Spatio-temporal Prediction of Wind Fields

PhD Thesis

Jethro Dowell

Wind Energy Systems Centre for Doctoral Training
Department of Electronic and Electrical Engineering
University of Strathclyde, Glasgow

September 18, 2015

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Abstract

Short-term wind and wind power forecasts are required for the reliable and economic operation of power systems with significant wind power penetration. This thesis presents new statistical techniques for producing forecasts at multiple locations using spatio-temporal information. Forecast horizons of up to 6 hours are considered for which statistical methods outperform physical models in general. Several methods for producing hourly wind speed and direction forecasts from 1 to 6 hours ahead are presented in addition to a method for producing five-minute-ahead probabilistic wind power forecasts. The former have applications in areas such as energy trading and defining reserve requirements, and the latter in power system balancing and wind farm control.

Spatio-temporal information is captured by vector autoregressive (VAR) models that incorporate wind direction by modelling the wind time series using complex numbers. In a further development, the VAR coefficients are replaced with coefficient functions in order to capture the dependence of the predictor on external variables, such as the time of year or wind direction. The complex-valued approach is found to produce accurate speed predictions, and the conditional predictors offer improved performance with little additional computational cost.

Two non-linear algorithms have been developed for wind forecasting. In the first, the predictor is derived from an ensemble of particle swarm optimised candidate solutions. This approach is low cost and requires very little training data but fails to capitalise on spatial information. The second approach uses kernelised forms of popular linear algorithms which are shown to produce more accurate forecasts than their linear equivalents for multi-step-ahead prediction.

Finally, very-short-term wind power forecasting is considered. Five-minute-ahead

parametric probabilistic forecasts are produced by modelling the predictive distribution as logit-normal and forecasting its parameters using a sparse-VAR (sVAR) approach. Development of the sVAR is motivated by the desire to produce forecasts on a large spatial scale, i.e. hundreds of locations, which is critical during periods of high instantaneous wind penetration.

Contents

List of Figures

List of Tables

Abbreviations and Mathematical Symbols

Abbriviations

AEMO	Australian Electricity Market Operator.
AIC	Akaike Information Criterion.
AR	Autoregressive.
ARCH	Autoregressive Conditional Heteroscedasticity.
ARIMA	Autoregressive Integral Moving Average.
ARMA	Autoregressive Moving Average.
BIC	Bayesian Information Criterion.
CDR	Continuous Directional Regime.
CRPS	Continuous Rank Probability Score.
CSWF	Cyclo-Stationary Wiener Filter.
EPSO	Ensemble Particle Swarm Optimisation.
FIR	Finite Impulse Response.
GARCH	Generalised Autoregressive Conditional Heteroscedasticity.
i.i.d	Independent and Identically Distributed.
LMS	Least Mean Squares.
MAE	Mean Absolute Error.
MLE	Maximum Likelihood Estimation.

MMSE	Minimum Mean Squared Error.
MOS	Model Output Statistics.
O&M	Operations and Maintenance.
PSC	Partial Spectral Coherence.
PSO	Particle Swarm Optimisation.
R&D	Research and Development.
RKHS	Reproducing Kernel Hilbert Space.
RLS	Recurrent Least Squares.
RMSE	Root Mean Squared Error.
sVAR	Sparse Vector Autoregressive.
VAR	Vector Autoregressive.

Mathematical Symbols and Notation

$\mathbb{N},\ \mathbb{R},\ \mathbb{C},\ \mathbb{Z}$	Set of natural, real, complex, integer numbers.
$\mathbb{R}^N,\ \mathbb{R}^{N\times M}$	Set of real valued vectors of length N, set of real valued $N \times M$ matrices.
$x,\ \boldsymbol{x},\ \mathbf{X}$	Scalar, vector, matrix.
$X \sim N(\mu, \sigma^2)$	X is normally distributed with mean μ and variance σ^2.
$X \sim U(a, b)$	X is uniformly distributed over the interval $[a, b]$.
$E\{\cdot\}$	Expectation operator.
$\text{trace}\{\cdot\}$	Trace operator, returns the sum of the diagonal elements of a square matrix.
$\arg\min_x\{f(x)\}$	The value of x that minimises $f(x)$.
$\lvert x\rvert,\ \lvert\lvert\boldsymbol{x}\rvert\rvert$	Absolute value of x, Euclidean norm of $\boldsymbol{x}$.
$\text{Re}\ x,\ \text{Im}\ x$	The real part of x, the imaginary part of x.
$\boldsymbol{x}^*$	Complex conjugate of $\boldsymbol{x}$.
$\boldsymbol{x}^{\text{T}}$	Transpose of $\boldsymbol{x}$.
$\boldsymbol{x}^{\text{H}}$	Hermitian (conjugate transpose) of $\boldsymbol{x}$.
$\mathbf{R}_{\boldsymbol{xx}}$	Covariance matrix of $\boldsymbol{x}$.
$\mathbf{R}_{\boldsymbol{xy}}$	Cross-covariance matrix of $\boldsymbol{x}$ and $\boldsymbol{y}$.
$\tilde{\mathbf{R}}_{\boldsymbol{xx}}$ and $\underline{\mathbf{R}}_{\boldsymbol{xx}}$	Complimentary and augmented covariance matrix of $\boldsymbol{x}$.
$x(t)$	The value of x at continuous time $t \in \mathbb{R}$.
$x[t]$ or x_t	The value of x at discrete time index $t \in \mathbb{Z}$.
τ	General lag parameter.
$x_{t\lvert t-\tau}$	The estimate of x_t based on information available at time $t - \tau$.
Δ	Specific prediction horizon/look-ahead time.

Preface

Since the industrial revolution in the first half of the 19^{th} century, demand for energy to power high-tech societies and lifestyles has increased exponentially. That demand has, to date, largely been met by burning fossil fuels, a by-product of which is the emission of carbon dioxide into the atmosphere. At present, the daily release of over 100 million tones of this invisible gas goes largely unnoticed, while the lives of many have never been more comfortable thanks to the abundance of on-demand energy and derived products. In 2011 cumulate anthropogenic CO_2 emissions reached over 2000Gt, half of which has been emitted since 1970. The concentration of CO_2 in the Earth's atmosphere is increasing, and recently passed 400ppm (parts per million), well above the 1850 level of 285ppm and the estimated *safe* upper limit of 350ppm. Emissions of CO_2 and other greenhouse gases are driving global climate change, the effects of which are beginning to be felt around the world.

Quoting from the 2014 Intergovernmental Panel on Climate Change synthesis report [1]: "Climate change will amplify existing risks and create new risks for natural and human systems. Risks are unevenly distributed and are generally greater for disadvantaged people and communities in countries at all levels of development." Those risks include increased frequency and duration of extreme weather events, ocean acidification, sea level rise, and increased/decreased precipitation depending on region. Large fractions of animal and plant species face extinction due to climate change. Food and water security are at risk without significant adaptation. Urban areas face increased risks to people, assets, economies and ecosystems, including risks from heat stress, storms and extreme precipitation, inland and coastal flooding, landslides, air pollution, drought, water scarcity, sea level rise and storm surges. Rural areas are expected to

experience major impacts on water availability and supply, food security, infrastructure and agricultural incomes, including shifts in the production areas of food and non-food crops around the world.

The need for action has never been more apparent.

In 1988 the UN and World Meteorological Organization established the Intergovernmental Panel on Climate Change to assess scientific information on all aspects of climate change and its impacts in order to formulate a realistic response. This lead to the adoption of the Kyoto Protocol in 1997, which set various targets for developed countries to reduce emissions. However, the 2009 UN climate summit in Copenhagen failed to produce any legally binding targets for global emission control and by the end of the first phase of the Kyoto Protocol in 2012, many large emitters had failed to ratify or removed themselves from the treaty, including the US, Canada, Russia and Japan. While many developed nations now have domestic emission targets, including the US and China, hopes for global commitment to curb greenhouse gas emissions rest with the 2015 climate summit in Paris later this year.

The European Union is one of the few original signatories of the Kyoto Protocol which has legally binding emissions reduction targets at present. The block has targets to reduce emissions by 20% compared to 1990 levels and to be generating 20% of electricity from renewables by 2020. Specific targets vary between member states depending on their ability to make reductions. The long-term goal for the EU is to reduce emissions by 80–95% by 2050.

The UK became the first country to set long-range carbon reduction targets in law with the Climate Change Act 2008. The Act outlines a framework for transitioning to a *low-carbon economy* requiring an 80% cut in carbon emissions by 2050 compared to 1990 levels. At present in the UK power generation accounts for around one quarter of greenhouse gas emissions. De-carbonising the UK power sector over the next three and a half decades will require a huge reduction in fossil fuel use and a large increase in renewable energy generation in combination with other low-carbon energy sources and energy efficiency measures.

The way energy systems are operated will have to change to accommodate high levels of variable, weather dependent renewable generation, and an important part of any solution will be forecasting.

Acknowledgements

I would first like to thank my supervisors Stephan Weiss and David Infield for their excellent tutelage, I have learned a tremendous about from them both. The Wind Energy Systems Centre for Doctoral Training has provided a stimulating and enjoyable place to work and for that I thank Drew Smith, Bill Leithead, David Infield, all the lecturers who taught the first year courses, and the other CDT students, particularly my 3rd cohort colleagues. Special thanks go to Pierre Pinson and the Technical University of Denmark for hosting and mentioning me during a 3 month visit at the beginning of 2014. The work presented in Chapter 5 of this thesis is the result of that fruitful visit and ongoing collaboration.

For their financial support I gratefully acknowledge the UK Engineering and Physical Sciences Research Council doctoral training centre grant number EP/G037728/1. In addition, for supporting my visit to the Technical University of Denmark, I thank COST Action ES 1002 "Weather Intelligence for Renewables".

For the provision of data, without which this work would not have been possible, I thank the UK Meteorological Office and the British Atmospheric Data Centre for their supply of the MIDAS dataset [2], the Royal Netherlands Meteorological Institute for their supply of the Hydra dataset [3], and the Australian Energy Market Operator and Stefanos Delikaraoglou for their supply of wind power data and its pre-processing [4].

Jethro Dowell

Glasgow, UK

September 18, 2015

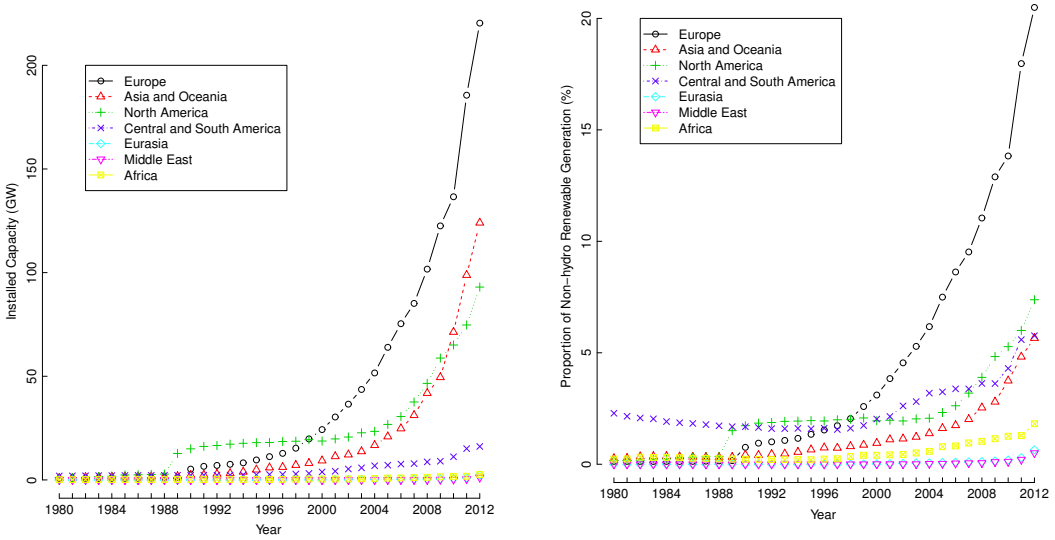
Chapter 1

Introduction

1.1 The Power System

Many of the world's power systems were developed over the past century: initially to power electric lighting, then to transmit electricity from a few large power stations to individual cities or industrial complexes, and later becoming increasingly interconnected eventually providing a reliable power supply on national and even continental scales. Unlike resources that can be easily stored, electricity supply must meet demand in real time. If more power is produced than consumed, the frequency of the AC power system increases, and vice versa. Even a small change in frequency is enough to damage synchronous machines and other equipment. Other limits on voltage, line and transformer capacity, reactive power and phase angle are imposed for similar reasons. Modern power systems are highly controlled and include protection systems to maintain safe operation and protect equipment.

Today, power system operators act in conjunction with electricity markets to provide secure and economic supply. Electricity networks form natural monopolies which were traditionally operated by vertically integrated public companies that generated, transmitted and distributed electricity to consumers. However, the liberalization of electricity markets, beginning in the UK in the 1980s, has seen the vertical disintegration and privatisation of the electricity industry, and the creation of new markets for energy, ancillary services and capacity [5].



(a) Global Non-hydro Renewable Capacity.

(b) Proportion of Non-hydro Renewable Generation.

Figure 1.1: Total capacity (a) and proportion of capacity (b) of global non-hydro renewable power by continent. Source: US Energy Information Administration.

During the same period, the mix of generation technologies began to change. Electricity generation has been dominated by large synchronous machines, driven by thermal power stations since the early 20^{th} century, and power systems have been designed to accommodate them: high voltage transmission systems carry power form large power stations to load centres, where it is distributed to customers at a lower voltage. However, the beginning of the 21^{st} century has seen the rapid growth of renewable electricity generation in developed countries motivated by the threat of climate change [6, 7], and the desire of states to reduce reliance on energy imports [8]. The growth of renewable generation capacity is illustrated in Figure 1.1, and wind capacity in Figure 1.2.

Weather dependent renewable generation such as wind and solar power are variable and often spread over large geographical areas, connecting to power systems at the distribution level. The rise in so-called *distributed generation* poses a challenge to power systems that were built for large synchronous machines connected close to load centres. Electricity markets and power system operators are having to adapt to ever increasing penetration of variable generation, and one key component of that transition is forecasting variable generation [9].

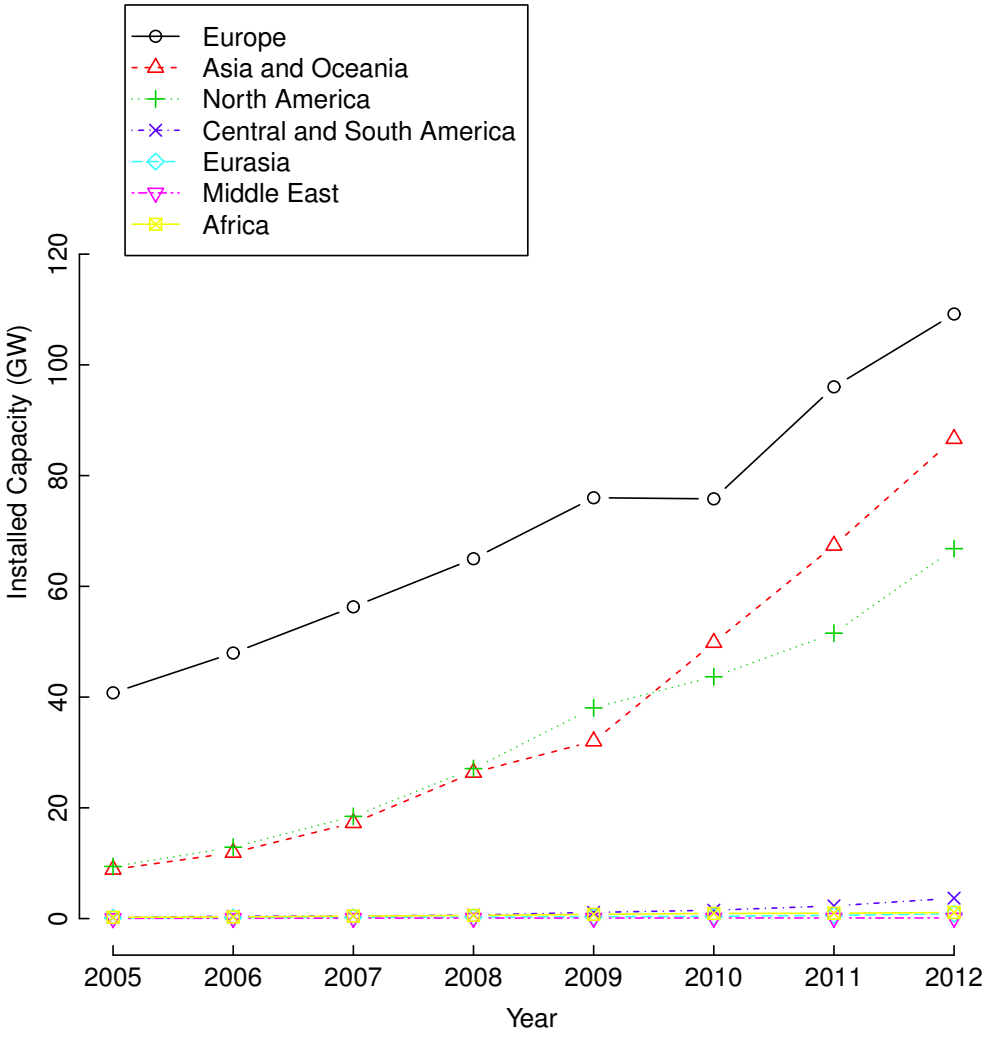


Figure 1.2: Global wind power capacity by continent. Source: US Energy Information Administration.

1.2 Wind Power Forecasting

Given that the wind, and therefore the power generated by wind turbines, is variable, and that electricity supply must meet demand in real time, the need for wind power forecasts is clear [10]. Furthermore, since the day-to-day running of today's liberalised energy industry is marketised, all participants in the industry are exposed to the effects of increasing the penetration of variable generation. With financial penalties for over-/under-delivering on generation, and repercussions for electricity price and balancing costs, forecasting is critical to economic operations, as well as technical [9]. Wind farm developers and operators also have an interest in future production to minimise lost energy capture when performing maintenance, protect assets against extreme weather events and identify locations with an abundant wind resource [11].

Power Systems

To optimally utilise variable renewable generation, such as wind power, power systems and the way they are operated are changing: transmission networks must connect distant renewable generation to load centres, distribution networks must accommodate small and medium scale generation, and operators must consider the stochastic nature of this new variable generation when performing scheduling tasks. Many decisions relating to power system operation are increasingly informed by forecasts on a variety of temporal and spatial scales, and the upper limit on the level of variable generation that can be accommodated by a given power system will ultimately be set by the skill of these forecasts, and their users. A recent survey of US power system operators identifies the growing importance of forecasting for reliable grid operations, with one of the key findings being that "wind power forecast[ing] is the most important pre-requisite for successfully integrating wind energy into power systems" [12].

More specifically, integrating wind power and maintaining security of supply requires careful management of transmission constraints and scheduling of conventional generation, which to be done most efficiently, requires accurate intra-day and day-ahead forecasts [13–16]. Furthermore, it is well established that moving to a probabilistic approach is of as much benefit as moving from naive to advanced point forecasting [17,18].

Reducing the requirement for fast-responding back-up generation is critical in realising the maximum de-carbonisation wind power offers and requires skilful forecasts and decision-making [19–21].

Electricity Markets

Electricity markets were designed for dealing with mainly dispatchable generation and fairly predictable demand allowing for extensive forward contracting accompanied by a real-time mechanism to facilitate power system operation. As recently as a decade ago their future evolution in many developed countries was expected to remain in this paradigm, as demonstrated by a 2005 paper describing US electricity markets and their future evolution that includes no mention of the potential role of renewable generation [22]. Meanwhile, Denmark was learning how to operate liberalised power markets with high volumes of wind power, occasionally approaching 100% instantaneous penetration [23].

Today, with many governments committed to reducing CO_2 emissions, electricity markets in developed countries are increasingly having to operate with, and plan for, high renewable energy penetration. Participants in existing markets rely on forecasts to make optimal trading decisions, while new market structures are being proposed to address some of the failings of markets designed for conventional generation [24–28].

Operations and Maintenance

Finally, maintenance costs contribute a significant portion to the cost of energy over the lifetime of a wind farm. Onshore, operations and maintenance (O&M) costs typically make up around 5% of the cost of energy, whereas offshore the figure can be much higher, from 20% to 30% or more, depending on the distance of the farm from shore. Wind forecasts allow non-essential maintenance to be scheduled to minimise lost energy capture onshore, and are essential for scheduling maintenance offshore where safety constraints on vessel operation and crew transfer are very restrictive.

1.3 Objective of Research

It is the objective of this research to develop new prediction techniques for application to short- and very-short-term wind power forecasting. Forecasts on this time scale are typically made using recent measurements as an input to a statistical model. Numerous such models are described in the literature, each with its own merits. However, spatial techniques, where measurements made at multiple locations are used as inputs, are underdeveloped and have many attractive benefits. Capturing spatial correlation has been shown to improve forecast skill in small scale studies and is here expanded and generalised to national-scale forecasting problems. Furthermore, by including wind direction, which can have a large influence on wind farm power generation and is often overlooked, it is believed that significant improvements in short-term wind power forecasting can be made.

Spatial models may also be built to directly forecast power production removing the need to model wind farm power curves. This is investigated in conjunction with a method for producing very-short-term forecasts with a much higher spatial dimension, a problem facing power systems operators with very high wind penetration.

Chapter 2

State-of-the-Art in Wind Power Forecasting

The history of wind power forecasting can be traced back to the late 1970s when it was identified as a key requirement for operating large scale wind power plants [29]. A good example of early work is by Brown *et al.* [30] who used wind speed forecasts and a wind turbine power curve to produce forecasts, published in 1984*. Brown identified the need to understand how wind might contribute to future *'multisource'* energy networks, and recognised at this early stage that "once a wind power generator is supplying power to an energy system, a method of forecasting wind power a few hours in advance is required to ensure efficient utilization of the power." Over the following 30 years research activity in this area has expanded, most significantly since the early 2000s, as wind power has been adopted around the world.

Wind power forecasting is regarded as a high priority research area that is expected to reduce energy and power system running costs, and improve power system reliability [12, 32]. The International Energy Agency highlights advances in forecasting in its 2013 technology roadmap [33] using Spain as an example where forecast errors from 1 to 48 hours ahead have reduced significantly between 2008 and 2013. However, it goes on to stress the importance of further research and development (R&D) in short-term forecasting saying: "Improving the accuracy of short-term wind forecast is needed for

*In 1984 the first European Wind Energy Conference was held, Vestas began serial production of a 75kW turbine, and California had installed 8469 turbines with a combined capacity of 609MW [31].

the operation of wind power plants, especially for electricity markets and the power system." As a result, both academic and commercial institutions are investing in forecasting R&D and the state-of-the-art is advancing rapidly. Energy forecasting more generally has grown into a broad and fast moving research area featured in many international conferences and publications. In 2012, point wind power and load forecasting challenges comprised the first Global Energy Forecasting Competition [34], and attracted a large number of entries. The 2014 competition expanded to include solar power and electricity price forecasting, and probabilistic forecasts.

The European Commission funded research project ANEMOS.plus—"Advanced Tools for the Management of Electricity Grids with Large-Scale Wind Generation" in partnership with the SafeWind project produced a broad literature review of the state-of-the-art in short term prediction of wind power in 2011, containing some 386 references, which serves as a starting point for this review [35]. Another extensive review was produced by the Argonne National Laboratory in 2009 [36]. A more compact review of short-term wind speed forecasting for power system operations is presented in [37, 38]. A brief history of the field is available in [39].

This thesis is primarily concerned with short-term statistical prediction, though an overview of longer-term and physical methods is offered since there is a degree of overlap.

2.1 Basis of the Forecasting Problem

2.1.1 Nature of Wind Power Generation

Before proceeding to prediction, it is important to understand the nature of the quantity that we wish to predict, and the inherent limitations of the problem. The type of measurements being considered are of importance: wind speed may be recorded at a single moment in time or averaged over a short period, 10 minute and 1 hour means are typical for meteorological records [40]; the height at which the measurements are made and local geography will influence the characteristics of the measured data; and the reliability and accuracy of the measurement equipment will impact on the predictability

of the resulting time series [41–43].

Long-term Wind Characteristics

The Weibull distribution [44] is considered the standard for describing wind speed over long periods of time (typically one or more years) [45], and sites are often characterised by the parameters of a Weibull distribution. The European Wind Atlas [46], for example, provides estimates of the wind resource across Europe in terms of the parameters of the Weibull distribution. The two parameter probability density function of a Weibull random variable x is given by

$$f(x;k,\lambda) = \frac{k}{\lambda}\Big(\frac{x}{\lambda}\Big)^{k-1} e^{-(\frac{x}{\lambda})^k} \qquad x > 0 \tag{2.1}$$

where $k > 0$ is the shape factor and $\lambda > 0$ is the scale factor.

However, other distributions, such as the Gamma distribution, may provide a better fit at some locations and others have been proposed more recently, such as the M-Rice wind speed frequency distribution [47]. Directional information is important to consider, particularly in situations where the layout of a wind farm means that the power produced depends on both wind speed and direction due to wake interactions within the farm or complex terrain. A number of bivariate distributions for characterising both wind speed and direction are compared in [48], and similarly analysed in [49].

In the special case of $k = 2$, the Weibull distribution reduces to the Rayleigh distribution. This result provides a link to directional representations of the wind since the Rayleigh distribution describes how the Euclidean norm of two perpendicular i.i.d. Gaussian variables is distributed. This is demonstrated in Appendix B.1. This result has been used by some authors to model wind speed and direction as perpendicular components in Cartesian space which has the pleasing result of supporting Gaussian processing while maintaining a representative marginal distribution of wind speed.

Variations in wind speed are observed on a variety of scales, as illustrated by the van der Hoven spectrum [50] in Figure 2.1. The spectrum can be divided into three regions: the macro-meteorological range, the spectral gap, and the micro-meteorological range. The macro-scale includes annual variation, synoptic variation (passing weather systems

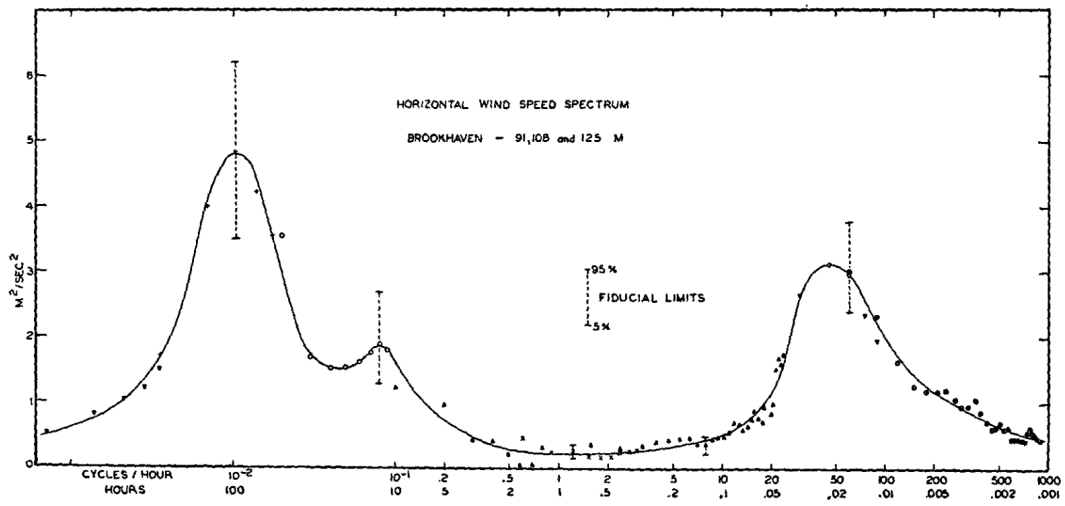


Figure 2.1: Van der Hoven spectrum from the original 1957 paper showing (from left to right) the synoptic, diurnal and turbulent peaks. Measurements were made at Brookhaven National Laboratory at roughly 100m above ground level.

with a frequency of approximately 4 days), and diurnal variation. The micro-scale contains the *turbulent peak* with a frequency of around 1 minute. There is very little energy in the spectral gap, the region between 10 minutes and 2 hours, which is often used implicitly to separate diurnal and higher-frequency fluctuations of turbulence.

What is Wind?

Wind, in its simplest sense, is the motion of gas particles that comprise our planet's atmosphere; so we could define the wind speed as the *speed* of those particles, in what ever direction they happen to be travelling. This, however, would be incredibly difficult to measure and not of very much use for our needs: a wind turbine with a rotor diameter of 100m in 10ms^{-1} wind could interact with of the order of 10^{30} particles per second that would neither be travelling at the same speed nor in the same direction.

A different approach is required. As we are motivated by the power produced by wind turbines, we need only concern ourselves with a wind speed (and later direction) representative of that which is seen by the wind turbine rotor. Furthermore, in most cases, it is the behaviour of large groups of wind turbines generating a significant amount of power that are of interest. Therefore, the high frequency (order $\leq$ min-

utes) changes in power output from individual turbines can be negated since they are smoothed out when considering the aggregate power of a wind farm, or group of wind farms.

Now that we have an idea of what wind speed might be, the practicality of measuring it requires some attention. The work-horse of wind measurements for a number of decades has been the cup anemometer. While other measurement devices are beginning to be used, LIDAR and sonic anemometers for example, the vast majority of current weather stations and met masts are equipped with cup anemometers and almost all historic data sets comprise cup anemometer measurements. A number of standards exist describing the procedures for calibrating and installing cup anemometers: ASTM D 5096–02, ISO 17713–1, and IEC 61400–12–1. The latter refers specifically to assessing the power performance of wind turbines. A review of these and other anemometry standards is offered in [51].

Cup anemometers have the advantage of being robust and relatively cheap; however, they suffer from a delayed response to changes in wind speed, and respond quicker to increases in wind speed than decreases. This results in an *over-speeding effect* resulting in an overestimation of the wind speed. Sufficiently fast responding anemometers should be used for the resolution of measurements being made. Other concerns include the effect of a vertical wind component and response in cold weather/icing.

Common time scales that are relevant for application to power system operations are 1 minute, 10 minute and 1 hour mean power generation, and it is the short-term prediction of wind speed and wind power on these time scales that is considered in this work. If the measured wind speed is not made at turbine hub height, the wind shear must be estimated and a correction applied [45]. This must be done carefully as it can be a significant source of error that should be avoided.

Wind Power Conversion

The amount of power generated by a wind turbine depends on the power in the air flow incident on its rotor and the efficiency of the conversion process. This power is

calculated using the power equation

$$P = \frac{1}{2}\rho\pi R^2 v^3 C_p \quad , \tag{2.2}$$

where ρ is the density of the air, R is the rotor radius, v is the wind speed, and C_p is the power coefficient. The power coefficient is a measure of the aerodynamic efficiency of the turbine and has an upper limit, called the *Betz limit*, of $C_{p,\max} = \frac{16}{27} \approx 0.593$. This limit is set by the need to allow air from which energy has been extracted to move away from the rotor, making way for new high-energy air. Modern wind turbines can achieve aerodynamic efficiency close to 0.55, but after mechanical and electrical losses this drops to less than 0.5 at the output terminal [10, 45].

The cubic relationship with the wind speed only forms part of the full *power curve*. The power curve for a modern, utility scale, variable speed, pitch regulated turbine is made up of 4 parts: 1) below *cut-in* wind speed (typically $\sim$ 3 to 4ms^{-1}) where the turbine does not operate, 2) between cut-in and *rated* wind speed where the turbine is operated to maximise C_p and energy capture as in Equation (2.2), 3) above rated wind speed where the power is limited to the turbine's rated power, the rating of the generator, drive train and so on, 4) above some *cut-out* wind speed (typically $\sim$ 25ms^{-1}) the turbine is shut-down to prevent damage [45]. A typical power curve is sketched in Figure 2.2.

In reality, however, the relationship between wind speed and power is difficult to model because the conversion process is effected by many external factors such as mechanical wear, blade erosion, yaw misalignment, and poor quality wind speed measurements, among others. As a result, power curve models can introduce uncertainty when producing power forecasts and in some cases, particularly very-short-term forecasting, it may be preferable to model power alone to remove the need for a power curve model all together.

2.1.2 Types of Forecast

Wind and wind power forecasts come in a variety of forms to satisfy the end user. Different decisions are made on different spatial and temporal scales, and forecasts must

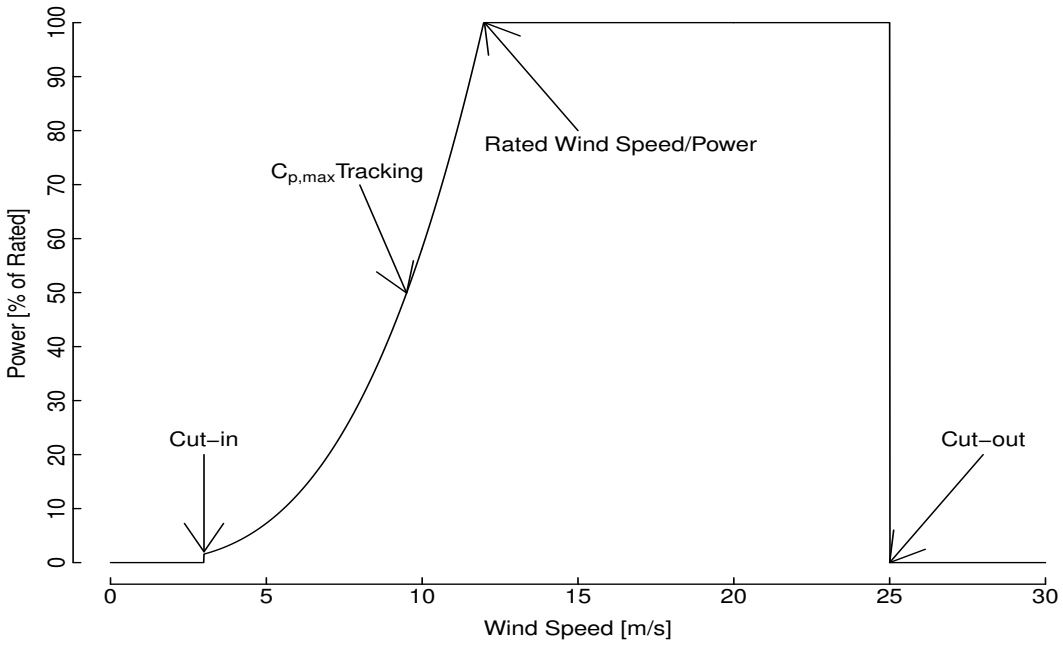


Figure 2.2: Typical power curve for a modern, utility scale, variable speed, pitch regulated wind turbine.

reflect this: distribution system operators may be concerned with individual wind farms connected to their system, a transmission system operator may be concerned with the aggregate wind power at different connection points; likewise, there is no need to know what the wind speed is for every minute of the next day when trading hourly periods of generation; similarly predicting the next hour's mean power does not help wind farm controllers. A brief list of forecast horizons and the decisions they may inform is given in Table 2.1.

Some decisions can benefit from probabilistic forecasts, such as interval or quantile forecasts, or from predictions of specific events, such as large changes in power output or crossing some threshold. These different types of forecast are described in the proceeding text.

Point Forecasts

Point forecasts are the simplest and most familiar type of forecast. They comprise a single prediction of some future observation, e.g. "the wind speed will be 10ms^{-1}

Table 2.1: Forecast horizons, common temporal resolutions and the decisions they inform.

	Forecast Horizon	Resolution	Decision
Ultra-short-term	< 1 minute	Seconds	Wind turbine control.
Very-short-term	<1 hour	1, 5, 10, 15 minutes	Balancing, wind farm control, some spot markets.
Short-term	1–48 hours	30 minutes, 1 hour	Generation scheduling, day-ahead markets, some spot markets.
Medium-term	1–10 days	1 hour, 3 hours	Generation scheduling, maintenance planning.
Long-term	months–years	Days–months	Maintenance planning, resource assessment/project financing.

one hour from now." Point forecasts, sometimes called deterministic forecasts, are favoured by many practitioners because of their ease of use: a non-expert can produce, communicate and interpret point forecasts with relative ease. Most media that provide weather forecasts for public consumption will offer a point forecasts for precisely this reason.

Probabilistic Forecasts

Point forecasts are inherently uncertain, and while they offer a 'best estimate' of some future quantity, they provide no information as to how confident one can be in that outcome being realised. Probabilistic forecasts offer more information than a point forecast by providing an estimate of the likelihood of a range of possible outcomes, information that is essential for optimal decision-making in many situations. Probabilistic forecasts are the optimal input to decision-making problems with non-symmetric cost functions. For example, if the penalty for a shortage of wind generation is different to the penalty for a surplus, then the optimal bid is not the expected power but a quantile [52].

Probabilistic forecasts come in a variety of forms: Quantile forecasts, for instance, estimate the probability that an observation will exceed some value, e.g. "there is a 90% chance that the wind speed will be greater than 5ms^{-1} one hour from now." Similarly, an interval forecast predicts the probability that an observation will fall within some interval. Information pertaining to the full range of possible outcomes is contained in a predictive distribution, where the full probability density function for a future observation is estimated, this may take the form of either a parametric or non-parametric distribution. Zhang *et al.* provide a more detailed review of these techniques in [53].

When multiple connected forecasts are required, such as the wind power generation at several wind farms in the same region, capturing dependence between observations is extremely important.* In these situations scenario forecasts capture both spatial structures and temporal structures necessary for multi-stage decision making problems.

*The 2008 financial crisis has been partly attributed to use of Gaussian copulas to calculate the risk associated with sub-prime mortgage derivatives that strictly assumed the independence of individual component mortgages, which were in fact deeply connected.

Over the past 20 years there has been a shift from deterministic to probabilistic forecasting in applications from economic and financial risk management to demographic and epidemiological projections [54]. The ability to quantify the confidence of a prediction is extremely valuable to decision makers and is now a common requirement of many forecasting tools, including those designed for wind power.

Ramp Forecasts

A large proportion of the errors in wind power forecasting are the result of incorrectly predicting large changes in power, called ramp events. In general the magnitude of such changes are forecast well, but the time at which they occur is not, resulting in large errors [38]. Increasingly, efforts are being made to better predict these events but they remain a significant challenge [55, 56]. Drew *et al.* have used reanalysis weather data to evaluate the effect of building the planned Round 3* offshore wind farms in the UK concluding that the magnitude of ramp events could increase 5-fold by 2025 [57].

Offshore Wind Power Forecasting

The properties of the wind in the offshore environment can be very different to those onshore. The reduced diurnal heating of the surface and the effect of low roughness over vast areas on the atmospheric boundary layer mean that the wind does not exhibit some properties which are familiar onshore [58, 59]. Therefore, authors have proposed methods specifically for offshore wind power forecasting, such as [60, 61]. Rogers *et al.* have produced a comparison of prediction accuracy on- and offshore in [62] concluding that the performance of offshore forecasting lies somewhere between onshore sites with simple terrain, which can be forecast with relatively high accuracy, and onshore sites in complex terrain that are more difficult to forecast.

*The Crown Estate, which owns the seabed around the UK, has leased areas of the seabed for wind farm development in 3 rounds, to date. Round 3 represents the largest areas with 24GW of capacity leased to developers. Construction of the first round 3 wind farm, Rampion Offshore Wind Farm, began in 2015.

2.2 Physical Models

Numerical weather prediction (NWP) forms the basis of most meteorological forecasts. NWP involves using observations to estimate the current state of the atmosphere and oceans in order to compute their future states. The atmospheric model is initialised and a set of linearised equations describing atmospheric physics, including the Navier-Stokes equation and ideal gas law, are solved on a 3-dimensional grid. Both the initialisation of atmospheric parameters and the linearisation of the governing equations are critical in producing meaningful forecasts.

A number of NWPs covering different regions of the planet are run in several countries around the world using measurements from weather satellites and radiosondes. Despite some of the most powerful supercomputers in the world being used for NWP, spatial and temporal resolution are limited. Weather forecasts are typically issued with forecast horizons of between 7 and 10 days; with spatial resolution ranging from 5km to 25km; and temporal resolution of either 1 or 3 hours. Longer term climate forecasts are made but at much lower resolution. Due to the vast computational expense of NWPs, forecasts are typically issued every 6 or 12 hours [63, 64].

In Europe, some national weather service providers run NWPs including the UK Met Office, Météo-France and Deutcher Wetterdienst, Germany. The European Centre for Medium-Range Weather Forecasts (ECMWF) is an intergovernmental organisation supported by 34 states formed in 1975 to produce medium-range numerical weather predictions for Europe. ECMWF runs many NWPs ranging from days ahead to months and seasons. The need for high resolution forecasts has led to the creation of a number of other European groups producing limited-area, high-resolution NWPs (ALADIN, COSMO, HIRLAM) [35].

Significant post-processing is required to derive wind power forecast from NWP outputs. Post-processing NWPs is a large area of research that has lead to the development of *model output statistics* (MOS) which aim to calibrate the forecast where possible for specific variables or locations. Modelling the wind power conversion process is necessary and can be a source of additional uncertainty. A review of how NWPs are used for short-term wind power forecasting can be found in [35].

A complete wind power forecasting system should utilise both physical and statistical modelling techniques which is the case for most operational commercial models [36]. Landberg and Troen [65, 66] developed a wind power forecasting tool called Prediktor which takes NWP wind speed and direction forecasts and transforms them to the local site before applying a power curve model. Statistical improvement can be achieved using MOS throughout the process. The Wind Power Prediction Tool (WPPT) has been developed by the Technical University of Denmark and is operated by the spin-off company ENFOR [67]. The WPPT uses adaptive recursive least squares estimation of the parameters of conditional parametric models to find the best connection between the NWP predicted wind speeds for the site and the measured power for each forecast horizon. The WPMS model, using neural networks, was developed in Germany and is used by E.On, RWE and National Wind Power in the UK [68, 69]. DNV–GL (formerly Garrad Hassan) [70] has a forecasting model called GH Forecaster, based on NWP forecasts from the UK Met Office. It uses multi-input linear regression techniques to convert from NWP to local wind speeds.

An important technique within NWP is ensemble forecasting, which involves running the NWP simulations multiple times with the estimates of the initial atmospheric conditions perturbed and/or different physical models [71, 72]. Ideally, the ensembles would be thought of as samples from a probability distribution function reflecting the uncertainty of the unperturbed forecast. The result is a probabilistic forecast from which the likelihood of different futures can be assessed, while, importantly, capturing spatial relationships, as discussed by Möller *et al.* [73], for example. University College Cork has developed wind power forecasting methodologies based on ensemble forecasts [74–77] and produced an operational forecasting system MSEPS (Multi-Scheme Ensemble Prediction System) based on a 2-step process: in the first step a physical reference power is computed, and in a second step the reference power is localised statistically and with the help of weather classes defined by the ensemble weather input.

A large volume of research aiming to improve wind power forecasts using NWPs has been undertaken over the last decade: Galanis *et al.* applies a Kalman filter to NWP data to improve wind and temperature forecasts [78]. Howard and Clark employ

a physical model to improve forecasts based on local terrain [79]. Khalid *et al.* use NWP predictions to supplement an autoregressive power prediction technique in [80] while Lee *et al.* use neural networks to produce power forecasts [81]. Raw ensemble forecasts are *un-calibrated* and measurements fall outside of the ensemble forecast, as a result, processes for calibrating ensemble forecasts have been developed. Specifically for wind power, Sloughter *et al.* proposes a method for the calibration of the wind speed output [82], while Pinson proposes the adaptive calibration of the bivariate (u, v)-wind to improve prediction [83]. A comparison of the COSMO and ECMWF model applications to short-term wind power forecasting can be found in [84].

In the wind power forecasting track of the 2014 Global Energy Forecasting Competition, the task was to produce day-ahead forecasts of the 1^{st}–99^{th} quantiles of wind power generation at 10 wind farms on a rolling bases using NWP forecasts and historic power generation as inputs. The top-ranking entries were dominated by machine learning techniques such as gradient boosting machines and various clustering and optimisation algorithms. They will be published in a forthcoming special issue of the International Journal of Forecasting.

A few examples of NWP based wind power forecasting systems have been highlighted here to show the breadth of techniques in the literature; for a more detailed survey see [36].

2.3 Statistical Methods

For short forecasting horizons statistical methods are superior to physical models for three reasons. Firstly, NWPs take several hours to produce and are typically only issued every 6 or 12 hours; so when the forecast is issued, the most recent input measurements will be several hours old. Secondly, NWP outputs are produced on spatial grids of varying resolution, not at specific points of interest, such as wind farms. As a result, spatial interpolation is required to produce forecasts at the location of interest which adds a further layer of complexity and potential source of error, particularly in complex terrain. Finally, users requiring very-short-term forecasts with temporal resolution finer than 1 hour suffer from the same interpolation problems as in the spatial case. For

these reasons a complete wind power forecasting tool will use statistical methods for the first 4–8 hours with a smooth transition to physical models for longer horizons.

There are two approaches to statistical wind power forecasting. The first is to forecast power directly, the second is to produce a wind speed (and direction) prediction and then combine that with a wind turbine or farm power curve model. The latter allows the stochastic wind speed to be separated from the power curve, though modelling the power curve presents its own challenges.

2.3.1 Linear Methods

The wind is frequently modelled as an autoregressive (AR) process, where the resulting prediction is a linear combination of past measurements (Appendix A.1), however, a classical AR process has zero mean and is homoscedastic — the wind speed and power do not fit these criteria. Therefore, either the wind speed/power time series must be transformed to meet these requirements or the AR model must be modified to accommodate them. Furthermore, wind and power time series are non-stationary, that is to say that their statistical properties change over time.

A popular approach is to fit a standard AR model after removing diurnal and seasonal trends to leave a residual series with the necessary properties [89–92]. Hill *et al.* go further by fitting a vector autoregressive model in order to utilise the spatial correlation between geographically separated sites, which is discussed in detail below. El-Fouly *et al.* employ an autoregressive approach modelling both wind speed and direction as independent variables to make predictions for a given day based on one-year and two-year-old measurements from the same day in previous years [93]. While the process of removing diurnal and seasonal trends helps, the synoptic trends, which contain far more energy than the diurnal as illustrated by Figure 2.1, are difficult to remove since their period is variable. The resulting de-trended time series are therefore still non-stationary, albeit to a lesser degree.

Others fit an autoregressive moving average (ARMA) model which models trends in the data, though this still assumes a constant variance [94–100]. Autoregressive integrated moving average (ARIMA) models, which attempt to remove the non-stationary

part of the wind with an initial differencing step, are described in [101,102]. Pinson *et al.* have proposed Markov-switching models in [61, 103] whereby multiple autoregressive models are fit, each representing different regimes, and either switched between (in the first case) or mixed (in the second). These sophisticated approaches consistently outperform simple AR based methods for the locations they are demonstrated on, however, few have been generalised to model and forecast multiple locations. A user interested in forecasts at multiple locations would have to identify the most suitable method for each location and run them in parallel.

Multi-scale analysis is used as an alternative to de-trending in [104] through a combination of second-order blind identification and autoregressive modelling. Many approaches based on wavelet analysis have been presented [105–109], as well as the empirical mode decomposition [110]. While such approaches are attractive and utilise powerful tools for analysing and modelling wind and power time series, in the forecasting paradigm they suffer from lag effects when combining forecasts on multiple scales which negatively impact accuracy.

Temperature and vertical wind component are included in the hypercomplex approach suggested in [111–113] where the 3D wind vector and air temperature or atmospheric pressure are modelled as quaternions and predictions are made using linear predictors estimated by stochastic gradient methods. These methods have been very successful for ultra-short-term forecasting (<1 second) with potential applications in wind turbine control, however, they have not been demonstrated on time scales relevant to power system operation. Exogenous variables, such as temperature and pressure, have been incorporated on longer time scales in conditional models and are surveyed in the next section.

2.3.2 Adaptive and Conditional Approaches

The characteristics of both the wind and wind power conversion process are not static; they vary slowly over time. The behaviour of the wind changes with season and the seasons themselves vary significantly from year to year. Further, the wind turbine power curve changes as the condition of the turbine varies with wear and maintenance [114].

Models used for forecasting should capture, or at least track, these changes in order to produce consistent and skilful forecasts. Linear models can often be updated recursively by algorithms such as the recursive least squares (RLS) and least mean squares (LMS) algorithms [115, 116], though these approaches will only track slow changes, and with some lag.

Conditions experienced by a wind farm may also change quickly, if the wind direction shifts, or if the local weather system changes. Wind farm power curves can depend strongly on wind direction and change over time, as illustrated by Jeon and Taylor in [117, 118], who use conditional kernel density estimation to produce wind power density and quantile forecasts from wind speed and direction forecasts. Regime-switching methods model these different behaviours separately and can utilise exogenous variables, such as wind direction, to switch between models [61, 103, 119–124].

Alternatively, regression coefficients may be replaced with coefficient functions to directly model a specific feature [125], such the motion of weather systems [126].

2.3.3 Machine Learning and Neural Networks

The complex non-linear nature of wind and wind power time series has motivated the application of machine learning algorithms to the prediction problem. In general, these algorithms attempt to *learn* the response of some unknown, potentially non-linear system linking some set of know inputs to known outputs. Once the system have been learnt, outputs may be estimated given some new input.

Artificial neural networks are statistical learning algorithms inspired by biological neural networks and many variations of the basic approach have been used to produce wind and wind power forecasts, for example [127–130]. Neural networks have been used to forecast wind time series in [131–134], for example, and are combined with a fuzzy logic model in [135]. Wan *et al.* produce 1-hour-ahead wind power interval forecasts using an extreme learning machine [136].

Neural networks are easily extended to the complex domain to capture directional information. Gautama *et al.* develops a test for detecting the complex-valued nature of time series in [137] and it is applied to hourly mean wind speeds by Goh *et al.*

in [132]. It is found that the wind time series being examined exhibits the properties of a complex-natured time series and that there are, therefore, advantages to modelling the wind as a complex-valued time series. Goh *et al.* go on to demonstrate that their neural network approach does indeed perform better when treating the wind a complex time series than as a Cartesian bivariate one.

Other learning algorithms have also been employed, often in combination with other statistical techniques: a genetic algorithm is used to train a fuzzy model for prediction in [138], and a clustering algorithm trains the fuzzy model in [139]. Markov chains [140] and data mining [141] have been utilised, among others, plus various hybrid approaches such as [142, 143].

Neural networks offer an easy to implement solution to a challenging problem but are not favoured by many in the statistical community since they are effectively a 'black box' offering little insight into why some variants perform well and others do not. The wind and wind power prediction problem is set in the real world and governed by physical processes that are well understood. It is preferable to use that understanding to design and develop predictors, rather than relying on abstract learning algorithms.

2.3.4 Spatio-temporal Prediction

Many forecast users, particularly utility-scale wind generators and power system operators, are not simply concerned with the production of individual wind farms but with their entire generation portfolio in the case of utilities, or all wind on the power system in the case of power system operators (aggregated by region or feed-in point to transmission system). In both cases it is desirable to model and forecast spatio-temporal features in order to improve the accuracy of point forecasts and to evaluate the uncertainty of aggregate generation forecasts. Spatio-temporal wind speed and direction forecasting offers the possibility of improved speed prediction if multiple locations are modelled by capturing the propagation of changes in wind speed to downwind sites, as demonstrated in [128].

Hering and Genton compare three spatial models in [122] that incorporate wind direction. In the first technique, also described in [121], different regimes are identified

associated with different wind directions. Independent prediction models are then applied to the wind speed depending on the current regime, determined by the most recent measurements. In the second, the sine and cosine of the wind direction are included in a linear predictor of wind speed, removing the need to identify specific regimes. The third model predicts perpendicular Cartesian components of the wind vector by fitting a bivariate skew-t regression model to the de-trended components. The third method produces less accurate wind speed forecasts, but it is the only method that forecasts wind direction. The same methods are applied to an economic dispatch model in [144] set in West Texas and demonstrate significant economic advantage vs. simple AR and persistence forecasts.

The three models described in [122] use information from three locations positioned favourably along the Colombia River Gorge. Likewise, [128] use a measurement station positioned up-wind in the prevailing wind direction to improve forecasts at a specific location. More general models, such as those described in [92], use measurements taken across an entire country with no specific knowledge of the topography at individual sites to improve predictions using vector-valued linear models. For large-scale spatial forecasting automated model fitting, numerical robustness and computational efficiency become serious considerations. This aspect is addressed by Sanandaji *et al.* who build a low-cost sparse spatio-temporal predictor inspired by techniques from compressive sensing [145] and by this thesis.

Capturing spatial information in probabilistic models is also an important problem. The potential benefits of using spatial information were realised in [146] and have been investigated more recently in [147, 148]. An alternative type of probabilistic forecast is proposed in [140] where the probability of the wind speed increasing, decreasing or staying the same in the next time step is predicted. Other contributions have sought to build efficient probabilistic spatial models with sparse Gaussian random fields [147, 149, 150] which can be sampled to produce ensemble-type forecasts.

2.4 Reference Models and Forecast Evaluation

In the most general sense, the *best* forecast is that which allows the end user to make decisions in some optimal fashion, often to minimise costs or maximise returns, as in energy trading, for example. Ideally, the predictor should optimise the cost function of the higher level problem. Probabilistic forecasts are optimal inputs to decision-making problems under uncertainty but modelling complex decisions can be challenging. Instead, many practitioners use point forecasts to inform their decisions. These forecasts are commonly evaluated based on analysis of forecast errors and compared to the performance of reference models. However, it should be noted that the forecast with the lowest average error does not necessarily produce the greatest economic return, as illustrated by Bessa *et al.* in [151].

The simplest reference model is the *persistence* forecast. This method supposes that the value of x_t at some future time $t + \Delta$ will be unchanged from time t, i.e.

$$\hat{x}_{t+\Delta} = x_t \quad , \tag{2.3}$$

where $\hat{}$ denotes a forecast. The forecast error is given by

$$e_t = \hat{x}_t - x_t \quad . \tag{2.4}$$

This simple technique performs (perhaps surprisingly) well for short-term wind and wind power prediction largely due to the relatively slow evolution of weather phenomena. The performance of the persistence forecast is sometimes considered a measure of the 'predictability' of a particular time series and is still used by some practitioners in the energy industry today for short-term forecasting. Any complex forecasting technique must demonstrate robust improvement over persistence to justify the additional cost and effort of its use.

Point forecasts are typically evaluated in terms of root mean squared error (RMSE)

and/or mean absolute error (MAE) given by

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N} e_t^2} \tag{2.5}$$

and

$$\text{MAE} = \frac{1}{N}\sum_{t=1}^{N} |e_t| \quad , \tag{2.6}$$

respectively. The RMSE is a common cost function in linear prediction problems since its minimisation presents a quadratic problem that can be solved either directly through differentiation, or iteratively by gradient decent, for example. However, MAE may be a more representative measure of utility in situations where the economic cost of a forecast error is proportional to the magnitude of the error, as opposed to its square. Both of these scores are often presented in terms of percentage improvement over persistence.

Further analysis can be performed to assess the quality of point forecasts. Linear predictors often assume normal i.i.d. errors and this assumption may be tested to validate the original assumption. Deviation from assumed properties may indicate the presence of systematic biases but lead to developments that could improve the predictor's performance.

In Chapters 3 and 4 of this thesis both wind speed and direction are forecast by modelling the wind as a complex random variable. Analysing the errors in this framework requires some thought since the complex prediction error alone is neither representative of the accuracy of the speed part of the prediction or the directional part, it is a combination of both. For example, a perfect prediction of the wind speed with an erroneous direction would be indistinguishable from a perfect direction prediction with erroneous speed without some additional information, as illustrated in Figure 2.3. Obviously, the speed and direction components can be separated and errors calculated, but predictors are formulated by minimising the complex prediction error, or a function thereof.

Probabilistic forecasts require slightly more involved validation since their skill is a combination of two properties: sharpness and reliability. A sharp forecast is one with narrow prediction intervals, a reliable one produces forecasts with the empirical

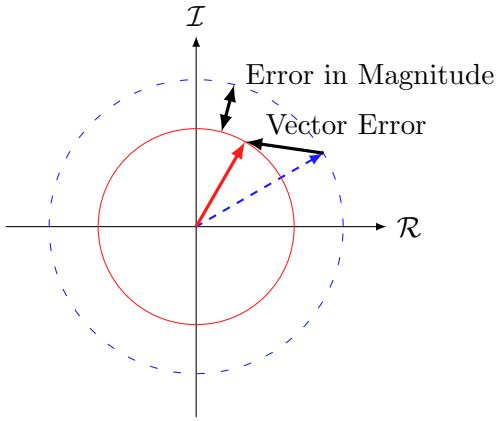


Figure 2.3: Illustration of the difference between the errors in magnitude associated with speed only prediction methods, and the vector error calculated after predicting the complex wind vector. The speed and vector prediction are illustrated by the dashed circle and arrow, respectively, and the actual wind speed and direction by the solid circle and arrow.

probabilities that match the nominal ones, for example the 25% quantile should be exceeded 25% of the time [152, 153]. Sharpness can be quantified by the continuous rank probability score (CRPS) or the log score. For a cumulative predictive distribution $\hat{F}_t(x)$ of random process X at time t, the CRPS is given by

$$\text{CRPS} = \frac{1}{T}\sum_{t=1}^{T}\int_0^1 \{\hat{F}_t(x) - \mathbf{1}(x \geq x_t)\}^2 dx \tag{2.7}$$

where x_t is the realisation of X at t and $\mathbf{1}(\cdot)$ is the indicator function. CRPS rewards sharpness and reduces to MAE when the forecast is deterministic. The log score is the mean negative log of the predictive distribution $\hat{f}_t(x_t)$ evaluated at the corresponding observation,

$$\text{Log Score} = \frac{1}{T}\sum_{t=1}^{T} -\ln\left(\hat{f}_t(x_t)\right) \quad . \tag{2.8}$$

Due to its logarithmic nature, the log score is not as robust as the CRPS: measurements in the tails of the predictive distribution are heavily penalised and the score returns ∞ if a single measurement falls where the predictive distribution is numerically zero.

A reliable or calibrated forecast $\hat{f}$ of a real process with observed distributions

$(g_t)_{t=1,2...}$ satisfies

$$\frac{1}{T}\sum_{t=1}^{T} g_t \circ \hat{f}_t^{-1}(p) \to p \tag{2.9}$$

for all $p \in (0,1)$; i.e. if the event $X = x$ is forecast with probability p, it must be observed with probability p. Reliability is usually evaluated using reliability diagrams, which are plots of nominal probabilities vs. observed outcomes [154, 155].

2.5 Summary and Discussion

Wind power forecasting has received a lot of attention from both academic and commercial enterprises over the past 30 years or so. A vast array of techniques have emerged capable of producing informative forecasts on a wide range of spatial and temporal scales. Forecasting beyond the next few hours typically requires large amounts of input data, often generated by physical models such as NWPs, whereas for shorter horizons statistical methods are preferred.

Until recently, the majority of short-term predictors were location specific, having been designed to perform well on a specific data set or at locations that exhibit certain properties, such as strong prevailing winds, reliable diurnal trends and so on. Others are more general, such as neural network based techniques, but can be computationally expensive and unreliable.

Ultimately, the needs of the end user will determine the appropriate forecasting methodology for a given problem, and it is perhaps for this reason that many utilities, power system operators and weather forecasting organisations employ staff specifically to meet their wind power forecasting needs, when they can be justified economically. In a recent survey of power system operators, respondents were asked to rate the importance of a variety of forecast products: next-hour forecasts were ranked as highly important by 70% of respondents, more than any other product in the survey, followed by ramp forecasts (62%) and ensemble forecasts (50%) [12]. However, a number of companies still rely on persistence for short-term forecasts because it is simple to implement and robust.

Several approaches have been proposed to capitalise on the spatial correlation be-

tween the wind speed and direction at geographically distributed sites, however, there is a lack of low-cost, easy to implement techniques that could be applied on a regional or national scale. Furthermore, there is increasing demand for very-short-term forecasts to aid the balancing of power systems with high wind penetration. When considering many 10s or 100s of locations, sparsity is desirable for numerical robustness and computational efficiency. All of these factors are combined in the smart grid paradigm of highly interconnected and communicative power systems which will require a suite of forecasts to operate most effectively and realise their potential.

This thesis aims to address some of the emerging and future needs for short- and very-short-term wind power forecasts on a large spatial scale. Throughout, scalability, ease of implementation, and computational efficiency guide the development of new statistical techniques for producing wind and wind power forecasts.

2.6 Main Contributions of this Thesis

The work presented in this thesis has contributed several new statistical methodologies to the wind and wind power forecasting community and literature. The focus is on statistical methods: in the Chapters 3 and 4, linear and non-linear techniques for short-term spatio-temporal wind speed and direction forecasting are developed, respectively. The aim is to produce low-cost predictors for hourly mean wind speed and direction up to 6 hours in advance. Spatial modelling can be used to capture spatio-temporal structures and produce accurate forecasts with relatively low computational demands compared to physical modelling. Wind speed and direction are modelled as the magnitude and phase of complex numbers and multiple spatial locations are used to build multi-channel filters* for prediction, inspired by techniques from signal processing. These filters are then conditioned on external variables, time of year [156–158] and wind direction [159], and methods for non-linear parameter estimation are investigated [160, 161].

In Chapter 4 very-short-term power forecasting is considered. A method for pro-

*Multi-channel filtering (signal processing) is very closely related to multi-variate time series analysis (statistics). The Wiener filter, which is used throughout this thesis, is mathematically equivalent to the maximum likelihood estimate of a vector autoregressive model.

ducing parametric probabilistic power forecasts is developed for the spatial case using the logit-normal distribution and vector autoregressive modelling [162]. Motivated by the need to produce such forecasts on a very large spatial scale, where fitting traditional VAR is impractical, a sparse parametrisation of VAR models is pursued. In addition, a novel exponential smoothing scheme is developed to better track changes in volatility.

Chapter 3

Linear Wind Prediction

The study of linear time series as produced a vast array of powerful tools for analysing, synthesising and predicting real world observations [85]. These tools are so revered, in fact, that often a lot of effort is put into *linearising* time series so that they approximate linearity and these tools can be employed. It suffices here to define a linear time series as an ordered list of numbers, each of which can be written as a weighted sum of the others. A technical definition is included in Appendix A.1. In this chapter, wind speed and direction are modelled as the magnitude and phase of a complex random variable and assumed linear. Linear tools are then used to predict wind time series.

The reach of linear methods extends to the analysis of multiple time series, such as measurements of the same variable made simultaneously at multiple locations. Understanding the relationship between such time series can allow inferences to be made about others out-with the original group (in techniques such as spatial interpolation or *kriging* [88]), or to inform the prediction of one times series using information from another [86]. Throughout this chapter, and thesis, the latter approach is employed and is called *spatio-temporal prediction*, since spatial information from the recent past is used to inform predictions.

In this chapter spatio-temporal predictors for hourly mean wind speed and direction are made from 1 to 6 hours ahead at multiple locations. Forecasts on this time scale are important when predicting the power produced by wind turbines for power system operation, energy trading and maintenance scheduling [9, 10, 163]. When working with

large volumes of data computational efficiency and the numerical properties of calculations are important considerations. Here, low-complexity predictors are developed that are simple to implement and fast to compute so that they may be easily employed as part of more computationally demanding problems, such as generation scheduling or energy trading, and by non-expert users.

Inspired by methodology developed in the field of signal processing, the wind speed and direction are modelled via the magnitude and phase of a complex-valued time series [132, 137, 164, 165]. A multichannel adaptive filter is set to predict this signal, based on its past values and the spatio-temporal correlation between wind signals measured at multiple geographical locations. Furthermore, complex-linear processing is significantly less computationally costly than bivariate approaches, such as those in [122].

In Section 3.1, motivated by the annual cycle of the seasons, a cyclo-stationary predictor is developed based on the Wiener filter — an optimal minimum mean squared error predictor, followed by a similar predictor conditioned on the wind direction, rather than season, in Section 3.2. Finally, the numerical properties of complex-linear and widely-linear predictors are compared in Section 3.3.

3.1 Seasonal Prediction

This section aims to produce a low-complexity predictor for the hourly mean wind speed and direction from 1 to 6 h ahead at multiple sites distributed around the UK. The wind speed and direction are modelled via the magnitude and phase of a complex-valued time series. A multichannel adaptive filter is set to predict this signal on the basis of its past values and the spatio-temporal correlation between wind signals measured at numerous geographical locations. The filter coefficients are determined by minimising the mean squared prediction error. To account for the seasonal variation of the wind time series and the underlying system, a cyclo-stationary Wiener solution is developed, which is shown to produce an accurate predictor [166]. An iterative solution, which provides lower computational complexity, increased robustness towards ill-conditioning of the data covariance matrices (since the need to invert the covariance matrices is avoided), and the ability to track time-variations in the underlying system, is also presented.

The approach is tested on wind speed and direction data measured at various sites across the UK. Results show that the proposed approaches are able to predict wind speed as accurately as state-of-the-art wind speed forecasting benchmarks while simultaneously providing valuable directional information.

3.1.1 Data Model and Adaptive Prediction

Complex Valued Wind Data Model

Wind speed and direction across M geographically separate sites are embedded in a vector-valued complex time series $\boldsymbol{x}[n] \in \mathbb{C}^M$, where the speed and direction of the wind form the magnitude and phase of the complex samples, and n is the discrete time index. The measured time series from individual spatial locations form the channels of a multichannel data model. Since the real and complex components of the wind signal are connected, i.e. $\boldsymbol{x}[n]$ is a complex process [165], it is both sensible and simple to pursue complex processing. As well as the mathematical economies of complex processing, the complex representation offers geometrical insight without the need for a bivariate coordinate system.

Based on the expectation operator $E\{\cdot\}$, we define the space-time covariance matrix

$$\mathbf{R}_{xx}[n,\tau] = E\{\boldsymbol{x}[n]\boldsymbol{x}^{\mathrm{H}}[n-\tau]\} \quad , \quad \mathbf{R}_{xx}[n,\tau] \in \mathbb{C}^{M\times M} \quad , \tag{3.1}$$

which contains auto-correlation sequences of the M wind signals on its main diagonal, and the cross-correlation sequences between different site measurements on the off-diagonals. The vector $\boldsymbol{x}^{\mathrm{H}}[n]$ denotes the conjugate transpose of $\boldsymbol{x}[n]$. In the case of wide-sense stationary data, the space-time covariance matrix will only depend on the lag parameter τ and takes on the Hermitian form $\mathbf{R}_{xx}[\tau] = \mathbf{R}^{\mathrm{H}}_{xx}[-\tau]$.

With respect to wind speed and wind direction, the former is likely non-stationary and non-linear, while the latter can be volatile and depend heavily on the physical characteristics of the measurement site. Furthermore, the seasonal and diurnal trends that characterise our human experience of the wind are themselves variable. Below, the potential non-linear nature of the wind is ignored and linear processing is pursued.

The assumption of stationarity is dropped for a quasi-stationary behaviour, whereby the space-time covariance matrix can be assumed to be stationary — and therefore only dependent on the lag parameter τ — for sufficiently short time windows [167].

Optimal Mean-Squared Error Predictor

We consider the problem of predicting Δ samples ahead, based on M spatial measurements in $\boldsymbol{x}[n]$ and a time window containing N past samples for each site. Therefore, the prediction error can be formulated as

$$\begin{aligned} \boldsymbol{e}[n] &= \boldsymbol{x}[n] - \sum_{\nu=0}^{N-1} \boldsymbol{W}^{\mathrm{H}}[n,\nu]\boldsymbol{x}[n-\Delta-\nu] && (3.2) \\ &= \boldsymbol{x}[n] - \boldsymbol{W}_n^{\mathrm{H}}\mathbf{x}_{n-\Delta} \quad , && (3.3) \end{aligned}$$

with

$$\boldsymbol{W}_n = \begin{bmatrix} \boldsymbol{W}[n,0] \\ \boldsymbol{W}[n,1] \\ \vdots \\ \boldsymbol{W}[n,N-1] \end{bmatrix} \in \mathbb{C}^{MN\times M} \quad , \quad \mathbf{x}_n = \begin{bmatrix} \boldsymbol{x}[n] \\ \boldsymbol{x}[n-1] \\ \vdots \\ \boldsymbol{x}[n-N+1] \end{bmatrix} \in \mathbb{C}^{MN} \quad . \quad (3.4)$$

The matrices $\boldsymbol{W}[n,\nu] \in \mathbb{C}^{M\times M}$ describe the predictor's reliance on all spatial measurements taken $\nu+\Delta$ samples in the past, at time instance n. Specifically, $[\boldsymbol{W}[n,\nu]]_{p,q}$ addresses the influence of the measurement at site p onto the prediction at the q^{th} location. In order to simply use the Hermitian transpose operator in (3.3), $\boldsymbol{W}[n,\nu]$ contains the complex conjugate prediction filter coefficients.

The error covariance matrix derived from (3.3), $\mathbf{R}_{ee}[n] = E\{\boldsymbol{e}[n]\boldsymbol{e}^{\mathrm{H}}[n]\} \in \mathbb{C}^{M\times M}$, is obtained by taking expectations over the ensemble, and in itself may be varying with time n. Note that in case of stationarity, the dependency of both $\boldsymbol{W}_n$ and $\mathbf{R}_{ee}[n]$ on n vanishes. We will carry forward n since it is well known that the wind signal is non-stationary and develop a cyclo-stationary solution in Section 3.1.1.

Calculating $\mathbf{R}_{ee}[n]$ using (3.3) yields a quadratic expression in $\boldsymbol{W}_n$,

$$\begin{aligned}\mathbf{R}_{ee}[n] &= E\left\{(\boldsymbol{x}[n]-\boldsymbol{W}_n^{\mathrm{H}}\mathbf{x}_{n-\Delta})(\boldsymbol{x}^{\mathrm{H}}[n]-\mathbf{x}_{n-\Delta}^{\mathrm{H}}\boldsymbol{W}_n)\right\} \quad , \\ &= \mathbf{R}_{xx}[n,0]-E\{\boldsymbol{x}[n]\mathbf{x}_{n-\Delta}^{\mathrm{H}}\}\boldsymbol{W}_n-\boldsymbol{W}_n^{\mathrm{H}}E\{\mathbf{x}_{n-\Delta}\boldsymbol{x}^{\mathrm{H}}[n]\}+\boldsymbol{W}_n^{\mathrm{H}}E\{\mathbf{x}_{n-\Delta}\mathbf{x}_{n-\Delta}^{\mathrm{H}}\}\boldsymbol{W}_n \quad , \\ &= \mathbf{R}_{xx}[n,0]-\mathbf{R}_{x\mathbf{x}}[n]\boldsymbol{W}_n-\boldsymbol{W}_n^{\mathrm{H}}\mathbf{R}_{x\mathbf{x}}^{\mathrm{H}}[n]+\mathbf{W}_n^{\mathrm{H}}\mathbf{R}_{\mathbf{xx}}[n]\boldsymbol{W}_n \quad , \end{aligned} \tag{3.5}$$

where

$$\mathbf{R}_{x\mathbf{x}}[n] = \left[\begin{array}{cccc} \mathbf{R}_{xx}[n,\Delta] & , & \mathbf{R}_{xx}[n,\Delta-1] & , \quad \ldots \quad , \quad \mathbf{R}_{xx}[n,\Delta-N+1] \end{array}\right] \quad , \tag{3.6}$$

$$\mathbf{R}_{\mathbf{xx}}[n] = \left[\begin{array}{ccc} \mathbf{R}_{xx}[n-\Delta,0] & \ldots & \mathbf{R}_{xx}[n-\Delta,N-1] \\ \mathbf{R}_{xx}[n-\Delta-1,-1] & & \mathbf{R}_{xx}[n-\Delta-1,N-2] \\ \vdots & \ddots & \vdots \\ \mathbf{R}_{xx}[n-\Delta-N+1,-N+1] & \ldots & \mathbf{R}_{xx}[n-\Delta-N+1,0] \end{array}\right] \quad . \tag{3.7}$$

We assume that $\boldsymbol{x}[n]$ is stationary over at least 2Δ samples, which is reasonable since $\Delta = 1, ..., 6$ hours and the annual cycle of seasons has a fundamental period of 8760 hours. As a result, $\mathbf{R}_{\mathbf{xx}}[n]$ is Hermitian and positive semi-definite [168]. The matrix $\mathbf{R}_{\mathbf{xx}}[n]$ admits a unique solution to minimise the mean square error,

$$\boldsymbol{W}_{n,\mathrm{opt}} = \arg\min_{\boldsymbol{W}_n} \mathrm{trace}\{\mathbf{R}_{ee}[n]\} \quad . \tag{3.8}$$

It can be shown that $\mathrm{trace}\{\mathbf{R}_{ee}[n]\}$ is quadratic in $\boldsymbol{W}_n$, such that the solution to (3.8) can be found by matrix- and complex-valued calculus [169]. Finding the minimum requires equating the gradient w.r.t. the unconjugated predictor coefficients in $\boldsymbol{W}_n^*$ to zero. We utilise results from [169] which show that for constant matrices $\mathbf{A}$ and $\mathbf{B}$ the expressions

$$\partial \mathrm{trace}\{\mathbf{A}\boldsymbol{W}_n^{\mathrm{H}}\mathbf{B}\}/(\partial \boldsymbol{W}_n^*) = \mathbf{B}\mathbf{A} \tag{3.9}$$

and

$$\partial \mathrm{trace}\{\mathbf{A}\boldsymbol{W}_n\mathbf{B}\}/(\partial \boldsymbol{W}_n^*) = \mathbf{0} \tag{3.10}$$

hold. Applying this, and using the product rule for differentiation of the quadratic term in (3.5), yields

$$\frac{\partial}{\partial \boldsymbol{W}_n^*}\text{trace}\{\mathbf{R}_{ee}[n]\} = -\mathbf{R}_{x\mathbf{x}}^{\text{H}}[n] + \mathbf{R}_{\mathbf{xx}}[n]\boldsymbol{W}_n \quad . \tag{3.11}$$

Finally, setting the gradient on the right had side of (3.11) equal to zero yields the optimal predictor coefficients that minimise trace$\{\mathbf{R}_{ee}[n]\}$,

$$\boldsymbol{W}_{n,\text{opt}} = \mathbf{R}_{\mathbf{xx}}^{-1}[n]\mathbf{R}_{x\mathbf{x}}^{\text{H}}[n] \quad , \tag{3.12}$$

which is the well-known Wiener-Hopf solution [170, 171].

Cyclo-stationary Solution

The cyclo-stationary solution is based on the assumption that windows of data of length $L+1$ are approximately stationary, and furthermore, that the statistics of that period are the same during the equivalent window in all years. The covariance matrix $\mathbf{R}_{xx}[n,\tau]$ is estimated by calculating the expectation using only data in the quasi-stationary window centred on n from each year of available training data. In the estimation of $\mathbf{R}_{xx}[n,\tau]$, assume cyclo-stationarity, i.e. $\mathbf{R}_{xx}[n,\tau] = \mathbf{R}_{xx}[n-kT,\tau]$, with $k \in \mathbb{N}$ and T the fundamental period, i.e. 1 year. On the basis of cyclo-stationarity and data available for K past years, the estimation of the covariance matrix for time n is calculated as

$$\hat{\mathbf{R}}_{xx}[n,\tau] = \frac{1}{K(L+1)}\sum_{k=1}^{K}\left(\sum_{\nu=\frac{-L}{2}}^{\frac{L}{2}}\boldsymbol{x}[n-kT-\nu]\boldsymbol{x}^{\text{H}}[n-kT-\nu-\tau]\right) + \frac{2}{L}\sum_{\nu=1}^{\frac{L}{2}}\boldsymbol{x}[n-\nu]\boldsymbol{x}^{\text{H}}[n-\nu-\tau] \quad . \tag{3.13}$$

The optimal prediction filter for time n can then be calculated as

$$\boldsymbol{W}_{n,\text{opt}} = \hat{\mathbf{R}}_{\mathbf{xx}}^{-1}[n]\hat{\mathbf{R}}_{x\mathbf{x}}^{\text{H}}[n] \quad . \tag{3.14}$$

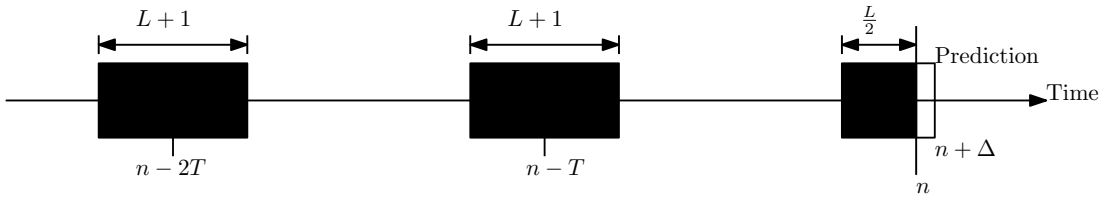


Figure 3.1: Diagram illustrating input data used for the estimation of the cyclo-stationary covariance matrix at time index n.

The input data for the estimation of the cyclo-stationary covariance matrix at time index n is illustrated in Figure 3.1.

Determining the window length L is a trade-off between consistency of performance and excess error caused by the inclusion of mismatched statistics. The window must be short enough to capture the common properties of the season but also long enough to smooth the effects of extreme events from individual years.

Iterative Prediction Filter

As an alternative to the Wiener-Hopf solution defined by (3.12), the quadratic MSE cost function has motivated lower-cost iterative approaches such as the method of steepest descent where

$$\boldsymbol{W}_{n+1} = \boldsymbol{W}_n - \mu \frac{\partial}{\partial \boldsymbol{W}_n^*} \text{trace}\{\mathbf{R}_{ee}[n]\} \quad , \tag{3.15}$$

i.e. the algorithm steps in the direction of the negative gradient of the cost function in proportion to the learning rate, μ. Amongst iterative schemes, Widrow's stochastic gradient technique called the least-mean square (LMS) algorithm [172] has proven simple and robust, whereby $\mathbf{R}_{ee}[n]$ is replaced by the poor instantaneous estimate $\hat{\mathbf{R}}_{ee}[n] = \boldsymbol{e}[n]\boldsymbol{e}^{\mathrm{H}}[n]$. The differentiation $\frac{\partial}{\partial \boldsymbol{W}_n^*}\text{trace}\{\boldsymbol{e}[n]\boldsymbol{e}^{\mathrm{H}}[n]\} = -\mathbf{x}_n\boldsymbol{e}^{\mathrm{H}}[n]$ leads to the straightforward update equation

$$\boldsymbol{W}_{n+1} = \boldsymbol{W}_n + \mu \mathbf{x}_n \boldsymbol{e}^{\mathrm{H}}[n] \quad . \tag{3.16}$$

Assuming a sufficiently small value of μ, the iterative nature of (3.16) averages out the gradient noise that results from the poor estimation of $\mathbf{R}_{ee}[n]$.

For the stationary case, selecting μ presents a trade-off between convergence speed

and mean squared error. For a large value of μ within the learning rate bounds [170,171], the filter coefficients will quickly converge towards the Wiener-Hopf solution, however the gradient noise contributes inaccuracy to $\boldsymbol{W}$ which negatively impacts the predictor's performance. Choosing a smaller μ reduces the effect of noise on the filter coefficients — reducing excess MSE — at the cost of convergence speed.

Under non-stationary conditions, the optimal filter coefficients are time dependent; the LMS algorithm will track this dependence, albeit with some lag [173]. Now the trade-off when choosing μ lies between accurate tracking and minimising lag. Convergence speed is still also a consideration. The tracking ability and relative simplicity of the LMS algorithm offer a powerful and computationally inexpensive predictor compared to other similar algorithms [174, 175].

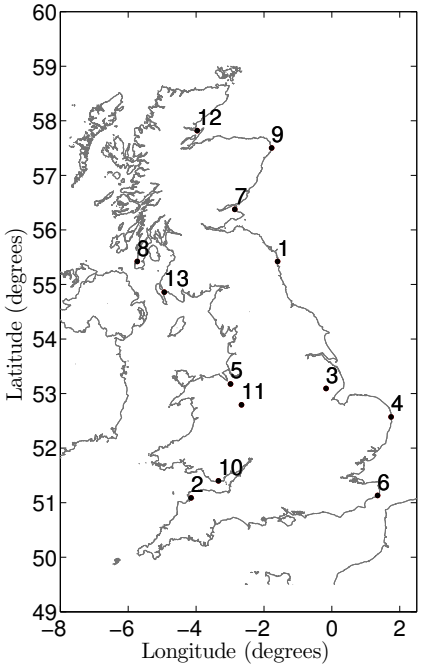
3.1.2 Testing and Results

Data Used for Testing

The proposed approaches are tested on wind data provided by the British Atmospheric Data Centre, which comprises of recordings over 6 years — from 00:00h on 1/3/1992 to 23:00h on 28/2/1998 — obtained from 13 sites across the UK as detailed in Figure 3.2. The measurements are taken in open terrain at a height of 10m, and comprise hourly averages that are quantised to a 10° angular granularity and integer multiples of one knot (0.515ms^{-1}) [2]. For this study, only sites with near complete continuous data are used and any prediction errors affected by a missing or erroneous data points are discarded. For the purposes of calculating data covariance matrices, missing and erroneous data was again discarded and the normalisation factors in (3.13) adjusted accordingly. Predictions and errors affected by missing data are thus mitigated. The performance of each filter is assessed by measuring the RMSE and improvement over the persistence method, which is a common benchmark for such forecasts [176].

Wiener Solution

The estimated stationary Wiener filter for the complete data set was calculated on the 5 year training data and tested on the remaining year of data for comparison to the cyclo-



1. Boulmer
2. Chivenor
3. Coningsby
4. Gorleston
5. Hawarden Airport
6. Langdon Bay
7. Leuchars
8. Machrihanish
9. Peterhead Harbour
10. Rhoose
11. Shawbury
12. Tain Range
13. West Freugh

Figure 3.2: Geographical distribution of 13 Met. Office stations supplying test data.

stationary Wiener and LMS approaches. In order to implement the cyclo-stationary approximation, the optimum window size L that best approximates stationarity with a sufficiently consistent estimation was found through numerical testing, shown in Figure 3.3, to be L equivalent to 15 weeks. Data windows from some sites are closer to being stationary than others and this is reflected in the final filter's performance.

While the notation in (3.13) suggests to re-calculate the Wiener filter coefficients at every time step, for the sake of computational complexity, the coefficient set was only updated once every 24 time steps, i.e. once a day. In tests this proved to be sufficiently short compared to the much longer data window L, and incurred no penalty in terms of performance.

Least Mean-Squares Algorithm

As discussed in Section 3.1.1, there are trade-offs to be made when choosing the filter length, N, and the learning rate, μ, of the LMS algorithm. The filter length and learning

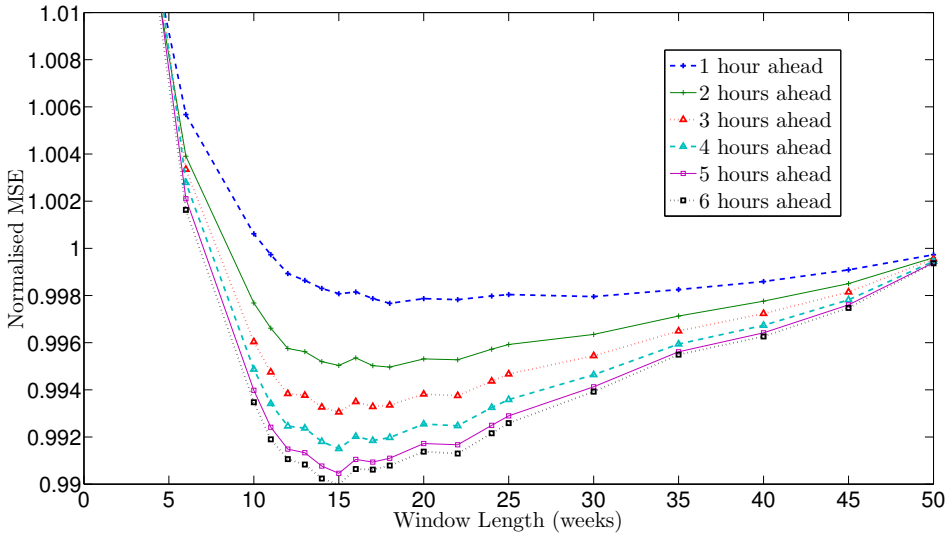


Figure 3.3: Cyclo-stationary filter performance depending on the window length L in terms of MSE averaged across all locations and normalised w.r.t. a stationary Wiener filter ($L = 52$ weeks) for all look-ahead times Δ.

rate are chosen to minimise excess MSE caused by poor tracking of the non-stationarity and the effects of noise.

In order to characterise the tracking ability of the LMS algorithm, the filter weights are initialised with the Wiener-Hopf solution of Section 3.1.1, using (3.12) and the training data. Based on the resulting tracking performance in Figure 3.4 for a combination of values for the filter length N and the learning rate μ, approximate optimal performance is determined for $N = 5$ and $\mu = 2.5 \times 10^{-5}$, which have been employed for all further tests with the LMS predictor below.

Figure 3.5 shows the typical variation of the filter coefficients during the year of test data. Clearly the LMS algorithm fails to match the tracking ability of the cyclo-stationary Wiener filter. This should not come as a surprise since each LMS update relies on only very recent information whereas the cyclo-stationary filter additionally relies on information from previous weeks and years.

		Speed RMSE (ms^{-1})				Velocity RMSE (ms^{-1})		
Site	Δ	CSWF	LMS	VAR(2)	Persistence	CSWF	LMS	Persistence
Boulmer	1	**1.09**	**1.09**	1.15	1.20	**1.54**	1.55	1.64
	2	1.50	**1.49**	1.57	1.71	**2.08**	2.09	2.35
	3	1.76	**1.74**	1.83	2.06	**2.41**	2.43	2.85
	4	1.97	**1.95**	2.03	2.34	**2.68**	2.70	3.27
	5	2.14	**2.12**	2.18	2.56	**2.91**	2.95	3.64
	6	2.30	**2.27**	2.30	2.73	**3.12**	3.17	3.95
Coningsby	1	**0.84**	0.85	0.87	0.93	**1.22**	**1.22**	1.32
	2	**1.17**	1.18	1.21	1.36	**1.67**	**1.67**	1.92
	3	**1.40**	1.41	1.45	1.68	**1.98**	1.99	2.38
	4	**1.58**	1.60	1.64	1.94	**2.23**	2.24	2.77
	5	**1.74**	1.76	1.80	2.17	**2.45**	2.46	3.09
	6	**1.87**	1.91	1.93	2.37	**2.63**	2.65	3.37
Leuchars	1	**1.08**	**1.08**	1.09	1.15	**1.52**	**1.52**	1.59
	2	**1.49**	**1.49**	1.50	1.63	**2.07**	**2.07**	2.25
	3	1.77	**1.75**	1.76	1.97	**2.43**	2.44	2.74
	4	2.00	**1.98**	**1.98**	2.25	**2.72**	2.74	3.14
	5	2.20	2.18	**2.17**	2.50	**2.98**	3.01	3.48
	6	2.37	2.36	**2.32**	2.71	**3.20**	3.24	3.77
Shawbury	1	1.01	**1.00**	1.03	1.10	1.43	**1.42**	1.56
	2	1.37	**1.33**	1.38	1.54	1.90	**1.89**	2.20
	3	1.62	**1.56**	1.60	1.85	**2.23**	**2.23**	2.67
	4	1.83	**1.77**	1.79	2.10	**2.50**	2.51	3.07
	5	2.01	**1.94**	1.95	2.32	**2.73**	2.76	3.41
	6	2.15	2.10	**2.09**	2.51	**2.93**	2.98	3.70
Mean Across All Sites	1	**1.11**	**1.11**	1.15	1.20	**1.64**	**1.64**	1.74
	2	1.51	**1.50**	1.52	1.66	**2.20**	**2.20**	2.42
	3	1.78	**1.77**	1.78	1.98	**2.59**	**2.59**	2.92
	4	2.01	2.00	**1.98**	2.24	**2.90**	2.91	3.34
	5	2.20	2.20	**2.14**	2.46	**3.16**	3.19	3.70
	6	2.36	2.37	**2.28**	2.64	**3.39**	2.43	4.01

Table 3.1: Root mean-squared speed and vector prediction errors (RMSE) from four selected sites, and the mean taken across all thirteen sites in the model for the cyclostationary Wiener filter (CSWF), least mean-squares algorithm (LMS), vector autoregressive (VAR(2)) and persistence.

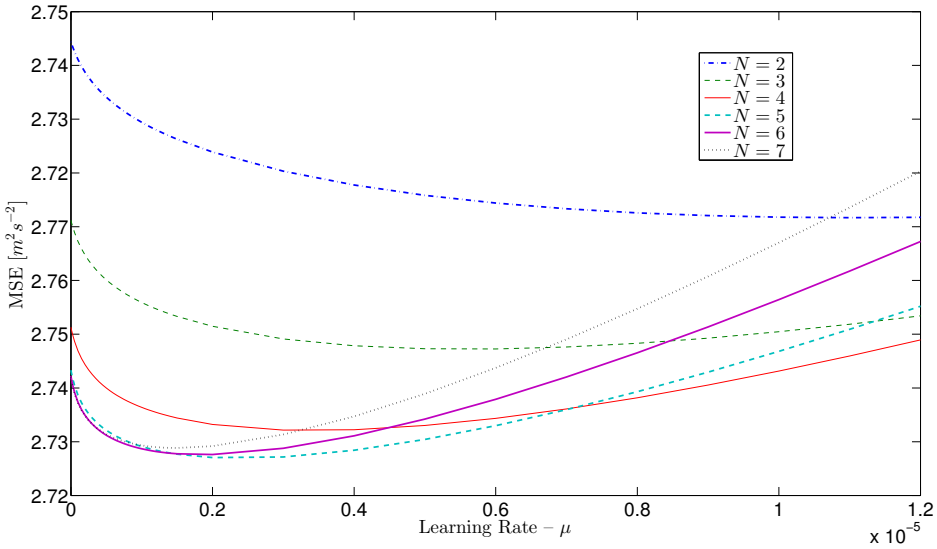


Figure 3.4: Tracking performance of LMS algorithm for different filter lengths when initialised with the stationary Wiener solution.

Results

Prediction results from the cyclo-stationary Wiener solution and LMS algorithm are shown in Table 3.1, while improvement over persistence is shown in Figure 3.6 along with the stationary Wiener filter for comparison. The LMS algorithm's tracking ability provides a clear improvement on the stationary Wiener filter though not as much as the cyclo-stationary Wiener solution. The largest improvements are seen at greater look-ahead times where the performance of the persistence method worsens.

Time series of 1-hour-ahead CSWF and LMS complex valued wind forecasts are illustrated in Figure 3.7, and 1- and 6-hour-ahead wind speed forecasts are illustrated in Figures 3.8a and 3.8b.

To compare the proposed approaches with others, it is noted that for non-site-specific spatial multichannel prediction to date only wind speed has been considered. Compared to the complex prediction error $e_l[n]$ of the predicted estimate $\hat{x}_l[n]$ at a site l, i.e. the lth component in (3.3), an error for the speed-only component, $e_{\mathrm{s},l}[n]$, can be defined as

$$e_{\mathrm{s},l}[n] = |x_l[n]| - |\hat{x}_l[n]| \quad . \tag{3.17}$$

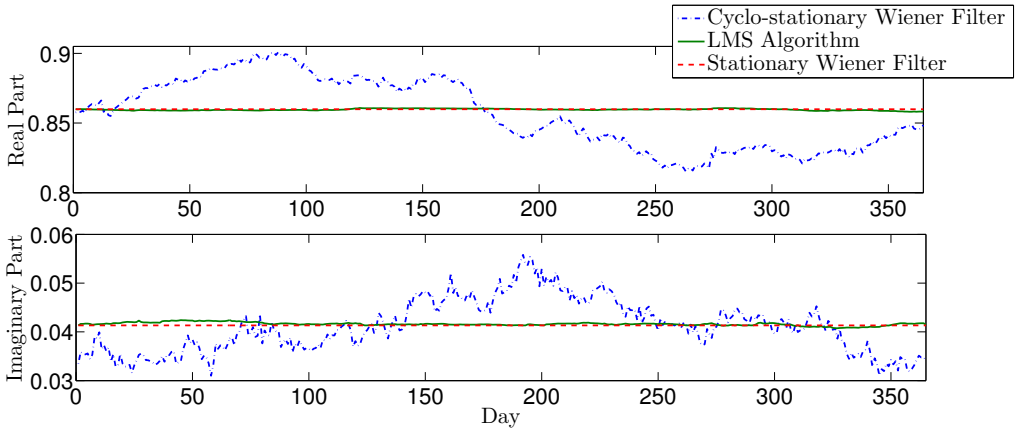


Figure 3.5: Variation of a typical Cyclo-stationary, LMS and stationary Wiener auto-correlative filter coefficient during prediction year.

However, note that due to Schwartz' inequality, $|e_{s,l}[n]| \leq |e_l[n]|$, such a comparison is difficult.

The accuracy of the cyclo-stationary Wiener filter's wind speed prediction for specific look-ahead periods, Δ, has been calculated and is compared to the mathematically similar vector autoregressive (VAR(2)) method of [92] in Figure 3.9a. The autoregressive coefficients of the VAR(2) model are static and calculated using the Yule-Walker approach on the de-trended test data [86]. The annual and diurnal trends were determined by fitting Fourier series to the test data for individual sites: a three term series for the annual trend and four two-term series for the diurnal trend, one for each season [92]. The improvement of the wind vector forecast over persistence is illustrated in Figure 3.9b.

We see that the performance of the speed part of the cyclo-stationary Wiener filter's prediction is comparable to the speed-only VAR(2) method overall, though the performance of both approaches varies from site to site. The directional Wiener filter shows greater improvement over persistence but it should be noted that this is improvement in the directional speed forecast error for both the CSWF and persistence and therefore cannot be directly compared to speed-only forecasts.

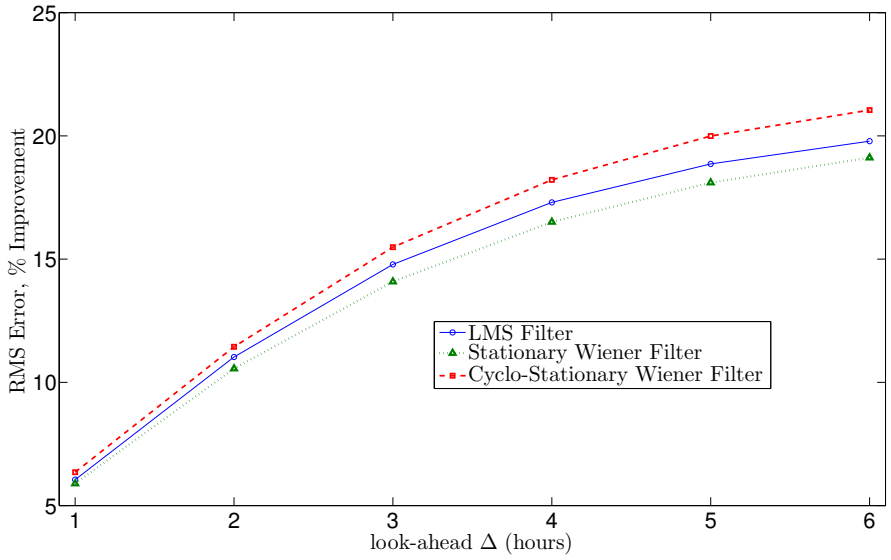


Figure 3.6: Improvement over persistence for the stationary Wiener filter, cyclo-stationary Wiener filter and LMS filter.

3.1.3 Summary

The aim of this work was to propose a low-cost spatio-temporal adaptive filter for predicting the hourly mean of both wind speed and direction, based on spatial information drawn from geographically separated sites. With prediction methods of comparable complexity to date either restricted to single-site data but with multiple parameters captured e.g. in complex valued time series, or restricted to speed prediction only when based on multiple measurement locations, the proposed method CSWF fills a gap in research.

We have developed a new cyclo-stationary Wiener filter which is motivated by the approximately annual cycles in the data, and leads to the estimation of a cyclo-stationary covariance matrix, which is assumed to be quasi-stationary over sufficiently small intervals. The calculation of this covariance matrix aims to keep the data window sufficiently short in order to discard out-dated samples from the estimation, while the cyclic inclusion of several years' of data enhances consistency of estimates. An iterative, stochastic gradient predictor has also been suggested, which utilises a multichannel least mean squares algorithm. The LMS is motivated by its significantly lower complexity

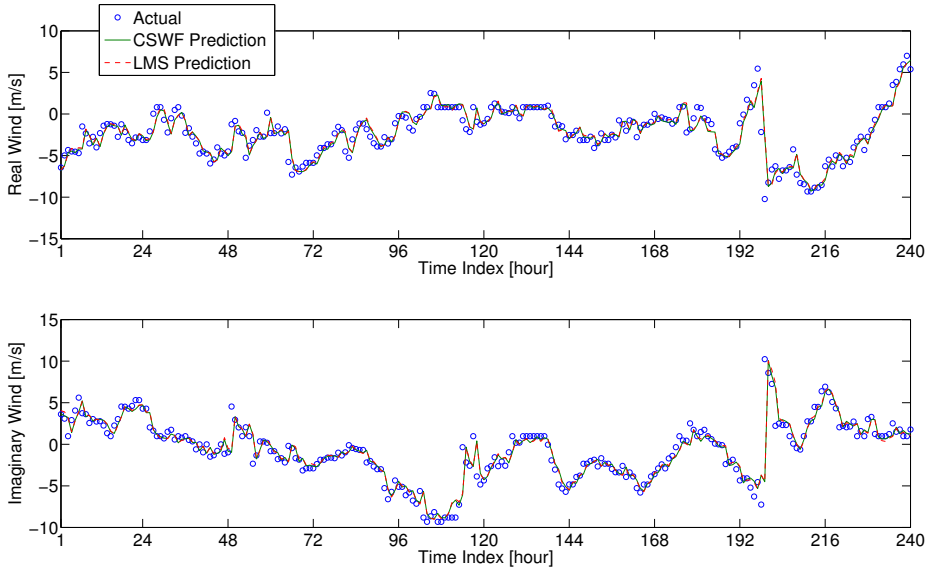
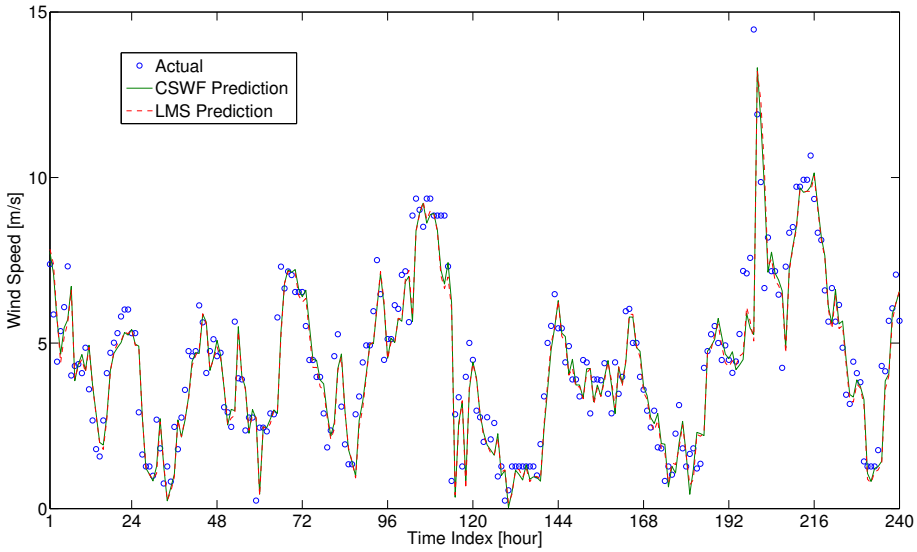


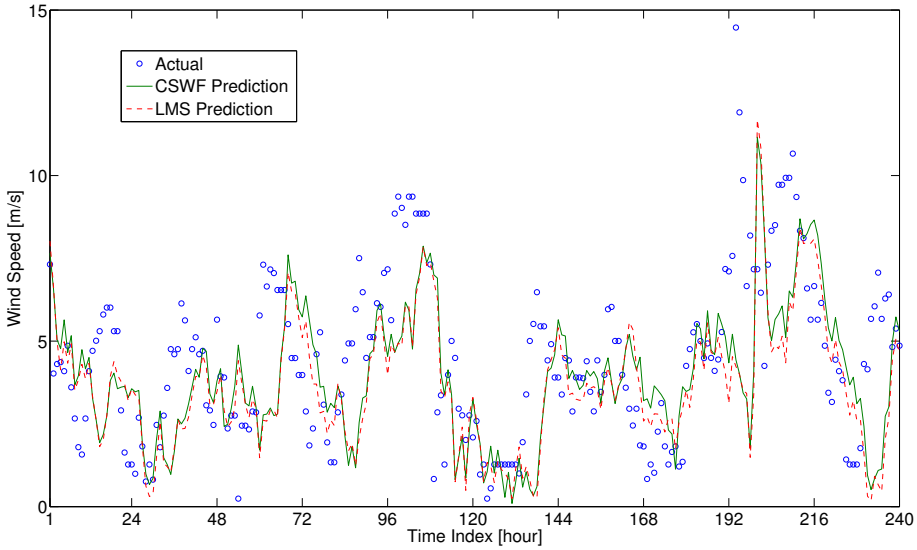
Figure 3.7: Real and imaginary parts of the 1-hour-ahead wind forecasts for Boulmer starting at 3pm on 2nd August 1998.

compared to the Wiener solution, its enhanced numerical stability due to the avoidance of matrix inversions, and its favourable tracking performance.

The proposed methods have been tested on wind measurements obtained at 13 locations in the UK over a period of 6 years. The results have been assessed against persistence, and generally show superior performance of the cyclo-stationary Wiener filter over a stationary version and the LMS, which supports the assumption of cyclo-stationarity of the data. The cyclo-stationary Wiener filter and LMS provide speed and vector predictions with greater accuracy than persistence and the simplicity of the LMS algorithm is found to come at only a small cost in vector prediction accuracy and no cost in speed prediction accuracy. Finally, the performance of the proposed methods is compared to the speed-only spatial prediction VAR(2), as described in [92]. The proposed filters are found to produce wind speed predictions of comparable accuracy to VAR(2) while, significantly, also providing directional information. The general applicability of the cyclo-stationary Wiener filter and the multichannel LMS prediction methods provide a valuable alternative to other statistical techniques which are often have to be tailored to local conditions or are computationally demanding.

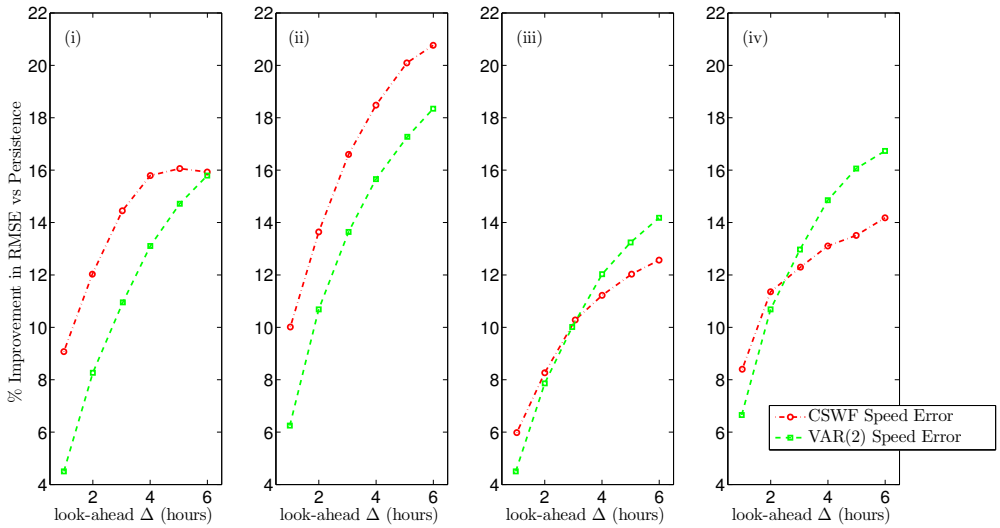


(a) 1-hour-ahead wind speed forecasts for Boulmer.

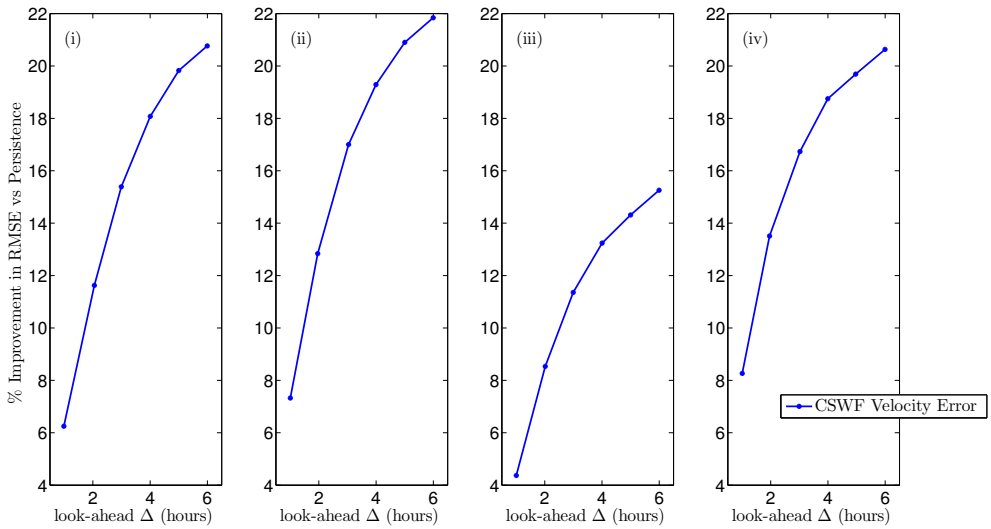


(b) 6-hour-ahead wind speed forecasts for Boulmer.

Figure 3.8: 1- and 6-hour-ahead wind speed forecasts for Boulmer starting at 3pm on 2nd August 1998.



(a) Wind speed forecast error improvement compared to persistence.



(b) Wind vector forecast error improvement compared to persistence.

Figure 3.9: Improvement over persistence of VAR(2) and cyclo-stationary Wiener predictions for four sites. Both the velocity and speed error of the Wiener predictions are shown for comparison to the speed-only VAR(2) method. Site (i): Boulmer, Site (ii): Coningsby, Site (iii): Leuchars, Site (iv): Shawbury.

3.2 Continuous Directional Regimes

In this section a conditional predictor is developed: the predictor's coefficients are replaced with coefficient functions which depend explicitly on the wind direction. This approach is motivated by the desire to capture causality in the spatio-temporal relationship between locations; the propagation of weather fronts across a region for example.

Direction has been captured in complex-valued neural networks, for example in [132, 133, 135], but these only model individual spatial locations. Others have developed regime-switching approaches which predict the wind speed depending on which direction-based regime the most recent measurements fall into, for example [121, 123, 124]. Two bivariate models are described in [122] that predict wind speed and direction; the first is regime-based and models the wind speed and direction, while the second models perpendicular Cartesian components of the wind speed. All of these predictors are trained on a continuous series of the most recent measurements made at multiple locations. Other regime-switching approaches, such as the Markov-switching autoregressive model proposed in [61], show that regime determination and selection may be data-driven and need not depend on an exogenous variable.

Furthering the development of complex-valued prediction, reported in Section 3.1 and [157], this work aims to extend the regime-switching type approaches, which commonly contain 2 or 3 fixed regimes (though the predictor for each regime is commonly adaptive) specific to the target prediction site. By introducing the concept of continuous directional regimes, an adaptive spatial predictor is developed which is optimised at regular intervals for the current wind conditions at multiple sites on a national scale based on the wind's behaviour during periods of similar conditions in the past.

The data model and approach to spatial prediction are introduced in Section 3.2.1 and the minimum mean squared error predictor and proposed continuous directional regime predictor are derived in Sections 3.2.1 and 3.2.1. The testing procedure and results are presented in Section 3.2.2 and a summary is provided in 3.2.3.

3.2.1 Data Model and Spatial Prediction

At discrete time n, the wind speed and direction at M locations are embedded as the magnitude and phase of a complex valued vector $\boldsymbol{x}[n] \in \mathbb{C}^M$. The spatial covariance matrix is defined based on the expectation operator, $E\{\cdot\}$, as $\mathbf{R}_{xx}[n,\tau] = E\{\boldsymbol{x}[n]\boldsymbol{x}^{\mathrm{H}}[n-\tau]\}$. Where $\boldsymbol{x}^{\mathrm{H}}[n]$ denotes the Hermitian transpose of $\boldsymbol{x}[n]$ and τ is a general lag parameter.

It is well known that wind speed and wind direction are likely non-stationary (has time-varying probability distribution) and otherwise non-linear; both can be volatile and, direction particularly, can depend heavily on the physical characteristics of the measurement site. Furthermore, the seasonal and diurnal trends that characterize our human experience of the wind are themselves variable. In the succeeding text, we ignore the potential non-linear nature of the wind and restrict ourselves to linear processing but drop the assumption of stationarity for a quasi-stationary behaviour, whereby the space-time covariance matrix can be assumed to be stationary—and therefore only dependent on the lag parameter τ—for sufficiently short time windows [167].

MMSE Predictor

Consider again the problem of predicting Δ samples ahead while minimising the mean-squared prediction error (MSE), based on M spatial measurements in $\boldsymbol{x}[n]$ and a time window containing N past samples for each site. The prediction error can be formulated as

$$\boldsymbol{e}[n] = \boldsymbol{x}[n] - \sum_{\nu=0}^{N-1} \boldsymbol{W}^{\mathrm{H}}[n,\nu]\boldsymbol{x}[n-\Delta-\nu] \tag{3.18}$$

$$= \boldsymbol{x}[n] - \boldsymbol{W}_n^{\mathrm{H}}\mathbf{x}_{n-\Delta} \quad , \tag{3.19}$$

with

$$\boldsymbol{W}_n = \begin{bmatrix} \boldsymbol{W}[n,0] \\ \boldsymbol{W}[n,1] \\ \vdots \\ \boldsymbol{W}[n,N-1] \end{bmatrix} \in \mathbb{C}^{MN\times M} \quad \text{and} \quad \mathbf{x}_n = \begin{bmatrix} \boldsymbol{x}[n] \\ \boldsymbol{x}[n-1] \\ \vdots \\ \boldsymbol{x}[n-N+1] \end{bmatrix} \in \mathbb{C}^{MN} \quad . \tag{3.20}$$

The matrices $\boldsymbol{W}[n,\nu] \in \mathbb{C}^{M\times M}$ describe the predictor's reliance on all spatial measurements taken $\nu+\Delta$ samples in the past, at time instance n.

Following the same analysis as Equations (3.5)–(3.12) yields the Wiener solution

$$\boldsymbol{W}_{n,\text{opt}} = \mathbf{R}_{\mathbf{xx}}^{-1}[n]\mathbf{R}_{x\mathbf{x}}^{\text{H}}[n] \quad . \tag{3.21}$$

Continuous Directional Regimes

To condition the predictor on wind direction, the time dependent covariance matrix $\mathbf{R}_{xx}[n,\tau]$ is estimated by including only historic data for periods when the wind direction was similar to that of, or in the same *directional regime* as, the most recent measurements. A *continuous directional regime* refers to the sliding range of angles, 2Θ, centred on the most recent measurement of wind direction. Each mini-series of $N+1$ samples (corresponding to the concatenation of $\boldsymbol{x}[n]$ and $\mathbf{x}_{n-\Delta}$ in (3.19)) contributing to the estimation of $\boldsymbol{W}_{n,\text{opt}}$ are assumed to be jointly stationary.

Each historic measurement $\boldsymbol{x}[i]$ that is in the same directional regime of the most recent measurement must be accompanied by its N preceding samples which may not lie within the current regime, therefore define $\tilde{\mathbf{R}}_{xx}[n,\delta,\tau] = \mathbf{R}_{xx}[n-\delta,\tau]$ before proceeding.

The estimation of the spatial covariance matrix can now be written as

$$\tilde{\mathbf{R}}_{xx}[n,\delta,\tau] = \frac{1}{|P[n]|}\sum_{i\in P[n]} \boldsymbol{x}[i-\delta]\boldsymbol{x}^{\text{H}}[i-\delta-\tau] \quad , \tag{3.22}$$

where $|P[n]|$ is the cardinality of the set $P[n]$ containing the time indexes p that satisfy

$$\left| \left(\overline{\arg \boldsymbol{x}[p]} - \overline{\arg \boldsymbol{x}[n]} \right) \mod (-\pi, \pi] \right| < \Theta \quad , \tag{3.23}$$

where $\arg \boldsymbol{x}[i] \in (-\pi, \pi]$ and $\overline{\arg \boldsymbol{x}[n]}$ denotes the circular mean of $\arg \boldsymbol{x}[n]$.

Using this covariance matrix, Equation (3.5) is rewritten as

$$\tilde{\mathbf{R}}_{ee}[n] = \tilde{\mathbf{R}}_{xx}[n,0,0] - \tilde{\mathbf{R}}_{\mathbf{x}\mathbf{x}}[n]\boldsymbol{W}_n - \boldsymbol{W}_n^{\mathrm{H}}\tilde{\mathbf{R}}_{x\mathbf{x}}^{\mathrm{H}}[n] + \mathbf{W}_n^{\mathrm{H}}\tilde{\mathbf{R}}_{\mathbf{x}\mathbf{x}}[n]\boldsymbol{W}_n \quad , \tag{3.24}$$

where

$$\tilde{\mathbf{R}}_{x\mathbf{x}}[n] = \left[\tilde{\mathbf{R}}_{xx}[n,0,\Delta] \ , \quad \tilde{\mathbf{R}}_{xx}[n,0,\Delta+1] \ , \quad \ldots \ , \quad \tilde{\mathbf{R}}_{xx}[n,0,\Delta+N-1] \right] \tag{3.25}$$

$$\tilde{\mathbf{R}}_{\mathbf{x}\mathbf{x}}[n] = \begin{bmatrix} \tilde{\mathbf{R}}_{xx}[n,\Delta,0] & \ldots & \tilde{\mathbf{R}}_{xx}[n,\Delta,N-1] \\ \tilde{\mathbf{R}}_{xx}[n,\Delta+1,-1] & & \tilde{\mathbf{R}}_{xx}[n,\Delta+1,N-2] \\ \vdots & \ddots & \vdots \\ \tilde{\mathbf{R}}_{xx}[n,\Delta+N-1,-N+1] & \ldots & \tilde{\mathbf{R}}_{xx}[n,\Delta+N-1,0] \end{bmatrix} \quad . \tag{3.26}$$

Finally the Wiener-Hopf solution (3.21) becomes

$$\tilde{\boldsymbol{W}}_{n,\mathrm{opt}} = \tilde{\mathbf{R}}_{\mathbf{x}\mathbf{x}}^{-1}[n]\tilde{\mathbf{R}}_{x\mathbf{x}}^{\mathrm{H}}[n] \quad , \tag{3.27}$$

By estimating the spatial covariance for a specific directional regime, the propagation of changes in wind speed and direction from upwind to downwind sites can be captured. The inclusion of mismatched information corresponding to periods during which the wind direction was significantly different to the present, which would have the effect of smoothing, or at least skewing the directional dependence of the predictor, is avoided.

The regime specific optimal predictor can be recalculated at each time step or at regular intervals to reduce computational expense at little cost in accuracy.

3.2.2 Testing and Results

The proposed method will be compared to the cyclo-stationary Wiener filter (CSWF) of Section 3.1 and [157], and persistence. Persistence predicts that the future wind speed will be the same as the most recent measurement. Like the direction based approach described in this section, the CSWF also make a quasi-stationary assumption but this time based on the cyclic seasonal behaviour of the wind; the space-time covariance is estimated using historic data from the same season as the current prediction.

The cyclo-stationary covariance matrix is estimated as

$$\hat{\mathbf{R}}_{xx}[n,\tau] = \tfrac{1}{K(L+1)} \sum_{k=1}^{K} \left(\sum_{\nu=\frac{-L}{2}}^{\frac{L}{2}} \boldsymbol{x}[n-kT-\nu]\boldsymbol{x}^{\mathrm{H}}[n-kT-\nu-\tau] \right) + \frac{2}{L} \sum_{\nu=1}^{\frac{L}{2}} \boldsymbol{x}[n-\nu]\boldsymbol{x}^{\mathrm{H}}[n-\nu-\tau] \quad , \quad (3.28)$$

where L is the length of each cyclo-stationary window, K is the number years of training data being used, and T is the period of the cyclo-stationary, i.e. 1 year. For the dataset in question, $L = 20$ weeks was found to be optimal and $K = 5$ to make use of all available training data.

Test Data

The data used for testing is from the Hydra dataset [3] of hourly mean potential wind at multiple locations across the Netherlands, shown in Figure3.10. Data from 2001–2005 inclusive is used as training data and data from 2006 is used for testing.

The measured wind speed has been corrected for the effects of shelter from buildings or vegetation. The resulting *potential* wind is an estimate of the wind speed that could have been measured at 10m height if the station's surroundings were free of obstacles and flat with a roughness length equal to that of grass onshore (0.03m) and water offshore (0.002m). For more information on this process see [3].

This transformation aids spatial prediction by removing biases present at individual measurement locations that would otherwise interfere with the spatio-temporal corre-

lation of the data. The procedure is simple to implement once information regarding the terrain surrounding a weather station is known.

In order to assess the performance of the proposed predictor on spatial datasets of different sizes, it is tested first on 4 central locations and then on larger datasets with sites added progressively beginning with those closest to the original 4, illustrated in Figure 3.11.

Results

The range of wind direction in a regime, Θ, is taken to be $\frac{2\pi}{3}$ since the performance at this range is found by numerical testing to yield better results than $\frac{\pi}{3}$ and π. Given the large range of wind direction, the improvement in prediction is perhaps best thought of as due to the exclusion of mismatched data, rather than the inclusion of well matched data. The number of historic samples, $|P|$, that contribute to the estimation of the covariance matrix for a given regime ranges from 33 230 to 37 043 depending on sites in the data model and the wind direction.

The order of regression N is taken to be 3 since any more significantly increases the computational complexity for negligible reduction in prediction error. For the same reason, the covariance matrix is only recalculated every 24 time steps, i.e. once per day.

The performance of the 1-hour-ahead ($\Delta = 1$) forecast in terms of root mean squared error (RMSE) for the CDR and CSWF predictors is plotted in Figure3.11 for data models containing information from between 4 and 27 sites. The proposed CDR predictions are consistently more accurate than the CSWF, but only by a small margin.

There is a clear reduction in RMSE at all prediction locations as the amount of spatial information is increased. Particularly large improvements are seen at specific sites when new data from nearby locations is added; for example, at site 260 when sites 240, 248 and 256 are added to the data model. Site 249 also sees marked improvement when a number of surrounding sites are included in the data model. Time series plots of the 1-hour-ahead predictions produced by predictors built using 4 and 27 sites are

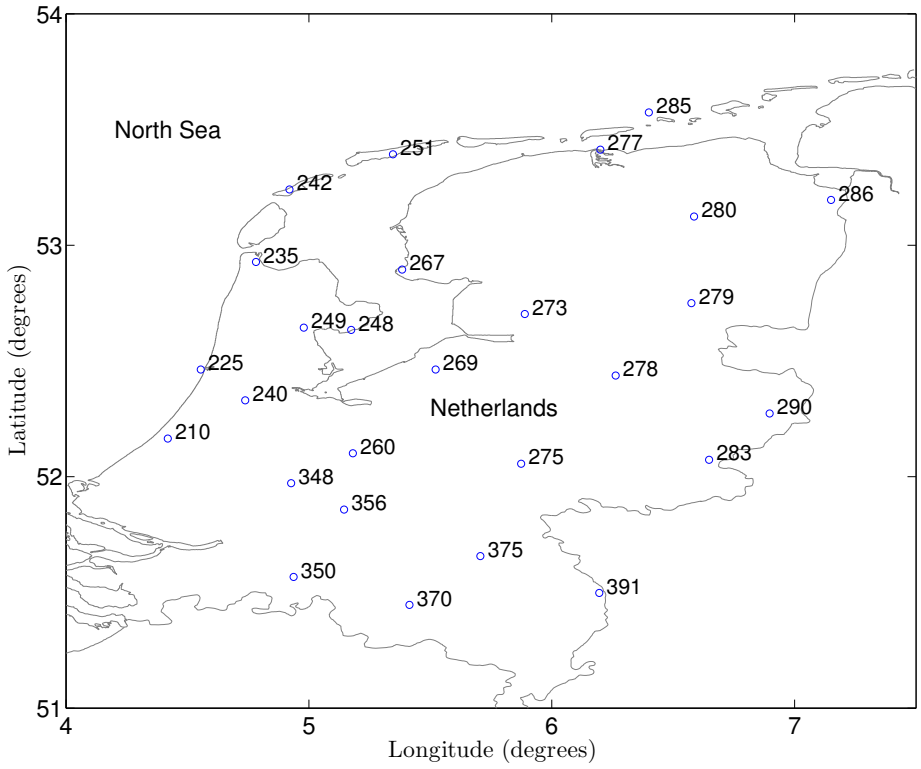


Figure 3.10: Map of the Netherlands showing the location of weather stations and their reference numbers.

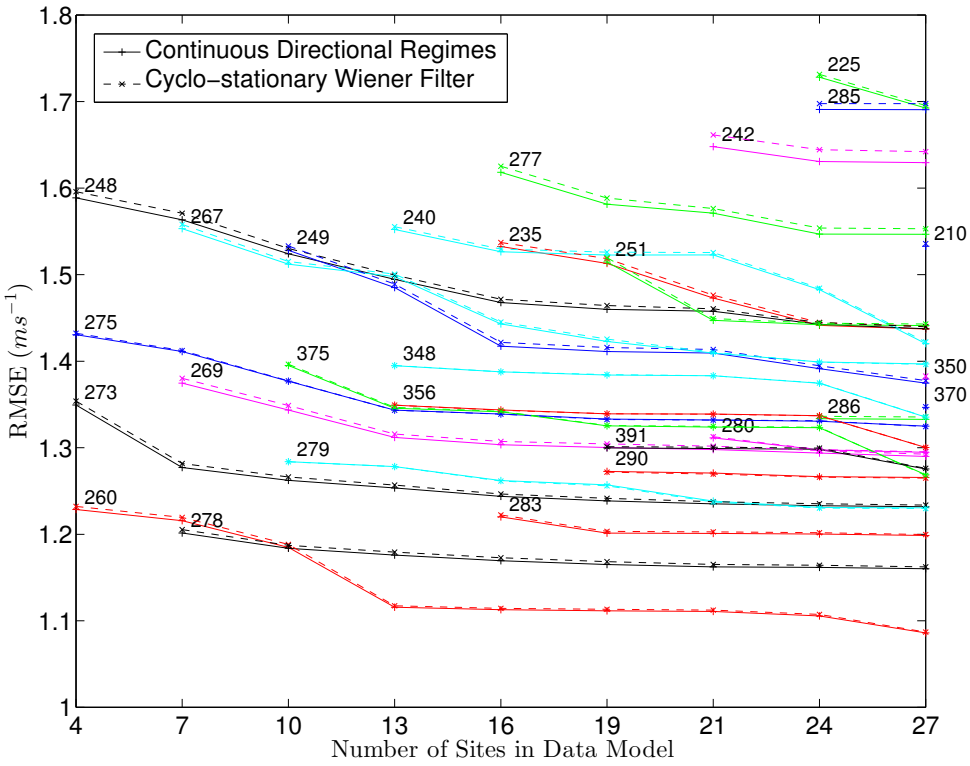


Figure 3.11: Root Mean Squared Error (RMSE) for 1-hour-ahead forecast at sites, labelled by station number, for data models containing 4 to 27 sites.

illustrated in Figures 3.12 and 3.13 showing complex-valued wind and wind speed, respectively.

The performance of the directional predictor is compared to persistence for look-ahead times from 1 to 6 hours in Table 3.2. The CDR is an improvement on persistence at all look-ahead times, with approximately twice the reduction in RMSE for the 27 site data model compared to that containing only 4 sites.

3.2.3 Summary

This work proposes a new spatio-temporal predictor for hourly mean wind speed and direction at multiple measurement locations. Inspired by approaches which define fixed, discrete *regimes* based on wind direction, an adaptive predictor based on continuous direction regimes (CDR) is derived and tested, and shown to produce accurate forecasts

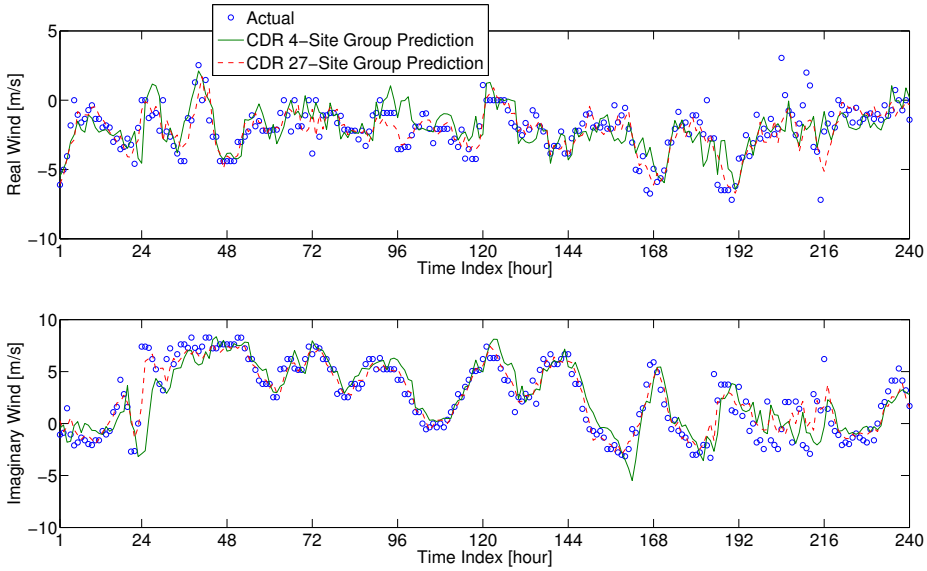


Figure 3.12: Real and imaginary parts of the 1-hour-ahead wind forecasts for site 248 (Wijdenes) starting at 3pm on 2nd August 2006. Predictions from a model containing 4 sites and a model containing 27 sites are displayed.

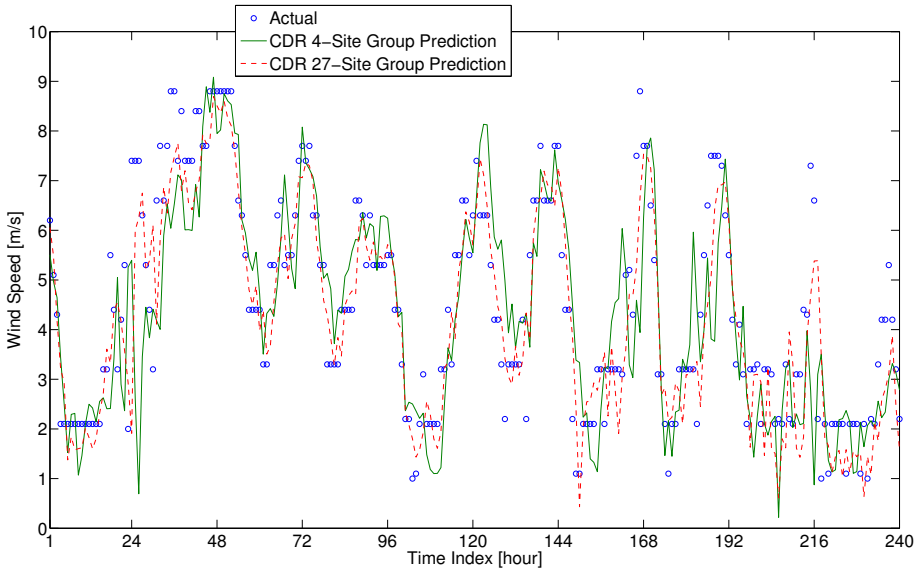


Figure 3.13: 1-hour-ahead wind speed forecasts for site 248 (Wijdenes) starting at 3pm on 2nd August 2006. Predictions from a model containing 4 sites and a model containing 27 sites are displayed.

Location	Δ	RMSE (ms^{-1}) Persistence	Data Model 4 Sites	Data Model 27 Sites
248: Wijdenes	1	1.68	1.59	1.44
	2	2.26	2.10	1.87
	3	2.69	2.50	2.22
	4	3.05	2.83	2.54
	5	3.36	3.12	2.84
	6	3.63	3.36	3.10
260: De Bilt	1	1.37	1.23	1.09
	2	1.73	1.53	1.34
	3	2.03	1.79	1.58
	4	2.28	2.01	1.80
	5	2.51	2.21	2.00
	6	2.70	2.37	2.18
273: Marknesse	1	1.55	1.35	1.23
	2	2.05	1.74	1.53
	3	2.46	2.09	1.82
	4	2.81	2.39	2.10
	5	3.11	2.66	2.37
	6	3.37	2.88	2.61
275: Deelen	1	1.66	1.43	1.32
	2	2.13	1.78	1.61
	3	2.51	2.09	1.88
	4	2.82	2.35	2.14
	5	3.10	2.59	2.37
	6	3.34	2.80	2.58

Table 3.2: Comparison of CDR Root Mean Squared Error at the 4 sites in the smallest data model to persistence and when included in a larger data model at look-ahead times from $\Delta = 1$ to 6 hours.

for look-ahead times of 1 to 6 hours.

The CDR is a spatial covariance-based minimum MSE predictor, it is innovative in its selection of training data in real time to exclude mismatched historic data based on the most recent measurements. This approach was motivated by the idea that spatial dependence between wind speed measurements at different locations depends on the wind direction. The spatial covariance matrix is estimated using only data from periods during which the wind direction was within a fixed range of its present direction from which the adaptive predictor is calculated.

The new predictor is tested on the Hydra dataset and compared to persistence and the cyclo-stationary Wiener filter, another spatial-covariance-based adaptive predictor. The CDR is found to produce forecasts which are a significant improvement on persistence and consistently more accurate than the CSWF, if only by a small margin. Furthermore, it is shown that the prediction error is reduced as more spatial information is added to the data model.

While it is relatively crude, the proposed method performs well and provide encouraging support for the future refinement of this type of approach, perhaps building-in constraints on wind speed or choosing specific measurement sites to improve prediction at some target location.

3.3 Augmented Wiener Filter

So far in this chapter, wind speed and direction have been forecast in a complex-linear framework with the goal of producing simple and efficient predictors that capitalise on spatial information. Various approaches have been developed to condition the predictor on physical conditions: season and average wind direction. In this final development, the structure of the input data is examined for the possibility of sacrificing some computational efficiency for improved performance.

Linear operations have been applied to complex quantities (then termed strictly linear or $\mathbb{C}$-linear) in exactly the same way as to real ones, but with some limitations

that must be appreciated. Consider the $\mathbb{C}$-linear transformation

$$y = kx, \qquad x, y, k \in \mathbb{C} \tag{3.29}$$

with $x = x_r + jx_i$ and $y = y_r + jy_i$ where $x_r, x_i, y_r, y_i \in \mathbb{R}$. Writing the product in terms of its real and imaginary parts

$$\begin{bmatrix} y_r \\ y_i \end{bmatrix} = \begin{bmatrix} \text{Re } k & -\text{Im } k \\ \text{Im } k & \text{Re } k \end{bmatrix} \begin{bmatrix} x_r \\ x_i \end{bmatrix} \tag{3.30}$$

and comparing that to the more general $\mathbb{R}^2$ transformation

$$\begin{bmatrix} y_r \\ y_i \end{bmatrix} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} x_r \\ x_i \end{bmatrix} \tag{3.31}$$

illustrates the limitation: the $\mathbb{R}^2$ transformation is only $\mathbb{C}$-linear iff $M_{11} = M_{22}$ and $M_{12} = -M_{21}$. The complex equivalent of (3.31) is the *widely linear* transformation

$$y = k_1 x + k_2 x^* \quad . \tag{3.32}$$

Detailed discussions on widely linear processing can be found in [164, 165, 177].

Wind measurements have been modelled as $\mathbb{C}$-linear and cyclo-stationary in Section 3.1 and [157], and as $\mathbb{C}$-linear and conditionally stationary in Section 3.2 and [159] with computational efficiency in mind. However, in order to quantify any improvements in performance that could be achieved by sacrificing complexity, this section explores a widely linear model, which comes at the expense of doubling the filter order.

The data model is described in Section 3.3.1 and the minimum mean squared error predictor is derived in 3.3.1, with the cyclo-stationary estimation of the covariance matrices outlined in 3.3.1. The data used for testing and test results are presented in Section 3.3.3 and a summary is provided in 3.3.4.

3.3.1 Data Model and Prediction

At discrete time n, the wind speed and direction at M locations are embedded as the magnitude and phase of a complex valued vector $\boldsymbol{x}[n] \in \mathbb{C}^M$. The spatial covariance matrix is defined based on the expectation operator, $E\{\cdot\}$, as $\mathbf{R}_{xx}[n,\tau] = E\{\boldsymbol{x}[n]\boldsymbol{x}^{\mathrm{H}}[n-\tau]\}$, where $\boldsymbol{x}^{\mathrm{H}}[n]$ denotes the Hermitian transpose of $\boldsymbol{x}[n]$ and τ is a general lag parameter.

Furthermore, in widely linear processing it is useful to also define the complementary covariance matrix based on the expectation operator as $\tilde{\mathbf{R}}_{xx}[n,\tau] = E\{\boldsymbol{x}[n]\boldsymbol{x}^{\mathrm{T}}[n-\tau]\}$. In addition, by considering the *augmented* vector $\underline{\boldsymbol{x}}[n]$, which is the concatenation of $\boldsymbol{x}[n]$ and its conjugate, the augmented covariance matrix is defined as $\underline{\mathbf{R}}_{xx}[n,\tau] = E\{\underline{\boldsymbol{x}}[n]\underline{\boldsymbol{x}}^{\mathrm{H}}[n-\tau]\}$,

$$\begin{aligned}\underline{\mathbf{R}}_{xx}[n,\tau] &= E\left\{\begin{bmatrix}\boldsymbol{x}[n]\\ \boldsymbol{x}^*[n]\end{bmatrix}\begin{bmatrix}\boldsymbol{x}^{\mathrm{H}}[n-\tau] & \boldsymbol{x}^{\mathrm{T}}[n-\tau]\end{bmatrix}\right\}\\ &= \begin{bmatrix}\mathbf{R}_{xx}[n,\tau] & \tilde{\mathbf{R}}_{xx}[n,\tau]\\ \tilde{\mathbf{R}}^*_{xx}[n,\tau] & \mathbf{R}^*_{xx}[n,\tau]\end{bmatrix} \quad . \end{aligned} \tag{3.33}$$

Notice that since $\underline{\mathbf{R}}_{xx}$ is positive semi-definite, and therefore has a non-negative determinant, the limit $|\mathbf{R}_{xx}|^2 \geq |\tilde{\mathbf{R}}_{xx}|^2$ follows and sets an upper bound for the determinant of $\tilde{\mathbf{R}}_{xx}$.

It is well known that wind speed and wind direction are likely non-stationary and non-linear, both can be volatile and, direction particularly, can depend heavily on the physical characteristics of the measurement site. Furthermore, the seasonal and diurnal trends that characterise our human experience of the wind are themselves variable. In the succeeding text, the potential non-linear nature of the wind is ignored and linearity is assumed. The assumption of stationarity is dropped for a quasi-stationary behaviour, whereby the space-time covariance matrix can be assumed to be stationary — and therefore only dependent on the lag parameter τ — for sufficiently short time windows [167].

MMSE Predictor

We consider the problem of predicting Δ samples ahead while minimising the mean-squared prediction error (MSE), based on M spatial measurements in $\boldsymbol{x}[n]$ and a time window containing N past samples for each site, plus the complex conjugates of the same. Therefore, the prediction error can be formulated as

$$\boldsymbol{e}[n] = \boldsymbol{x}[n] - \sum_{\nu=0}^{N-1} \left(\boldsymbol{P}^{\mathrm{H}}[n,\nu]\boldsymbol{x}[n-\Delta-\nu] + \boldsymbol{Q}^{\mathrm{H}}[n,\nu]\boldsymbol{x}^*[n-\Delta-\nu] \right) \quad (3.34)$$

$$= \boldsymbol{x}[n] - \boldsymbol{W}_n^{\mathrm{H}} \underline{\mathbf{x}}_{n-\Delta} \quad , \quad (3.35)$$

with

$$\boldsymbol{W}_n = \begin{bmatrix} \boldsymbol{P}[n,0] \\ \boldsymbol{P}[n,1] \\ \vdots \\ \boldsymbol{P}[n,N-1] \\ \boldsymbol{Q}[n,0] \\ \boldsymbol{Q}[n,1] \\ \vdots \\ \boldsymbol{Q}[n,N-1] \end{bmatrix}, \quad \underline{\mathbf{x}}_n = \begin{bmatrix} \boldsymbol{x}[n] \\ \boldsymbol{x}[n-1] \\ \vdots \\ \boldsymbol{x}[n-N+1] \\ \boldsymbol{x}^*[n] \\ \boldsymbol{x}^*[n-1] \\ \vdots \\ \boldsymbol{x}^*[n-N+1] \end{bmatrix} \quad (3.36)$$

The matrices $\boldsymbol{P}[n,\nu], \boldsymbol{Q}[n,\nu] \in \mathbb{C}^{M\times M}$ describe the predictor's reliance on all spatial measurements and their conjugates, respectively, taken $\nu+\Delta$ samples in the past, at time instance n.

The error covariance matrix derived from (3.35), $\mathbf{R}_{ee}[n] = E\{\boldsymbol{e}[n]\boldsymbol{e}^{\mathrm{H}}[n]\} \in \mathbb{C}^{M\times M}$, is obtained by taking expectations over the ensemble, and in itself may be varying with time n. Note that in case of stationarity, the dependency of both $\boldsymbol{W}_n$ and $\mathbf{R}_{ee}[n]$ on n vanishes. We will carry forward n since it is well known that the wind signal is non-stationary and develop an approximately stationary solution in Section 3.3.1. Calculating $\mathbf{R}_{ee}[n]$ using (3.35) yields a quadratic expression in $\boldsymbol{W}_n$,

$$\begin{aligned}
\mathbf{R}_{ee}[n] &= E\left\{(\boldsymbol{x}[n] - \boldsymbol{W}_n^{\mathrm{H}}\underline{\mathbf{x}}_{n-\Delta})(\boldsymbol{x}^{\mathrm{H}}[n] - \underline{\mathbf{x}}_{n-\Delta}^{\mathrm{H}}\boldsymbol{W}_n)\right\} \quad , \\
&= \mathbf{R}_{xx}[n,0] - E\{\boldsymbol{x}[n]\underline{\mathbf{x}}_{n-\Delta}^{\mathrm{H}}\}\boldsymbol{W}_n - \boldsymbol{W}_n^{\mathrm{H}}E\{\underline{\mathbf{x}}_{n-\Delta}\boldsymbol{x}^{\mathrm{H}}[n]\} + \boldsymbol{W}_n^{\mathrm{H}}E\{\underline{\mathbf{x}}_{n-\Delta}\underline{\mathbf{x}}_{n-\Delta}^{\mathrm{H}}\}\boldsymbol{W}_n \quad , \\
&= \mathbf{R}_{xx}[n,0] - \underline{\mathbf{R}}_{x\underline{\mathbf{x}}}[n]\boldsymbol{W}_n - \boldsymbol{W}_n^{\mathrm{H}}\underline{\mathbf{R}}_{x\underline{\mathbf{x}}}^{\mathrm{H}}[n] + \boldsymbol{W}_n^{\mathrm{H}}\underline{\mathbf{R}}_{\mathbf{xx}}[n]\boldsymbol{W}_n \quad , \qquad (3.37)
\end{aligned}$$

where

$$\begin{aligned}
\underline{\mathbf{R}}_{x\mathbf{x}}[n] &= \big[\ \mathbf{R}_{xx}[n,\Delta]\ ,\ \ \mathbf{R}_{xx}[n,\Delta{+}1]\ ,\ \ \ldots\ ,\ \ \mathbf{R}_{xx}[n,\Delta{+}N{-}1]\ , \\
&\qquad \tilde{\mathbf{R}}_{xx}[n,\Delta]\ ,\ \ \tilde{\mathbf{R}}_{xx}[n,\Delta{+}1]\ ,\ \ \ldots\ ,\ \ \tilde{\mathbf{R}}_{xx}[n,\Delta{+}N{-}1]\ \big] \quad , \qquad (3.38) \\
\underline{\mathbf{R}}_{\mathbf{xx}}[n] &= \begin{bmatrix} \mathbf{R}_{\mathbf{xx}}[n] & \tilde{\mathbf{R}}_{\mathbf{xx}}[n] \\ \tilde{\mathbf{R}}_{\mathbf{xx}}^{*}[n] & \mathbf{R}_{\mathbf{xx}}^{*}[n] \end{bmatrix}, \qquad (3.39) \\
\mathbf{R}_{\mathbf{xx}}[n] &= \begin{bmatrix} \mathbf{R}_{xx}[n{-}\Delta,0] & \ldots & \mathbf{R}_{xx}[n{-}\Delta,N{-}1] \\ \mathbf{R}_{xx}[n{-}\Delta{-}1,-1] & & \mathbf{R}_{xx}[n{-}\Delta{-}1,N{-}2] \\ \vdots & \ddots & \vdots \\ \mathbf{R}_{xx}[n{-}\Delta{-}N{+}1,-N{+}1] & \ldots & \mathbf{R}_{xx}[n{-}\Delta{-}N{+}1,0] \end{bmatrix}. \qquad (3.40)
\end{aligned}$$

We assume that $\boldsymbol{x}[n]$ is stationary over at least 2Δ samples. As a result, $\underline{\mathbf{R}}_{\mathbf{xx}}[n]$ is Hermitian and therefore positive semi-definite [168]. This property together with full rank of $\underline{\mathbf{R}}_{\mathbf{xx}}[n]$ admits a unique solution to minimises the mean square error,

$$\boldsymbol{W}_{n,\mathrm{opt}} = \arg\min_{\boldsymbol{W}_n} \mathrm{trace}\{\mathbf{R}_{ee}[n]\} \quad . \qquad (3.41)$$

It can be shown that $\mathrm{trace}\{\mathbf{R}_{ee}[n]\}$ is quadratic in $\boldsymbol{W}_n$, such that the solution to (3.41) can be found by matrix- and complex-valued calculus [169]. Finding the minimum requires equating the gradient with respect to the unconjugated predictor coefficients in $\boldsymbol{W}_n^*$ to zero. We utilise results from [169] which show that for constant matrices $\mathbf{A}$ and $\mathbf{B}$ the expressions

$$\partial\mathrm{trace}\{\mathbf{A}\boldsymbol{W}_n^{\mathrm{H}}\mathbf{B}\}/(\partial\boldsymbol{W}_n^{*}) = \mathbf{B}\mathbf{A} \qquad (3.42)$$

and

$$\partial \text{trace}\{\mathbf{A}\boldsymbol{W}_n\mathbf{B}\}/(\partial \boldsymbol{W}_n^*) = \mathbf{0} \tag{3.43}$$

hold. Applying this, and using the product rule for differentiation of the quadratic term in (3.37), yields

$$\frac{\partial}{\partial \boldsymbol{W}_n^*}\text{trace}\{\mathbf{R}_{ee}[n]\} = -\underline{\mathbf{R}}_{x\mathbf{x}}^{\text{H}}[n] + \underline{\mathbf{R}}_{\mathbf{xx}}[n]\boldsymbol{W}_n \quad . \tag{3.44}$$

Finally, setting the gradient on the right-hand side of (3.44) equal to zero yields the optimum predictor coefficients that minimise $\text{trace}\{\mathbf{R}_{ee}[n]\}$,

$$\boldsymbol{W}_{n,\text{opt}} = \underline{\mathbf{R}}_{\mathbf{xx}}^{-1}[n]\underline{\mathbf{R}}_{x\mathbf{x}}^{\text{H}}[n] \quad , \tag{3.45}$$

which is the well-known Wiener-Hopf solution [170, 171].

If the process $\boldsymbol{x}[n]$ is uncorrelated with its conjugate, i.e. $\tilde{\mathbf{R}}_{xx} = \mathbf{0}$, all the matrices $\boldsymbol{Q}[n,\nu] = \mathbf{0}$ and the prediction problem reduces to the $\mathbb{C}$-linear case.

Cyclo-Stationary Covariance Matrix

The cyclo-stationary covariance matrix (and its associated complementary covariance matrix) is formulated based on the assumption that windows of data of length L are approximately stationary, and furthermore, that the statistics of that period are the same during the equivalent window in all years. The covariance matrix $\mathbf{R}_{xx}[n,\tau]$ is estimated by calculating the expectation using only data in the quasi-stationary window centred on n from each year of available training data. In the estimation of $\mathbf{R}_{xx}[n,\tau]$, assume cyclo-stationarity, i.e. $\mathbf{R}_{xx}[n,\tau] = \mathbf{R}_{xx}[n-kT,\tau]$, with $k \in \mathbb{N}$ and T the fundamental period, i.e. 1 year. On the basis of cyclo-stationarity and data available

for K past years, the estimation of the covariance matrix for time n is performed as

$$\hat{\mathbf{R}}_{xx}[n,\tau] = \tfrac{1}{K(L+1)} \sum_{k=1}^{K} \left(\sum_{\nu=\frac{-L}{2}}^{\frac{L}{2}} \boldsymbol{x}[n-kT-\nu]\boldsymbol{x}^{\mathrm{H}}[n-kT-\nu-\tau] \right) + \frac{2}{L} \sum_{\nu=1}^{\frac{L}{2}} \boldsymbol{x}[n-\nu]\boldsymbol{x}^{\mathrm{H}}[n-\nu-\tau] \quad , \quad (3.46)$$

and the complementary covariance matrix is calculated in the same way but with the Hermitian transpositions replaced by standard transpositions. The widely linear optimal prediction filter for time n can then be calculated by replacing the quantities in the Wiener solution (3.45) by their estimates derived from (3.46) inserted into (3.38)–(3.40).

3.3.2 Testing and Results

Test Data

The proposed approach is tested on wind data provided by the British Atmospheric Data Centre, which comprises of recordings over 6 years — from 00:00h on 1/3/1992 to 23:00h on 28/2/1998 — obtained from 13 sites across the UK. The measurements are taken in open terrain at a height of 10m and sampled at hourly intervals, comprise hourly averages that are quantised to a 10° angular granularity and integer multiples of one knot ($0.515\mathrm{ms}^{-1}$) [2].

Widely linear processing is advantageous for improper signals, or cross-improper in the multichannel case, i.e. if $\tilde{\mathbf{R}}_{xx} \neq \mathbf{0}$. The statistical hypothesis test for the impropriety of complex vectors described in [178] has been applied to the test data. The test unambiguously rejected the hypothesis $H_0 \ : \ \tilde{\mathbf{R}}_{xx} = \mathbf{0}$ in favour of $H_1 \ : \ \tilde{\mathbf{R}}_{xx} \neq \mathbf{0}$ indicating that the data is improper and therefore that widely linear processing is appropriate. The test for impropriety is described in Appendix B.3.

Cyclo-stationary Estimation

In the estimation of the cyclo-stationary covariance matrix, (3.46), $K = 5$ to make use of all available training data and the optimal window length L is chosen heuristically to be 15 weeks. The filter length is chosen to be $2N = 6$ since the gains from increasing it further are negligible.

3.3.3 Results

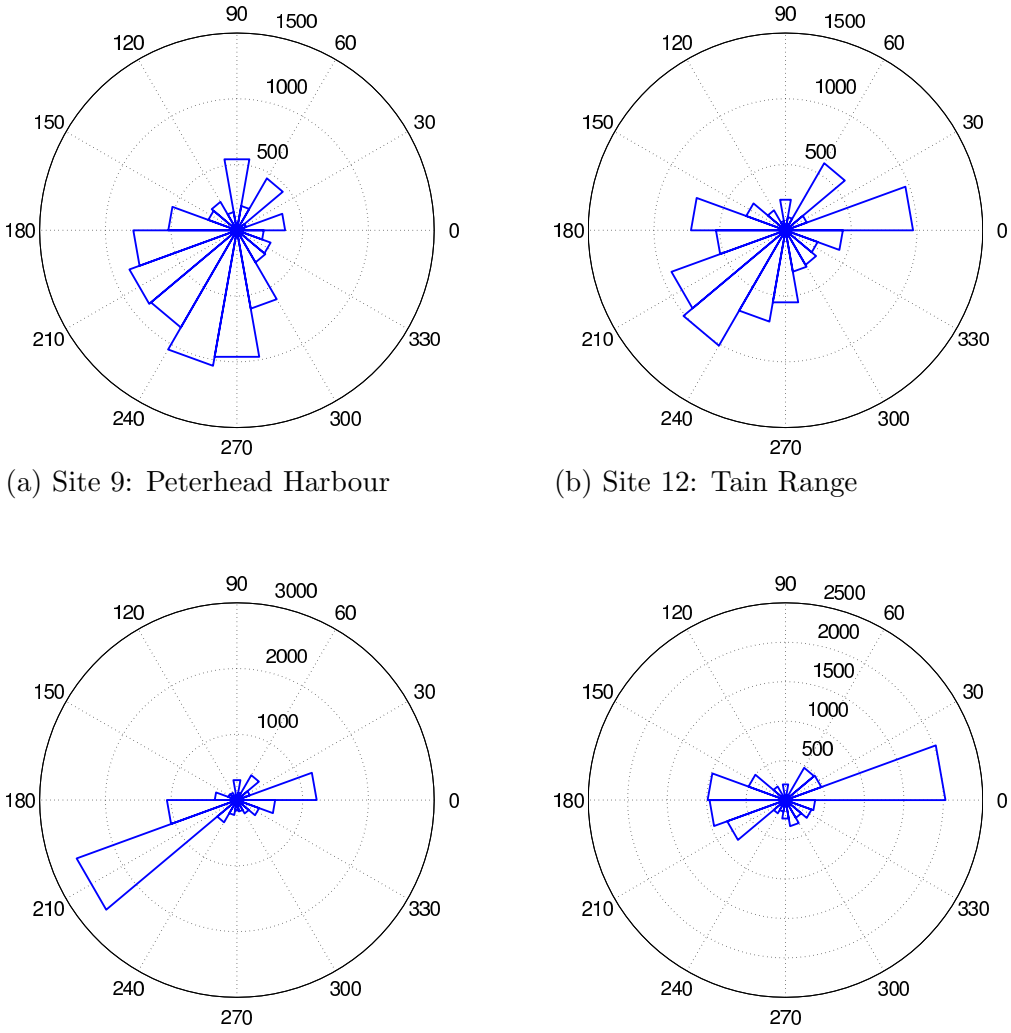
The widely linear predictor yields improved prediction performance in terms of root mean squared error for all 13 channels and at all look-ahead times. As one would expect the new predictor yields greater improvement over its $\mathbb{C}$-linear equivalent at sites with larger complementary correlation and lower directional variance.

Results from two channels that showed the least improvement (9 and 12) and the two that showed the most (7 and 10) are detailed in Table 3.3. The distribution of the arguments of these four channels are illustrated by the histograms in Figure 3.14. The sites in Figures 3.14a and 3.14b show little improvement and have arguments, or wind directions, spread evenly over a wide range of angles, whereas the sites in Figures 3.14c and 3.14d demonstrate large improvement and have very narrow distributions, corresponding to low directional variance and high complementary correlation. Plots of the 1- and 6-hour-ahead predictions are illustrated in Figures 3.15 and 3.16.

The complementary autocorrelation coefficients for all 13 channels are plotted in Figure 3.17. The two channels showing the least improvement over the $\mathbb{C}$-linear predictor have relatively small complementary auto-correlation coefficients where as those showing large improvement have relatively large values. This illustration serves as a crude indication of significance but should not be interpreted as the cause of the difference in performance.

3.3.4 Summary

In this section a multichannel widely linear cyclo-stationary Wiener filter for the prediction of hourly mean wind speed and direction from 1 to 6 hours ahead has been derived and tested. The performance of the proposed filter is compared to that of its



(a) Site 9: Peterhead Harbour

(b) Site 12: Tain Range

(c) Site 7: Leuchars

(d) Site 10: Rhoose

Figure 3.14: Circular histograms of hourly-mean wind direction measurements at 4 selected sites from the 1 year of data used for testing.

Site	Δ	RMSE (ms^{-1}) CSWF	RMSE (ms^{-1}) WLCSWF	% Improvement
9: Peterhead Harbour	1	1.73	1.72	0.5
	2	2.35	2.33	0.8
	3	2.77	2.74	1.0
	4	3.11	3.08	1.0
	5	3.40	3.37	1.1
	6	3.66	3.62	1.0
12: Tain Range	1	1.97	1.96	0.4
	2	2.47	2.45	0.6
	3	2.78	2.75	0.8
	4	3.03	3.00	1.1
	5	3.23	3.19	1.2
	6	3.41	3.36	1.3
7: Leuchars	1	1.52	1.50	1.1
	2	2.07	2.02	2.0
	3	2.43	2.36	2.9
	4	2.72	2.63	3.4
	5	2.98	2.87	3.7
	6	3.20	3.07	3.9
10: Rhoose	1	1.67	1.66	0.4
	2	2.18	2.15	1.3
	3	2.59	2.54	1.8
	4	2.98	2.91	2.3
	5	3.31	3.21	2.7
	6	3.59	3.48	3.0

Table 3.3: Root Mean Squared Errors (RMSE) for the cyclo-stationary (CSWF) and widely linear (WLCSWF) Wiener filters at look-ahead times (Δ) from 1–6 hours.

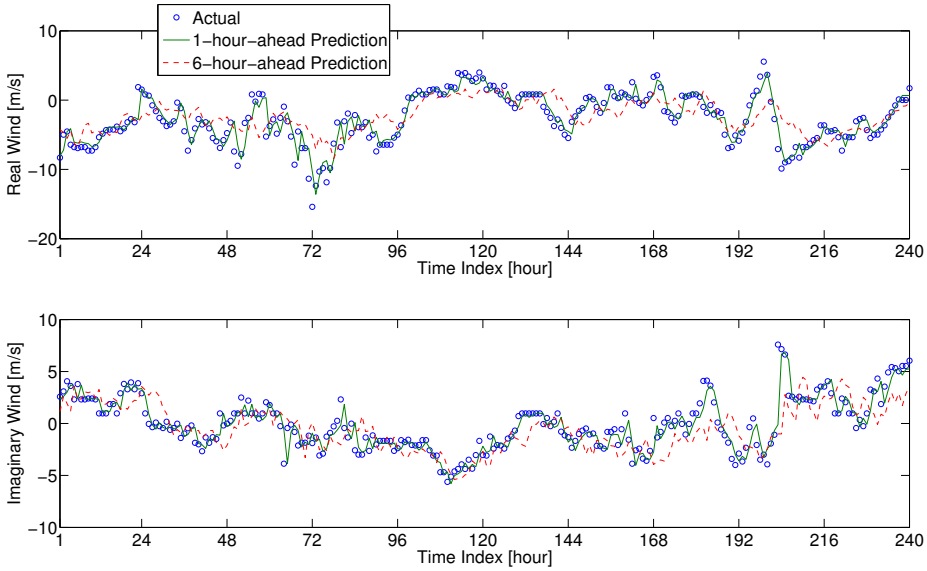


Figure 3.15: Real and imaginary parts of the 1-hour-ahead wind forecasts for Boulmer starting at 3pm on 2nd August 1998.

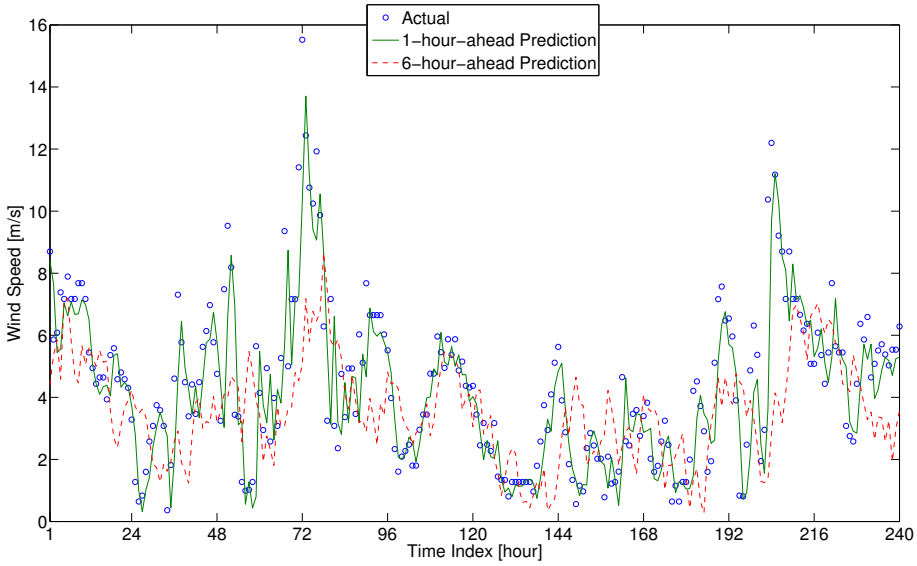


Figure 3.16: 1-hour-ahead wind speed forecasts for Boulmer starting at 3pm on 2nd August 1998.

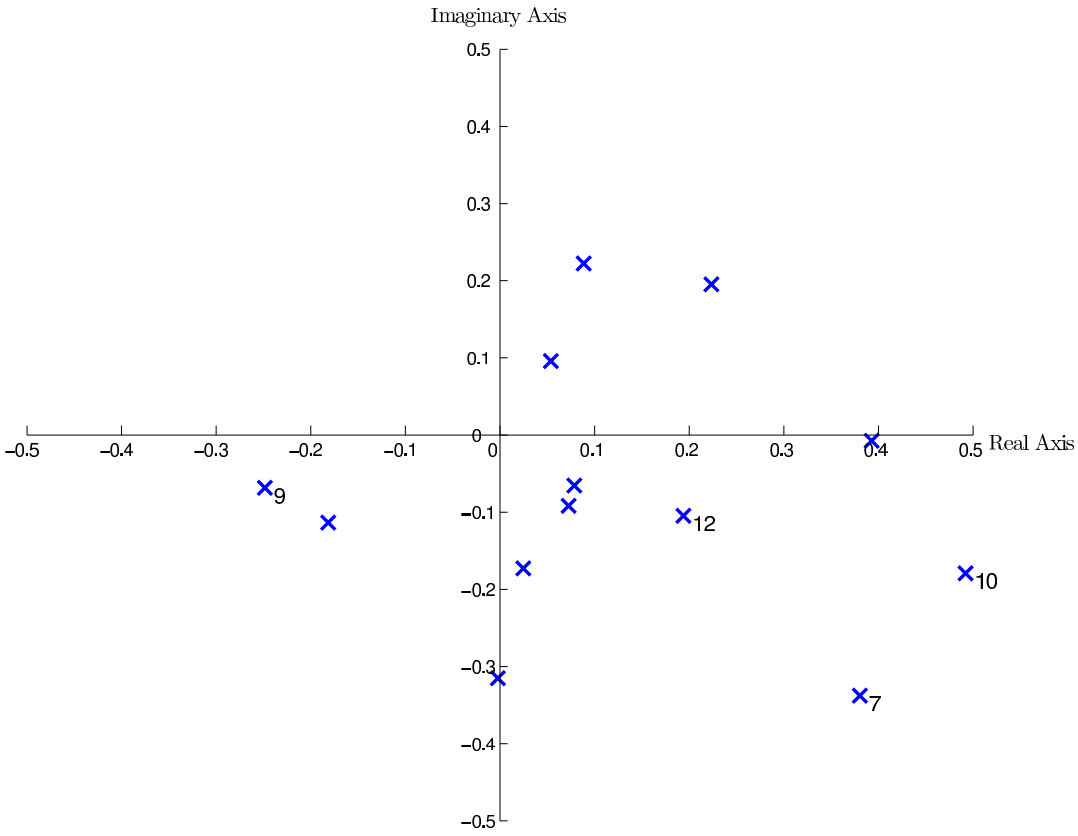


Figure 3.17: Scatter plot of complementary auto-correlation coefficients at zero lag for the 13 measurement locations. The four examples used in Figure3.14 and Table 3.3 are labelled by site number.

$\mathbb{C}$-linear equivalent to quantify the benefits of increasing computational complexity to accommodate the widely linear model.

The widely linear model captures information contained in the complementary auto- and cross-covariance which is inaccessible in a strictly linear formulation. In addition, the cyclo-stationary estimation of the covariance matrices captures the seasonal behaviour of the wind which would otherwise lead to the inclusion of mismatched data in the estimation of the covariance matrices.

The predictors are tested on wind measurements made at 13 locations distributed geographically around the UK over a period of 6 years. The widely linear predictor shows improved prediction versus its $\mathbb{C}$-linear equivalent at all 13 locations. The locations which exhibit greater improvement are those with the least directional variation

and associated high complementary auto-correlation.

3.4 Conclusions

This chapter has introduced linear spatial prediction and developed predictors of wind speed and direction using a complex-valued data model. Linear statistical techniques facilitate spatial modelling at relatively low computational expense and with robust performance. In addition, the linear framework is highly flexible and allows for predictors to be conditioned on exogenous variables. Two such conditional approaches have been developed, one based on the time of year, or season, and the second on average wind direction over a region. It is demonstrated that the conditional approaches outperform the non-conditional predictors, as well as an iterative scheme and persistence.

The seasonal characteristics of the wind have been captured by excluding mismatched data when training a Wiener filter based predictor for a given time of year. For the test dataset, well-matched data was determined do be contained in a time window of length 15 weeks centred on the time of year for which the predictor is being implemented. The windows of training data are assumed quasi-stationary and multiple years of historic data are used to ensure stable parameter estimates. The coefficients of the final predictor vary cyclically with the season and quasi-stationarity of the wind time series hence the name cyclo-stationary Wiener filter.

The performance of the CSWF has been compared to an adaptive approach, the LMS algorithm, a static/stationary Wiener filter and persistence. The LMS algorithm tracks changes in wind dynamics, albeit slowly and with some lag, and the static Wiener filter does not capture any changes in dynamics. The CSWF performed better than all the benchmarks of the one-year test period, with the LMS outperforming the static Wiener filter and persistence. Capturing the seasonal variation in wind dynamics has a positive effect on the performance of the predictors. Furthermore, directly modelling the seasonal variation by selecting only relevant training data is shown to be superior to tracking variation with an adaptive algorithm.

The second method was developed based on conditioning the linear predictor on wind direction, rather than the time of year. In this case, training data are considered

to be well-matched if they correspond to a time when the average wind direction over the region was similar to that at the target time. The proposed predictor depends on direction in a continuous way, and is termed 'continuous directional regimes' — this is a generalisation of so-called 'regime switching' approach, which trained a small number of predictors for specific wind directions.

The CDR predictor demonstrates very similar performance to the CSWF, offering minor improvement at a few sites, and negligible difference in performance at others. In addition, this study demonstrated that performance of both methods is enhanced by the inclusion of more spatial information; in other words, the more locations from a given region in the data model, the more accurate the forecast will be at all sites.

Finally, the complex-linear Wiener filter has been extended to the widely linear Wiener filter to capture additional covariate information available only when both the standard input and it's conjugate are considered. This extension increases the complexity of problem but is shown to yield improved performance at all sites, particularly those with distinct directional regimes where the complex-linear assumption of rotational invariance is least valid. Sites with distinct directional regimes benefit from the full widely-linear treatment, whereas those without gain little from the added complexity.

In this chapter the flexibility of linear time series methods has been exploited to develop several predictors that are able to capitalise on long-term physical characteristics of wind time series. However, it has also exposed the limitations of such approaches: the performance of a linear predictor depends on the relevance of the data used to train it to the moment at which it is being tested. In many real world applications, not least wind modelling, where the assumption of linearity is stretched, breaking a large problem down into smaller problems which better approximate linearity is wise. Any linear approach will at some point reduce to a trade off between extracting only *useful* covariate information and including a sufficient number of training samples to generate reliable parameter estimates.

With that in mind, the next chapter explores two non-linear methods: the first is designed to quickly track changes in wind dynamics, and the second attempts to

identify and exploit non-linear *features* in the wind time series.

Chapter 4

Non-linear Wind Prediction

In the previous chapter making linear assumptions about the wind time series allowed the formulation of optimal predictors in the mean-squares sense. However, the wind is known to be non-linear meaning that assuming linearity excludes potentially useful information from the forecasting problem, and possibly introduces systematic biases. In this chapter, two non-linear predictors are developed in an attempt to better capture the dynamics of the wind.

The first is based on particle swarm optimisation, a *social* algorithm inspired by the behaviour of swarms in nature [179, 180]. While the prediction is still a weighted (linear) sum of the most recent observations, the weights are determined by the swarm, which searches for optimal solutions based on the predictor's recent performance. Each particle in the swarm is a candidate solution that moves around the *problem-space* influenced by its own performance and that of the best performing particle in the swarm. PSO is suited to problems which are irregular, noisy and change over time, unlike conventional recursive approaches based on gradient descent or quasi-Newton methods. It is demonstrated that the ensemble mean of multiple PSO predictors produces more consistent performance than individual PSO predictors.

The second non-linear approach is based on kernel methods, a relatively new class of learning algorithm developed initially for classification problems, but with many other useful applications, including regression and prediction. Kernel methods are characterised by the use of kernel functions and the so-called *kernel trick*: samples of

data are projected into a high-dimensional feature space, via some non-linear function, where linear processing may be more effective than in the original problem space. The kernel trick allows this processing to take place without exact knowledge of the non-linear function since that information is not needed to compute inner-products in the feature space, only the kernel function is required. Here, kernelised forms of the LMS and RLS algorithms are used to produce spatio-temporal wind speed forecasts.

The particle swarm optimisation-based approach is presented in Section 4.1, followed by kernel methods in Section 4.2. Some general conclusions are then drawn in Section 4.3.

4.1 Particle Swarm Optimised FIR Prediction

This section describes an ensemble particle-swarm-optimised filtering technique for 1-hour-ahead prediction of hourly mean wind speed and direction. The performance of the method is assessed by testing it on data from 13 locations around the UK where it performs comparably to linear techniques but is able to provide significant improvement at a subset of locations.

The non-stationarity of the wind can be partially attributed to diurnal and seasonal cycles, which have been modelled in [92, 157], among others, by de-trending and developing conditional predictors, such as the cyclo-stationary Wiener filter of Chapter 3. In addition, synoptic variation (passing weather systems) contributes further non-stationary features. Linear filters satisfy the requirement for low complexity but are limited by their delayed response to changes in wind regime, which occur as a result of changing atmospheric conditions.

Therefore, in this section a prediction method that lifts the linear assumption of the previous chapter is pursued. In particular, particle swarm optimisation has been applied to FIR filters for prediction [142, 179, 181, 182]. These adaptive filters exhibit a good response to sudden changes in wind regime while retaining the ability to track cyclic non-stationarities that have been captured in the linear case. Furthermore, an ensemble of particle swarm optimised FIR filters is found to produce the most consistent 1-hour-ahead prediction.

The wind model and particle swarm optimisation (PSO) algorithm are described in Sections 4.1.1 and 4.1.2, and the application of PSO to the wind model for prediction is detailed in Sections 4.1.2 and 4.1.2. Results from testing the proposed algorithm are presented and discussed in Section 4.1.3 before some conclusions and suggestions for future work are presented in Section 4.1.4.

4.1.1 Wind Model

The hourly mean wind speed and direction at discrete time index t are modelled as the magnitude and phase of a complex random variable, $y[t]$, which is the weighted linear combination of N past measurements of $y[t]$ and some error of unknown statistics, $\epsilon[t]$. The past measurements of $y[t]$ and the complex prediction coefficients, $w_\tau[t]$, are arranged as vectors $\boldsymbol{y}_t$ and $\boldsymbol{w}[t]$ of size N, respectively,

$$y[t] = \sum_{\tau=1}^{N} w_\tau[t] y[t-\tau] + \epsilon[t] = \boldsymbol{w}^{\mathrm{T}}[t]\boldsymbol{y}_t + \epsilon[t] \quad , \tag{4.1}$$

where the coefficients of $\boldsymbol{w}[t]$ form a time dependent FIR filter of length N, and $(\cdot)^{\mathrm{T}}$ denotes the transpose operator.

We choose $\boldsymbol{w}[t]$ to make a prediction, $\hat{y}[t]$, of $y[t]$ by minimising the prediction error $\epsilon[t]$. The prediction problem can now be written thus:

$$\hat{y}[t] = \boldsymbol{w}^{\mathrm{T}}[t]\boldsymbol{y}_t \quad , \tag{4.2}$$

$$\epsilon[t] = y[t] - \hat{y}[t] \quad . \tag{4.3}$$

By making assumptions about the statistical properties of $\epsilon[t]$, one could proceed to formulate a number of linear predictors for $y[t]$, however, it is our goal to proceed without making such assumptions.

4.1.2 Prediction Based on Particle Swarm Optimisation

Review of Particle Swarm Optimisation

The particle swarm optimisation algorithm [179, 183], is a powerful and intuitive tool inspired by the social behaviour of swarms in nature. A group of candidate solutions, or particles, are flown through a given problem space with their velocities influenced by both their own performance, evaluated by some cost function, and that of the most successful member of the swarm.

Particle accelerations are randomly perturbed to produce the swarm-like behaviour observed in nature and to allow for the problem space to be appropriately explored. The swarm is accelerated towards the best known minima of the cost function while continuously searching for a better solution.

Algorithm: The i^{th} particle occupies the position $p_i[t]$ at time t in a problem space governed by cost function $C(p)$, has velocity $v_i[t]$, memory of its own previous best position, $p_{i,best}$, and knowledge of the previous best position of any particle $p_{g,best}$. A maximum particle velocity v_{max} is set to prevent divergence.

1. Initialise particles with random positions and velocities in the problem space for time step $t = 0$. Assign $p_{i,best} := p_i[0]$ for all particles and set $p_{g,best} := \arg\min_{p_{i,best}} \big(C(p_{i,best})\big)$.

Repeat:

2. For each particle, calculate $C(p_i[t])$:

if $C(p_i[t]) < C(p_{i,best})$ **then** $p_{i,best} := p_i[t]$,

if $C(p_i[t]) < C(p_{g,best})$ **then** $p_{g,best} := p_i[t]$.

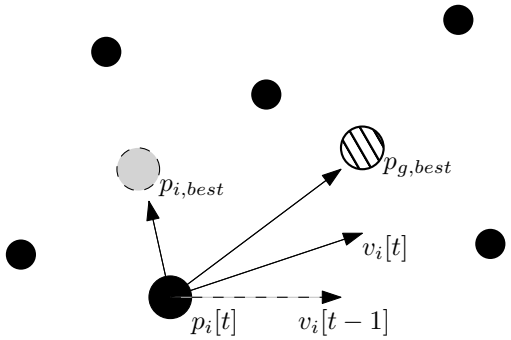


Figure 4.1: Illustration of the PSO algorithm. The velocity of particle i at time t is determined by its previous velocity $v_i[t-1]$, memory of its previous best position $p_{i,best}$, and the locations and the previous best position of any particle in the swarm $p_{g,best}$.

3. Update velocity, $v_i[t]$, and position of each particle:

$$\begin{aligned} v_i[t+1] &= c_0 v_i[t] + r_1 c_1 (p_{i,best} - p_i[t]) + r_2 c_2 (p_{g,best} - p_i[t]) \quad , \\ &\textbf{if} \;\; v_i[t+1] > v_{max} \;\; \textbf{then} \;\; v_i[t+1] := v_{max} \quad , \\ p_i[t+1] &= p_i[t] + v_i[t+1] \quad , \end{aligned}$$

where $r_1, r_2 \sim U(0,1)$ are random weights, c_0 is the inertial weight, c_1 is the cognition acceleration, and c_2 is the social acceleration.

4. Advance one time step and return to Step 2.

The velocity update for the i^{th} particle is illustrated in Figure 4.1.

PSO for FIR Prediction

The algorithm described in Section 4.1.2 is applied to the FIR predictor described by (4.2). Each particle in the swarm is a candidate for the FIR filter $\boldsymbol{w}[t]$ and at each time step the best performing particle is selected to make the next prediction. The problem space is therefore the N-dimensional complex space $\mathbb{C}^N$. Each particle, $p_i[t]$, is a candidate for $\boldsymbol{w}[t]$ and is therefore a complex vector of length N.

The cost function to be minimised is the absolute value of the prediction error, $|\epsilon[t]|$. When a new measurement is received, the potential past performance of all the

particles can be evaluated and the best performing particle selected to make the next prediction. Note that the progression of the algorithm would be exactly the same if the cost function were $\epsilon[t]^2$ since only the ranking of particles is of consequence, not the relative values of the cost function.

In addition to the basic algorithm, a maximum particle speed, $v_{\max}$, is enforced to restrict the step-size of particles in the problem space in order to control the resolution of the optimization and prevent it from diverging, akin to [182, 184]. If a particle's speed exceeds $v_{\max}$, it is reduced to $v_{\max}$.

Since the wind signal is non-stationary, the optimal solution is not static in the problem space and the PSO must be adjusted to allow for out-of-date solutions to be forgotten. Therefore, the particles are given a finite memory of the previous best locations $p_{i,best}$ and $p_{g,best}$.

Finally, due to the stochastic nature of the algorithm, the most consistent prediction is produced by generating an ensemble of FIR filters, each individual filter optimised by a separate particle swarm, and taking the mean prediction to be the *ensemble* prediction. Therefore, an ensemble of particle swarm optimised FIR (EPSO-FIR) filters is constructed.

The k^{th} member of the ensemble comprising K members optimises $\boldsymbol{w}_k[t]$ to produce the prediction $\hat{y}_k[t]$, as in (4.2). The ensemble prediction, $\tilde{y}[t]$,

$$\tilde{y}[t] = \frac{1}{K}\sum_{k=1}^{K}\hat{y}_k[t] \quad , \tag{4.4}$$

is the mean of the individual members' predictions.

Parameter Choice

The parameters of the PSO have been chosen heuristically, after extensive tests, to produce appropriate swarm behaviour and to minimise the root mean-squared error over the prediction period. Each parameter was perturbed in turn and the prediction error and behaviour of the swarm evaluated visually until a satisfactory parameter set was arrived at. Table 4.1 details the parameter values.

Parameter	Value
c_0	1.5
c_1	0.5
c_2	0.5
$v_{\max}$	0.05
No. of Particles	25
Memory	48
Ensemble Size	20

Table 4.1: List of parameter values used in PSO algorithm.

The coefficients of the velocity equation are chosen to produce swarm-like behaviour to enable the PSO algorithm to function as intended. This requires a balance between cognition and social acceleration to maintain a healthy particle distribution, and a sufficiently large inertial weight to ensure that the problem space is adequately explored. The maximum velocity is chosen to limit the distance each particle can travel in a single time step.

Each particle is given a memory of 48 time steps, i.e. 48 hours, since this is the time scale that the weather systems which govern the wind regime move across the UK, and is therefore an important component scale related to the wind signal's non-stationarity. An ensemble of 20 particle swarm optimised filters is found to produce consistent performance with little to be gained from using a larger ensemble.

4.1.3 Results

In this section the proposed method is applied to wind measurements in order to produce 1-hour-ahead forecasts. The performance of the ensemble of particle swarm optimised FIR filters (EPSO-FIR) is compared to the complex LMS algorithm (CLMS), [172,185], and a single channel cyclo-stationary Wiener filter (CSWF) described in [156] as examples of state-of-the-art linear predictors.

All quoted errors are root mean-squared error (RMSE),

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} \epsilon[t]\epsilon^*[t]} \quad , \tag{4.5}$$

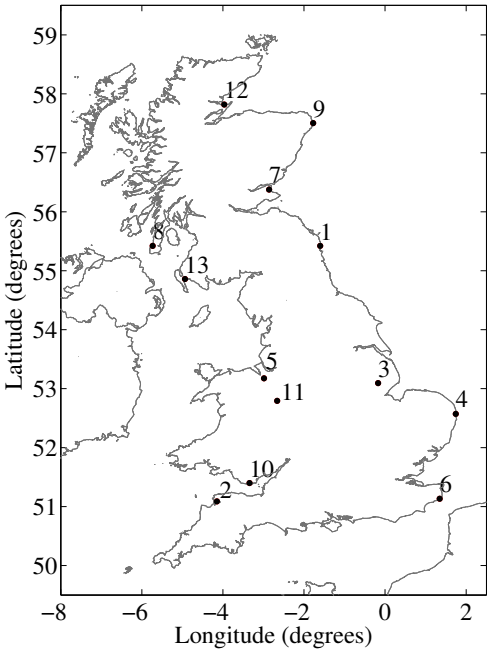


Figure 4.2: Locations of the 13 meteorological stations from which measurements have been used. Numbering corresponds to Table 4.3.

where the prediction error $\epsilon[t]$ is the difference between the predicted and measured wind velocity, i.e. not wind speed or direction independently.

Description of Data

The proposed approach is tested on wind data provided by the British Atmospheric Data Centre, which comprise hourly measurements made from 00:00h on 1/3/1997 to 23:00h on 28/2/1998 at 13 sites across the UK detailed in Figure 4.2. The measurements are taken in open terrain at a height of 10m, and comprises hourly averages that are quantised to a 10° angular granularity and integer multiples of one knot (0.515ms^{-1}) [2].

Prediction

Some example time series from individual and ensemble predictions are illustrated in Figures 4.3 and 4.4. The individual filters are able to track large and fast changes in the wind speed well but do not do so consistently. This tracking is often accompanied

Site	Individual PSO-FIR	Ensemble PSO-FIR
Boulmer	2.0568–2.8172	1.6347
Cheivenor	1.7105–2.3543	1.3361
Langdon Bay	1.9056–2.4388	1.6326
Peterhead Harbour	2.1990–2.9896	1.8113
Roose	1.8201–2.1254	1.5887

Table 4.2: Comparison of the RMSE (ms^{-1}) from individual predictors and the RMSE from the corresponding ensemble prediction

#	Site	CLMS	CSWF	EPSO-FIR
1	Boulmer	1.6252	**1.6238**	1.6347
2	Chivenor	1.7812	1.7790	**1.3361**
3	Coningsby	1.2939	**1.2932**	1.3231
4	Gorleston	1.6071	**1.6090**	1.6462
5	Hawarden Airport	1.5984	**1.5948**	1.6401
6	Langdon Bay	1.7399	1.7423	**1.6326**
7	Leuchars	1.5783	**1.5717**	1.6026
8	Machrihanish	2.0591	**2.0532**	2.0945
9	Peterhead Harbour	**1.7801**	n/a*	1.8113
10	Rhoose	1.7596	1.7578	**1.5887**
11	Shawbury	1.5326	**1.5314**	1.5701
12	Tain Range	2.0262	**2.0224**	2.1034
13	West Freugh	**1.8260**	1.8289	1.8626

Table 4.3: Comparison of 1 hour ahead RMSE (ms^{-1}) for the complex LMS (CLMS) algorithm, cyclo-stationary Wiener filter (CSWF) and the ensemble of particle swarm optimised FIR filters (EPSO-FIR). The RMSE for the best performing method is highlighted in bold.
* Implementation of the CSWF was not possible for Peterhead Harbour due to insufficient training data

by a significant over-shoot as the filter fails to anticipate the sudden change in gradient.

The inconsistent behaviour of the individual filters is lost when an ensemble of predictions is averaged, resulting in an overall reduction in error but a systematic lag in response to large changes in wind speed.

The benefit of taking the mean prediction from an ensemble of PSO optimised predictors is significant. The RMSE, measured over the entire year of predictions, for the ensemble prediction is substantially lower than that for the individual predictors. Some examples are given in Table 4.2.

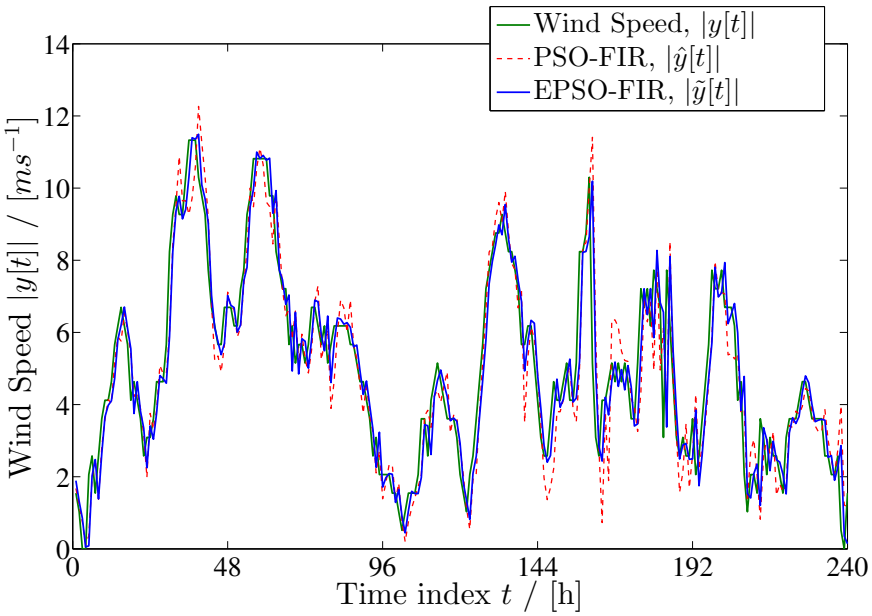
The results from the EPSO-FIR prediction and the two linear methods are listed in Table 4.3. The EPSO-FIR is out performed by the other two methods at 10 of the 13 locations by approximately 4%, however, it performs substantially better than both the CLMS and CSWF at three sites with a 15% reduction in RMSE, notably the three most southerly sites in the data set, see Figure 4.2.

The results provide evidence that PSO can afford a significant performance advantage for at least some sites in the current setting of the method. Whether there are any anomalies in those three sites that favour PSO over our previous techniques is difficult to established based on only three sites, and will be the subject of future investigation.

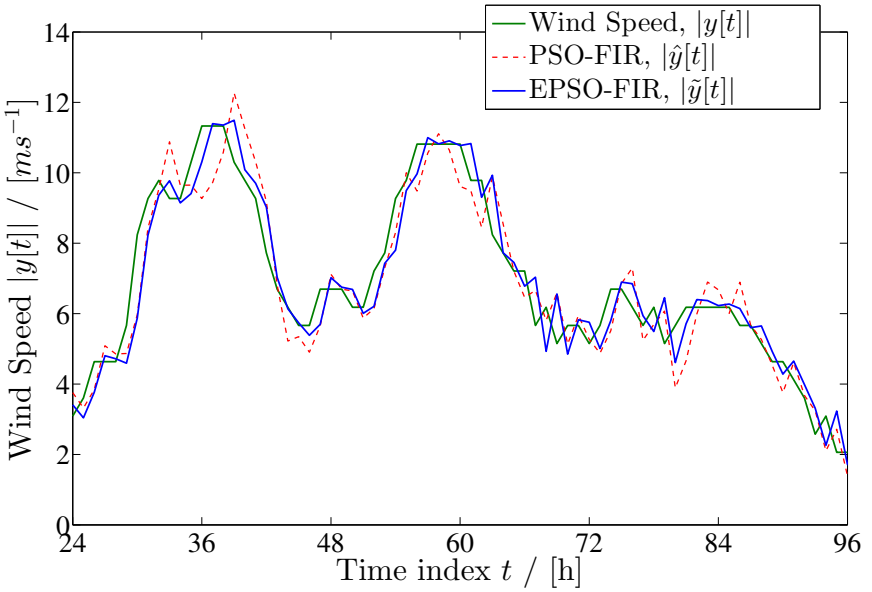
The behaviour of the particle swarm is good: the distribution of particles is such that a sensible region of the problem space is explored. The algorithm converges quickly and tracks the non-stationary wind signal well. The density evolution of the real part of the first element of the PSO particles is shown in Figure 4.5. Also of note is that the EPSO-FIR requires very little training data, approximately $2N$ samples to populate the filter and converge, compared to the CLMS which, depending on the choice of learning rate and training strategy, requires several months of data, and the CSWF which needs several years worth of training data in order to capture the seasonal trends in the wind data.

4.1.4 Summary and Future Work

The proposed ensemble particle swarm optimised FIR predictor offers similar performance to linear techniques of higher complexity, which require substantially more train-

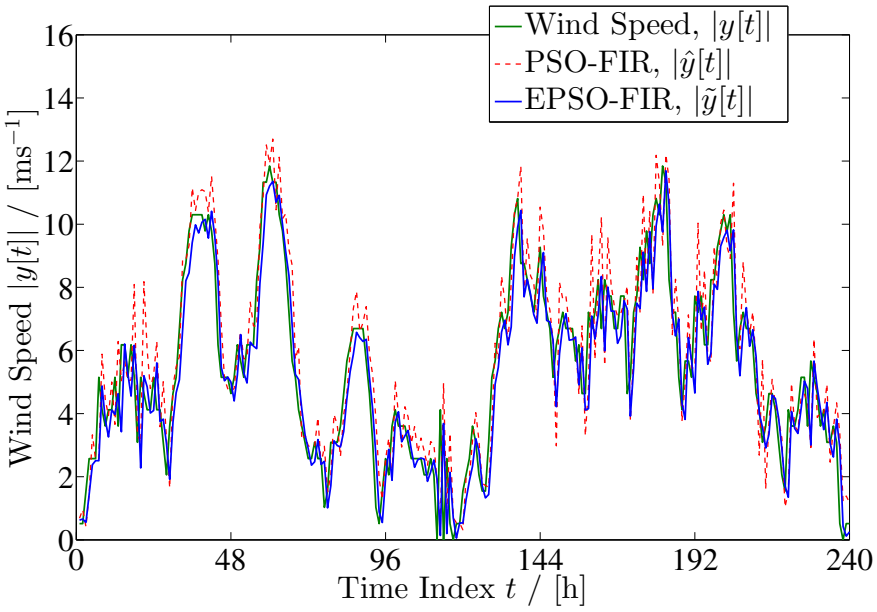


(a) Chivenor, 31/05/97–09/06/97

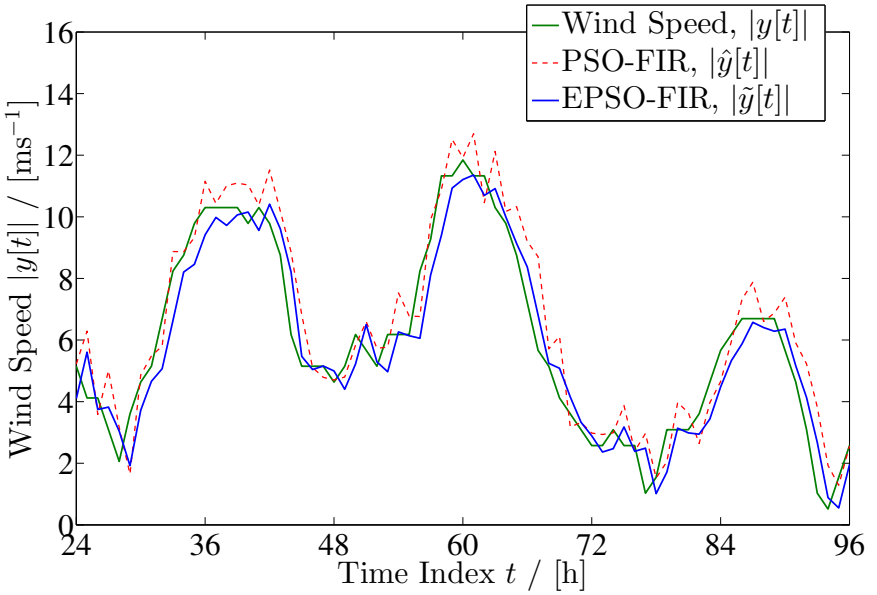


(b) Chivenor, 01/06/97–03/06/97

Figure 4.3: Wind speed, individual PSO-FIR prediction and ensemble prediction (EPSO-FIR) for Chivenor (a) shows 31/05/97–09/06/97, while (b) shows 01/06/97–03/06/97.



(a) Rhoose, 31/05/97–09/06/97



(b) Rhoose, 01/06/97–03/06/97

Figure 4.4: Wind speed, individual PSO-FIR prediction and ensemble prediction (EPSO-FIR) for Rhoose. (a) shows 31/05/97–09/06/97, while (b) shows 01/06/97–03/06/97.

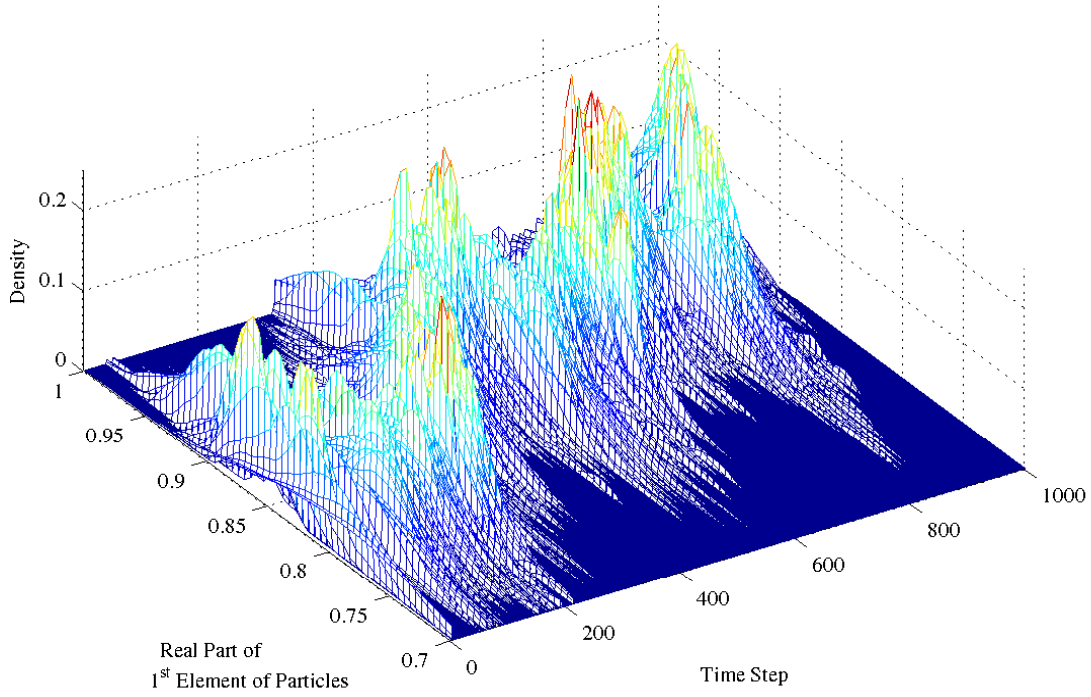


Figure 4.5: An example of the density evolution of the real part of the first element of the particles in a swarm for the first 1000 time steps of prediction.

ing data, and still has great potential for further development. The PSO algorithm is found to be efficient and converges quickly, tracking the non-stationary wind signal well.

The potential for tracking large changes in wind speed is of great interest since this is a weakness of the simple linear and many substantially more complex techniques which are currently employed for short-term wind prediction.

The performance of this early-development approach is encouraging, and the method warrants further investigation. Both the complex LMS and cyclo-stationary Wiener filter saw significant improvement when expanded to process information from multiple sites simultaneously, taking advantage of the spatial correlation between different locations, as in Chapter 3 and [157]. An attempt was made to incorporate spatial information by simply concatenating the input vector with measurements from other locations akin to the linear approaches. However, this negatively impacted the performance of both individual ensemble members and the EPSO-FIR overall.

Other techniques that have combined PSO with multi-scale analysis such as wavelet decomposition, [142, 181]. These show some promise and could be combined with with other appropriate PSO variations [182, 183, 186].

4.2 Kernel Methods

To date, the majority of statistical methods used for wind speed prediction have been linear despite the well established non-linear nature of the wind. Here a relatively new and exciting class of learning algorithms called *kernel methods* is explored. Kernel methods enable the linear processing of non-linear 'features' in some high-dimensional feature space. This approach retains many desirable properties of linear processing (fast learning algorithms, unique optimal solution) while making it possible to capture some non-linearities.

Over the last decade, many kernel methods have been developed and now represent a distinct class of learning algorithms. Such methods are based on the so-called 'kernel trick', a result which allows the inner product of a non-linear function defined by a Mercer kernel (Mercer's Theorem [187]) to be calculated while the function itself remains unknown [188]. This has advantageous properties in function estimation and classification; support vector machines, for example, rely on kernel methods.

A direct application of non-linear function estimation is regression, where some non-linear mapping is followed by linear processing in a high (or infinite) dimensional feature space. The kernel trick removes the need to identify the mapping associated with the a given Mercer kernel which may not be available or be difficult to calculate; the only challenge is selecting an appropriate kernel for the problem at hand.

Several linear methods have been 'kernelised' including the popular least means squares (LMS) [189] and recursive least squares (RLS) [190] algorithms, plus extensions, [191–194] for example. Reported applications to forecasting include high frequency wind prediction [194, 195] and load forecasting [196], among others.

In this section two kernel methods are employed to produce short-term wind forecasts: a simple kernelised LMS (KLMS) algorithm and the kernel RLS (KRLS) of [190] are studied. The theory of kernel methods is briefly introduced in Section 4.2.1 and

the prediction problem is stated in Section 4.2.2 followed by descriptions of the KLMS and KRLS algorithms. A case study is then presented in Section 4.2.3, including how the algorithms and benchmarks were implemented, and their performance is evaluated. Finally, conclusions are drawn in Section 4.2.4.

4.2.1 Kernel Methods

Kernel methods are a class of learning algorithm which use Mercer kernels in order to produce non-linear versions of conventional linear learning algorithms. The kernel trick allows the inner product of two input vectors in some high-dimensional Hilbert space $\mathcal{H}$ (often called the feature space) to be calculated without explicit knowledge of the feature vectors (the non-linear projection of the input vectors in $\mathcal{H}$).

First the Mercer kernel is introduced: a continuous, symmetric, positive-definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R},\ \mathcal{X} \in \mathbb{R}^n$ (or $\mathbb{C}^n$). Mercer's theorem states that any Mercer kernel $k(\cdot,\cdot)$ can be expressed as the inner product of some fixed non-linear function $\{\phi(\boldsymbol{x}) : \mathcal{X} \to \mathcal{H}_1,\ \boldsymbol{x} \in \mathcal{X}\}$,

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle_{\mathcal{H}_1} \quad , \tag{4.6}$$

where $\mathcal{H}_1$ is a real- or complex-valued reproducing kernel Hilbert space, for which $k(\cdot,\cdot)$ is a reproducing kernel and $\langle \cdot,\cdot \rangle_{\mathcal{H}_1}$ is the corresponding inner product in $\mathcal{H}_1$.

Equation (4.6) states that if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are mapped onto $\mathcal{H}_1$ by $\phi(\boldsymbol{x}_i)$ and $\phi(\boldsymbol{x}_j)$, respectively, then the inner product of these functions can be calculated by evaluating the kernel $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, even if the mapping $\phi(\cdot)$ is unknown. This result is known as the Kernel trick.

The Gaussian kernel is frequently used in real world applications with particular success in time series prediction problems [197]. It is the expansion function for an infinite dimensional feature space and is given by

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-||\boldsymbol{x}_i - \boldsymbol{x}_j||^2\right) \tag{4.7}$$

and used throughout this study [188]. While other kernels could be chosen, the Gaussian

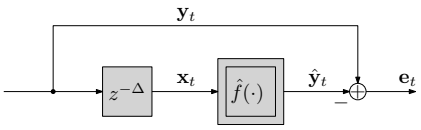


Figure 4.6: Non-linear system $f(\cdot)$ set-up to predict the vector quantity $\boldsymbol{y}_t$ some Δ samples ahead based on input $\boldsymbol{x}_t$.

kernel has a physical interpretation as a measure of similarity, which is fitting here, and has out-performed other candidate kernels (triangular and polynomial) in similar work [190, 194]. The choice, or construction, of kernels is very much an open problem and the subject of ongoing research.

4.2.2 Prediction Algorithms

Prediction Set-up

The prediction problem is outlined in Fig. 4.6, whereby the purpose of a potentially non-linear function $f(\cdot)$ is to look Δ samples ahead in time, estimating $\boldsymbol{y}_t \in \mathbb{R}^n$ (or $\mathbb{C}^n$) from the input vector $\boldsymbol{x}_t \in \mathbb{R}^m$ (or $\mathbb{C}^m$) containing space-time data comprising measurements $\boldsymbol{y}_{t-\Delta},\ \boldsymbol{y}_{t-\Delta-1}, ...$ The predictor is written

$$\hat{\boldsymbol{y}}_t = f(\boldsymbol{x}_t) \quad . \tag{4.8}$$

The aim in the context of a prediction problem is to find an estimate $\hat{f}(\cdot)$ of $f(\cdot)$ which minimises the estimation error in the mean squared sense, i.e.

$$J = \sum_t |\boldsymbol{e}_t|^2 = \sum_t |\boldsymbol{y}_t - \hat{f}(\boldsymbol{x}_t)|^2 \tag{4.9}$$

The linear approximation of this problem is given by $\hat{f}(\boldsymbol{x}_t) = \mathbf{A}\boldsymbol{x}_t$ where $\mathbf{A} \in \mathbb{R}^{n\times m}$ (or $\mathbb{C}^{n\times m}$) is a coefficient matrix whose entries are to be determined. Many estimation schemes based on this approximation have been studied.

Alternatively, the approximation can be stated in terms of the mapping $\phi(\cdot)$ to place us in a non-linear setting, writing $\hat{f}(\boldsymbol{x}_t) = \mathbf{A}\phi(\boldsymbol{x}_t)$ with $\mathbf{A} \in \mathbb{R}^{n\times l}$ (or $\mathbb{C}^{n\times l}$). The properties of Mercer kernels make it possible to derive estimation schemes for $f(\cdot)$ in a

high l-dimensional feature space without performing calculations in such a space. This combines simple implementation of linear methods with the advantageous properties of working with a non-linear mapping.

In the remainder of this section two popular linear algorithms, the least mean squares (LMS) and recursive least squares (RLS), are presented in their conventional linear and kernelised forms and discussed. Only wind speed, which is real valued, is considered therefore in the proceeding sections all quantities are real valued and the conjugations necessary in the complex case are not included.

Kernel LMS

The LMS algorithm comprises an update scheme based on gradient decent for the coefficient matrix $\mathbf{A}$ given by

$$\begin{aligned} \mathbf{A}_0 &= \mathbf{0}_{n\times m} && (4.10)\\ \boldsymbol{e}_t &= \boldsymbol{y}_t - \mathbf{A}_t\boldsymbol{x}_t && (4.11)\\ \mathbf{A}_{t+1} &= \mathbf{A}_t + \mu\boldsymbol{e}_t\boldsymbol{x}_t^{\mathrm{T}} \quad , && (4.12) \end{aligned}$$

where the prediction $\hat{\boldsymbol{y}}_t = \mathbf{A}_t\boldsymbol{x}_t$ and μ is the positive learning rate which controls the trade-off between confidence in individual samples and convergence speed.

When kernelised, the update step (4.12) becomes

$$\mathbf{A}_{t+1} = \mathbf{A}_t + \mu\boldsymbol{e}_t\phi^{\mathrm{T}}(\boldsymbol{x}_t) \tag{4.13}$$

however to express the algorithm in terms of inner products it is more convenient to write

$$\mathbf{A}_{t+1} = \mu\sum_{i=1}^{t}\boldsymbol{e}_i\phi^{\mathrm{T}}(\boldsymbol{x}_i) \tag{4.14}$$

which allows the prediction to be written

$$\hat{\boldsymbol{y}}_t = \mu\sum_{i=1}^{t}\boldsymbol{e}_i\phi^{\mathrm{T}}(\boldsymbol{x}_i)\phi(\boldsymbol{x}_t) \quad . \tag{4.15}$$

Finally, if $\phi(\cdot)$ is chosen to be an expansion function of some reproducing kernel Hilbert space the inner product in (4.15) can be computed using the corresponding generating kernel $k(\cdot,\cdot)$, i.e.

$$\hat{\boldsymbol{y}}_t = \mu \sum_{i=1}^{t} \boldsymbol{e}_i k(\boldsymbol{x}_i, \boldsymbol{x}_t) \quad . \tag{4.16}$$

Notice however that as t increases so does the number of terms in the sum required to produce an estimate. This quickly becomes impractical and must be avoided. We therefore impose a sparsity constraint, by retaining a finite *dictionary*, D, of input vectors. At each time step t the input vector is compared to the dictionary D_{t-1}; if the minimum distance between $\boldsymbol{x}_t$ and D_{t-1} exceeds some sparsity parameter ν it is added to the dictionary, i.e. $D_t := D_{t-1} \cup \{\boldsymbol{x}_t\}$, else $D_t := D_{t-1}$. The estimation is now

$$\hat{\boldsymbol{y}}_t = \mu \sum_{i \in D_{t-1}} \boldsymbol{e}_i k(\boldsymbol{x}_i, \boldsymbol{x}_t) \quad . \tag{4.17}$$

Kernel RLS

The RLS algorithm attempts to minimise the cost function (4.9) at each time step, rather than the mean squared error as in the LMS algorithm. The cost function is rewritten as

$$J(\boldsymbol{w}) = \sum_{i=1}^{t} (\boldsymbol{y}_i - \mathbf{A}\phi(\boldsymbol{x}_i))^2 = |\mathbf{Y}_t - \boldsymbol{\Phi}_t^{\mathrm{T}} \boldsymbol{w}|^2 \tag{4.18}$$

where $\mathbf{Y}_t$ and $\boldsymbol{\Phi}_t$ are output and projected input data matrices, and $\boldsymbol{w}$ is a weight vector. As before, working in some high dimensional feature space is undesirable, so writing the optimal weight vector as $\boldsymbol{w}_t = \sum_{i=1}^{t} \alpha_i \phi(\boldsymbol{x}_i) = \boldsymbol{\Phi}_t \boldsymbol{\alpha}$ the kernel trick allows the cost function to be expressed as

$$J(\boldsymbol{\alpha}) = |\mathbf{Y}_t - \mathbf{K}_t \boldsymbol{\alpha}|^2 \quad , \tag{4.19}$$

where $[\mathbf{K}_t]_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j);\ i,j = 1, ..., t$, is called the kernel matrix.

In theory the minimiser, $\boldsymbol{\alpha} = \mathbf{K}_t^{-1} \mathbf{Y}_t$ could be computed recursively using the conventional RLS algorithm, however, as with the kernelised LMS algorithm, the complexity of the calculation would increase with each new sample, in addition to possible

over-fitting when the number of samples becomes large. We therefore sparsify the algorithm by retaining a finite dictionary of samples and replacing $\mathbf{K}_t$ with the dictionary kernel matrix $[\tilde{\mathbf{K}}_t]_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j);\ \boldsymbol{x}_i, \boldsymbol{x}_j \in D$, and $\boldsymbol{\alpha}_t$ with the reduced weights $\tilde{\boldsymbol{\alpha}}_t$. The dictionary is updated in the same manor as described for the KLMS algorithm.

The KRLS algorithm closely resembles the classical RLS algorithm with the exception that if the dictionary changes size, so must the reduced weight vector $\tilde{\boldsymbol{\alpha}}_t$. and precision matrix $\mathbf{P}_t$. The full derivation of the KRLS algorithm can be found in the original paper [190]; pseudo code is presented here without proof.

Initialisation: $\tilde{\mathbf{K}}_1 = [k(\boldsymbol{x}_1, \boldsymbol{x}_1)]$, $\tilde{\mathbf{K}}_1^{-1} = [1/\tilde{\mathbf{K}}_1]$, $\tilde{\boldsymbol{\alpha}}_1 = [\boldsymbol{y}_1/\tilde{\mathbf{K}}_1]$, $\mathbf{P}_1 = [1]$, sparsity parameter: ν.

for t=2,3,...

Compute: $\tilde{\boldsymbol{k}}_{t-1}(\boldsymbol{x}_t)$ where $[\tilde{\boldsymbol{k}}_{t-1}(\boldsymbol{x}_t)]_i = k(\boldsymbol{x}_i, \boldsymbol{x}_t), i \in D$

Make Prediction: $\hat{\boldsymbol{y}}_t = \tilde{\boldsymbol{k}}_{t-1}^{\mathrm{T}}(\boldsymbol{x}_t)\tilde{\boldsymbol{\alpha}}_t$

Compute: $\boldsymbol{a}_t = \tilde{\mathbf{K}}_{t-1}^{-1}\tilde{\boldsymbol{k}}_{t-1}(\boldsymbol{x}_t)$

Compute: $\delta_t = k(\boldsymbol{x}_t, \boldsymbol{x}_t) + \tilde{\boldsymbol{k}}_{t-1}^{\mathrm{T}}(\boldsymbol{x}_t)\boldsymbol{a}_t$

Case 1: The new sample is approximately linearly independent with respect to the current dictionary, satisfying $\delta_t > \nu$, therefore the new sample, $\boldsymbol{x}_t$, is added to the dictionary. The matrices $\tilde{\mathbf{K}}_t^{-1}$, $\mathbf{P}_t$ and $\tilde{\boldsymbol{\alpha}}_t$ are updated as follows:

$$\tilde{\mathbf{K}}_t^{-1} = \frac{1}{\delta_t}\begin{bmatrix} \tilde{\mathbf{K}}_{t-1}^{-1} + \boldsymbol{a}_t\boldsymbol{a}_t^{\mathrm{T}} & -\boldsymbol{a}_t \\ -\boldsymbol{a}_t^{\mathrm{T}} & 1 \end{bmatrix} \tag{4.20}$$

$$\mathbf{P}_t = \begin{bmatrix} \mathbf{P}_{t-1} & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & 1 \end{bmatrix} \tag{4.21}$$

$$\tilde{\boldsymbol{\alpha}}_t = \begin{bmatrix} \tilde{\boldsymbol{\alpha}}_{t-1} - \frac{\boldsymbol{a}_t}{\delta_t}(\boldsymbol{y}_t - \tilde{\boldsymbol{k}}_{t-1}^{\mathrm{T}}(\boldsymbol{x}_t)\boldsymbol{a}_t) \\ \frac{1}{\delta_t}(\boldsymbol{y}_t - \tilde{\boldsymbol{k}}_{t-1}^{\mathrm{T}}(\boldsymbol{x}_t)\boldsymbol{a}_t) \end{bmatrix} \tag{4.22}$$

Case 2: The new sample is approximately linearly dependent with respect to the current dictionary, satisfying $\delta_t \leq \nu$, so the new sample is not added to the dictionary and $\tilde{\mathbf{K}}_t^{-1} = \tilde{\mathbf{K}}_{t-1}^{-1}$. The matrices $\mathbf{P}_t$ and $\tilde{\boldsymbol{\alpha}}_t$ are updated as follows:

$$\mathbf{P}_t = \mathbf{P}_{t-1} - \frac{\mathbf{P}_{t-1}\boldsymbol{a}_t\boldsymbol{a}_t^{\mathrm{T}}\mathbf{P}_{t-1}}{1 + \boldsymbol{a}_t^{\mathrm{T}}\mathbf{P}_{t-1}\boldsymbol{a}_t} \tag{4.23}$$

$$\boldsymbol{q}_t \stackrel{\text{def}}{=} \frac{\mathbf{P}_{t-1}\boldsymbol{a}_t}{1 + \boldsymbol{a}_t^{\mathrm{T}}\mathbf{P}_{t-1}\boldsymbol{a}_t} \tag{4.24}$$

$$\tilde{\boldsymbol{\alpha}}_t = \tilde{\boldsymbol{\alpha}}_{t-1} + \tilde{\mathbf{K}}_t^{-1}\boldsymbol{q}_t\left(\boldsymbol{y}_t - \tilde{\boldsymbol{k}}_{t-1}^{\mathrm{T}}(\boldsymbol{x}_t)\boldsymbol{a}_t\right) \tag{4.25}$$

end for

Sparsity

Both the KLMS and KRLS algorithms retain a dictionary of input samples as a sparse representation of the complete history of input samples up to some time. It is an important property of the dictionary that it is finite, and it can be shown to be so with only mild conditions on the data and kernel function. If $\boldsymbol{x} \in \mathcal{X}$ and $\phi(\boldsymbol{x}) \in \mathcal{H}$ then if $\mathcal{X}$ is compact, and the sparsity parameter is positive ($\nu > 0$), the dictionary will be finite. For a rigorous proof, see [190].

It should be noted that for these algorithms to be fully adaptive they should incorporate some forgetting mechanism whereby out-of-date dictionary elements are 'forgotten' in order to track changing dynamics of the process being modelled. A sophisticated multi-kernel LMS algorithm is developed in [194] which includes this feature, as well as the ability to combine multiple kernel functions. The drawback, of course, is the need to determine the parameters for each of these mechanisms.

4.2.3 Case Study

The kernelised LMS and RLS algorithms are implemented to produce 1 to 6 hour ahead predictions of wind speed at six locations in the same region. Their performance

is compared to several conventional benchmarks, including the conventional LMS and RLS algorithms.

Test Data

The data used for testing is from the Hydra dataset of hourly mean potential wind at multiple locations across the Netherlands. Six locations within 150km of each other are considered here with measurements from 2001 used as a training set and data from 2002 used for testing. The measured wind speed has been corrected for the effects of shelter from buildings or vegetation. The resulting *potential* wind is an estimate of the wind speed that could have been measured at 10m height if the station's surroundings were free of obstacles and flat with a roughness length equal to that of grass onshore (0.03m) and water offshore (0.002m). For more information on this process see [3]. In addition, the data have been normalised so that they occupies the range $[0, 1]$ by division their maximum value.

This transformation aids spatial prediction by removing biases present at individual measurement locations that would otherwise interfere with the spatio-temporal correlation of the data. The procedure is simple to implement once information regarding the terrain surrounding a weather station is known.

Implementation

The KLMS and KRLS are employed to predict the wind speed at the six locations which are embedded in the vector $\boldsymbol{y}_t \in \mathbb{R}^6$. The prediction of $\boldsymbol{y}_t$ made with measurements available at time $t - \Delta$ (or Δ steps ahead) is denoted $\hat{\boldsymbol{y}}_{t|t-\Delta}$. The input vector $\boldsymbol{x}_{t|t-\Delta}$ for a $\Delta = 1$ step-ahead prediction is the concatenation of p lagged values of $\boldsymbol{y}_t$, i.e. $\boldsymbol{x}_{t|t-1} = (\boldsymbol{y}_{t-1}^{\mathrm{T}}, ..., \boldsymbol{y}_{t-p}^{\mathrm{T}})^{\mathrm{T}}$, and for horizons of $\Delta > 1$ where not all lags are available,

predictions are used

$$\boldsymbol{x}_{t|t-\Delta} = \begin{bmatrix} \hat{\boldsymbol{y}}_{t-1|t-\Delta} \\ \vdots \\ \hat{\boldsymbol{y}}_{t-\Delta+1|t-\Delta} \\ \boldsymbol{y}_{t-\Delta} \\ \vdots \\ \boldsymbol{y}_{t-p} \end{bmatrix} \quad . \tag{4.26}$$

This is the so-called *direct forecasting* approach where forecast are used as input data to produce multi-step-ahead forecasts; the alternative would be to build separate predictors for each forecast horizon and run them in parallel. In this study only forecast horizons up to 6 hours are of concern, or $\Delta = 1, ..., 6$ and the forecasts for each horizon are made in parallel by distinct predictors. The number of temporal lags is a trade-off between accuracy and computational expense, and is chosen as $p = 6$ for the KLMS and $p = 3$ for the KRLS since the improvement in accuracy was negligible for greater values.

The sparsification parameter and LMS learning rate are determined heuristically by exhaustive search to minimise the residual error on the training data. The sparsification parameter is $\nu = 0.02$ for the KRLS and $\nu = 0.1$ for the KLMS. The KLMS learning rate was chosen to be $\mu = 0.01$.

Benchmarks

An important benchmark is the persistence forecast which supposes that the future wind speed will be the same as the most recent measurement. While its implementation is trivial its performance is still considered acceptable by many practitioners, particularly in situations where more complex approaches offer only modest gains. The persistence forecast Δ-hours ahead is given by

$$\hat{\boldsymbol{y}}_{t|t-\Delta} = \boldsymbol{y}_{t-\Delta} \quad . \tag{4.27}$$

In order to compare the kernelised algorithms to conventional techniques and highlight the value of spatial information two non-recursive linear time series models are used as further benchmarks in addition to the conventional LMS and RLS algorithms. The first is the non-spatial autoregressive (AR) model which is given by

$$\hat{y}_{t|t-\Delta} = \sum_{i=1}^{\Delta-1} a_i \hat{y}_{t-i|t-\Delta} + \sum_{i=\Delta}^{p} a_i y_{t-i} \tag{4.28}$$

for each location. The number of lags p is determined by the Akaike information criterion, and the parameters a_i are determined by maximum likelihood estimation assuming independent identically distributed (i.i.d.) Gaussian prediction errors [198].

The second is the vector generalisation of AR, the vector autoregressive model (VAR). As in the multivariate kernelised algorithms, the measurements at multiple locations are embedded in the vector $\boldsymbol{y}_t$ and the model is written

$$\hat{\boldsymbol{y}}_{t|t-\Delta} = \mathbf{A}\boldsymbol{x}_{t|t-\Delta} \quad , \tag{4.29}$$

where $\boldsymbol{x}_t$ is given by Equation (4.26). Once again the number of lags p is determined by the Akaike information criterion and assuming i.i.d. Gaussian errors the coefficient matrix $\mathbf{A} \in \mathbb{R}^{n \times np}$ is determined by maximum likelihood estimation [198].

The conventional LMS algorithm, with update scheme given in equations (4.10)–(4.12) and learning rate $\mu = 0.0005$, and conventional RLS with update scheme

$$\begin{aligned}
\boldsymbol{e}_t &= \boldsymbol{y}_t - \mathbf{A}_t \boldsymbol{x}_t & (4.30)\\
\boldsymbol{k}_t &= \frac{\boldsymbol{x}_t^{\mathrm{T}} \mathbf{Q}_t}{1/\lambda + \boldsymbol{x}_t^{\mathrm{T}} \mathbf{Q}_t \boldsymbol{x}_t} & (4.31)\\
\mathbf{Q}_{t+1} &= \mathbf{Q}_t - \mathbf{Q}_t \boldsymbol{x}_t \boldsymbol{k}_t & (4.32)\\
\mathbf{A}_{t+1} &= \mathbf{A}_t + \boldsymbol{e}_t \boldsymbol{k}_t \quad , & (4.33)
\end{aligned}$$

and forgetting factor $\lambda = 0.9995$ are also included for comparison with their kernelised versions. The look-ahead indexing has been dropped here to avoid notational clutter but the principle still applies.

The AR and VAR methods are non-recursive, that is to say that their parameters

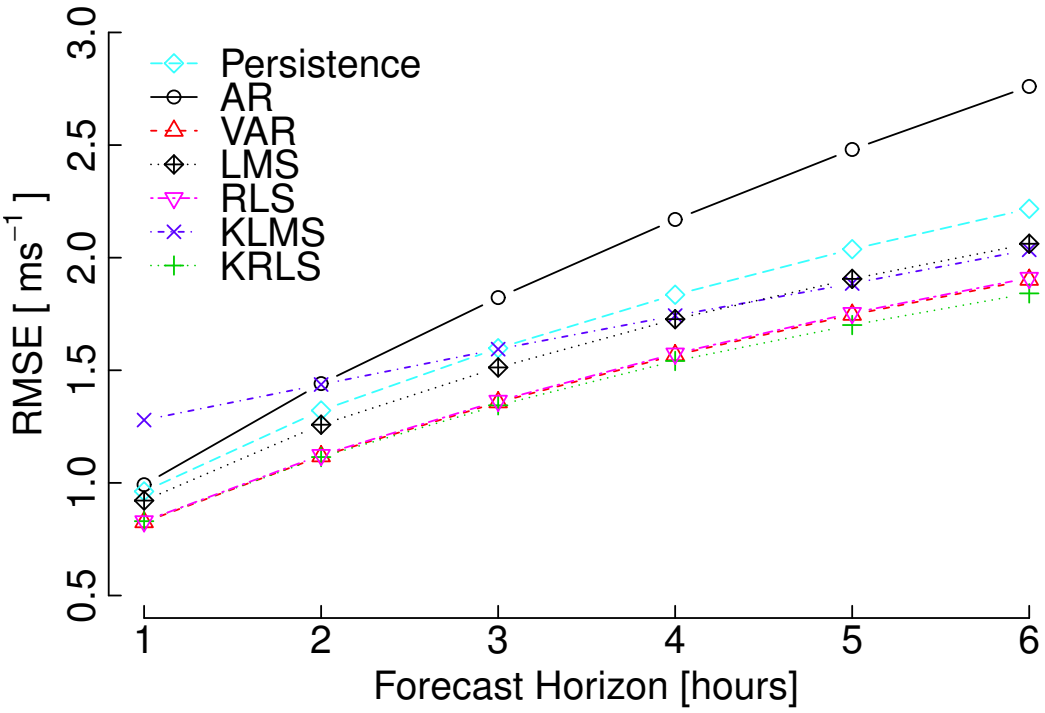


Figure 4.7: Mean root mean squared error across all 6 sites for Kernelised algorithms and benchmarks.

are estimated directly from the training data and are then fixed throughout the test period. The other methods, with the exception of persistence, are recursive and as such are initialised and then run sequentially through the training and test data, updating their parameter estimates at each step.

Results

Performance is evaluated in terms of root mean squared error, which is given by the expression

$$\text{RMSE}_\Delta = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{y}_{t|t-\Delta} - y_t)^2} \tag{4.34}$$

for the samples $y_1, ..., y_T$ in the test dataset at each location and for each forecast horizon $\Delta = 1, ..., 6$.

The performance of the kernelised algorithms and benchmarks is illustrated in Figure 4.7 in terms of mean RMSE accross the six sites in the dataset. An example fore-

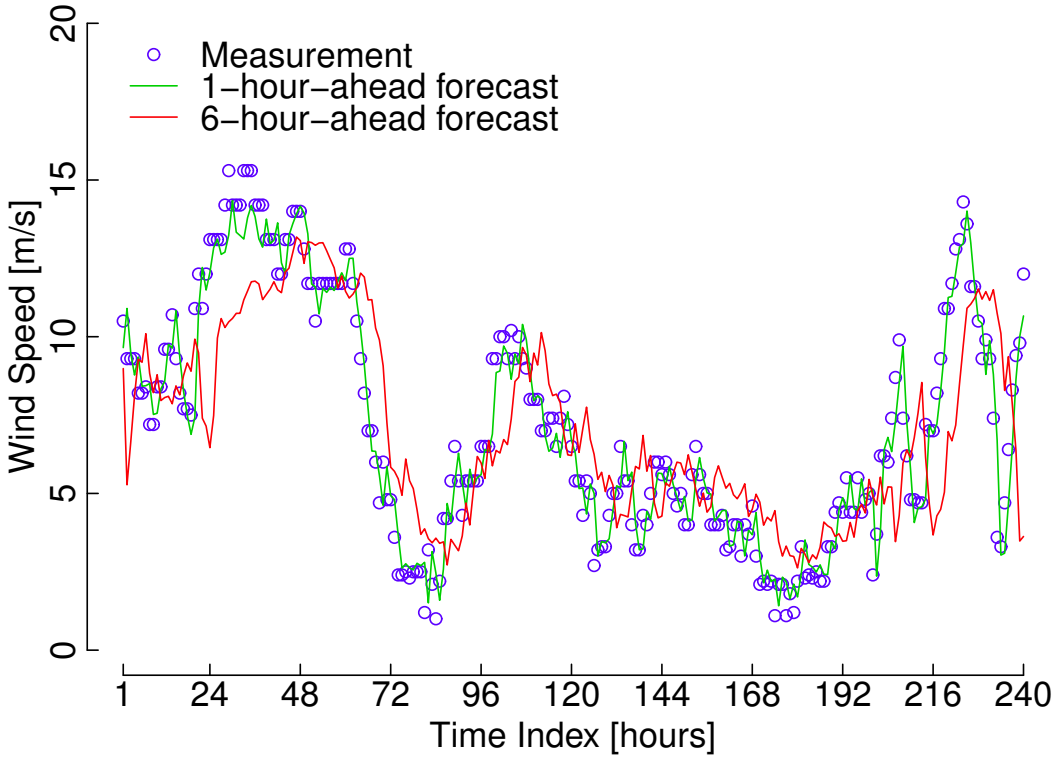


Figure 4.8: Forecast time series 1- and 6-hours-ahead produced by KRLS algorithm beginning on 10^{th} February 2002.

cast time series is illustrated in Figure 4.8. The most simplistic approaches, persistence and AR, perform significantly worse at all forecast horizons than the more sophisticated VAR, RLS and KRLS. Both the LMS and its kernelised version (KLMS) have intermediate performance, reflective of their complexity, though the KLMS performs particularly poorly for the 1 and 2 hour ahead predictions.

The improvement over persistence is shown in Figure 4.9 for the VAR, RLS and KRLS predictions. All three exhibit similar performance for 1 and 2 hour ahead forecasts, but the KRLS outperforms the two linear methods for the longer horizons. It is also notable that the KLMS improves relative to the LMS at longer forecast horizons. In both cases the kernelised versions of linear algorithms offer improved 5 and 6 hour ahead predictions. The performance of the AR model is particularly poor, especially when compared to persistence, possible due to over-fitting.

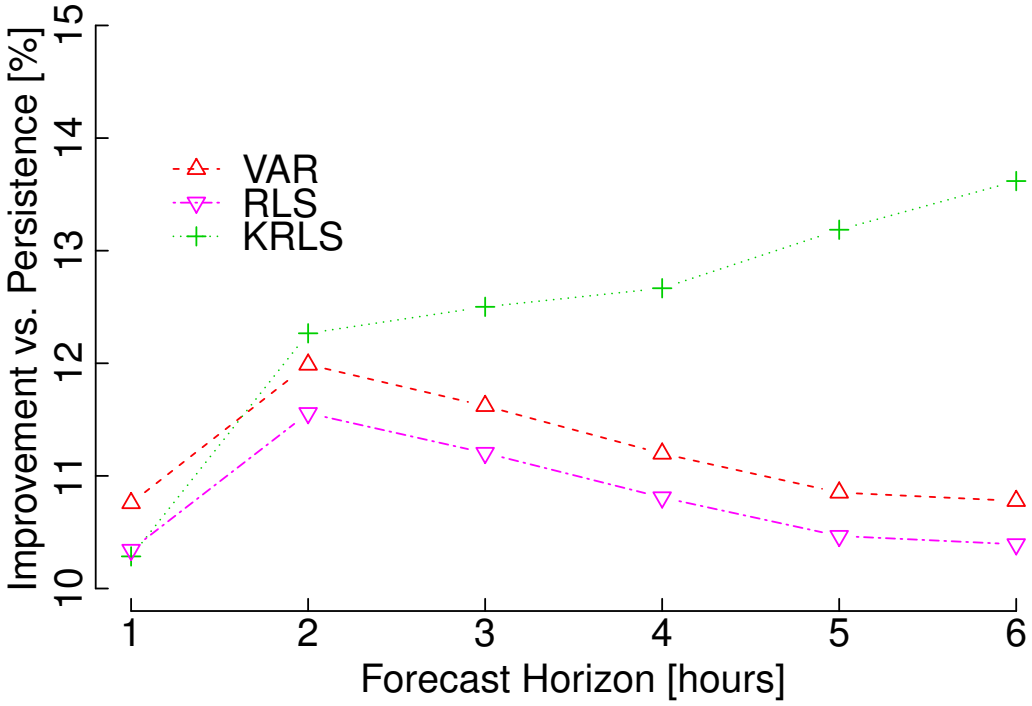


Figure 4.9: Percentage improvement vs. persistence for the VAR and KRLS forecasts.

4.2.4 Summary and Future Work

For the short-term spatio-temporal prediction of wind speed, two examples from a new class of learning algorithms called kernel methods have been investigated. The kernel least mean squares and kernel recursive least squares algorithms are non-linear extensions of their conventional linear forms and have been applied to a dataset comprising wind speed measurements made at six locations in the Netherlands over a period of 2 years. The KRLS in particular shows significant improvement over several established linear benchmarks, especially for longer forecast horizons.

The case study was restricted to a modest number of spatial locations since the kernelised algorithms struggled with large numbers ($m \gg 6$) of sites despite the spatial model with $m = 6$ performing much better than non-spatial, i.e. $m = 1$. The same is characteristic of linear models but is observed at a larger spatial dimension. Others have developed variations on kernelised algorithms, such as the multi-kernel approach of Tobar *et al.* in [194]. Alternative kernel structures that discriminate between temporal

and spatial input data may address this issue and yield improved performance. Alternatively, the proposed approach could be implemented on subsets of a larger spatial dataset.

4.3 Conclusions

The techniques presented in this chapter have demonstrated the advantages of non-linear methods for wind prediction. Two learning algorithms have been employed to produce wind predictions based on recent measurements; both exhibit desirable properties, and common among them is the relatively small amount of training data required to begin producing skilful forecasts.

The ensemble particle swarm optimised FIR filter offers comparable performance to linear methods over all, and significantly improved performance at a subset of locations. The algorithm demonstrates fast convergence and tracking ability compared to linear methods of comparable complexity. In fact, the amount of training data required is tiny compared to algorithms such as the LMS or RLS — a property that could be exploited in a hybrid approach, using the PSO to quickly find an approximate solution before switching to a more robust method, since the PSO itself struggles with the more complex spatial prediction problem.

Kernel methods have been shown to improve over linear methods for multiple step-ahead spatio-temporal prediction. The performance of the KRLS predictor 3–6 hour ahead is particularly good, offering significant improvement over the linear benchmarks for these horizons, whereas the KLMS does not beat the conventional LMS algorithm. The KLMS offers potential for development, see [194] for example, and possible application to directly producing short-term wind power forecasts.

Chapter 5

Very-Short-Term Wind Power Forecasting

In this chapter, a spatio-temporal method for producing very-short-term parametric probabilistic wind power forecasts at a large number of locations is presented. The large-scale integration of wind power presents operational challenges for both power systems [10] and electricity markets [9] due to the stochastic nature of the wind itself. Power systems with a high wind penetration, or areas of concentrated wind generation (e.g. offshore), require skilled very-short-term forecasts to operate effectively, and spatial information is highly desirable. In addition, probabilistic forecasts are widely regarded as necessary for optimal power system management as they quantify the uncertainty associated with point forecasts.

Very-short-term forecasts are required for applications including balancing and the optimal operation of reserves [163, 199], and wind farm control [200, 201]. Furthermore, the stochastic nature of the wind and complexity of the problem calls for a spatio-temporal probabilistic treatment in order to make optimal decisions under inherent uncertainty. Here, a parametric framework is used based on the logit-normal distribution, the parameters of which are forecast. The location parameter for multiple wind farms is modelled as a vector-valued spatio-temporal process, and the scale parameter is tracked by modified exponential smoothing. A state-of-the-art technique for fitting sparse vector autoregressive models is employed to model the location parameter and

demonstrates numerical advantages over conventional vector autoregressive models.

This chapter describes a single predictor for very-short-term probabilistic forecasting on large, previously intractable, spatial scales. The model fitting procedure is completely data driven making it ideal for smart grid applications where many generators share a single, highly interconnected power system and capturing spatial dependence is desirable. Two state-of-the-art statistical techniques are combined: a parametric probabilistic framework based on the logit-normal distribution, as in [202, 203], and model the location parameter of that distribution as a sparse vector autoregressive process [204]. Further, a novel exponential smoothing scheme is described featuring dynamic forgetting factor to track the scale parameter and compare it to the boundary weighted scheme described in [203].

The method is tested on a dataset of 5 minute mean wind power generation at 22 wind farms in Australia. five-minute-ahead forecasts are produced and evaluated in terms of point and probabilistic forecast skill scores and calibration. Conventional autoregressive and vector autoregressive models serve as benchmarks.

The framework for producing spatial probabilistic forecasts based on the logit-normal distribution and transformation is outlined in Section 5.1. The spatio-temporal modelling of the location parameter and the procedure for fitting sparse vector autoregressive models are described in Section 5.2. The tracking of the scale parameter is addressed in Section 5.3. In Section 5.4 the proposed method is tested on to a case study of 22 wind farms in southeastern Australia and results are presented and discussed. Conclusions are drawn in Section 5.5.

5.1 Spatial Probabilistic Forecast Framework

The power generated by a wind farm at any given time is bounded between zero, when no turbines are operating, and nominal, when all turbines are generating their rated power output. As a result, wind power cannot be directly modelled using conventional unbounded Gaussian distributions. Truncated Gaussian, censored Gaussian and generalised logit-normal distributions have all been proposed to model the conditional density of wind power motivated by the desire to work in a linear Gaussian frame-

work [203, 205]. In what follows, data are normalised by their corresponding nominal power such that they occupy the range $[0, 1]$.

In the proceeding derivation wind power observations are assumed to be logit-normal distributed and data are transformed along the lines of [203]. The complete distribution is a discrete-continuous mixture of the logit-normal distribution with the possibility of probability masses on the bounds of the interval $[0, 1]$.

The logit-normal transformation is given by

$$y = \gamma(x) = \ln\Big(\frac{x}{1-x}\Big) \ , \quad x \in (0,1) \ , \tag{5.1}$$

with inverse

$$x = \gamma^{-1}(y) = \Big(1 + e^{-y}\Big)^{-1} \ , \quad y \in \mathbb{R} \quad . \tag{5.2}$$

Assuming that the variable X is logit-normal distributed, the transformed variable $Y = \gamma(X)$ is normally distributed. The logit-normal distribution has density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{x(1-x)} \exp\Big[-\frac{1}{2}\Big\{\frac{\gamma(x)-\mu}{\sigma}\Big\}^2\Big] \ , \tag{5.3}$$

where location and scale parameters μ and σ^2 are directly connected to the mean and variance of $Y \sim N(\mu, \sigma^2)$. The location parameter can be interpreted as the expected value of wind power, and the scale parameter as a measure of spread.

Consider now the stochastic process $\{X_t\}$ and its transformation $\{Y_t\}$ with realisations $\{x_t\}$ and $\{y_t\}$, respectively. The full predictive distribution of X_t, including probability masses on the bounds, is given by the sum of the logit-normal distribution, $L(\mu_t, \sigma_t^2)$, and probability masses w_t^0 and w_t^1 corresponding to zero and nominal power, respectively. It is written as

$$X_t \sim \delta_0 w_t^0 + \delta_1 w_t^1 + (1 - w_t^0 - w_t^1) L(\mu_t, \sigma_t^2) \ , \tag{5.4}$$

with

$$\begin{aligned} w_t^0 &= \Phi\Big\{\frac{\gamma(\eta)-\mu_t}{\sigma_t}\Big\}\,, \\ w_t^1 &= 1-\Phi\Big\{\frac{\gamma(1-\eta)-\mu_t}{\sigma_t}\Big\}\,, \end{aligned} \tag{5.5}$$

where δ_x is the Dirac delta function at x, Φ is the cumulative distribution function of a standard normal variable, and η is the order of the measurement precision. Wind power values less than η, or greater than $1-\eta$, are considered to be 0 or 1, respectively. The key result is that the predictive density of $\{X_t\}$ is parametrized by the conditional mean and variance of $\{Y_t\} \sim N(\mu_t, \sigma_t^2)$ only.

In order to calculate density forecasts for the wind power at some future time, $\{X_{t+k}\}$, it is only necessary to forecast the location and scale parameters of the predictive distribution, which are the mean and variance of the transformed process $\{Y_{t+k}\}$. Finally, $\{Y_{t+k}\}$ is be modelled as an autoregressive process (AR), or a vector autoregressive process (VAR) in the spatial case. Indeed, the spatial case is the main focus of this chapter.

The wind power measurements from multiple wind farms are logit-normal transformed and embedded in a vector-valued time series, and the expected future value for each vector element provides the forecast of the location parameter for the predictive distribution at the corresponding site. The scale parameter could be similarly modelled, but for simplicity it is assumed to be slowly varying and is tracked by an exponential smoothing scheme on a site-by-site basis.

For a vector-valued process, such as a series of measurements made at multiple locations, dependencies between vector elements may exist on a range of scales. Such spatio-temporal dependence can be captured by VAR models and produce more skilful forecasts than independent AR models. However, as the spatial dimension becomes large, VAR models quickly become difficult to estimate as the number of parameters increases with the square of the dimension, and useful spatial information is increasingly diluted. It is therefore advantageous pursue a sparse parametrisation of VAR models whereby coefficients linking sites that exhibit spatial co-dependence are retained in

the model, and those that do not are omitted. The resulting sparse-VAR (sVAR) is a refined parametrisation of the full VAR model and requires a fewer training data compared to the full VAR equivalent.

5.2 From VAR to sVAR

5.2.1 Definitions

First consider the problem of calculating the predictive density for the wind power generation at a single wind farm. The power measured at the wind farm at time t is contained in the time series $\{x_t\}$. The logit-normal transformation of $\{x_t\}$ is $\{y_t\}$ and this series is modelled as an autoregressive process of order p, denoted AR(p). The expression relating the future observation y_{t+k} to previous measurements is written

$$y_{t+k} = \sum_{\tau=1}^{p} a_\tau y_{t-\tau+1} + \epsilon_{t+k} \quad , \tag{5.6}$$

where a_τ is the autoregressive coefficient for the τ^{th} lag, and ϵ_t is additive Gaussian noise with finite variance σ^2. The expected value of y_{t+k} is

$$\hat{\mu}_{t+k} = \sum_{\tau=1}^{p} a_\tau y_{t-\tau+1} \tag{5.7}$$

which along with σ^2 parametrises the predictive distribution of $\{Y_{t+k}\} \sim N(\hat{\mu}_{t+k}, \sigma^2)$ conditional on the p previous measurements.

Next consider the problem of calculating the predictive density for the wind power generation at M spatially separate wind farms. The power measured at each wind farm at time t is contained in the vector valued time series $\{\boldsymbol{x}_t\}$ where each $\boldsymbol{x}_t \in [0,1]^M$. The logit-normal transformation and predictive distributions of $\{\boldsymbol{x}_t\}$ are all calculated by applying Equations (5.1)–(5.5) element-wise. It is then possible proceed working with the transformed vector-valued time series $\{\boldsymbol{y}_t\}$, where $\boldsymbol{y}_t \in \mathbb{R}^M$.

The new time series is modelled as a vector autoregressive process of order p,

VAR(p), expressed as

$$\boldsymbol{y}_{t+k} = \sum_{\tau=1}^{p} \mathbf{A}_\tau \boldsymbol{y}_{t-\tau+1} + \boldsymbol{\epsilon}_{t+k} \quad , \tag{5.8}$$

with matrices $\mathbf{A}_\tau \in \mathbb{R}^{M\times M}$ containing the VAR coefficients, and zero-mean Gaussian noise $\boldsymbol{\epsilon}_t \in \mathbb{R}^M$ with non-singular covariance matrix $\Sigma_{\boldsymbol{\epsilon}}$. The expected value of $\boldsymbol{y}_{t+k}$ is given by

$$\hat{\boldsymbol{\mu}}_{t+k} = \sum_{\tau=1}^{p} \mathbf{A}_\tau \boldsymbol{y}_{t-\tau+1} \quad . \tag{5.9}$$

Typically the VAR coefficients and the noise covariance matrix are determined by maximum likelihood estimation, yielding the Yule-Walker equations for the case when the VAR(p) process is Gaussian and no constraints are placed on the parameters. However, estimating all pM^2 VAR coefficients quickly becomes impractical for models of large spatial dimension and can lead to noisy coefficient estimates and unstable predictions, particularly when insufficient training data are available. A recently proposed method for the sparse estimation of the coefficient matrices offers an alternative to the conventional VAR that can overcome these drawbacks.

5.2.2 sVAR Fitting

A 2-stage procedure for fitting a sparse vector autoregressive model has been proposed by Davis *et al.* in [204]. The first stage selects symmetric pairs of coefficients to be included in the sparse model based on the corresponding pair of time series' conditional dependence. The second stage refines the initial selection based on ranking individual coefficients by their t-statistic. At each stage the set of coefficients selected is that which minimises the Bayesian information criterion (BIC). This approach is detailed in the remainder of this section, for further discussion see Davis *et al.* [204].

Stage 1

The goal of stage 1 is to determine the order of temporal regression, p, and choose N pairs of off-diagonal coefficients to be retained in the sparse model. This is achieved by eliminating pairs of series which are determined to be conditionally uncorrelated and setting the corresponding VAR coefficients (at all lags) to zero. All diagonal coefficients,

i.e. those containing auto-covariate information, are retained in stage 1.

Let $\{\boldsymbol{y}_{t,i}\}$ denote the i^{th} marginal series of the process $\{\boldsymbol{y}_t\}$. If two distinct time series $\{\boldsymbol{y}_{t,i}\}$ and $\{\boldsymbol{y}_{t,j}\}$ $(i \neq j)$ are conditionally uncorrelated then their partial spectral coherence $PSC_{ij}(\omega) = 0$ for $\omega \in (-\pi, \pi]$. The PSC can be computed efficiently from the spectral density matrix $f^Y(\omega)$ of the process $\{\boldsymbol{y}_t\}$, where the (i,j)th element of $f^Y(\omega)$ is the usual (cross-)spectrum between $\{\boldsymbol{y}_{t,i}\}$ and $\{\boldsymbol{y}_{t,j}\}$. The PSC is the negative rescaled inverse of the spectral density matrix, as demonstrated in [206]. Let $g^Y(\omega) = f^Y(\omega)^{-1}$, then

$$PSC_{ij}(\omega) = -\frac{g^Y_{ij}(\omega)}{\sqrt{g^Y_{ii}(\omega) g^Y_{jj}(\omega)}} \ , \quad \omega \in (-\pi, \pi] \ , \tag{5.10}$$

where $g^Y_{ij}(\omega)$ denotes the (i,j)th entry of $g^Y(\omega)$.

In practice, however, the estimated PSC will not be exactly zero for a finite number of samples. We therefore rank each pair of time series by a summary statistic, $\hat{S}_{ij}$, calculated from the estimated PSC, which is denoted $\hat{PSC}_{ij}(\omega)$, taken to be the supremum of the squared PSC estimate, i.e.,

$$\hat{S}_{ij} = \sup_{\omega} |\hat{PSC}_{ij}(\omega)|^2 \quad . \tag{5.11}$$

Large values of $\hat{S}_{ij}$ indicate pairs of series which are likely to be conditionally correlated; therefore, consider the constrained VAR models containing the top N pairs of off-diagonal coefficients plus the M diagonal coefficients, all other coefficients are zero. This reduces the number of parameters to be estimated from pM^2 to $(M + 2N)p$.

Finally, the maximum likelihood estimates of constrained VAR models are calculated for predetermined sets of values for p and N. When VAR parameters are constrained the parameter estimates and covariance matrix $\Sigma_{\boldsymbol{\epsilon}}$ are commingled and their estimates must be updated iteratively until convergence, see [86] for details. The pair of parameters $(\tilde{p}, \tilde{N})$ that minimise the BIC are taken forward to stage 2. The BIC is given by

$$\text{BIC}(p, N) = -2 \log L(\hat{A}_1, ..., \hat{A}_p) + \log T \cdot (M + 2N)p \tag{5.12}$$

where $L(\hat{A}_1, ..., \hat{A}_p)$ is the maximised likelihood of the constrained VAR model, given

by [86, Chapter 5], and T is the length of the training data.

Stage 2

The first stage selects VAR coefficients based on conditional correlation according to the BIC, however, it is unable to discriminate between the $2\tilde{p}$ coefficients associated with each pair of series, nor between the $\tilde{p}$ diagonal coefficients associated with each individual series. The aim of the second stage is therefore to refine the selection of coefficients made by stage 1.

Begin by ranking the non-zero VAR coefficient estimates from the stage 1 model $[\mathbf{A}_\tau]_{ij},\ \tau = 1, ..., \tilde{p}$ by their t-statistic, which is

$$\Lambda_{i,j,\tau} = \frac{[\mathbf{A}_\tau]_{ij}}{\text{s.e.}([\mathbf{A}_\tau]_{ij})} \quad . \tag{5.13}$$

The standard error, s.e.$(\cdot)$, of $[\mathbf{A}_\tau]_{ij}$ is computed from the asymptotic distribution of the constrained maximum likelihood estimator of the stage 1 model, see [86].

Large values of $\Lambda_{i,j,\tau}$ imply significance in the model so the n coefficients with the largest t-statistic values are retained. Once again, the BIC for a set of constrained VAR models containing n parameters is calculated and the value $n = \tilde{n}$ which gives the minimum BIC value is determined. Here, the BIC is given by

$$\text{BIC}(p, \tilde{n}) = -2 \log L(\hat{A}_1, ..., \hat{A}_p) + \log T \cdot \tilde{n} p \quad . \tag{5.14}$$

The resulting sVAR model has an autoregressive order of $\tilde{p}$ and contains $\tilde{n}$ non-zero coefficients; it is denoted sVAR($\tilde{p}$,$\tilde{n}$).

5.2.3 Implementation of sVAR

The spectral density matrix used in the calculation of partial spectral coherence must be estimated from available training data. The periodogram smoothed by a modified Daniell kernel is used here, as in [204], though alternative spectral density estimates could be employed.

The BIC is a smooth convex function of the number of parameters being estimated

which allows for efficient implementation of the sVAR procedure: once the turning point of the function has been found, the minimum is known and the fitting algorithm can advance. Since the parameter estimation and BIC calculation are relatively expensive this represents a significant speed-up over a naive approach.

It is well documented that the properties of meteorological time series, including wind speed, change slowly over time with changes of season and climate; therefore, it is appropriate to allow the parameters of time series models to track this variation, if it is not modelled directly. The same applies to wind power as a weather-dependant process. Recursively updating AR parameters is frequently practised and can easily extend to VAR models; however, it is not possible to modify the sparsity structure of the proposed sVAR model in a simple way. Indeed, the idea of slowly varying parameters conflicts with abruptly choosing to include or remove a coefficient.

In order to capture these gradual changes the sVAR is trained on a window of the most recent measurements, and then re-trained in the same way periodically, i.e., at any time t, the model is trained on past observations between $t - L$ and $t - 1$, where L is the training window length. For comparison, the AR and VAR benchmarks are trained in the same fashion. Note that the parameters of an sVAR (with a fixed sparsity structure) could be updated in a recursive framework (such as a least squares update [115]) in the same way as a conventional AR or VAR model, but this would distract from our main investigation so is not done here.

The scale parameter should also be allowed to track changes in dynamics resulting meteorological variation, and that is the subject of the next section.

5.3 Dynamic Tracking of Scale Parameter

The scale parameter $\sigma^2_{t+k,i}$ of $\{\mathbf{Y}_{t+k,i}\}$ is estimated recursively by exponential smoothing for each site $i \in \{1, ..., M\}$ independently, i.e. assuming no spatial dependence. To avoid notational clutter the second index is dropped in this section.

Here, two variations on exponential smoothing are implemented and their performance compared. First, the boundary weighted forgetting factor, which down-weighs observations when the location parameter is close to the bounds of the $[0, 1]$ interval,

akin to [203]. The logit-normal transformation is particularly sensitive in these regions and this approach is designed to robustify the smoothing scheme. The second scheme employs a 'dynamic forgetting factor' motivated by regime-switching type behaviour often exhibited by weather-dependent processes.

Boundary Weighted Forgetting Factor

In a modification to standard exponential smoothing, observations are down weighted by a factor ω_t when the expected power $\gamma^{-1}(\hat{\mu}_{t+k})$ is close to the bounds due to the sensitivity of the logit-normal transformation in these regions [203]. The factor ω_t is given by

$$\omega_t = 4\gamma^{-1}(\hat{\mu}_{t+k})\big(1 - \gamma^{-1}(\hat{\mu}_{t+k})\big). \tag{5.15}$$

and the smoothing scheme is written

$$\hat{\sigma}^2_{t+k} = \lambda^*_t \hat{\sigma}^2_t + (1 - \lambda^*_t)(y_t - \hat{\mu}_t)^2 \tag{5.16}$$

where $\lambda^*_t = 1 - (1 - \lambda)\omega_t$.

Dynamic Forgetting Factor

The behaviour of wind power generation can switch quickly between periods of smooth generation and periods of volatile generation. In the event of such a switch it is necessary to briefly but dramatically reduce the forgetting factor in order to *forget* out of date, mismatched information. Therefore, when the difference between the squared residual, ϵ^2_t, and estimated scale parameter $\hat{\sigma}^2_t$ is large, the forgetting factor is reduced. The dynamic forgetting factor is given by the logit function as follows,

$$\lambda^*_t = \lambda - \frac{b}{1 + \exp[c(a - \mathcal{E}_t)]} \quad , \tag{5.17}$$

where $\mathcal{E}_t = |\hat{\sigma}^2_t - \epsilon^2_t|$. The parameters a and b control the threshold location and the minimum value that λ^*_t can take, respectively, and c controls the gradient of the transition.

5.4 Application and Case Study

5.4.1 Dataset

The proposed approach is tested on 5 minute mean wind power data provided by the Australian Energy Market Operator [4], which comprises recordings of wind farm power generation at 22 wind farms in south-eastern Australia. Data from 2012 and 2013 are available comprising 210 528 measurements at each site; all have been normalised by the nominal power of the corresponding wind farm so that they occupy the range $[0, 1]$. Wind farm locations are plotted in Figure 5.1. The 2012 data are used as a training set on which the implementation of the fitting procedure is optimised by cross-validation, and the parameters of the exponential smoothing scheme are chosen. The 2013 data are then used to evaluate the performance of the predictor, the results of which are presented and discussed in Section 5.4.3. The results comprise the analysis of more than 2.3 million individual forecasts. The complete dataset as used in this paper is available to download from [207]. In this study, only predictions for $t + 1$ (one step ahead) are considered, though cases with forecast for $t + k$ could be similarly produced.

5.4.2 Implementation

The size of data window, L, used to train the AR, VAR and sVAR is determined heuristically, by cross-validation using the training dataset. The chosen window length is that which minimises the point prediction root-mean-squared error (RMSE) since this is the cost function minimised in the predictors' estimation. A new model is fit for each calendar month to be forecasts to track changes in the time series dynamics (as discussed in Section 5.2.3); this choice is somewhat arbitrary but provides a satisfactory trade-off between accuracy and computational expense. Results of the window length selection procedure are illustrated in Figure 5.2. The optimal window length is $L = 60$ days for the AR model and $L = 150$ days for the sVAR. As already mentioned, the conventional VAR model is extremely data-hungry and computationally expensive to fit and as a result a VAR model cannot be fit with more than $L = 270$ days of training data on the computer being used (64-bit operating system, 8GB of RAM, Intel Core

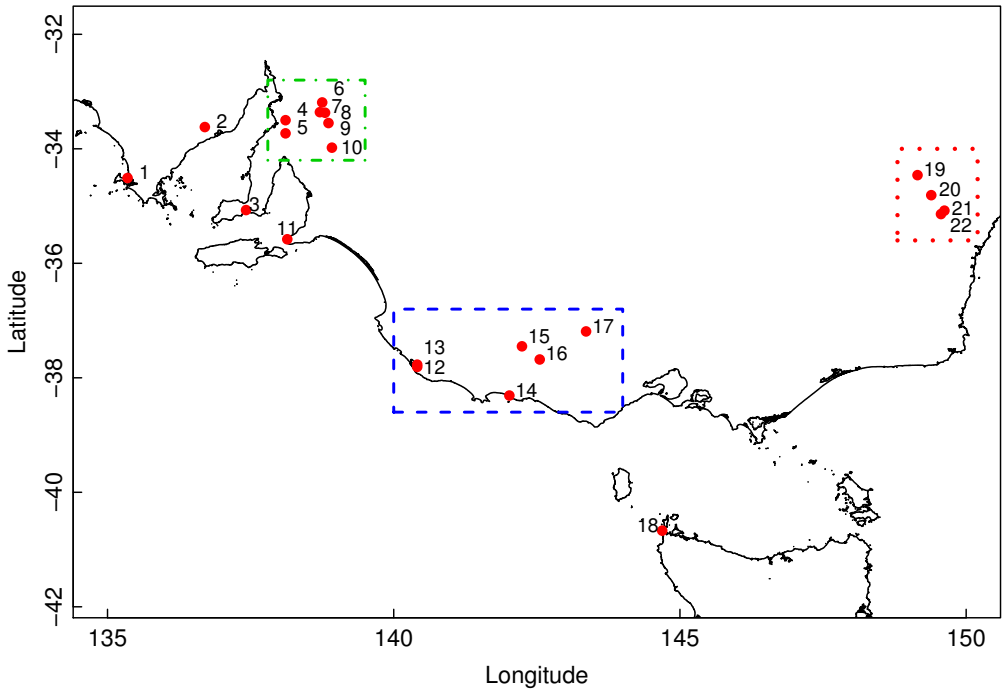


Figure 5.1: Location of 22 sites located in south-eastern Australia used in the data model. Boxed regions correspond to those in Figure 5.3.

i7-2600 3.4GHz processor). Each VAR model is therefore trained on the maximum $L = 270$ days of data.

The optimal window length is directly related to the number of parameters being estimated in each of the three models. The AR has pM parameters so only requires a modest amount of training data, whereas the VAR has pM^2 parameters and as a result requires much more training data to produce reliable parameter estimates. The sVAR offers a compromise: increase the number of parameters to take advantage of spatial information, but only include those parameters deemed significant.

The basic forgetting factor for both exponential smoothing schemes is chosen such that the effective memory is 2000 samples ($\lambda = 0.9995$). The parameters of the dynamic forgetting factor exponential smoothing scheme are chosen by expert judgement such that the forgetting factor does not drop bellow 0.5 ($b = 0.4995$), such that the forgetting factor is reduced when the squared residuals exceed 0.1 ($a = 0.1$), and such that the gradient of the logit function is sharp ($c = 50$).

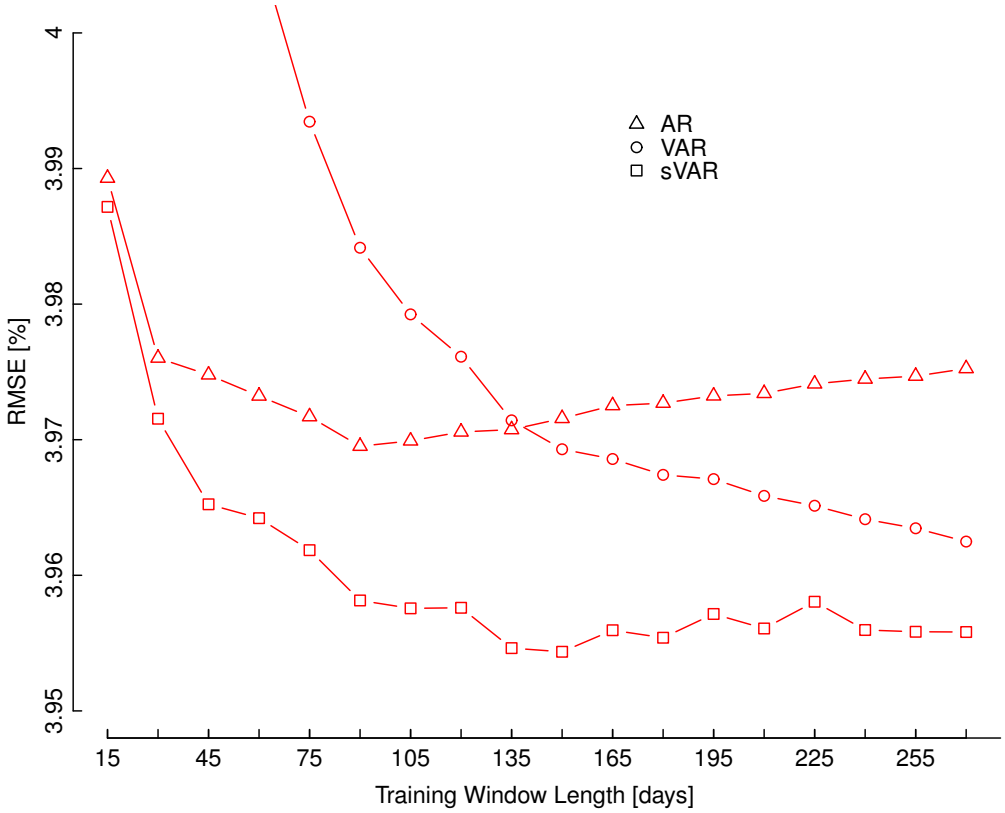


Figure 5.2: Variation of root mean squared error (RMSE) of AR, VAR and sVAR models with training window length.

5.4.3 Results

The proposed technique is implemented on the test dataset in the manor determined by the cross-validation exercise described above.

The 2-stage method for fitting an sVAR model results in the inclusion of 5%–10% of the possible pM^2 parameters. The number of lags is typically $\tilde{p} = 3$. A superposition of the VAR coefficient matrices, taking the absolute value of each element, from one sVAR model is illustrated in Figure 5.3. There is a strong diagonal structure with off-diagonal coefficients appearing in blocks corresponding to groups sites that are close to one another geographically, precisely the sites one would expect to display spatio-temporal dependence.

The 10 minute-ahead sVAR forecasts made over a 24 hour period, and the behaviour of the variable forgetting factor are presented in Figure 5.4. Prediction intervals from 10%–90% are illustrated by shading. The variable forgetting factor behaves as intended, decreasing to allow fast learning when the behaviour switches, and then returning to normal. The width of the prediction intervals behave accordingly and widen quickly during volatile periods, and narrowing during periods of relative calm.

Both point and probabilistic forecast scores are used to quantify the skill of the proposed and benchmark methods. Point forecasts are assessed using the familiar root mean squared error, RMSE $= \sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_t - \hat{x}_t)^2}$, and mean absolute error, MAE $= \frac{1}{T}\sum_{t=1}^{T}|x_t - \hat{x}_t|$, where $\hat{x}_t = \gamma^{-1}(\hat{\mu}_t)$ is the predicted value of x_t.

The skill of the distributional forecasts is quantified by the continuous rank probability score (CRPS) and log score [152]. The CRPS is given by

$$\text{CRPS} = \frac{1}{T}\sum_{t=1}^{T}\int_0^1 \{F(x|\hat{\mu}_t, \hat{\sigma}_t) - \mathbf{1}(x \geq x_t)\}^2 dx \tag{5.18}$$

where F is the cumulative form of the predictive distribution and $\mathbf{1}(\cdot)$ is the indicator function. CRPS rewards sharpness and reduces to MAE when the forecast is deterministic.

The log score is the mean negative log of the predictive distribution evaluated at the corresponding observation, Log Score $= \frac{1}{T}\sum_{t=1}^{T} -\ln\big(f(x_t|\hat{\mu}_t, \hat{\sigma}_t)\big)$. Due to its

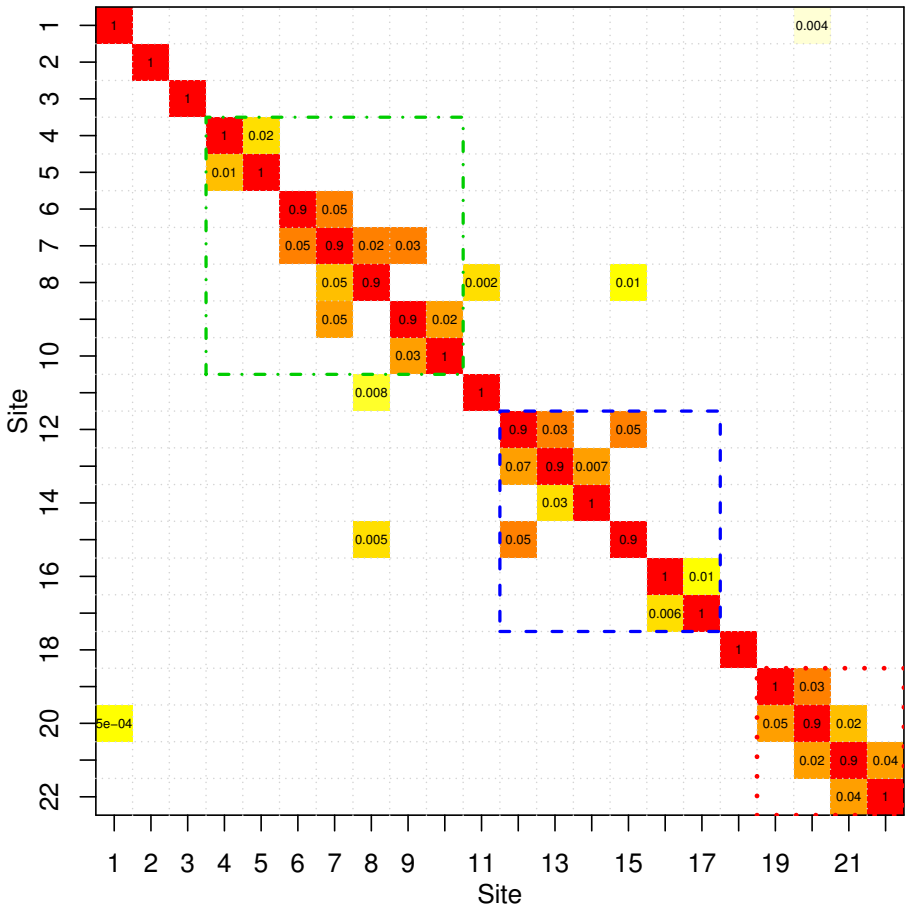
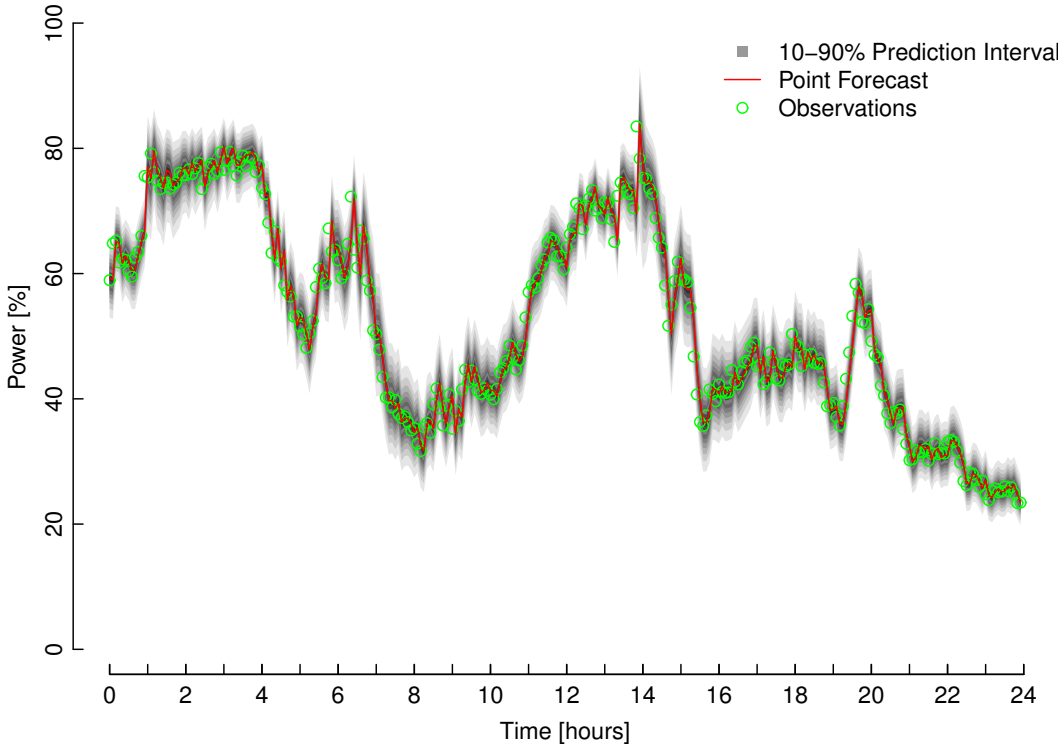
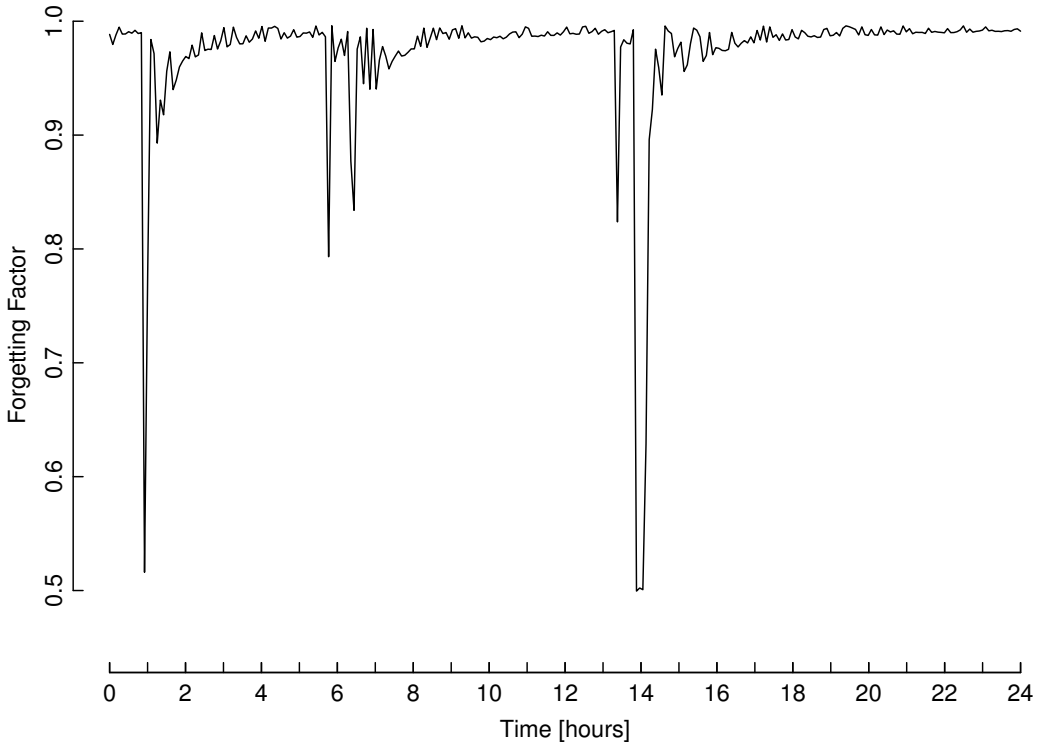


Figure 5.3: Superposition of January 2013 sVAR coefficient matrices taking absolute values and displaying 1 s.f. Blank entries correspond to coefficients not included in the sparse model and are therefore equal to zero at all lags. Boxed regions correspond to those in Figure 5.1.



(a) Point and probabilistic forecasts made 5 minutes (1 step) ahead.



(b) Value of the dynamic forgetting factor, λ_t^*.

Figure 5.4: Probabilistic forecasts and value of the dynamic forgetting factor at site 9 for July 11$^{\text{th}}$ 2013.

logarithmic nature, the log score is not as robust as the CRPS: measurements in the tails of the predictive distribution are heavily penalised and the score returns ∞ if a single measurement falls where the predictive distribution is numerically zero.

Point and probabilistic forecast skill scores are listed in Table 5.1 and probabilistic scores are broken-down by calendar month in Table 5.2. The persistence point forecast, which is simply $\hat{x}_{t+k} = x_t$, is also included in Table 5.1. Point forecast scores show that the sVAR improves on all the benchmarks in terms RMSE, and all but persistence in terms of MAE. Persistence does not offer probabilistic information, which is required for optimal decision making under uncertainty, hence the move to more sophisticated approaches.

With the boundary weighted tracking of the scale parameter, the sVAR performs very well in terms of CRPS but has a poor log score, when compared to the other models. The high log score is an effect of the very sharp predictive distribution close to the upper and lower bounds where measurements are more likely to be found in the tails of the distribution. The AR and VAR models, with their higher variance and broader predictive distributions, are not exposed to this affect as frequently and this is reflected in their comparatively low log scores.

When the scale parameter is tracked by the proposed dynamic forgetting factor scheme, all three models see significant improvement in both CRPS and log score compared to the boundary weighted scheme. Notably, the improved behaviour of the predictive distributions close to the bounds has brought the log score of the sVAR in line with the AR and VAR models. In this case, the sVAR performs marginally better than the two benchmarks in terms of both CRPS and log score.

Reliability (or calibration) of probabilistic forecasts is critical and can be assessed with quantile-quantile reliability diagrams, such as in Figure 5.5. A calibrated forecast with nominal proportion α should cover the observation $\alpha\%$ of the time. In Figure 5.5 nominal quantiles from 5% to 95% in steps of 5% are evaluated.

The forecasts produced by the sVAR with the boundary weighted scale factor smoothing is reliable and the best calibrated of the six forecasts, followed by the sVAR with dynamic smoothing. The boundary weighted smoothing scheme results in better

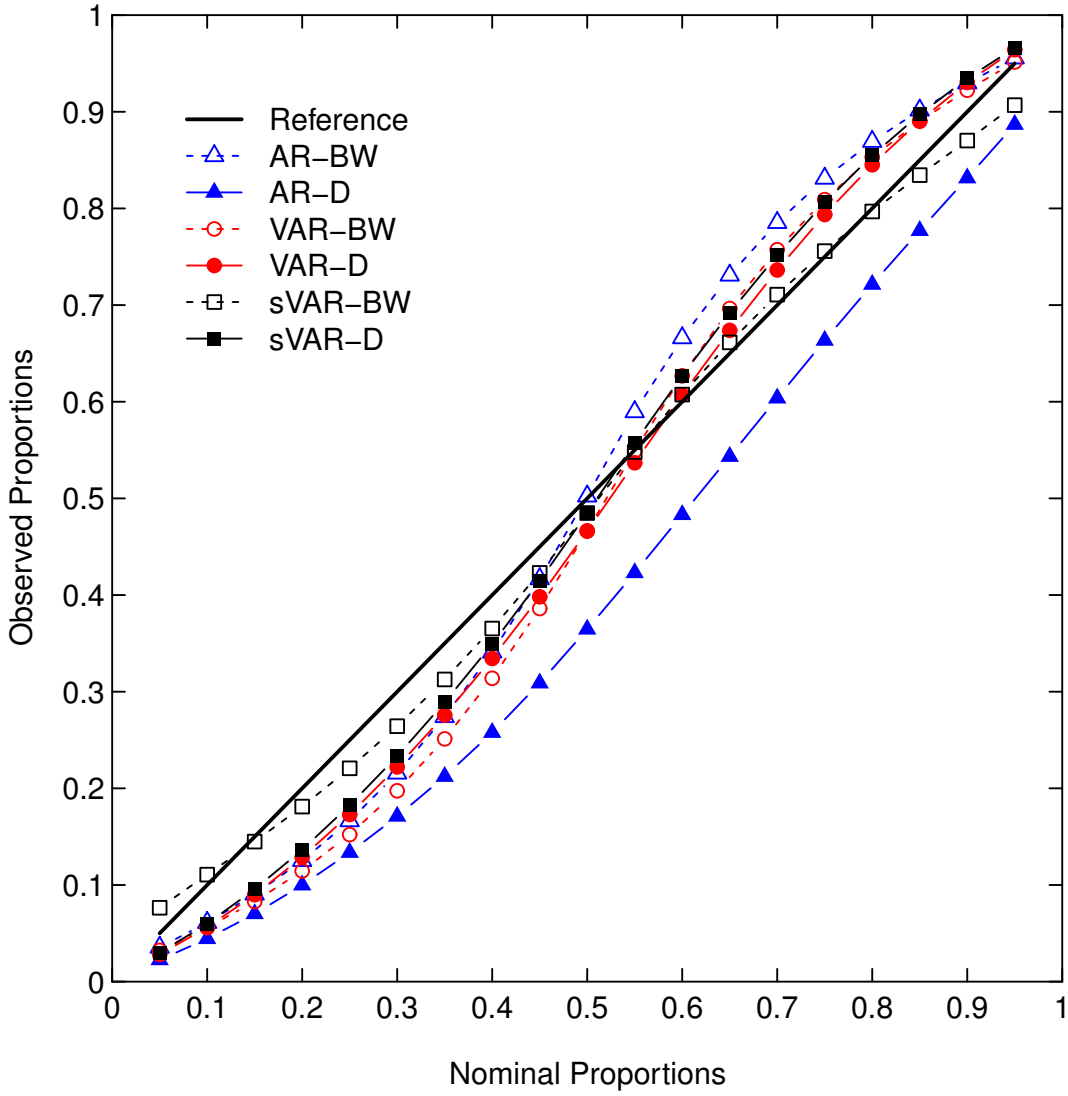


Figure 5.5: Reliability diagram for the AR, VAR and sVAR models with boundary weighted (BW) and dynamic (D) forgetting factors.

		Persistence	AR	VAR	sVAR
	RMSE	3.956	3.970	3.962	3.954
	MAE	2.308	2.347	2.358	2.343
BW λ	CRPS	n/a	1.843	1.837	1.801
	Log Score	n/a	5.080	5.067	5.909
Dynam. λ	CRPS	n/a	1.751	1.751	1.745
	Log Score	n/a	4.634	4.629	4.622
Δ% vs BW λ	CRPS Δ%	n/a	*5.0%*	*4.7%*	*3.0%*
	Log Score Δ%	n/a	*8.8%*	*8.6%*	*21.8%*

Table 5.1: Mean skill scores (RMSE, MAE and CRPS as % of nominal power) across all sites with % improvement (Δ%) for dynamic vs boundary weighted (BW) forgetting factor.

calibration than the dynamic smoothing scheme for the sVAR and AR models, but the opposite is true for the conventional VAR. The calibration of forecasts produced by the AR model with dynamic smoothing is particularly poor.

5.4.4 Discussions

It has been demonstrated that the proposed approach produces forecasts that are a non-negligible improvement on two competitive benchmarks in terms of several skill scores and reliability, while also offering attractive numerical properties through sparse parametrisation. The sVAR makes it possible to model data of high spatial dimension that would be impractical, or impossible, with a conventional VAR approach. In addition, the data-driven detection of dependence structures means that the benefits of a spatial treatment can be realised without knowledge of precise locations, or in situations where many generators are located in a small area, as is commonplace in the smart grid paradigm. This technique is equally applicable to other forecasting problems where VARs have been used, such as wind speed [157] and solar power forecasting [208], including short-term forecasting at other temporal resolutions, e.g. hourly.

However, the sVAR comes with some limitations: Regression parameters are commonly updated by a process of recursive estimation [115], or replaced with coefficient functions of some covariate such as wind direction [122, 209]. While in principle these

Month		AR	VAR	sVAR
January	CRPS	1.910	**1.896**	1.897
	Log Score	4.788	4.788	**4.781**
February	CRPS	1.826	1.819	**1.812**
	Log Score	4.752	4.755	**4.749**
March	CRPS	1.796	1.780	**1.779**
	Log Score	4.685	4.691	**4.681**
April	CRPS	**1.375**	1.383	1.380
	Log Score	4.351	4.355	**4.337**
May	CRPS	**1.617**	1.637	1.634
	Log Score	4.570	**4.565**	**4.565**
June	CRPS	1.486	1.500	**1.483**
	Log Score	4.434	4.435	**4.425**
July	CRPS	**1.544**	1.567	1.548
	Log Score	4.460	4.449	**4.436**
August	CRPS	1.831	1.840	**1.829**
	Log Score	4.712	4.697	**4.686**
September	CRPS	1.717	1.710	**1.700**
	Log Score	4.606	4.595	**4.594**
October	CRPS	2.001	1.999	**1.990**
	Log Score	4.759	**4.739**	**4.739**
November	CRPS	2.020	**2.007**	2.009
	Log Score	4.790	4.778	**4.777**
December	CRPS	1.883	**1.871**	1.875
	Log Score	4.703	4.697	**4.692**
All	CRPS	1.751	1.751	**1.745**
	Log Score	4.634	4.629	**4.622**

Table 5.2: Mean probabilistic forecast skill scores with dynamic forgetting factor broken down by calendar month (CRPS as % of nominal power). The best scores are highlighted in bold.

techniques could be applied to an sVAR model, they would not be able to capture possible changes in the sparsity structure.

Computational cost is of interest: while the MLE of a single constrained VAR model takes around 2 minutes, compared to 4 for the full VAR, the calculation is repeated making the total time to fit an sVAR an order of magnitude larger than the conventional VAR. However, the stopping criterion described in Section 5.2.3 may be refined, and other speed-ups are possible such as parallelising the fitting procedure. There exist alternative methods for fitting sparse regression models, such as quasi-MLE, [210], and penalised linear regression (e.g. *lasso* [211]) which can be implemented by very efficient algorithms which are available in common software packages. However, reformulating the problem as one of linear regression comes at a cost as both the temporal ordering of samples and any error cross-covariance between spatial locations is negated.

Furthermore, retaining full covariance information may offer opportunities for future development. While the deterministic part of the forecast methodology described in this paper utilises spatial information, the scale parameter, and by extension the predictive distribution, for each location are calculated independently. A more general probabilistic forecast could consider the full joint predictive distribution taking into account the full covariance structure of observations.

The framework facilitated by the logit-normal transformation allows us to work in the familiar Gaussian domain, however, a generalisation of this transformation has been proposed in [203] for wind power forecasting. By including a shape parameter to control the skewness of the transformation, the properties of the transformed data may be improved. The optimal shape parameter to fit the marginal distribution of the data can be calculated by standard techniques, however, the same is not true of the conditional distributions, which are of concern here. In [203] the optimal shape parameter for the conditional distributions of a univariate time series is determined by an iterative process, which would be extremely time consuming in the spatial case, particularly if individual shape parameters were assigned to each location. Furthermore, the effects of using different shaped transformations on the spatio-temporal dependencies of the transformed data are unknown. For these reason, the generalised logit-normal

transformation is left for future investigation.

5.5 Conclusions

This chapter describes a large-scale spatial technique for producing very-short-term probabilistic forecasts of wind power generation at multiple locations. A parametric framework for distributional forecasts based on the logit-normal transformation and distribution is combined with a spatio-temporal model for the distribution's location parameter, and two competing smoothing schemes for it's scale parameter are presented. The location parameter is first modelled as a vector autoregressive process, and then as a sparse vector autoregressive process (sVAR), dramatically reducing the number of coefficients requiring estimation, and by extension the computational expense of model fitting and the volume training data required.

In a case study, the sVAR technique has been used to produce 5 minute ahead probabilistic forecasts of wind power at 22 wind farms in south-eastern Australia for a test period of 1 year. The performance of the sVAR is compared to conventional VAR and AR models yielding improvement in terms of both deterministic and probabilistic skill scores, as well as in the reliability of the distributional forecasts.

This work was motivated by the desire to produce accurate very-short-term forecasts at multiple wind farms, ultimately on a national scale, i.e., at 100s of wind farms. Future work should extend to spatial dimensions of this order, other forecast horizons, and consider building an adaptive sVAR, possibly with a dynamic sparsity structure. The parametric framework could also be extended by moving to the generalised logit-normal distribution and transformation which would require the development of an efficient method for determining the optimal shape parameter(s) with respect to conditional distributions.

Chapter 6

Conclusions

As the penetration of wind power increases on power systems around the world so does the importance of having access to accurate wind power forecasts on a wide range of spatial and temporal scales. This thesis has described a range of new techniques for short-term wind and wind power forecasting. The considered approaches have all been developed to utilise readily available spatial information in a computationally efficient way. Capturing the relationships between geographically separate locations has been shown to increase forecast skill while providing forecasts at multiple locations simultaneously. Furthermore, improving the scalability of such predictors has been made possible by employing a sparse predictor model.

6.1 Summary of Contributions

Wind speed and direction forecasts are often combined with wind farm power curves to produce wind power forecasts. In Chapter 3 wind speed and direction were modelled as the magnitude and phase of complex numbers and are forecast by highly efficient complex-linear multi-channel predictors. Inspired by developments in the field of signal processing, these methods are easy to implement, fast to compute and produce point predictions with lower error than their non-spatial equivalent and persistence, the standard benchmark. The wind speed part of the forecast is of comparable accuracy to the equivalent real-valued speed-only methods, indicating that incorporating direction does not come at the expense of accurate speed prediction.

To refine the forecasts, the coefficient matrices of the predictor have been 1) estimated recursively and 2) conditioned on the time of year or mean wind direction. The recursive approach provides a step-improvement in performance compared to the static approach, and the conditional predictors provide the same improvement again. Only a negligible difference in performance between the predictors conditioned on time of year or wind direction was observed.

Finally, the cost of assuming that complex-valued wind time series are *proper* (in order to produce the most computationally efficient predictor possible) is quantified by producing and testing a widely-linear multi-channel filter. The RMSE for the widely-linear filter was found to be 0.4–1.1% lower than the complex-linear filter for 1-hour-ahead prediction, and 1.0–3.9% lower for 6-hour-ahead prediction, depending on the variability of wind direction at each test site.

Two quite different non-linear methods were employed to forecast the wind in Chapter 4. The first, an ensemble of particle swarm optimised prediction filters, is extremely fast to train and requires only a short training period before its predictions are reliable, though it was unable to make use of spatial information. While this approach showed no improvement in general over the spatial linear methods of Chapter 3, the performance at 3 sites was significantly better. However, due to the non-physical nature of the algorithm, it is difficult to identify the reason(s) for this.

The second part of Chapter 4 examined kernel methods, a class of learning algorithm that facilitates linear processing in some high-dimensional feature space at very low computational expense using the kernel trick. The kernelised LMS and RLS algorithms were investigated with the latter showing most promise. Both approaches utilise a *dictionary* of features against which input data is compared, a kind of pattern recognition. While the performance of the KRLS is comparable to the linear methods of Chapter 3 for 1 and 2 hour-ahead forecasts, the 3–6 hour-ahead forecast are a significant improvement.

Finally, a method for producing spatial forecasts on a large-scale (many 10s or hundreds) is developed in Chapter 5, specifically for producing very-short-term probabilistic wind power forecasts. Five-minute-ahead probabilistic forecasts are produced

using only power measurements as inputs to a sparse vector autoregressive model. The doubly-bounded and non-Gaussian nature of wind power predictive distributions is accommodated by the logit-normal transformation and distribution which makes parametric probabilistic forecasting straightforward. The sparse-VAR approach enables large numbers of locations to be modelled by identifying and retaining VAR parameters that capture spatial dependence, and excluding those which do not. In the presented case study, the approach only provides a modest improvement in point forecast accuracy over persistence, but provides sharp and reliable distributional forecasts.

6.2 Future Work

While several methods have been developed for producing spatio-temporal forecasts during this research programme there are of course many possible developments and extensions that could improve their performance and usefulness to end users. This section outlines some thoughts on possible future work.

Linear Methods: While sophisticated approaches tailored to specific applications may be able to offer superior performance, linear methods for producing short-term forecasts will remain important because they are simple to implement and robust. They are also easily extended to take advantage of exogenous information, such as time of year or wind direction as demonstrated in this thesis. However, there are many other sources of potentially useful information, and no clear way of choosing between them (and the danger of over parametrisation if they are all included). Possible candidates include time of day, to capture the diurnal trend, atmospheric stability, or air temperature and pressure. The latter two can serve as a proxy for air density which could have practical benefits in short-term wind power forecasting since the power in the wind is proportional to both the cubed wind speed and air density. A simple and robust technique for identifying useful information and model selection could be a powerful and valuable tool.

Kernel Methods: The application of established kernel methods in this thesis has found them to be capable of producing high quality wind forecasts capitalising on spacial information, particularly for multiple step-ahead prediction. However, their

performance suffers dramatically if the spatial dimension is increased beyond a small number of sites, suggesting that they struggle to discriminate between useful and redundant inputs. Advances in kernel methods may be able to address this drawback, similarly, effectively partitioning large scale spatial problems into numerous small-scale problems could also provide a solution.

Sparse-VAR Power Forecasting: The sVAR, as described in Chapter 5, offers a powerful data-driven method for *sparsifying* conventional VAR models, but the time consuming process of iteratively estimating multiple constrained-VAR models is a significant drawback. If model selection could be performed by a method other than minimising the BIC this could be avoided. Furthermore, the skill of the probabilistic forecast could be improved if the generalised logit-normal transformation and distribution were utilised; however, this would require the development of an efficient method for determining the optimal shape parameter(s) with respect to conditional distributions, and consideration for the effect of using different shape parameters for different locations on spatio-temporal dependencies.

Appendix A

Definitions

A.1 Linear and Non-linear Stochastic Processes

A stochastic process $(Y_t, t \in \mathbb{Z})$ is said to be a *linear process* if for every $t \in \mathbb{Z}$

$$Y_t = \sum_{j=0}^{\infty} a_j x_{t-j} \tag{A.1}$$

where $a_0 = 1$, $(x_t, t \in \mathbb{Z})$ is i.i.d. with $E\{x_t\} = 0$, $E\{x_t^2\} < \infty$, and $\sum_{j=0}^{\infty} a_j < \infty$. That is to say that a stochastic process Y_t is linear if it can be expressed as the finite linear combination of past values of a zero mean, finite variance explanatory variable x_t. Any process which does not meet these conditions is described as non-linear.

A.2 Stationarity

A stochastic process is said to be **strict-sense stationary** if all of its stochastic properties are invariant to shifts of the time origin. That is to say that the joint distribution of the random vector $[x(t_1), x(t_2), \ldots, x(t_N)]$ is the same as the joint distribution of $[x(t_1 + \tau), x(t_2 + \tau), \ldots, x(t_N + \tau)]$ for any dimension N and shift τ.

A stochastic process is said to be **wide-sense stationary** if its mean is invariant to shifts of the time origin and the covariance function depends only on the time difference

$\tau = t_1 - t_2$ such that

$$r_x(t_1, t_2) = r_x(t_1, t_1 - \tau) = r_x(\tau) \quad . \tag{A.2}$$

In general, a strict-sense stationary process is wide-sense stationary, but a wide-sense stationary process is not necessarily strict-sense stationary. However, since a Gaussian random process is defined by only its mean and variance, if it is wide-sense stationary, it is also strict-sense stationary.

Two stochastic processes are said to be **jointly stationary** if both processes are individually stationary, and their cross correlation is invariant under shifts of the time origin.

A stochastic process is said to be **cyclo-stationary** if its stochastic properties vary periodically with shifts of the time origin. For example, the mean of a cyclo-stationary process must satisfy

$$\mu(t) = \mu(t + T) \tag{A.3}$$

where T is the period of the cyclo-stationary variation.

A.3 Covariance Matrices

The covariance matrix, $\mathbf{R}_{xx}$ of vector-valued stochastic process $\boldsymbol{x}[t] \in \mathbb{C}^N$ is a measure of how much the elements of $\boldsymbol{x}[t]$ vary with respect to each other over time, for a given time lag, τ. It is defined by

$$\mathbf{R}_{xy}[t, \tau] = E\{(\boldsymbol{x}[t] - E\{\boldsymbol{x}[t]\})^{\mathrm{H}}(\boldsymbol{x}[t - \tau] - E\{\boldsymbol{x}[t - \tau]\})\} \tag{A.4}$$

however an number of simplifications can often be made. The expectation of $\boldsymbol{x}[t]$ is often zero, or is made zero by some transformation, and if $\boldsymbol{x}[t]$ is wide-sense stationary, the dependence of $\mathbf{R}_{xx}$ on t can be dropped, yielding the more common expression

$$\mathbf{R}_{xx}[\tau] = E\{\boldsymbol{x}[t]^{\mathrm{H}}\boldsymbol{x}[t - \tau]\} \tag{A.5}$$

For a complex processes, $\mathbf{R}_{xx}$ takes the anti-symmetric Hermitian form $\mathbf{R}_{xx}^{\mathrm{H}}[\tau] = \mathbf{R}_{xx}[-\tau]$.

The cross-covariance matrix, $\mathbf{R}_{xy}$ of vector-valued stochastic processes $\boldsymbol{x}[t], \boldsymbol{y}[t] \in \mathbb{C}^N$ is a measure of how much the elements of $\boldsymbol{x}[t]$ vary with the elements of $\boldsymbol{y}[t]$ over time, again at a given time lag, τ. It is defined by

$$\mathbf{R}_{xy}[t,\tau] = E\{(\boldsymbol{x}[t] - E\{\boldsymbol{x}[t]\})^{\mathrm{H}}(\boldsymbol{y}[t-\tau] - E\{\boldsymbol{y}[t-\tau]\})\} \tag{A.6}$$

again, the inner expectations are often zero. However, $\boldsymbol{x}[t]$ and $\boldsymbol{y}[t]$ must be jointly stationary for the dependence on t to be dropped. In the jointly stationary case

$$\mathbf{R}_{xy}[\tau] = E\{\boldsymbol{x}[t]^{\mathrm{H}}\boldsymbol{y}[t-\tau]\} \quad . \tag{A.7}$$

Again from complex processes, $\mathbf{R}_{xy}$ takes the anti-symmetric Hermitian form $\mathbf{R}_{xy}^{\mathrm{H}}[\tau] = \mathbf{R}_{xy}[-\tau]$.

Appendix B

Mathematical Results

B.1 Weibull & Rayleigh Distributions

The Weibull distribution is well established as the standard distribution for wind speed for sufficiently long periods of time. It was inspired in Weibull's original 1951 paper [44] buy the desire for a simple description of problems where the occurrence of an event in any part of an object may be said to have occurred in the object as a whole. For example, if a single link in a chain fails, the chain as a whole is said to have failed. The resulting distribution was found to fit data in a number of situations better than previously existing distributions.

The two parameter probability density function of a Weibull random variable x is given by

$$f(x;k,\lambda) = \frac{k}{\lambda}\Big(\frac{x}{\lambda}\Big)^{k-1} e^{-(\frac{x}{\lambda})^k} \qquad x > 0 \tag{B.1}$$

where $k > 0$ is the shape factor and $\lambda > 0$ is the scale factor. For the case when $k = 2$, the Weibull distribution reduces to the Rayleigh distribution which has the pleasing property of being distribution of the hypotenuse of two perpendicular zero mean Gaussian random variables. This is of interest since many approaches to directional wind forecasting treat the wind as a bivariate random variable in Cartesian space.

Derivation

Let X and Y be perpendicular i.i.d. Gaussian random variables, and $Z^2 = X^2 + Y^2$. The probability density functions of X and Y are

$$P_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \tag{B.2}$$

and

$$P_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{y^2}{2\sigma^2}} \tag{B.3}$$

respectively, and as X and Y are independent, their joint probability density function is

$$P_{XY}(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad . \tag{B.4}$$

The joint probability that X lies between x and $x+dx$, and that Y lies between y and $y+dy$ is therefore

$$P_{XY}(x < X \leq x+dx, y < Y \leq y+dy) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} dxdy \quad . \tag{B.5}$$

Making a transformation to polar coordinates, $z = \sqrt{x^2+y^2}$ and $dxdy = zdzd\theta$,

$$P_{XY}(x < X \leq x+dx, y < Y \leq y+dy) = P_{Z\Theta}(z < Z \leq z+dz, \theta < \Theta \leq \theta+d\theta) \quad , \tag{B.6}$$

$$P_{Z\Theta}(z < Z \leq z+dz, \theta < \Theta \leq \theta+d\theta) = \frac{1}{2\pi\sigma^2} e^{-\frac{z^2}{2\sigma^2}} zdzd\theta \quad . \tag{B.7}$$

Finally, since Z and Θ are independent, and $P_{Z\Theta}$ does not depend on θ, Θ has the uniform distribution

$$P_\Theta(\theta) = \frac{1}{2\pi} \quad , \quad 0 < \theta \leq 2\pi \tag{B.8}$$

and Z has the distribution

$$P_Z(z) = \frac{z}{\sigma^2} e^{-\frac{z^2}{2\sigma^2}} \quad , \quad z \geq 0 \tag{B.9}$$

which is the Rayleigh distribution.

B.2 Equivalence of VAR and Wiener Filter

The multi-channel predictor with constant coefficients described in this section produces each prediction in the same way as the VAR method [86, 92] but for complex-valued signals. The only difference being the different organisation of data and prediction coefficient matrices. Here, this equivalence is demonstrated using.

The VAR(p) method considering m separate sites is formulated, equation (3) in [92], as follows:

$$\boldsymbol{y}(t) = \boldsymbol{\Phi}_1\boldsymbol{y}(t-1) + \boldsymbol{\Phi}_2\boldsymbol{y}(t-2) + ... + \boldsymbol{\Phi}_p\boldsymbol{y}(t-p) + \boldsymbol{e}(t) \tag{B.10}$$

Where $\boldsymbol{y}(t)$ is a $m \times 1$ vector containing the wind speed for each site at time t and $\boldsymbol{e}(t)$ is a $p \times 1$ vector containing the prediction error. The prediction can therefore be expressed as $\boldsymbol{y}(t) - \boldsymbol{e}(t)$. Each $\boldsymbol{\Phi}_i$ is a $m \times m$ matrix containing the coefficients of the model relating the influences between the wind speed at the m sites.

The Yule-Walker equations are obtained by post–multiplying (B.10) by $\boldsymbol{y}^{\mathrm{H}}(t-h)$ and taking the expectation of both sides [86].

$$E\{\boldsymbol{y}(t)\boldsymbol{y}^{\mathrm{H}}(t-h)\} = \sum_{k=1}^{p} \boldsymbol{\Phi}_k E\{\boldsymbol{y}(t-k)\boldsymbol{y}^{\mathrm{H}}(t-h)\} \tag{B.11}$$

Now, using the covariance matrix defined by

$$\boldsymbol{\Gamma}_y(h) = E\{\boldsymbol{y}(t)\boldsymbol{y}^{\mathrm{H}}(t-h)\}, \tag{B.12}$$

(B.11) becomes

$$\boldsymbol{\Gamma}_y(h) = \sum_{k=1}^{p} \boldsymbol{\Phi}_k \boldsymbol{\Gamma}_y(h-k). \tag{B.13}$$

The Wiener–Hopf equation is obtained by operating on the integral describing the prediction, $\hat{y}_{t+\lambda}$, of the zero–mean stochastic process x_t at time $t + \lambda$.

$$\hat{y}(t+\lambda) = \int_{-\infty}^{t} w(t,s)x(s)ds \tag{B.14}$$

Choosing w such that the prediction error is minimised in the mean-squares sense and applying the orthogonality principal, that is requiring the estimation error to be perpendicular to all data used to generate the estimate, yields the *Wiener-Hopf* equation.

$$R_{yx}(t+\lambda) = \int_{0-}^{\infty} w(\tau) R_x(t-\tau) d\tau, \quad \forall t > 0 \tag{B.15}$$

For a discrete time, stationary random vector, $\boldsymbol{y}[n]$, equation (B.14) for a filter with p coefficients can be written as

$$\hat{\boldsymbol{y}}[n] = \sum_{k=1}^{p} \boldsymbol{w}[k]\boldsymbol{x}[n-k] \tag{B.16}$$

The *Wiener filter* is the optimum filter in the minimum mean–square error sense, here the prediction error is

$$\boldsymbol{e}[n] = \boldsymbol{x}[n] - \hat{\boldsymbol{y}}[n]. \tag{B.17}$$

The filter is optimised when the mean–squared error,

$$||\boldsymbol{e}[n]||^2 = E\{\boldsymbol{e}[n]\boldsymbol{e}^{\mathrm{H}}[n]\}, \tag{B.18}$$

is minimised. By the orthogonality principal, which states that this error is minimised when orthogonal to each of the data vectors, $\boldsymbol{x}[t-l], l = 1, 2, ...$ The inner product, denoted by $< \boldsymbol{A}, \boldsymbol{B} >= E\{\boldsymbol{A}\boldsymbol{B}^{\mathrm{H}}\}$, of the error and data vectors is therefore zero [212].

$$\Big\langle \boldsymbol{x}[n] - \sum_{k=1}^{p} \boldsymbol{w}[k]\boldsymbol{x}[n-k], \boldsymbol{x}[n-l] \Big\rangle = 0 \tag{B.19}$$

for $l = 0, 1, 2, ...$; or

$$\Big\langle \boldsymbol{x}[n], \boldsymbol{x}[n-l] \Big\rangle = \Big\langle \sum_{k=1}^{p} \boldsymbol{w}[k]\boldsymbol{x}[n-k], \boldsymbol{x}[n-l] \Big\rangle \tag{B.20}$$

and using definition of the inner product

$$E\{\boldsymbol{x}[n]\boldsymbol{x}^{\mathrm{H}}[n-l]\} = \sum_{k=1}^{p} \boldsymbol{w}[k] E\{\boldsymbol{x}[n-k]\boldsymbol{x}^{\mathrm{H}}[n-l]\} \tag{B.21}$$

$$\mathbf{\Gamma}_x[l] = \sum_{k=1}^{p} \boldsymbol{w}[k]\mathbf{\Gamma}_x[l-k]. \tag{B.22}$$

Comparing (B.22) to (B.13) we see that the Wiener–Hopf equation for a discrete time signal and p coefficient filter reduces to the Yule–Walker equation for a VAR(p) process.

B.3 A Test for Impropriety of Complex-Valued Gaussian Vectors

Here a description of how to implement a test for the impropriety of complex-valued Gaussian vectors is given. Its derivation follows that of the test described in [178]; for a full derivation and discussion of associated statistics please refer to the original paper, and references therein.

Let $\boldsymbol{x} = [x_1, x_2, ..., x_p]$ denote a complex-valued zero mean random vector of length p. The covariance matrix of $\boldsymbol{x}$ is defined as $\mathbf{R}_{\boldsymbol{xx}} = E\{\boldsymbol{x}\boldsymbol{x}^{\mathrm{H}}\}$, which is Hermitian and positive semi-definite. In addition, there is a complimentary covariance matrix associated with $\boldsymbol{x}$ which is defined as $\tilde{\mathbf{R}}_{\boldsymbol{xx}} = E\{\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$, which is complex and symmetric.

The complete second order structure of $\boldsymbol{x}$ is contained in the covariance matrix of the so-called augmented vector

$$\underline{\boldsymbol{x}} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{x}^* \end{bmatrix} \tag{B.23}$$

which is called the augmented covariance matrix, is give by

$$\underline{\mathbf{R}}_{\boldsymbol{xx}} = E\{\underline{\boldsymbol{x}}\,\underline{\boldsymbol{x}}^{\mathrm{H}}\} = E\left\{ \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{x}^* \end{bmatrix} \begin{bmatrix} \boldsymbol{x}^{\mathrm{H}}, & \boldsymbol{x}^{\mathrm{T}} \end{bmatrix} \right\} = \begin{bmatrix} \mathbf{R}_{\boldsymbol{xx}} & \tilde{\mathbf{R}}_{\boldsymbol{xx}} \\ \tilde{\mathbf{R}}^*_{\boldsymbol{xx}} & \mathbf{R}^*_{\boldsymbol{xx}} \end{bmatrix} \quad . \tag{B.24}$$

If the complimentary covariance of $\boldsymbol{x}$ is $\tilde{\mathbf{R}}_{\boldsymbol{xx}} = \mathbf{0}$, $\boldsymbol{x}$ is said to be a *proper complex-valued random vector*; conversely, if $\tilde{\mathbf{R}}_{\boldsymbol{xx}} \neq \mathbf{0}$, $\boldsymbol{x}$ is said to be an *improper complex-valued random vector*.

Hypothesis tests have been developed to determine if a random vector is proper or not, with the hypothesis $H_0 : \tilde{\mathbf{R}}_{\boldsymbol{xx}} = \mathbf{0}$ being tested against $H_1 : \tilde{\mathbf{R}}_{\boldsymbol{xx}} \neq \mathbf{0}$. Tests based

on i.i.d. samples of $\boldsymbol{x}$ are described in [213,214], but here the real valued representation described in [178] is followed since it allows for easy reference to the rich discussion in the original paper and to accommodate links available in the statistical literature.

Real Valued Notation

The complex process $\boldsymbol{x}$ can be written $\boldsymbol{x} = \boldsymbol{a} + i\boldsymbol{b}$, where $\boldsymbol{a} = \mathrm{Re}(\boldsymbol{x})$ and $\boldsymbol{b} = \mathrm{Im}(\boldsymbol{x})$. Defining

$$\hat{\mathbf{x}} = \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix} \in \mathbb{R}^{2p \times 1} \quad , \tag{B.25}$$

the covariance matrix $\boldsymbol{\Gamma} = E\{\hat{\mathbf{x}}\hat{\mathbf{x}}^{\mathrm{H}}\}$ can be partitioned into four $p \times p$ matrices

$$\boldsymbol{\Gamma} = \begin{bmatrix} \mathbf{R}_{\boldsymbol{aa}} & \mathbf{R}_{\boldsymbol{ab}} \\ \mathbf{R}_{\boldsymbol{ba}} & \mathbf{R}_{\boldsymbol{bb}} \end{bmatrix} \quad . \tag{B.26}$$

Now let $\mathcal{H}^+$ denote the set of $2p \times 2P$ positive definite matrices of "complex" from

$$\begin{bmatrix} \mathbf{K}_1 & -\mathbf{K}_2 \\ \mathbf{K}_2 & \mathbf{K}_1 \end{bmatrix} \tag{B.27}$$

where $\mathbf{K}_1$ is symmetric and $\mathbf{K}_2$ is anti-symmetric ($\mathbf{K}_2 = -\mathbf{K}_2^{\mathrm{T}}$), and therefore $\mathbf{K}_1 + i\mathbf{K}_2$ is Hermitian. Also let $\mathcal{R}$ denote the set of matrices

$$\begin{bmatrix} \mathbf{J}_1 & \mathbf{J}_2 \\ \mathbf{J}_2 & -\mathbf{J}_1 \end{bmatrix} \tag{B.28}$$

where $\mathbf{J}_1$ and $\mathbf{J}_2$ are symmetric. Then $\boldsymbol{\Gamma}$ can be written as

$$\boldsymbol{\Gamma} = \dot{\boldsymbol{\Gamma}} + \ddot{\boldsymbol{\Gamma}} \tag{B.29}$$

where

$$\dot{\mathbf{\Gamma}} = \frac{1}{2}\begin{bmatrix} \mathbf{R}_{aa} + \mathbf{R}_{bb} & \mathbf{R}_{ab} - \mathbf{R}_{ba} \\ \mathbf{R}_{ba} - \mathbf{R}_{ab} & \mathbf{R}_{bb} + \mathbf{R}_{aa} \end{bmatrix} \in \mathcal{H}^{+} \tag{B.30}$$

and

$$\ddot{\mathbf{\Gamma}} = \frac{1}{2}\begin{bmatrix} \mathbf{R}_{aa} - \mathbf{R}_{bb} & \mathbf{R}_{ab} + \mathbf{R}_{ba} \\ \mathbf{R}_{ba} + \mathbf{R}_{ab} & \mathbf{R}_{bb} - \mathbf{R}_{aa} \end{bmatrix} \in \mathcal{R} \quad . \tag{B.31}$$

Now

$$\mathbf{R}_{xx} = E\{\boldsymbol{x}\,\boldsymbol{x}^{\mathrm{H}}\} = (\mathbf{R}_{aa} + \mathbf{R}_{bb}) + i(\mathbf{R}_{ba} - \mathbf{R}_{ab}) \tag{B.32}$$

and

$$\tilde{\mathbf{R}}_{xx} = E\{\boldsymbol{x}\,\boldsymbol{x}^{\mathrm{T}}\} = (\mathbf{R}_{aa} - \mathbf{R}_{bb}) + i(\mathbf{R}_{ba} + \mathbf{R}_{ab}) \tag{B.33}$$

If $\tilde{\mathbf{R}}_{\mathbf{xx}} = \mathbf{0}$, then $\mathbf{R}_{aa} = \mathbf{R}_{bb} = \mathrm{Re}(\mathbf{R}_{\mathbf{xx}})/2$, and $\mathbf{R}_{ab} = -\mathbf{R}_{ba} = -\mathrm{Im}(\mathbf{R}_{\mathbf{xx}})/2$. So a test for propriety should be based on the hypothesis test $H_0 : \ddot{\mathbf{\Gamma}} = 0$ versus $H_1 : \ddot{\mathbf{\Gamma}} \neq 0$.

Test Statistics

Let $\hat{\mathbf{x}}_1, ..., \hat{\mathbf{x}}_n$ be n independent, normally distributed random vectors with zero mean and covariance matrix $\mathbf{\Gamma}$. The matrix $\mathbf{\Lambda} = n\mathbf{\Gamma}$ can be written as $\mathbf{\Lambda} = \dot{\mathbf{\Lambda}} + \ddot{\mathbf{\Lambda}}$ in the same way as equations (B.29)–(B.31). Furthermore, it can be represented in the form

$$\mathbf{\Lambda} = \boldsymbol{C}\boldsymbol{D}\boldsymbol{C}^{\mathrm{T}} \text{ with } \boldsymbol{D} = \begin{bmatrix} \boldsymbol{I}_p + \boldsymbol{L} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_p - \boldsymbol{L} \end{bmatrix} \tag{B.34}$$

where $\boldsymbol{L} = \mathrm{daig}(l_1, ..., l_p)$, $1 > l_1 \geq ... \geq l_p \geq 0$, and $\pm l_1, ..., \pm l_p$ are the eigenvalues of $\dot{\mathbf{\Lambda}}^{-1}\ddot{\mathbf{\Lambda}}$.

It can be shown that the hypothesis test can be restated as $H_0 : \det(\boldsymbol{D}) = 1$. A test follows as: accept H_0 iff

$$T_1(n, p) \equiv \det(D) = \prod_{k=1}^{p}(1 - l_k^2) \geq c_1 \tag{B.35}$$

where c_1 is a constant which determines the size, α, (false alarm probability) of the

test, such that $\Pr\left(\prod_{k=1}^{p}(1-l_k^2) < c_1|\mathbf{\Gamma}\right) = 1-\alpha$ A second test can be similarly derived: accept H_0 if

$$T_2(n,p) \equiv \frac{1}{2}\text{trace}(\dot{\mathbf{\Lambda}}^{-1}\ddot{\mathbf{\Lambda}}\dot{\mathbf{\Lambda}}^{-1}\ddot{\mathbf{\Lambda}}) = \sum_{k=1}^{p} l_k^2 \leq c_2 \tag{B.36}$$

where c_2 is a constant which determines the size (false alarm probability) of the test, such that $\Pr\left(\sum_{k=1}^{p} l_k^2 > c_2|\mathbf{\Gamma}\right) = 1-\alpha$.

Given a specified size of the test α, the intervals $[0, c_1]$ and $[c_2, p]$ need to be determined. The null hypothesis is rejected if the sample value of the test statistic falls within each interval, so we must find the critical values c_1 and c_2.

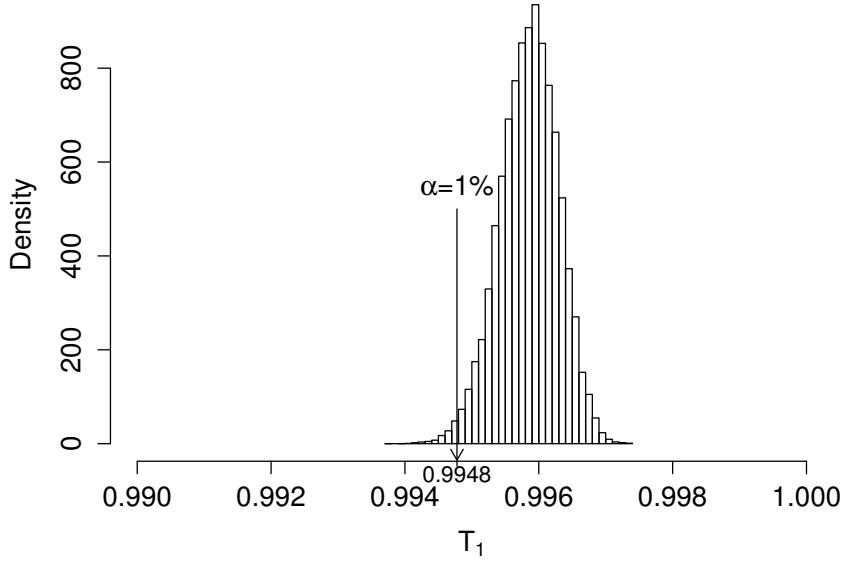
This is done numerically: for a given (n, p) combination, n independent $2p$ vectors are sampled from a normal distribution with zero mean and covariance matrix $\mathbf{\Gamma} = \boldsymbol{I}_{2p}$. Then, $\mathbf{\Delta}$, $\dot{\mathbf{\Delta}}$ and $\ddot{\mathbf{\Delta}} = \mathbf{\Delta} - \dot{\mathbf{\Delta}}$ are calculated, followed by the test statistics $T_1(n,p)$ and $T_2(n,p)$. This is repeated until smooth empirical cumulative distributions of T_1 and T_2 are available. From these the critical values of c_1 and c_2 can be "looked-up" for a given test size α.

Alternatively, a χ^2 approximation for the distribution of T_1 can be made, but is not detailed here. See [178, 215] for details.

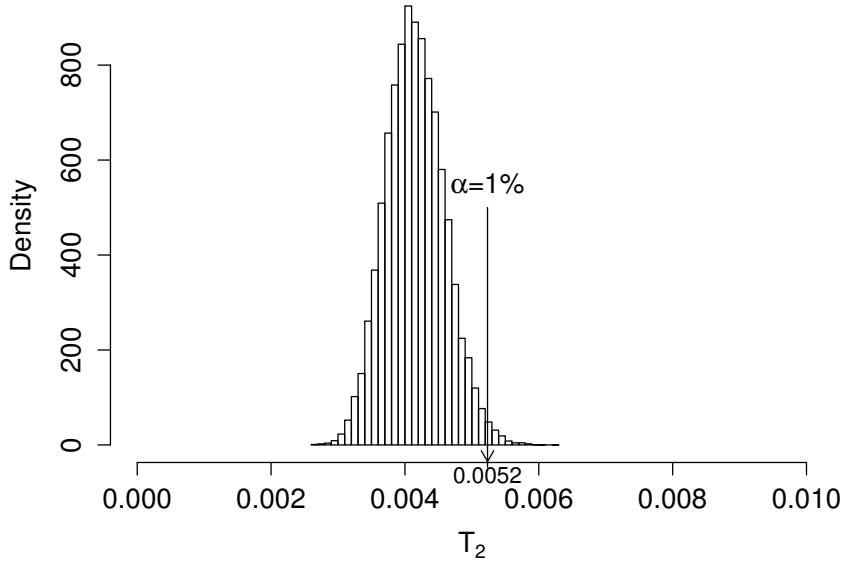
Example

The test has been applied to the MIDAS test data used in Chapter 3, Section 3.3.2. The training data comprise $n = 43\,824$ samples from $p = 13$ complex-valued time series. In order to produced smooth empirical distributions of T_1 and T_2 and determine the critical values of c_1 and c_2, the random sampling procedure described above is repeated 50 000 times. The resulting distributions are illustrated in Figure B.1 and the critical values for a range of α and listed in Table B.1.

The values of the test statistics from the MIDAS training dataset are $T_1 = 0.0982 \ll c_1$ and $T_2 = 1.78443 \gg c_2$, so H_0 is rejected for H_1 with negligible probability of a false positive.



(a) Distribution of T_1.



(b) Distribution of T_2.

Figure B.1: Distribution of impropriety test statistics T_1 and T_2 showing critical values for $\alpha = 1\%$ chance of false positive.

(1-α)%	90%	95%	99%
c_1	0.9953	0.9951	0.9948
c_2	0.0047	0.0049	0.0052

Table B.1: Critical values c_1 and c_2 of test statistics T_1 and T_2, respectively.

B.4 Maximum Likelihood Estimation of Constrained VAR Models

The coefficient matrices of a sparse vector autoregressive model sVAR(p,n), given by

$$\boldsymbol{y}_{t+k} = \sum_{\tau=1}^{p} \mathbf{A}_\tau \boldsymbol{y}_{t-\tau+1} + \boldsymbol{\epsilon}_{t+k} \quad , \tag{B.37}$$

$$\boldsymbol{y}_i, \boldsymbol{\epsilon}_i \in \mathbb{R}^M, \quad \mathbf{A}_\tau \in \mathbb{R}^{M\times M} \quad ,$$

can be expressed as

$$\boldsymbol{\alpha} = \mathrm{vec}\,(\mathbf{A}_1, ..., \mathbf{A}_p) = \mathbf{R}\boldsymbol{\gamma} \quad , \tag{B.38}$$

where $\alpha = \mathrm{vec}\,(\mathbf{A}_1, ..., \mathbf{A}_p)$ is the $pM^2 \times 1$ vector of coefficients obtained by column-stacking the matrices $\mathbf{A}_1, ..., \mathbf{A}_p$; $\mathbf{R}$ is a $pM^2 \times n$ matrix of known constants with rank n; and $\boldsymbol{\gamma}$ is an $n \times 1$ vector of unknown parameters. The matrix $\mathbf{R}$ is the *constrain matrix* and determines which VAR coefficients are zero and which are to be estimated, one entry in each column is 1, specifying a VAR coefficient, and all others are 0. The vector $\boldsymbol{\gamma}$ contains the VAR coefficients to be estimated that are mapped onto the matrices $\mathbf{A}_1, ..., \mathbf{A}_p$ by $\mathbf{R}$.

The maximum likelihood estimation of the VAR coefficients for a constrained model of the form in Equation (B.38) is given in by Lütkepohl in [86, Chapter 5]. The coefficients $\boldsymbol{\alpha}$ and noise covariance matrix $\boldsymbol{\Sigma}$ are the solutions to the following equations,

$$\hat{\boldsymbol{\alpha}} = \mathbf{R}\{\mathbf{R}^{\mathrm{T}}(\mathbf{L}\mathbf{L}^{\mathrm{T}} \otimes \hat{\boldsymbol{\Sigma}}^{-1})\mathbf{R}\}^{-1}\mathbf{R}^{\mathrm{T}}(\mathbf{L} \otimes \boldsymbol{\Sigma}^{-1})\mathcal{Y} \quad , \tag{B.39}$$

$$\boldsymbol{\Sigma} = \frac{1}{T-p} \sum_{t=p+1}^{T} (\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t)(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t)^{\mathrm{T}} \tag{B.40}$$

where $\otimes$ is the *Kronecker* product and

$$\mathbf{L}_t = (\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, ..., \boldsymbol{y}_{t-p+1})^{\mathrm{T}} \quad , \tag{B.41}$$

$$\mathbf{L} = (\mathbf{L}_0, \mathbf{L}_1, ..., \mathbf{L}_{T-1}) \quad , \tag{B.42}$$

$$\mathcal{Y} = \mathrm{vec}(\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_T) \quad , \tag{B.43}$$

$$\hat{\boldsymbol{y}}_t = \sum_{\tau=1}^{p} \hat{\mathbf{A}}_\tau \boldsymbol{y}_{t-\tau+1} \quad . \tag{B.44}$$

In the unconstrained case, i.e. $\mathbf{R} = \mathbb{I}_{pM^2}$, the maximum likelihood estimator for the VAR coefficients does not involve the noise covariance matrix $\boldsymbol{\Sigma}$. However, under the parameter constraints (B.38) the parameter estimates (B.39) are commingled with the estimate of the covariance matrix $\hat{\boldsymbol{\Sigma}}$. Therefore, the estimators $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\Sigma}}$ are updated iteratively until convergence to obtain the maximum likelihood estimate of the VAR coefficients.

Appendix C

Publications Arising from this Thesis

A number of publications have been produced during the development of this thesis:

Journal Papers

1. J. Dowell and P. Pinson, *"Very-Short-Term Wind Probabilistic Wind Power Forecasts by Sparse Vector Auto-regression"*, IEEE Transactions on Smart Grid, available online, 2015.

2. V. M. Catterson, D. McMillan, I. Dinwoodie, M. Revie, J. Dowell, J. Quigley, K. Wilson, *"An economic impact metric for evaluating wave height forecasters for offshore wind maintenance access,"* Wind Energy, available online, 2014.

3. J. Dowell, S. Weiss, D. Hill and D. Infield, *"Short-Term Spatio-Temporal Prediction of Wind Speed and Direction"*, Wind Energy, 17, pp. 1945–1955, 2013.

Book Chapters

1. R. Bessa, J. Dowell, P. Pinson, Chapter: Renewable Energy Forecasting, in *"International Handbook of Smart Grid Development,"* Wiley, 2015.

Conference Papers

1. J. Dowell, S. Weiss, D. Infield, *"Kernel Methods for Short-term Spatio-temporal Wind Prediction,"*, IEEE PES General Meeting, Denver, CO, 2015.

2. J. Dowell, S. Weiss and D. Infield, *"Spatio-Temporal Prediction of Wind Speed and Direction by Continuous Directional Regime"*, Probabilistic Methods Applied to Power Systems Conference, Durham, UK, July, 2014. *Best student paper award.

3. J. Dowell, S. Weiss, D. Infield and S. Chandna, *"A Widely Linear Wiener Filter for Wind Prediction"*, IEEE Statistical Signal Processing Workshop, Gold Coast, Australia, July, 2014.

4. J. Dowell and S. Weiss *"Short-term Wind Prediction Using an Ensemble of Particle Swarm Optimised FIR Filters"*, Intelligent Signal Processing Conference, London, UK, December, 2013.

5. J. Dowell, A. Zitrou, L. Walls, T. Bedford and D. Infield, *"Analysis of Wind and Wave Data to Assess Maintenance Access to Offshore Wind Farms"*, European Safety and Reliability Conference, Amsterdam, Netherlands, September, 2013.

6. J. Dowell, S. Weiss, D. Hill and D. Infield, *"A Cyclo-stationary Complex Multichannel Wiener Filter for the Prediction of Wind Speed and Direction"*, European Signal and Image Processing Conference, Marrakech, Morocco, September, 2013.

7. J. Dowell, S. Weiss, D. Hill and D. Infield, *"Improved Spatial Modelling of Wind Fields"*, EWEA Annual Event, Vienna, Austria, February, 2013.

Bibliography

[1] IPCC, *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, R. Pachauri and L. Meyer, Eds., Geniva, Switzerland, 2014.

[2] UK Meteorological Office, "MIDAS Land Surface Stations data (1853–current), NCAS British Atmospheric Data Centre, accessed December 2008." [Online]. Available: http://badc.nerc.ac.uk

[3] Royal Netherlands Meteorological Institute. (2005, October) Hydra potential wind data set. [Online]. Available: http://www.knmi.nl/samenw/hydra

[4] Australian Energy Market Operator, "AEMO 5 minute wind power data, 2011–2012." [Online]. Available: http://www.aemo.com.au/

[5] M. Shahidehpour, H. Yamin, and Z. Li, *Market Operations in Electric Power Systems: Forecasting, Scheduling, and Risk Management.* Wiley-IEEE Press, 2002.

[6] R. J. Barthelmie and S. C. Pryor, "Potential contribution of wind energy to climate change mitigation," *Nature Climate Change*, vol. 4, pp. 684–688, 2014.

[7] Working Group III of the Intergovernmental Panel on Climate Change, *IPCC Special Report on Renewable Energy Sources and Climate Change Mitigation*, O. Edenhofer, R. Pichs-Madruga, Y. Sokona, K. Seyboth, P. Matschoss, S. Kadner, T. Zwickel, P. Eickemeier, G. Hansen, S. Schlömer, and C. von Stechow, Eds. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, 2011.

[8] A. C. Marques, J. A. Fuinhas, and J. P. Manso, "Motivations driving renewable energy in european countries: A panel data approach," *Energy Policy*, vol. 38, no. 11, pp. 6877–6885, 2010, energy Efficiency Policies and Strategies with regular papers. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0301421510005252

[9] J. Morales, A. Conejo, H. Madsen, P. Pinson, and M. Zugno, *Integrating Renewable in Electricity Markets*. Springer, 2014.

[10] T. Ackermann, Ed., *Wind power in power systems*, 2nd ed. John Wiley & Sons: New York, 2012.

[11] R. Girard, K. Laquaine, and G. Kariniotakis, "Assessment of wind power predictability as a decision factor in the investment phase of wind farms," *Applied Energy*, vol. 101, no. 0, pp. 609–617, 2013.

[12] L. Jones, "Strategies and decision support systems for integrating variable energy resources in control centres for reliable grid operations," Alstom Grid Inc., Tech. Rep., 2011. [Online]. Available: http://www1.eere.energy.gov/wind/pdfs/reliable_grid_operations.pdf

[13] S. J. Watson, L. Landberg, and J. A. Halliday, "Application of wind speed forecasting to the integration of wind energy into a large scale power system," *IEE Proceedings—Generation, Transmission and Distribution*, vol. 141, no. 4, pp. 357–362, 1994.

[14] S. Faias, J. De Sousa, F. Reis, and R. Castro, "Assessment and optimization of wind energy integration into the power systems: Application to the portuguese system," *IEEE Transactions on Sustainable Energy*, vol. 3, no. 4, pp. 627–635, 2012.

[15] W. Mahoney, K. Parks, G. Wiener, Y. Liu, W. Myers, J. Sun, L. Delle Monache, T. Hopson, D. Johnson, and S. Haupt, "A wind power forecasting system to optimize grid integration," *Sustainable Energy, IEEE Transactions on*, vol. 3, no. 4, pp. 670–682, 2012.

[16] C. Lowery and M. O'Malley, "Impact of wind forecast error statistics upon unit commitment," *Sustainable Energy, IEEE Transactions on*, vol. 3, no. 4, pp. 760–768, 2012.

[17] P. Pinson, "Estimation of the uncertaintyin wind power forecasting," Ph.D. dissertation, L'Ecole des Mines de Paris, March 2006.

[18] E. McGarrigle and P. Leahy, "Quantifying the value of improved wind energy forecasts in a pool-based electricity market," *Renewable Energy*, vol. 80, pp. 517–524, 2015.

[19] E. Denny and M. O'Malley, "Wind generation, power system operation, and emissions reduction," *IEEE Transactions on Power Systems*, vol. 21, no. 1, pp. 341–347, 2006.

[20] A. Papavasiliou, S. Oren, and R. O'Neill, "Reserve requirements for wind power integration: A scenario-based stochastic programming framework," *Power Systems, IEEE Transactions on*, vol. 26, no. 4, pp. 2197–2206, Nov 2011.

[21] M. Matos and R. Bessa, "Setting the operating reserve using probabilistic wind power forecasts," *Power Systems, IEEE Transactions on*, vol. 26, no. 2, pp. 594–603, May 2011.

[22] R. Baldick, U. Helman, B. F. Hobbs, and R. P. ONeill, "Design of efficient generation markets," *Proceedings o the IEEE*, vol. 93, no. 11, pp. 1998–2012, November 2005.

[23] P. E. Morthorst, "Wind power and the conditions at a liberalized power market," *Wind Energy*, vol. 6, no. 3, pp. 297–308, 2003. [Online]. Available: http://dx.doi.org/10.1002/we.92

[24] R. Barthelmie, F. Murray, and S. Pryor, "The economic benefit of short-term forecasting for wind energy in the UK electricity market," *Energy Policy*, vol. 36, no. 5, pp. 1687–1696, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0301421508000323

[25] E. Bitar, R. Rajagopal, P. Khargonekar, K. Poolla, and P. Varaiya, "Bringing wind energy to market," *IEEE Transactions on Power Systems*, vol. 27, no. 3, pp. 1225–1235, 2012.

[26] T. Jónsson, P. Pinson, H. Nielsen, H. Madsen, and T. Nielsen, "Forecasting electricity spot prices accounting for wind power predictions," *Sustainable Energy, IEEE Transactions on*, vol. 4, no. 1, pp. 210–218, 2013.

[27] P. Pinson and M. O'Malley, "Foreword for the special section on wind and solar energy: Uncovering and accommodating their impact on electricity markets," *IEEE Transactions on Power Systems*, vol. 30, no. 3, pp. 1557–1559, May 2015.

[28] N. Zhang, C. Kang, Q. Xia, Y. Ding, Y. Huang, R. Sun, J. Huang, and J. Bai, "A convex model of risk-based unit commitment for day-ahead market clearing considering wind power uncertainty," *Power Systems, IEEE Transactions on*, vol. 30, no. 3, pp. 1582–1592, May 2015.

[29] L. Wendell, H. Wegley, and M. Verholek, "Report from a working group meeting on wind forecasts for WECS operation. pnl-2513," Pacific Northwest Laboratory, Tech. Rep., 1978.

[30] B. G. Brown, R. W. Katz, and A. H. Murphy, "Time series models to simulate and forecast wind speed and wind power," *J. Climate Appl. Meteorology*, vol. 23, pp. 1184–1195, 1984.

[31] L. Perry, D. Edwards, and K. Gray, "Wind energy development in california: Status report," California Energy Commission, Tech. Rep., April 1985.

[32] TPWind, "Strategic research agenda," European Wind Energy Technology Platform, Tech. Rep., 2008.

[33] C. Philibert, H. Holttinen, and H. Chandler, "Technology road map: Wind energy," International Energy Agency, Tech. Rep., 2013.

[34] T. Hong, P. Pinson, and S. Fan, "Global energy forecasting competition 2012," *International Journal of Forecasting*, vol. 30, pp. 357–363, 2014.

[35] G. Giebel, R. Brownsword, G. Kariniotakis, M. Denhard, and C. Draxl, *The State-of-the-Art in Short-Term Prediction of Wind Power*. ANEMOS.plus, 2011, project funded by the European Commission under the 6th Framework Program, Priority 6.1: Sustainable Energy Systems.

[36] C. Monteiro, R. Bessa, V. Miranda, A. Botterud, J. Wang, and G. Conzelmann, "Wind power forecasting: State-of-the-art 2009," Argonne National Laboratory ANL/DIS-10-1, Tech. Rep., 2009.

[37] A. M. Foley, P. G. Leahy, A. Marvuglia, and E. J. McKeogh, "Current methods and advances in forecasting of wind power generation," *Renewable Energy*, vol. 37, no. 1, pp. 1–8, 2012.

[38] X. Zhu and M. G. Genton, "Short-term wind speed forecasting for power system operations," *International Statistical Review*, vol. 80, no. 1, pp. 2–23, 2012. [Online]. Available: http://dx.doi.org/10.1111/j.1751-5823.2011.00168.x

[39] A. Costa, A. Crespo, J. Navarro, G. Lizcano, H. Madsen, and E. Feitosa, "A review on the young history of the wind power short-term prediction," *Renewable and Sustainable Energy Reviews*, vol. 12, no. 6, pp. 1725–1744, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364032107000354

[40] S. Burt, *The Weather Observer's Handbook*. Cambridge University Press, 2012.

[41] *Measurement Procedure for Cup Anemometer Calibrations*, MEASNET Std.

[42] *Power Performance Measurement Procedure, 1997*, MEASNET Std.

[43] *IEC 61400-12 - Wind Turbine Generation System - Part 12: Wind Turbine Performance Testing, 1998*, IEC Std.

[44] W. A. Weibull, "A statistical distribution function of wide applicability," *Journal of Applied Mechanics*, vol. 18, no. 3, pp. 293–297, 1951.

[45] T. Burton, N. Jenkins, D. Sharpe, and E. Bossanyi, *Wind Energy Handbook*, 2nd ed. Wiley, 2011.

[46] I. Troen and E. L. Petersen, "European wind atlas," Risø National Laboratory, Roskilde, Denmark, Tech. Rep., 1989, published for the Commission of the European Communities.

[47] R. Baïle, J. F. Muzy, and P. Poggi, "An M-Rice wind speed frequency distribution," *Wind Energy*, vol. 14, no. 6, pp. 735–748, 2011.

[48] E. Erdem and J. Shi, "Comparison of bivariate distribution construction approaches for analysing wind speed and direction data," *Wind Energy*, vol. 14, no. 1, pp. 27–41, 2011.

[49] B. Stephen, S. Galloway, D. McMillan, L. Anderson, and G. Ault, "Statistical profiling of site wind resource speed and directional characteristics," *Renewable Power Generation, IET*, vol. 7, no. 6, pp. 583–592, 2013.

[50] I. van der Hoven, "Power spectrum of horizontal wind speed in the frequency range from 0.0007 to 900 cycles per hour," *Journal of Meteorology*, vol. 14, pp. 160–164, 1957.

[51] R. V. Coquilla, "Review of anemometer calibration standards," in *In Proceedings of CANWEA 2009*. Otech Engineering, Inc., Davis, CA, 2009.

[52] P. Pinson, C. Chevallier, and G. Kariniotakis, "Trading wind generation from short-term probabilistic forecasts of wind power," *IEEE Transaction on Power Systems*, vol. 22, no. 3, pp. 1148–1156, 2007.

[53] Y. Zhang, J. Wang, and X. Wang, "Review on probabilistic forecasting of wind power generation," *Renewable and Sustainable Energy Reviews*, vol. 32, no. 0, pp. 255–270, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364032114000446

[54] T. Gneiting, "Editorial: probabilistic forecasting," *Journal of the Royal Statistical Society: Series A*, vol. 171, pp. 319–321, 2008.

[55] M. Ahlstrom, J. Blatchford, M. Davis, J. Duchesne, D. Edelson, U. Focken, D. Lew, C. Loutan, D. Maggio, M. Marquis, M. McMullen, K. Parks, K. Schuyler,

J. Sharp, and D. Souder, "Atmospheric pressure," *Power and Energy Magazine, IEEE*, vol. 9, no. 6, pp. 97–107, 2011.

[56] A. Suzuki, P. Shaw, C. Collier, J. Parkes, and L. Landberg, "Use of offsite data to improve short term ramp forecasting," in *Proceedings of EWEA 2013 Annual Conference*, 2013.

[57] D. R. Drew, D. J. Cannon, D. J. Brayshaw, J. F. Barlow, and P. J. Coker, "The impact of future offshore wind farms on wind power generation in great britain," *Resources*, vol. 4, no. 1, pp. 155–171, 2015.

[58] C. Graham, "The parameterisation and prediction of wave height and wind speed persistence statistics for oil industry operational planning purposes," *Coastal Engineering*, vol. 6, no. 4, pp. 303–329, 1982. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0378383982900059

[59] S. Kuwashima and N. Hogben, "The estimation of wave height and wind speed persistence statistics from cumulative probability distributions," *Coastal Engineering*, vol. 9, no. 6, pp. 563–590, 1986.

[60] P. Pinson and H. Madsen, "Ensemble-based probabilistic forecasting at horns rev," *Wind Energy*, vol. 12, no. 2, pp. 137–155, 2009.

[61] ——, "Adaptive modelling and forecasting of offshore wind power fluctuations with Markov-switching autoregressive models," *Journal of Forecasting*, vol. 31, no. 4, pp. 281–313, 2012. [Online]. Available: http://dx.doi.org/10.1002/for.1194

[62] K. Rogers, J. Collins, J. Parkes, and L. Landberg, "Wind power forecasting offshore, more or less accurate than onshore?" GL Harrad Hassam, Garrad Hassan and Partners Ltd St Vincent's Works Silverthorne Lane Bristol, UK BS2 0QD, Tech. Rep., 2012, avalialbe online at www.gl-garradhassan.com. [Online]. Available: http://www.gl-garradhassan.com/assets/downloads/Wind_Power_Forecasting_Offshore_-_More_

[63] J. Coiffier, *Fundamentals of Numerical Weather Prediction*. Cambridge University Press, 2011.

[64] P. Bauer, A. Thorpe, and G. Brunet, "The quiet revolution of numerical weather prediction," *Nature*, vol. 525, pp. 47–55, 2015.

[65] I. Troen and L. Landberg, "Short-term prediction of local wind conditions," in *Proceedings of the European Community Wind Energy Conference*, Madrid, Spain, September 1990, pp. 76–78.

[66] L. Landberg, "Short-term prediction of local wind conditions," Ph.D. dissertation, RisøNational Laboratory, Roskilde, Denmark, 1994.

[67] [Online]. Available: http://www.enfor.eu/

[68] B. Ernst, K. Rohrig, H. Regber, and P. Schorn, "Managing 3000 mw wind power in a transmission system operation center," in *Proceedings of the European Wind Energy Conference*, Copenhagen, Denmark, 2–6 June 2001, pp. 890–893, iSBN 3-936338-09-4.

[69] B. Lange, K. Rohrig, F. Schlögl, Ü. Cali, R. Mackensen, and L. Adzic, "Lessons learnt from the development of wind power forecast systems for six european transmission system operators," in *Proceedings of the EWEC*, Brussels,Belgium, 30 March–3 April 2008, (only abstract available).

[70] G. Gow, "Short term wind forecasting in the uk," in *Proceedings of the First IEA Joint Action Symposium on Wind Forecasting Techniques.* Norrköping, Sweden: FOI — Swedish Defence Research Agency., December 2002, pp. 3–10.

[71] T. N. Palmer, "The economic value of ensemble forecasts as a tool for risk assessment: From days to decades," *Quarterly Journal of the Royal Meteorological Society*, vol. 128, no. 581, pp. 747–774, 2002. [Online]. Available: http://dx.doi.org/10.1256/0035900021643593

[72] T. Gneiting and A. E. Raftery, "Weather forecasting with ensemble methods," *Science*, vol. 310, pp. 248–249, 2005.

[73] A. Möller, A. Lenkowski, and T. L. Thorarinsdottir, "Multivariate probabilistic forecasting using ensemble bayesian model averaging and copulas," *Quarterly Journal of the Royal Meteorological Society*, vol. 139, no. 673, pp. 982–991, 2013.

[74] C. Möhrlen and J. Jørgensen, "Verification of ensemble prediction systems for a new market: Wind energy," ECMWF Special Project Interim Reports 1–4, Tech. Rep., 2004.

[75] S. Lang, C. Möhrlen, J. Jørgensen, B. Ó. Gallachóir, and E. McKeogh, "Application of a multi-scheme ensemble prediction system for wind power forecating in ireland and comparison with validation results from denmark and germany," in *Proceedings of the European Wind Energy Conference and Exhibition EWEC*, Athens, Greece, 27 February–2 March 2006.

[76] ——, "Aggregate forecasting of wind power generation on the irish grid using a multi-scheme ensemble prediction system," in *Proceedings, Renewable Energy in Maritime Island Climates, 2nd Conference*, Dublin, Ireland, 26–28 2006.

[77] ——, "Forecasting total wind power generation on the republic of ireland grid with a multi-scheme ensemble prediction system," in *Proc. Global Wind Energy Conference GWEC*, Adelaide, Australia, 2006.

[78] G. Galanis, P. Louka, P. Katsafados, G. Kallos, and I. Pytharoulis, "Applications of Kalman filters based on non-linear functions to numerical weather predictions," *Ann. Geophys.*, vol. 24, pp. 1–10, 2006.

[79] T. Howard and P. Clark, "Correction and downscaling of NWP wind speed forecasts," *Meteorological Applications*, vol. 14, pp. 105–116, 2007.

[80] M. Khalid and A. V. Savkin, "A method for short-term wind power prediction with multiple observation points," *IEEE Transactions on Power Systems*, vol. 27, no. 2, pp. 579–586, May 2012.

[81] D. Lee and R. Baldick, "Short-term wind power ensemble prediction based on gaussian processes and neural networks," *Smart Grid, IEEE Transactions on*, vol. 5, no. 1, pp. 501–510, Jan 2014.

[82] J. M. Sloughter, T. Gneiting, and A. E. Raftery, "Probabilistic wind speed forecasting using ensembles and bayesian model averaging," *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 25–35, 2010.

[83] P. Pinson, "Adaptive calibration of (u,v)-wind ensemble forecasts," *Quarterly Journal of the Royal Meteorological Society*, vol. 138, no. 666, pp. 1273–1284, 2012.

[84] S. Alessandrini, S. Sperati, and P. Pinson, "A comparison between the ecmwf and cosmo ensemble prediction systems applied to short-term wind power forecasting on real data," *Applied Energy*, vol. 107, no. 0, pp. 271–280, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306261913001499

[85] G. R. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, revised ed., E. Robinson, Ed. Oakland, California: Holden Day, 1976.

[86] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. Heidelberg: Springer-Verlag, 2005.

[87] P. Young, *Recursive Estimation and Time-Series Analysis*, 2nd ed. Springer, 2011.

[88] M. Sheman, *Spatial Statistical and Spatio-Temporal Data Covariance Functions and Directional Properties*. Wiely, 2011.

[89] Y. Makarov, D. Hawkins, E. Leuze, and J. Vidov, "California ISO wind generation forecasting service design and experience," California Independent System Operator Corporation, Tech. Rep., 2002.

[90] A. Balouktsis, D. Tsanakas, and G. Vachtsevanos, "Stochastic simulation of hourly and daily average wind speed sequences," *Wind Engineering*, vol. 10, no. 1, pp. 1–11, 1986.

[91] A. Daniel and A. Chen, "Stochastic simulation and forecasting of hourly average wind speed sequences in jamaica," *Solar Energy*, vol. 46, no. 1, pp. 1–11, 1991.

[92] D. C. Hill, D. McMillan, K. R. W. Bell, and D. Infield, "Application of auto-regressive models to UK wind speed data for power system impact studies," *IEEE Transactions on Sustainable Energy*, vol. 3, no. 1, pp. 134–141, 2012.

[93] T. H. M. El-Fouly, E. F. El-Saadany, and M. A. Salama, "One day ahead prediction of wind speed and direction," *IEEE Transactions on Energy Conversion*, vol. 23, no. 1, pp. 191–201, 2008.

[94] A. Fellows and D. Hill, "Wind and load forecasting for integration of wind power into a meso-scale electrical grid," in *Proceedings of the European Community Wind Energy Conference*, Madrid, Spain, September 1990, pp. 636–640.

[95] C. Tantareanu, "Wind prediction in short term: A first step for a better wind turbine control," *Nordvestjysk Folkecenter for Vedvarende Energi*, 1992, iSBN 87-7778-005-1.

[96] Z. Huang and Z. Chalabi, "Use of time-series analysis to model and forecast wind speed," *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 56, pp. 311–322, 1994.

[97] L. Kamal and Y. Jafri, "Time series models to simulate and forecast hourly averaged wind speed in quetta, pakistan," *Solar Energy*, vol. 61, no. 1, pp. 23–32, 1997.

[98] J. Torres, M. D. B. A. Garcia, and A. D. Francisco, "ARMA models in navarre (spain)," *Solar Energy*, vol. 79, no. 1, pp. 65–77, 2005.

[99] M. Schwartz and M. Milligan, "Statistical wind forecasting at the US national renewable energy laboratory," in *First IEA Joint Action Symposium on Wind Forecasting Techniques*. Norrkping, Sweden: Published by FOI - Swedish Defence Research Agency., December 2002, pp. 115–124B.

[100] R. Karki, S. Thapa, and R. B. Billinton, "A simplified risk-based method for short-term wind power commitment," *IEEE Transactions on Sustainable Energy*, vol. 3, no. 3, pp. 498–505, July 2012. [Online]. Available: http://dx.doi.org/10.1109/TSTE.2012.2190999

[101] J. C. Palomares-Salas, J. J. G. de la Rosa, J. G. Ramiro, J. Melgar, A. Aguera, and A. Moreno, "ARIMA vs. neural networks for wind speed forecasting," in *Computational Intelligence for Measurement Systems and Applications, 2009. CIMSA '09. IEEE International Conference on*, 2009, pp. 129–133.

[102] R. G. Kavasseri and K. Seetharaman, "Day-ahead wind speed forecasting using f-arima models," *Renewable Energy*, vol. 34, no. 5, pp. 1388 – 1393, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0960148108003327

[103] P. Pinson, L. E. A. Christensen, H. Madsen, P. E. Sørensen, M. H. Donovan, and L. E. Jensen, "Regime-switching modelling of the fluctuations of offshore wind generation," *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 96, no. 12, pp. 2327–2347, 2008.

[104] U. Firat, S. N. Engin, M. Saraclar, and A. B. Ertuzun, "Wind speed forecasting based on second order blind identification and autoregressive model," in *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on*, 2010, pp. 686–691.

[105] K. Hunt and G. Nason, "Wind speed modelling and short-term prediction using wavelets," *Wind Engineering*, vol. 25, no. 1, pp. 55–61, 2001.

[106] C. Lei and L. Ran, "Short-term wind speed forecasting model for wind farm based on wavelet decomposition," in *Electric Utility Deregulation and Restructuring and Power Technologies, 2008. DRPT 2008. Third International Conference on*, 2008, pp. 2525–2529.

[107] A. A. Khan and M. Shahidehpour, "One day ahead wind speed forecasting using wavelets," in *Power Systems Conference and Exposition, 2009. PSCE '09. IEEE/PES*, 2009, pp. 1–5.

[108] L. Ling-ling, J.-H. Li, P.-J. He, and C.-S. Wang, "The use of wavelet theory and ARMA model in wind speed prediction," in *Electric Power Equipment — Switching Technology (ICEPE-ST), 2011 1st International Conference on*, 2011, pp. 395–398.

[109] J.-L. Tong, Z.-B. Zhao, and W.-Y. Zhang, "A new strategy for wind speed forecasting based on autoregression and wavelet transform," in *Remote Sensing, Environment and Transportation Engineering (RSETE), 2012 2nd International Conference on*, 2012, pp. 1–4.

[110] Y. Wang, S. Wang, and N. Zhang, "A novel wind speed forecasting method based on ensemble empirical mode decomposition and GA-BP neural network," in *Proceedings of the IEEE Power & Energy Society General Meeting*, Vancouver, July 2013.

[111] C. Took and D. P. Mandic, "The quaternion LMS algorithm for adaptive filtering of hypercomplex processes," *IEEE Transactions on Signal Processing*, vol. 57, no. 4, pp. 1316–1327, Apr 2009. [Online]. Available: http://dx.doi.org/10.1109/TSP.2008.2010600

[112] C. Jahanchahi, C. Took, and D. Mandic, "On hr calculus, quaternion valued stochastic gradient, and adaptive three dimensional wind forecasting," in *Neural Networks (IJCNN), The 2010 International Joint Conference on*, Barcelona, 2010, pp. 1–5.

[113] C. Took, G. Strbac, K. Aihara, and D. Mandic, "Quaternion-valued short-term joint forecasting of three-dimensional wind and atmospheric parameters," *Renewable Energy*, vol. 36, no. 6, pp. 1754–1760, 2011.

[114] S. Gill, B. Stephen, and S. Galloway, "Wind turbine condition assessment through power curve copula modelling," *IEEE Transactions on Sustainable Energy*, vol. 3, no. 1, pp. 94–101, 2012.

[115] L. Ljung and T. Söderström, *Theory of Recursive Identification*. MIT Press, 1983.

[116] H. Madsen, *Time Series Analysis*. Chapman & Hall, 2007.

[117] J. Jeon and J. W. Taylor, "Using conditional kernel density estimation for wind power density forecasting," *Journal of the American Statistical Association*, vol. 107, no. 497, pp. 66–79, 2012.

[118] J. W. Taylor and J. Jeon, "Forecasting wind power quantiles using conditional kernel estimation," *Renewable Energy*, vol. 80, no. 0, pp. 370–379, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0960148115001123

[119] T. S. Nielsen, H. A. Nielsen, and H. Madsen, "Prediction of wind power using time-varying coefficient functions," in *Proceedings of the XV IFAC World Congress*, 2002.

[120] G. Sideratos and N. Hatziargyriou, "Wind power forecasting focused on extreme power system events," *Sustainable Energy, IEEE Transactions on*, vol. 3, no. 3, pp. 445–454, 2012.

[121] T. Gneiting, K. Larson, K. Westrick, M. G. Genton, and E. Aldrich, "Calibrated probabilistic forecasting at the stateline wind energy center," *Journal of the American Statistical Association*, vol. 101, no. 475, pp. 968–979, 2006.

[122] A. S. Hering and M. G. Genton, "Powering up with space-time wind forecasting," *Journal of American Statistical Association*, vol. 105, pp. 92–104, 2010.

[123] J. Tastu, P. Pinson, E. Kotwa, H. Madsen, and H. A. Nielsen, "Spatio-temporal analysis and modeling of short-term wind power forecast errors," *Wind Energy*, vol. 14, no. 1, pp. 43–60, 2011.

[124] G. Reikard, "Regime-switching models and multiple causal factors in forecasting wind speed," *Wind Energy*, vol. 13, no. 5, pp. 407–418, 2010. [Online]. Available: http://dx.doi.org/10.1002/we.361

[125] J. Fan and W. Zhang, "Statistical methods with varying coefficient models," *Stat Interface*, vol. 1, no. 1, pp. 179–195, 2008.

[126] P. Ailliot, V. Monbet, and M. Prevosto, "An autoregressive model with time-varying coefficients for wind fields," *Environmetrics*, vol. 17, no. 2, pp. 107–117, 2006. [Online]. Available: http://dx.doi.org/10.1002/env.753

[127] L. Lin, J. Eriksson, H. Vihriälä, and L. Söderlund, "Predicting wind behaviour with neural networks," in *Proceedings of the EUWEC*, Göteborg (SE), 1996, pp. 655–658.

[128] M. C. Alexiadis, P. S. Dokopoulos, and H. S. Sahsamanoglou, "Wind speed and power forecasting based on spatial correlation models," *IEEE Transactions on Energy Conversion*, vol. 14, no. 3, pp. 836–842, 1999.

[129] A. Togelou, G. Sideratos, and N. Hatziargyriou, "Wind power forecasting in the absence of historical data," *Sustainable Energy, IEEE Transactions on*, vol. 3, no. 3, pp. 416–421, 2012.

[130] G. Sideratos and N. Hatziargyriou, "Probabilistic wind power forecasting using radial basis function neural networks," *Power Systems, IEEE Transactions on*, vol. 27, no. 4, pp. 1788–1796, 2012.

[131] D. Mandic, S. L. Goh, and K. Aihara, "Sequential data fusion via vector spaces: Complex modular neural network approach," in *Machine Learning for Signal Processing, 2005 IEEE Workshop on*, Mystic, CT, 2005, pp. 147–151.

[132] S. L. Goh, M. Chen, D. H. Popović, K. Aihara, D. Obradovic, and D. P. Mandic, "Complex-valued forecasting of wind profile," *Renewable Energy*, vol. 31, no. 11, pp. 1733–1750, 2006.

[133] S. Li, D. Wunsch, E. O'Hair, and M. Giesselmann, "Using neural networks to estimate wind turbine power generation," *Energy Conversion, IEEE Transactions on*, vol. 16, no. 3, pp. 276–282, 2001.

[134] T. Kitajima and T. Yasuno, "Output prediction of wind power generation system using complex-valued neural network," in *SICE Annual Conference, Proceedings of*, Taipei, 2010, pp. 3610–3613.

[135] K. Subramanian, R. Savitha, and S. Suresh, "Complex-valued neuro-fuzzy inference system for wind prediction," in *Neural Networks (IJCNN), The 2012 International Joint Conference on*, Brisbane, QLD, 2012, pp. 1–7.

[136] C. Wan, Z. Xu, P. Pinson, Z. Y. Dong, and K. P. Wong, "Probabilistic forecasting of wind power generation using extreme learning machine," *Power Systems, IEEE Transactions on*, vol. 29, no. 3, pp. 1033–1044, May 2014.

[137] T. Gautama, D. Mandic, and M. Hulle, "A non-parametric test for detecting the complex-valued nature of time series," *Int J Knowledge-based Intell Eng Systems*, vol. 8, no. 2, pp. 99–106, 2004.

[138] I. G. Damousis, M. C. Alexiadis, J. B. Theocharis, and P. S. Dokopoulos, "A fuzzy model forwind speed prediction and power generation in wind parks using spatial correlation," *IEEE Transactions on Energy Conversion*, vol. 19, no. 2, pp. 352–361, 2004.

[139] G. Zhang, H. Li, and M. Gan, "Design a wind speed prediction model using probabilistic fuzzy system," *IEEE Transactions on Industrial Informatics*, vol. 8, no. 4, pp. 819–827, November 2012.

[140] M. Yoder, A. S. Hering, W. C. Navidi, and K. Larson, "Short-term forecasting of categorical changes in wind power with markov chain models," *Wind Energy*, pp. n/a–n/a, 2013. [Online]. Available: http://dx.doi.org/10.1002/we.1641

[141] A. Kusiak, H. Zheng, and Z. Song, "Short-term prediction of wind farm power: A data mining approach," *Energy Conversion, IEEE Transactions on*, vol. 24, no. 1, pp. 125–136, 2009.

[142] J. Catalão, H. M. I. Pousinho, and V. Mendes, "Hybrid wavelet-PSO-ANFIS approach for short-term wind power forecasting in portugal," *Sustainable Energy, IEEE Transactions on*, vol. 2, no. 1, pp. 50–59, 2011.

[143] Y. Liu, J. Shi, Y. Yang, and W.-J. Lee, "Short-term wind-power prediction based on wavelet transform, support vector machine and statistic-characteristics analysis," *IEEE Transactions on Industry Applications*, vol. 48, no. 4, pp. 1136–1141, 2012.

[144] L. Xie, Y. Gu, X. Zhu, and M. Genton, "Short-term spatio-temporal wind power forecast in robust look-ahead power system dispatch," *Smart Grid, IEEE Transactions on*, vol. 5, no. 1, pp. 511–520, Jan 2014.

[145] B. M. Sanandaji, A. Tascikaraoglu, K. Poolla, and P. Varaiya, "Low-dimensional models in spatio-temporal wind speed forecasting," in *American Control Conference*, Chicago, IL, July 2015.

[146] G. Papaefthymiou and P. Pinson, "Modeling of spatial dependence in wind power forecast uncertainty," in *Probabilistic Methods Applied to Power Systems*, 2008.

[147] J. Tastu, P. Pinson, and H. Madsen, *Lecture Notes in Statistics: Modeling and Stochastic Learning for Forecasting in High Dimension.* Springer, 2014, ch. Space-time trajectories of wind power generation: Parameterized precision matrices under a Gaussian copula approach, to appear.

[148] J. Tastu, P. Pinson, P.-J. Trombe, and H. Madsen, "Probabilistic forecasts of wind power generation accounting for geographically dispersed information," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 480–489, 2014.

[149] M. Wytock and J. Zico Kolter, "Large-scale probabilistic forecasting in energy systems using sparse gaussian conditional random fields," in *Proceedings of the IEEE Conference on Decision and Control*, 2013.

[150] ——, "Sparse gaussian conditional random fields: Algorithms, theory, and application to energy forecasting," in *Proceedings of the International Conference on Machine Learning*, 2013.

[151] R. J. Bessa, V. Miranda, A. Botterud, and J. Wang, "'good' or 'bad' wind power forecasts: a relative concept," *Wind Energy*, vol. 14, no. 5, pp. 625–636, 2011.

[152] T. Gneiting, F. Balabdaoui, and A. E. Raftery, "Probabilistic forecasts, calibration and sharpness," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, pp. 243–268, 2007.

[153] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.

[154] P. Pinson, P. McSharry, and H. Madsen, "Reliability diagrams for nonparametric density forecasts of continuous variables: accounting for serial correlation," *Quarterly Journal of the Royal Meteorological Society*, vol. 136, no. 646, pp. 77–90, 2010.

[155] J. Brøcker and L. A. Smith, "Increasing the reliability of reliability diagrams," *Weather Forecasting*, vol. 22, no. 3, pp. 651–661, 2007.

[156] J. Dowell, S. Weiss, D. Hill, and D. Infield, "A cyclo-stationary complex multichannel Wiener filter for the prediction of wind speed and direction," in *Proceedings of the European Signal and Image Processing Conference (EUSIPCO)*, Mareakech, Morocco, 2013.

[157] ——, "Short-term spatio-temporal prediction of wind speed and direction," *Wind Energy*, vol. 17, no. 12, pp. 1945–1955, 2014.

[158] J. Dowell, S. Weiss, D. Infield, and S. Chandna, "A widely linear wiener filter for wind prediction," in *IEEE Statistical Signal Processing Workshop*, Gold Coast, Australia, 2014.

[159] J. Dowell, S. Weiss, and D. Infield, "Spatio-temporal prediction of wind speed and direction by continuous directional regime," in *Probabilistic Methods Applied to Powers Systems Conference*, Durham, UK, 2014.

[160] J. Dowell and S. Weiss, "Short-term wind prediction using an ensemble of particle swarm optimised FIR filters," in *Proceeding of the Intelligent Signal Processing Conference*, London, UK, 2–3 December 2013.

[161] J. Dowell, S. Weiss, and D. Infield, "Kernel methods for short-term spatio-temporal wind prediction," in *IEEE PES General Meeting*, Denver, CO, 2015.

[162] J. Dowell and P. Pinson, "Very-short-term probabilistic wind power forecasts by sparse vector autoregression," *IEEE Transactions on Smart Grid*, 2015, submitted.

[163] P. Sørensen, N. A. Cutululis, A. Vigueras-Rodriguez, L. E. Jensen, J. Hjerrild, M. H. Donovan, and H. Madsen, "Power fluctuations from large wind farms," *IEEE Transactions on Power Systems*, vol. 22, no. 3, pp. 958–965, 2007.

[164] D. Mandic and S. L. Goh, *Complex Valued Nonlinear Adaptive Filters: Noncircularity, Widely Linear and Neural Modes.* John Wiley & Sons, 2009.

[165] P. J. Schreier and L. L. Scharf, *Statistical Signal Processing of Complex-Valued Data: The Theory of Improper and Noncircular Signals.* Cambridge University Press, 2010.

[166] W. A. Gardner, A. Napolitano, and L. Paura, "Cyclostationarity: half a century of research," *Signal Processing*, vol. 86, pp. 639–697, 2006.

[167] A. Ispas, M. Dörpinghaus, G. Ascheid, and T. Zemen, "Characterization of non-stationary channels using mismatched Wiener filtering," *IEEE Transactions on Signal Processing*, vol. 61, no. 2, pp. 274–288, 2013.

[168] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed. Baltimore: The John Hopkins University Press, 1996.

[169] A. Hjørungnes and D. Gesbert, "Complex-valued matrix differentiation: Techniques and key results," *IEEE Transactions on Signal Processing*, vol. 55, no. 6, part I, pp. 2740–2746, June 2007.

[170] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*. Englewood Cliffs, New York: Prentice Hall, 1985.

[171] S. Haykin, *Adaptive Filter Theory*, 4th ed. Prentice Hall, 2002.

[172] B. Widrow, J. McCool, and M. Ball, "The complex LMS algorithm," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 719–720, 1975.

[173] B. Widrow, J. McCool, M. Larimore, and C. Johnson, "Stationary and nonstationary learning characteristics of the LMS adaptive filter," *Proceedings of the IEEE*, vol. 64, no. 8, pp. 1151–1162, aug. 1976.

[174] N. J. Bershad and O. M. Macchi, "Adaptive Recovery of a Chirped Sinusoid in Noise, Part 2: Performance of the LMS Algorithm," *IEEE Transaction on Signal Processing*, vol. 39, no. 3, pp. 595–602, March 1991.

[175] O. M. Macchi and N. J. Bershad, "Adaptive Recovery of a Chirped Sinusoid in Noise, Part 1: Performance of the RLS Algorithm," *IEEE Transaction on Signal Processing*, vol. 39, no. 3, pp. 583–594, March 1991.

[176] T. Nielsen, A. Joensen, H. Madsen, L. Landberg, and G. Giebel, "A new reference model for wind power forecasting," *Wind Energy*, vol. 1, pp. 29–36, 1998.

[177] B. Picinbono and P. Chevalier, "Widely linear estimation with complex data," *Signal Processing, IEEE Transactions on*, vol. 43, no. 8, pp. 2030–2033, 1995.

[178] A. Walden and P. Rubin-Delanchy, "On testing for impropriety of complex-valued Gaussian vectors," *IEEE Transactions on Signal Processing*, vol. 57, no. 3, pp. 825–834, March 2009.

[179] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Neural Networks, 1995. Proceedings., IEEE International Conference on*, vol. 4, 1995, pp. 1942–1948.

[180] Y. Shi and R. Eberhart, "A modified particle swarm optimizer," in *Proceedings of IEEE International Conference on Evolutionary Computation*, 1998, pp. 69–73.

[181] W. Xiao-lu, L. Jian, and L. Jian-jun, "Wavelet transform and pso support vector machine based approach for time series forecasting," in *Artificial Intelligence and Computational Intelligence, 2009. AICI '09. International Conference on*, vol. 1, 2009, pp. 46–50.

[182] A. Y. Alanis, C. Simetti, L. J. Ricalde, and F. Odone, "A wind speed neural model with particle swarm optimization kalman learning," in *World Automation Congress (WAC), 2012*, 2012, pp. 1–5.

[183] R. Eberhart and Y. Shi, "Particle swarm optimization: developments, applications and resources," in *Evolutionary Computation, 2001. Proceedings of the 2001 Congress on*, vol. 1, 2001, pp. 81–86 vol. 1.

[184] R. Kiran, S. Jetti, and G. Venayagamoorthy, "Online training of a generalized neuron with particle swarm optimization," in *Neural Networks, 2006. IJCNN '06. International Joint Conference on*, 2006, pp. 5088–5095.

[185] J. Dowell, S. Weiss, D. Hill, and D. Infield, "Improved spatial modelling of wind fields," in *EWEA Annual Conference*, Vienna, Austria, September 2013.

[186] X. Hu, Y. Shi, and R. Eberhart, "Recent advances in particle swarm," in *Evolutionary Computation, 2004. CEC2004. Congress on*, vol. 1, 2004, pp. 90–97.

[187] J. Mercer, "Functions of positive and negative type, and their connection with the theory of integral equations," *Philosophical Transactions of the Royal Society of London, Series A*, vol. 209, no. 441–458, pp. 415–446, 1909. [Online]. Available: http://rsta.royalsocietypublishing.org/content/209/441-458/415.short

[188] B. Schølkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001.

[189] W. Liu, P. Pokharel, and J. Principe, "The kernel least-mean-square algorithm," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 543–554, Feb 2008.

[190] Y. Engel, S. Mannor, and R. Meir, "The kernel recursive least-squares algorithm," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2275–2285, Aug 2004.

[191] W. Liu, I. M. Park, Y. Wang, and J. Principe, "Extended kernel recursive least squares algorithm," *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 3801–3814, Oct 2009.

[192] S. Van Vaerenbergh, M. Lazaro-Gredilla, and I. Santamaria, "Kernel recursive least-squares tracker for time-varying regression," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 8, pp. 1313–1326, Aug 2012.

[193] F. A. Tobar, A. Kuh, and D. P. Mandic, "A novel augmented complex valued kernel LMS," in *7th Sensor Array and Multichannel Signal Processing Workshop*. Hoboken, NJ, USA: IEEE, June 2012.

[194] F. Tobar, S.-Y. Kung, and D. Mandic, "Multikernel least mean square algorithm," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 2, pp. 265–277, Feb 2014.

[195] A. Kuh, C. Manloloyo, R. Corpuz, and N. Kowahl, "Wind prediction using complex augmented adaptive filters," in *International Conference on Green Circuits and Systems*, June 2010, pp. 46–50.

[196] C. Liu and F. Liu, "The short-term load forecasting using the kernel recursive least-squares algorithm," in *3rd International Conference on Biomedical Engineering and Informatics*, vol. 7, Oct 2010, pp. 2673–2676.

[197] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *The Annals of Statistics*, vol. 36, no. 3, pp. 1171–1220, 2008.

[198] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[199] V. Akhmatov, "Influence of wind direction on intense power fluctuations in large offshore windfarms in the North Sea," *Wind Engineering*, vol. 31, no. 1, pp. 59–64, 2007.

[200] J. Kristoffersen and P. Christiansen, "Horns rev offshore wind farm: its main controller and remote control system," *Wind Engineering*, vol. 27, pp. 351–359, 2003.

[201] P. E. Sørensen, A. D. Hansen, K. Thomsen, T. Buhl, P. E. Morthorst, L. H. Nielsen, F. Iov, F. Blaabjerg, H. A. Nielsen, H. Madsen *et al.*, "Operation and control of large wind turbines and wind farms. final report," RisøNational Laboratory, Tech. Rep., 2005.

[202] A. Lau and P. McSharry, "Approaches for multi-step density forecasts with application to aggregated wind power," *The Annals of Applied Statistics*, vol. 4, no. 3, pp. 1311–1341, 2010.

[203] P. Pinson, "Very-short-term probabilistic forecasting of wind power with generalized logit-normal distributions," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, pp. 555–576, 2012.

[204] R. A. Davis, P. Zang, and T. Zheng, "Sparse vector autoregressive modelling," *arXiv:1207.0520*, 2012.

[205] Y. Zhang, J. Wang, and X. Luo, "Probabilistic wind power forecasting based on logarithmic transformation and boundary kernel," *Energy Conversion and Management*, vol. 96, no. 0, pp. 440–451, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0196890415002228

[206] R. Dahlhaus, "Graphical interaction models for multivariate time series," *Metrika*, vol. 51, pp. 157–172, 2000.

[207] http://www.jethrodowell.com.

[208] R. Bessa, A. Trindade, and V. Miranda, "Spatial-temporal solar power forecasting for smart grids," *Industrial Informatics, IEEE Transactions on*, vol. PP, pp. n/a–n/a, 2014.

[209] P. Pinson, H. A. Nielsen, H. Madsen, and T. S. Nielsen, "Local linear regression with adaptive orthogonal fitting for the wind power application," *Statist. Comput.*, vol. 18, pp. 59–71, 2008.

[210] T. McElroy and D. Findley, "Fitting constrained vector autoregression models," in *Empirical Economic and Financial Research*. Springer, 2015, pp. 451–470.

[211] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B*, vol. 58, pp. 267–288, 1996.

[212] T. K. Moon and W. C. Stirling, *Mathematical Methods and Algorithms for Singal Processing*. Prentice Hall, 2000.

[213] E. Ollila and V. Koivunen, "Generalized complex elliptical distributions," in *Sensor Array and Multichannel Signal Processing Workshop Proceedings*, July 2004, pp. 460–464.

[214] P. Schreier, L. Scharf, and A. Hanssen, "A statistical test for impropriety of complex random signals," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 3, May 2006, pp. III–III.

[215] G. E. P. Box, "A general distribution theory for a class of likelihood criteria," *Biometrika*, vol. 36, no. 3/4, pp. 317–46, 1949.

Bibliography