{360}\theta_2\right) \tag{50}$$

and

$$v_2 = -U_2 \cos\left(\frac{2\pi}{360}\theta_2\right) \tag{51}$$

Now, we have four signal channels u_1, v_1, u_2, v_2 and hence, four columns in the SVD matrix i.e. N_0=4 instead of two that we had before when wind speed was only used. For the prediction part of the MCP algorithm we need to be careful when we want to normalise the results back to actual wind speed ones for both the reference and target sites. This is the reason for which a scaling methodology was used carefully which was explained in section 3.6.1.

6.2 PCA-MCP methodology for wind speed and direction combination

Initially, as described in detail in Chapter 4, the method was used to predict wind speeds up to 24 hours ahead i.e. was used for wind speed forecasting. It predicted wind speeds based on a set of previous measurements which were used to construct an attractor in an optimally defined phase space as a 'training set'. Current wind measurements were then projected to onto that phase space to find most similar previous measurements. By tracing the evolution of these similar previous data, it became possible not only to forecast the wind speed but also to obtain a measure of the expected forecasting uncertainty [97].

6.3 Data and analysis setup

6.3.1 Dataset used in the PCA-MCP analysis

The PCA-MCP data came from the same source [75] as in section 4.2 for the forecasting purposes. For the purposes of this analysis 8 sites in Scotland, UK sites were used and as in section 4.2, the sites used anemometers of 10m high above ground and the data records used spanned from 2000-2010 with hourly mean wind readings with the wind speed stored to the nearest knot (1kn=0.5144 m/s) and the wind direction in degree to the nearest 10°. Table 9 indicates the position and characterisation of the 8 sites used and Figure 31 their position in the map of the UK.

Station name	Latitude	Longitude	Characterisation
1) Stornoway	58.2138	-6.31772	coastal, exposed
2) Blackford Hill	55.9228	-3.18750	inland, exposed
3)Machrihanish	55.4408	-5.69571	coastal, exposed
4) Salsburgh	55.8615	-3.87409	inland, exposed
5) Prestwick Gannet	55.5153	-4.58343	coastal, sheltered
6) Gogarbank	55.9284	-3.34294	inland, sheltered
7) Port Ellen	55.6813	-6.24866	coastal, exposed
8) Bishopton	55.9068	-4.53122	inland, sheltered

Table 9. Summary of Met.Office stations used in the analysis with latitude and longitude in the decimal degrees and characterisation.

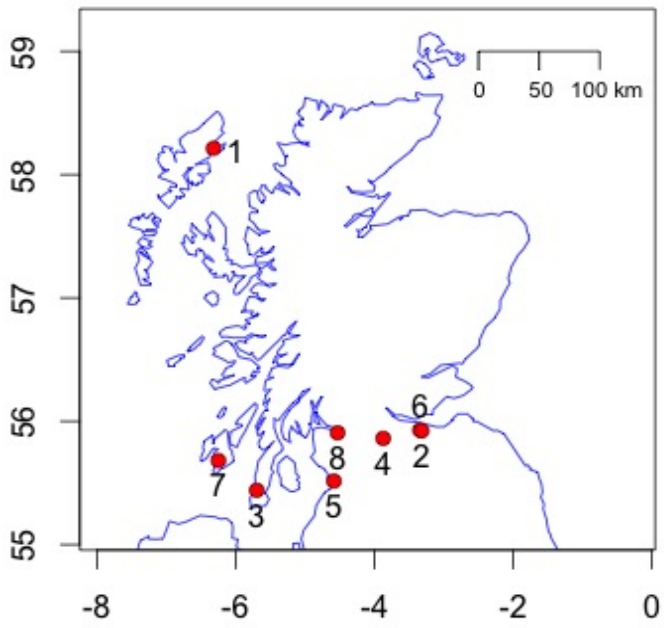


Figure 31. Map of the data used for the PCA-MCP analysis

6.3.2 PCA-MCP parameter analysis setup

The parameter analysis setup for the dataset described in the previous section 6.3.1 is described in Table 10. The principal components used were determined after the first application of PCA based on the singular values spectrum result and the window length (M_w) which was used in the setup of the time-delay matrix was determined according to the truncations (M_t) of the relevant components. Hence, both parameters were determined at the 'Correlation' part of the process as mentioned in Table 3 and were therefore carefully chosen after extensive trial and error attempts performed for the PCA-MCP analysis which due to brevity are not presented here. All 8 stations were used for all possible permutations of pairs as reference and target stations i.e. the number of models used were 8x8=64-8=56 in order to investigate how the method reacts when each site has been used in the analysis i.e. as a reference or target site. The method was also applied for all one-year periods in the span of 2000-2010 used in turn as training (concurrent) years.

Training periods (concurrent data)	2000, …, 2010
Window length for training (M_w)	1 (24h) and 2 days (48h)
Number of principal components retained for prediction (truncations) (M_t)	3,6,9,12
Prediction period	2000 - 2010

Table 10. Parameter settings used for the PCA-MCP analysis.

6.4 Comparison of PCA-MCP with simple linear regression

The next step in the validation of the PCA-MCP methodology is to compare it with an established MCP method such as standard linear regression. The linear regression in this case was established with the following linear model:

$$U_2 = \beta_0 + \beta_1 U_1 + \varepsilon \tag{52}$$

where, as denoted in section 6.1 of this chapter, U_1, U_2 are the reference and target wind speed respectively, β_0 is the intercept, β_1 is the slope and ε is the error term. Hence, the target wind speed U_2 is the dependent variable and the reference wind speed U_1 the independent variable in the linear model. After performing linear regression in R, a new variable $U_{2,pred}$ is created being the target wind speed prediction denoted as:

$$U_{2,pred} = \hat{\beta}_0 + \hat{\beta}_1 * \sqrt{u_{1,past}^2 + v_{1,past}^2} \tag{53}$$

was created with $\hat{\beta}_0$ the estimate of the intercept, $\hat{\beta}_1$ the estimate of the reference wind speed in the linear model of equation (52) times the magnitude of the wind vector combination of the actual wind speed of the reference historical data $u_{1,past}, v_{1,past}$.

The linear regression error E_{lr} was then calculated which will be depicted in graphs in the following sections of this chapter. It is of the form:

$$E_{lr} = \left\langle U_{2,pred} \right\rangle - \left\langle \sqrt{u_{2,past}^2 + v_{2,past}^2} \right\rangle \tag{54}$$

i.e. the difference of the mean value of the wind vector combination magnitude to actual wind speed of the target historical data $u_{2,past}, v_{2,past}$ from the mean value of the predicted target wind speed $U_{2,pred}$.

6.5 PCA-MCP sensitivity analysis

The performance of the PCA-MCP method in comparison with the linear regression as an alternative MCP method was essential to be quantified as part of the validation of the new method. This was achieved with the use of different statistical sensitivity analysis measures as explained in the following section 6.5.1.

6.5.1 Error and uncertainty measures

Initially, the error was quantified as the difference of the prediction and actual wind speed distribution for the reference; target and linear regression was calculated as follows:

$$e_{ref}(u)du = \left(P_{ref,pred}(u) - P_{ref,actual}(u)\right)du \tag{55}$$

and

$$e_{tar}(u)du = \left(P_{tar,pred}(u) - P_{tar,actual}(u)\right)du \tag{56}$$

where e_{ref}, e_{tar} are the difference of the reference and target probability density function prediction and the probability density function of the actual reference and target data respectively and

$$e_{lr}(u)du = \left(P_{lr,pred}(u) - P_{lr,actual}(u)\right)du \quad (57)$$

is the difference of the linear regression probability density function prediction and the probability density function of the actual reference data.

Then, the Mean Absolute Error (MAE) was calculated [107] as:

$$MAE_{ref} = \int_{u=0}^{\infty} \left| e_{ref}(u) \right| du \quad (58)$$

and

$$MAE_{tar} = \int_{u=0}^{\infty} \left| e_{tar}(u) \right| du \quad (59)$$

and finally,

$$MAE_{lr} = \int_{u=0}^{\infty} \left| e_{lr}(u) \right| du \quad (60)$$

where MAE_{ref}, MAE_{tar} and MAE_{lr} are the sums of the absolute error as defined in equations (58), (59) and (60) for the reference, target site and linear regression and are a measure of the goodness-of-fit of the predictions. In the following, the distribution error was calculated on probabilities in N wind speed bins of width, $\Delta u = 1 m/s$ the *MAE* for each distribution was calculated as:

$$MAE_{j} = \sum_{i=1}^{N} \left| e_j(u_i) \right| \Delta u;\ j = \{ref, tar, lr\} \quad (61)$$

(paragrafos!!)Since we use an error of two distributions of unit area, an error measure corresponding to the bias as defined in equation (33) of Chapter 4.3,

$$Bias = \int_{u=0}^{\infty} e(u)du \quad (62)$$

is always exactly zero. Since the bias as defined in equation (33) is identical to the difference between the predicted mean wind speed and the actual mean wind speed, this is used here as the measure for the bias.

To summarise the performance of the PCA-MCP against the standard linear regression MCP, a performance index (PI), was defined as the ratio of the MAE_{tar} over MAE_{lr}:

$$PI = \frac{MAE_{tar}}{MAE_{lr}} \tag{63}$$

6.6 PCA-MCP results

6.6.1 A 'good' example

Using the model of Stornoway as reference site and Salsburgh as target site for 2007 with truncation $M_t = 12$ and window $M_w = 24$h the PCA-MCP results are shown in the following graphs. The singular values (Λ) spectrum graph used to determine step 3 of the PC-MCP algorithm in Table 3 for the PCA analysis is presented in Figure 32. As it can be seen, the lambda values have rapid cut off after 5 and 9 and between 12 and 14. Hence, the choice of truncation $M_t = 12$ for this example is justified.

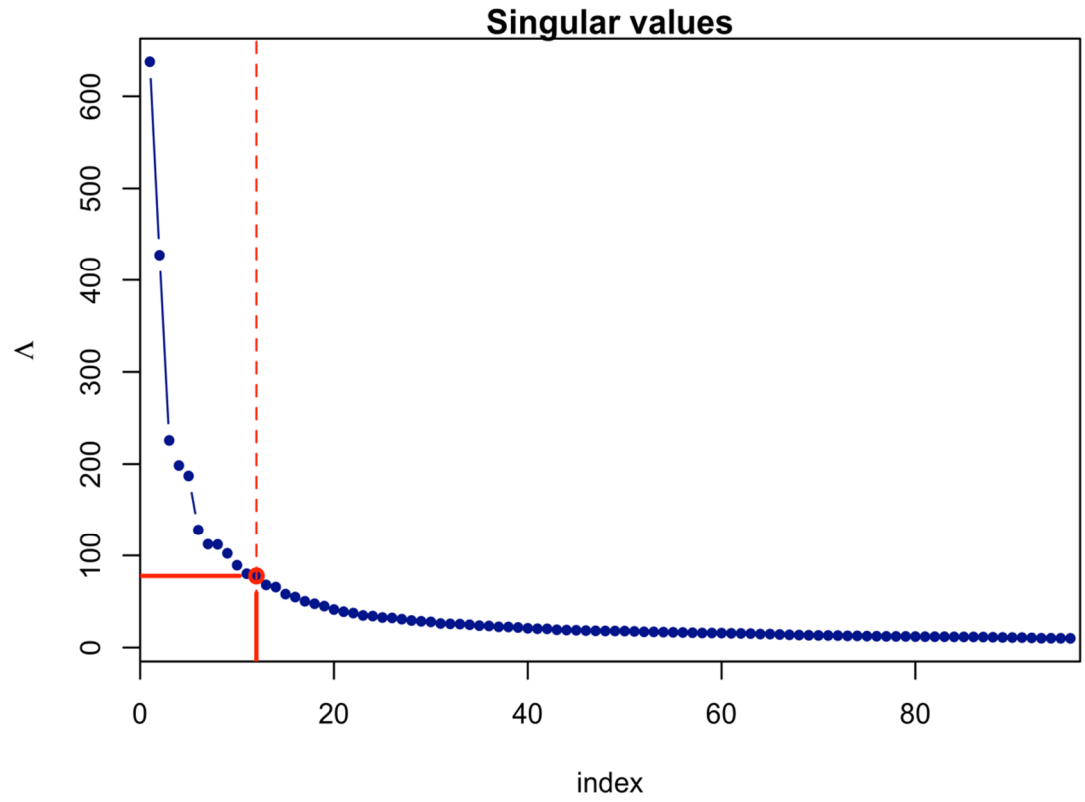


Figure 32. Singular values spectrum for Stornoway and Salsburgh stations for the training year 2007, window length $M_w = 24$h, truncation $M_t = 12$.

As it can be seen from the wind speed histograms of Figure 33 and Figure 34, the predicted wind speeds match to a big extent with the actual wind speeds for both the reference and target sites. Thus, PCA performed well for these specific data. However, a slight overprediction of PCA-MCP can be observed for both reference and target sites.

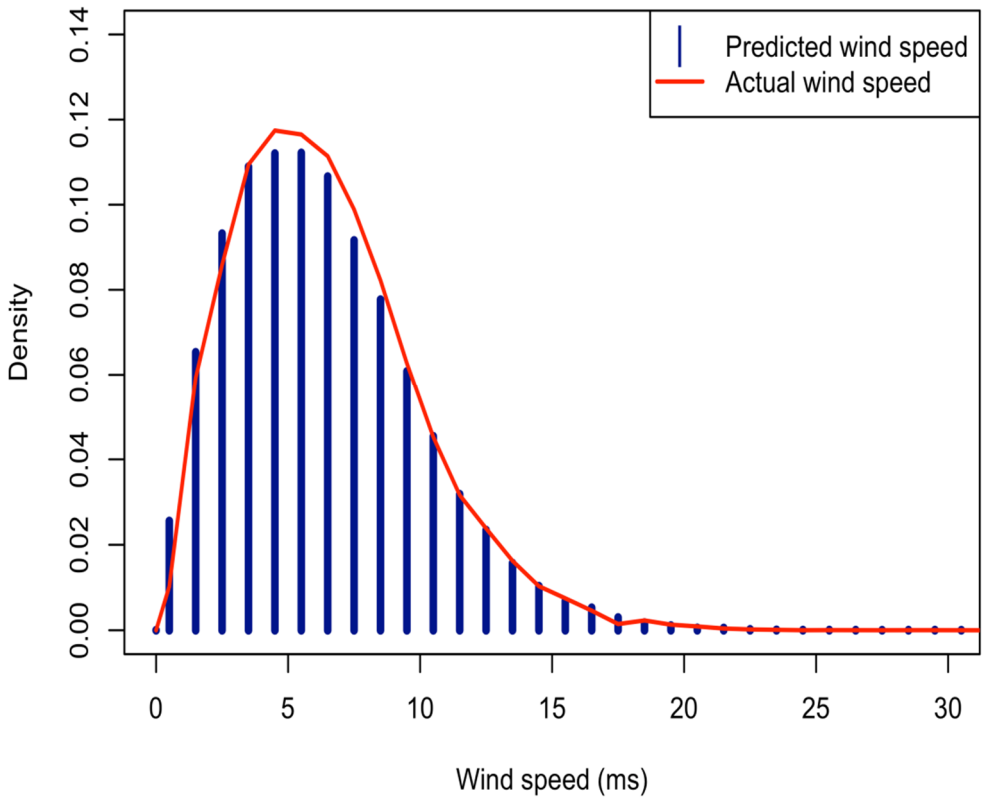


Figure 33. Actual and predicted wind speed for Stornoway (reference) for training year 2007, window length $M_w = 24$h, truncation $M_t = 12$.

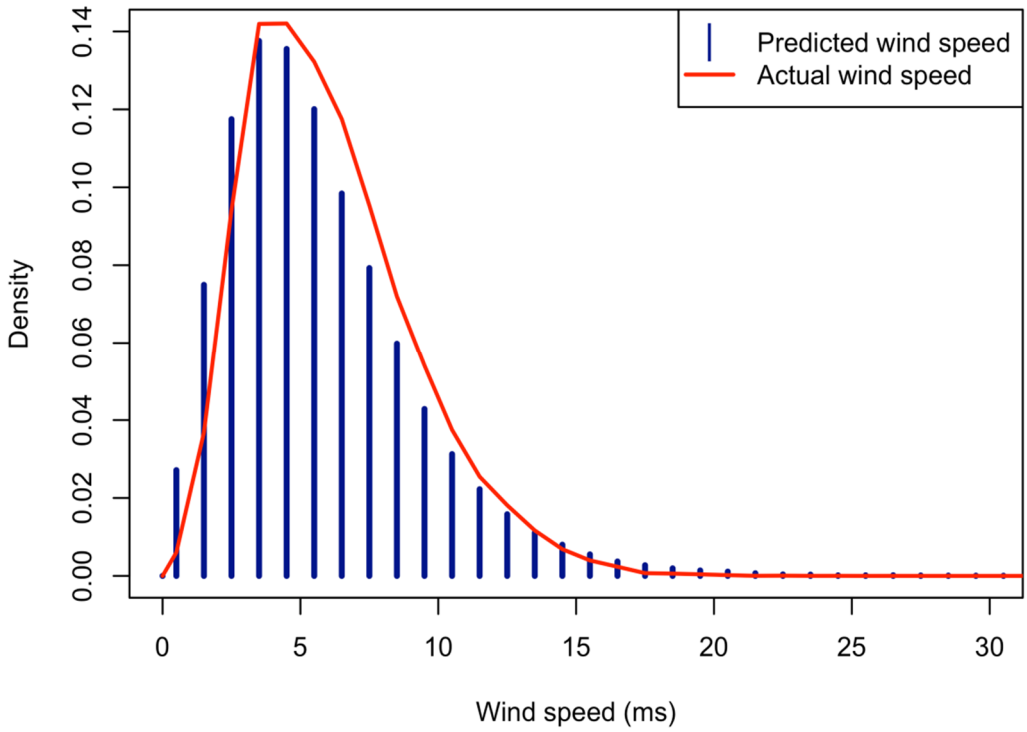


Figure 34. Actual and predicted wind speed for Salsburgh (target) for training year 2007, window length $M_w = 24$h, truncation $M_t = 12$.

Next, the wind roses which describe the wind distribution are shown. Regarding the reference site, the actual and predicted wind roses seem to be quite similar to each other as seen in Figure 35 and Figure 36. On the other hand, the target site wind roses indicate some differences and as it can be seen from the predicted target site wind rose in Figure 38, it indicates overprediction of the wind speeds in the southeast direction in comparison with the actual target data in Figure 37 which shows the prevailing wind in the southwest. It can be concluded from Figure 37 and Figure 38 that PCA-MCP seems to predict quite well the wind speed distributions for both the actual and predicted data but not so good the wind direction even in the 'good' example described here. As it will be discussed in Chapter 7, this is one of the issues to be considered in the future validation work of the PCA-MCP method.

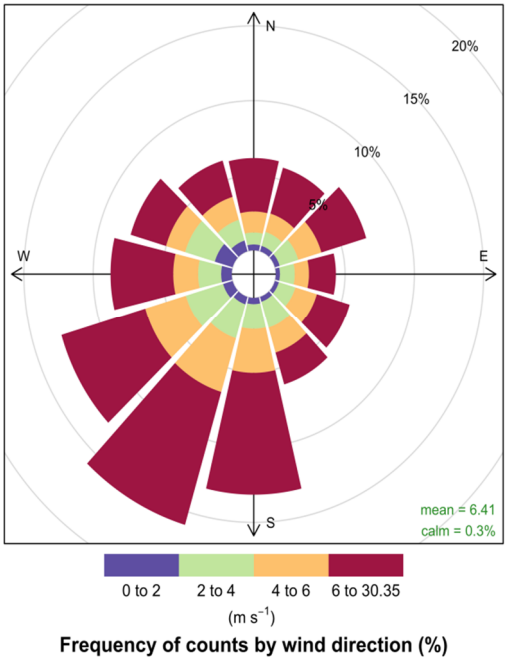


Figure 35. Wind rose for Stornoway (reference) actual data for training year 2007, window length $M_w = 24$h, truncation $M_t = 12$.

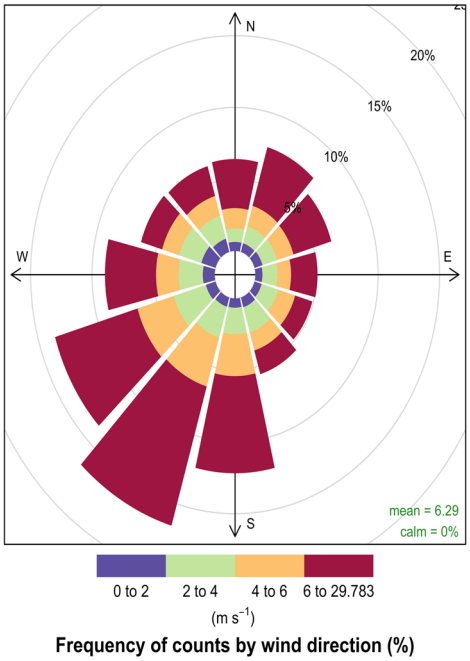


Figure 36. Wind rose for Stornoway (reference) predicted data for training year 2007, window length $M_w = 24$h, truncation $M_t = 12$.

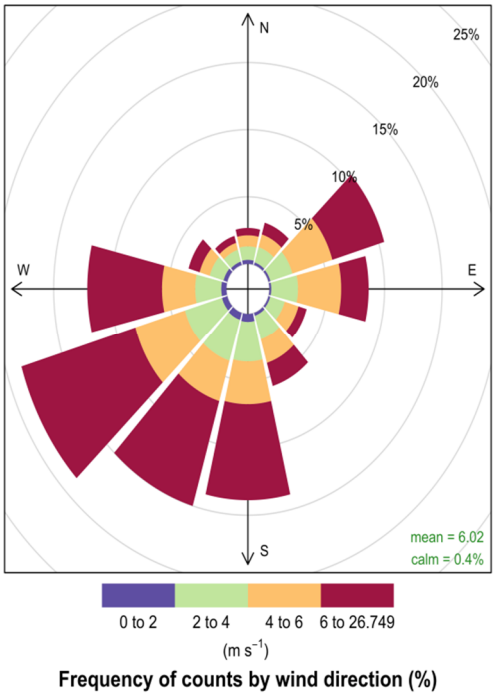


Figure 37. Wind rose of Salsburgh (target) actual data for training year 2007, window length $M_w = 24$h, truncation $M_t = 12$.

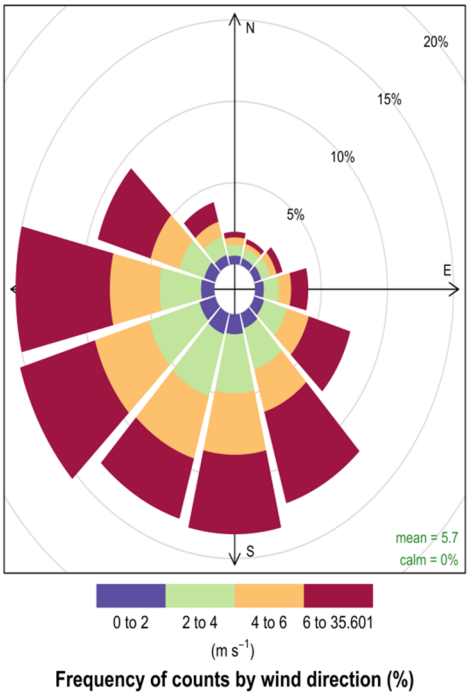


Figure 38. Wind rose of Salsburgh (target) predicted data for training year 2007, window length $M_w = 24$h, truncation $M_t = 12$.

6.6.2 A 'bad' example

Using Blackford Hill as reference and Machrihanish as target site for 2003 with truncation $M_t = 9$ and window length $M_w = 48$h we obtain the following PCA-MCP results. It can be seen from Figure 39 which depicts the singular values spectrum that the lambda values have a rapid cut off around 30. However, the split between the singular values sections is not so clear as it was in Figure 32. It can be concluded that $M_t = 9$ would be a suitable choice for $M_w = 24$h in the previous 'good' case but choosing it for this case of $M_w = 48$ seems to have resulted in the loss of important variance. The choice of another truncation other than 9 would thus be the most suitable one to test the bad performance of the PCA-MCP method.

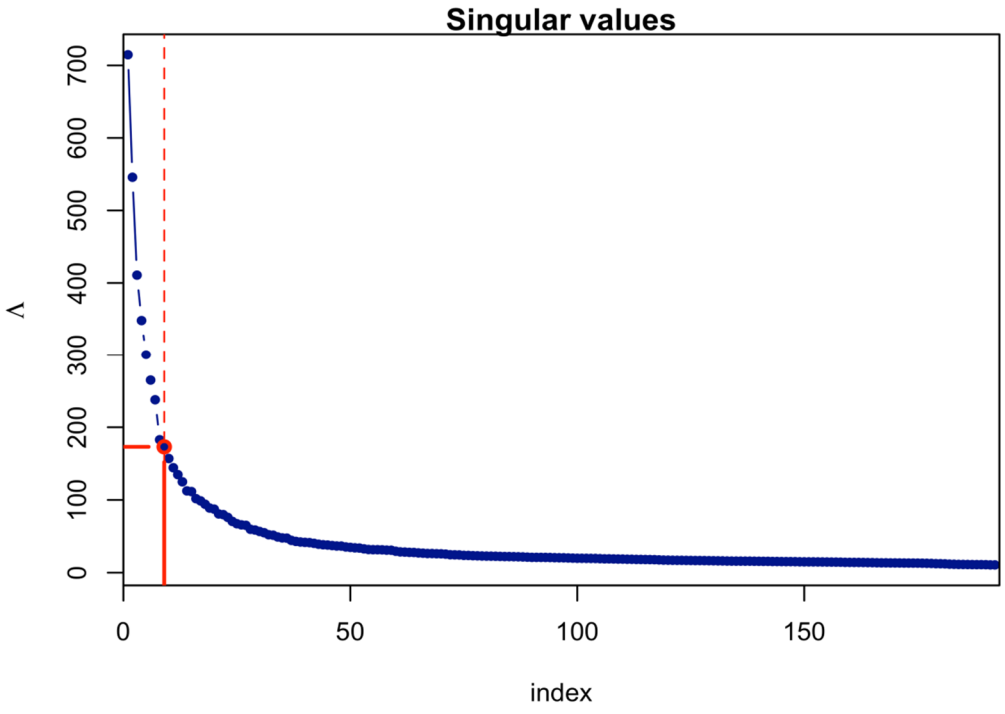


Figure 39. Singular values spectrum for Blackford Hill and Machrihanish stations for the training year 2003, window length $M_w = 48$h, truncation $M_t = 9$.

Next, the wind speed histograms in Figure 40 and Figure 41 are shown. In this case, it can be observed that the predicted wind speeds with the actual wind speeds for both the reference and target sites have some differences, especially for the target site. In

Figure 40, the predicted wind speed seems to be underpredicting the wind speeds for the very low wind speeds (< 1m/s) and overpredicting for the low wind speeds (< 5m/s). Figure 41 indicates an overprediction of the target wind speeds to begin with (< 3m/s) followed by a very big underprediction for wind speeds ranging from 3m/s up to 11m/s and then followed again by an overprediction for the large wind speeds. Thus, a very big deviation in the target site prediction and actual wind speed data can be observed which shows a poor PCA-MCP performance. This could be due to an old anemometer, or the Blackford Hill anemometer being of poor response due to false instrumentation. However this big deviation didn't show in Figure 40 because it predicted well internally within the station but when combined with the Machrihanish site which has better quality data , this error is apparent. Also comparing the actual data of the two sites wind roses of Figure 42 and Figure 44, they are very different.

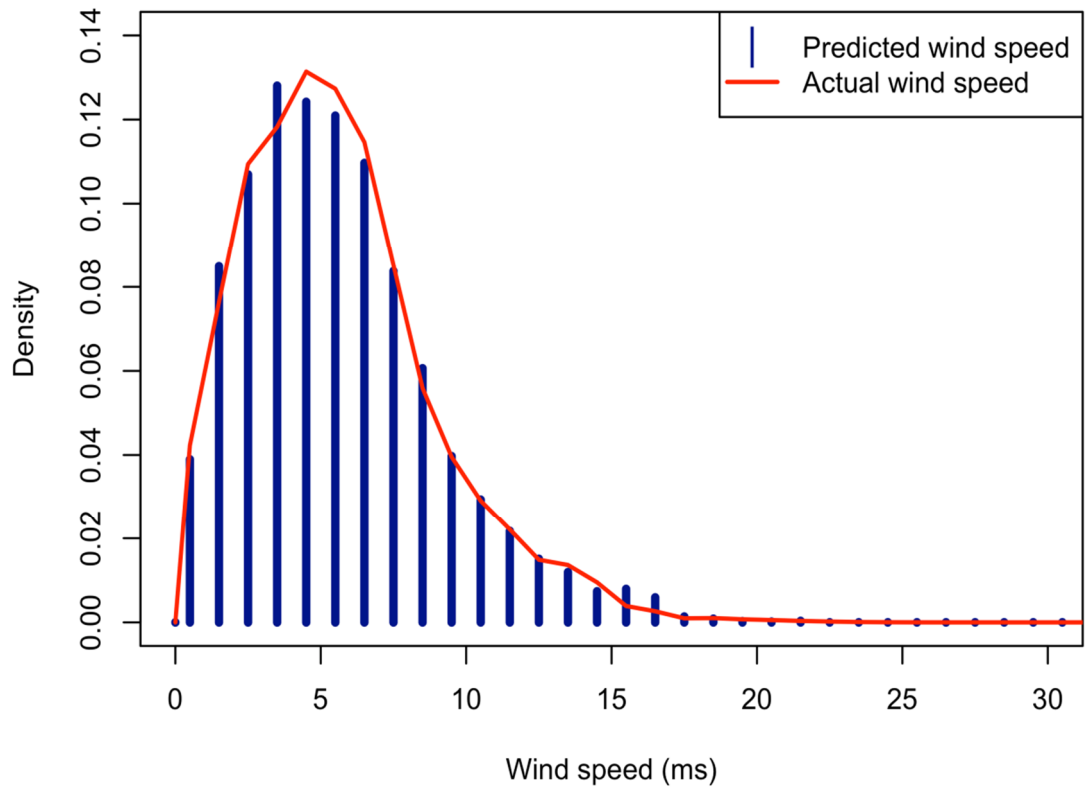


Figure 40. Actual and predicted wind speed for Blackford Hill (reference) for training year 2003, window length $M_w = 48$h, truncation $M_t = 9$.

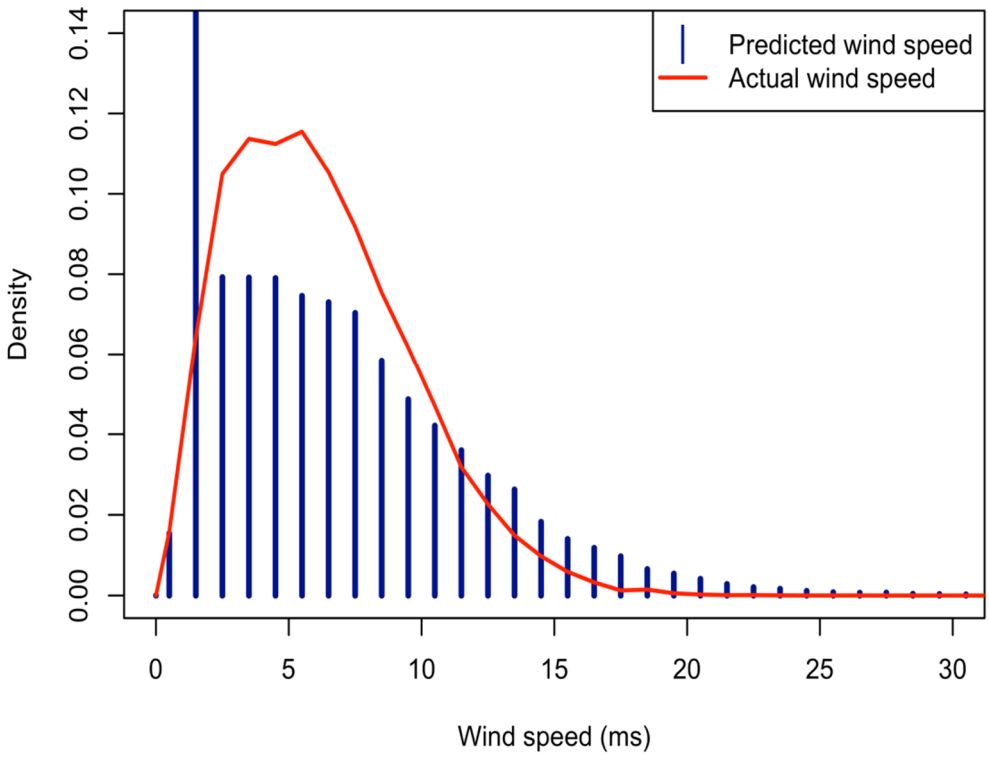


Figure 41. Actual and predicted wind speed for Machrihanish (target) for training year 2003, window length $M_w = 48$h, truncation $M_t = 9$.

The wind roses of the reference data, Figure 42 and Figure 43 are similar with very little differences as described in the histogram of Figure 40. However, Figure 44 and Figure 45 regarding the actual and predicted target site data indicate a lot of differences which verify the under and over predictions in the target predicted wind speed as described in the histogram of Figure 41. The prevailing winds in Figure 45 seem to be coming from the southwest whereas in the actual target data shown in Figure 44 come from southeast. This example was one of a poor performance for the PCA-MCP method because the actual and predicted data especially for the target site deviated to a big extent. More measures to validate the method's performance will be shown in the following sections of this chapter.

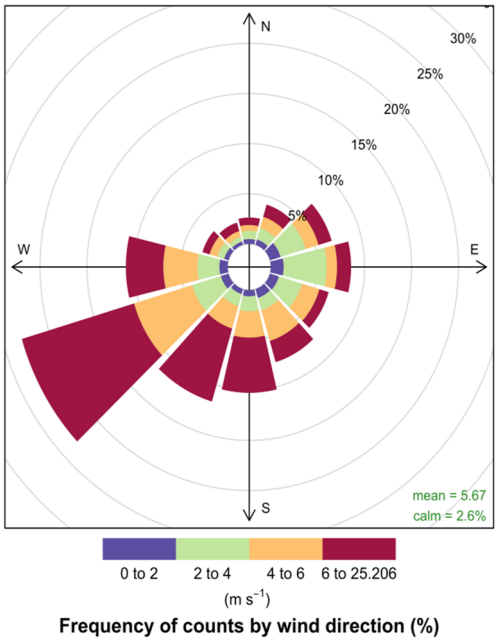


Figure 42. Wind rose for Blackford Hill (reference) actual data for training year 2003, window length $M_w = 48$h, truncation $M_t = 9$.

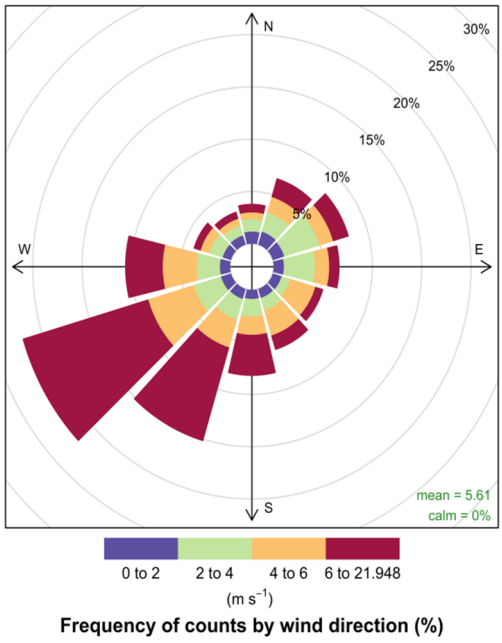


Figure 43. Wind rose for Blackford Hill (reference) predicted data for training year 2003, window length $M_w = 48$h, truncation $M_t = 9$.

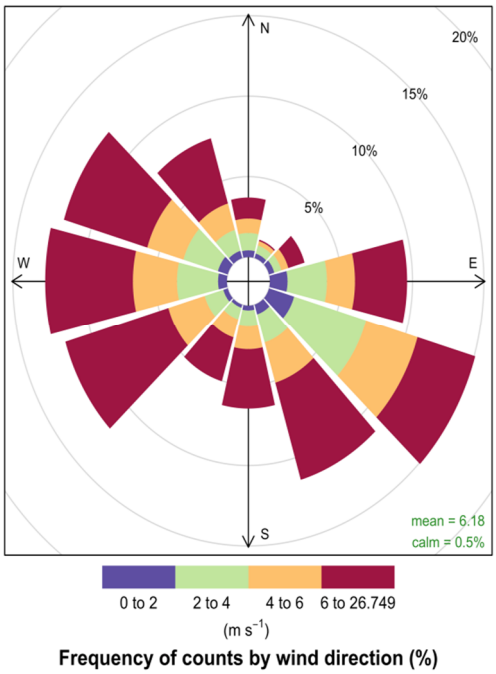


Figure 44. Wind rose of Machrihanish (target) actual data for training year 2003, window length $M_w =$ 48h, truncation $M_t = 9$.

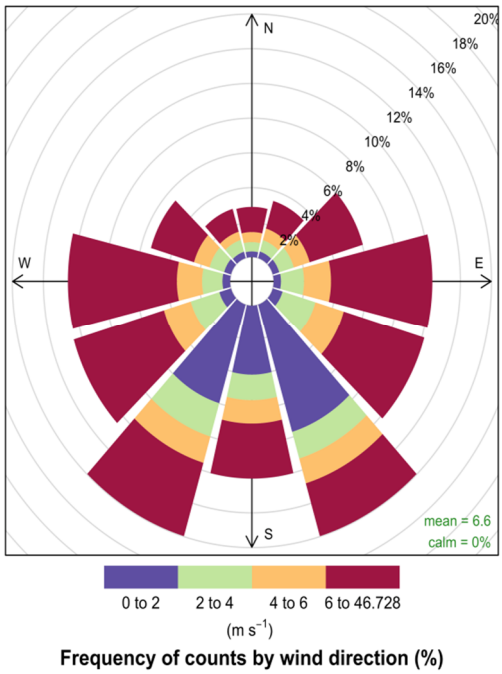


Figure 45. Wind rose of Machrihanish (target) predicted data for training year 2003, window length $M_w = 48$h, truncation $M_t = 9$.

6.6.3 Overall PCA-MCP performance and evaluation

The following graphs depict the measures used in the previous section 6.5.1 validating the overall performance of PCA-MCP as a method alongside with a comparison with linear regression.

First, the graphs of the MAE_{ref}, MAE_{tar} and MAE_{lr} from equations (58), (59) and (60) against all 8 reference stations for each of the 8 target sites are depicted from Figure 46 to Figure 57. As it can be seen from the first four graphs Figure 46 to Figure 49, MAE_{lr} is ranging from 0.7 to 1.2 thus is the highest when compared to MAE_{tar} and MAE_{ref} which are relatively low, i.e. up to 0.4.

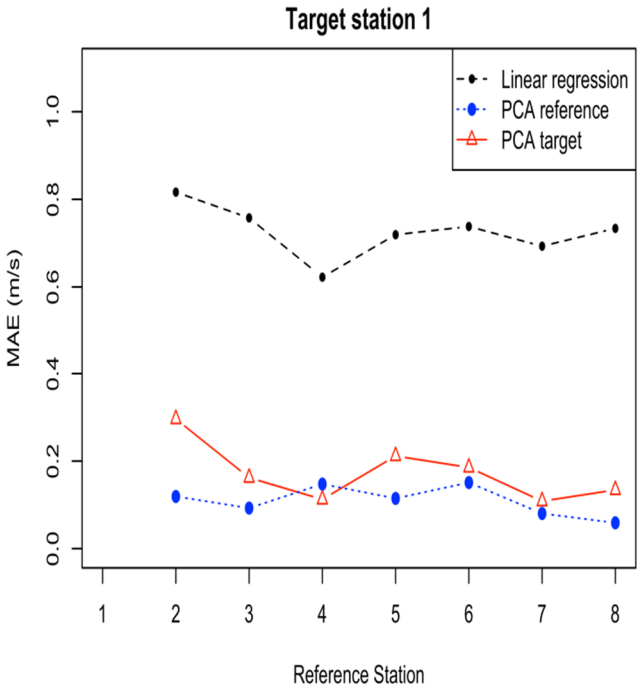


Figure 46. Stornoway MAE for all reference stations.

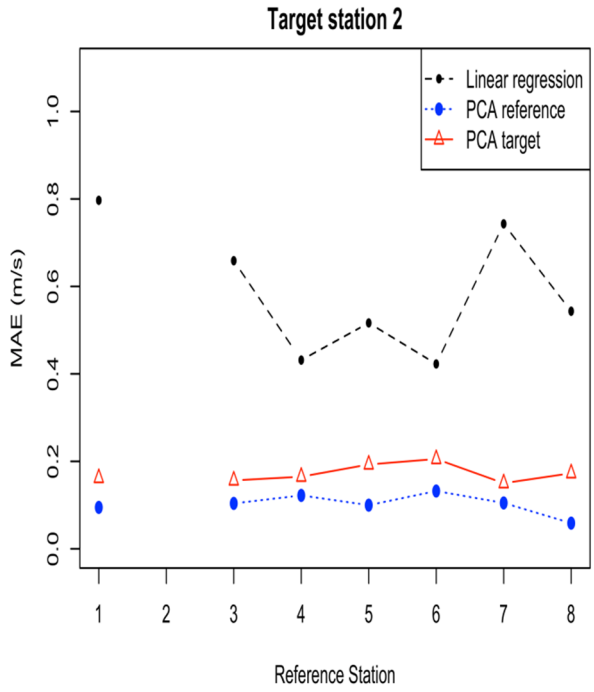


Figure 47. Blackford Hill MAE for all reference stations.

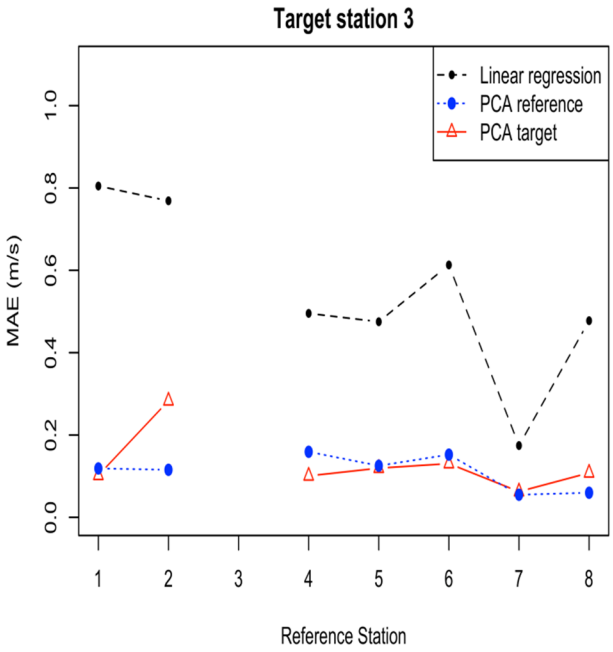


Figure 48. Machrihanish MAE for all reference stations.

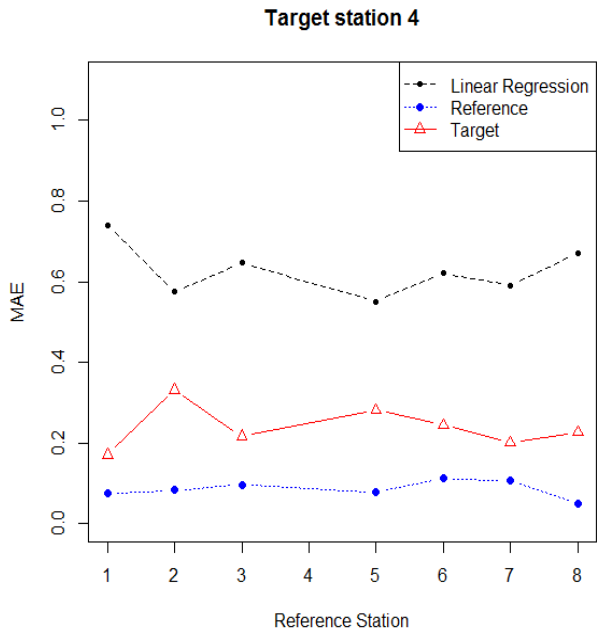


Figure 49. Salsburgh MAE for all reference stations.

The first 4 target stations bias graphs indicate in general the existence of negative bias, more specifically looking at Figure 50. The reason behind this could be that PCA-MPC method predicts slower wind speeds than simple linear regression does. Two reasons could be behind this; possibly the calibration used in the PCA-MCP analysis is not yet optimal i.e. using the mean and standard deviation rations as expressed in Chapter 3.6.2 equation (30) and equation (31). Secondly, the calibration was performed in order to minimise the distribution error i.e. calibrate so as to expect the smallest error in the wind speed distribution.

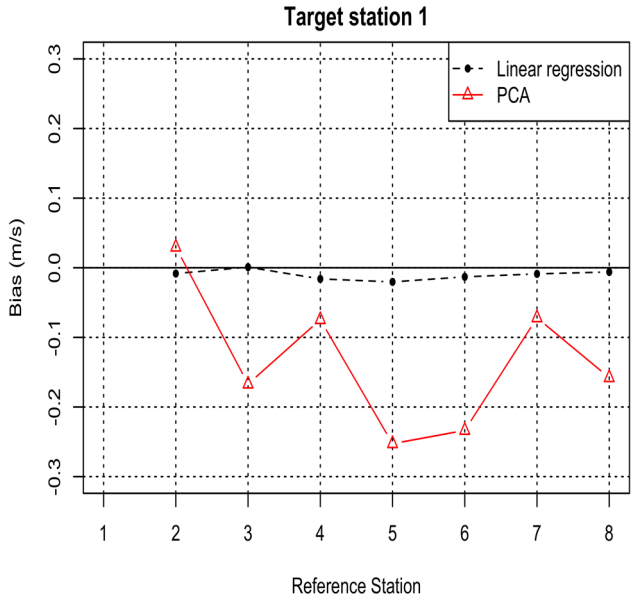


Figure 50. Stornoway Bias for all reference stations.

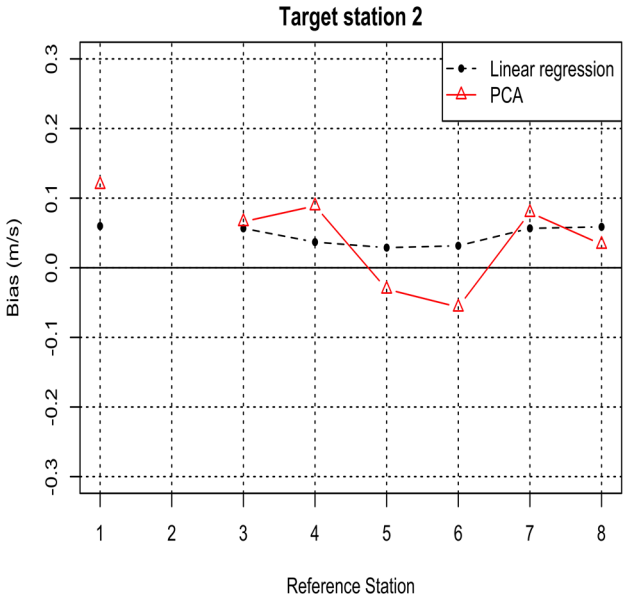


Figure 51. Blackford Hill Bias for all reference stations.

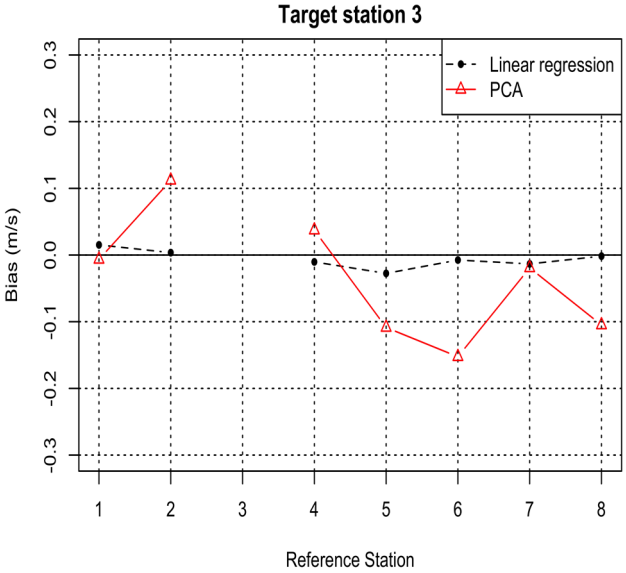


Figure 52. Machrihanish Bias for all reference stations.

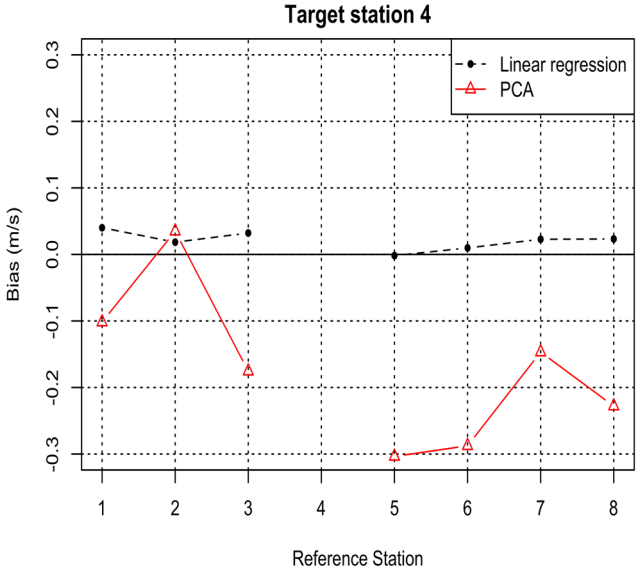


Figure 53. Salsburgh Bias for all reference stations.

The next four graphs, Figure 54 to Figure 57 follow the same pattern as in Figure 47 to Figure 49 i.e. the MAE_{lr} is higher when compared to MAE_{tar} and MAE_{ref} . More specifically, the MAE_{lr} is ranging from 0.7 to 1 and it can be observed that it is the highest for Stornoway (Figure 46), Gogarbank (Figure 55) and Bishopton (Figure 57). Examining the MAE_{tar} , it ranges between 0.1 and 0.3 and the lowest one, MAE_{ref} ranges from 0.1 to 0.2. It can be thus

concluded that linear regression performs worse than PCA-MCP since MAE_{lr} has the highest values, when examining all target stations with respect to all reference stations.

PCA-MCP seems robust since the MAE graphs are fairly flat for most of the stations in comparison with simple linear regression. This suggests that a good performance of the standard linear regression MCP relies strongly on having chosen a good reference site (which may not always be obvious in advance or even possible), whereas the PCA-MCP method is fairly insensitive to a particular choice of reference site. Being able to calculate the MAE_{ref} as part of the PCA-MCP prediction also provides a tool to estimate the actual MAE_{tar}.

From the last four bias graphs for stations 5-8, it can be seen that the bias is above zero in most cases. Stations 5,6 and 8 are low wind speed stations whereas stations 1,2,3,4,7 are high wind speed stations. This indicates that somehow bias is related with whether the prediction comes from a low or high wind speed site, i.e. bias seems to be correlated with the predicted wind speed. Linear regression as a method is unbiased by default since the linear estimation tries to minimise bias and hence it is only logical to result in minimal bias. In the PCA-MCP case, it was chosen to minimise the distribution error rather than the bias thus the bias graphs are a flipside of MAE graphs. In general, for wind resource purposes it is preferable to calibrate more correctly wind speed distribution rather than bias.

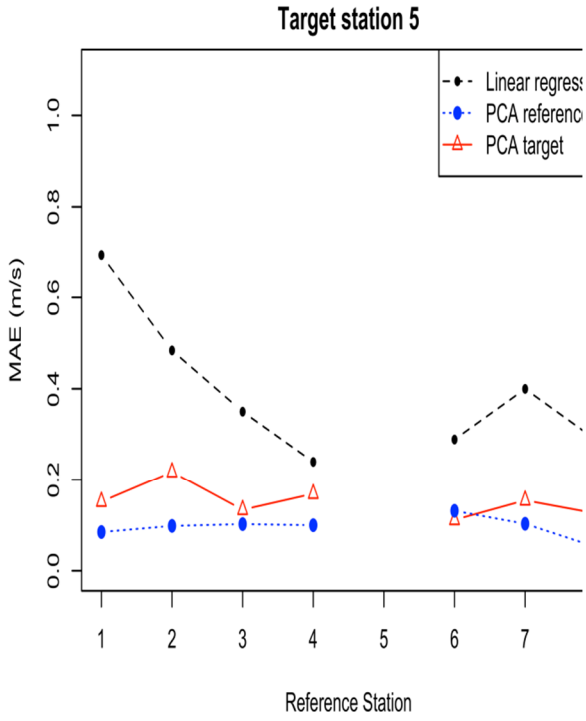


Figure 54. Prestwick Gannet MAE for all reference stations.

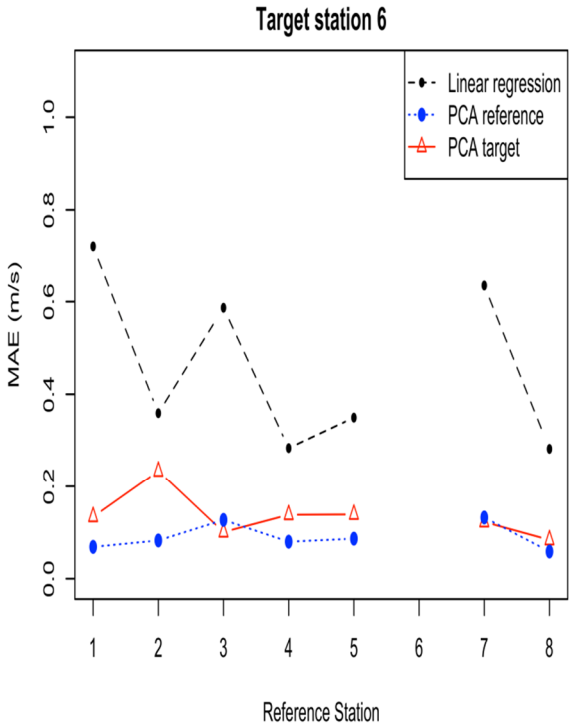


Figure 55. Gogarbank MAE for all reference stations.

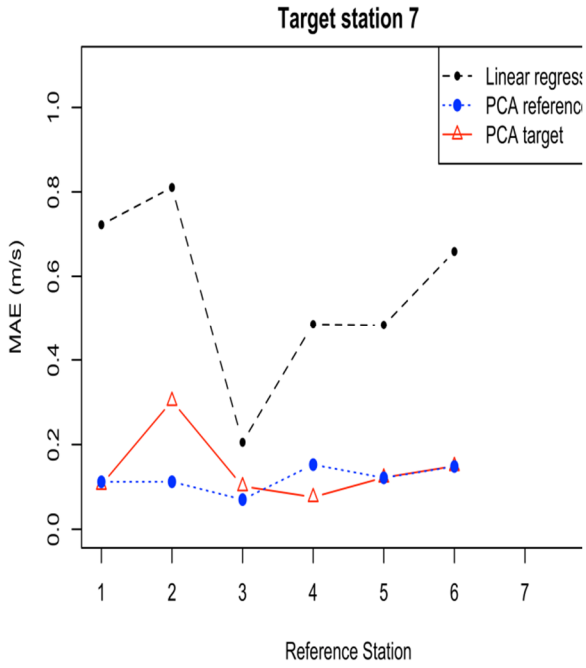


Figure 56. Port Ellen MAE for all reference stations.

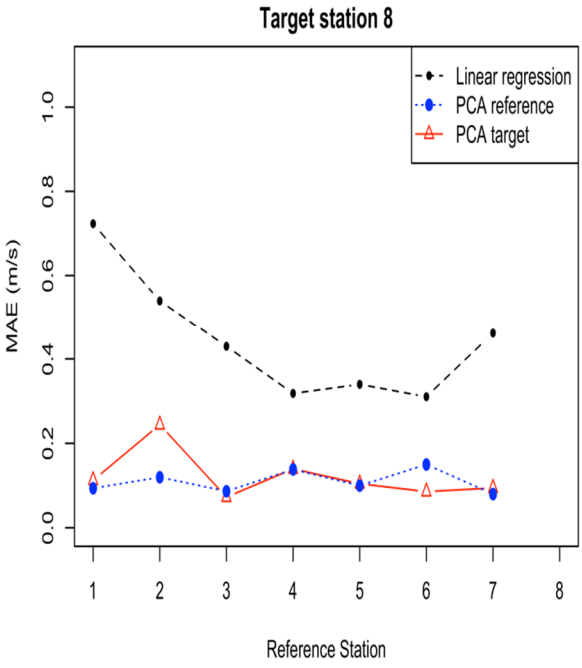


Figure 57. Bishopton MAE for all reference stations.

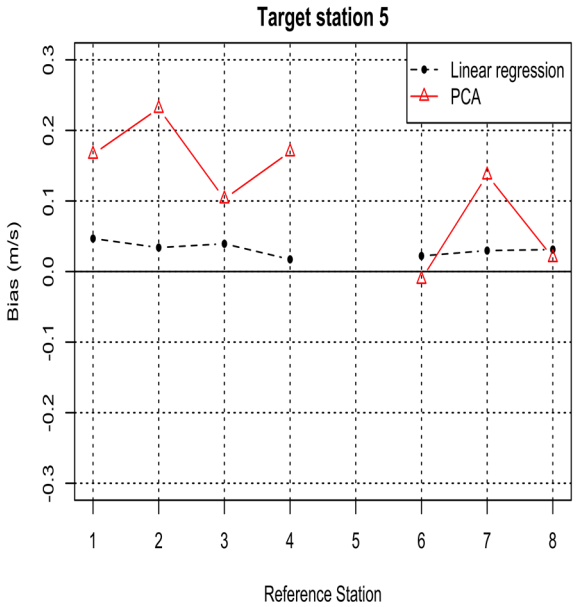


Figure 58. Prestwick Gannet Bias for all reference stations.

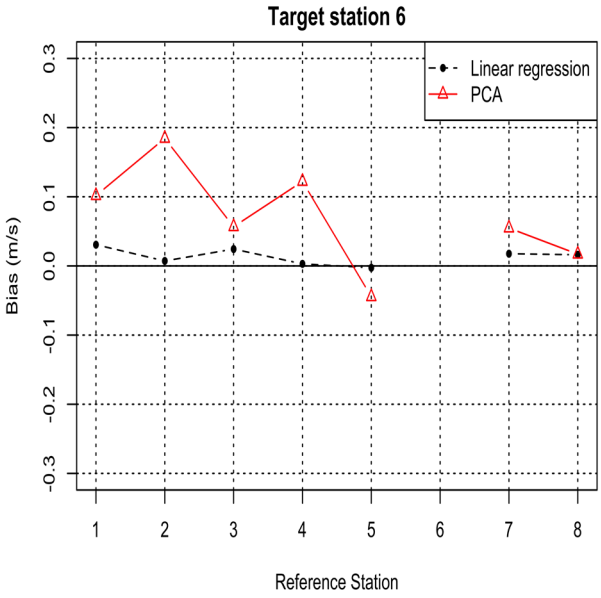


Figure 59. Gogarbank Bias for all reference stations.

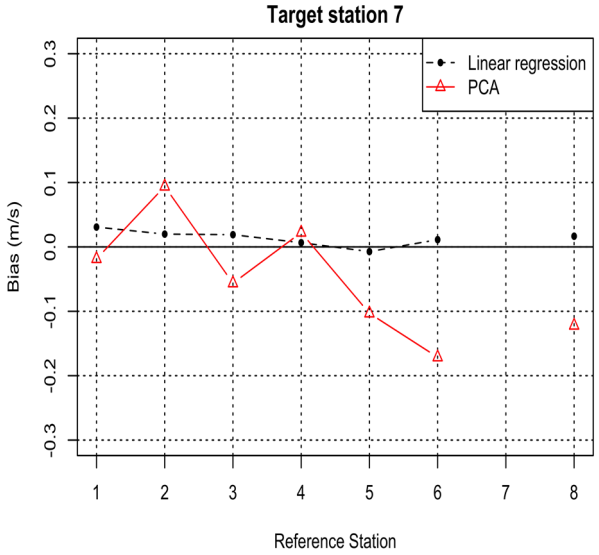


Figure 60. Port Ellen Bias for all reference stations.

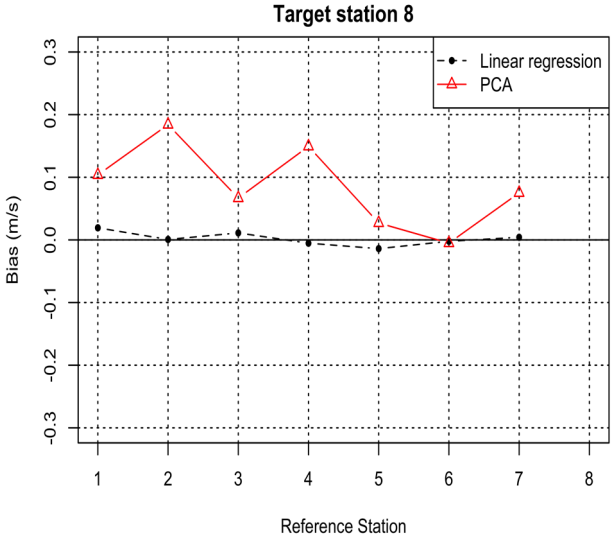


Figure 61. Bishopton Bias for all reference stations.

6.6.4 Evaluation of the 'good' and 'bad' PCA examples

Going back to the 'good' and 'bad' examples of sections 6.6.1 and 6.6.2, the following graphs were created to explore the MAE_{ref}, MAE_{tar} and MAE_{lr} for these specific cases for all training years. In Figure 62 it can be seen that for the 'good' example of Salsburgh as a target and for Stornoway (reference station 1) and for window length 24h and truncation 12

the MAE_{ref}, MAE_{tar} are quite low, up to 0.2, for all training years when compared to the other reference stations. However, the MAE_{lr} is quite high (almost 0.7). As it is shown from Figure 63, for Machrihanish as a target and Blackford Hill (reference station 2) for window length of 48h and truncation 9, the MAE_{ref}, MAE_{tar} for all years are higher when compared to the other reference stations and close to 0.6 i.e. a much higher value than the one in Figure 62. Again, the MAE_{lr} value is high (almost 0.8) though not the highest among all reference stations.

Hence, the model of Figure 63 verifies the 'bad' example of PCA-MCP performance in comparison with the use of the other stations as reference ones and as an overall MAE value when compared with the 'good' example model of Figure 62. From both Figure 62 and Figure 63 it is observed that reference station 2 i.e. Blackford Hill seems to contain the biggest MAE_{ref}, MAE_{tar} when compared to the other reference stations though this is the case for these specific models. It can be also seen that the values of MAE_{ref} are the lowest and the values of MAE_{lr} are the highest amongst most training years and reference stations in both graphs.

Regarding Figure 64, it shows similar a behaviour as previous bias figures i.e. that the bias is consistently negative for PCA-MCP and it can be also seen that Figure 65 has a similar pattern when compared with Figure 52 both regarding station 3. It can be concluded that good choice of window and truncation can affect the bias can since similar spreads can be seen for bias values of both PCA-MCP and linear regression in Figure 65. Longer window lengths and smaller truncations as well as the calibration methods have to be explored more. By finding the optimum parameter settings, the minimisation of the distribution error and bias will be achieved.

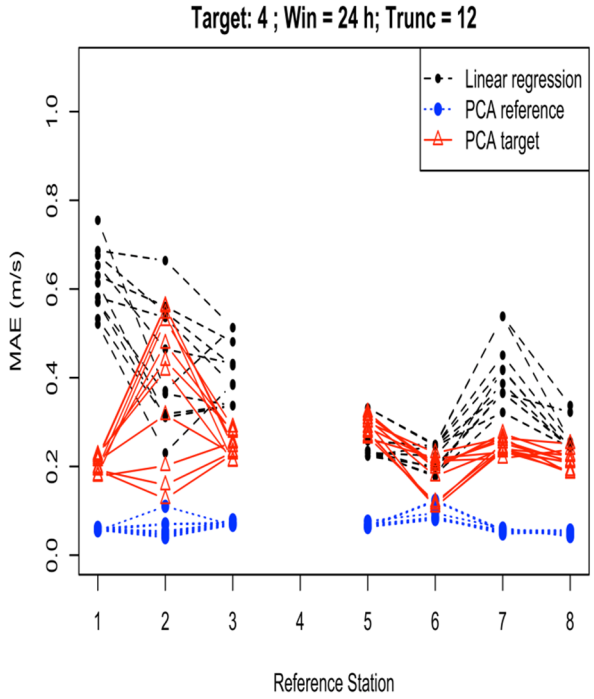


Figure 62. MAE of Salsburgh (target) for window length $M_w = 24$h, truncation $M_t = 12$, for all training years and reference stations.

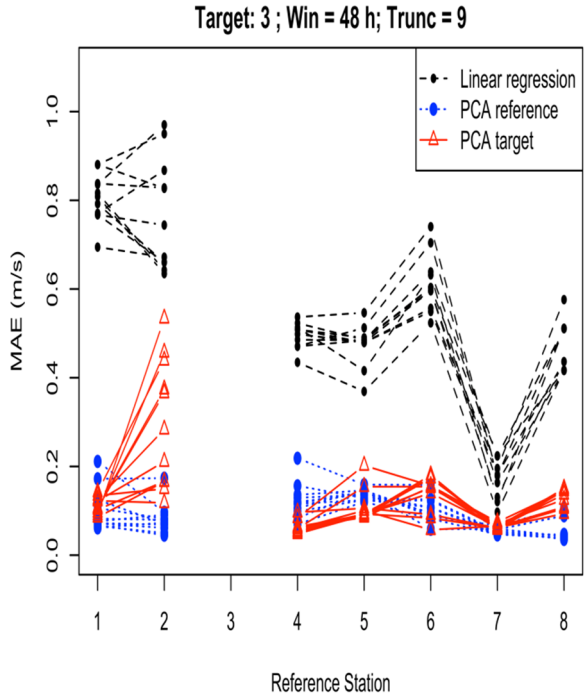


Figure 63. MAE of Machrihanish (target) for window length $M_w = 48$h, truncation $M_t = 9$, for all training years and reference stations.

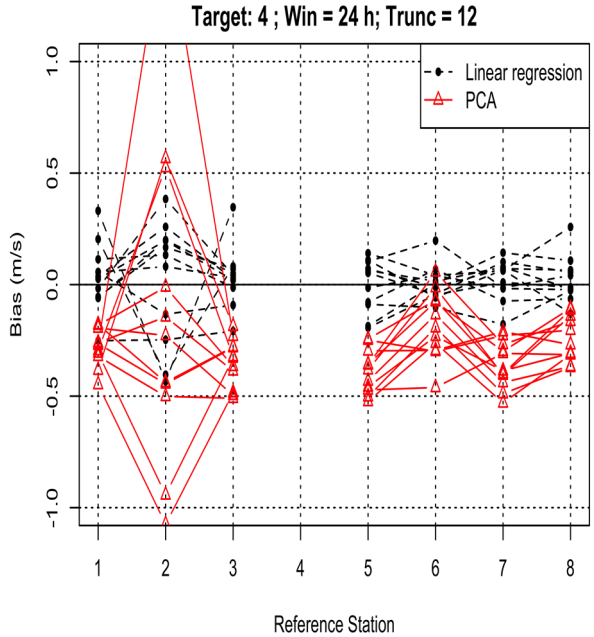


Figure 64. Bias of Salsburgh (target) for window length $M_w = 24$h, truncation $M_t = 12$, for all training years and reference stations.

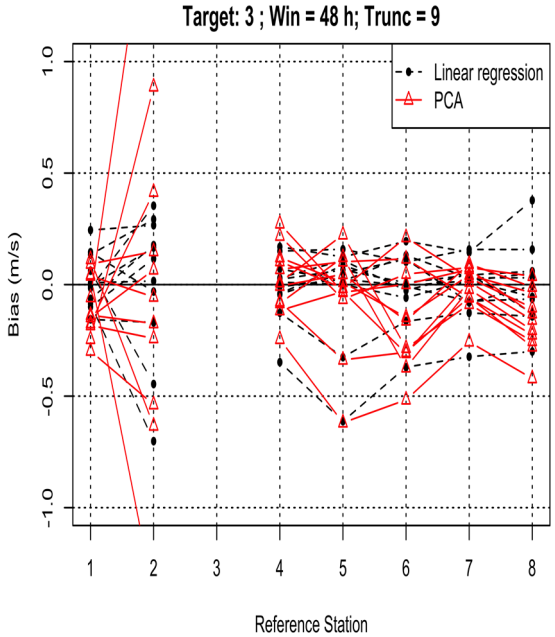


Figure 65. Bias of Machrihanish (target) for window length $M_w = 48$h, truncation $M_t = 9$, for all training years and reference stations.

Figure 66 and Figure 67 show the MAE_{ref}, MAE_{tar} and MAE_{lr} for the reference stations of the 'good' and 'bad' PCA-MCP performance models for all training years and target stations. The conclusions for the target site 4, Salsburgh and Stornoway as reference in Figure 66 and for target site 3, Machrihanish and Blackford Hill as reference in Figure 67 are the same as the aforementioned ones drawn from Figure 62 and Figure 63 regarding the justification of them as 'good' and 'bad' examples. Again, MAE_{lr} is high for almost all cases. Regarding Figure 69 the findings are similar to the Figure 65 conclusions described above.

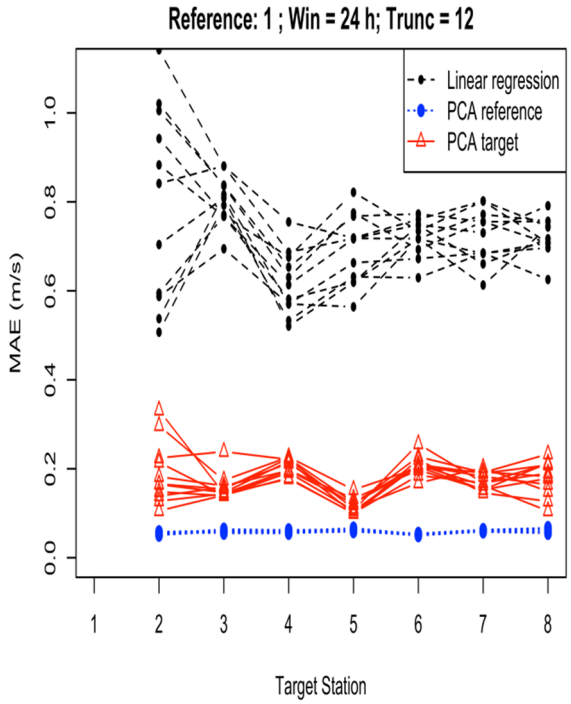


Figure 66. MAE of Stornoway (reference) for window length $M_w = 24$h, truncation $M_t = 12$, for all training years and target stations.

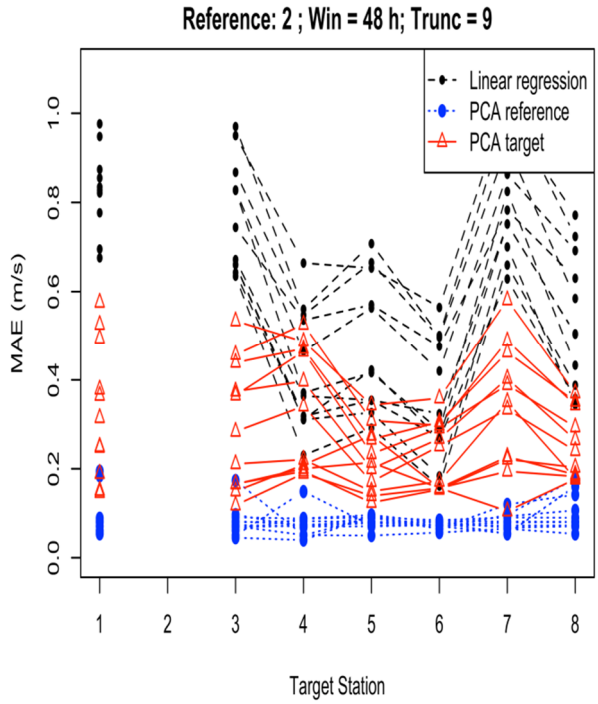


Figure 67. MAE of Blackford Hill (reference) for window length $M_w = 48$h, truncation $M_t = 9$, for all training years and target stations.

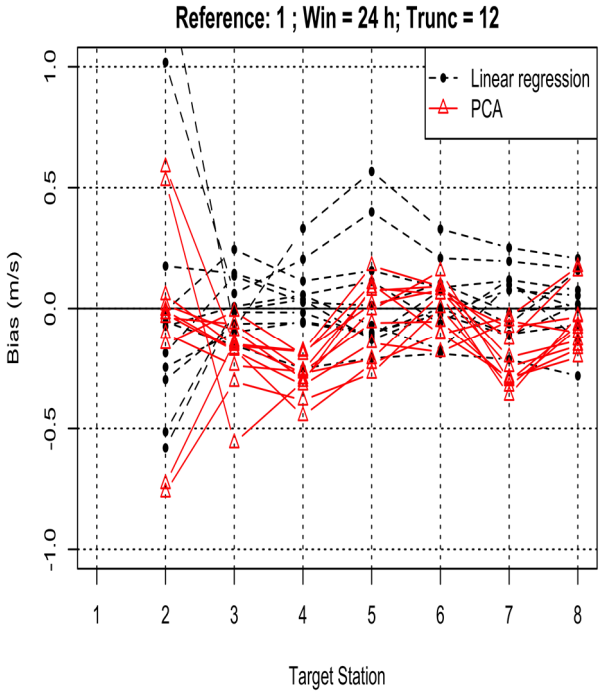


Figure 68. Bias of Stornoway (reference) for window length $M_w = 24$h, truncation $M_t = 12$, for all training years and target stations.

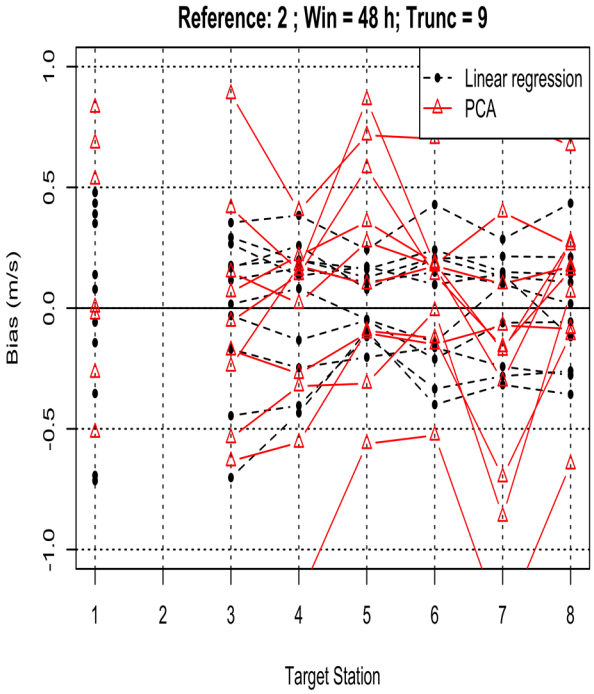


Figure 69. Bias of Blackford Hill (reference) for window length $M_w = 48$h, truncation $M_t = 9$, for all training years and target stations.

Finally, Figure 70 and Figure 71 indicate both the 'good' and 'bad' models of Chapter 6.6.1 and 6.6.2 for the MAE_{ref}, MAE_{tar} and MAE_{lr} for all training years for all eight window length and truncation combinations. In general, Figure 70 and Figure 71 both depict high MAE_{lr} throughout all training years whereas the MAE_{ref}, MAE_{tar} are lower. Regarding Figure 70, both MAE_{ref}, MAE_{tar} values are ranging below 0.2 for most years and looking specifically at 2007 which was examined in section 6.6.1 it is relatively low (below 0.2). On the contrary, in Figure 71 the MAE_{ref}, MAE_{tar} values generally range more and are higher. Looking specifically at 2003 which was examined in section 6.6.2, the MAE_{tar} is low, up to 0.4. Overall Figure 71 clearly indicates the poorer performance of PCA-MCP in comparison with Figure 70 throughout most training years.

In Figure 72 the general trend of the bias is similar for both PCA-MCP and linear regression. Positive bias is occurring in early years from 2000 to 2003 and negative bias

in the years after 2003. This also holds for Figure 73, however taking into account that station 2 has bad quality data which explains the 2003 spike.

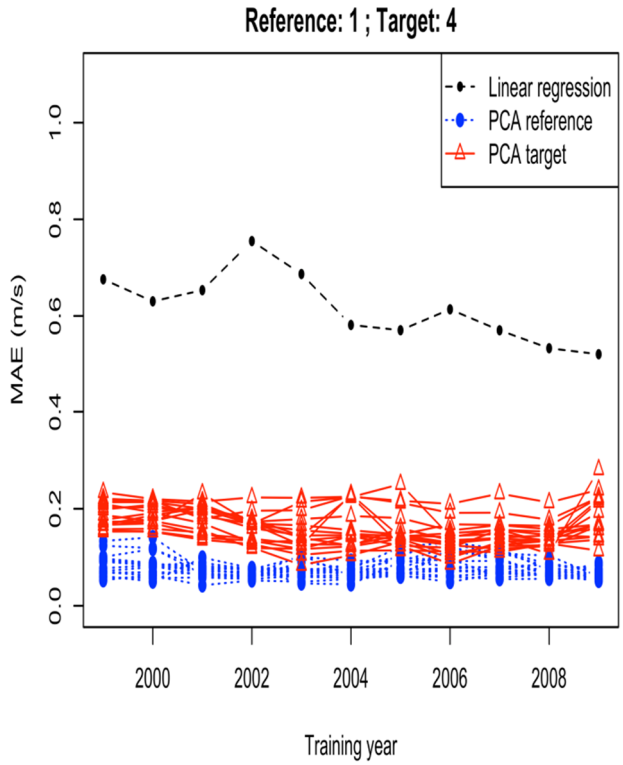


Figure 70. MAE of Stornoway (reference) and Salsburgh (target) for all training years and parameter combinations i.e. window lengths $M_w = 24\text{h}, 48\text{h}$, truncations $M_t = 3,6,9,12$.

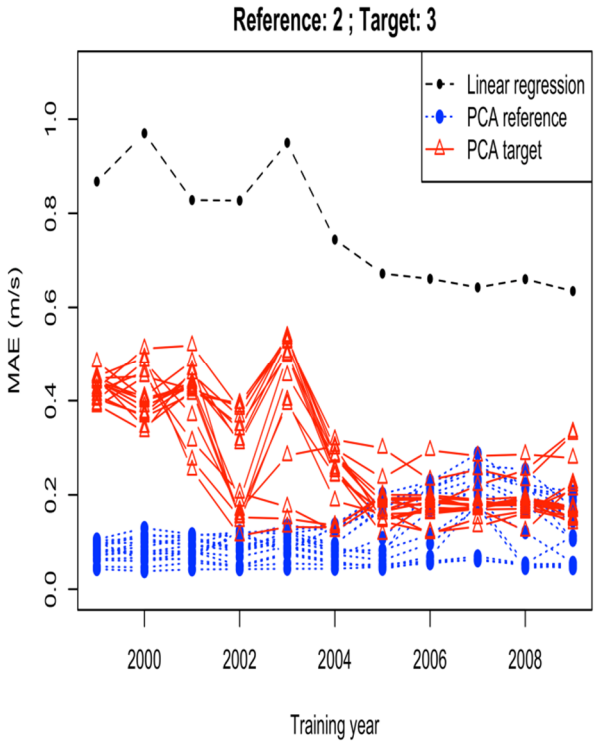


Figure 71. MAE of Blackford Hill (reference) and Machrihanish (target) for all training years and parameter combinations i.e. window lengths $M_w = 24\text{h}, 48\text{h}$, truncations $M_t = 3,6,9,12$.

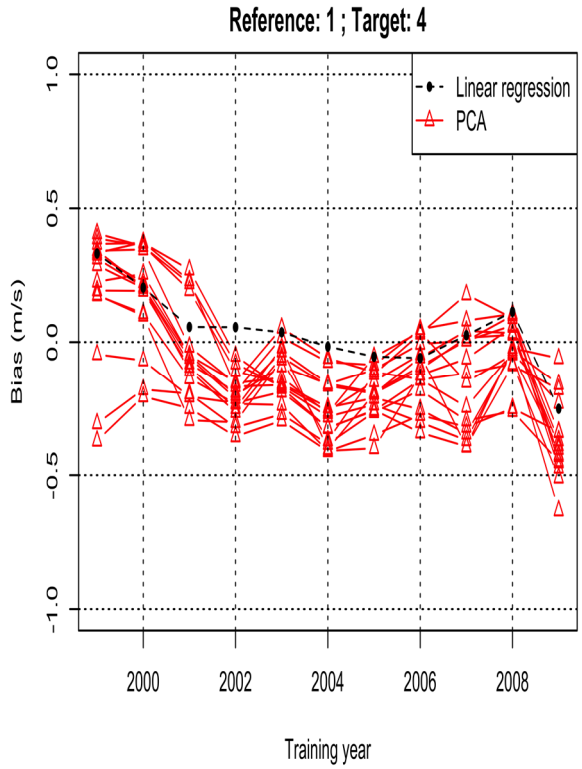


Figure 72. Bias of Stornoway (reference) and Salsburgh (target) for all training years and parameter combinations i.e. window lengths $M_w = 24\text{h}, 48\text{h}$, truncations $M_t = 3,6,9,12$.

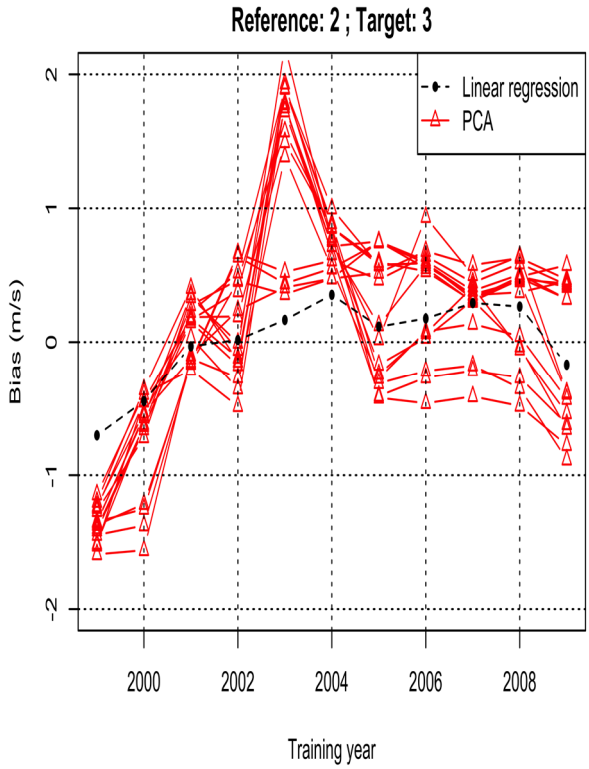


Figure 73. Bias of Blackford Hill (reference) and Machrihanish (target) for all training years and parameter combinations i.e. window lengths $M_w = 24\text{h}, 48\text{h}$, truncations $M_t = 3,6,9,12$.

When using as training the earlier years would support the earlier suggestion that the Blackford Hill performed poorly. In general, the inclusion of the year 2010 did not affect the PCA-MCP performance since the MAE was kept relatively low for 2010 and at a similar level when compared to the other training years.

6.6.5 Further PCA-MCP validation

Next, the parameter analysis setup measures of Table 10 are examined. Figure 74a) shows the MAE_{tar} for the two different window lengths $M_w = 24\text{h}$ and $M_w = 48\text{h}$. It indicates most of the MAE_{tar} values are concentrated up to 0.3 which shows that the MAE_{tar} for both window length choices was low however they do not lie along the line i.e. there is some scatter. Their choice, as mentioned in section 6.3.2, was undertaken

after extensive parameter testing on the PCA-MCP algorithm prior to choosing these two window length values. As shown in Figure 74a) since most observations lie below the diagonal line, M_w = 48h was the best window length choice. Figure 74b) also indicates that the two window lengths, M_w = 24h, 48h are highly correlated.

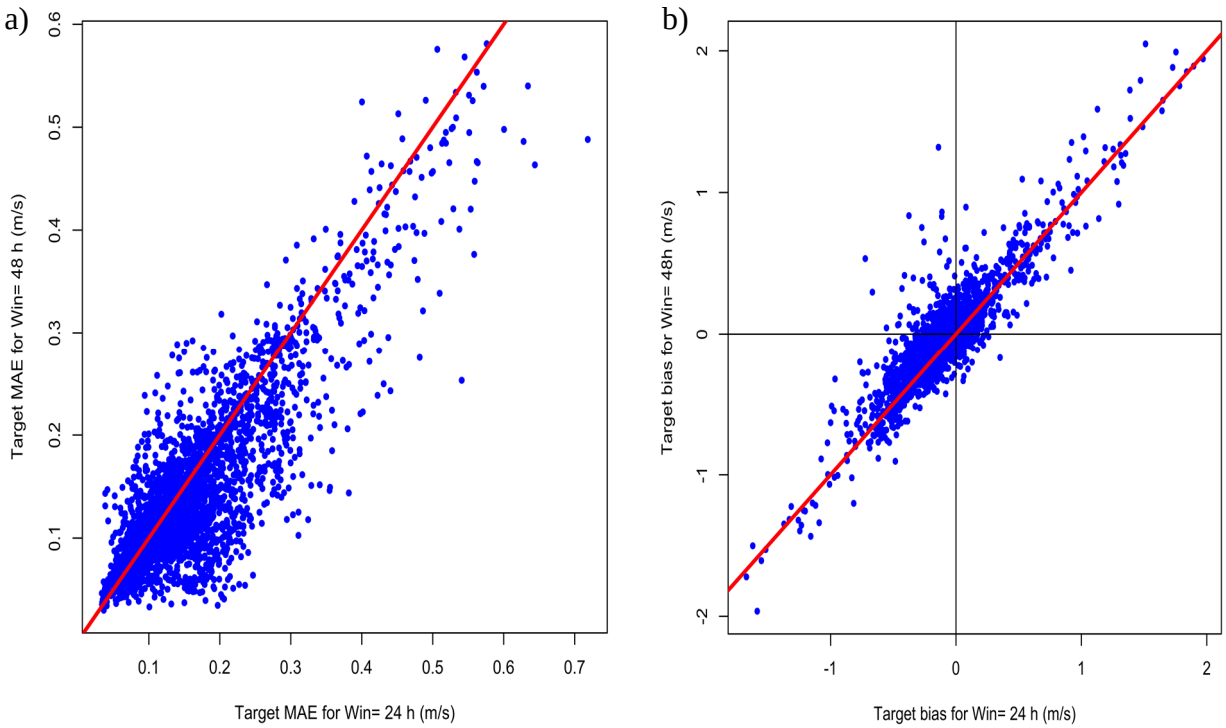


Figure 74. Target MAE and target Bias for window length of 24h and 48h

The next figure, Figure 75 shows the MAE_{tar} for the four different truncations combination M_t = 3,6,9,12. Similarly to the window length choice as mentioned in section 6.3.2, their choice was concluded too after extensive parameter testing on the PCA-MCP algorithm. Most MAE_{tar} values lie below 0.3 which shows that the MAE_{tar} for all truncation combinations was low. There is more scatter in the MAE_{tar} of the truncation choices M_r = 3,6 since the MAE values do not lie close to the diagonal. On the contrary, it can be seen that for truncations M_t = 9,12 the MAE_{tar} values lie closer to the diagonal and the same holds for truncations M_t = 6,9. Hence, the choice of higher truncations seems to be the most appropriate for the PCA-MCP performance.

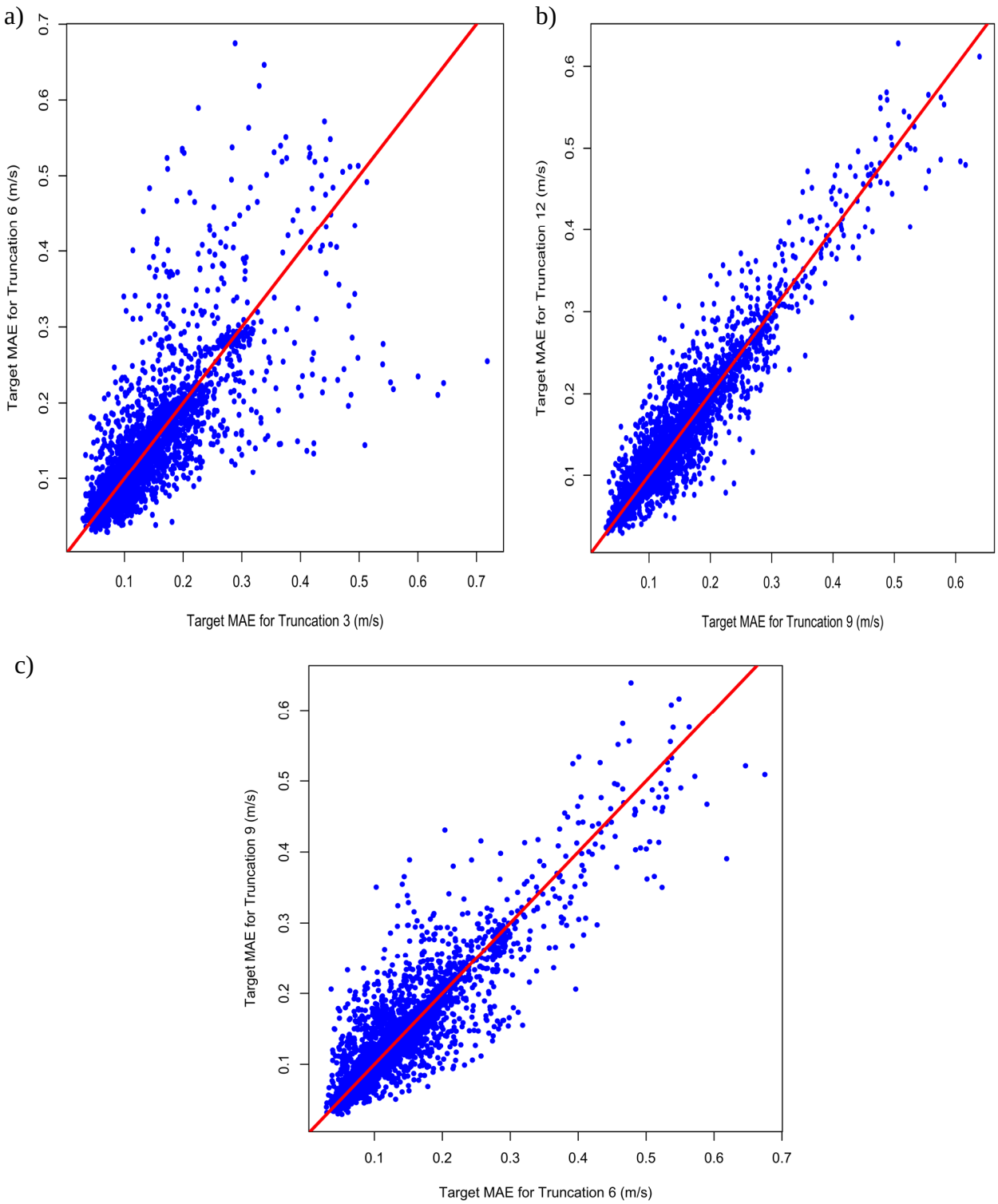


Figure 75. Target MAE for all truncation combinations.

Figure 76 shows that the target bias for all truncation combinations has a similar shape to MAE ones. It can be drawn that as the truncation gets higher the bias becomes more variable for smaller truncations.

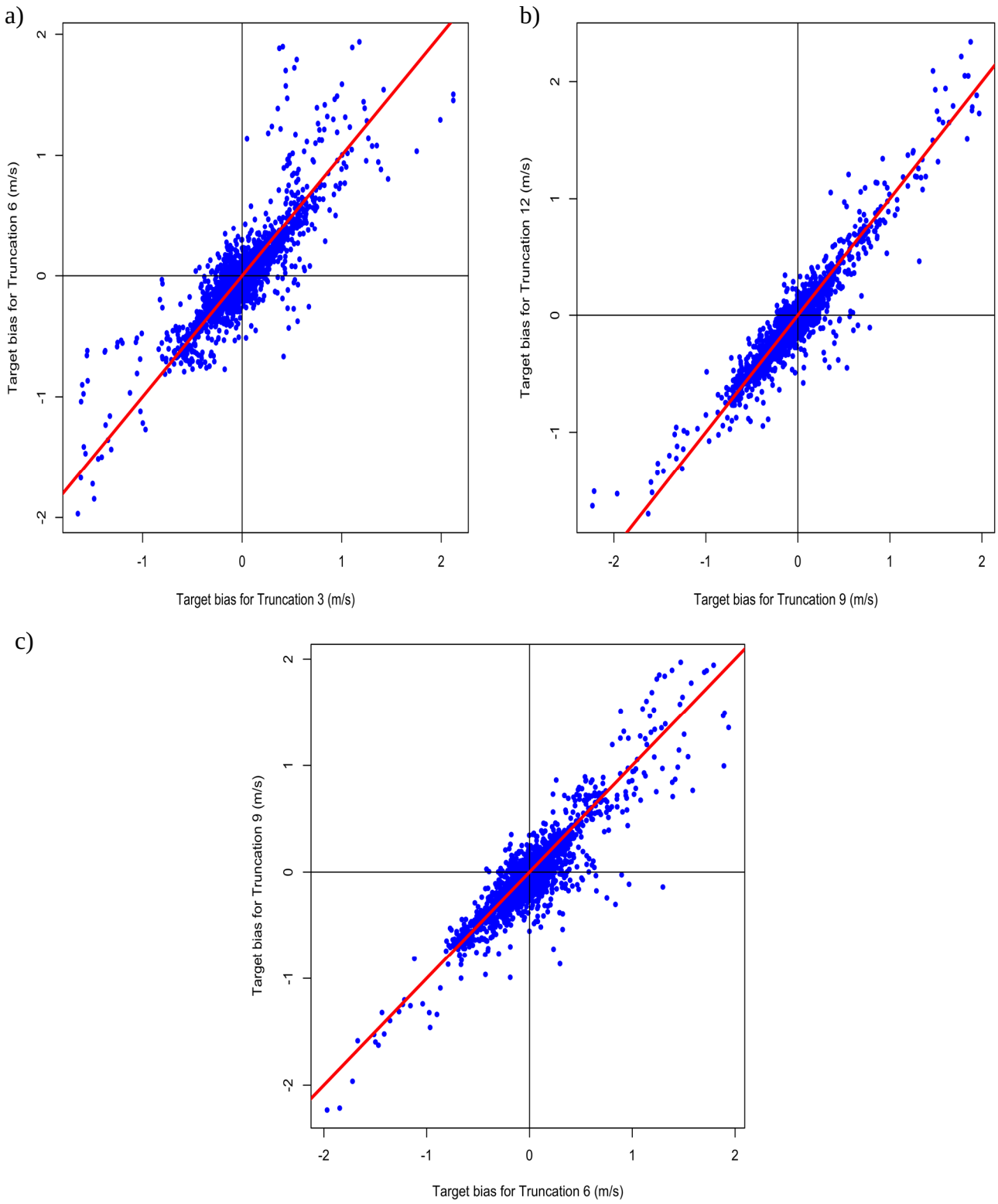


Figure 76. Target Bias for all truncation combinations.

Next, the MAE_{tar} versus the MAE_{ref} is presented in Figure 77a) and the MAE_{tar} versus MAE_{lr} in Figure 77b). From Figure 77a) we can see that there is no proportionality between MAE_{tar} and MAE_{ref} hence in this case it is not clear that knowing the MAE_{tar}, it quantifies us the predictability and thus, a different way to predict uncertainty should be investigated. In Figure 77b), since almost all observations lie below the line, this verifies the overall overperformance of the PCA-MCP method against linear regression with a few cases being above the line which indicate a poorer PCA-MCP performance.

From Figure 77c) it is depicted that in general linear regression has smaller bias than PCA-MCP because it can be seen from the left side of graphs that more data lie below the red line whereas for the right side of the graph more date lie above red line and thus more data clash towards the zero linear regression bias line. Hence, the earlier observations from previous graphs can be verified.

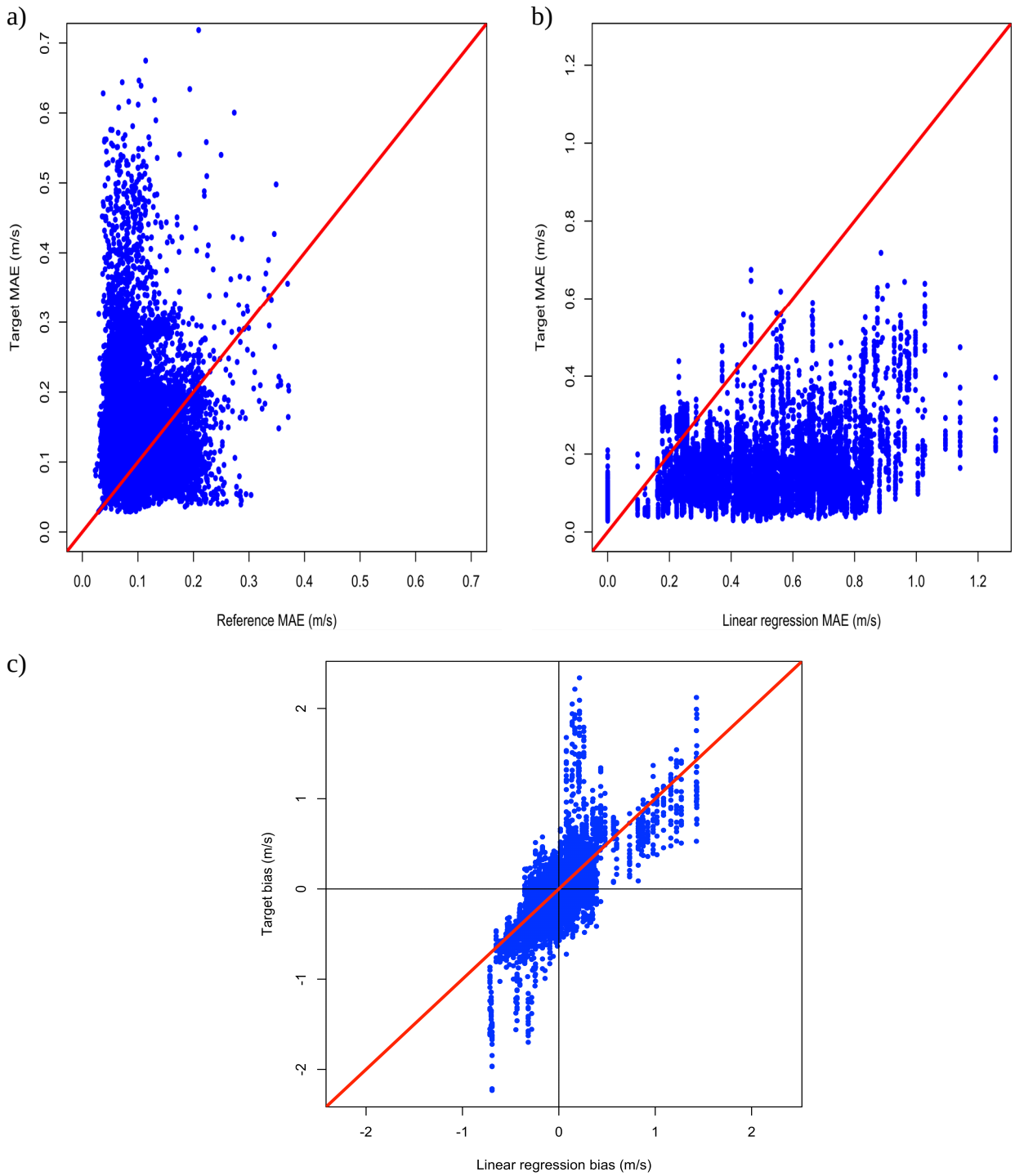


Figure 77. Target MAE and linear regression MAE versus reference MAE. Target Bias versus linear regession Bias

As far as the performance index (PI) shown in equation (63) is concerned, the following graphs were created. Figure 78 shows a histogram of the PI for all possible permutations of pairs of the 8 Met.Office stations. As it can be seen, the PI is below 1

in almost all cases, and most often, it is between 0.1 and 0.5. This means that in the majority of cases, the error of the resource prediction is between 10% and 50% of that made using the standard linear regression.

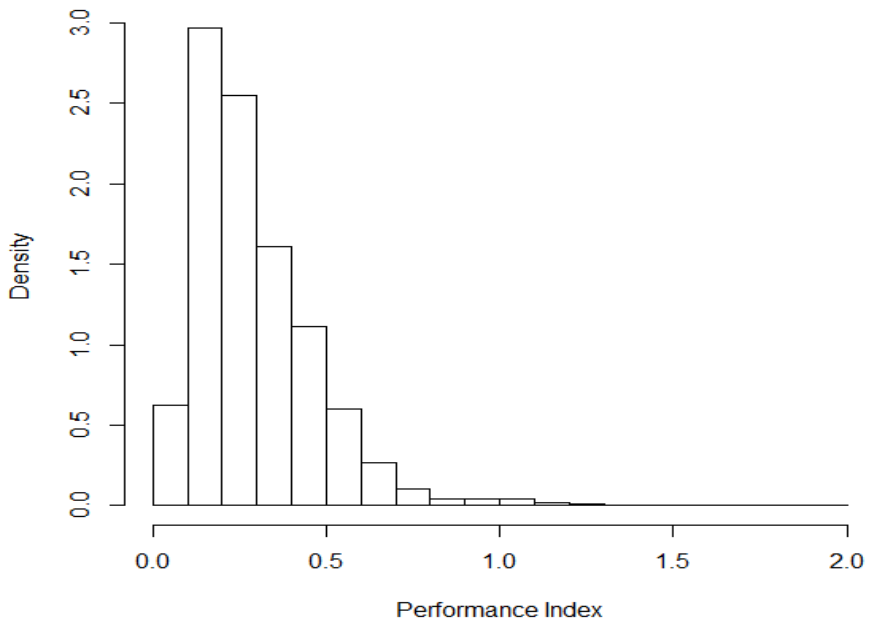


Figure 78. Performance Index histogram.

Figure 79 indicates the PI against Figure 79a) MAE_{ref}, Figure 79b) MAE_{tar} and Figure 79c) MAE_{lr}. The MAE_{ref}, MAE_{tar} range from 0.04 to 0.3 and from 0.05 to 0.7 respectively. However, MAE_{lr} ranges from 0.1 to 1.3 and hence, it is the highest. Consequently, this graph is another justification of the better performance of PCA-MCP when compared with linear regression. It is also observed that the PI is less than 1 for most cases of MAE_{tar} which means that PCA-MCP performs well. The general mean of the PI was also found to be 0.29, thus the overall MAE_{tar} was found to be only the 29% of the MAE_{lr}

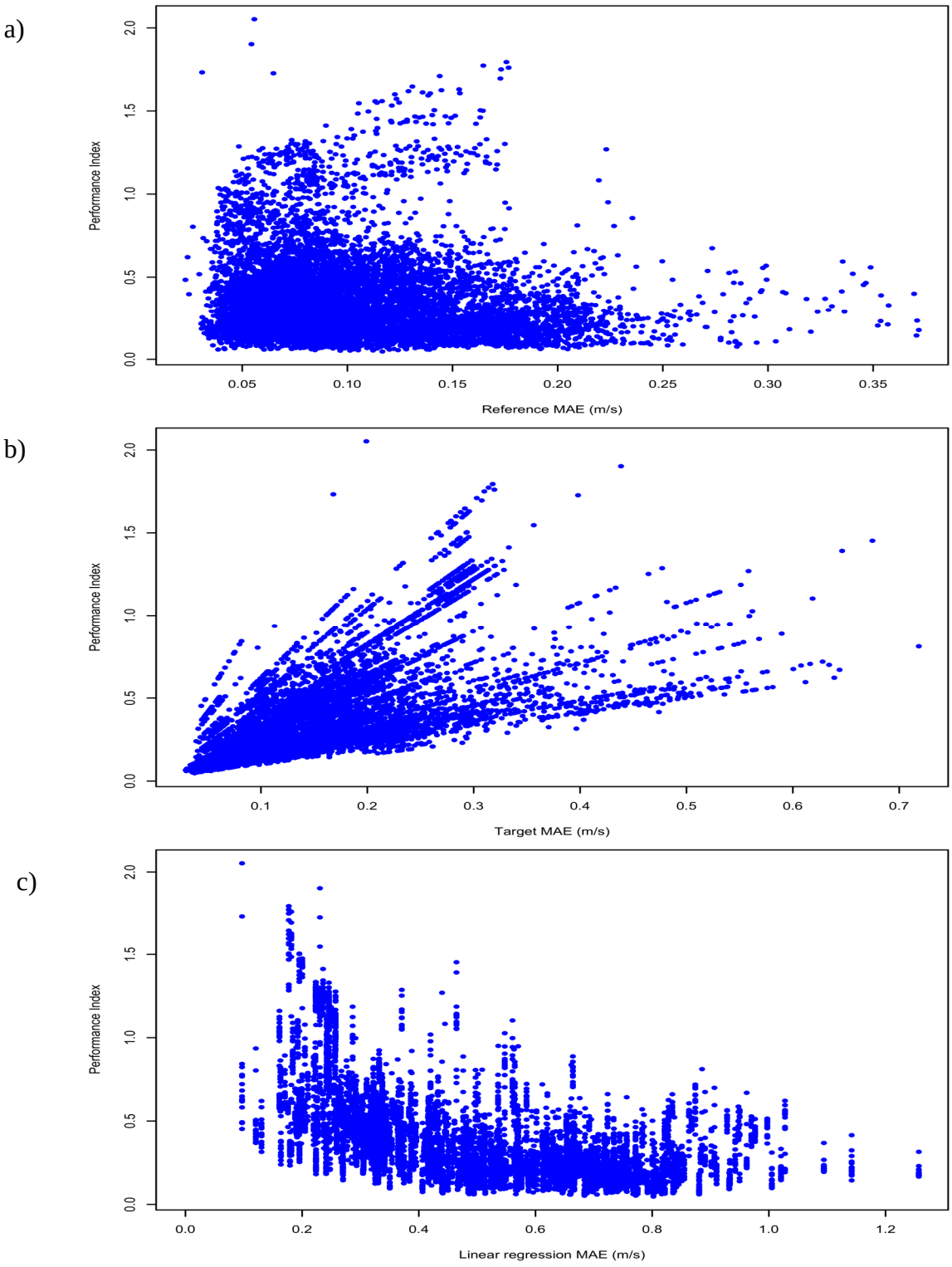


Figure 79. Performance Index against reference, target and linear regression MAE.

In Figure 80b) the linear regression bias has a pear shape and this can be interpreted that if linear regression has extremely low bias, the PCA-MCP cannot improve because linear regression is performing well. However, in the lower parts of the pear shape it can be seen that there is improvement of the PCA-MCP method. In other words, the worse linear regression is performing, the better the PCA-MCP performance.

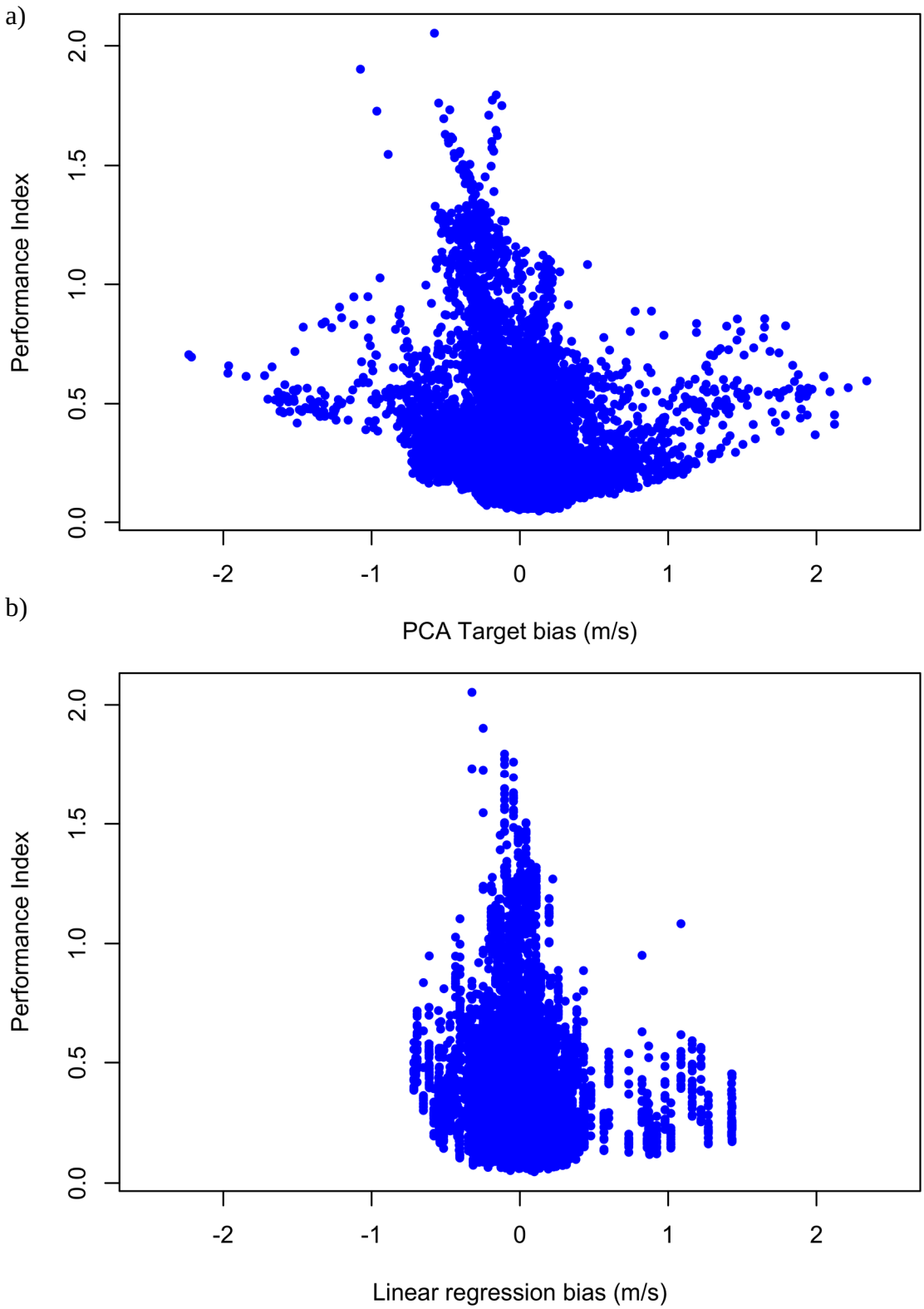


Figure 80. Performance Index again reference and linear regression MAE

Overall, a superiority of PCA-MCP over the standard linear regression can be verified from the graphs of Chapter 6. The final chapter, Chapter 7 will give the overall conclusions and discussion of this research summarizing the most important PCA-MCP's findings but also noting the method's limitations and room for improvement.

Chapter 7 Conclusions of PCA as a wind energy resource tool

This is the final chapter in which the key findings and limitations of this research as well as the next steps and future work will be explored.

7.1 Summary of key findings

From the previous chapters of this thesis, the key findings can be summarized as following. Wind speed can indeed be treated as a dynamical system and it was proven that the time series analysis technique works. Furthermore, it was concluded that PCA can be used successfully for wind forecasting purposes and for wind resource assessment purposes as an MCP method. The PCA-MCP methodology was proven to be in most cases superior to simple linear regression and can be used successfully as measure of the predictions uncertainty.

7.1.1 Strengths and current limitations of PCA as a wind forecasting method

As demonstrated in more detail in Chapter 4, the main conclusions for the application of PCA as a forecasting method that can be made are firstly that PCA is capable of identifying weather regimes by being able to represent the wind measurements in the form of an attractor with a clear structure. Furthermore, it was demonstrated that this can be done both, by just using wind speed measurements and by using multivariate measurements, such as wind speed and wind direction combined.

Applying the PCA to wind forecasting demonstrated that the method is a reliable forecasting method for forecasting wind speeds hours ahead to day ahead. By combining the PCA prediction with persistence prediction at very short time scales, it was possible to eliminate the weakness of applying PCA to a coarsely sampled wind record. It was specifically found that persistence is much better than PCA at short lead times up to 6 hours but that PCA outperforms persistence at longer lead times T.

Using a single point of overlapping values in the forecasting analysis i.e. $n_x = 1$ rather than fitting a short time series of point ($n_x > 1$) overlap seems to yield the best improvement (around 11.2%) of the PCA forecasting results. In other words, the PCA results were 11.2% closer to the actual results in comparison with the persistence method. Thus it was determined that that the best overlapping values was $n_x = 1$. The overall PCA improvement raised from below 8% for only $n_n = 2$ nearest neighbours to above 11% for $n_n = 5$ but then dropped again to around 9%. Using too few or too many neighbours might not have been appropriate since with too few the information used for the analysis might be too little whereas on the contrary, using too many might initially show that we can obtain more information; however, these neighbours might actually lie very far apart from each other in the phase space. As far as the reduced dimensions M_t are concerned, for M_t=16 the PCA improvement seemed to be consistently high for 5.6%. There is clearly a distinct optimum which needs to be determined and this could be possible by optimising the parameters through experience at each site individually.

One of the most useful aspects of PCA over some other forecasting techniques is that it is based on an ensemble forecast using ensembles of similar past events. This allows an estimation of the forecast accuracy at the time when the forecast is made. The analysis showed that this estimated forecast uncertainty is a reliable predictor of the actual forecasting error. This knowledge will be useful for the wind farm operators to evaluate their forecasts and will help with their decision making. Regarding the limitations of PCA, gaps in wind data are a common phenomenon which in the case of PCA was overcome with the linear interpolation of the data. The missing values were treated in a similar way for the PCA-MCP case. The linear interpolation for gap of length N_G from time $T+1$ to $T+N_g$ was performed by

$$U(T+i) = U(T) + \frac{i}{(N_G+1)}\left(U(T+N_g+1) - U(T)\right) \tag{64}$$

7.1.2 Strengths and current limitations of the PCA-MCP method

In the PCA-MCP approach, the formalism determines the shape and coefficients of the best relationship between the target and a reference site by treating the measurements as representative of the joint dynamical system, rather than one as input and the other as output. When applied on a variety of station pairs some several hundred kilometers apart, it was shown that it is almost always superior to the basic standard MCP using linear regression. More specifically, for the majority of cases, the error of the resource prediction was found to be between 10% and 50% of that made using the standard linear regression. Moreover, the mean target MAE was found to be only the 29% of the linear regression MAE.

PCA-MCP seems robust since the MAE graphs are fairly flat for most of the stations in comparison with simple linear regression. This suggests that a good performance of the standard linear regression MCP relies strongly on having chosen a good reference site (which may not always be obvious in advance or even possible), whereas the PCA-MCP method is fairly insensitive to a particular choice of reference site. Being able to calculate the MAE_{ref} as part of the PCA-MCP prediction also provides a tool to estimate the actual MAE_{tar}.

As was found from the first four MAE graphs, MAE_{lr} is ranging from 0.7 to 1.2 thus is the highest when compared to MAE_{tar} and MAE_{ref} which are relatively low, i.e. up to 0.4. The first 4 target stations bias graphs indicated in general the existence of negative bias and the reason behind this could be that PCA-MPC method predicts slower wind speeds than simple linear regression does. Two reasons could be behind this; possibly the calibration used in the PCA-MCP analysis is not yet optimal i.e. using the mean and standard deviation rations as expressed in Chapter 3.6.2 equation (30) and equation (31). Secondly, the calibration was performed in order to minimise the distribution error i.e. calibrate so as to expect the smallest error in the wind speed distribution.

The next four graphs, followed the same pattern as the first four MAE graphs i.e. the MAE_{lr} was higher when compared to MAE_{tar} and MAE_{ref} . More specifically, the MAE_{lr} was ranging from 0.7 to 1. Examining the MAE_{tar} , it ranges between 0.1 and 0.3 and the lowest one, MAE_{ref} ranges from 0.1 to 0.2. It can be thus concluded that linear regression performed worse than PCA-MCP. From the last four bias graphs for stations 5-8, it can be seen that the bias was above zero in most cases. Stations 5,6 and 8 are low wind speed stations whereas stations 1,2,3,4,7 are high wind speed stations. This indicates that somehow bias is related with whether the prediction comes from a low or high wind speed site, i.e. bias seems to be correlated with the predicted wind speed.

As far as the PCA-MCP limitations are concerned, the initial idea that this reference estimate might give clues about the quality of the target prediction could not be substantiate as verified in the results of Chapter 6. The dataset used was originating from Scotland, U.K. i.e. a 'coastal', mid-latitude European climate. It would be therefore useful to test how the method performs for completely different types of climate so that its sensitivity against climate change would be investigated. It can be concluded that good choice of window and truncation can affect the bias can since similar spreads can be seen for bias values of both PCA-MCP and linear regression. Longer window lengths and smaller truncations as well as the calibration methods have to be explored more. Sensitivity parameters such as the truncation value and window length have proven to be important factors not only for the PCA-MCP analysis, but also for PCA as a forecasting technique, thus, a careful consideration of these parameters when applying PCA should be undertaken By finding the optimum parameter settings, the minimisation of the distribution error and bias will be achieved.

7.1.3 Future work

To encounter the limitations described in the previous section 7.1.2, future improvement steps should be considered. The next stage of the work is to subject it to a systematic analysis to identify if it is possible to judge the quality of the prediction at the target site from the information available to the analyst. Other possible quantities to test in the next stage of development are the estimates returned from applying the

truncated PCA-MCP predictor to the data from the reference site for the training period and thus predicting the target wind speed for the training period. Also, quantify the uncertainty based on both calibration stage and return of prediction for the target and reference sites.

Since it performed well against spatially distant stations such as for example for station 1, Stornoway, it would be interesting to investigate PCA-MCP for offshore wind resource assessment. Moreover, further validation of the PCA-MCP method should be performed. This could be achieved with the use of various datasets, multiple reference and/or target sites, different training periods and different choice of parameter settings such as from half day to 7 days of window length or truncations varying from for example, 9 to 12. However, when the method would be tested for different sites and datasets, different parameter settings may be chosen.

Further investigation should also be undertaken on how to treat more effectively the missing data values. For example, a possible solution to be considered would be to use the PCA forecasting methodology in order to fill in the data missing values for the PCA-MCP method, also explained in Appendix B. Wind direction calculation should also be worth being considered for systematic evaluation in comparison with other wind direction calculation methods as proposed by literature [18].

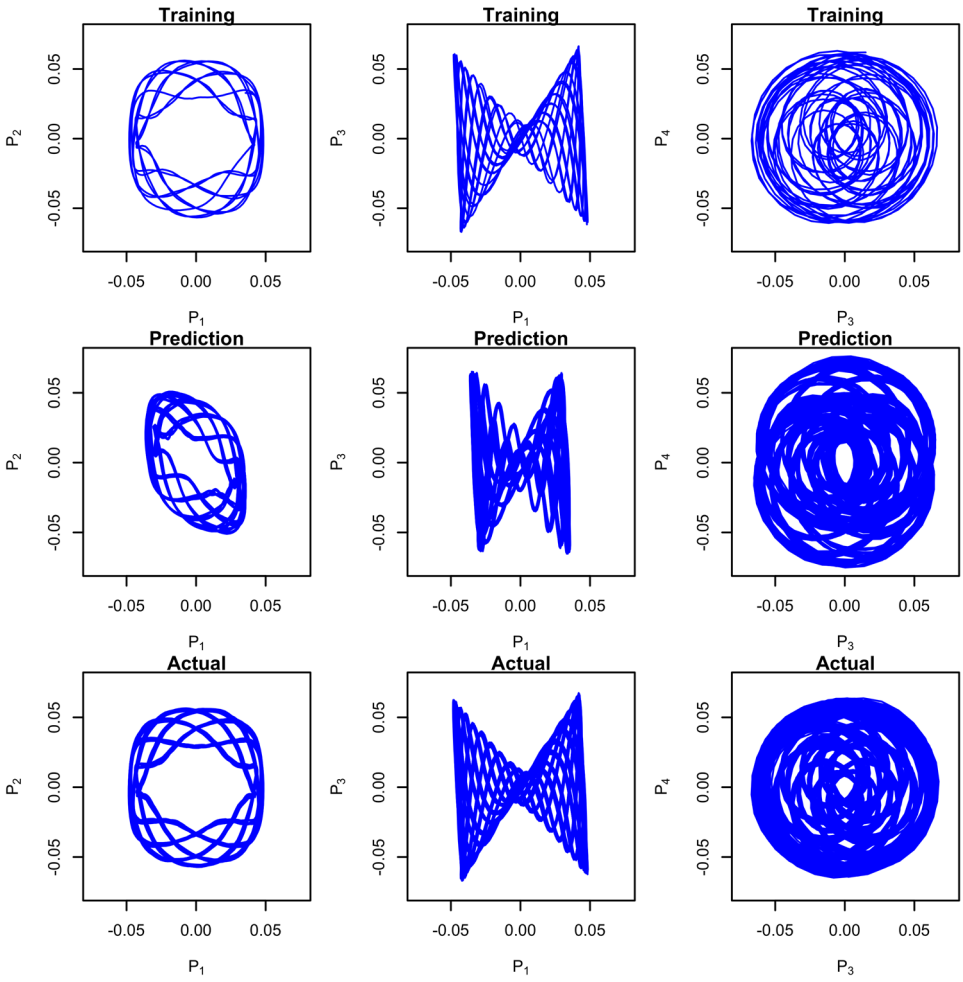
Appendix

Appendix A: Supplementary results of Chapter 5

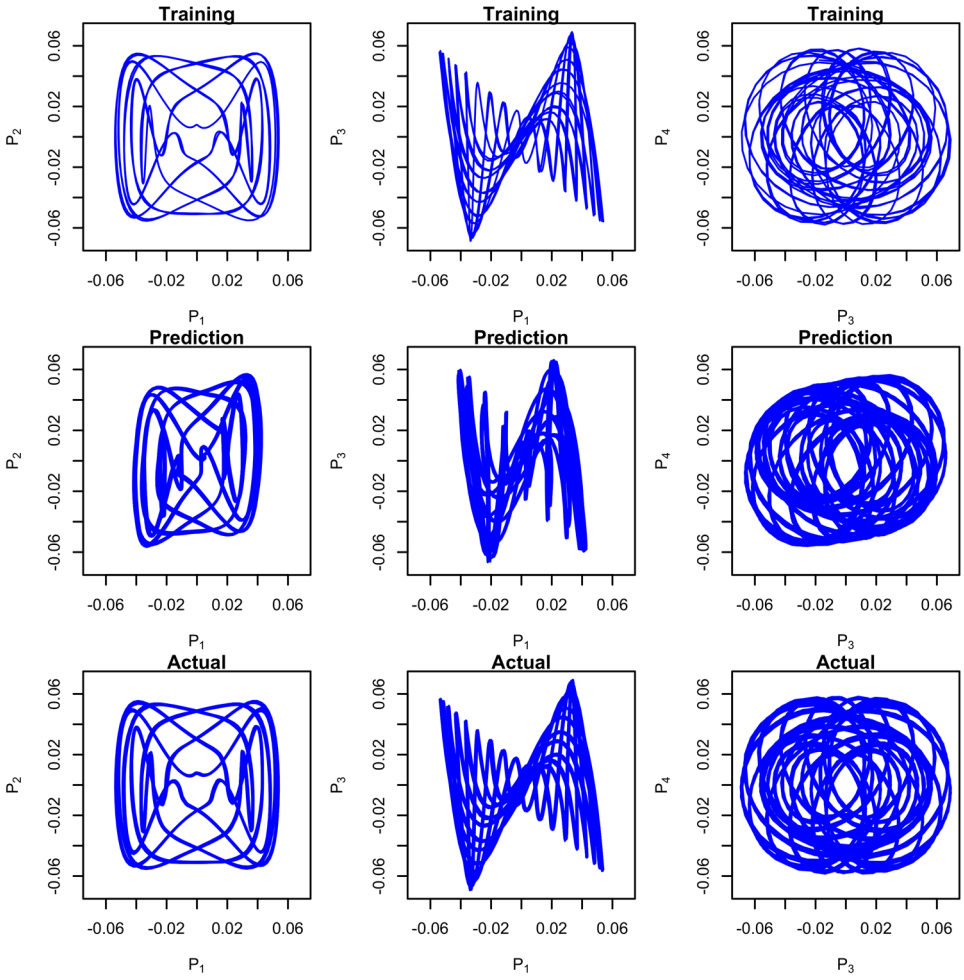
The supplementary PCA pendulum application results of Chapter 5 are presented here with respect to the reference case of Table 5 with parameter values: $\delta\phi$ =π/9, A_1 =4, A_2 =0.3, f_1 =0.5, f_2 =0.3. Here, three representative values of each parameter are shown.

A.1 For the A_1 values

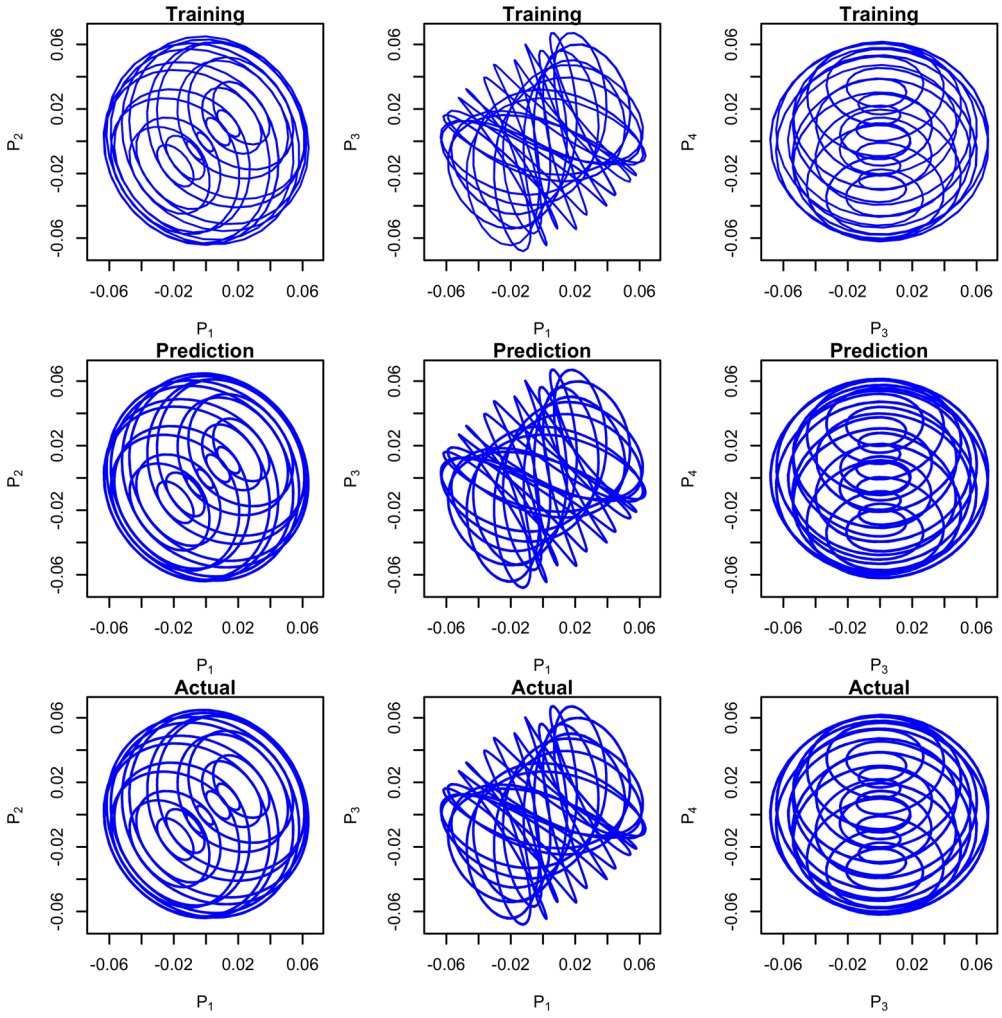
A_1=0.9



A_1=2

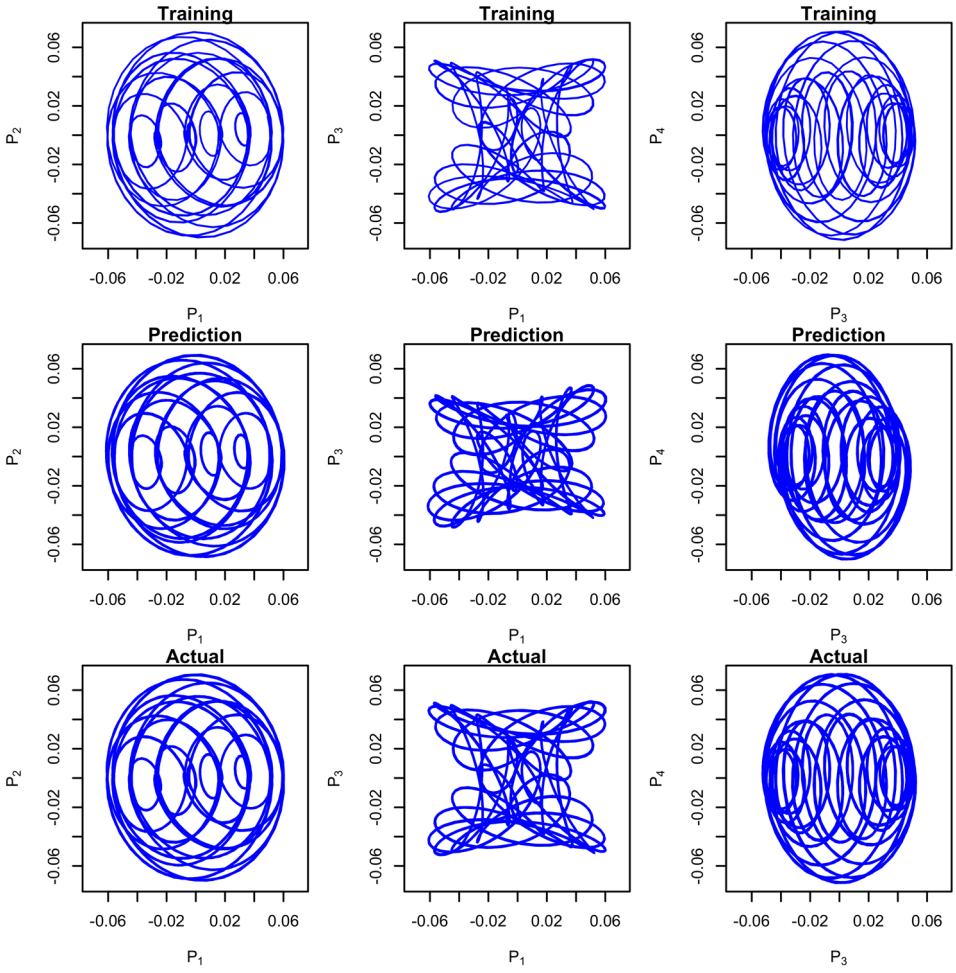

Training
Training
Training
Prediction
Prediction
Prediction
Actual
Actual
Actual
P1
P2
P3
P4
-0.06
-0.02
0.02
0.06

A_1=10

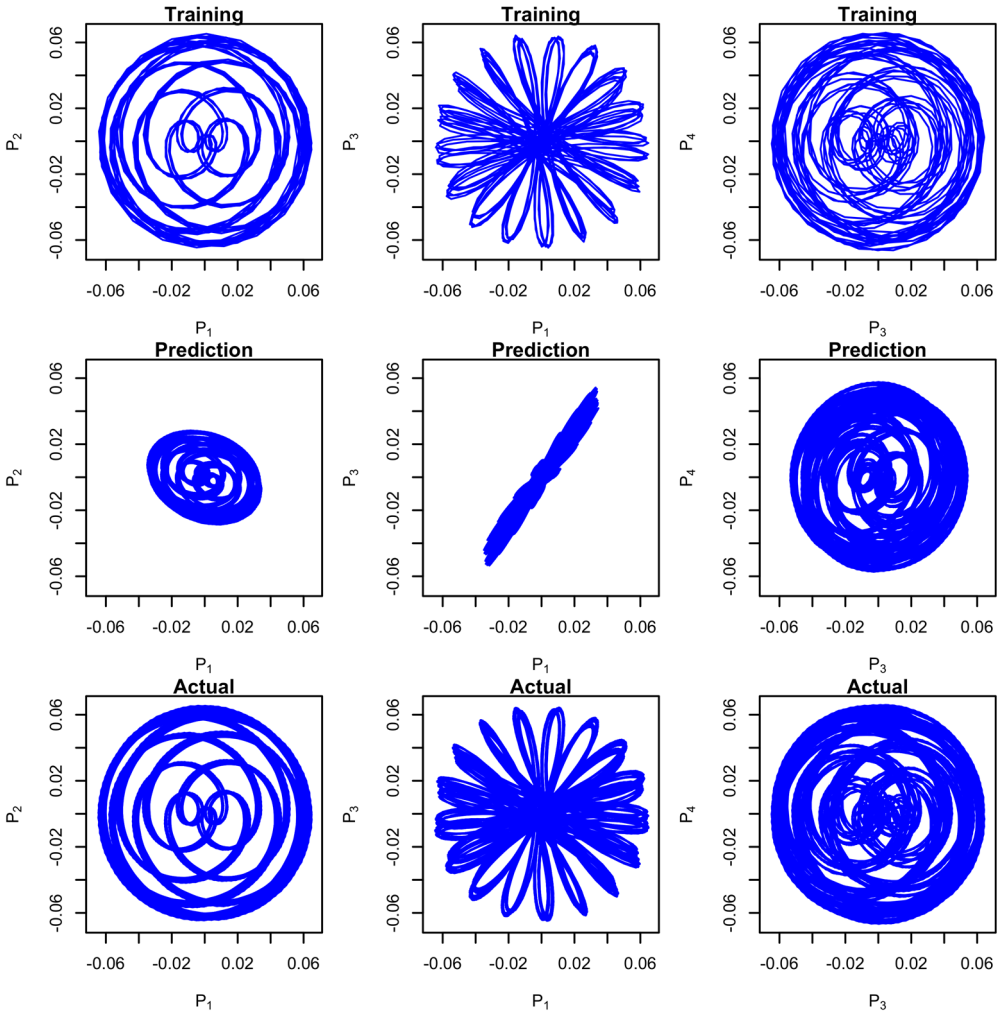

Training
Training
Training
Prediction
Prediction
Prediction
Actual
Actual
Actual
P1
P2
P3
P4
-0.06
-0.02
0.02
0.06

A.2 For the A_2 values

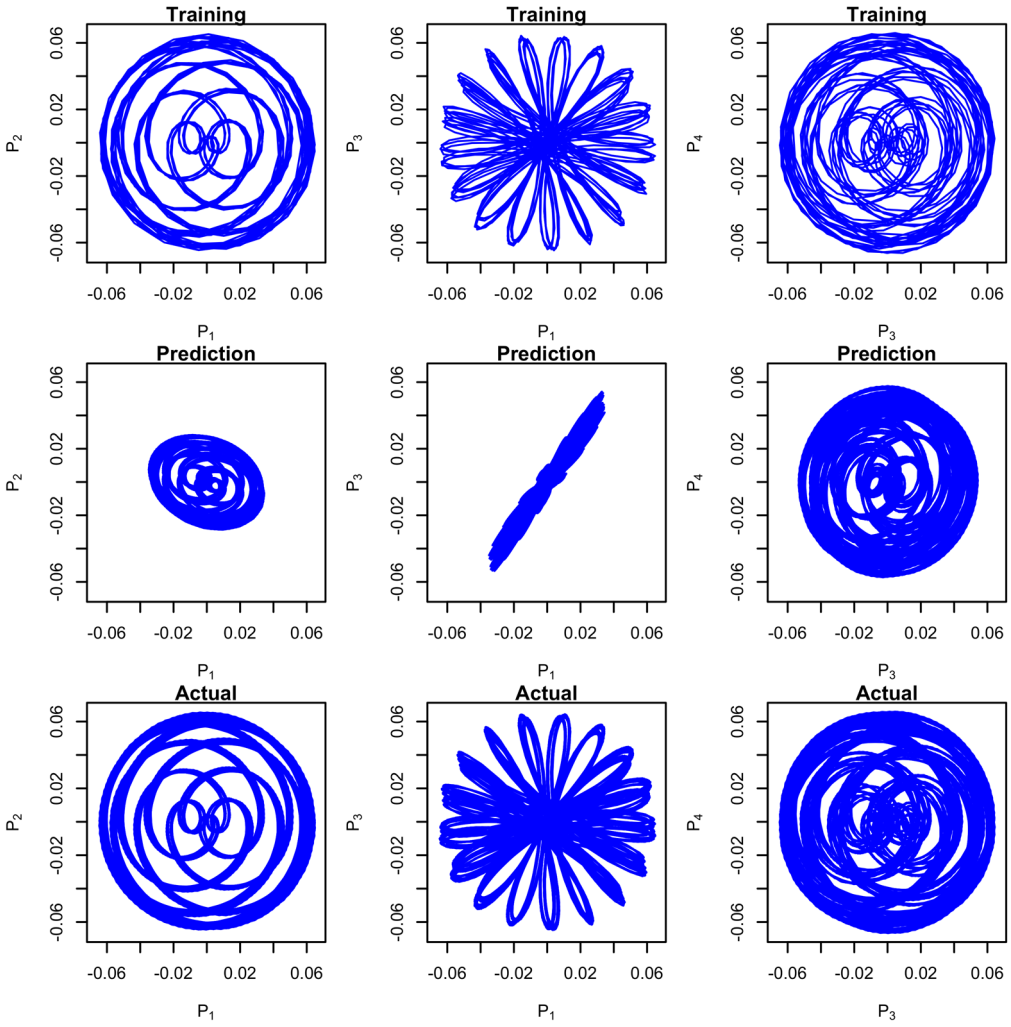
A_2=0.1



A_2=1

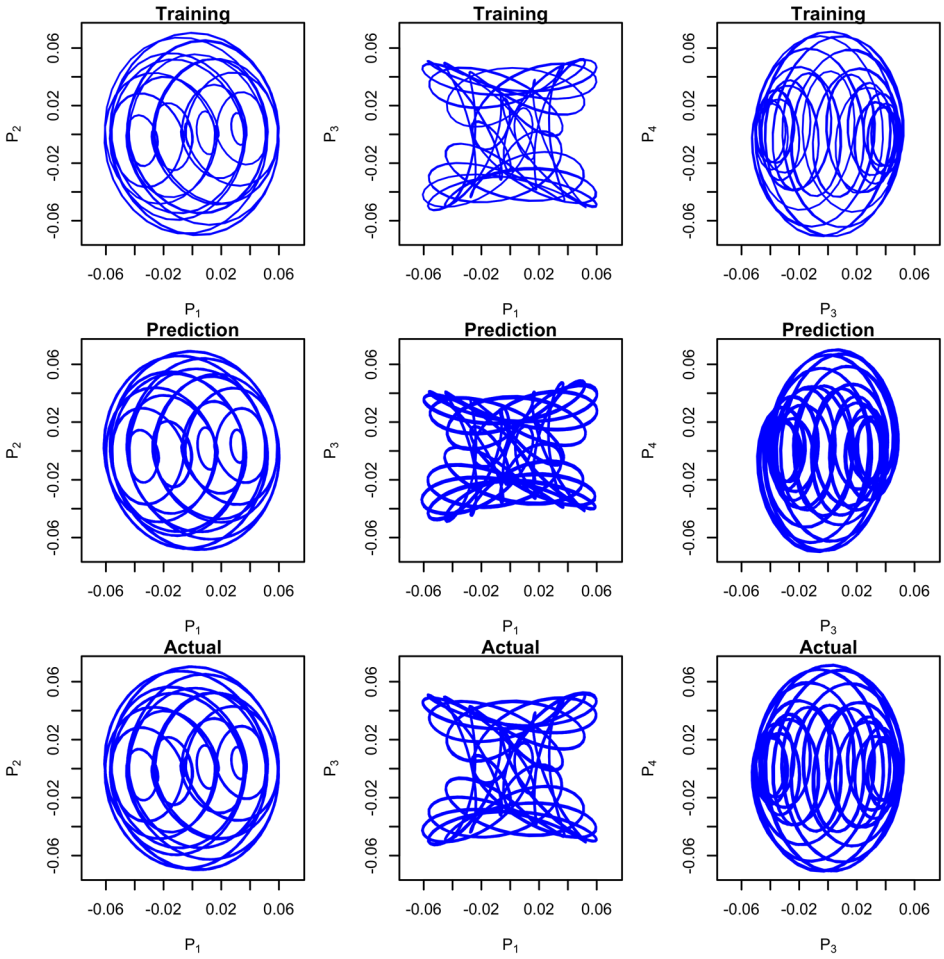

Training
Training
Training
Prediction
Prediction
Prediction
Actual
Actual
Actual
P1
P2
P3
P4
-0.06
-0.02
0.02
0.06

A_2=10

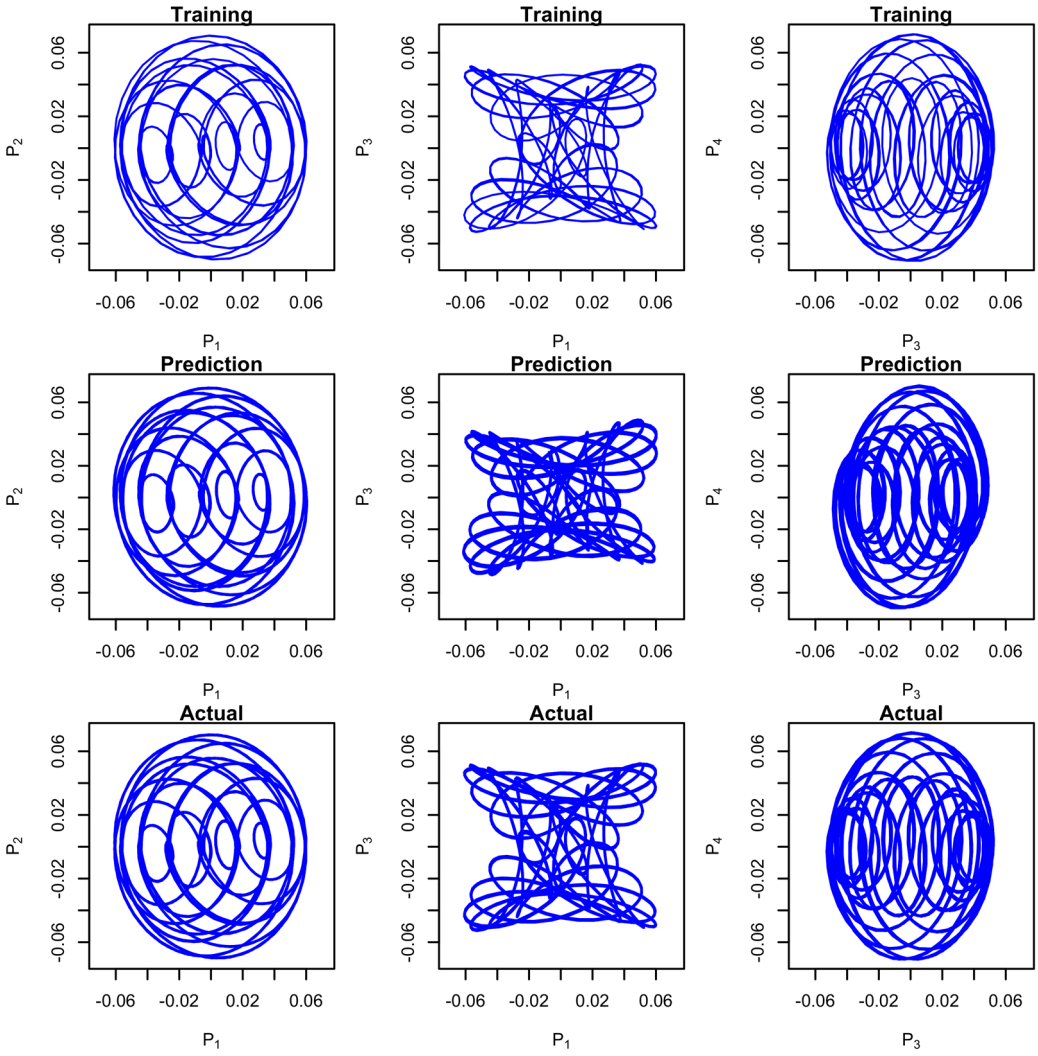

Training
Training
Training
Prediction
Prediction
Prediction
Actual
Actual
Actual
P1
P2
P3
P4
-0.06
-0.02
0.02
0.06

A.3 For the $\delta\varphi$ values

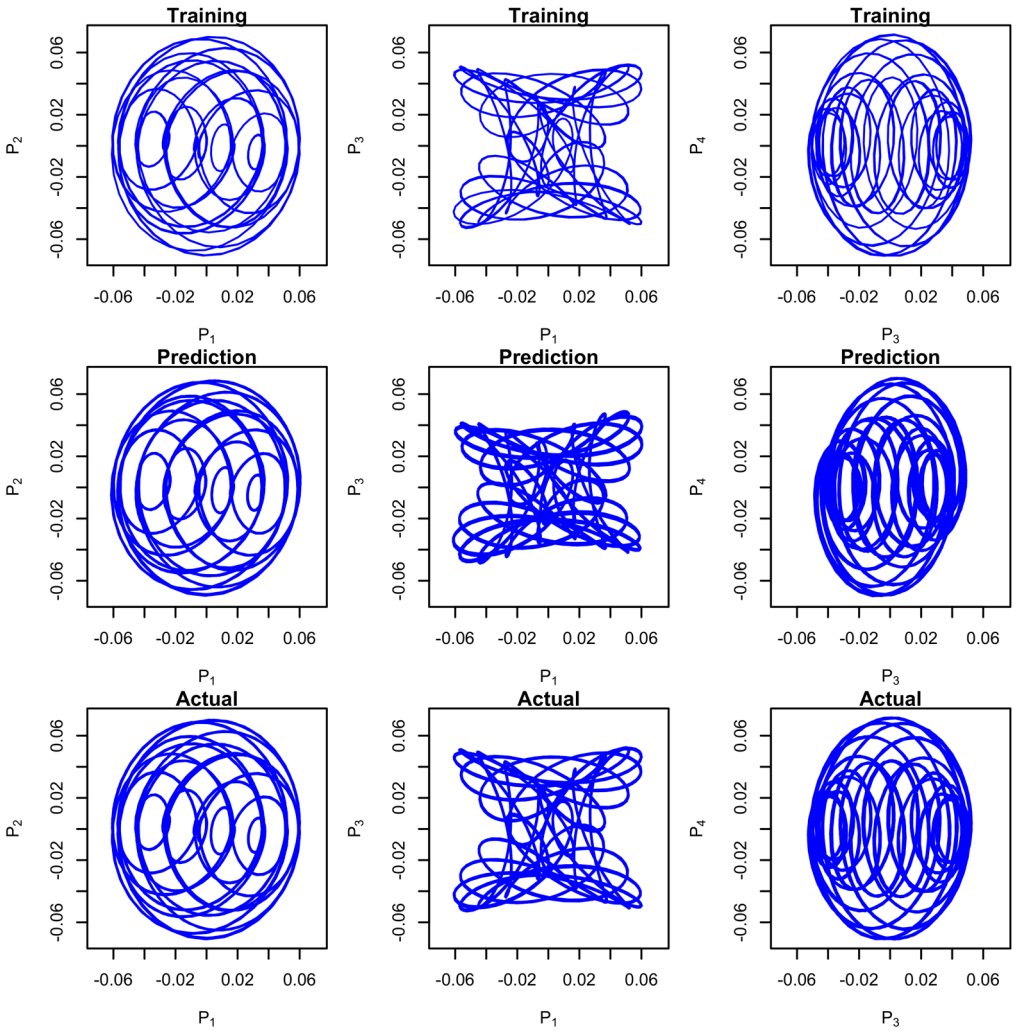
$$\delta\varphi = \frac{\pi}{12}$$



$$\delta\varphi = \frac{\pi}{2}$$

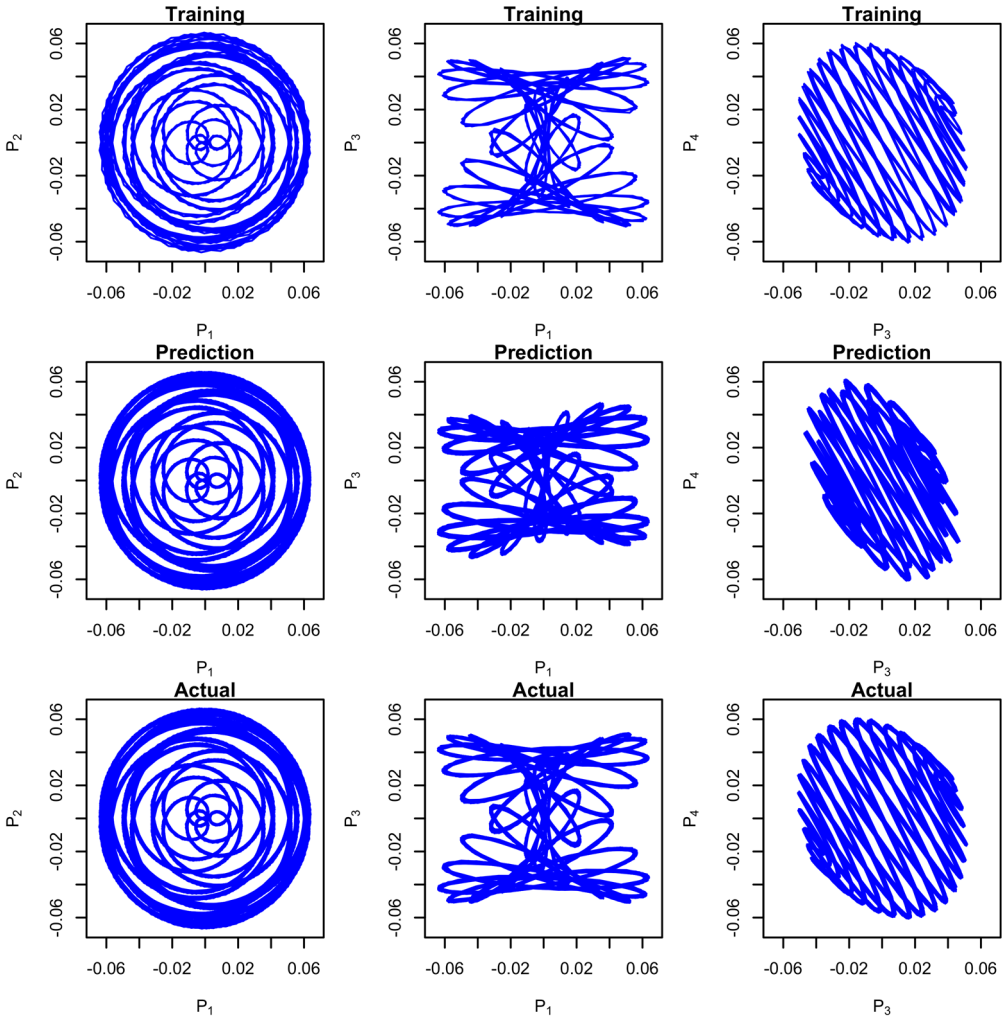


$\delta\varphi = 1.5\pi$

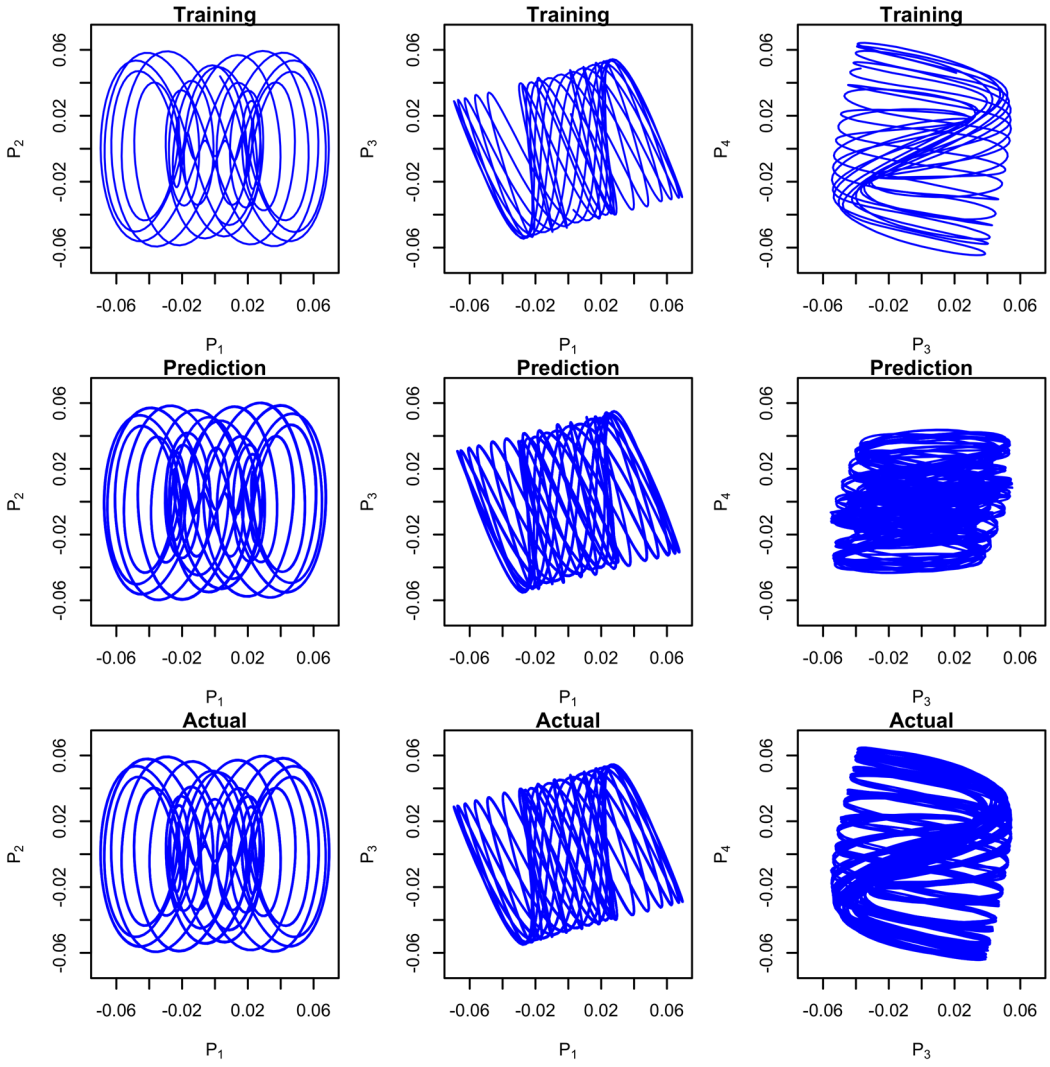


A.4 For the f_1 values

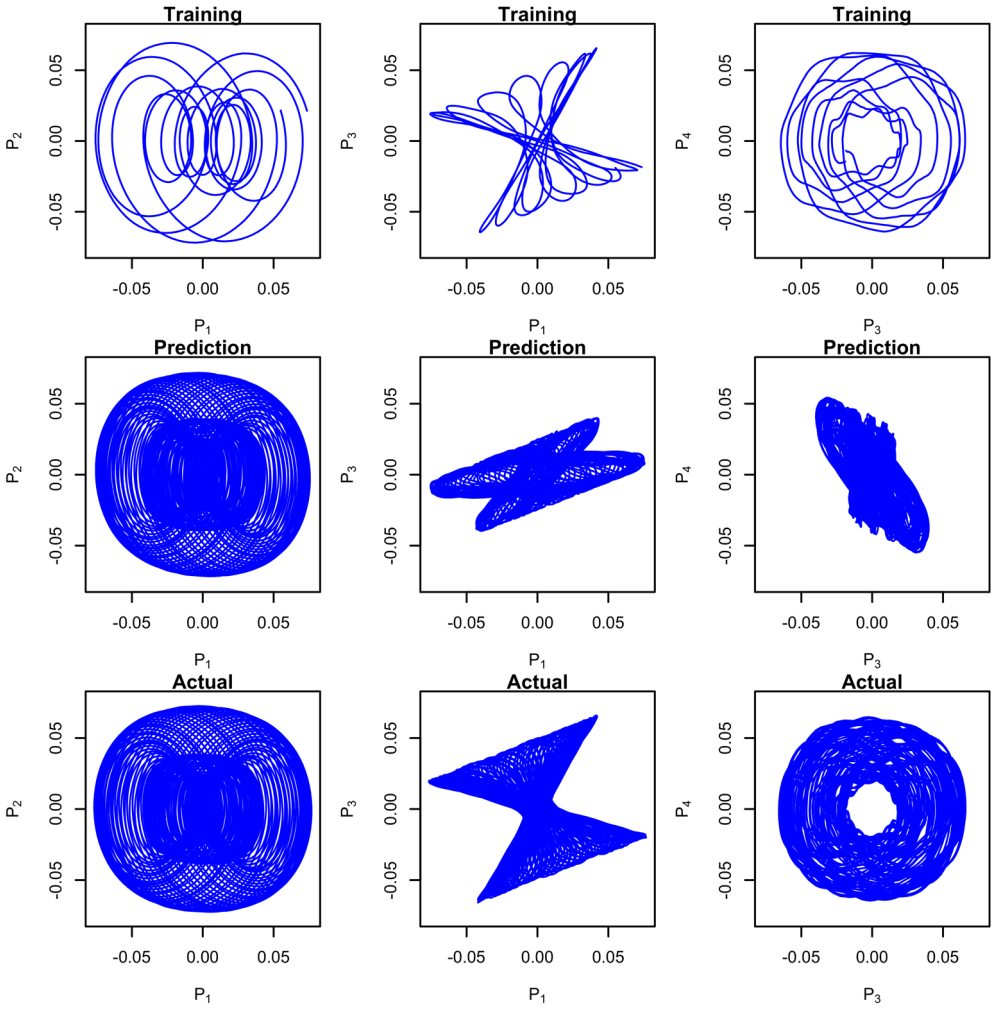
f_1=0.2



f_1=1

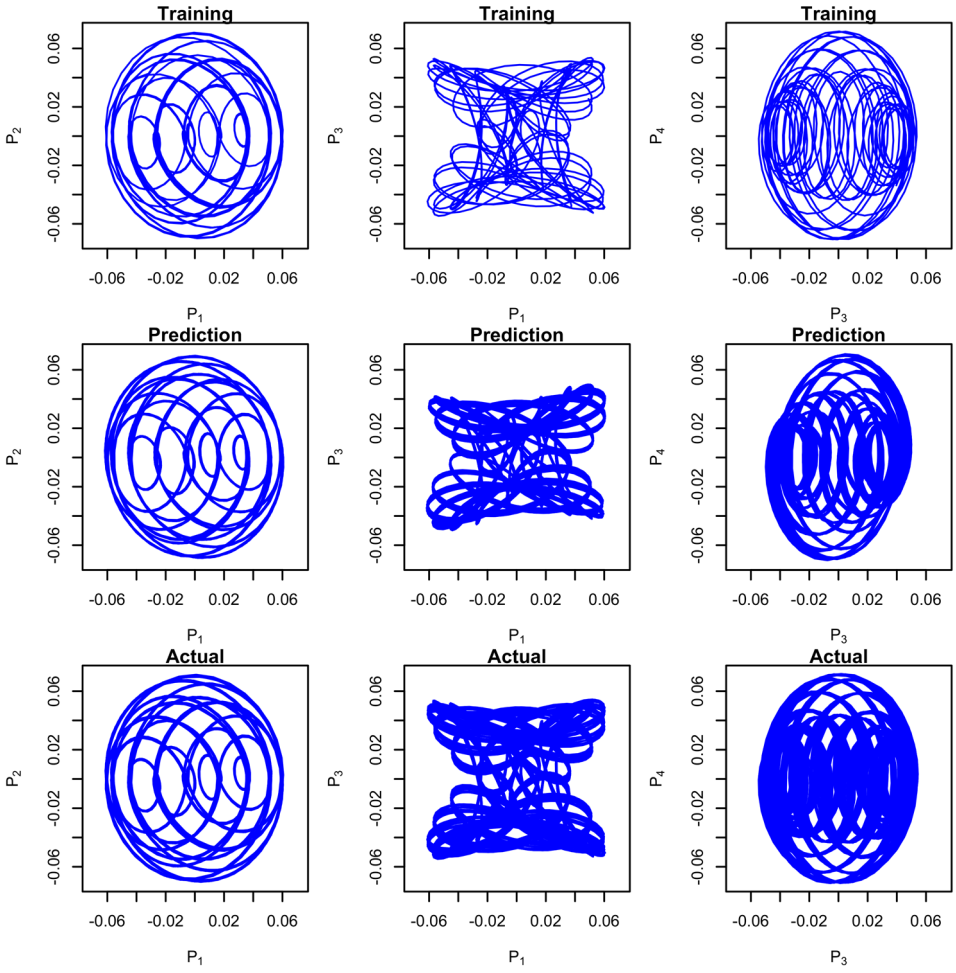

Training
Training
Training
Prediction
Prediction
Prediction
Actual
Actual
Actual
P1
P2
P3
P4
-0.06
-0.02
0.02
0.06

f_1=3.14

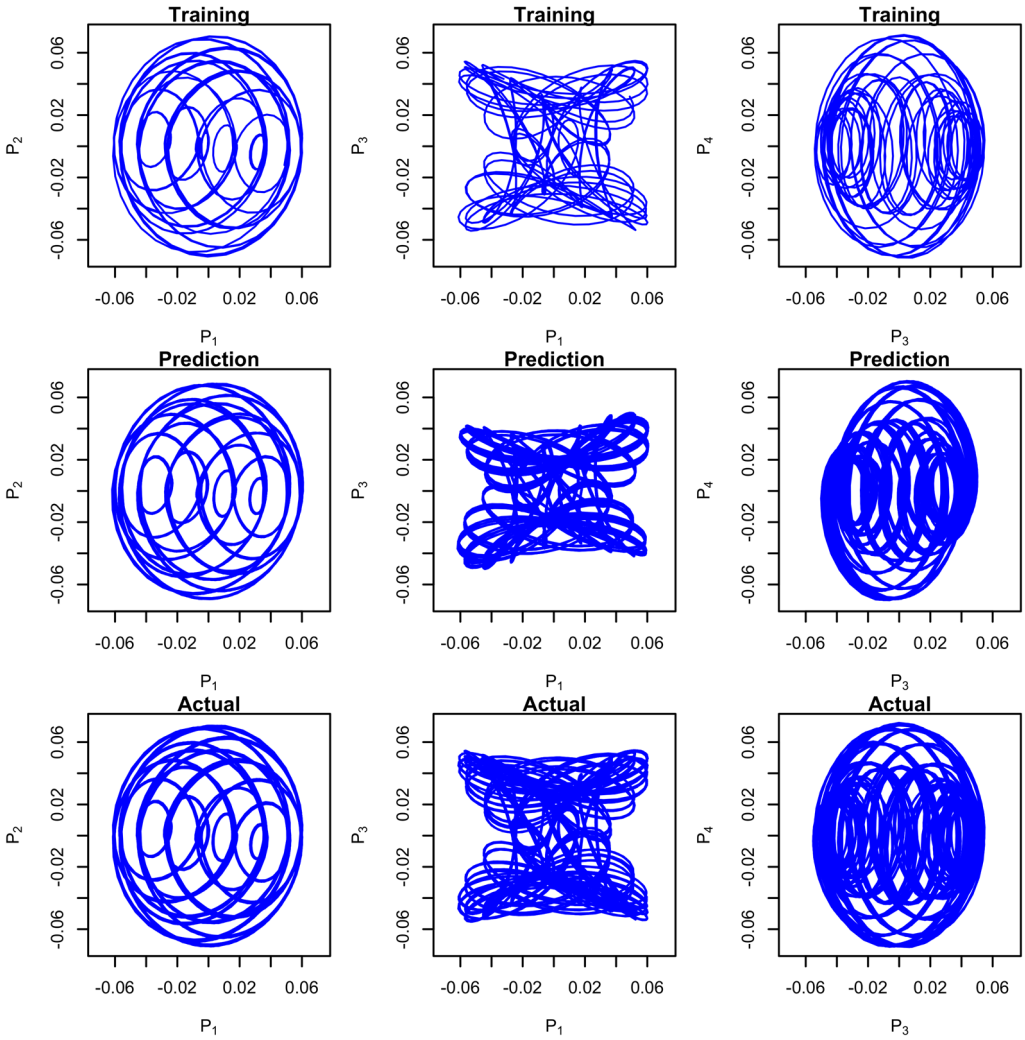

Training
Training
Training
Prediction
Prediction
Prediction
Actual
Actual
Actual
P1
P2
P3
P4
-0.05
0.00
0.05

A.5 For the f_2 values

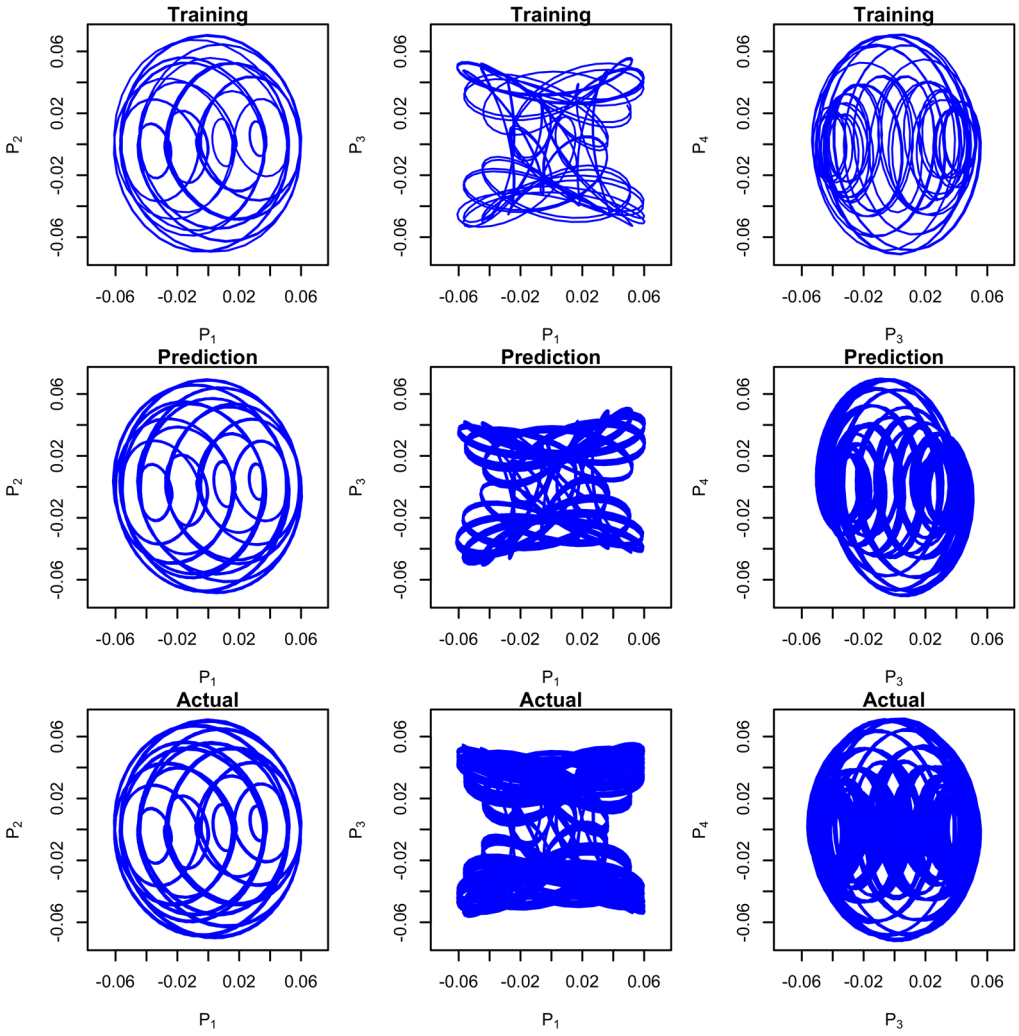
f_2=0.9



f_2=2


Training
Training
Training
Prediction
Prediction
Prediction
Actual
Actual
Actual
P1
P2
P3
P4
-0.06
-0.02
0.02
0.06

f_2=10


Training
Training
Training
Prediction
Prediction
Prediction
Actual
Actual
Actual
P1
P2
P3
P4
-0.06
-0.02
0.02
0.06

Appendix B: Data quality

The data used for the analysis originated from the BADC MIDAS dataset [75] and were linearly interpolated as explained in Chapter 7.1.1 and equation (64). Their quality is depicted in Table B.1 for the years 2000-2010. While short gaps of a few hours, up to about a day, are unlikely to affect the results, extended gaps of many days should be expected to affect the results. Strategies to avoid this, could be to leave long gaps in the data and then delete any rows in the delay matrix which contain missing values or to use a more advanced interpolation. One possibility could be to use the PCA forecasting methodology from Chapter 4, to forecast from the last measurement before the gap and backcast from the measurement after the gap to fill in the gap. This would be an option to be explored in future.

Station	Wind speed data	Valid Data	Missing Data	Gaps (< 4h long)	Gaps (> 4h long)
1-Stornoway	96432	90746 (94.1%)	5686 in 889 gaps	580	309
2-Blackford Hill	96432	76931 (79.8%)	19501 in 300 gaps	273	27
3-Machrihanish	96432	92883 (96.3%)	3549 in 2085 gaps	1985	100
4-Salsburgh	96432	96101 (99.6%)	331 in 124 gaps	117	7
5-Prestwick Gannet	96432	94453 (97.9%)	1979 in 640 gaps	524	116
6-Gogarbank	96432	94691 (98.2%)	1741 in 364 gaps	288	76
7-Port Ellen	96432	94051 (97.5%)	2381 in 780 gaps	693	87
8-Bishopton	96432	94500 (98%)	1932 in 478 gaps	402	76

Table B1. Quality of data used for the analysis

Blackford Hill was different from the other stations in that it had a very long period of no data from the middle of 2003 to the middle of 2004.

Appendix C: The PCA forecasting algorithm

Appendices C and D contain all the R scripts [31] used. The lines of the scripts starting with '#' are the comments of the code and not actual part of the algorithms.

C.1 Preparation of data and setting of parameters

```
# Load the wind speed data and extract the training and prediction years for the chosen site

# Do the following only if starting from fresh:
# load("wind_yr.RData")
print("Data loaded")
# select station
istn <- which(wind$stshort == "Ggb")
print(wind$stname[istn])
# save plots to file if idev is equals to jpg
# idev <- "jpg"
# Select parameters
# Training
# Select the year for training
yeart1 <- 2008
yeart2 <- 2009
# Choose delay parameters
tau <- 1
win <- 48
# Choose reduced dimension
# dimred <- 15
# Forecasting
# Year in which wind speed forecasting is carried out
yearp <- 2010
# Prediction horizon
horizon <- 25
# Number of nearest neighbours
```

```
# nnearest <- 2
# Overlap for finding place on attractor
# overlap <- 1
# Number of predictions to be carried out  (doing every single hour would take far too long)
Npred <- 100
# Training
print("Training ...")
source("Forecasting_CS_3a.R")
# Prediction
print("Predicting ...")
source("Forecasting_CS_3b.R")
# Postprocessing
print("Postprocessing ...")
source("Forecasting_CS_3c.R")
```

C.2 Training the PCA forecasting model

```
# TRAINING
# Find the entries in the record corresponding to that year
idxt <- (wind$year >= (yeart1-1900) & wind$year <= (yeart2-1900))
# extract the wind speed
ut <- wind$u[idxt,istn]
ut <- ut[!is.na(ut)]
# extract the wind direction
dirt <- wind$udir[idxt,istn]
dirt <- dirt[!is.na(dirt)]
# Prepare the delay matrix
Nrec = length(ut)
# Remove mean and scale by standard deviation
umean <- mean(ut, na.rm = TRUE)
usd <- sd(ut, na.rm = TRUE)
dmean <- mean(dirt, na.rm = TRUE)
dsd <- sd(dirt, na.rm = TRUE)
```

```
y1 <- (ut-umean)/usd
y2 <- (dirt-dmean)/dsd
# Create delay matrix
rowdel <- Nrec-tau*win
tarr <- matrix(nrow = rowdel,ncol=2*win)
for (i in 1:(tau*win)){
        tarr[,i] <- y1[i:(i+rowdel-1)]
        tarr[,(i+win)] <- y2[i:(i+rowdel-1)]
}
# Carry out PCA
svd(tarr) -> dtmp
lambda <- dtmp$d
svec <- t(dtmp$v)
pc <- dtmp$u
# Plot key results from Training
if (idev == "jpg") {
        jpeg(paste("Spectrum_T_",wind$stshort[istn],"_",yeart1,"_",yeart2,".jpeg",sep=""
        ), width = 600, height = 480, units = "px", pointsize = 12, quality = 100, bg =
        "white", res = NA, restoreConsole = TRUE)
}
par(mfrow=c(1,1))
par(mai=c(.8,0.8,.2,.2))
plot(lambda, main = "Training set singular values")
if (idev == "jpg") {dev.off()}
print("PCA completed")
dfull <- dim(svec)
##set new lambda
lambdafull <- matrix(data = 0, nrow = dfull[1], ncol = dfull[2])
for (i in 1:dfull[1]){lambdafull[i,i]<- lambda[i]}
lambdared <- matrix(data = 0, nrow = dimred, ncol = dimred)
for (i in 1:dimred){lambdared[i,i]<- lambda[i]}
lambdai<-matrix(0,ncol=dimred,nrow=dfull[2])
for (i in 1:dimred){lambdai[i,i]=1/lambda[i]}
svecR <- svec[1:dimred,]
```

```
pcR <- pc[,1:dimred]
# To show that it works plot a section of the time series, overlay the full reconstruction, and the reconstruction using the reduced dimension
if (idev == "jpg") {
    jpeg(paste("Training_example",dimred,".jpeg",sep=""), width = 600, height = 480, units = "px", pointsize = 12, quality = 75, bg = "white", res = NA, restoreConsole = TRUE)}
par(mfrow=c(1,1))
par(mai=c(.8,0.8,.2,.2))
ptime <- 1:(2*win)
 ydel1recfull <- pc[1:(win+1),]%*%lambdafull%*%svec
y1recfull <- c(ydel1recfull[1,1:(win-1)],ydel1recfull[,win])
 utrecfull <- y1recfull*usd + umean
ydel1rec <- pcR[1:(win+1),]%*%lambdared%*%svecR
y1rec <- c(ydel1rec[1,1:(win-1)],ydel1rec[,win])
utrec <- y1rec*usd + umean
plot(ptime,ut[ptime],"l", xlab = "time (h)", ylab = "u (m/s)")
lines(ptime,utrecfull,col = "green", lty = 3)
lines(ptime,utrec,col = "red", lty = 1)
 if (idev == "jpg") {dev.off()}
```

C.3 Predicting with the PCA forecasting model

```
# PREDICTION
# Extract the wind speeds and directions for the prediction year
# Number of overlapping time points for finding nearest neighbours
idxp <- wind$year == (yearp - 1900)
tyearp <- seq(1,length(idxp[idxp]))
upi <- approx(tyearp,wind$u[idxp,istn],tyearp)
up <- upi$y
dirpi <- approx(tyearp,wind$udir[idxp,istn],tyearp)
dirp <- dirpi$y
# number of wind speed measurements (a full year 8760)
Npy <- length(up)
```

```
# maximum number of predictions is the number of measurements minus the history needed to create time delay matrix for finding place on attractor with specific overlap minus the prediction horizon

Npredmax <- Npy - tau*win - overlap - horizon

# spread out the actual predictions over the year

istep <- floor(Npredmax/Npred)

# rescale wind speeds and

yp1 <- (up-umean)/usd

yp2 <- (dirp-dmean)/dsd

delaylength <- tau*win + overlap - 1

idelay2 <- seq((-delaylength+1), 0, by = 1)

# prepare matrices for predictions and errors as well as vectors of actual current observation

upredicted <- matrix(nrow = Npred, ncol = horizon)

dupredicted <- matrix(nrow = Npred, ncol = horizon)

prederr <- matrix(nrow = Npred, ncol = horizon)

uactual <- matrix(nrow = Npred, ncol = horizon)

upersist <- matrix(nrow = Npred, ncol = horizon)

persisterr <- matrix(nrow = Npred, ncol = horizon)

uobs <- array(dim = Npred)

dirobs <- array(dim = Npred)

upast <- array(dim = c(Npred,delaylength))

tarrcur <- matrix(nrow = overlap, ncol = 2*win)

DistVector <- array(dim = c(nnearest,dimred))

print("Start the loop")

# Prediction loop:

for (i0 in 1:Npred){

    # start at beginning of the year,  select i0 + idelay2 section; make delay matrix, project onto reduced EOF from training set, find nearest neighbours and predict; compare;  repeat for all possible sections in 2010

    Source ("Forecasting_CS_3b1.R")

    # Plot intermediate results

    #    if (i0 == 3){

        #      source("Forecasting_CS_3b2.R")

    #    }
```

```
} # end of the prediction 'for (i0 in 1:Npred)' loop
# save original prediction
upredicted_std <- upredicted
```

C.4 Post-processing of the PCA forecasting model

```
#  POSTPROCESSING
ptime <- (1:horizon) - 1
# trying various correction filters
filter <- array(0,dim=horizon)
filterlength <- horizon - 1
filter[1:filterlength] <- seq(1,1/filterlength,by= - 1/filterlength )
for (i0 in 1:Npred){
      upredicted[i0,] <- upredicted_std[i0,] - filter*(upredicted_std[i0,1] - uobs[i0])
}
# Calculate mean errors and uncertainties
prederr <- upredicted - uactual
persisterr <- upersist - uactual
# only select the cases in the relevant wind speed range
idx <- (uobs >= 4)
#MSE
meanperr <- colMeans(abs(prederr[idx,]),na.rm=TRUE)
sderr <- apply(abs(prederr[idx,]),MARGIN=2,FUN=sd, na.rm=TRUE)
meanuncert <- colMeans(dupredicted[idx,],na.rm=TRUE)
sduncert <- apply(abs(dupredicted[idx,]),MARGIN=2,FUN=sd, na.rm=TRUE)
persistmean <- colMeans(abs(persisterr[idx,]),na.rm=TRUE)
persistsd <- apply(abs(persisterr[idx,]),MARGIN=2,FUN=sd, na.rm=TRUE)
#Bias
biasperr <- colMeans((prederr[idx,]),na.rm=TRUE)
bsderr <- apply((prederr[idx,]),MARGIN=2,FUN=sd, na.rm=TRUE)
biasuncert <- colMeans(dupredicted[idx,],na.rm=TRUE)
bsduncert <- apply((dupredicted[idx,]),MARGIN=2,FUN=sd, na.rm=TRUE)
biaspersistmean <- colMeans((persisterr[idx,]),na.rm=TRUE)
bpersistsd <- apply((persisterr[idx,]),MARGIN=2,FUN=sd, na.rm=TRUE)
```

```
#RMSE
rmseperr <- colSums(prederr[idx,]^2/length(which(idx)),na.rm=TRUE)
rmseuncert <- colSums(dupredicted[idx,]^2/length(which(idx)),na.rm=TRUE)
rmsepersistmean <- colSums(persisterr[idx,]^2/length(which(idx)),na.rm=TRUE)
performanceindex <- round(mean((persistmean-meanperr))/mean(persistmean)*100,1)
# Plot the annually averaged statistics MSE
ymax <- max(c(meanuncert+sduncert,meanperr,persistmean))*1.01
if ((idev == "jpg")) {
    jpeg(filename = paste("meanPred_err_h_", dimred, "_", nnearest, "_", overlap, "_", horizon, ".jpeg", sep = ""), width = 600, height = 480, units = "px", pointsize = 12, quality = 100, bg = "white", res = NA, restoreConsole = TRUE)
}
par(mfrow=c(1,1))
par(mai=c(1.2,1.2,0.8,.2))
mtext <- paste(wind$stshort[istn],": W=", win,", D=", dimred,", o=", overlap, ", nn=", nnearest, "; N=", Npred)
plot(ptime,meanperr,"l", col="red", xlab="prediction time (h)", ylab = expression(paste("MAE ",symbol(delta)," u (m/s)")), ylim = c(0,ymax), yaxs="i", lwd = 2)
points(ptime,meanuncert ,pch=21,col="blue")
# lines(ptime, meanperr+sderr )
# lines(ptime,meanperr-sderr )
lines(ptime,meanuncert + sduncert ,lty = "dotted",col="blue")
lines(ptime,meanuncert - sduncert, lty="dotted",col="blue")
lines(ptime,persistmean,col="darkgreen", lty = 4, lwd = 2)
legend(horizon*0.7,0,c("predicted", "actual","persistence"), lty = c(0,1,4), col =c("blue","red","dark green"), pch =c(21,NA,NA), yjust = 0)
grid(col="darkgrey")
if (idev == "jpg") {dev.off()}
# Plot the annually averaged statistics BIAS
ymax <- max(c(biasuncert+bsduncert,biasperr,biaspersistmean))*1.01
if ((idev == "jpg")) {
    jpeg(filename = paste("biasPred_err_h_", dimred, "_", nnearest, "_", overlap, "_", horizon, ".jpeg", sep = ""), width = 600, height = 480, units = "px", pointsize = 12, quality = 100, bg = "white", res = NA, restoreConsole = TRUE)
}
```

```
par(mfrow=c(1,1))

par(mai=c(1.2,1.2,0.8,.2))

mtext <- paste(wind$stshort[istn],": W=", win,", D=", dimred,", o=", overlap, ", nn=", nnearest, "; N=", Npred)

plot(ptime,biasperr,"l", col="red", xlab="prediction time (h)", ylab = expression(paste("Bias ",symbol(delta)," u (m/s)")), ylim = c(0,ymax), yaxs="i", lwd = 2)

# points(ptime,biasuncert ,pch=21,col="blue")

#lines(ptime, biasperr+bsderr )

#lines(ptime,biasperr-bsderr )

# lines(ptime,biasuncert + bsduncert ,lty = "dotted",col="blue")

# lines(ptime,biasuncert - bsduncert, lty="dotted",col="blue")

lines(ptime,biaspersistmean,col="darkgreen", lty = 4, lwd = 2)

legend(horizon*0.7,0,c("actual","persistence"), lty = c(1,4), col =c("red","dark green"), yjust = 0)

grid(col="darkgrey")

if (idev == "jpg") {dev.off()}

# Plot the annually averaged statistics RMSE

ymax <- max(c(rmseuncert,rmseperr,rmsepersistmean))*1.01

if ((idev == "jpg")) {

    jpeg(filename = paste("rmsePred_err_h_", dimred, "_", nnearest, "_", overlap, "_", horizon, ".jpeg", sep = ""), width = 600, height = 480, units = "px", pointsize = 12, quality = 100, bg = "white", res = NA, restoreConsole = TRUE)

}

par(mfrow=c(1,1))

par(mai=c(1.2,1.2,0.8,.2))

mtext <- paste(wind$stshort[istn],": W=", win,", D=", dimred,", o=", overlap, ", nn=", nnearest, "; N=", Npred)

plot(ptime,rmseperr,"l", col="red", xlab="prediction time (h)", ylab = expression(paste("RMSE ", symbol(delta)," u (m/s)")), ylim = c(0,ymax), yaxs="i", lwd = 2)

#points(ptime,rmseuncert ,pch=21,col="blue")

# lines(ptime,meanuncert + sduncert ,lty = "dotted",col="blue")

# lines(ptime,meanuncert - sduncert, lty="dotted",col="blue")

lines(ptime,rmsepersistmean,col="darkgreen", lty = 4, lwd = 2)
```

```
legend(horizon*0.7,0,c("actual","persistence"), lty = c(1,4), col =c("red","dark green"), yjust = 0)
grid(col="darkgrey")
if (idev == "jpg") {dev.off()}
```

Appendix D: The PCA-MCP algorithm

D.1 Main R script of the PCA-MCP model: preparation of data and setting of parameters

```
#load data
load("Winds_8st_2000-2011.RData")
require("openair")
source("PCAMCP_function.R")
source("simplelrerror_function.R")
source("lin_reg_pred.R")
tau <-1
Ytrain<-seq(099,109)
Nyear<-length (Ytrain)
trunca<-c(3,6,9,12)
wina<-c(1,2,3,4)*24
Ntrunc<-length(trunca)
Nwin <- length(wina)
#store reference,target and lr MAE,Bias and performance index in matrices
Meanerror<-array(dim=c(8,8,Nyear,Nwin,Ntrunc,3))
abserrorref <- array(dim=c(8,8,Nyear,Nwin,Ntrunc))
biasref <- array(dim=c(8,8,Nyear,Nwin,Ntrunc))
abserrortar <- array(dim=c(8,8,Nyear,Nwin,Ntrunc))
biastar <- array(dim=c(8,8,Nyear,Nwin,Ntrunc))
abserrorlin <- array(dim=c(8,8,Nyear,Nwin,Ntrunc))
biaslin <- array(dim=c(8,8,Nyear,Nwin,Ntrunc))
Perfindex <- array(dim=c(8,8,Nyear,Nwin,Ntrunc))
for (st1 in 1:1) {
      for (st2 in 8:8) {
```

```
# Find records where we have both sites
idx<-(!is.na(uwind[st1,])&
!is.na(uwind[st2,])&!is.na(udir[st1,])&!is.na(udir[st2,]) )
# Extract the wind speed & wind direction
#wind speed for referene site
uabs <- uwind[st1,idx]
#wind speed for target
uabs2 <- uwind[st2,idx]
#wind direction for GGB
dabs<-udir[st1,idx]
#wind direction for BFH
dabs2<-udir[st2,idx]
date <- timeseql[idx]
#create vector combination consisting of wind speed and direction
#positive u coming from west to east, negative u coming from east to west
#positive v coming from north to south, negative u coming from south to
north
#u1 and v1 is the vector combination of wind speed and direction for GGB
u1<-uabs*sin(dabs/180*pi)
v1<-uabs*cos(dabs/180*pi)
#u2 and v2 is the vector combination of wind speed and direction for BFH
u2<-uabs2*sin(dabs2/180*pi)
v2<-uabs2*cos(dabs2/180*pi)

# Prepare past data (as we always use the same years
jdx <- date$year >= 099 & date$year <= 110
upast1<-u1[jdx]
vpast1<-v1[jdx]
referencepast <-sqrt(upast1^2+vpast1^2)
upast2<-u2[jdx]
vpast2<-v2[jdx]
targetpast <-sqrt(upast2^2+vpast2^2)
past<-cbind(u1[jdx],v1[jdx])
        for (iyear in 1:1) {
```

```
jdx<-((date$year==Ytrain[iyear]) | (date$year==Ytrain[iyear+1]))
u109 <- u1[jdx]
v209 <- v1[jdx]
u309<-u2[jdx]
v409<-v2[jdx]
Nrec <- length(u109)
nchan<- 4
signalfull <- matrix(nrow=Nrec,ncol=4)
signalfull[,1] <- u109
signalfull[,2] <- v209
signalfull[,3] <- u309
signalfull[,4] <- v409
xbreaks <- seq(0,500)

for (iwin in seq (1,2)) {
        win <- wina[iwin]
        for(itrunc in 1:2) {
                trunc <- trunca[itrunc]
                casename<paste("St",st1,"St",st2,"Y",
                Ytrain[iyear]+1900,"win",win,"Trunc",trunc,
                sep="")
                PCAMCP(signalfull,tau,win,trunc,past,casename)
                ->prediction2
                targetpred <-sqrt(prediction2[,3]^2+
                prediction2[,4]^2)
                referencepred<- sqrt(prediction2[,1]^2
                +prediction2[,2]^2)
                #histogram of reference and target site
                # predictions saved as jpeg
                jpeg(filename=paste("./MCP graphs/",
                casename,"Histogram actual and pred ref.jpeg",
                sep=""))
                histreferencepast<-hist(referencepast,
```

```
xbreaks ,plot=FALSE)

histreferencepred<-hist(referencepred, xbreaks , prob=TRUE,xlab="wind speed (m/s)",main="",xlim=c(0,30))

lines(histreferencepast$mid, histreferencepast$density, col="red", lwd=4) legend(20,.1,c("Predicted ws",

"Actual ws"),col=c("black","red"), pch=c(22,NA),lty=c(0,1))

dev.off()

jpeg(filename = paste("./MCP graphs/", casename,"Histogram actual and pred target.jpeg",

sep=""))

histtargetpast<-hist(targetpast, xbreaks, plot=FALSE)

histtargetpred<-hist(targetpred, xbreaks, prob=TRUE,xlab="wind speed (m/s)", main="",
xlim=c(0,30))

lines(histtargetpast$mid,histtargetpast$density,

col="red", lwd=4)

legend(20,.1,c("Predicted ws", "Actual ws"), col=c("black","red"), pch=c(22,NA),lty=c(0,1))

dev.off()

#Meanerror is the bias of reference,

#target and linear regression

Meanerror[st1,st2,iyear,iwin,itrunc,1] <- mean(referencepred) - mean(referencepast)

Meanerror[st1,st2,iyear,iwin,itrunc,2] <- mean(targetpred) - mean(targetpast)

Meanerror[st1,st2,iyear,iwin,itrunc,3] <- simplelrerror(uabs,uabs2,referencepast,targetpast)

histreferencepast<-hist(referencepast, xbreaks ,plot=FALSE)

histreferencepred<-hist(referencepred, xbreaks , plot=FALSE)

histtargetpast<-hist(targetpast, xbreaks , plot=FALSE)
```

```
 histtargetpred<-hist(targetpred, xbreaks , plot=FALSE)

#abserroref,abserrortar are the MAE

#of reference and target

#biasref,biastar are the Bias of reference and target

errorref<-histreferencepred$density-histreferencepast$density

abserrorref[st1,st2,iyear,iwin,itrunc] <-sum(abs(errorref))

biasref[st1,st2,iyear,iwin,itrunc]<-sum((errorref))

errortar<-histtargetpred$density-histtargetpast$density

abserrortar[st1,st2,iyear,iwin,itrunc] <-sum(abs(errortar))

biastar[st1,st2,iyear,iwin,itrunc] <-sum((errortar))

#  If abserrorref is somehow correlated with

# abserrortar, then we can say that the abserrorref
# is somehow a measure of the 'predictability'

# given a linear regression histogram

linreghist <-lin_reg_pred(uabs,uabs2,referencepast)

errorlin<-linreghist-histreferencepast$density

#MAE and Bias for simple linear regression

abserrorlin[st1,st2,iyear,iwin,itrunc]<-sum(abs(errorlin))

biaslin[st1,st2,iyear,iwin,itrunc]<-sum((errorlin))

#Performance index, ratio of absolute abserrortar
#over abserrorlin (we want it to be less than 1)

Perfindex[st1,st2,iyear,iwin,itrunc] <- abserrortar[st1,st2,iyear,iwin,itrunc] / abserrorlin[st1,st2,iyear,iwin,itrunc]

#converting back to degrees wind direction for
#wind rose purposes

jpeg(filename = paste("./MCP graphs/",casename,"Windrose actual target.jpeg",sep=""))

windrose1 <- data.frame(cbind(targetpast, atan2(upast2,vpast2)/pi*180))
```

```
                                    names(windrose1) <- c("U","dir")
                                    windRose(windrose1,ws="U",wd="dir")
                                    dev.off()
                                    jpeg(filename = paste("./MCP
                                    graphs/",casename,"Windrose actual
                                    ref.jpeg",sep=""))
                                    windrose2 <- data.frame(cbind(referencepast,
                                    atan2(upast1,vpast1)/pi*180))
                                    names(windrose2) <- c("U","dir")
                                    windRose(windrose2,ws="U",wd="dir")
                                    dev.off()
                                    jpeg(filename = paste("./MCP
                                    graphs/",casename,"Windrose predicted target
                                    .jpeg",sep=""))
                                    windrose3 <- data.frame(cbind(targetpred,
                                    atan2(prediction2[,3],prediction2[,4])/pi*180))
                                    names(windrose3) <- c("U","dir")
                                    windRose(windrose3,ws="U",wd="dir")
                                    dev.off()
                                    jpeg(filename = paste("./MCP
                                    graphs/",casename,"Windrose predicted
                                    ref.jpeg",sep=""))
                                    windrose4 <- data.frame(cbind(referencepred,
                                    atan2(prediction2[,1],prediction2[,2])/pi*180))
                                    names(windrose4) <- c("U","dir")
                                    windRose(windrose4,ws="U",wd="dir")
                                    dev.off()
                                }
                            }
                        }
                    }
                }
            }
```

D.2 Training the PCA-MCP method

```
# create function which performs PCA
```

```
training <- function(signalfull,tau,win,nchan,casename){
	dims<-dim(signalfull)
	Nrec<-dims[1]
	nchan<-dims[2]
	smean <-colMeans(signalfull)
	#sstdev <-var(signalfull)
	sstdev <- apply(signalfull,2,sd)
	signalv <- matrix(nrow=Nrec,ncol=nchan)
	# training rescaled data
	for (i in 1:nchan){
		signalv[,i] <- (signalfull[,i]-smean[i])/sstdev[i]}
	# create the time-delay matrix
	Ntd <- Nrec- (win-1)*tau
	Ncd= nchan*win
	Ntd->n
	Ncd->m
	tarr <- array(0, c(Ntd,Ncd))
	for (i in 1:win) {
		for (j in 1:nchan){
			signalv[(1+(i-1)*tau):(Ntd+(i-1)*tau),j] -> tarr[,(i+(j-1)*win)]
		}
	}
	# carry out a Singular Value Decomposition and split the output into singular
	# values (lambda), singular vectors (svec) and principal components, then plot a
	# selection of them
	svd(tarr) -> dtmp
	#singular vectors
	svec <- dtmp$v
	#principal components
	pc <- dtmp$u
	#singular values
	lambda <- dtmp$d
	jpeg(filename = paste("./MCP graphs/",casename,"lambda.jpeg",sep=""))
	plot(lambda)
```

```
        dev.off()
        training<-dtmp
}
```

D.3 Preparing the PCA-MCP method for prediction

```
#truncate to the relevant components
preparation<-function(fullpca,trunc){
        svec <- fullpca$v
        pc <- fullpca$u
        lambdafull <- fullpca$d
        dims <- dim(pc)
        m <- dims[2]
        Sm<-t(svec[1:(m/2),1:trunc])
        Sp<-t(svec[1:m,1:trunc])
        pct <- pc[,1:m]
        lambdaa <- diag(lambdafull[1:trunc])
        lambdai <- diag(1/lambdafull[1:trunc])
        preparation <- list(Sm,Sp,pct,lambdaa,lambdai)
}
```

D.4 Predicting with the PCA-MCP method

```
#prepare time-delay matrix for historical data
prediction <- function(past,tau,win,pca_t,sstdev,smean,sstdevcorr,smeancorr){
        dims<-dim(past)
        nchan2<-dims[2]
        Nrec2<-dims[1]
        nchan<-2*nchan2
        Ntd <- Nrec2- (win-1)*tau
        Ncds= nchan2*win
        Ncd= nchan*win
        signalpast<-array(0,dims)
        for (j in 1:nchan2){
                signalpast[,j]<-(past[,j]-smean[j])/sstdev[j]
        }

        tarr2 <- array(dim=c(Ntd,Ncds))
        for (i in 1:win) {
                for (j in 1:nchan2){
```

```
                signalpast[(1+(i-1)*tau):(Ntd+(i-1)*tau),j] -> tarr2[,(i+(j-1)*win)]
        }
    }

    Sm <- pca_t[[1]]
    Sp <- pca_t[[2]]
    lambdaa <- pca_t[[4]]
    lambdai <- pca_t[[5]]
    #get new pc's
    #Pp <-tarr2%*%t(Sm)%*%lambdai
    #Yp<-Pp%*%lambdaa%*%Sp
    Yp <- tarr2 %*% t(Sp[,1:(nchan2*win)]) %*% Sp
    # new full matrix with the first two channels reproducted as a combination of
    #signalpast and signalv
    signalpred<-array(0,c(Ntd,nchan))
    for (j in 1:nchan){
        signalpred[,j]<-Yp[,(j-1)*win+1]
    }
    #going back to wind speeds, normalised ones
    #wind speeds reference and target sites, scaling training period
    scaledpred<-array(0,c(Ntd,nchan))
    for (j in 1:nchan){
        scaledpred[,j]<-signalpred[,j]*sstdevcorr[j]+smeancorr[j]
    }
    prediction<-scaledpred
}
```

D.5 Performing simple linear regression for comparison with the PCA-MCP method

```
#Perform linear regression and calculate error
simplelrerror<-function (uabs,uabs2,referencepast,targetpast) {
    lm(uabs2~uabs)->simpleMCP
    coef(simpleMCP)->coef
    coef[1]+coef[2]*referencepast->up
    lrerror<-mean(up)-mean(targetpast)
    simplelrerror<-lrerror
}
```

D.6 Calibration of the PCA-MCP method predictions

```
PCAMCP<-function(signalfull,tau,win,trunc,past,casename) {
    source("training_function.R")
```

```
        source("preparation_function.R")
        source("prediction_function.R")
        source("simplelrerror_function.R")
        dims<-dim(signalfull)
        nchan<-dims[2]
        smean<-colMeans(signalfull)
        #sstdev<-var(signalfull)
        sstdev <- apply(signalfull,2,sd)
        training(signalfull,tau,win,nchan,casename)->fullpca
        preparation(fullpca,trunc)->pca_t
        prediction(signalfull[,(1:(nchan/2))],tau,win,pca_t,sstdev,smean,sstdev,smean)   ->
        prediction1
        smean2<-colMeans(prediction1)
        #sstdev2<-var(prediction1)
        sstdev2 <- apply(prediction1,2,sd)
        smeancorr<-smean*(smean/smean2)
        sstdevcorr<-sstdev*(sstdev/sstdev2)
        prediction(past,tau,win,pca_t,sstdev,smean,sstdevcorr,smeancorr) -> prediction2
        PCAMCP<-prediction2
}
```

D.7 Calculating simple linear regression prediction

```
lin_reg_pred<-function (uabs,uabs2,referencepast) {
        lm(uabs2~uabs)->simpleMCP
        coef(simpleMCP)->coef
        coef[1]+coef[2]*referencepast->up
        zdx <- (up < 0)
        up[zdx] <- 0
        xbreaks <- seq(0,500)
        histlregpred<-hist(up, xbreaks , plot=FALSE)
        return(histlregpred$density)
}
```

D.8 Depicting PCA-MCP results in comparison with simple linear regression for different parameter settings

```
if (!exists("abserrorlin")) {
         load("trial3.RData")
        st1 <- seq(1,8)
        st2 <- seq(1,8)
```

```
        yr <- seq(2000,2010)
        wn <- seq(1,4)
        trn <- seq(1,4)
}
# choose which target station, truncation and window you want to plot:
# Then plot for all reference stations and years
ist1 <- st1
ist2 <- 3
iwn <- 2
itrn <- 2
maintext <- paste("Target:",ist2,"; Win = ", wina[iwn], "; Trunc = ", trunca[itrn])
par(mar=c(4,4,2,1))
matplot(st1,abserrorlin[,ist2,,iwn,itrn], "b", lty = 2, pch=20, xlim=c(0,8), ylim=c(0,1.2), xlab = "Reference Station", ylab = "MAE", main = maintext)
matplot(st1,abserrorref[,ist2,,iwn,itrn], "b", lty = 3, pch=21, add=TRUE)
matplot(st1,abserrortar[,ist2,,iwn,itrn], "b", lty = 1, pch=24, add=TRUE)
legend(0,1.2,legend=c("Linear Regression", "Reference", "Target"), lty = c(2,3,1), pch = c(20,21,24))
# choose which reference station, truncation and window you want to plot:
# Then plot for all target stations and years
ist1 <- 5
ist2 <- st2
iwn <- 2
itrn <- 2
maintext <- paste("Reference:",ist1,"; Win = ", wina[iwn], "; Trunc = ", trunca[itrn])
par(mar=c(4,4,2,1))
matplot(st1,abserrorlin[ist1,ist2,,iwn,itrn], "b", lty = 2, pch=20, xlim = c(0,8), ylim=c(0,1.2), xlab = "Target Station", ylab = "MAE", main = maintext)
matplot(st1,abserrorref[ist1,ist2,,iwn,itrn], "b", lty = 3, pch=21, add=TRUE)
matplot(st1,abserrortar[ist1,ist2,,iwn,itrn], "b", lty = 1, pch=24, add=TRUE)
legend(0,1.2,legend=c("Linear Regression", "Reference", "Target"), lty = c(2,3,1), pch = c(20,21,24))
# choose which pair of Reference station, truncation and window you want to plot:
# Then plot for all target stations and years
```

```
ist1 <- 3

ist2 <- 6

iwn <- c(1,4)

itrn <- c(1,4)

maintext <- paste("Reference:",ist1, "Target:", ist2)

par(mar=c(4,4,2,1))

matplot(yr,abserrorlin[ist1,ist2,,1,1], "b", lty = 2, pch=20, xlim= c(1999,2009), ylim=c(0,1.2), xlab = "Training year", ylab = "MAE", main = maintext, col="black")

matplot(yr,abserrorref[ist1,ist2,,iwn,1], "b", lty = 3, pch=21, col = "blue", add=TRUE)

matplot(yr,abserrortar[ist1,ist2,,iwn,1], "b", lty = 1, pch=24, col="red", add=TRUE)

matplot(yr,abserrorref[ist1,ist2,,iwn,2], "b", lty = 3, pch=19, col="blue", add=TRUE)

matplot(yr,abserrortar[ist1,ist2,,iwn,2], "b", lty = 1, pch=17, col="red", add=TRUE)

legend(1999,1.2,legend=c("Linear Regression", "Reference", "Target"), lty = c(2,3,1), pch = c(20,21,24), col=c("black","blue","red"))

# if you want to see if there is any correlation between the reference error and the target error:

par(mfrow=c(2,1))

matplot(abserrorref,abserrortar,pch=20, col="blue", xlab="Reference MAE", ylab="Target MAE", xlim=c(0,max(abserrorref,na.rm=TRUE)), ylim=c(0,max(abserrortar,na.rm=TRUE)), xaxs="i", yaxs = "i")

lines(c(0,1.1),c(0,1.1), col="red", lwd= 2)

# if you want to see if there is any correlation between the linear regr error and the target error:

matplot(abserrorlin,abserrortar,pch=20, col="blue", xlab="Linear regression MAE", ylab="Target MAE", xlim=c(0,max(abserrorlin,na.rm=TRUE)), ylim=c(0,max(abserrortar,na.rm=TRUE)), xaxs="i", yaxs = "i")

lines(c(0,1.1),c(0,1.1), col="red", lwd= 2)

# Overall performance profile

hist(Perfindex,breaks = seq(0,max(Perfindex[is.finite(Perfindex)])+0.1,by =0.1), prob=TRUE, xlab = "Performance Index", main="")

matplot(abserrorref[is.finite(Perfindex)],Perfindex[is.finite(Perfindex)],pch=20, col="blue", xlab="Reference MAE", ylab="Performance Index", xlim=c(0,max(abserrorref,na.rm=TRUE)),
ylim=c(0,max(Perfindex[is.finite(Perfindex)])), xaxs="i", yaxs = "i")

lines(c(0,1.1),c(0,1.1), col="red", lwd= 2)

matplot(abserrortar[is.finite(Perfindex)],Perfindex[is.finite(Perfindex)],pch=20, col="blue", xlab="Target MAE", ylab="Performance Index",
```

```
xlim=c(0,max(abserrortar,na.rm=TRUE)),
ylim=c(0,max(Perfindex[is.finite(Perfindex)])), xaxs="i", yaxs = "i")

lines(c(0,1.1),c(0,1.1), col="red", lwd= 2)

matplot(abserrorlin[is.finite(Perfindex)],Perfindex[is.finite(Perfindex)],pch=20,
col="blue", xlab="Linear Regression MAE", ylab="Performance Index",
xlim=c(0,max(abserrorlin,na.rm=TRUE)),
ylim=c(0,max(Perfindex[is.finite(Perfindex)])), xaxs="i", yaxs = "i")

lines(c(0,1.1),c(0,1.1), col="red", lwd= 2)

# Mean and Median of Performance index

mean(Perfindex[is.finite(Perfindex)])

median(Perfindex[is.finite(Perfindex)])

if (!exists("abserrorlin")) {
        load("trial3.RData")
        st1 <- seq(1,8)
        st2 <- seq(1,8)
        yr <- seq(2000,2010)
        wn <- seq(1,4)
        trn <- seq(1,4)
}

# plot the mean absolute error for the three measures (Lin.Reg;PCA ref and PCA target)
averaging over all years, all windows and all truncations

# Plotting for each target station against the reference stations as the x-axis

par(mfrow=c(1,1))

ist2 <-1

{if (ist2 == 1) { stx <- seq(2,8)}

if (ist2 == 8) {stx <- seq(1,7)}

if (ist2 >1 & ist2< 8) {stx <- c(seq(1,ist2-1),seq(ist2+1,8))}

matplot(stx,apply(abserrorlin[stx,ist2,,1,],c(1),mean,na.rm=TRUE),"b",pch=20,lty="das
hed", ylim= c(0,1.1), xlab = "Reference Station", ylab = "MAE", main=paste("Target
station",ist2), col="black", xlim=c(1,8))

matplot(stx,apply(abserrorref[stx,ist2,,,],c(1),mean,na.rm=TRUE),"b",pch=19,lty="dott
ed",add=TRUE, col="blue")

matplot(stx,apply(abserrortar[stx,ist2,,,],c(1),mean,na.rm=TRUE),"b",pch=24,lty="solid
",add=TRUE, col="red")

legend("topleft",legend=c("Linear Regression", "Reference", "Target"), lty = c(2,3,1),
pch = c(20,19,24), col=c("black","blue","red"))
```

```
ist2 <-2

{if (ist2 == 1) { stx <- seq(2,8)}

if (ist2 == 8) {stx <- seq(1,7)}

if (ist2 >1 & ist2< 8) {stx <- c(seq(1,ist2-1),seq(ist2+1,8))}

matplot(stx,apply(abserrorlin[stx,ist2,,1,],c(1),mean,na.rm=TRUE),"b",pch=20,lty="dashed", ylim= c(0,1.1), xlab = "Reference Station", ylab = "MAE", main=paste("Target station",ist2), col="black", xlim=c(1,8))

matplot(stx,apply(abserrorref[stx,ist2,,,],c(1),mean,na.rm=TRUE),"b",pch=19,lty="dotted",add=TRUE, col="blue")

matplot(stx,apply(abserrortar[stx,ist2,,,],c(1),mean,na.rm=TRUE),"b",pch=24,lty="solid",add=TRUE, col="red")

legend("topright",legend=c("Linear Regression", "Reference", "Target"), lty = c(2,3,1), pch = c(20,19,24), col=c("black","blue","red"))

ist2 <-3

{if (ist2 == 1) { stx <- seq(2,8)}

if (ist2 == 8) {stx <- seq(1,7)}

if (ist2 >1 & ist2< 8) {stx <- c(seq(1,ist2-1),seq(ist2+1,8))}

matplot(stx,apply(abserrorlin[stx,ist2,,1,],c(1),mean,na.rm=TRUE),"b",pch=20,lty="dashed", ylim= c(0,1.1), xlab = "Reference Station", ylab = "MAE", main=paste("Target station",ist2), col="black", xlim=c(1,8))

matplot(stx,apply(abserrorref[stx,ist2,,,],c(1),mean,na.rm=TRUE),"b",pch=19,lty="dotted",add=TRUE, col="blue")

matplot(stx,apply(abserrortar[stx,ist2,,,],c(1),mean,na.rm=TRUE),"b",pch=24,lty="solid",add=TRUE, col="red")

legend("topright",legend=c("Linear Regression", "Reference", "Target"), lty = c(2,3,1), pch = c(20,19,24), col=c("black","blue","red"))

ist2 <-4

{if (ist2 == 1) { stx <- seq(2,8)}

if (ist2 == 8) {stx <- seq(1,7)}

if (ist2 >1 & ist2< 8) {stx <- c(seq(1,ist2-1),seq(ist2+1,8))}

matplot(stx,apply(abserrorlin[stx,ist2,,1,],c(1),mean,na.rm=TRUE),"b",pch=20,lty="dashed", ylim= c(0,1.1), xlab = "Reference Station", ylab = "MAE", main=paste("Target station",ist2), col="black", xlim=c(1,8))

matplot(stx,apply(abserrorref[stx,ist2,,,],c(1),mean,na.rm=TRUE),"b",pch=19,lty="dotted",add=TRUE, col="blue")

matplot(stx,apply(abserrortar[stx,ist2,,,],c(1),mean,na.rm=TRUE),"b",pch=24,lty="solid",add=TRUE, col="red")
```

```
legend("topright",legend=c("Linear Regression", "Reference", "Target"), lty = c(2,3,1), pch = c(20,19,24), col=c("black","blue","red"))

ist2 <-5

{if (ist2 == 1) { stx <- seq(2,8)}

if (ist2 == 8) {stx <- seq(1,7)}

if (ist2 >1 & ist2< 8) {stx <- c(seq(1,ist2-1),seq(ist2+1,8))}

matplot(stx,apply(abserrorlin[stx,ist2,,1,],c(1),mean,na.rm=TRUE),"b",pch=20,lty="dashed", ylim= c(0,1.1), xlab = "Reference Station", ylab = "MAE", main=paste("Target station",ist2), col="black", xlim=c(1,8))

matplot(stx,apply(abserrorref[stx,ist2,,,],c(1),mean,na.rm=TRUE),"b",pch=19,lty="dotted",add=TRUE, col="blue")

matplot(stx,apply(abserrortar[stx,ist2,,,],c(1),mean,na.rm=TRUE),"b",pch=24,lty="solid",add=TRUE, col="red")

legend("topright",legend=c("Linear Regression", "Reference", "Target"), lty = c(2,3,1), pch = c(20,19,24), col=c("black","blue","red"))

ist2 <-6

{if (ist2 == 1) { stx <- seq(2,8)}

if (ist2 == 8) {stx <- seq(1,7)}

if (ist2 >1 & ist2< 8) {stx <- c(seq(1,ist2-1),seq(ist2+1,8))}

matplot(stx,apply(abserrorlin[stx,ist2,,1,],c(1),mean,na.rm=TRUE),"b",pch=20,lty="dashed", ylim= c(0,1.1), xlab = "Reference Station", ylab = "MAE", main=paste("Target station",ist2), col="black", xlim=c(1,8))

matplot(stx,apply(abserrorref[stx,ist2,,,],c(1),mean,na.rm=TRUE),"b",pch=19,lty="dotted",add=TRUE, col="blue")

matplot(stx,apply(abserrortar[stx,ist2,,,],c(1),mean,na.rm=TRUE),"b",pch=24,lty="solid",add=TRUE, col="red")

legend("topright",legend=c("Linear Regression", "Reference", "Target"), lty = c(2,3,1), pch = c(20,19,24), col=c("black","blue","red"))

ist2 <-7

{if (ist2 == 1) { stx <- seq(2,8)}

if (ist2 == 8) {stx <- seq(1,7)}

if (ist2 >1 & ist2< 8) {stx <- c(seq(1,ist2-1),seq(ist2+1,8))}

matplot(stx,apply(abserrorlin[stx,ist2,,1,],c(1),mean,na.rm=TRUE),"b",pch=20,lty="dashed", ylim= c(0,1.1), xlab = "Reference Station", ylab = "MAE", main=paste("Target station",ist2), col="black", xlim=c(1,8))

matplot(stx,apply(abserrorref[stx,ist2,,,],c(1),mean,na.rm=TRUE),"b",pch=19,lty="dotted",add=TRUE, col="blue")
```

```
matplot(stx,apply(abserrortar[stx,ist2,,,],c(1),mean,na.rm=TRUE),"b",pch=24,lty="solid
",add=TRUE, col="red")

legend("topleft",legend=c("Linear Regression", "Reference", "Target"), lty = c(2,3,1),
pch = c(20,19,24), col=c("black","blue","red"))

ist2 <-8

{if (ist2 == 1) { stx <- seq(2,8)}

if (ist2 == 8) {stx <- seq(1,7)}

if (ist2 >1 & ist2< 8) {stx <- c(seq(1,ist2-1),seq(ist2+1,8))}

matplot(stx,apply(abserrorlin[stx,ist2,,1,],c(1),mean,na.rm=TRUE),"b",pch=20,lty="das
hed", ylim= c(0,1.1), xlab = "Reference Station", ylab = "MAE", main=paste("Target
station",ist2), col="black", xlim=c(1,8))

matplot(stx,apply(abserrorref[stx,ist2,,,],c(1),mean,na.rm=TRUE),"b",pch=19,lty="dott
ed",add=TRUE, col="blue")

matplot(stx,apply(abserrortar[stx,ist2,,,],c(1),mean,na.rm=TRUE),"b",pch=24,lty="solid
",add=TRUE, col="red")

legend("topright",legend=c("Linear Regression", "Reference", "Target"), lty = c(2,3,1),
pch = c(20,19,24), col=c("black","blue","red"))

}}}}

par(mfrow=c(3,1))

# need to adjust where to put legend depending on the lines; options "right"; "left",
"topright", "topleft")

legend("left",legend=c("Linear Regression", "Reference", "Target"), lty = c(2,3,1), pch
= c(20,19,24), col=c("black","blue","red"))

#  Illustrating the effect of two window choices

plot(abserrortar[,,,1,],abserrortar[,,,2,], xlab=paste("MAE for Win =",wina[1]),
ylab=paste("MAE for Win =",wina[2]),pch=20)

lines(c(0,1),c(0,1), col="red")

#  Illustrating the effect of first two Truncation choices

plot(abserrortar[,,,,1],abserrortar[,,,,2], xlab=paste("MAE for Truncation",trunca[1]),
ylab=paste("MAE for Truncation",trunca[2]), pch=20)

lines(c(0,1),c(0,1), col="red")

# Illustrating the effect of last two Truncation choices

plot(abserrortar[,,,,3],abserrortar[,,,,4], xlab=paste("MAE for Truncation",trunca[3]),
ylab=paste("MAE for Truncation",trunca[4]), pch=20)

lines(c(0,1),c(0,1), col="red")

par(mfrow=c(3,1))
```

```
# need to adjust where to put legend depending on the lines; options "right"; "left", "topright", "topleft")

legend("left",legend=c("Linear Regression", "Reference", "Target"), lty = c(2,3,1), pch = c(20,19,24), col=c("black","blue","red"))

#  Illustrating the effect of two window choices

plot(abserrortar[,,,1,],abserrortar[,,,2,], xlab=paste("Target MAE for Win =",wina[1]), ylab=paste("Target MAE for Win =",wina[2]),pch=20)

lines(c(0,1),c(0,1), col="red")

#  Illustrating the effect of first two Truncation choices

plot(abserrortar[,,,,1],abserrortar[,,,,2], xlab=paste("Target MAE for Truncation",trunca[1]), ylab=paste("

Target MAE for Truncation",trunca[2]), pch=20)

lines(c(0,1),c(0,1), col="red")

plot(abserrortar[,,,,3],abserrortar[,,,,4], xlab=paste("Target MAE for Truncation",trunca[3]), ylab=paste("Target MAE for Truncation",trunca[4]), pch=20)

lines(c(0,1),c(0,1), col="red") }
```

References

[1] Renewable UK. Wind: State of the Industry 2011. 2011 [cited; Available from: http://www.renewableuk.com/en/publications/index.cfm/Wind-SOI-2011]

[2] EWEA European Wind Energy Association. Wind in Power 2014 European Statistics. 2015 [cited; Available from: http://www.ewea.org/fileadmin/files/library/publications/statistics/EWEA-Annual-Statistics-2014.pdf]

[3] Renewable UK. Why Renewables Matter. 2015 [cited; Available from: http://www.renewableuk.com/en/publications/index.cfm/Why-Renewables-Matter]

[4] Scottish Renewables. Renewables in Numbers. 2015 [cited; Available from: https://www.scottishrenewables.com/sectors/renewables-in-numbers/]

[5] Früh WG. How much can regional aggregation of wind farms and smart grid demand management facilitate wind energy integration? World Renewable Energy Congress-XIII – "Renewable Energy in the Service of Mankind"; 2014; London, UK; 2014.

[6] Hernandez-Escobedo Q, Manzano-Agugliaro F, Gazquez-Parra JA, Zapata-Sierra A. Is the wind a periodical phenomenon? The case of Mexico. Renewable and Sustainable Energy Reviews 2011;15:721-728.

[7] Hernández-Escobedo Q, Manzano-Agugliaro F, Zapata-Sierra A. The wind power of Mexico. Renewable and Sustainable Energy Reviews 2010;14:2830-2840.

[8] Angelis-Dimakis A, Biberacher M, Dominguez J, Fiorese G, Gadocha S, Gnansounou E, et al. Methods and tools to evaluate the availability of renewable energy sources. Renewable and Sustainable Energy Reviews 2011;15:1182-1200.

[9] Weekes SM, Tomlin AS. Evaluation of a semi-empirical model for predicting the wind energy resource relevant to small-scale wind turbines. Renewable Energy 2013;50:280-288.

[10] Messac A, Chowdhury S, Zhang J. Characterizing and mitigating the wind resource-based uncertainty in farm performance. Journal of Turbulence 2012;13:1–26.

[11] Landberg L, Myllerup L, Rathmann O, Petersen EL, Jorgensen BH, Badger J, et al. Wind resource estimation - An overview. Wind Energy 2003;6:261-271.

[12] Clive PJM. Non-linearity in MCP with Weibull distributed wind speeds Wind Engineering 2008;32:319-324.

[13] National Center for Atmospheric Research (NCAR). [cited; Available from: http://rda.ucar.edu/]

[14] Watson SJ. Quantifying the variability of wind energy. Wiley Interdisciplinary Reviews: Energy and Environment 2013;3:330-342.

[15] Bechrakis DA, Deane JP, McKeogh EJ. Wind resource assessment of an area using short term data correlated to a long term data set. Solar Energy 2004;76:725-732.

[16] Velázquez S, Carta JA, Matías JM. Comparison between ANNs and linear MCP algorithms in the long-term estimation of the cost per kWh produced by a wind turbine at a candidate site: A case study in the Canary Islands. Applied Energy 2011;88:3869-3881.

[17] Met Office. Small-scale wind energy - technical report. 2008 [cited; Available from: https://www.carbontrust.com/media/85174/small-scale-wind-energy-technical-report.pdf]

[18] Carta JA, Velázquez S, Cabrera P. A review of measure-correlate-predict (MCP) methods used to estimate long-term wind characteristics at a target site. Renewable and Sustainable Energy Reviews 2013;27:362-400.

[19] Justus CG, Mani K, Mikhail AS. Interannual and Month-to-Month Variations of Wind Speed. Journal of Applied Meteorology 1979;18:913-920.

[20] Gerdes G, Strack M. Long-term Correlation of Wind Measurement Data. DEWI Magazin. 1999.

[21] Ramsdell JV, Houston S, Wegley HL. Measurement strategies for estimating long-term average wind speeds. Solar Energy 1980;25:495-503.

[22] Bueno C, Carta JA. Wind powered pumped hydro storage systems, a means of increasing the penetration of renewable energy in the Canary Islands. Renewable and Sustainable Energy Reviews 2006;10:312-340.

[23] Brower MC. Wind resource assessment, 1st ed. New Jersey: Wiley; 2012.

[24] Lackner MA, Rogers AL, Manwell JF. Uncertainty analysis in MCP-Based wind resource assessment and energy production estimation. Journal of Solar Energy Engineering-Transactions of the Asme 2008;130.

[25] Hiester TR, Pennell WT. The siting handbook for large wind energy systems, 1st ed. New York: WindBook; 1981.

[26] Baker RW, Walker SN, Wade JE. Annual and seasonal variations in mean wind speed and wind turbine energy production. Solar Energy 1990;45:285-289.

[27] Burton T, Sharpe D, Jenkins N, Bossanyi E. Wind energy handbook, 1st ed. West Sussex: John Wiley & Sons; 2001.

[28] Albers A, Klug H. High Quality Wind Speed Measurements for Site Assessment. DEWI Magazin. 1999.

[29] Lo S-F, Wu C-Y. Evaluating the performance of wind farms in China: An empirical review. International Journal of Electrical Power & Energy Systems 2015;69:58-66.

[30] Früh WG. Long-term wind resource and uncertainty estimation using wind records from Scotland as example. Renewable Energy 2013;50:1014-1026.

[31] The R project for statistical computing. [cited; Available from: http://www.r-project.org/]

[32] Dinler A. A new low-correlation MCP (measure-correlate-predict) method for wind energy forecasting. Energy 2013;63:152-160.

[33] Weekes SM, Tomlin AS. Data efficient measure-correlate-predict approaches to wind resource assessment for small-scale wind energy. Renewable Energy 2014;63:162-171.

[34] Klyatis LM, Teskin OI, Fulton JW, Iest. Multi-variate Weibull model for predicting system-reliability, from testing results of the components. Annual Reliability and Maintainability Symposium - 2000 Proceedings. New York: I E E E; 2000. p. 144-149.

[35] Woods JC, Watson SJ. A new matrix method of predicting long-term wind roses with MCP. Journal of Wind Engineering and Industrial Aerodynamics 1997;66:85-94.

[36] Rogers AL, Rogers JW, Manwell JF. Comparison of the performance of four measure–correlate–predict algorithms. Journal of Wind Engineering and Industrial Aerodynamics 2005;93:243-264.

[37] Derrick A. Development of the measure-correlate-predict strategy for site assessment. EWEC; 1993; 1993.

[38] Carta JA, Velázquez S. A new probabilistic method to estimate the long-term wind speed characteristics at a potential wind energy conversion site. Energy 2011;36:2671-2685.

[39] Weekes SM, Tomlin AS. Comparison between the bivariate Weibull probability approach and linear regression for assessment of the long-term wind energy resource using MCP. Renewable Energy 2014;68:529-539.

[40] Perea AR, Amezcua J, Probst O. Validation of three new measure-correlate-predict models for the long-term prospection of the wind resource. Journal of Renewable and Sustainable Energy 2011;3:20.

[41] Saavedra-Moreno B, Salcedo-Sanz S, Carro-Calvo L, Gascón-Moreno J, Jiménez-Fernández S, Prieto L. Very fast training neural-computation techniques for real measure-correlate-predict wind operations in wind farms. Journal of Wind Engineering and Industrial Aerodynamics 2013;116:49-60.

[42] Zhang J, Chowdhury S, Messac A, Hodge B-M. Assessing Long-Term Wind Conditions by Combining Different Measure-Correlate-Predict Algorithms. ASME, International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. Portland, Oregon; 2013.

[43] Probst O, Cardenas D. State of the Art and Trends in Wind Resource Assessment. Energies 2010;3:1087-1141.

[44] Mortimer A. A new correlation /prediction method for potential wind farm sites. BWEA; 1994; 1994.

[45] Wind Atlas and Application Program (WAsP). [cited; Available from: http://www.wasp.dk]

[46] García-Rojo R. Algorithm for the estimation of the long-term wind climate at a meteorological mast using a joint probabilistic approach. Wind Engineering 2004;28:213-224.

[47] Mardia KV. Families of bivariate distributions, 1st ed. USA: Lubrecht & Cramer Ltd.; 1970.

[48] Breslow PB, Sailor DJ. Vulnerability of wind power resources to climate change in the continental United States. Renewable Energy 2002;27:585-598.

[49] Watson K, Hodgson. Wind speed variability across the UK between 1957 and 2011. Wind Energy 2015;18:21-42.

[50] Burch SF, Newton M, Ravenscroft F, Whittaker J. Computer Modelling of the UK Wind Resource; Final Overview Report: Harwell: Energy Technology Support Unit.; 1992.

[51] Sexton DMH, Murphy J. UKCP09: probabilistic projections of wind speed. Exeter, U.K.; 2010.

[52] WindSim [cited; Available from: https://www.windsim.com/]

[53] Zephy Science, ZephyTOOLS. [cited; Available from: http://www.zephy-science.com/]

[54] Megajoule, Windiee. [cited; Available from: http://www.megajoule.pt/index.php]

[55] DNVGL-Garrad-Hassan, WindFarmer. [cited; Available from: http://www.gl-garradhassan.com/en/software/windfarmer/16206.php]

[56] EMD International A/S, WindPRO. [cited; Available from: http://www.emd.dk/windpro/]

[57] AWS Truepower, openWIND. [cited; Available from: http://www.awsopenwind.org/]

[58] European Centre for Medium-range Weather Forecasting (ECMWF). [cited; Available from: http://www.ecmwf.int]

[59] MERRA: Modern-Era Retrospective Analysis For Research and Applications. [cited; Available from: http://gmao.gsfc.nasa.gov/research/merra/]

[60] Virtual Met Mast (VMM). [cited; Available from: http://www.metoffice.gov.uk/renewables/vmm]

[61] Mengelkamp H-T, Kapitza H, Pflüger U. Statistical-dynamical downscaling of wind climatologies. Journal of Wind Engineering and Industrial Aerodynamics 1997;67–68:449-457.

[62] Chávez-Arroyo R, Lozano-Galiana S, Sanz-Rodrigo J, Probst O. On the Application of Principal Component Analysis for Accurate Statistical-dynamical Downscaling of Wind Fields. Energy Procedia 2013;40:67-76.

[63] Frank HP, Rathmann O, Mortensen NG, Landberg L, Landberg. The Numerical Wind Atlas-the KAMM/WAsP Method. Roskilde, Denmark; June, 2011.

[64] MM5 (Fifth-Generation Penn State/NCAR Mesoscale Model). [cited; Available from: http://www.mmm.ucar.edu/mm5/]

[65] Bowen AJ, Mortensen NG. Exploring the limits of WAsP the Wind Atlas Analysis and Application program. European Union Wind Energy Conference. Gotenborg, Sweden; 1996.

[66] Jimenez B, Durante F, Lange B, Kreutzer T, Tambke J. Offshore Wind Resource Assessment with WAsP and MM5: Comparative Study for the German Bight. Wind Energy 2007:121–134.

[67] Walmsley JL, Troen IB, Lalas DP, Mason PJ. Surface-layer flow in complex terrain: comparison of models and full-scale observations. Boundary-Layer Meteorology 1990;52:259–281.

[68] Quine CPW, I. M. S. Using the relationship between rate of tatter and topographic variables to predictsite windiness in upland Britain. . Forestry 1994;67:245–256.

[69] Suárez JC, Gardiner BA, Quine CP. A comparison of three methods for predicting wind speeds in complex forested terrain. Meteorological Applications 1999;6:329–342.

[70] ReSoft WindFarm. [cited; Available from: http://www.resoft.co.uk/English/index.htm]

[71] SgurrEnergy Ltd. MCP Comparison: Discrete and Linear. Internal report. Glasgow, UK; 2008.

[72] Frandsen S, Christensen CJ. Accuracy of estimation of energy production from wind power plants. Wind Engineering 1992;16:257-268.

[73] Jung S, Arda Vanli O, Kwon S-D. Wind energy potential assessment considering the uncertainties due to limited data. Applied Energy 2013;102:1492-1503.

[74] Gass V, Strauss F, Schmidt J, Schmid E. Assessing the effect of wind power uncertainty on profitability. Renewable and Sustainable Energy Reviews 2011;15:2677-2683.

[75] British Atmospheric Data Centre (BADC). [cited; Available from: http://www.badc.ac.uk]

[76] Rodrigo JS. State-of-the-Art of Wind Resource Assessment. 2010 [cited; Available from: http://www.waudit-itn.eu/download.php?id=103&parent=79]

[77] Brown University. A tutorial on learning dynamical systems. [cited; Available from: http://www.cs.brown.edu/research/ai/dynamics/tutorial/Documents/DynamicalSystems]

[78] Kantz H, Schreiber T. Nonlinear time series analysis, 2nd ed: Cambridge University Press; 2004.

[79] Lorenz EN. Deterministic Nonperiodic Flow. Journal of the Atmospheric Sciences 1963;20:130-141.

[80] Takens F. Detecting strange attractors in turbulence. Dynamical systems and turbulence, Warwick, 1980: lecture notes in Mathematics 898. Berlin: Springer-Verlag; 1981.

[81] Read PL. Phase portrait reconstruction using multivariate singular systems analysis. Physica D: Nonlinear Phenomena 1993;69:353-365.

[82] Everitt B, Dunn G. Applied multivariate data analysis, 2nd ed: Wiley; 2001.

[83] Palmer NT. A nonlinear dynamical perspective on climate change. Weather 1993;48:314-326.

[84] Bakalian F, Ritchie H, Thompson K, Merryfield W. Exploring Atmosphere–Ocean Coupling Using Principal Component and Redundancy Analysis. Journal of Climate 2010;23:4926-4943.

[85] Allen MR, Smith LA. Investigating the origins and significance of low frequency models of climate variability. Geophysical Research Letters 1994;21:883-886.

[86] Broomhead DS, Jones R, King JP, Pike ER. Singular Systems Analysis with application to dynamical systems: Adam Hilger; 1987.

[87] Golyandina N, Zhaglijavsky A. Singular spectrum analysis for time series. 2013.

[88] Früh WG. Methods to describe barotropic vortices by global fields and vortex characteristics. Nonlinear Processes in Geophysics 2002;9:189-200.

[89] Allen MR, Smith LA. Optimal filtering in singular spectrum analysis. Physics Letters A 1997;234:419-428.

[90] Allen MR, Read PL, Smith LA. Temperature time-series? 1992;355.

[91] Allen MR, Mutlow CT, Blumberg GMC, Christy JR, McNider RT, Llewellyn-Jones DT. Global change detection. Nature 1994;370:24-25.

[92] Hardy DM, Walton JJ. Principal Components Analysis of Vector Wind Measurements. Journal of Applied Meteorology 1978;17:1153-1162.

[93] Benestad RE. A comparison between two empirical downscaling strategies. International Journal of Climatology 2001;21:1645–1668.

[94] Dreveton C, Guillou Y. Use of a Principal Components Analysis for the Generation of Daily Time Series. Journal of Applied Meteorology 2004;43:984-996.

[95] Martinez Y, Yu W, Lin H. A New Statistical–Dynamical Downscaling Procedure Based on EOF Analysis for Regional Time Series Generation. Journal of Applied Meteorology and Climatology 2012;52:935-952.

[96] Foley AM, Leahy PG, Marvuglia A, McKeogh EJ. Current methods and advances in forecasting of wind power generation. Renewable Energy 2012;37:1-8.

[97] Skittides C, Früh W-G. Wind forecasting using Principal Component Analysis. Renewable Energy 2014;69:365-374.

[98] Monteiro C, Bessa R, Miranda V, Botterud A, Wang J, Conzelmann G. Wind Power Forecasting:State-of-the-Art 2009. 2009 [cited; Available from: http://www.ipd.anl.gov/anlpubs/2009/11/65613.pdf]

[99] Giebel G, Brownsword R, Kariniotakis G, Denhard M, Draxl C. The state-of-art in short-term prediction of wind power, a literature overview. 2011 [cited; 2nd:[Available from: http://orbit.dtu.dk/fedora/objects/orbit:83397/datastreams/file_5277161/content]]

[100] Wang X, Guo P, Huang X. A Review of Wind Power Forecasting Models. Energy Procedia 2011;12:770-778.

[101] Zhou J, Shi J, Li G. Fine tuning support vector machines for short-term wind speed forecasting. Energy Conversion and Management 2011;52:1990-1998.

[102] Erdem E, Shi J. ARMA based approaches for forecasting the tuple of wind speed and direction. Applied Energy 2011;88:1405-1414.

[103] De Giorgi M, Ficarella A, Tarantino M. Error analysis of short term wind power prediction models. Applied Energy 2011;88:1298-1311.

[104] De Giorgi M, Ficarella A, Tarantino M. Assessment of the benefits of numerical weather predictions in wind power forecasting based on statistical methods. Energy 2011;36:3968-3978.

[105] Fruh WG. Evaluation of simple wind power forecasting methods applied to a long-term wind record from Scotland. International Conference on Renewable Energies and Power Quality (ICREPQ'12). Santiago de Compostela, Spain; 2012.

[106] Shi J, Guo J, Zheng S. Evaluation of hybrid forecasting approaches for wind speed and power generation time series. Renewable and Sustainable Energy Reviews 2012;16:3471-3480.

[107] Madsen H, Pinson P, Kariniotakis G, Nielsen HA, Nielsen TS. Standardizing the performance evaluation of short-term wind power prediction models. Wind Engineering 2005;29:475–489.

[108] Van Der Hoven I. Power spectrum of horizontal wind speed in the frequency range from 0.0007 to 900 cycles per hour. Journal of Climatology 1956; 14:160-164.