Probabilistic Wind Power Forecasts: From Aggregated Approach to Spatiotemporal Models

Ada Lau

Mathematical Institute
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

2011

Abstract

Wind power is one of the most promising renewable energy resources to replace conventional generation which carries high carbon footprints. Due to the abundance of wind and its relatively cheap installation costs, it is likely that wind power will become the most important energy resource in the near future. The successful development of wind power relies heavily on the ability to integrate wind power efficiently into electricity grids. To optimize the value of wind power through careful power dispatches, techniques in forecasting the level of wind power and the associated variability are critical. Ideally, one would like to obtain reliable probability density forecasts for the wind power distributions. As wind is intermittent and wind turbines have non-linear power curves, this is a challenging task and many ongoing studies relate to the topic of wind power forecasting. For this reason, this thesis aims at contributing to the literature on wind power forecasting by constructing and analyzing various time series models and spatiotemporal models for wind power production. By exploring the key features of a portfolio of wind power data from Ireland and Denmark, we investigate different types of appropriate models. For instance, we develop anisotropic spatiotemporal correlation models to account for the propagation of weather fronts. We also develop two-stage models to accommodate the probability masses that occur in wind power distributions due to chains of zeros. We apply the models to generate multi-step probability forecasts for both the individual and aggregated wind power using extensive data sets from Ireland and Denmark. From the evaluation of probability forecasts, valuable insights are obtained and deeper understanding of the strengths of various models could be applied to improve wind power forecasts in the future.

Probabilistic Wind Power Forecasts: From Aggregated Approach to Spatiotemporal Models

Ada Lau

Mathematical Institute

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

2011

Acknowledgements

I wish to thank my parents for providing all the best to me and supporting me in every moment of my life. Their encouragement helps me to accomplish my DPhil study at Oxford.

I thank Prof. Sam Howison for his advice and support during my DPhil study. I thank Prof. Tilmann Gneiting, Dr. Pierre Pinson, Dr. Max Little and Prof. James Taylor for their patience in reading some of my work, and for their insightful comments and feedback. I am grateful to Dr. Patrick McSharry for helping me to access valuable wind power data for my extensive research. The wind power data in Ireland is kindly provided by Eirgrid plc, and the wind power data in Denmark is obtained through SafeWind.

Last but not least, I would like to give thanks to the Oxford-Man Institute of Quantitative Finance, the China Oxford Scholarship Fund, and the Oxford and Cambridge Society of Hong Kong which have generously supported my DPhil research in Oxford. The Oxford-Man Institute of Quantitative Finance has offered me an inspiring and supportive environment to conduct my DPhil research.

Abstract

Wind power is one of the most promising renewable energy resources to replace conventional generation which carries high carbon footprints. Due to the abundance of wind and its relatively cheap installation costs, it is likely that wind power will become the most important energy resource in the near future. The successful development of wind power relies heavily on the ability to integrate wind power efficiently into electricity grids. To optimize the value of wind power through careful power dispatches, techniques in forecasting the level of wind power and the associated variability are critical. Ideally, one would like to obtain reliable probability density forecasts for the wind power distributions. As wind is intermittent and wind turbines have non-linear power curves, this is a challenging task and many ongoing studies relate to the topic of wind power forecasting. For this reason, this thesis aims at contributing to the literature on wind power forecasting by constructing and analyzing various time series models and spatiotemporal models for wind power production. By exploring the key features of a portfolio of wind power data from Ireland and Denmark, we investigate different types of appropriate models. For instance, we develop anisotropic spatiotemporal correlation models to account for the propagation of weather fronts. We also develop two-stage models to accommodate the probability masses that occur in wind power distributions due to chains of zeros. We apply the models to generate multi-step probability forecasts for both the individual and aggregated wind power using extensive data sets from Ireland and Denmark. From the evaluation of probability forecasts, valuable insights are obtained and deeper understanding of the strengths of various models could be applied to improve wind power forecasts in the future.

Contents

List of Figures x

List of Tables xiv

1 Introduction and Motivation 1

1.1 Background and literature review . . . 2
1.1.1 Various approaches of wind power forecasts . . . 3
1.1.2 Categories of wind power forecasts . . . 4
1.1.3 Horizons for wind power forecasts . . . 6
1.2 Major challenges in wind power forecasting . . . 7
1.2.1 Boundedness and skewness . . . 7
1.2.2 Discrete probability masses . . . 8
1.2.3 Non-stationarity and long memory . . . 8
1.3 Exploratory analysis of wind data . . . 9
1.3.1 Individual wind power at a single wind farm . . . 9
1.3.1.1 Distributions and autocorrelations . . . 9
1.3.1.2 Modified logit transformation . . . 11
1.3.2 Aggregated wind power . . . 14
1.3.2.1 Distributions and autocorrelations . . . 14
1.3.2.2 Seasonality . . . 16
1.3.2.3 Logit transformation . . . 18
1.4 Rationale in choosing the various models . . . 20
1.4.1 Gaussian models . . . 20
1.4.2 Kriging models . . . 20
1.4.3 Two-stage models . . . 21

1.5 Contributions in our study . . . 22
1.5.1 Multi-step probability forecasts . . . 22
1.5.2 Spatiotemporal correlation analysis and modeling . . . 23
1.5.3 Two-stage model for wind power . . . 24
1.5.4 Statement of originality . . . 24
1.6 Roadmap of this thesis . . . 25

2 Models for Aggregated Wind Power 26
2.1 Gaussian model for transformed data . . . 28
2.1.1 Logit transformation . . . 28
2.1.2 ARIMA-GARCH models . . . 28
2.1.3 Multi-step ahead forecasts . . . 30
2.1.4 Inverse transformation . . . 30
2.2 Exponential smoothing and parametric distributions . . . 31
2.2.1 Exponential smoothing . . . 32
2.2.1.1 Smoothing the location parameter only . . . 32
2.2.1.2 Smoothing both the location and scale parameters 35
2.2.2 Mapping to parametric distributions . . . 38
2.2.2.1 Truncated normal distribution . . . 40
2.2.3 Estimation of smoothing parameters . . . 44

3 Spatiotemporal Kriging and Correlation Models 45
3.1 Spatial kriging . . . 45
3.2 Spatiotemporal kriging . . . 46
3.2.1 Optimal linear predictor . . . 46
3.3 Correlation structures of wind power data . . . 49
3.3.1 Autocorrelation . . . 49
3.3.2 Spatial correlation . . . 50
3.3.3 Spatiotemporal correlation . . . 52
3.4 Covariance models . . . 56
3.4.1 Properties of covariance models . . . 56
3.4.1.1 Positive definiteness, stationarity and full-symmetry 56
3.4.1.2 Properties of wind power covariances . . . 57
3.4.1.3 Nugget effect . . . 58

3.4.2 Construction of correlation models 60
3.4.3 Construction of anisotropic correlation models 60
3.4.4 Calculation of distance on a sphere 62
3.5 Spatiotemporal correlation models 63
3.5.1 Examples of spatiotemporal correlation models 64
3.5.2 Our anisotropic spatiotemporal correlation model 67
3.5.2.1 Idea from Lagrangian approach 67
3.5.2.2 Features of anisotropic correlations 68
3.5.2.3 Model specification 72
3.5.2.4 Discussions and remarks 73
3.5.3 Estimation of correlation models 76
3.6 Generating spatiotemporal forecasts 77
3.6.1 Normalizing the wind power data 77
3.6.2 Kriging predictor 77
3.6.3 Homoscedastic or heteroscedastic variances 79
3.6.4 Generating valid forecasts in $[0, 1]$ 81

4 Two-stage Model with Latent Gaussian Processes 83
4.1 Introduction 83
4.1.1 Discrete-continuous model 84
4.1.2 Two-stage model 86
4.2 Two-stage model: Model specification 88
4.2.1 Model for Gaussian process $W(\mathbf{s}, t)$ 88
4.2.2 Model for the transformation $F(\cdot)$ 90
4.2.3 Model for Gaussian process $Z(\mathbf{s}, t)$ 92
4.3 Two-stage model: Parameter estimation 95
4.3.1 Regression coefficients for the dynamics of $W(\mathbf{s}, t)$ 95
4.3.2 Parameters in correlation model for $W(\mathbf{s}, t)$ 97
4.3.3 Smoothing parameters for the dynamics of $F(y|\ell, s^2)$ . . . 99
4.3.3.1 Minimizing CRPS 99
4.3.3.2 Rescaling the location and scale parameters . . . 100
4.3.4 Parameters in correlation model for $Z(\mathbf{s}, t)$ 102
4.3.5 Summary of two-stage model 106

4.4 Generating forecasts through simulations . . . 106
4.4.1 Drawing samples of $W(\mathbf{s}, t)$ . . . 107
4.4.2 Drawing samples of $Z(\mathbf{s}, t)$ . . . 108
4.4.2.1 Spatiotemporal correlation matrices . . . 108
4.4.2.2 Conditional mean and covariance . . . 109
4.4.3 Samples of aggregated forecasts . . . 112

5 Application to Irish Wind Data 113
5.1 Wind data . . . 113
5.1.1 Research questions using this data set . . . 115
5.1.2 Training and testing data . . . 116
5.2 Aggregated Forecasts . . . 117
5.2.1 Competing models . . . 118
5.2.2 Benchmark models . . . 118
5.2.3 Aggregated point forecasts . . . 121
5.2.4 Aggregated probability forecasts . . . 123
5.3 Forecasts using spatiotemporal kriging . . . 127
5.3.1 List of models . . . 127
5.3.2 Parameter estimation . . . 131
5.3.2.1 Modified logit transformation . . . 131
5.3.2.2 Parameters in correlation models . . . 133
5.3.3 Individual forecast evaluations . . . 135
5.3.3.1 Forecasts at all 64 wind farms . . . 135
5.3.3.2 Forecasts at a single wind farm . . . 141
5.3.3.3 Comparison of forecast performances at different locations . . . 149
5.3.4 Aggregated forecast evaluations . . . 151
5.3.4.1 Significance of forecast improvements . . . 156
5.4 Forecasts using two-stage model . . . 158
5.4.1 List of models . . . 159
5.4.2 Parameter estimation . . . 161
5.4.2.1 Regression coefficients for $W(\mathbf{s}, t)$ . . . 161
5.4.2.2 Correlation model for $W(\mathbf{s}, t)$ . . . 162

5.4.2.3 Smoothing parameters for $F(y|\ell, s^2)$ 163
5.4.2.4 Correlation model for $Z(\mathbf{s}, t)$ 163
5.4.3 Individual forecast evaluations 164
5.4.4 Aggregated forecast evaluations 168

6 Application to Danish Wind Data **172**
6.1 Wind data 172
6.1.1 Training and testing data 178
6.1.2 Research questions using this data set 178
6.2 Forecasts using spatiotemporal models 179
6.2.1 List of models 179
6.2.2 Individual forecasts evaluations 181
6.2.3 Aggregated forecasts evaluations 185
6.3 Comparison of forecast performances at different locations 188
6.4 Robustness Analysis 189
6.4.1 Methodology 190
6.4.2 Spatially out-of-sample forecast performances 191
6.5 Subsampling analysis 194
6.5.1 Methodology 194
6.5.2 Results using different subsampling approaches 197

7 Conclusion and Discussion **199**
7.1 Summary of contributions 199
7.1.1 Aggregated wind power forecasts 199
7.1.1.1 List of major results 201
7.1.2 Spatiotemporal wind power forecasts 202
7.1.2.1 Correlation modeling 202
7.1.2.2 Two-stage approach 202
7.1.2.3 List of major results 203
7.2 Comparing the Irish and Danish results 204
7.3 Discussions and future developments 205
7.3.1 Aggregated wind power forecasts 205
7.3.2 Spatiotemporal wind power forecasts 205
7.4 Conclusion 206

Appendices

A ARIMA-GARCH Processes **207**
A.1 ARMA process . . . 207
A.2 GARCH process . . . 208
A.3 Optimal multi-step forecasts . . . 210
A.3.1 Cost function . . . 210
A.3.2 Optimal forecast for ARMA process . . . 211
A.3.3 Optimal forecast for ARIMA process . . . 212
A.4 Variance of forecast errors . . . 214
A.4.1 Conditional homoscedastic white noise . . . 214
A.4.2 Conditional heteroscedastic white noise . . . 216

B Exponential Smoothing Methods **218**
B.1 Model for conditional mean . . . 218
B.2 Model for conditional mean and variance . . . 220
B.3 Optimal forecast and forecast variance . . . 223
B.3.1 Conditional homoscedastic white noise . . . 223
B.3.2 Conditional heteroscedastic white noise . . . 224

C Positive Definite Covariance Functions **226**

D Techniques of Forecast Evaluations **228**
D.1 Scoring rules . . . 229
D.2 Diagnostic diagrams . . . 232

E Remarks on Alternative Models **233**
E.1 Diurnal seasonality and seasonal exponential smoothing . . . 233
E.2 Exponential smoothing with logit transformation . . . 236
E.3 Exponential smoothing with sliding window . . . 238
E.4 Double kernel density benchmark . . . 240
E.5 ARIMA$(p, 1, q)$-GARCH(r, s) benchmark . . . 243

F Computational Remarks **245**

F.1 Computational time: parameter estimation 245

F.2 Computational time: samplings 245

References **246**

List of Figures

1.1 Number of turbines in 64 wind farms 5
1.2 Time series and densities of normalized individual wind power . . 10
1.3 Autocorrelations of individual wind power 11
1.4 Density of residuals with modified logit transformation 12
1.5 Time series and density of normalized aggregated wind power . . 14
1.6 Time series and autocorrelations of normalized aggregated wind power 15
1.7 Long term seasonality of aggregated wind power 16
1.8 Diurnal seasonality of aggregated wind power 17
1.9 Three transformations of normalized aggregated wind power . . . 19
1.10 Road Map of this thesis 25

2.1 First differences of logit transformed Irish wind power 29
2.2 Mean and variance of truncated normal distribution 42
2.3 Relation between the variance of wind power and the power curve 43

3.1 Autocorrelation of wind power 49
3.2 Spatial correlation of Irish wind power 51
3.3 Spatial correlation of Irish wind speed 51
3.4 Spatial correlation of Danish wind power 52
3.5 Wind power generation driven by wind farm no. 1 54
3.6 Cross correlations between Irish wind power 55
3.7 Wind power correlations between wind farms 1 and 55 70
3.8 Wind power correlations between wind farms 42 and 55 70

3.9 Difference between correlations at positive and negative temporal lag . . . 71
3.10 Temporal lags at which maximum correlation is attained. 71
3.11 Definition of azimuth angle 72
3.12 Shape of decay of correlation 75
3.13 Estimated parameters versus number of temporal lag 80
3.14 Forecast results with number of samples 82

4.1 Results with and without rescaling 103
4.2 Decay of autocorrelations of $z(\mathbf{s}, t+h|t)$ 105
4.3 Decay of cross correlations of $z(\mathbf{s}, t+h|t)$ 105
4.4 Flow diagram of sample drawings 111

5.1 Locations of wind farms in Ireland 115
5.2 Weights in EWMA empirical conditional probability forecasts . . 120
5.3 RMSE of aggregated point forecasts 122
5.4 Mean CRPS of aggregated probability forecasts 123
5.5 QQ-plot of PIT values 125
5.6 Estimated variance of data in the testing set 126
5.7 Histograms of PIT values conditioned on large variances 126
5.8 In-sample MSE of conditional forecasts 128
5.9 Median distance between $F(z)$ and $\Phi(z)$ 132
5.10 Modified logit transformation of wind power 132
5.11 Correlations fits of the models 134
5.12 RMSE of Individual point forecasts 136
5.13 Mean MQE versus quantile values 138
5.14 Mean MQE versus quantile values around 50% 138
5.15 Mean CRPS of Individual probability forecasts 140
5.16 Wind power time series at Anarget 141
5.17 RMSE at Anarget (Farm No. 30) 142
5.18 MQE at Anarget (Farm No. 30) 143
5.19 MQE around 50% at Anarget (Farm No. 30) 143
5.20 Mean CRPS at Anarget (Farm No. 30) 144
5.21 Wind power time series at Kilbranish 145

5.22 RMSE at Kilbranish (Farm No. 25) . . . 146
5.23 MQE at Kilbranish (Farm No. 25) . . . 147
5.24 MQE around 50% at Kilbranish (Farm No. 25) . . . 147
5.25 Mean CRPS at Kilbranish (Farm No. 25) . . . 148
5.26 Differences of RMSE between Model 12 and 13 . . . 150
5.27 RMSE of aggregated point forecasts . . . 152
5.28 MQE of aggregated quantile forecasts . . . 153
5.29 MQE of aggregated quantile forecasts around 50% . . . 153
5.30 Mean CRPS of aggregated probability forecasts . . . 155
5.31 QQ-plot of PIT values for one-step ahead aggregated probability forecasts . . . 155
5.32 Expected log likelihood of $W(\mathbf{s}, t)$ . . . 162
5.33 Box-plots of RMSE of individual point forecasts . . . 165
5.34 MQE of individual quantile forecasts . . . 166
5.35 MQE of individual quantile forecasts around 50% . . . 166
5.36 Box-plots of mean CRPS of individual probability forecasts . . . 167
5.37 RMSE of aggregated point forecasts . . . 169
5.38 MQE of aggregated quantile forecasts . . . 170
5.39 MQE of aggregated quantile forecasts around 50% . . . 170
5.40 Mean CRPS of aggregated probability forecasts . . . 171

6.1 Locations of 49 wind farms in Denmark . . . 173
6.2 Wind power variance across wind farms . . . 176
6.3 Annual variations of the mean and variance . . . 176
6.4 Wind power time series: Farm No. 1 . . . 177
6.5 Wind power time series: Farm No. 46 . . . 177
6.6 Monthly individual point forecasts . . . 182
6.7 Monthly individual probability forecasts . . . 182
6.8 QQ-plot of PIT values for one-step ahead individual probability forecasts . . . 184
6.9 Monthly aggregated point forecasts . . . 186
6.10 Monthly aggregated probability forecasts . . . 186

6.11 QQ-plot of PIT values for one-step ahead aggregated probability forecasts . . . 187
6.12 Percentage differences in RMSE . . . 189
6.13 Out-of-sample point forecasts . . . 193
6.14 Out-of-sample probability forecasts . . . 193
6.15 Groups of wind farms . . . 196
6.16 Eastern and western wind farms . . . 196
6.17 Mean CRPS in different subsampling approaches . . . 198

E.1 Relative performances of the diurnal seasonality model . . . 234
E.2 Performances of the seasonal exponential smoothing method . . . 235
E.3 Comparing the two approaches . . . 237
E.4 Mean CRPS of forecasts with various lengths of sliding window . 239
E.5 Mean CRPS of various aggregated benchmarks . . . 242
E.6 Variation of mean CRPS with $N_{r.v.}$ . . . 242
E.7 Relative performances of an ARIMA$(p, 1, q)$-GARCH(r, s) model . 244
E.8 Relative performances of an ARIMA(3,1,1)-GARCH(1,1) model . 244

List of Tables

3.1 Anisotropic correlations with wind farm number 1 55
3.2 Summary of the properties of a correlation model 59
3.3 Properties of various correlation models 75

4.1 Summary of two-stage model 106

5.1 Summary of wind power data from 64 wind farms in Ireland . . . 114
5.2 Summary of aggregated point forecasts under RMSE 122
5.3 Summary of aggregated probability forecasts under CRPS 124
5.4 Fit of the correlation models 134
5.5 Diebold-Mariano test statistic for RMSE 158
5.6 Regression coefficients for $W(\mathbf{s}, t)$ 162
5.7 Smoothing parameters for $F(y|\ell, s^2)$ with rescaling 163

6.1 Summary of wind power data from 49 wind farms in Denmark . . 174
6.2 Different subsampling approaches 195

Chapter 1

Introduction and Motivation

> *The most reliable way to forecast the future is to try to understand the present.*
>
> – **John Naisbitt, author of Megatrends**
>
> *If Noah had IBM's Deep Thunder, he would have been piloting a 'smart' ark and wouldnt have needed that dove to find land.*
>
> – **Herman K. Trabish**

IBM's Deep Thunder is the 'brother' of Deep Blue, the famous chess computer that defeated world champion Garry Kasparov in 1997. They are both developed from IBM's Deep Computing initiative. Deep Thunder is a project that utilizes high-performance computing to provide short-term local weather forecasts in real time at high resolution down to 1 km, which aims to address weather-sensitive business problems and improve operational efficiency[1].

In 2009, IBM applied Deep Thunder's weather-forecasting capability to help optimizing wind farm operations and maintenance. They built a virtual 42-MW wind farm comprised of 25 turbines. Through this simulation farm, they showed that wind farm optimization could save a significant amount of money. "In our demonstration, from the ISO (Independent transmission System Operators) perspective, looking out 84 hours in advance and doing optimization on it, it was several hundred thousand dollars," said Jay Mashburn, IBMs Principle Consultant for Wind Power[2].

[1] Information according to IBM: http://www.research.ibm.com/weather/DT.html.

[2] Extracted from the article 'IBM Wants to Make Wind Farms and Solar Power Plants Smarter' by Herman K. Trabish, available at http://www.greentechmedia.com/articles/read/ibm-wants-to-make-wind-farms-and-solar-power-plants-smarter/

1.1 Background and literature review

To begin with the complicated yet interesting topic of wind power forecasts, we start by introducing its background in terms of motivation. We then comment on various approaches of wind power forecasts, and give an overview of the problem in terms of different forecast categories and forecast horizons.

The problem of forecasting wind power can be traced back to 1980s (Brown *et al.*, 1984), when the governments of United States and some European countries such as Denmark and Germany started to develop and commercialize wind farm projects. The Kyoto Protocol was adopted in 1997 and the world started to place more attention on issues related to global warming and climate change. Wind power rapidly became one of the main source of renewable energy that plays an important role in achieving a reduction in carbon emissions. For instance, the Dutch government set up a target of installing a total amount of 1000 MW wind power in the Netherlands by 2000 (van Wijk *et al.*, 1992), and the Danish government aims at having 50% of energy demand met by wind power by 2025 (Tastu *et al.*, 2011). In addition, the European Union targets to generate at least 20% of the power from renewables by 2020[1], in which wind power would correspond to a large proportion due to its abundance, availability and relatively low cost. It is thus inevitable that national power grids would be connected extensively to wind farms and incorporate a significant amount of wind power generation.

A major challenging problem for power grids is to optimally plan and schedule on the dispatch of power generated by coal, gas and other non-renewable resources, as compared with some cheaper or more freely available renewables such as wind energy (Parsons *et al.*, 2004). This is particularly challenging when we consider power systems with limited interconnections, such as Ireland (Fox *et al.*, 2007). Since wind is intermittent and wind power cannot be stored efficiently, there are risks of power shortages during periods of low wind speeds. Wind turbines may also need to be shut down when wind speeds are too high, leading to an abrupt drop of power supply. It is extremely important for power

[1]Relevant details can be found from the European Union Committee report (2007-08), which can be downloaded at http://www.publications.parliament.uk/pa/ld200708/ldselect/ldeucom/175/175.pdf.

system operators to quantify the uncertainties of wind power generation in order to plan for system reserve efficiently (Doherty & O'Malley, 2005). In addition, wind farm operators require accurate estimations of the uncertainties of wind power generation to reduce penalties and maximize revenues from the electricity market (Pinson *et al.*, 2007b). Reliable online updating of wind power forecasts are essential for power grids to control the operations of generators and make optimal decisions on power dispatch. Specific attention is given to forecasting short-term wind power generation at a horizon of several hours to a few days ahead. Costa *et al.* (2008) provide an overview of the history of short-term wind power prediction. Extensive reviews of the short-term state-of-the-art wind power prediction are contained in Giebel *et al.* (2003) and Landberg *et al.* (2003).

1.1.1 Various approaches of wind power forecasts

For the forecasting of wind power generation, there could be two general approaches according to whether one directly models the wind power time series, or models the wind speed and transform the results to wind power. In this thesis, we consider only the direct approach as described below.

Direct approach: Of course, the simplest way is to directly consider the statistical modeling of wind power time series itself. In this case, the difficulty arises from the fact that wind power time series are highly nonlinear and non-Gaussian. In particular, wind power time series at individual wind farms often contain long chains of zeros. Occasionally, there could even be sudden jumps from maximum capacity to a low value due to gusts of wind since turbines have to be shut down temporarily[1]. Nevertheless, it has been shown that statistical time series models, or even persistence models, may outperform sophisticated meteorological forecasts for short forecast horizons within six hours (Fox *et al.*, 2007; Milligan *et al.*, 2004).

[1]However, we do not really observe many such occasions in either the Irish data in Chapter 5 or the Danish data in Chapter 6.

Indirect approach: An indirect approach of wind power forecasts is to focus on the modeling of wind speed and then transform the data into wind power through a power curve (Sanchez, 2006). An advantage is that wind speed time series are smoother and easier to describe by linear models. However, a major difficulty is that the shape of the power curve may vary with time, and it is also difficult to quantify the uncertainties in calibrating the nonlinear power curve. Another indirect approach is to transform meteorological forecasts into wind power forecasts, where ensemble forecasts are generated from sophisticated numerical weather prediction (NWP) methods (Pinson & Madsen, 2009; Taylor *et al.*, 2009). This approach is able to produce reliable wind power forecasts up to ten days ahead, but it requires the computation of a large number of scenarios as well as expensive high quality meteorological data.

1.1.2 Categories of wind power forecasts

Depending on the nature of the wind power time series, we could also identify three categories of wind power forecasts as below. In our thesis, we model and generate all of the following types of forecasts.

Aggregated forecasts: One could sum up the wind power generated from a portfolio of wind farms with hundreds of turbines located around a certain region. Aggregated wind power time series are smoother and easier to forecast because abrupt changes are less likely, and probability masses are in general avoided.

Individual forecasts: Individual wind power refers to wind power obtained from a single wind farm, which typically consists of 5-20 turbines. Figure 1.1 shows a boxplot of the number of turbines in 64 Irish wind farms. The time series could be very fluctuated, and abundant values at zero and maximum capacity could be obtained across a long period of time, say, hours or even days. State-of-the-art wind power prediction systems mainly provide forecasts for a single individual wind farm (Tastu *et al.*, 2011).

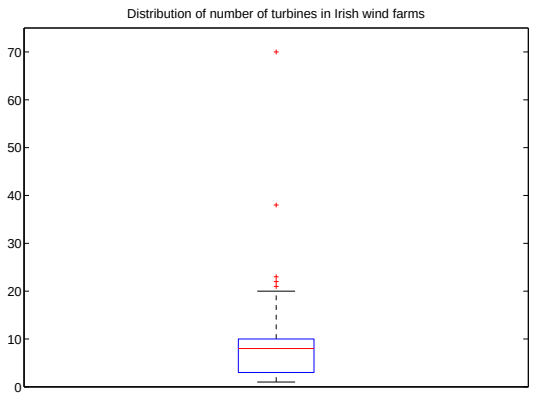


Figure 1.1: This boxplot shows the number of turbines in 64 Irish wind farms. The median is 8 and is shown by the red line. Minimum number of turbines is one and maximum is 70. The edges of the box are the 25% and 75% quantiles respectively.

Spatiotemporal forecasts: This is a multivariate problem where we consider all the individual wind power time series simultaneously and account for their interactions and correlations. We are interested in the contemporary forecasts at each individual wind farm. Since power grids often have connections to a portfolio of wind farms situated at different locations, it is important to utilize the spatial information to produce more sophisticated and reliable forecasts. This relies on the use of spatiotemporal models which are built to describe the dynamics of spatiotemporal stochastic processes. Spatiotemporal models have been applied extensively to environmental statistics such as modeling of tropical ocean surface winds (Wikle *et al.*, 2001), aggregated precipitation (Sanso & Guenni, 1999) and temperatures in north Atlantic ocean (Higdon, 1998; Lemos & Sans, 2009). The ultimate aim of building spatiotemporal models is to describe the spatiotemporal dynamics of the process as well as to capture the spatial and temporal variability at different scales.

To describe the spatiotemporal dynamics, the most common approach is to assume that the conditional distribution of the process is Gaussian, at least up to suitable transformations. Since a Gaussian process is totally char-

acterized by its mean and covariance structures, the problem boils down to the modeling of a global mean field and the spatiotemporal covariance of the residuals. In addition, one may want to build a more sophisticated model to capture spatial and temporal variability at different scales (Cressie, 1993). This leads to the development of hierarchical spatiotemporal models which provides a larger flexibility of modeling a complex process by breaking it down into stages of conditionally specified models (Banerjee *et al.*, 2003; Wikle *et al.*, 1998). The model is naturally estimated by a Bayesian approach, although one could also estimate the parameters via maximum likelihood or the method of moments. Those models are usually written in a Gaussian state space framework (Huang & Cressie, 1996; Stroud *et al.*, 2001). Although Gaussian processes have a lot of nice properties, it is clear that many spatiotemporal processes are non-Gaussian (Diggle *et al.*, 1998). Diggle & Ribeiro (2007) provide a detailed description of a generalized linear model framework in a geostatistical context, which is capable of describing non-Gaussian spatial processes.

1.1.3 Horizons for wind power forecasts

Different horizons for wind power forecasts are considered in different problems. In general, the forecast techniques need to be adapted for different scales of horizons. In our thesis, we consider only short-term forecasts with pure statistical models.

Short-term: Short-term wind power forecasts become one of the most important issues in the literature recently due to the increase of availability in high frequency wind data at a scale of minutes or even seconds (Potter & Negnevitsky, 2006). Short-term usually means a look-ahead horizon of several minutes to several hours. This is an important horizon for optimal planning of power dispatch (Landberg, 1999). Due to the persistence of wind speed, statistical models are most suitable for short-term forecasting (Kariniotakis *et al.*, 1996).

Long-term: Long-term forecasts refer to horizons in the order of days. In this case, pure statistical models are less useful[1], and usually meteorological information such as wind speed, wind direction and pressure is applied to the numerical weather prediction (NWP) models to generate superior forecasts. Long-term forecasts are useful for planning connection or disconnection of wind turbines or conventional generators, thus achieving low spinning reserve and optimal operating costs (Barbounis *et al.*, 2006; Ma *et al.*, 2009).

1.2 Major challenges in wind power forecasting

1.2.1 Boundedness and skewness

Various characteristics in a typical wind power time series make it a challenging problem to generate competitive wind power forecasts. Most obviously, wind power is a non-negative variable, which means that the distribution is bounded below by zero. Wind power is also bounded above by the maximum capacity of the wind turbines. In addition, the distribution of wind power data is highly non-Gaussian. It is significantly right-skewed (positively skewed) and the tail on the right is heavier than a Gaussian distribution. This is partly due to the fact that wind power mainly depends on wind speed, and the wind speed has a skewed distribution which is sometimes described by the Rayleigh distribution or the Weibull distribution (Gipe, 2004). In addition, wind power is related to wind speed through a cubic power curve[2], and this is a non-linear transformation which could account for the skewness in wind power generation (Pinson *et al.*, 2008). Appropriate transformations are required to normalize the data.

[1] Nevertheless, statistical models are still important (Sideratos & Hatziargyriou, 2007), and statistical approaches are needed to calibrate the ensemble predictions of physical variables such as wind speed (Taylor *et al.*, 2009).

[2] More details on the relations between wind speed and wind power could be found in Wagner & Mathur (2009).

1.2.2 Discrete probability masses

Another major challenge in modeling wind power is that there is usually a large number of zeros in the wind power data, due to various reasons such as the absence of wind, the occurrence of severe weather conditions or maintenance of the turbines. This introduces a probability mass at zero wind power, which gives a discontinuity in the distribution function. This effect is difficult to handle via smooth transformations due to the discontinuity. Occasionally, we also observe a number of maximum wind power generation at certain sites, indicating that the wind turbines are running at their maximum capacities. Similarly, this introduces a probability mass at maximum wind power, which causes the same problem as the zeros. This feature is similar to rainfall data, where one could have long periods of dry days with no rainfall observation (Little *et al.*, 2009). A possible approach is to handle these probability masses explicitly in the model, which is proposed in Chapter 4.

1.2.3 Non-stationarity and long memory

Apart from problems that arise from the distribution of wind power, another challenge in forecasting wind power is due to the non-stationarity of the time series. Wind power, being driven by wind, exhibits seasonality across the year as well as diurnal cycles within a day. It is also highly persistent, as demonstrated from the slow decay of the autocorrelations and the fact that persistent forecasts are very difficult to beat at short horizons. In fact, there exists evidence that wind speed data has long memory effects (Haslett & Raftery, 1989). More problems could arise when one considers spatiotemporal wind power forecasts. In such cases, wind power time series could even be non-stationary in both temporal and spatial dimensions. For instance, the variance of wind power may not be constant at different locations, and covariance structures of the spatiotemporal process may not simply depend on the difference between spatial coordinates.

1.3 Exploratory analysis of wind data

As we have described some features of wind power data above, it would be useful to look into more details about typical wind power data sets. We do some simple exploratory analysis on the distributions, autocorrelations and seasonalities of wind power in this section. These analyses are based on the Irish data in Chapter 5 and the Danish data in Chapter 6.

In particular, we consider some transformations for the wind data so as to normalize it to an approximately Gaussian distribution. This will be applied extensively in some of the models concerning the Irish and Danish data, so it is useful to mention it here. Of course, we cannot cover all properties of wind power data in this section. Some particular concerns, for example, spatiotemporal correlation structures of wind power, will be covered in more detail when we focus on the topic in later chapters.

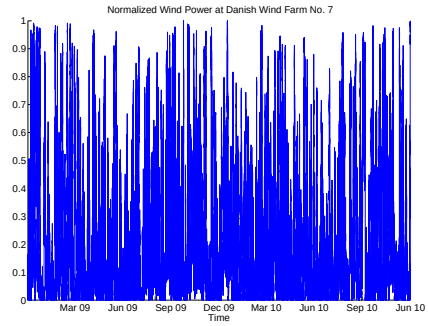
1.3.1 Individual wind power at a single wind farm

1.3.1.1 Distributions and autocorrelations

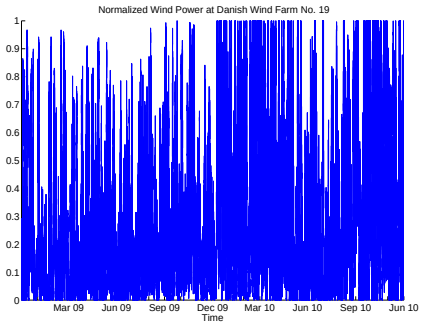
First, let us consider individual wind power generation at a single wind farm. In order to facilitate comparisons between wind farms with different capacities, in this thesis, we always normalize the individual wind power by dividing by the maximum capacity[1] of the wind farm. We denote the normalized individual wind power at location s and time t by $y(\mathrm{s}, t)$, which has values between $[0, 1]$.

Examples of time series for $y(\mathrm{s}, t)$ at low and high variances are shown in Figure 1.2 (a)-(b). In general, wind power with higher variances are associated with higher mean level. We also see that individual wind power in general contains abundant zeros. Occasionally, some wind power time series have a large numbers of ones, i.e. maximum output, but the problem is usually less severe. Two examples of the corresponding densities of individual wind power are shown in Figure 1.2 (c)- (d). It is clear that there is a probability mass at zero, which sometimes occurs at one as well. Such probability masses in wind power distributions have been shown in Carlin & Haslett (1982).

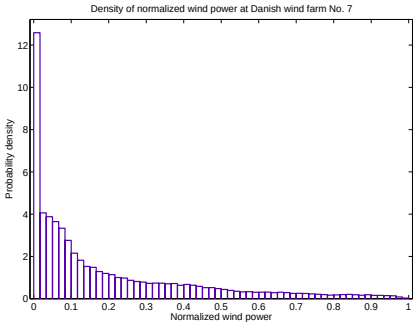
[1]The capacity of a wind farm is the maximum output of a wind farm when all turbines operate at their maximum nominal power.



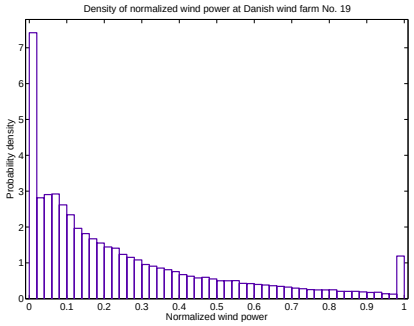
(a)Time series of normalized individual wind power at Danish wind farm No. 7, where the index is according to Table 6.1. It has a low variance of about 5%.



(b)Time series of normalized individual wind power at Danish wind farm No. 19, where the index is according to Table 6.1. It has a high variance of about 7%.



(c)Density of normalized individual wind power at Danish wind farm No. 7, where the index is according to Table 6.1. It has a probability mass at zero. This density corresponds to the time series as shown in Figure 1.2(a).



(d)Density of normalized individual wind power at Danish wind farm No. 19, where the index is according to Table 6.1. It has a probability mass at both zero and one. This density corresponds to the time series as shown in Figure 1.2(b).

Figure 1.2: (a) and (b): Time series of normalized individual wind power. (c) and (d): Corresponding densities of normalized individual wind power.

Figure 1.3 shows a typical plot of autocorrelations of individual wind power, which is seen to be correlated even after 24 hours, i.e. 96 steps in our 15-minute data. Nevertheless, this autocorrelation is weaker than that of the aggregated wind power, as will be shown in Figure 1.6(c) in the next section.

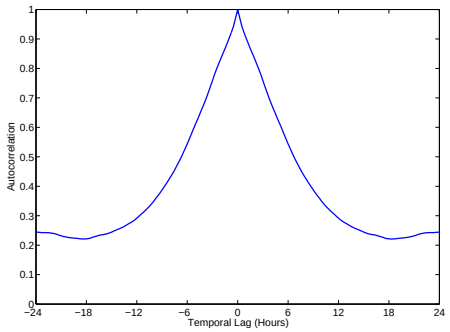


Figure 1.3: Autocorrelations of individual wind power, which is seen to be correlated after even 24 hours. Nevertheless, this autocorrelation is weaker than that of the aggregated wind power. Autocorrelations are symmetric with respect to temporal lags.

1.3.1.2 Modified logit transformation

As seen in Figure 1.2, the density of wind power at a single wind farm is obviously far from being Gaussian. Although transformations could not remove the discrete probability masses, they are very useful and are mainly motivated by three reasons:

1. Wind power generation depends on wind speed through a cubic power curve (Wagner & Mathur, 2009). The different inflection points in the power curve at low and high wind power can be nicely modeled by using suitable transformations such as the (modified) logit transformation. A simplified version of a power curve and some relevant details are discussed in Figure 2.3

2. The transformed wind power will no longer be confined as non-negative. This helps to generate valid forecasts through inverse transformation

3. The residuals could be better approximated by Gaussian distributions, which is an important assumption in some of the models and forecasts. We attempt to make the conditional distribution of the transformed individual wind power y close to a Gaussian distribution[1].

[1]This will be necessary before we apply spatiotemporal kriging to the data in Chapter 3. Nevertheless, note that for the two-stage model in Chapter 4, we will no longer need this

We will discuss in the next section that for aggregated wind power, the logit transformation is a nice candidate for achieving an approximate Gaussian distribution. As a result, let us consider a similar transformation applied to individual wind power. Unfortunately, we could not simply apply the logit transformation on individual wind power since there are many zeros and ones in $y(s,t)$ where the logit transformation is undefined. To resolve the problem, we introduce a tuning parameter $\delta > 0$ and define the modified logit transformation as

$$z = \log\left(\frac{y+\delta}{1-y+\delta}\right), \qquad 0 \le y \le 1 \tag{1.1}$$

Figure 1.4 shows an example of the density of residuals of individual wind power y_t at Irish wind farm No. 1, where the residuals are calculated as $\epsilon_t = y_t - y_{t-1}$. We see that the modified logit transformation reduces the skewness and kurtosis of the distribution of residuals. Note that the y-axis of the two plots are different.

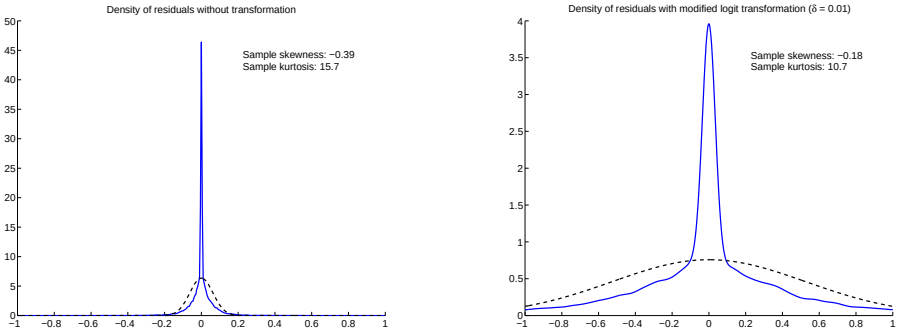


Figure 1.4: Density of residuals of individual wind power at Irish wind farm No. 1, where the residuals are calculated as $\epsilon_t = y_t - y_{t-1}$. Left: without transformation. Right: with modified logit transformation with $\delta = 0.01$. The modified logit transformation reduces the skewness and the kurtosis of the distribution. The dashed lines are the fitted Gaussian distributions. Note that the y-axis of the two plots are different.

To estimate the optimal parameter δ, we maximize the likelihood of the residuals of the transformed data $\epsilon_t = z_t - z_{t-1}$ so that they are better approximated by a Gaussian distribution $N(\mu, \sigma^2)$ with mean $\mu = \mathrm{E}[\epsilon_t]$ and variance $\sigma^2 = \mathrm{Var}[\epsilon_t]$.

transformation because in that case, we will construct an approach that directly models the continuous distribution (using truncated normal distributions) and handles the problem of probability masses.

We consider the two sample Kolmogorov-Smirnov test (Lilliefors, 1967; Massey, 1951) with test statistics

$$T_{KS} = \max(|F(\epsilon_t) - \Phi(\epsilon_t)|) \tag{1.2}$$

where F and Φ are the empirical cumulative distribution functions of ϵ_t and the normal distribution $N(\mu, \sigma^2)$ respectively. For each wind farm j, we calculate the test statistic T_{KS}^j and we estimate the value of δ in (1.1) such that the median test statistic is minimized[1]. Note that we have also considered maximizing the joint likelihood at all wind farms, but due to the large number of wind farms, the likelihood function is not smooth and it is difficult to estimate a global maximum. On the other hand, minimizing the median test statistics in (1.2), i.e. median distance between the empirical distribution $F(\epsilon_t)$ and the target distribution $\Phi(\epsilon_t)$, provides a more robust approach. We have obtained the distributions of T_{KS}^j under some other transformations, and the modified logit transformation in (1.1) gives the best results[2].

[1]The median test statistic is used as it provides a more robust measure than the mean.

[2]For example, we have considered 49 individual wind power time series (1 year) from Denmark. The mean and median of T_{KS}^j at all wind farm $j = 1 - 49$ using the square root transformation is .0552 and .0546 respectively, while the mean and median using the modified logit transformation is .0423 and .0396 respectively.

1.3.2 Aggregated wind power

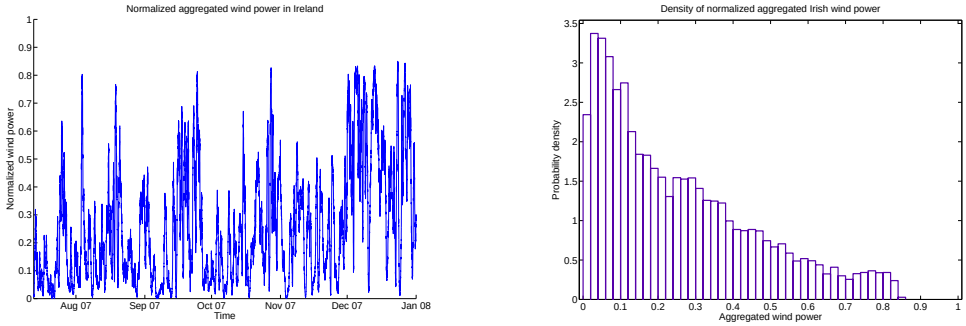
1.3.2.1 Distributions and autocorrelations

Aggregated wind power is obtained by adding up all the individual wind power in a portfolio of wind farms. Again, in order to facilitate comparisons between data sets with different capacities, we normalize the aggregated wind power by dividing by the total capacity. Thus the normalized aggregated wind power, denoted by $y_A(t)$, is bounded within $[0, 1]$. Expressing the normalized individual wind power generated by a wind farm with capacity ω_j and at location s_j by $y(\mathrm{s}_j, t)$, we have

$$y_A(t) = \frac{\sum_{j=1}^{N} \omega_j y(\mathbf{s}_j, t)}{\sum_{j=1}^{N} \omega_j} \tag{1.3}$$

where N is the total number of wind farms in a portfolio.

Figure 1.5 (a) shows a typical example of the time series for normalized aggregated wind power. One can see that the maximum value is smaller than 0.9, and in general it does not reach 1. In fact, there are no zeros either. This is due to diversity resulting from geographical spread. A bigger contrast with the individual wind power time series could be seen by plotting the density of the aggregated wind power as shown in Figure 1.5 (b).

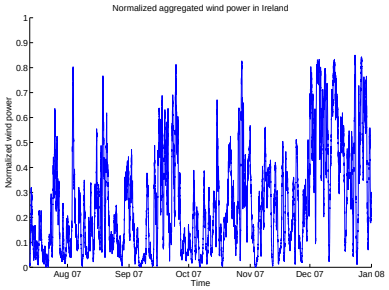


(a) An example of normalized aggregated wind power from 64 wind farms in Ireland. One can see that the maximum value is smaller than 0.9, and in general it does not reach 1. In fact, there are no zeros either.

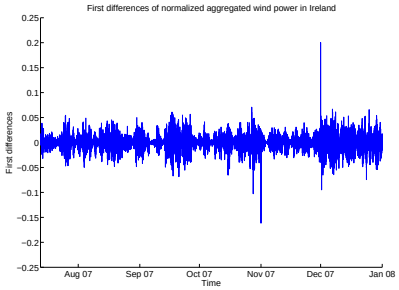
(b) An example of the density of normalized aggregated wind power in Ireland. The density is non-Gaussian since the data is bounded between zero and one, and it is right-skewed (positively-skewed). This is a big contrast with the densities of individual wind power in Figure 1.2 (c) and (d).

Figure 1.5: (a): Time series and (b) density of normalized aggregated wind power.

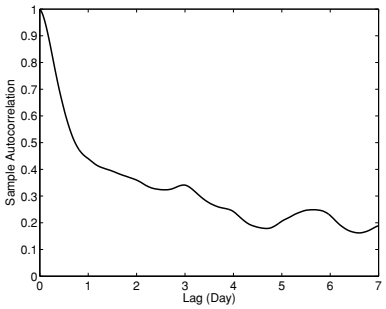
As seen in Figure 1.5 (a), aggregated wind power is nonstationary. The mean level is larger in December and January. Volatility clusterings exist and there are occasional spikes. To remove the non-stationarity, we consider the first differences as shown in Figure 1.6(b). There is still evidence of volatility clusterings and occasional spikes. However, the autocorrelations in the differenced time series are significantly reduced. Figure 1.6(c) shows the autocorrelations of aggregated wind power, which decay slowly with significant autocorrelations even at long lags of 7 days. On the other hand, the autocorrelations for the first differences are much smaller as shown in Figure 1.6(d).



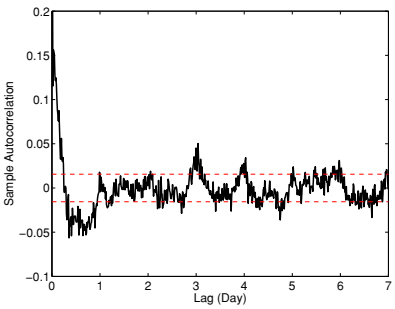
(a)Normalized aggregated wind power from 64 wind farms in Ireland.



(b)First differences of normalized aggregated wind power in (a). There is still evidence of volatility clusterings. Spikes occur when there are drastic increases or decreases in wind power.



(c)Autocorrelations of aggregated wind power up to a lag of 7 days. The autocorrelations decay very slowly and are significant even at long lags of 7 days.



(d)Autocorrelations of the first differences of aggregated wind power up to a lag of 7 days, where the dashed lines are the bartlett intervals for 2 standard deviations. The autocorrelations are significantly reduced.

Figure 1.6: (a) and (b): Time series of normalized aggregated wind power and its first differences. (c) and (d): Autocorrelations of the time series in (a) and (b) respectively.

1.3.2.2 Seasonality

Since our aim is to generate short-term forecasts, we do not focus on modeling any long term seasonality, which often appears in wind data due to the changing wind patterns throughout the year. For example, we can model a cycle of T_0 days by regressing the wind power with M harmonics of sines and cosines with periods $T = j/(T_0 \times 96), j = 1, \ldots, M$, where 96 is the number of data per day in our Irish and Danish wind data. For illustration purpose, let us choose $T_0 = 90$ and $M = 16$. This gives a fitted time series as shown in Figure 1.7 with $R^2 = .395$. One may then consider to model the deseasonalized data, but studies show that results may be worse than those obtained by modeling the seasonality directly (Jorgenson, 1967). We have also tried to account for these long term seasonalities, but forecast results are poor as we only consider short horizons in this thesis.

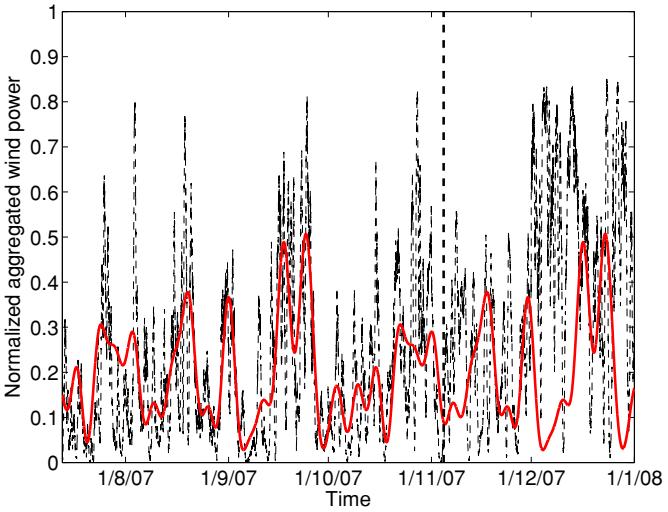


Figure 1.7: Long term seasonality appears in wind power due to the changing wind patterns throughout the year. In this example, we regress the aggregated wind power in Ireland in the training set with 16 harmonics of sines and cosines with periods $T = j/(90 \times 96), j = 1, \ldots, 16$. The fit gives an $R^2 = .395$. The thin dashed line is the observed wind power and the solid red line is the fitted time series. The vertical dashed line dissects the data into a training set and a testing set. We have also tried to account for this long term seasonalities, but forecast results are poor as we only consider short horizons.

On the other hand, one could be more interested in the diurnal cycle since it plays a more important role in intraday forecasts. Diurnal cycles may appear in wind data due to different temperatures and air pressures during the day and the night, and wind speeds are sometimes larger during the day when convection currents are driven by the heating of the sun (Weisser & Foxon, 2003). For example, with the Irish and Danish wind data in Chapters 5 and 6, we investigate the variations of the mean aggregated wind power within a day and observe the patterns as shown in Figure 1.8. We see that wind power generation is slightly higher at noon to afternoon, and lower at midnight. We have tried to account for this diurnal seasonality and generate forecasts based on deseasonalized data (See Appendix E.1). However, results are not encouraging due to the short horizons that we consider in this thesis. Thus we decide to exclude the modeling of any diurnal cycles here.

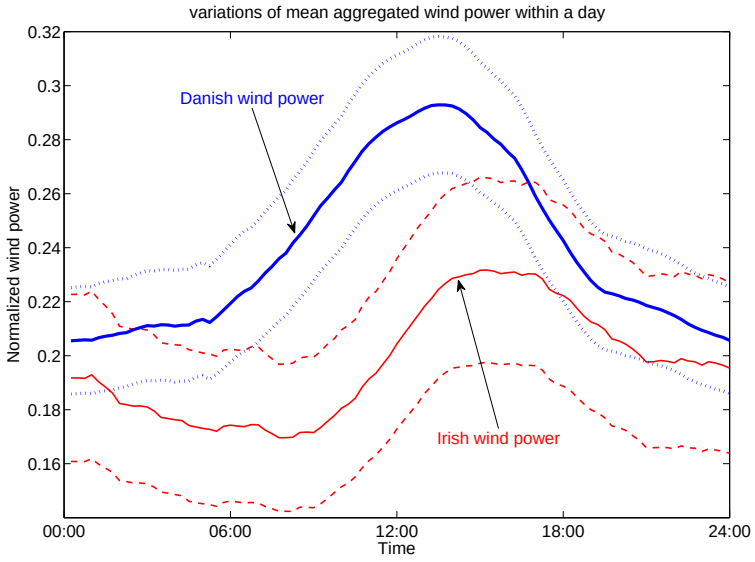


Figure 1.8: Diurnal seasonality appears in wind power due to different temperatures and air pressures during the day and the night, and wind speeds are sometimes larger during the day when convection currents are driven by the heating of the sun. In this example, we calculate the mean of the aggregated wind power at different times within a day. The solid lines represent the mean wind power and the dotted lines are 95% confidence intervals. We consider 11008 observations in the Irish wind power and 35040 observations in the Danish wind power.

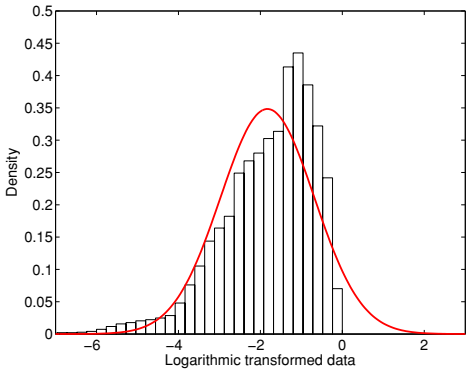
1.3.2.3 Logit transformation

For wind speed data, we have seen examples of normalizing the data by the logarithmic transformation and the square root transformation (Taylor *et al.*, 2009). However, due to the cubic power curve that transforms wind speed to wind power, such transformations are no longer ideal. These are demonstrated in Figures 1.9 (a) - (d) for the normalized wind power in Ireland.

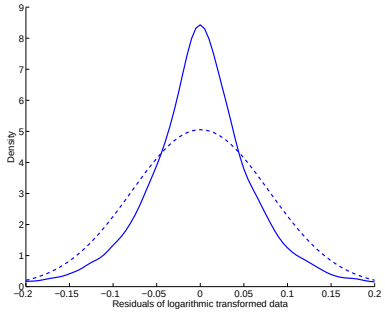
As mentioned previously, the logit transformation is a better candidate for normalizing aggregated wind power $y_A(t)$. This can be traced back to the work of Johnson (1949), and recently Bremnes (2006) applies this transformation to model wind power in Norway. The logit transformation is given by

$$z_A = \log\left(\frac{y_A}{1-y_A}\right), \qquad 0 < y_A < 1 \tag{1.4}$$

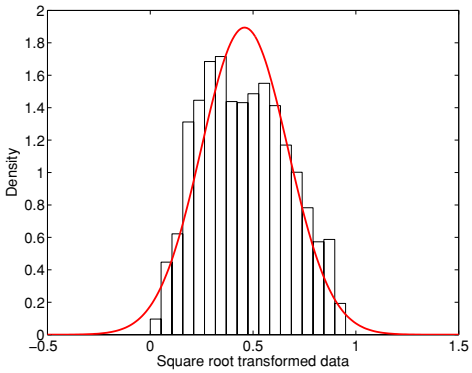
and the transformed data $z_A(t)$, as well as the residuals $z_A(t) - z_A(t-1)$, follow a distribution which can be better approximated by a Gaussian distribution as shown in Figures 1.9 (e) - (f). In contrast with individual wind power, we do not encounter any values of zero or one for $y_A(t)$, and so (1.4) is well defined. In Chapter 2, we apply this transformation and build a Gaussian model to generate multi-step probability forecasts for aggregated wind power generation.



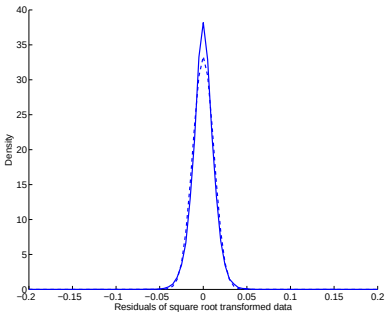
(a)Logarithmic transformation



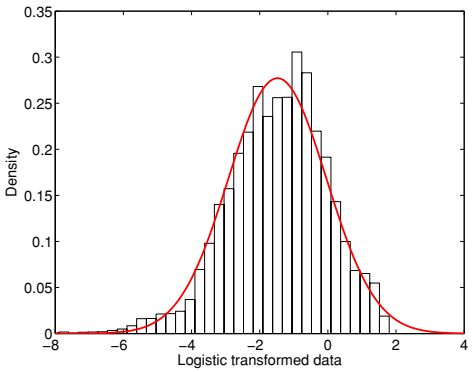
(b)Logarithmic transformation (Residuals)



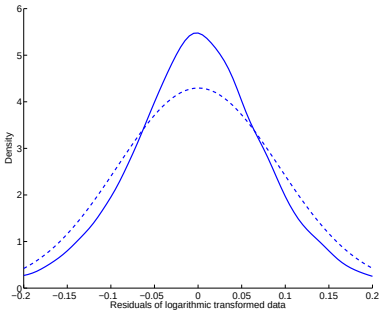
(c)Square root transformation



(d)Square root transformation (Residuals)



(e)Logit transformation



(f)Logit transformation (Residuals)

Figure 1.9: Distributions of various transformed wind power (left) and their corresponding residuals (right). The data used are the normalized aggregated wind power in Ireland. (a) Logarithmic transformation, (c) square root transformation and (e) logit transformation. Logarithmic and square root transformations are commonly applied to convert wind speed into an approximate Gaussian distribution, but is inappropriate for wind power. However, for wind power, logit transformation is a better candidate. The red solid lines in the left panels and the dashed lines in the right panels are the fitted Gaussian distributions by maximizing the likelihood. Note that the y-axis of the plots are different.

1.4 Rationale in choosing the various models

With the preliminary analysis of wind power data as discussed above, one needs to think of the potential models that could describe and forecast the data successfully. In addition, this is a practical problem where in reality, forecasts need to be made at real time and efficient online updating is required. As a result, one should bear in mind that the models should be efficiently estimated and parameters must be robust across different data sets.

1.4.1 Gaussian models

With such considerations, we first consider a Gaussian model for aggregated wind power in Chapter 2. Fortunately, the much smoother distribution of the aggregated wind power as well as their residuals could be nicely approximated by a Gaussian distribution through the application of the logit transformation in (1.4). This allows us to build Gaussian models based on the transformed data, and the main advantage of Gaussian models is the analytical tractability which is essential for our study of multi-step ahead forecasts.

Our another approach to modeling aggregated wind power is to explore the exponential smoothing methods. This choice is driven by the fact that exponential smoothing is a robust methodology for forecasting. The smoothing parameters are able to control the rate of decay of information with time, thus putting more weights on recent observations. This is a reasonable assumption for wind power given that it is highly persistent, while the autocorrelations decay smoothly with temporal lags. Another reason for the use of exponential smoothing lies on the fact that these methods are closely related to the state-space models (Hyndman *et al.*, 2008), and thus provide an efficient way for us to generate multi-step ahead forecasts as in the Gaussian models above.

1.4.2 Kriging models

As we move on to consider spatiotemporal forecasts for a portfolio of wind power, we face the problem of multivariate modeling. More importantly, the distribution of individual wind power is no longer close to Gaussian, and there exists

probability masses which are more difficult to model. Applying the results in spatial statistics, the first idea that we attempt is to generalize the conventional best linear unbiased predictor (BLUP), or the kriging predictor in the language of geostatistics, to include the temporal dimension. This results in our use of spatiotemporal kriging predictor for forecasting a portfolio of individual wind power. Unfortunately, this relies on the assumption that individual wind power follows a Gaussian process. Following the successful logit transformation for aggregated wind power, we modify it so as to avoid infinities and apply a modified logit transformation on individual wind power. Although the distributions are still non-Gaussian, and in particular, we only mitigate the problem of probability masses in the distribution, spatiotemporal kriging provides an efficient way to linearly interpolate nearby observations and obtain an estimate of wind power at a site based on the mean and covariance structure of the Gaussian field.

To calculate the spatiotemporal kriging predictor, we attempt to construct covariance models for the transformed wind power. As we shall see, covariances of wind power is anisotropic due to the fact that wind is driven by weather fronts which in general moves along a certain direction. Thus, we implement this feature in our covariance model by transforming the temporal lag appropriately. This turns out to be an important improvement in generating forecasts, especially for wind power generated at wind farms that lie along the propagation of the weather front.

1.4.3 Two-stage models

Motivated by the need to handle the probability masses in the individual wind power distribution appropriately, and realizing that this problem could not be perfectly handled by the help of transformations, we turn to the approach of explicitly modeling these probability masses by a two-stage approach. This leads to the idea of a two-stage model, where we model the probability of having zero or maximum wind power in the first stage. Conditional on different regimes, we then model the distribution of the wind power in the second stage. In this model, probability of having different regimes of wind power is driven by a latent Gaussian process W, whereas the probability of observing a particular value of

wind power y with $y \in (0, 1)$ is driven by another independent Gaussian process Z. To account for the covariance structures among individual wind power time series, we impose appropriate covariance models for the latent Gaussian processes W and Z. We consider the same classes of covariance models that have been applied in the spatiotemporal kriging models and this constitutes a nice comparison of the performances of these covariance models in different applications.

In the two-stage models, we also need to describe the dynamics of the distribution of wind power in $(0, 1)$. Due to the robustness of exponential smoothing and its advantage of modeling a large number of time series with relatively few parameters, we adopt a smoothing approach to evolve these individual wind power distributions. In particular, we simultaneously smooth through the mean level and the variances of the individual wind power, as this has been successfully applied in the case of aggregated wind power forecast.

1.5 Contributions in our study

With the above rationale in mind, we construct appropriate models to forecast aggregated and individual wind power. Our work in this thesis leads to various contributions in the following aspects:

1.5.1 Multi-step probability forecasts

Most of the early literature on wind research focuses on point forecasts (Holttinen, 2005; Madsen *et al.*, 2005; Milligan *et al.*, 2004). However, due to the need of quantifying the uncertainties in wind power generation, there is an increasing demand for quantile or even probability forecasts (Juban *et al.*, 2007; Pinson *et al.*, 2007a). In this thesis, focuses are put on generating full probability forecasts and we provide detailed analysis on the forecast performances based on appropriate metrics tailored for probability forecasts. More importantly, our models are purely statistical, as opposed to many of the existing literature on probability forecasts where numerical weather predictions (NWP) are fed into the models (Bremnes, 2004). In the case of using NWP, one could utilize multi-step ahead NWP's as the model's inputs and obtain forecasts at larger horizons. In our

case, we directly iterate our models and obtain forecasts according to the data generating process (DGP).

In conclusion, there are very few studies of multi-step probability forecasts using purely statistical models. A few investigations consider multi-step point forecasts with ARIMA models (Marcellino *et al.*, 2006). Some others consider single-step ahead probability forecasts with statistical models (Gneiting *et al.*, 2006), while some papers focus on multi-step probability forecasts apply numerical weather predictions (Bremnes, 2004). Early studies of multi-step probability forecasts can be found in Davies *et al.* (1988) and Moeanaddin & Tong (1990), where the probability densities are estimated using recursive numerical quadratures. Another example is the work by Manzan & Zerom (2008), where forecast densities for the U.S. Industrial Production series are obtained by non-parametric bootstrap approaches. However, they only evaluate forecasts up to two-step ahead horizons, while in this thesis, we consider horizons up to 96 steps ahead for aggregated forecasts and 12 steps ahead for spatiotemporal forecasts.

1.5.2 Spatiotemporal correlation analysis and modeling

In spatiotemporal forecasts, a critical component is the correlation model which describes the spatiotemporal correlations between wind power at different locations and time. Using a similar idea as the moving Lagrangian reference frame, we successfully construct a anisotropic correlation model by transforming the temporal lags so as to account for the general movement of the weather front. This has shown to be better than transforming the spatial coordinates in some cases. We construct this model by analyzing the empirical spatiotemporal correlations of real wind data, and this is a valuable study because there are very little research on large spatiotemporal data sets due to the difficulty in obtaining decent data. A recent study by Tastu *et al.* (2011) analyzes 22 wind farms in western Denmark with hourly wind data over the first seven months in 2004. In this thesis, we analyze 15-minute wind data from 49 Danish wind farms between 1-Jan-2009 to 31-Dec-2010, as well as 15-minute wind data from 64 Irish wind farms between 13-Jul-2007 to 01-Jan-2008. These data sets are much larger than those being studied in most existing papers, and reveal important statistics on

the features of spatiotemporal correlations in a generic wind power data set. As a result, it is extremely useful to extend the modeling techniques in this thesis to other wind power time series.

1.5.3 Two-stage model for wind power

Another major contribution in this thesis is our work on forecasting wind power using the two-stage model, which we generalize the work of Berrocal *et al.* (2008). We successfully demonstrate the importance of modeling the discrete probability masses at zero and one in wind power distributions by using the two-stage model. We show that the quantile forecasts at low quantile values (i.e. when wind power is close to zero) as well as those at high quantile values (i.e. when wind power is close to one) are significantly improved when we apply the two-stage model instead of the simpler but less realistic kriging approach. According to our best knowledge, this is the first piece of research to apply a two-stage model driven by latent Gaussian process to generate spatiotemporal wind power forecasts. With the outstanding forecast performances of this model and the improvements in computational power for drawing Monte-Carlo simulations, we envisage this idea of two-stage modeling to be utilized in many other portfolios of wind power so as to improve probability forecasts and enhance the planning of power dispatch.

1.5.4 Statement of originality

Unless otherwise noted, all the results in this thesis are original. The work of Chapter 2 is described in the paper co-authored with Patrick McSharry (Lau & McSharry, 2010).

1.6 Roadmap of this thesis

To summarize the structure of our work, a flow chart showing the road map of this thesis is illustrated in Figure 1.10. The light rectangular blocks indicate topics that are discussed in this thesis. The darker diamonds are decision nodes where different possibilities have been considered.

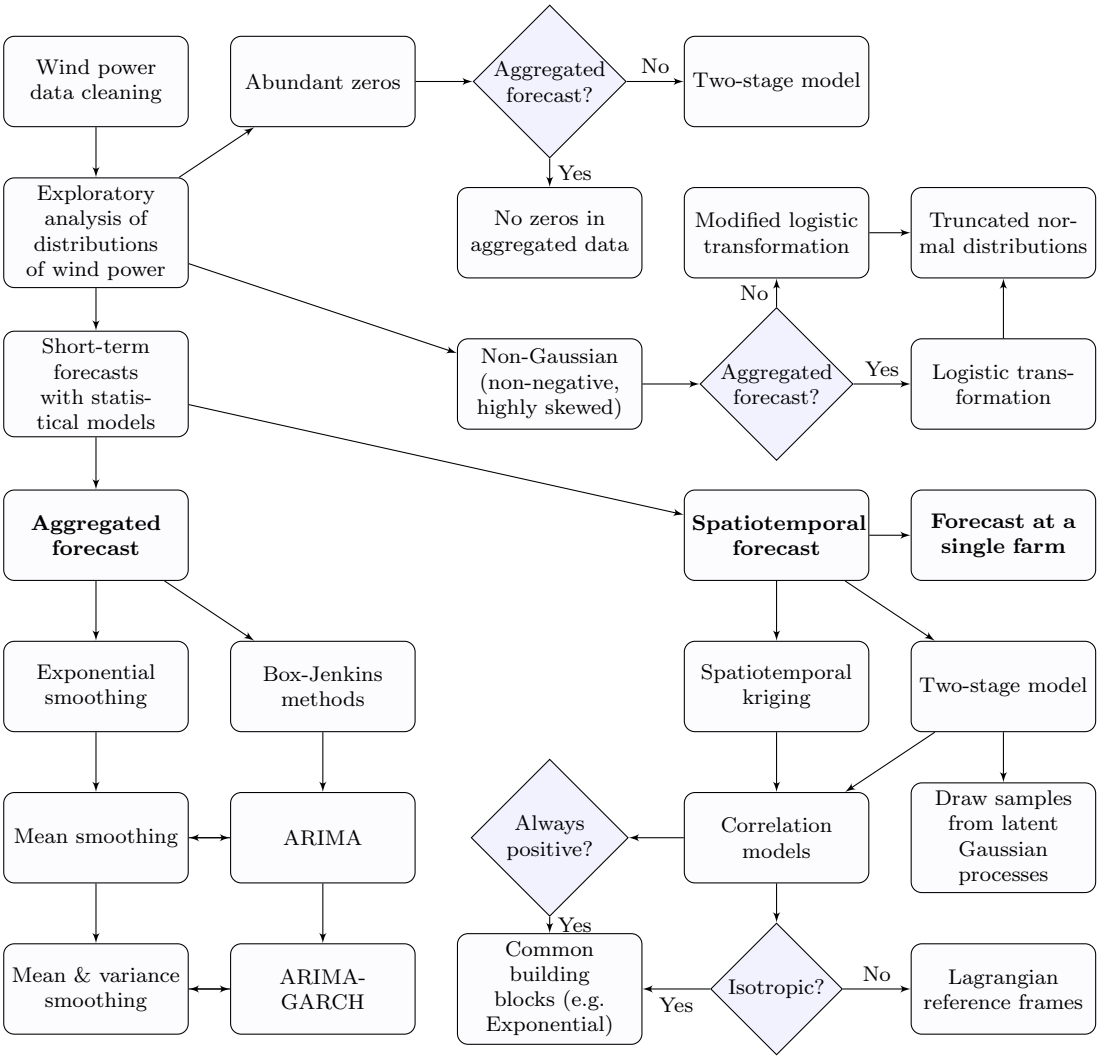


Figure 1.10: Flow chart showing the road map of this thesis. The light rectangular blocks indicate topics that are discussed in this thesis. The darker diamonds are decision nodes where different possibilities have been considered.

Chapter 2

Models for Aggregated Wind Power

Aggregated wind power generation is the total wind power obtained by summing up all the individual power generated in a portfolio of wind farms. We usually consider aggregated wind power when the portfolio of wind farms belongs to the same power grid, or when they are located in the same region or country. The forecast of aggregated wind power is important as it is the total level that could be dispatched to customers through the grid. Compared with individual wind power at a single wind farm, aggregated power is easier to forecast because the individual noises are smoothed out. More importantly, aggregated power do not have the problem of abundant zeros that leads to a probability mass in the distribution. It is because it is extremely unlikely that all wind farms in the portfolio (with hundreds of turbines located at different sites) generate no power simultaneously.

In this chapter, we first study some possible models for aggregated wind power generation. Although aggregated wind power do not have the problem of probability mass at zero, two challenges still exist. The first challenge is that aggregated wind power is double-bounded. It is bounded below by zero and bounded above by the maximum total capacity of the turbines. As a result, we have to ensure that the model forecasts lie between this permissible range. The second challenge is that wind power distribution is usually right-skewed. Together with the fact that it is bounded, its distribution is clearly non-Gaussian. Thus, one may need

to apply suitable transformations to the data if certain model assumptions are to be satisfied.

In the following, we consider two different approaches for modeling aggregated wind power[1]. In the first approach, we apply a transformation to the data so that we obtain an approximately Gaussian distribution. We then model the transformed data with Gaussian models. Final forecasts are generated by inverse transforming the model forecasts, which would be guaranteed to lie within the permissible range. In the second approach, we do not apply any initial transformations to the data. Instead, we apply a very robust and adaptive algorithm, namely the exponential smoothing method, to obtain a preliminary estimation of the location and scale parameters of the wind power distribution. These serve as rough estimates of the level and variance of wind power. To ensure that the final forecasts are in the valid range, we put a constraint on the distribution of wind power. We assume that it belongs to a parametric family governed by the location and scale parameters, which are estimated by exponential smoothing as mentioned. In other words, the second approach differs from the first by assuming a parametric distribution for wind power.

Note that an important objective for both approaches is to generate multi-step ahead forecasts for wind power. For a general model driven by an arbitrary process, multi-step ahead forecasts could be obtained by Monte-Carlo simulations. However, for the sake of efficient online updating of forecasts, we do not consider Monte-Carlo simulations at this stage. Instead, we aim at developing models which could be easily iterated to obtain multi-step ahead forecasts. A most popular choice is the Gaussian model due to its analytical tractability. This is applied in the first approach. For the second approach, we see that in fact one has to assume Gaussianity in the exponential smoothing methods in order to iterate the forecasts into further horizons. We will discuss these in more details in the following sections.

[1]Forecast approaches developed in this chapter and the corresponding forecast results in Chapter 5 have been published in Lau & McSharry (2010).

2.1 Gaussian model for transformed data

2.1.1 Logit transformation

In the first approach, we consider the transformation of wind power data into an approximately Gaussian distribution so that we could describe the transformed data by a simple Gaussian model. As described in (1.3), we first normalize the aggregated wind power by dividing by the total capacity of all wind farms in the portfolio, and we denote the normalized aggregated wind power by $y_A(t)$. We have shown that the logit transformation in (1.4), i.e.

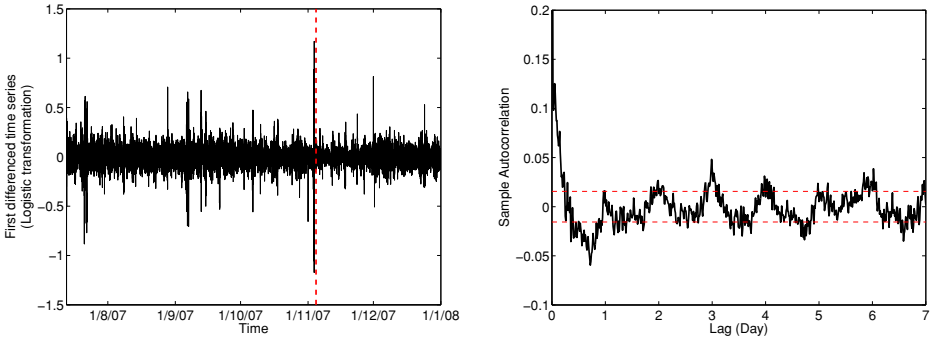
$$z_A(t) = \log\left(\frac{y_A(t)}{1 - y_A(t)}\right), \qquad 0 < y_A(t) < 1 \tag{2.1}$$

is capable of transforming the wind power into an approximately Gaussian distribution, and this transformation maps the support from $(0, 1)$ to the entire real axis. An example of the distribution of the transformed data is illustrated in Figure 1.9(e).

2.1.2 ARIMA-GARCH models

After obtaining a Gaussian distribution, simple Gaussian models could be applied on the data. In particular, we look into the conventional ARIMA-GARCH models with Gaussian innovations. We consider modeling the transformed series $z_A(t)$ obtained in (2.1). Since we are not considering individual wind power in this chapter, for notational brevity we denote the transformed data by z_t in the following.

As wind power is non-stationary, so could be the transformed data. Most non-stationarities could be removed by taking differences, and so we consider the first differences $w_t = z_t - z_{t-1}$. When compared with the original first differences $y_t - y_{t-1}$ in Figure 1.6(b), the first differences of logit transformed data w_t has less evidence of volatility clustering and smaller autocorrelations. These are illustrated in Figure 2.1 (a) and (b) respectively.



(a) Time series of the first differences of logit transformed Irish wind power. The variance is not changing as rapidly as before, and the amount of volatility clustering is reduced. The data is dissected by the dashed line into a training set and a testing set, which will be mentioned later in Chapter 5.

(b) Autocorrelations of the first differences of logit transformed Irish wind power. The dashed lines are the bartlett intervals for 2 standard deviations.

Figure 2.1: (a) Time series and (b) autocorrelations of the first differences of logit transformed Irish wind power.

As a result, we consider modeling the first differences of logit transformed data w_t by an ARIMA(p, q)-GARCH(r, s) process. This is equivalent to modeling z_t by an ARIMA$(p, 1, q)$-GARCH(r, s) model[1]

$$\begin{aligned} w_t &= \mu + \sum_{i=1}^{p} \phi_i w_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t, \qquad \varepsilon_t | \mathcal{F}_{t-1} \overset{i.i.d.}{\sim} N(0, \sigma^2_{\varepsilon;t}) \\ \sigma^2_{\varepsilon;t} &= \omega + \sum_{i=1}^{r} \alpha_i \varepsilon^2_{t-i} + \sum_{j=1}^{s} \beta_j \sigma^2_{\varepsilon;t-j} \end{aligned} \tag{2.2}$$

where $w_t = z_t - z_{t-1}$, and $\mu, \phi_i, \theta_j, \omega, \alpha_i, \beta_j$ are constant coefficients satisfying the usual conditions for stationarity (Tsay, 2005). The information set $\mathcal{F}_t$ consists of all the past observations of z up to time t. The ARIMA-GARCH model in (2.2) is capable of describing heteroscedastic volatility. To compare the results with a homoscedastic model, we also consider a plain ARIMA$(p, 1, q)$ model for z_t with constant conditional variance $\text{Var}[\varepsilon_t | \mathcal{F}_{t-1}] = \sigma^2_{\varepsilon;t} = \sigma^2_{\varepsilon}$. We select the models by minimizing the Bayesian Information Criteria (BIC). Parameters are estimated by maximizing the Gaussian likelihood of the transformed data.

[1] We have also considered modeling z_t by ARMA(p, q)-GARCH(r, s) models, but they are not selected based on the BIC values.

2.1.3 Multi-step ahead forecasts

For $h = 1, 2, \ldots$, the optimal h-step ahead forecasts can be easily obtained by iterating the model in (2.2). More details on optimal forecasts could be found in Appendices A.3 and A.4. For instance, the h-step ahead mean forecast, $\hat{z}_{t+h|t}$, is given by (A.14). The h-step ahead variance forecast for the differenced series $W_{t+h|t}$, $\hat{\sigma}^2_{\varepsilon;t+h|t}$, is given by (A.25). Note that this is not to be confused with the variance forecast for $Z_{t+h|t}$, i.e. $\text{Var}[z_{t+h}|\mathcal{F}_t] = \hat{\sigma}^2_{t+h|t}$, given in (A.26). In fact, one could obtain the variance forecast for $Z_{t+h|t}$ from $\{\hat{\sigma}^2_{\varepsilon;t+j|t}\}_{j=1}^{h}$ by expressing the model in a moving average (MA) representation (Tsay, 2005). With the mean and the variance forecasts, the corresponding h-step ahead probability forecast for Z_{t+h} is then given by the Gaussian distribution

$$f_{Z_{t+h|t}} \sim N(\hat{z}_{t+h|t}, \hat{\sigma}^2_{t+h|t}) \tag{2.3}$$

2.1.4 Inverse transformation

Finally, recall that our aim is to forecast the aggregated wind power Y_t, which is given by inversely transforming the forecast distribution for Z_t through the relation

$$y_t = \frac{\exp(z_t)}{1 + \exp(z_t)} \tag{2.4}$$

To restore the probability density of Y_{t+h}, we compute the Jacobian of the transformation in (2.1) to get

$$|J| = \left|\frac{dz}{dy}\right| = \frac{1}{y(1-y)} \tag{2.5}$$

The probability density of Y_{t+h} is then given by

$$\begin{aligned} f_{Y_{t+h|t}}(y_{t+h}) &= |J| f_{Z_{t+h|t}}(z_{t+h}) \\ &= \frac{1}{y_{t+h}(1-y_{t+h})} \phi_{\hat{z}_{t+h|t}, \hat{\sigma}^2_{t+h|t}} \left(\log\left(\frac{y_{t+h}}{1-y_{t+h}} \right) \right) \\ &= \frac{1}{y_{t+h}(1-y_{t+h})} \frac{1}{\sqrt{2\pi\hat{\sigma}^2_{t+h|t}}} \exp\left[\frac{-\left(\log\left(\frac{y_{t+h}}{1-y_{t+h}} \right) - \hat{z}_{t+h|t} \right)^2}{2\hat{\sigma}^2_{t+h|t}} \right] \end{aligned} \tag{2.6}$$

where ϕ_{μ,σ^2} is the density of Gaussian distribution with mean μ and variance σ^2. In other words, (2.6) is the h-step ahead conditional density of Y_{t+h} given the conditional point forecast of $\hat{z}_{t+h|t}$ at time t.

2.2 Exponential smoothing and parametric distributions

In our second approach for aggregated wind power forecasts, we deal with the normalized aggregated wind power y_t directly and do not immediately apply any transformations to the data[1]. Instead, we consider smoothing the level and variance of the time series y_t, and estimate the location and scale parameters of the wind power distribution using the smoothed time series. We apply the exponential smoothing methods and consider two similar cases as in the first approach with the Gaussian models.

The reason that we advocate exponential smoothing methods is that they can be easily implemented as an automated forecast algorithm. The estimation of smoothing parameters are very efficient and stable, and they generate extremely competitive and robust out-of-sample forecasts despite their simplicity. In wind power forecasts, it is common that frequent online updating of a large number of forecasts at different wind farms are required. This situation is well fitted to exponential smoothing, and so we believe that it is an approach that worth studying in detail. Apart from the exponential smoothing methods, it would also be interesting to investigate other alternatives such as optimal adaptive filters based on Kalman form. They would require a larger amount of computational power due to recursive optimization of parameters, but it could be beneficial if the data is highly non-stationary. Indeed, we find that the aggregated wind power could be regarded as quasi-stationary within a time scale of about 10 days. For more details please refer to Appendix E.3.

[1] One could replace the ARIMA-GARCH models in Section 2.1 with exponential smoothing methods, and apply the same technique of logit transformation. We find that this generates poorer out-of-sample forecasts compared with our approach described here. Exponential smoothing is very robust and it is better to apply it directly to the data without transformations. For more details please refer to Appendix E.2.

In the first case, we only smooth through the location parameter and do not smooth through the scale parameter. It means that we assume the scale parameter in the parametric wind power distribution to be constant. Note that the scale parameter is not to be confused with the actual variance of the wind power. In fact, we will see that under this approach, assuming a constant scale parameter do not correspond to a constant variance due to the choice of the parametric distribution for wind power. In the second case, we simultaneously smooth through both the location parameter and the scale parameter. We then proceed to obtain the dynamics of these parameters according to the smoothing equations, which are essential for multi-step ahead forecasts. Once the h-step ahead location and scale parameters are estimated, the final step is to determine a parametric distribution for wind power, which depends on these two parameters.

2.2.1 Exponential smoothing

Exponential smoothing methods have been widely and successfully adopted in areas such as inventory forecasting (Brown & Meyer, 1961), electricity forecasting (Taylor, 2003) and volatility forecasting (Taylor, 2004). A comprehensive review of exponential smoothing is given by Gardner (2006). Hyndman *et al.* (2008) provide a state space framework for exponential smoothing , which further strengthens its value as a statistical model instead of an *ad hoc* forecasting procedure. In the following, we focus on the basic version of exponential smoothing, which is also called the simple exponential smoothing as this targets at smoothing the level. We will not consider smoothing the trends and the seasonalities, and the interested reader could refer to the above literature for more details on this topic.

2.2.1.1 Smoothing the location parameter only

In the first case, we consider smoothing only the location parameter, denoted by ℓ_t. It means that we consider the scale parameter of the data to be constant. The smoothing equation for the location parameter implies that the location parameter follows an ARIMA(0,1,1) process (see Appendix B.1). This is reasonable as we have seen that the wind power data are non-stationary. By simple exponential

smoothing, the smoothed series of the location parameter ℓ_t is given by S_t, which is updated according to

$$S_t = \alpha y_t + (1 - \alpha) S_{t-1} \tag{2.7}$$

where y_t is the observed wind power at time t and $0 < \alpha < 1$ is a smoothing parameter. We initialize the series by setting $S_1 = y_1$[1]., and the one-step ahead forecast is $\hat{\ell}_{t+1|t} = S_t$. Iterating (2.7) gives $\hat{\ell}_{t+h|t} = S_t$. However, the forecast errors $y_t - \hat{\ell}_{t|t-1}$ are in general highly correlated, with a significant lag one sample autocorrelation up to about 0.25. A simple way to improve the forecast is to add a parameter ϕ_s to account for autocorrelations in the forecast equation (Taylor, 2003). We call this the simple exponential smoothing with error correction. The updating equation is still given by (2.7), but the forecast equation is modified as

$$\hat{\ell}_{t+1|t} = S_t + \phi_s (y_t - S_{t-1}) \tag{2.8}$$

where $|\phi_s| < 1$. Note that it is now possible to obtain negative values for $\hat{\ell}_{t+1|t}$ in (2.8) and in such cases $\hat{\ell}_{t+1|t}$ is obviously not the true conditional mean of wind power. Nevertheless, recall that $\hat{\ell}_{t+1|t}$ essentially serves as the location parameter of a parametric distribution, which could be negative even though the range of the parametric distribution is bounded on $[0, 1]$ for wind power. Following the taxonomy introduced by Hyndman *et al.* (2008), we denote (2.7) and (2.8) as the ETS$(A, N, N|EC)$ method, where ETS stands for both an abbreviation for exponential smoothing as well as an acronym for error, trend and seasonality respectively. The A inside the bracket stands for additive errors in the model, the first N stands for no trend, the second N stands for no seasonality and EC stands for error correction. By directly iterating (2.7) and (2.8) and substituting

[1]The smoothing procedure is insensitive to the initial value due to the usual sizes of our data sets, which contain over 10^4 observations. However, the initial value should be chosen more carefully when one considers a small data set of only hundreds of observations

$\hat{y}_{t+j|t} = \hat{\ell}_{t+j|t} = \hat{S}_{t+j-1|t} + \phi_s(\hat{y}_{t+j-1|t} - \hat{S}_{t+j-2|t})$[1], we have

$$\hat{\ell}_{t+h|t} = S_t + \frac{\alpha\phi_s(1-\phi_s^{h-1})}{1-\phi_s}(y_t - S_{t-1}) + \phi_s^h(y_t - S_{t-1}) \tag{2.9}$$

for $h > 1$.

Recall that in this case, we assume that the scale parameter is constant and do not smooth through it. Nevertheless, we still need to obtain h-step ahead forecasts for the scale parameters $\hat{s}^2_{t+h|t}$. To achieve this, we need to identify an underlying model corresponding to our updating and forecast equations (2.7) and (2.8), such that we know how the dynamics of the errors propagate. Ledolter & Box (1978) show that exponential smoothing methods produce optimal point forecasts if and only if the underlying data generating process is within a subclass of ARIMA(p, d, q) processes. It can be checked that the ETS$(A, N, N|EC)$ method is optimal for the ARIMA(1,1,1) model, in the sense that the forecasts in (2.8) are the minimum mean square error (MMSE) forecasts according to the ARIMA(1,1,1) model. Some related discussions could be found in Appendix B.1.

Expressed in the form of an ARIMA(1,1,1) model with Gaussian innovations, the ETS$(A, N, N|EC)$ method can be written as

$$w_t = \phi_s w_{t-1} + \varepsilon_t + (\alpha - 1)\varepsilon_{t-1}, \qquad \varepsilon_t \overset{i.i.d.}{\sim} N(0, s_\epsilon^2) \tag{2.10}$$

where $w_t = y_t - y_{t-1}$, ε_t is the Gaussian innovation with mean zero and constant variance s_ϵ^2, and α, ϕ_s are the smoothing parameters in (2.7) and (2.8). It can also be verified that (2.9) is identical to the h-step ahead forecasts obtained from the ARIMA(1,1,1) model in (2.10). It then follows from the ARIMA(1,1,1) model that the h-step ahead forecast variance is given by

$$\hat{s}^2_{t+h|t} = \hat{s}^2_\epsilon \sum_{j=1}^{h} \Omega^2_{h-j} \tag{2.11}$$

where $\Omega_0 = 1, \Omega_k = \phi_s^k + \alpha(1-\phi_s^k)/(1-\phi_s)$ for $k \geq 1$, and $\hat{s}^2_\varepsilon$ is the estimated constant variance of the innovations. This is given in (B.18) in Appendix B.

[1]Note that we take $\hat{y}_{t+j|t} = \hat{\ell}_{t+j|t}$ at this stage so as to simplify the iterations. In fact, $\hat{y}_{t+j|t}$ depends on the specific parametric distributions $f(y|\ell, s^2)$ that we choose.

2.2.1.2 Smoothing both the location and scale parameters

Next, we consider that the scale parameter follows a heteroscedastic process, as this can be helpful in explaining the changing variances in wind power as shown in Figure 1.6(b). In this case, apart from smoothing the location parameter ℓ_t, we also simultaneously smooth the scale parameter s_t^2. In fact, we smooth the variance of innovations $s_{\varepsilon;t}^2$ and obtain the scale parameter s_t^2 for the wind power distribution Y_t as a function of $s_{\varepsilon;t}^2$, similar to the expression in (2.11) for the homoscedastic case.

Equipped with the one-step ahead forecast of the location parameter $\hat{\ell}_{t|t-1}$, we may calculate the squared difference between $\hat{\ell}_{t|t-1}$ and the observed wind power y_t, that is, $(y_t - \hat{\ell}_{t|t-1})^2$, as the estimated scale parameter $s_{\varepsilon;t}^2$ at time t. Applying simple exponential smoothing, denote the smoothed series of $s_{\varepsilon;t}^2$ by V_t, which is updated according to

$$V_t = \gamma(y_t - \hat{\ell}_{t|t-1})^2 + (1-\gamma)V_{t-1} \tag{2.12}$$

where y_t is the observed wind power at time t, $\hat{\ell}_{t|t-1}$ is obtained by (2.8) and $0 < \gamma < 1$ is a smoothing parameter. We initialize the series by setting V_1 to be the variance of the data in the training set[1]. The one-step ahead forecast is given by $\hat{s}_{\varepsilon;t+1|t}^2 = V_t$. Again, the forecast errors are highly correlated and it is better to include an additional parameter ϕ_v in the forecast equation to account for autocorrelations. The modified forecast equation is then given by

$$\hat{s}_{\varepsilon;t+1|t}^2 = V_t + \phi_v \left[(y_t - \hat{\ell}_{t|t-1})^2 - V_{t-1}\right] \tag{2.13}$$

where $|\phi_v| < 1$.

Unfortunately, a major drawback of introducing the second term in the forecast equation (2.13) is that negative values of $\hat{s}_{\varepsilon;t+1|t}^2$ may occur[2]. As a result, we modify our approach and consider smoothing the logarithmic transformed scale parameter $\log s_{\varepsilon;t}^2$ such that positive values of $s_{\varepsilon;t}^2$ are guaranteed. The smoothed

[1] In fact, the forecasts are not sensitive to the choice of initial values due to the size of the data set.

[2] Nevertheless, in the two data sets that we analyzed in Chapters 5 and 6, we do not actually encounter the problem of negative $\hat{s}_{\varepsilon;t+1|t}^2$.

series for $\log s^2_{\varepsilon;t}$ is then given by $\log V_t$. Denoting forecast errors as $\varepsilon_t = y_t - \hat{\ell}_{t|t-1}$ and the standardized forecast errors as $e_t = \varepsilon_t/\sqrt{V_t}$, the estimated logarithmic variance at time t is now chosen to be $g(e_t)$ so that

$$g(e_t) = \theta\left(|e_t| - \mathrm{E}[|e_t|]\right) \tag{2.14}$$

where $\theta > 0$ is a constant parameter. This ensures that $g(e_t)$ is positive for large values of e_t and negative if e_t is small, which is equivalent to saying that the next-period variance is larger if e_t ia greater than $\mathrm{E}[|e_t|]$ (See (2.15) in the following). As we assume Gaussian innovations, e_t follows a standard normal distribution and so we have $\mathrm{E}[|e_t|] = \sqrt{2/\pi}$ (Hamilton, 1994).

The updating equation and the forecast equation are now written as

$$\begin{aligned} \log V_t &= \gamma g(e_t) + (1-\gamma)\log V_{t-1} \\ \log \hat{s}^2_{\varepsilon;t+1|t} &= \log V_t + \phi_v\left[g(e_t) - \log V_{t-1}\right] \end{aligned} \tag{2.15}$$

which are analogous to (2.12) and (2.13), except that a logarithmic transformation is taken and $(y_t - \hat{\ell}_{t|t-1})^2$ is replaced by $g(e_t)$. We initialize the series by setting $\log V_1 = 0$[1].

Now, let us summarize our exponential smoothing method for both ℓ_t and s^2_t by combining (2.7), (2.8) and (2.15):

$$\begin{aligned} S_t &= \alpha y_t + (1-\alpha)S_{t-1} \\ \hat{\ell}_{t+1|t} &= S_t + \phi_s(y_t - S_{t-1}) \\ \log V_t &= \gamma g(e_t) + (1-\gamma)\log V_{t-1} \\ \log \hat{s}^2_{\varepsilon;t+1|t} &= \log V_t + \phi_v\left[g(e_t) - \log V_{t-1}\right] \end{aligned} \tag{2.16}$$

where $g(e_t)$ is given in (2.14) and $e_t = (y_t - \hat{\ell}_{t|t-1})/\sqrt{V_t}$. There are four smoothing parameters $\alpha, \gamma, \phi_s, \phi_v$ and a parameter θ for the estimated logarithmic variance $g(e_t)$. We adopt the taxonomy similar to that for simple exponential smoothing with error correction as described in Section 2.2.1.1, and denote (2.16) as the

[1]In fact, the smoothing procedure is insensitive to the initial value due to the large size of our data sets, which contain over 10^4 observations. However, the initial value should be chosen more carefully when one considers a small data set of only hundreds of observations

ETS$(A,N,N|EC)$-$(A,N,N|EC)$ method where the second bracket indicates the exponential smoothing method applied on the scale parameter.

The h-step ahead forecasts for the location parameters $\hat{\ell}_{t+h|t}$ can still be obtained using (2.9). Again, for the purpose of generating h-step ahead forecasts for the scale parameters $\hat{s}^2_{t+h|t}$, we need to identify an underlying model for this smoothing approach. Analogous to matching the ARIMA(1,1,1) model with the ETS$(A,N,N|EC)$ method, we aim at identifying the ETS$(A,N,N|EC)$-$(A,N,N|EC)$ method in (2.16) with an ARIMA-GARCH model. Using (2.10) as the ARIMA(1,1,1) model for y_t and writing $\varepsilon_t = y_t - \hat{\ell}_{t|t-1}$, the last equation in (2.16) can be written as

$$\begin{aligned}
\log \hat{s}^2_{\varepsilon;t+1|t} &= \log V_t + \phi_v[g(e_t) - \log V_{t-1}] \\
&= \gamma g(e_t) + (1-\gamma)\log V_{t-1} + \phi_v[g(e_t) - \log V_{t-1}] \\
&= (\gamma+\phi_v)g(e_t) - \phi_v \log V_{t-1} \\
&\quad +(1-\gamma)\{\log s^2_{\varepsilon;t} - \phi_v[g(e_{t-1}) - \log V_{t-2}]\} \\
&= (\gamma+\phi_v)g(e_t) - \phi_v g(e_{t-1}) + (1-\gamma)\log s^2_{\varepsilon;t} \qquad (2.17)
\end{aligned}$$

where we have used the updating equation in (2.15). This is an exponential GARCH model, i.e. the EGARCH(2,1) model[1] for the conditional variance of innovations ε_t (Nelson, 1991).

In summary, the ETS$(A,N,N|EC)$-$(A,N,N|EC)$ method in (2.16) is optimal for the ARIMA(1,1,1)-EGARCH(2,1) model, which can be written as

$$\begin{aligned}
w_t &= \phi_s w_{t-1} + \varepsilon_t + (\alpha-1)\varepsilon_{t-1}, \qquad \varepsilon_t|\mathcal{F}_{t-1} \overset{i.i.d.}{\sim} N(0, s^2_{\varepsilon;t}) \\
\log s^2_{\varepsilon;t} &= (1-\gamma)\log s^2_{\varepsilon;t-1} + (\gamma+\phi_v)g(e_{t-1}) - \phi_v g(e_{t-2}) \qquad (2.18)
\end{aligned}$$

where $w_t = y_t - y_{t-1}$ and $g(e_t)$ is given in (2.14). Now, equipped with the ARIMA(1,1,1)-EGARCH(2,1) model in (2.18), the h-step ahead forecasts for the scale parameter $\hat{s}^2_{\varepsilon;t+h|t}$ can be easily obtained (Tsay, 2005). Consequently, the

[1] In the conventional EGARCH models for asset prices, $g(e_t) = \theta_0 e_t + \theta\,(|e_t| - \mathrm{E}[|e_t|])$ is asymmetric in e_t so as to account for the increase in volatility when prices drop. In our case, we choose the symmetric version of $g(e_t)$ with $\theta_0 = 0$ since there is no reason to expect volatility to increase when wind power generation drops.

h-step ahead forecasts $\hat{s}^2_{t+h|t}$ can be expressed as a function of $\{\hat{s}^2_{\varepsilon;t+j|t}\}_{j=1}^{h}$, which is analogous to (2.11).

2.2.2 Mapping to parametric distributions

In this section, we would like to use a parametric distribution to describe the wind power distributions. In particular, this parametric distribution will depend on the location parameter ℓ_t and scale parameter s_t^2. Before we describe this in more details, let us consider the situation when the density of Y_t follows a distribution f that depends on the conditional mean and conditional variance of the distribution. For instance, the one-step ahead probability density is written as $f_{t+1|t}(y; \hat{\mu}_{t+1|t}, \hat{\sigma}^2_{t+1|t})$ where $\hat{\mu}_{t+1|t} = \mathrm{E}[y_{t+1}|\mathcal{F}_t]$ is the conditional mean and $\hat{\sigma}^2_{t+1|t} = \mathrm{Var}[y_{t+1}|\mathcal{F}_t] = \mathrm{Var}[\varepsilon_{t+1}|\mathcal{F}_t] = \hat{\sigma}^2_{\varepsilon;t+1|t}$ is the conditional variance[1].

Given any $f_{t+1|t}$ and a model M for the dynamics, we can evolve the density function to obtain the h-step ahead density so that

$$\begin{aligned} f_{t+1|t}(y; \hat{\mu}_{t+1|t}, \hat{\sigma}^2_{t+1|t}) \quad &\xrightarrow{M} \quad f_{t+h|t}(y; \hat{\mu}_{t+h|t}, \hat{\sigma}^2_{t+h|t}) \\ \hat{\mu}_{t+h|t} \quad &= \quad p_M^{(h)}(\hat{\mu}_{t+1|t}, \ldots, \hat{\mu}_{t+h-1|t}; y_1, \ldots, y_t) \\ \hat{\sigma}^2_{t+h|t} \quad &= \quad q_M^{(h)}(\hat{\sigma}^2_{\varepsilon;t+1|t}, \ldots, \hat{\sigma}^2_{\varepsilon;t+h|t}) \end{aligned} \tag{2.19}$$

where $\xrightarrow{M}$ denotes the process of evolving the dynamics and generating h-step ahead probability forecasts under a certain model M. This could be achieved by Monte Carlo simulations in practice. Here $p_M^{(h)}$ and $q_M^{(h)}$ stand for functions that give the conditional mean and the conditional variance of y_t, with parameters that depend on the model M and the forecast horizon h.

It is difficult to obtain any closed form for $f_{t+h|t}$ if the distribution of innovations ε_t is non-Gaussian. Thus we propose to use a two-step approach to approximate $f_{t+h|t}$. In the first step, we attempt to model the dynamics of the conditional mean $\hat{\ell}_{t+h|t}$ and the conditional variance $\hat{s}^2_{t+h|t}$ of the data using a Gaussian model G. We label them as $\hat{\ell}_{t+h|t}$ and $\hat{s}^2_{t+h|t}$ (instead of $\hat{\mu}_{t+h|t}$ and

[1]Note that $\hat{\sigma}^2_{t+h|t}$ denotes the conditional variance of the data y_{t+h}, while $\hat{\sigma}^2_{\varepsilon;t+h|t}$ denotes the conditional variance of the innovation ε_{t+h}, so that in general $\hat{\sigma}^2_{t+h|t}$ is a function of $\hat{\sigma}^2_{\varepsilon;t+j|t}$ with $j = 1, \ldots, h$.

$\hat{\sigma}^2_{t+h|t}$) to remind us that they may not be the true conditional mean and conditional variance respectively. In fact, they serve only as proxies and are actually the location and scale parameters of the final distribution. The first step is then expressed as

$$\begin{aligned} \textit{Step 1:} \qquad \hat{\ell}_{t+h|t} &= p_G^{(h)}(\hat{\ell}_{t+1|t}, \ldots, \hat{\ell}_{t+h-1|t}; y_1, \ldots, y_t) \\ \hat{s}^2_{t+h|t} &= q_G^{(h)}(\hat{s}^2_{\varepsilon;t+1|t}, \ldots, \hat{s}^2_{\varepsilon;t+h|t}) \end{aligned} \tag{2.20}$$

where $p_G^{(h)}$ and $q_G^{(h)}$ stand for functions that give the conditional mean and the conditional variance of y_{t+h}, with parameters that depend on the Gaussian model G and horizon h. In model G, the innovations are additive and are assumed to be i.i.d. Gaussian distributed. For example, G can be the conventional ARIMA-GARCH model with Gaussian innovations. Recall that in an ARIMA(1,1,1) model, the h-step ahead variance is given by (2.11), which is just the explicit form of (2.20) with constant $\hat{s}^2_\epsilon$ since

$$\hat{s}^2_{t+h|t} = \hat{s}^2_\epsilon \sum_{j=1}^{h} \Omega^2_{h-j} = q_G^{(h)}(\hat{s}^2_\epsilon) \tag{2.21}$$

As Gaussianity could be violated in reality, $\hat{\ell}_{t+h|t}$ and $\hat{s}^2_{t+h|t}$ obtained from model G may not be the true conditional mean $\hat{\mu}_{t+h|t}$ and conditional variance $\hat{\sigma}^2_{t+h|t}$ respectively. They only serve as proxies to the true values and are actually the location and scale parameters of the final distribution.

Although model G may not describe real situations, we rely on a second step for remedial adjustments such that the final probabilistic forecast is an approximation to reality. In the second step, we assume that the h-step ahead density $f_{t+h|t}$ can be approximated by a parametric function D, which is characterized by a location parameter and a scale parameter. In particular, the location parameter and the scale parameter are obtained from the proxy for conditional mean $\hat{\ell}_{t+h|t}$ and the proxy for conditional variance $\hat{s}^2_{t+h|t}$ respectively, which are estimated using the Gaussian model G. Thus in the second step, we take

$$\textit{Step 2:} \qquad f_{t+h|t}(y; \hat{\mu}_{t+h|t}, \hat{\sigma}^2_{t+h|t}) \approx D(y; \hat{\ell}_{t+h|t}, \hat{s}^2_{t+h|t}) \tag{2.22}$$

as the h-step ahead probabilistic forecast where D is a function depending on two parameters only. As a result, the two-step approach may be able to give a good estimation of $f_{t+h|t}$ if (2.22) is a close approximation. In (2.22), the true conditional mean $\hat{\mu}_{t+h|t}$ and conditional variance $\hat{\sigma}^2_{t+h|t}$ are generated by $p^{(h)}_M(\cdot)$ and $q^{(h)}_M(\cdot)$ under the true model M, while the corresponding proxy values $\hat{\ell}_{t+h|t}$ and $\hat{s}^2_{t+h|t}$ are generated by $p^{(h)}_G(\cdot)$ and $q^{(h)}_G(\cdot)$ under a Gaussian model G. Empirical studies will be needed to determine the appropriate Gaussian model G as well as the best choice D in order to approximate the final density $f_{t+h|t}$.

2.2.2.1 Truncated normal distribution

For normalized aggregated wind power y_t, a potential candidate for the parametric distribution D is the truncated normal distribution bounded within $[0, 1]$. This is a nice assumption because when the scale parameter s^2 is small (for instance, when wind is stable and our forecast horizon is only one-step ahead), then the truncations at 0 and 1 do not really affect the distribution, and the errors just behave like normally distributed. However, when the scale parameter s^2 is large, the truncations at 0 and 1 ensure that the errors are bounded and the forecast distribution lies on the range $[0, 1]$. Truncated normal distributions have also been applied successfully in modeling bounded, non-negative data in some other literature (Gneiting *et al.*, 2006; Sanso & Guenni, 1999). Other interesting candidates for the parametric distribution D are the logit normal distributions and the generalized logit normal distributions (Pinson, 2010). However, the generalized logit normal distributions depend on an extra shape parameter ν and slight modifications of our approach are needed. This would be an interesting extension to be investigated.

In the following, let us consider D to be the truncated normal distribution governed by the location parameter ℓ and the scale parameter s^2, with lower and upper truncations at 0 and 1 respectively. The corresponding normal distribution without truncations would then be $N(\ell, s^2)$. If the data is indeed Gaussian, the true conditional mean would be ℓ, and the conditional variance would be s^2. As in (2.22), we approximate the density function by the truncated normal distribution

D, and so the h-step ahead forecast density is obtained as

$$\begin{aligned} f_{t+h|t}(y;\hat{\mu}_{t+h|t},\hat{\sigma}^2_{t+h|t}) &\approx D(y;\hat{\ell}_{t+h|t},\hat{s}^2_{t+h|t}) \\ &= \frac{1}{\hat{s}_{t+h|t}} \frac{\varphi\left(\frac{y-\hat{\ell}_{t+h|t}}{\hat{s}_{t+h|t}}\right)}{\Phi\left(\frac{1-\hat{\ell}_{t+h|t}}{\hat{s}_{t+h|t}}\right) - \Phi\left(\frac{-\hat{\ell}_{t+h|t}}{\hat{s}_{t+h|t}}\right)} \end{aligned} \tag{2.23}$$

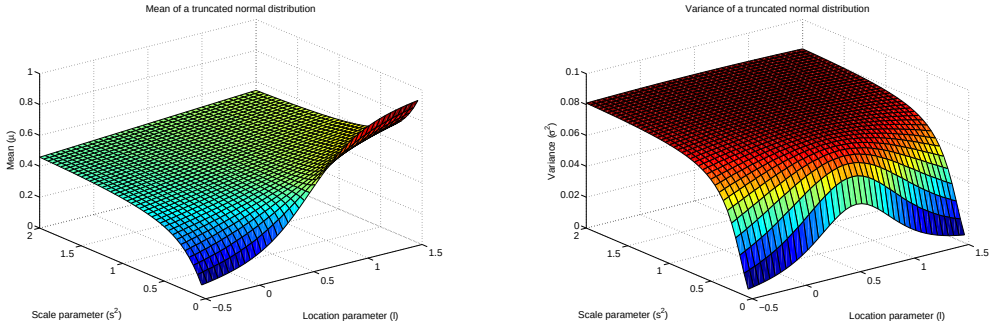
for $y \in [0,1]$, where φ and Φ are the standard normal density and distribution function respectively. It can be verified that the mean μ and the variance σ^2 of a truncated normal distribution $f(y|\ell, s^2)$, with truncations at 0 and 1, are given by (Olsen, 1980)

$$\mu = \ell + s\left(\frac{\varphi\left(\frac{-\ell}{s}\right) - \varphi\left(\frac{1-\ell}{s}\right)}{\Phi\left(\frac{1-\ell}{s}\right) - \Phi\left(\frac{-\ell}{s}\right)}\right) \tag{2.24}$$

$$\sigma^2 = s^2\left[1 + \frac{\frac{-\ell}{s}\varphi\left(\frac{-\ell}{s}\right) - \frac{1-\ell}{s}\varphi\left(\frac{1-\ell}{s}\right)}{\Phi\left(\frac{1-\ell}{s}\right) - \Phi\left(\frac{-\ell}{s}\right)} - \left(\frac{\varphi\left(\frac{-\ell}{s}\right) - \varphi\left(\frac{1-\ell}{s}\right)}{\Phi\left(\frac{1-\ell}{s}\right) - \Phi\left(\frac{-\ell}{s}\right)}\right)^2\right] \tag{2.25}$$

The dependence of the mean μ and the variance σ^2 of the truncated normal distribution (with truncations at 0 and 1) are shown in Figures 2.2(a) and 2.2(b) respectively. We see that when the scale parameter s^2 is large, then the truncated normal distribution simply converges to the uniform distribution $U[0,1]$. As a result, the mean converges to $1/2$ and the variance converges to $1/12$ for large values of s^2. In Figure 2.2(b), we also see an advantage of modeling wind power with the truncated normal distribution. It is because when the scale parameter s^2 is small and close to zero, the variance σ^2 is highly dependent on the location parameter ℓ. As seen from Figure 2.2(a), when the location parameter gives a mean which is close to zero or one, the wind power distribution has a low variance. On the other hand, when the mean is close to 0.5, i.e. in the middle of the range of wind power, the variance can be significantly larger (Lange, 2005). The advantage here is that this dependence of variance on the mean helps to explain the uncertainty in wind power through the power curve, i.e .the curve that relates wind speed and wind power. In the power curve, the cut-off region that corresponds to low wind power and the saturation region that corresponds

to high wind power both give a relatively low variance due to the flat shapes. However, for the rated region where the turbines provide their rated power, the power curve is cubic. This nonlinear transformation between wind speed and wind power increases forecast errors and hence the variance of the forecast distribution, as illustrated in Figure 2.3.



(a) This figure shows the dependence of the mean of a truncated normal distribution $f(y|\ell, s^2)$ on location parameters ℓ and scale parameters s^2. The distribution has lower and upper truncations at 0 and 1 respectively. For very large scale parameters, the mean converges to $\mu = 1/2$ as the distribution converges to the uniform distribution.

(b) This figure shows the dependence of the variance of a truncated normal distribution $f(y|\ell, s^2)$ on location parameters ℓ and scale parameters s^2. The distribution has lower and upper truncations at 0 and 1 respectively. For very large scale parameters, the variance converges to $\sigma^2 = 1/12 \approx 0.08$ as the distribution converges to the uniform distribution.

Figure 2.2: Mean and variance of truncated normal distribution.

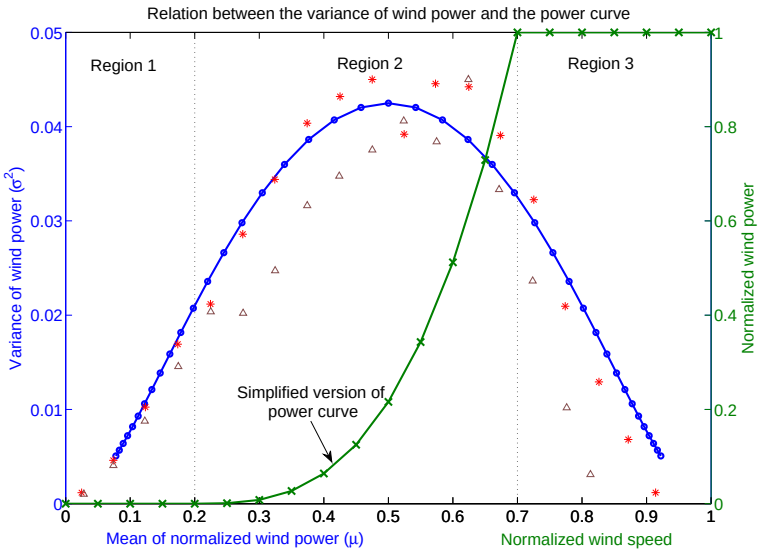


Figure 2.3: The dependence of variance on mean helps to explain the uncertainty in wind power through the power curve. Here we fix the scale parameter to be $s^2 = 0.05$, and calculate the mean and variance of the truncated normal distribution with different values of location parameters as shown by the blue line with circles. A simplified version of power curve is plotted as the green line with crosses. Region 1 is the cut-off region that corresponds to low wind speeds, and Region 3 is the saturation region that corresponds to high wind speeds. They both give a relatively low variance in wind power due to the flat shapes of the power curve. Region 2 is the rated region where the turbines provide their rated power and the power curve is cubic. This nonlinear transformation between wind speed and wind power increases forecast errors and hence the variance of the forecast distribution. This feature is verified using the Irish wind power. We estimate the conditional mean and conditional variance of normalized wind power by a moving average with window size of 6 hours. We then calculate the mean of the mean and mean of the variances, conditional on the mean wind power. Results obtained using the Irish wind power and the Danish wind power are shown as brown triangles and red asterisks respectively. Note that the empirical variances are rescaled to clarify this shape, as they do not necessarily lie on the blue line with fixed scale parameter $s^2 = 0.05$.

In general, given a forecast density, one could obtain the mean by numerical integration. For the case of truncated normal distribution, we could also use (2.24) to calculate the expected value of wind power. For instance, the h-step ahead point forecast would be given by

$$\hat{y}_{t+h|t} = \hat{\ell}_{t+h|t} - \hat{s}_{t+h|t} \left(\frac{\varphi\left(\frac{1-\hat{\ell}_{t+h|t}}{\hat{s}_{t+h|t}}\right) - \varphi\left(\frac{-\hat{\ell}_{t+h|t}}{\hat{s}_{t+h|t}}\right)}{\Phi\left(\frac{1-\hat{\ell}_{t+h|t}}{\hat{s}_{t+h|t}}\right) - \Phi\left(\frac{-\hat{\ell}_{t+h|t}}{\hat{s}_{t+h|t}}\right)} \right) \tag{2.26}$$

where $\hat{\ell}_{t+h|t}$ and $\hat{s}^2_{t+h|t}$ are the location and scale parameters of the truncated normal distribution in (2.23). Note that due to the truncation, the distribution may not be symmetric and so the expected value is in general different from the location parameter, that is, $\hat{y}_{t+h|t} \neq \hat{\ell}_{t+h|t}$. In fact, referring to (2.20), $\hat{\ell}_{t+h|t} = p_G^{(h)}(\hat{\ell}_{t+1|t}, \ldots, \hat{\ell}_{t+h-1|t}; y_1, \ldots, y_t)$ is obtained according to a Gaussian model G, which may not give the true conditional mean $\hat{y}_{t+h|t}$ of the data, and may even be negative. Since the final density $f_{t+h|t}$ is only obtained when an appropriate function D is chosen, we see that D transforms the conditional mean from $\hat{\ell}_{t+h|t}$ for Gaussian data to the optimal forecast $\hat{y}_{t+h|t}$ for our data. This is analogous to calculating optimal point forecasts when the loss function is asymmetric (Christoffersen & Diebold, 1997; Patton & Timmermann, 2007). Since the normalized aggregated wind power is bounded within $[0, 1]$, the loss function is always asymmetric unless the conditional mean is $\hat{\ell}_{t+h|t} = 0.5$. When the conditional mean is not the optimal forecast, an additional term is added to compensate for the asymmetric loss. Christoffersen & Diebold (1997) suggest an approximation to calculate the optimal forecast for conditionally Gaussian data by assuming $\hat{y}_{t+h|t} = G(\mu_{t+h|t}, \sigma^2_{t+h|t})$, where $\mu_{t+h|t}, \sigma^2_{t+h|t}$ are the conditional mean and conditional variance. Their method involves expanding G into a Taylor series.

2.2.3 Estimation of smoothing parameters

Since maximum likelihood estimators are well known to have nice asymptotic properties, we estimate all smoothing parameters as mentioned in Sections 2.2.1.1 and 2.2.1.2 by maximizing the likelihood of the truncated normal distribution $f_{t+1|t}(y_{t+1}; \hat{\ell}_{t+1|t}, \hat{s}^2_{t+1|t})$. For instance, in the ETS$(A, N, N|EC)$ method, we obtain the location parameter $\hat{\ell}_{t+h|t}$ from (2.9) and the scale parameter $\hat{s}^2_{t+h|t}$ from (2.11), and then generate the h-step ahead probability forecasts as the truncated normal distribution $f_{t+h|t}$ as given in (2.23).

Chapter 3

Spatiotemporal Kriging and Correlation Models

3.1 Spatial kriging

In geostatistics, obtaining accurate spatial prediction for physical variables, such as the deposition of minerals in soil, is one of the major problems. One of the pioneer works was done by Krige (1951), who proposed empirical methods to predict ore-grade distributions. Later on, Matheron (1963) developed the method of obtaining optimal spatial linear predictors and named the method as kriging. In other words, the kriging predictor is simply the common name for the optimal linear predictor for a spatial process $Z(\mathbf{s})$ at spatial locations $\mathbf{s}$. It is a linear combination of the observations $z(\mathbf{s_i})$ for some $\mathbf{s_i}$'s in the neighborhood of $\mathbf{s}$. It is optimal since the predictor has the minimum mean squared prediction errors among all linear predictors. If $Z(\mathbf{s})$ is a Gaussian process, then the optimal linear predictor coincides with the optimal predictor. If $Z(\mathbf{s})$ is not Gaussian, the kriging predictor is only guaranteed to be optimal in the class of linear predictors. Cressie (1993) provides an excellent overview of the theory of kriging, which also gives details on many variations on the topic, such as indicator kriging, trans-Gaussian kriging and Bayesian kriging.

3.2 Spatiotemporal kriging

Due to the analytical tractability of Gaussian models, in many statistical applications, one assumes that the process is Gaussian. If the data is obviously non-Gaussian, then various techniques such as transformations would be applied to normalize the data before proceeding to the Gaussian model. This also generalizes to spatiotemporal processes, where the most common treatment is to assume Gaussianity. Under this assumption, one of the approaches to generate spatiotemporal forecasts is to obtain the kriging predictor as in the case for spatial predictions in geostatistics, except that one includes an additional temporal dimension. As discussed in Chapter 1, the continuous part of wind power distribution is approximately Gaussian under the modified logit transformation in (1.1). This partly justifies the assumption of Gaussianity. Of course, there is still a problem with the transformation as it could not accommodate the discrete probability masses at zero and one. Nevertheless, due to the relative efficiency in obtaining the kriging predictors, the method of spatiotemporal kriging is attractive. In this chapter, we will describe in details on how to generate spatiotemporal forecasts using the kriging approach. To handle the problem of discrete probability masses in wind power distributions, we will further consider another more sophisticated latent Gaussian model in Chapter 4.

3.2.1 Optimal linear predictor

In this section, we review some of the theories in kriging, and give the details of extending the spatial kriging predictor for the case of spatiotemporal predictions. Suppose that we are interested in modeling the spatiotemporal process $Z(\mathbf{s}, t)$ at m fixed locations $\mathbf{s}_1, \ldots, \mathbf{s}_m$, and $t = 1, 2, \ldots$ is a discrete temporal index. We write $Z(\mathbf{s}_j, t) = Z_j(t)$ and consider the vector $\mathbf{Z}(t) = (Z_1(t), \ldots, Z_m(t))^T$. One may regard $\mathbf{Z}(t)$ as a multivariate time series. For current time t, the target is to obtain forecasts of $\mathbf{Z}(t+h)$ at a future time $t+h$ where h is the forecast horizon. Of course, this forecast is conditional on the history of observations, denoted by $\Omega_t = (\mathbf{Z}(1), \ldots, \mathbf{Z}(t))^T$. For horizons $h = 1, 2, \ldots$, the forecasts $\hat{\mathbf{Z}}(t+h)$ could be obtained as the optimal linear predictor. It is linear and thus we could express it

in terms of a linear combination of the history of observations $\mathbf{Z}(k)$ where $k \leq t$. We denote this optimal linear predictor by $p(Z(\mathbf{s}))$, which is given by

$$p(\mathbf{Z}(t+h)) = K + \sum_{k=1}^{t} \Gamma^{k,h}\mathbf{Z}(k) \tag{3.1}$$

where $K = (K_1, \ldots, K_m)^T$ is a constant and $\Gamma^{k,h}$ is an $m \times m$ matrix. We denote the ij^{th} element of $\Gamma^{k,h}$ by $\lambda_{ij}^{k,h}$. This means that the historical observation $Z_i(k)$ contributes to the prediction of $Z_j(t+h)$ through the coefficient $\lambda_{ij}^{k,h}$ (Myers, 1982).

As the optimal linear predictor is optimal in the sense that it minimizes the mean squared prediction errors, let us consider a squared error loss function $L(\mathbf{Z}(t+h), p(\mathbf{Z}(t+h))) = \sum_{j=1}^{m}[Z_j(t+h) - p(Z_j(t+h))]^2$. In other words, the optimal predictor $p(\mathbf{Z}(t+h))$ minimizes the expected loss $\mathrm{E}[L]$. As in the case when $\mathbf{Z}(t)$ is univariate, the expected squared error loss can be decomposed into the sum of the variance and the bias of prediction (Cressie, 1993). To ensure that the predictor is unbiased, we set

$$K = \boldsymbol{\mu} - \sum_{k=1}^{t} \Gamma^{k,h}\boldsymbol{\mu} \tag{3.2}$$

where we assume that $\mathbf{Z}(t)$ is stationary and $\mathrm{E}[\mathbf{Z}(t)] = (\mu_1, \ldots, \mu_m)^T = \boldsymbol{\mu}$. The coefficient matrices $\Gamma^{k,h}$ are then obtained by solving the systems of equations (Myers, 1982)

$$\frac{\partial \Phi}{\partial \lambda_{ij}^{k,h}} = 0 \qquad k = 1, \ldots, t; \quad i, j = 1, \ldots, m \tag{3.3}$$

where $\Phi = \mathrm{E}[L(\mathbf{Z}(t+h), p(\mathbf{Z}(t+h)))]$ is the expected loss function. Analogous to the case in simple kriging, the solutions $\boldsymbol{\Gamma}^h = (\Gamma^{1,h}, \ldots, \Gamma^{t,h})^T$ are given by (Cressie, 1993)

$$\boldsymbol{\Gamma}^h = \boldsymbol{\Sigma}^{-1}\mathbf{c} \tag{3.4}$$

where $\boldsymbol{\Sigma}$ and $\mathbf{c}$ are simply some covariance matrices.

Writing (3.4) explicitly, we have

$$\begin{pmatrix} \Gamma^{1,h} \\ \vdots \\ \Gamma^{t,h} \end{pmatrix}_{(tm)\times m} = \begin{pmatrix} \mathcal{C}(1,1) & \cdots & \mathcal{C}(1,t) \\ \vdots & \ddots & \vdots \\ \mathcal{C}(t,1) & \cdots & \mathcal{C}(t,t) \end{pmatrix}^{-1}_{(tm)\times(tm)} \begin{pmatrix} \mathcal{C}(1,(t+h)) \\ \vdots \\ \mathcal{C}(t,(t+h)) \end{pmatrix}_{(tm)\times m} \tag{3.5}$$

where $\mathcal{C}(\cdot,\cdot)$'s are $m \times m$ covariance matrices in the form

$$\mathcal{C}(t_i,t_j) = \begin{pmatrix} C_{11}(t_i,t_j) & \cdots & C_{1m}(t_i,t_j) \\ \vdots & \ddots & \vdots \\ C_{m1}(t_i,t_j) & \cdots & C_{mm}(t_i,t_j) \end{pmatrix}_{m\times m} \tag{3.6}$$

Here $C_{pq}(t_i,t_j) = \mathrm{Cov}(Z_p(t_i), Z_q(t_j))$ is the covariance between $Z_p(t_i)$ and $Z_q(t_j)$. Note that $\boldsymbol{\Sigma}$ in (3.4) is always symmetric. However, in general, each block matrix $\mathcal{C}(t_i,t_j)$ may not be symmetric, and we only have $\mathcal{C}(t_j,t_i) = \mathcal{C}^T(t_i,t_j)$.

Finally, substituting (3.2) and (3.4) into the expression of the optimal linear predictor in (3.1), we have

$$p(\mathbf{Z}(t+h)) = \boldsymbol{\mu} + \mathbf{c}^T\boldsymbol{\Sigma}^{-1}(\Omega_t - \boldsymbol{\mu}) \tag{3.7}$$

where $(\Omega_t - \boldsymbol{\mu})$ represents $(\mathbf{Z}(1) - \boldsymbol{\mu}, \ldots, \mathbf{Z}(t) - \boldsymbol{\mu})^T$. The corresponding kriging variance is

$$\sigma_h^2 = \mathrm{Tr}\{\mathcal{C}(t+h,t+h)\} - \mathrm{Tr}\left\{\mathbf{c}^T\boldsymbol{\Sigma}^{-1}\mathbf{c}\right\} = \mathrm{Tr}\{V\} \tag{3.8}$$

where $V = \mathcal{C}(t+h,t+h) - \mathbf{c}^T\boldsymbol{\Sigma}^{-1}\mathbf{c}$. This is analogous to the case in simple kriging. Note that (3.8) is the total summation of kriging variance for each component in $\mathbf{Z}(t+h)$. The variance of the j^{th} component could be written as $\sigma_{h,j}^2 = V_{jj}$ (Myers, 1982).

3.3 Correlation structures of wind power data

Wind power is of course driven by wind, which is simply the movement of air. In general, wind blows from high pressure to low pressure regions. Its speed and direction depend on a lot of factors such as the geometry of the surrounding landscape, the location of the site (e.g. coastal, inland, etc) and the season within a year. In other words, wind power is a random variable that is largely driven by a physical process, i.e. wind, which in turn is largely determined by geographical properties. As a result, one would expect to observe some typical correlation structures in an arbitrary wind power data set.

3.3.1 Autocorrelation

Before we consider the more complicated spatiotemporal correlation structure of a portfolio of wind power, we first look into the autocorrelation of a wind power time series at a single wind farm. As we have seen in Chapter 1, wind power time series are highly autocorrelated, and the decay of autocorrelation is so slow that it resembles a long memory process (Haslett & Raftery, 1989). A typical plot of autocorrelation is given in Figure 3.1. Note that the corresponding plot for aggregated wind power gives an even slower decay in autocorrelation, as shown in Figure 1.6(c).

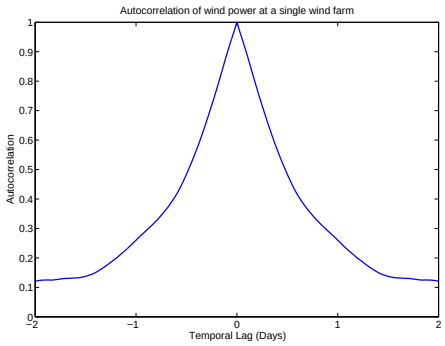


Figure 3.1: This is the plot of autocorrelation function of wind power at a single wind farm. The particular wind farm chosen in this example is wind farm number 1 in the Danish data in Table 6.1. We see that the decay of autocorrelations is very slow, and is still over 0.1 at a temporal lag of 2 days (i.e. 192 temporal steps in the 15-minute data).

3.3.2 Spatial correlation

Next, we consider how the correlation between different wind farms decay with spatial distances. Figure 3.2 shows a plot using the wind data from 64 wind farms in Ireland (Table 5.1), where the temporal lag is zero. It is observed that spatial correlations are quite strong, and contemporary wind power at different sites are correlated even when they are separated by over 300 km. Also, there seems to be a nugget effect at spatial lag $\mathbf{h} = \mathbf{0}$, i.e. a discontinuous jump of correlation to unity when $\mathbf{h}$ approaches zero. This could due to measurement errors as well as the fact that we do not have observations at such fine resolutions. As wind power is driven by wind speed, we also look into the spatial correlations of hourly wind speed data at 12 stations in Ireland to support our results. This is the Haslett & Raftery (1989) wind data[1], and plots of correlations with distances are shown in Figure 3.3[2]. We see that the rate of decay of spatial correlations are comparable to that observed in wind power.

The decay of spatial correlations could also depend on the locations of wind farms. For instance, Figure 3.4 shows a corresponding plot using the wind data from 49 Danish wind farms (Table 6.1). As the landscape in Denmark is relatively smooth and we consider a longer time series (2 years), the spatial correlations are less disperse. Denmark is a smaller country and the wind farms are closer to each other. As a result, we do not have observations at spatial lag beyond 300 km. Nevertheless, we use the same scale as in Figure 3.2 to enhance comparisons.

[1]This data set is freely downloadable from the internet at `http://lib.stat.cmu.edu/datasets/`. The latitudes and longitudes of the stations could be obtained from Table 2 in Gneiting *et al.* (2007b).

[2]This is the same as Figure 2 in Gneiting *et al.* (2007b).

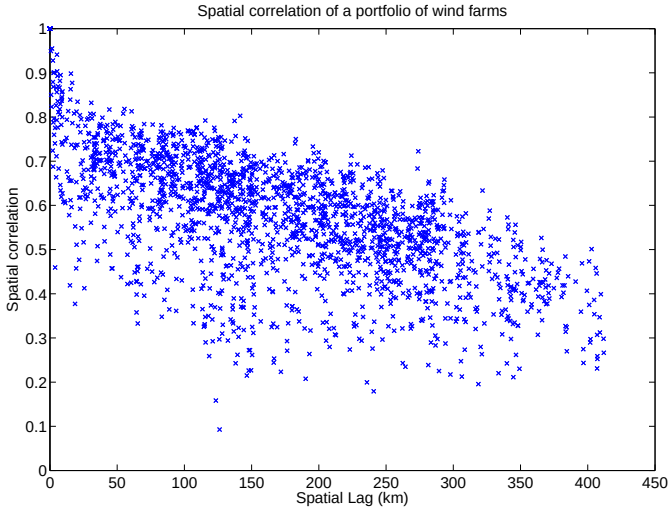


Figure 3.2: The figure shows the spatial correlations of wind power generated at 64 wind farms in Ireland (Table 5.1), where the temporal lag is zero. It is observed that spatial correlations are quite strong, and contemporary wind power at different sites are correlated even when they are separated by over 300 km. Also, there seems to be a nugget effect at spatial lag $\mathbf{h} = \mathbf{0}$. This could due to measurement errors as well as the fact that we do not have observations at that small resolution.

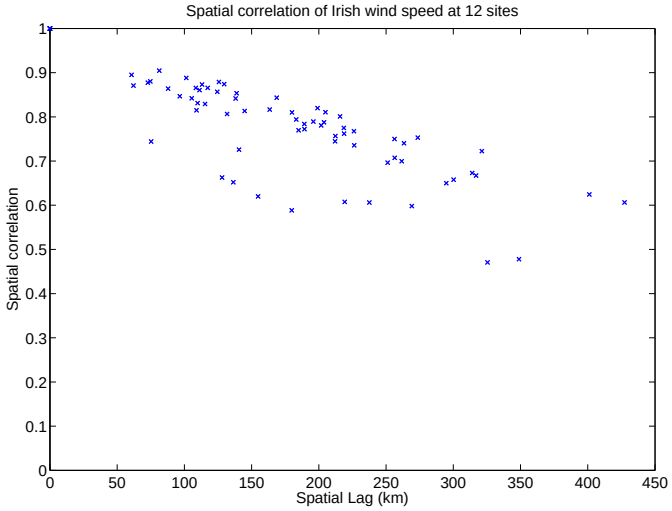


Figure 3.3: The figure shows the spatial correlations of wind speed at 12 stations in Ireland. These are the Haslett & Raftery (1989) hourly wind speed data which are freely downloadable from the internet (See footnote 1 on Page 50). We see that the rate of decay of spatial correlations are comparable to that observed in wind power.

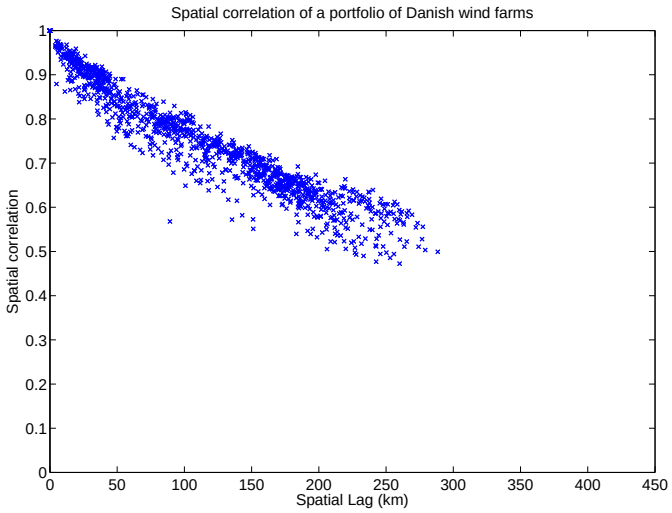


Figure 3.4: The figure shows the spatial correlations of wind power generated at 49 wind farms in Denmark (Table 6.1), where the temporal lag is zero. As the landscape in Denmark is relatively smooth and we consider a longer time series (2 years), the spatial correlations are less disperse. Denmark is a smaller country and the wind farms are closer to each other. As a result, we do not have observations at spatial lag beyond 300 km. Nevertheless, we use the same scale as in Figure 3.2 to enhance comparisons.

3.3.3 Spatiotemporal correlation

After considering both pure temporal and pure spatial correlations, we proceed to look into the spatiotemporal cross correlations between wind power generated at different wind farms. These correlations attain more interesting features. They are also more important in terms of spatiotemporal forecasts because they contain most of the information within the whole spatiotemporal correlation function.

As mentioned above, wind power is driven by wind and is largely affected by a number of physical factors. These physical factors are collectively represented and reflected in the general weather pattern that occurs within the particular region. Similar to the existence of prevailing ocean currents, we also observe prevailing wind patterns in most places on Earth. These prevailing wind patterns are due to the movement of the weather front across the region, which usually propagates along a certain directions with possible seasonal fluctuations throughout the year.

For example, the wind pattern in Ireland is predominantly westerly (Fox *et al.*, 2007, P. 144). This means that the current wind power generated at a wind farm located on the west coast would be highly correlated to that generated at a wind farm located on the east coast at a certain future time. This correlation is induced from the movement of the weather front, and this leads to anisotropy in the spatiotemporal correlation structure because space is no longer homogeneous in all directions. For instance, let us consider 15 sites in the Irish wind data with locations as shown in Figure 3.5. This demonstrate a significant dependence of spatiotemporal correlations on the orientation of the spatial lag $\mathbf{h} = \mathbf{s}_i - \mathbf{s}_j$. Wind farm number 1 is situated to the west or north-west of the other 14 wind farms, and the cross correlations between its wind power generation and those from the other farms show a clear pattern on their spatial locations. In particular, their correlations attain maximum at a temporal lag $u = t_i - t_j \neq 0$, on contrary to the symmetrical case when maximum correlations are observed at $t_i = t_j$, i.e. $u = 0$. In the case of Irish wind data, for temporal lag $u > 0$, we have

$$\text{Corr}(Z(\mathbf{s}_p, t), Z(\mathbf{s}_q, t+u)) > \text{Corr}(Z(\mathbf{s}_p, t), Z(\mathbf{s}_q, t-u)) \tag{3.9}$$

where $\mathbf{s}_p$ is located to the west or north-west of $\mathbf{s}_q$. Physically, this means that if $\mathbf{s}_p$ is located to the west of $\mathbf{s}_q$, then the wind power generated at $\mathbf{s}_p$ at time t_i is more correlated to the wind power generated at $\mathbf{s}_q$ at a later time $t_j = t_i + u$. In other words, wind power generated at $\mathbf{s}_p$ is driving wind power generated at $\mathbf{s}_q$.

Table 3.1 shows the correlation of wind power generation between wind farm number 1 and the other 14 wind farms in Figure 3.5. The temporal lags are at $u = t_i - t_j = \pm 24$, that is, ± 6 hours. Clearly, wind power generated at wind farm number 1 is driving the others, and the correlations at negative temporal lags are larger due to the time needed for the propagation of weather front from the north-west to the south-east. Figure 3.6(a) shows a typical plot of the cross correlations between wind power generated at two different wind farms. In this example, the two wind farms are chosen to be wind farm number 1 and 55 in Figure 3.5, and so we observe that they lie along the general direction of wind propagation in Ireland. As a result, there is a significant shift in the correlation plot, and the correlations no longer attain maximum at equal time. On the other

hand, Figure 3.6(b) shows the corresponding plot between wind power generated at wind farm number 22 and 39 in Figure 3.5. In this case, we do not see an obvious anisotropy because the two wind farms do not lie along the prevailing direction of propagating weather fronts.

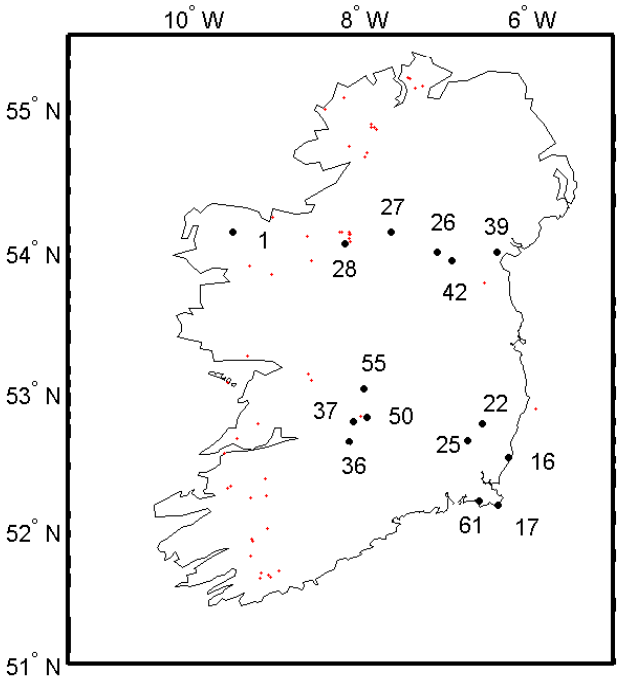
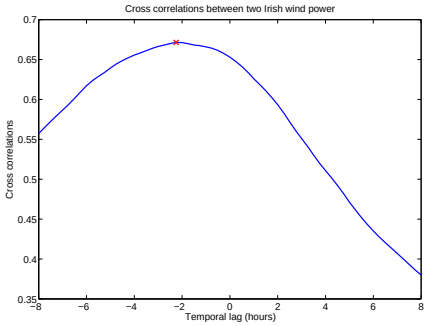


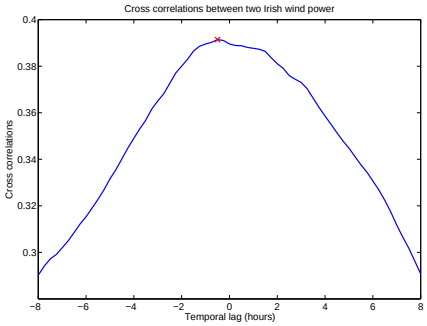
Figure 3.5: Locations of 15 wind farms in Ireland denoted by black circles. These are the wind farms that we consider in Table 3.1, which compares the correlations of the wind farms at temporal lags $u = \pm 24$, i.e. ± 6 hours. Wind power generation at wind farm no. 1 drives the wind power generation at other wind farms due to the westerly movement of the weather front, which results in a prevailing wind direction from the west. The wind farms are numbered according to Table 5.1. The tiny red dots indicate the locations of other wind farms in Table 5.1.

Farm No. j	$\|\mathbf{s}_1 - \mathbf{s}_j\|$ **(km)**	$\theta(\mathbf{s}_1, \mathbf{s}_j)(^\circ)$	$K(\mathbf{s}_1, \mathbf{s}_j, 24)$	$K(\mathbf{s}_1, \mathbf{s}_j, -24)$
16	135.3	51.9	0.29	0.50
17	90.5	90.8	0.35	0.57
22	133.5	39.6	0.32	0.56
25	183.9	52.4	0.31	0.53
26	275.2	174.1	0.43	0.59
27	179.1	48.8	0.35	0.54
28	120.8	59.2	0.49	0.57
36	98.4	173.4	0.28	0.42
37	91.3	94.4	0.43	0.56
39	133.8	52.6	0.16	0.42
42	272.9	175.3	0.40	0.61
50	202.9	181.2	0.19	0.40
55	200.9	180.5	0.44	0.62
61	276.7	120.3	0.38	0.59

Table 3.1: Correlations between wind power generation of wind farm number 1 and those lying to the east of it. The locations of these wind farms are shown in Figure 3.5. $\|\mathbf{s}_1 - \mathbf{s}_j\|$ is the great circle distance between locations $\mathbf{s}_1$ and $\mathbf{s}_j$. $\theta(\mathbf{s}_1, \mathbf{s}_j)$ is the azimuth angle, i.e., the angle of $\mathbf{s}_j$ as measured from $\mathbf{s}_1$ which increases in a clockwise direction from the north. The temporal lags are at $u = t_i - t_j = \pm 24$, that is, ± 6 hours. We denote $\mathrm{Corr}(Z(\mathbf{s}_p, t), Z(\mathbf{s}_q, t-u))$ in (3.9) by $K(\mathbf{s}_p, \mathbf{s}_q, u)$. Wind power generated at wind farm number 1 is driving the other wind farms since the correlations at negative temporal lags, $K(\mathbf{s}_1, \mathbf{s}_j, -24)$, are significantly larger than $K(\mathbf{s}_1, \mathbf{s}_j, 24)$. This means that wind power generated at wind farm number 1 at time t is more correlated with wind power generated at other wind farms 6 hours later, i.e. at time $t + 24$, because wind in Ireland generally propagates from west to east. The wind farms are numbered according to Table 5.1.



(a) This Figure shows the cross correlations between wind power generated at wind farm number 1 and 55 in Figure 3.5. We observe that they lie along the general direction of wind propagation in Ireland, and so there is a significant shift in the correlation plot. Also, the correlations no longer attain maximum (indicated by red cross) at equal time. In general, the further apart are the wind farms, the larger is the shift of the location of maximum correlation from zero.



(b) This Figure shows the cross correlations between wind power generated at wind farm number 22 and 39 in Figure 3.5. In this case, we do not see an obvious anisotropy because the two wind farms do not lie along the general direction of the propagation of weather front. The correlations attain maximum (indicated by red cross) at approximately equal time.

Figure 3.6: Cross correlations between Irish wind power.

3.4 Covariance models

In this section, we provide a review on the various properties of covariance models as well as a brief discussion on their constructions. This would equip us with the necessary background before we move on to study some classes of covariance models, and build our own covariance model for wind data.

3.4.1 Properties of covariance models

3.4.1.1 Positive definiteness, stationarity and full-symmetry

Consider a spatiotemporal process $Z(\mathbf{s}, t)$ with a covariance function $C(\mathbf{s}_i, t_i; \mathbf{s}_j, t_j)$= $\text{Cov}(Z(\mathbf{s}_i, t_i), Z(\mathbf{s}_j, t_j))$. Let $\tilde{Z}(\mathbf{s}, t)$ be any linear combination of $Z(\mathbf{s}, t)$ at k locations $\mathbf{s}_j$ and time t_j, $j = 1, \ldots, k$. In other words, we have $\tilde{Z} = \sum_{j=1}^{k} a_j Z(\mathbf{s}_j, t_j)$. Since the variance of $\tilde{Z}$ must be non-negative, we have

$$\text{Var}(\tilde{Z}) = \sum_{i=1}^{k} \sum_{j=1}^{k} a_i a_j C(\mathbf{s}_i, t_i; \mathbf{s}_j, t_j) \geq 0 \tag{3.10}$$

Any valid covariance function must satisfy the condition in (3.10), which is called positive semi-definiteness. In other words, a valid covariance model $\tilde{C}$ must be positive semi-definite[1]. Bochner (1955) gives the theorem stating that a continuous function C on $\mathbb{R}^d$ is positive definite if and only if the Fourier transform of C is a non-negative finite measure on $\mathbb{R}^d$. Cressie & Huang (1999) propose a method to construct valid classes of covariance models based on Bochner's theorem by considering spectral densities. Gneiting (2002b) proposes another simpler approach which directly builds a valid covariance model in the space-time domain.

A spatiotemporal covariance function $C(\mathbf{s}_i, t_i; \mathbf{s}_j, t_j)$ is stationary in space and time if it only depends on the relative lags, i.e.

$$\begin{aligned} C(\mathbf{s}_i, t_i; \mathbf{s}_j, t_j) &= C(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j) \\ &= C(\mathbf{h}, u) \end{aligned} \tag{3.11}$$

[1] We will use the word positive definite in the followings as this is simply a stronger condition where (3.10) is strictly positive.

where $\mathbf{h} = \mathbf{s}_i - \mathbf{s}_j$ is the spatial lag and $u = t_i - t_j$ is the temporal lag. If $C(\mathbf{h}, u)$ is fully symmetric, it further satisfies (Gneiting *et al.*, 2007b)

$$C(\mathbf{h}, u) = C(-\mathbf{h}, u) = C(\mathbf{h}, -u) = C(-\mathbf{h}, -u) \tag{3.12}$$

Note that a covariance must be symmetric, and so we always have $C(\mathbf{h}, u) = C(-\mathbf{h}, -u)$ in (3.12).

3.4.1.2 Properties of wind power covariances

For physical data such as wind power, the assumption of stationarity in (3.11) is usually invalid, especially in terms of spatial stationarity. For example, one would expect a larger variability in wind power generation at some windy coastal locations, while the variance at flat plateaus or offshore sites should be smaller in general. Nevertheless, one could decompose the covariance function into a product of variances and correlation function, i.e.

$$C(\mathbf{s}_i, t_i; \mathbf{s}_j, t_j) = \sigma(\mathbf{s}_i, t_i)\sigma(\mathbf{s}_j, t_j)K(\mathbf{s}_i, t_i; \mathbf{s}_j, t_j) \tag{3.13}$$

where $\sigma^2(\mathbf{s}_j, t_j)$ is the variance of $Z(\mathbf{s}_j, t_j)$ at location $\mathbf{s}_j$ and time t_j. We could then separately model the variances of individual wind power $\sigma^2(\mathbf{s}_j, t_j)$ as well as the correlation function $K(\mathbf{s}_i, t_i; \mathbf{s}_j, t_j)$. In general, it would be more justifiable to assume a stationary correlation function for wind power, i.e.

$$K(\mathbf{s}_i, t_i; \mathbf{s}_j, t_j) = K(\mathbf{h}, u) \tag{3.14}$$

where $\mathbf{h} = \mathbf{s}_i - \mathbf{s}_j$ and $u = t_i - t_j$. As long as we have a relatively stable weather pattern drives the wind resources, we may assume a stationary correlation function in (3.14). For the non-stationary variances $\sigma^2(\mathbf{s}, t)$, we could then describe their dynamics using a univariate model. On the other hand, we could further simplify the situation and assume that $\sigma^2(\mathbf{s}, t) = \sigma^2(\mathbf{s})$, i.e. the variance depends only on location $\mathbf{s}$ but are constant through time. In any case, modeling the variances would be a relatively straightforward task, and we would describe more about this in Section 3.6.3. The main challenge then boils down to be the modeling of the stationary spatiotemporal correlation function $K(\mathbf{h}, u)$.

In our modeling of wind power data, we only assume stationarity of the correlation functions $K(\mathbf{h}, u)$, and we could not assume full-symmetry as in (3.12). As described in (3.9), wind power generation is in general driven by the weather pattern, and one would expect there exists anisotropy in space as a result of the movement of the weather front. This will violate the full-symmetric property, and in general we have

$$K(\mathbf{h}, u) \neq K(\mathbf{h}, -u) \tag{3.15}$$

Moreover, because of spatial anisotropy, we also have

$$K(\mathbf{h}, u) \neq K(\|\mathbf{h}\|, |u|) \tag{3.16}$$

where $\|\cdot\|$ denotes the Euclidean distance. In other words, correlation functions of wind power depend on the direction between the wind farms as well as their mutual distance. (Gneiting *et al.*, 2007b) show that a non fully-symmetric correlation function must also be nonseparable. This means that

$$K(\mathbf{h}, u) \neq K_{\mathbf{s}}(\mathbf{h}) K_t(u) \tag{3.17}$$

where $K_{\mathbf{s}}(\mathbf{h})$ is a purely spatial correlation function, and $K_t(u)$ is a purely temporal correlation function.

3.4.1.3 Nugget effect

Apart from the above properties of covariance functions, there is another feature that one should be aware. This feature is known as the nugget effect (Cressie, 1993), which could be observed from the plot of empirical correlations. A nugget effect is simply a discontinuity in the correlation function. Gneiting *et al.* (2007b) classify the nugget effect into three cases, namely purely spatial, purely temporal, or spatiotemporal, depending on the location of discontinuity. In most cases, one considers a purely spatial nugget effect so that the correlation model is written as

$$K(\mathbf{h}, u) = K_{ST}(\mathbf{h}, u) + \delta_{\mathbf{h}=0} K_T(u) \tag{3.18}$$

where $\delta_{\mathbf{h}=0}$ is the indicator function. The extra term $\delta_{\mathbf{h}=0}K_T(u)$ in (3.18) means that there is an additional kink in the correlation at zero spatial lag, i.e. a discontinuity at $\mathbf{h} = 0$. The nugget effect could exist due to measurement errors or small scale variability, which cannot be observed due to the limited resolution in the data (Cressie, 1993).

In the following section, we start by investigating some classes of spatiotemporal correlation models that have been studied in the literature. In particular, some of them have been applied to model the Haslett & Raftery (1989) wind speed data in Ireland, while some have been used to model wind speed on the Pacific Ocean. We then proceed to construct our spatiotemporal correlation model for wind power data.

We consider spatiotemporal correlation structures of the data and attempt to build a model that could explain the features. Table 3.2 summarizes the main properties that we would consider in the construction of our spatiotemporal correlation model, which will be described in details in Section 3.5.2. Note that the first two properties, positive definiteness and symmetry, must always be satisfied by any covariance models.

Property	**Equation**	**Our Model**
Positive definiteness	$\sum_{i=1}^{k}\sum_{j=1}^{k} a_i a_j K(\mathbf{s}_i, t_i; \mathbf{s}_j, t_j) > 0$	Yes
Symmetry	$K(\mathbf{s}_i, t_i; \mathbf{s}_j, t_j) = K(\mathbf{s}_j, t_j; \mathbf{s}_i, t_i)$	Yes
Stationarity	$K(\mathbf{s}_i, t_i; \mathbf{s}_j, t_j) = K((\mathbf{s}_i - \mathbf{s}_j, t_i - t_j))$	Yes
Separability	$K(\mathbf{s}_i, t_i; \mathbf{s}_j, t_j) = K_{\mathbf{s}}(\mathbf{s}_i, \mathbf{s}_j)K_t(t_i, t_j)$	No
Isotropy	$K(\mathbf{s}_i, t_i; \mathbf{s}_j, t_j) = K(\lVert\mathbf{s_i} - \mathbf{s_j}\rVert, \lvert t_i - t_j\rvert)$	No
Full symmetry	$K(\mathbf{s}_i, t_i; \mathbf{s}_j, t_j) = K(\mathbf{s}_i, t_j; \mathbf{s}_j, t_i)$	No
Nugget effect	$\lim_{\mathbf{s}_i \to \mathbf{s}_j} K(\mathbf{s}_i, t_i; \mathbf{s}_j, t_j) < K(\mathbf{s}_j, t_i; \mathbf{s}_j, t_j)$	Yes

Table 3.2: Summary of the properties of a correlation model $K(\mathbf{s}_i, t_i; \mathbf{s}_j, t_j)$. An equation is valid if the corresponding property is satisfied by the model. The third column 'Our Model' indicates whether or not our correlation model for wind power generation, to be described in Section 3.5.2 satisfies the corresponding property.

3.4.2 Construction of correlation models

In this section we describe in general how valid spatiotemporal correlation models could be constructed. The major challenge is to ensure that the model is positive definite, thus representing a valid correlation function. We only consider the case where the correlation model is stationary and satisfies (3.14). There are some classes of well-known correlation models, for instance, the exponential, power-exponential, spherical, and the Cauchy models (Cressie, 1993). The main idea of constructing valid correlation models lie on the useful fact that appropriate combinations of valid correlation models are also itself a valid model. For instance, we could pick two valid correlation models $K_1(\mathbf{h}, u)$ and $K_2(\mathbf{h}, u)$, and construct a new valid model $K(\mathbf{h}, u)$ by

$$K(\mathbf{h}, u) = K_1(\mathbf{h}, u)K_2(\mathbf{h}, u) \qquad \text{or} \qquad K(\mathbf{h}, u) = a_1K_1(\mathbf{h}, u)+a_2K_2(\mathbf{h}, u) \quad (3.19)$$

where $a_1, a_2 > 0$. In other words, sum or product of positive definite functions are also positive definite, provided that the coefficients are non-negative (Ma, 2005). In general, as we have to satisfy the constraint $K(\mathbf{0}, 0) = 1$, only convex combinations of valid correlations are valid. For N valid correlation models $K_j(\mathbf{h}, u)$, $j = 1, \ldots, N$, the convex combination

$$K(\mathbf{h}, u) = \sum_{j=1}^{N} a_j K_j(\mathbf{h}, u) \quad (3.20)$$

where $a_j \geq 0$, $j = 1, \ldots, N$ and $\sum_{j=1}^{N} a_j = 1$ gives a valid correlation model. Using these sum and product properties, one could start by considering some simplest models as the basic building blocks.

3.4.3 Construction of anisotropic correlation models

Based on a valid isotropic spatiotemporal correlation model $K(\mathbf{h}, u) = K(\|\mathbf{h}\|, |u|)$, a simplest way to obtain an anisotropic version is to consider an appropriate transformation for the coordinates (Ma, 2003). This is similar to adopting a moving Lagrangian reference frame (Gneiting *et al.*, 2007b; May & Julien, 1998), which is attached to the center of mass of the moving object (e.g. Weather front). In such

case, we may express the spatiotemporal correlation function $K(\mathbf{h}, u)$ in terms of a purely spatial correlation function $K_S(\tilde{\mathbf{h}})$ so that

$$\begin{aligned} K(\mathbf{h}, u) &= K_S(\tilde{\mathbf{h}}) \\ &= K_S(\mathbf{h} - \mathbf{v}u) \end{aligned} \tag{3.21}$$

where $\mathbf{v}$ (in the same dimension as $\mathbf{h}$) is the velocity of the reference frame. In particular, if K_S is an isotropic correlation function, then we have

$$K(\mathbf{h}, u) = K_S(\|\mathbf{h} - \mathbf{v}u\|) \tag{3.22}$$

We will apply (3.22) in Section 3.5 to account for spatial anisotropy in the wind power data due to the movement of the weather front, and we assume that the weather front propagates at a constant velocity $\mathbf{v}$.

More generally, one may replace the constant velocity in (3.21) by a distribution of velocities denoted by the random variable $\mathbf{V}$. By the property in (3.19) one would obtain a valid spatiotemporal correlation model by taking the expectation with respect to the distribution of $\mathbf{V}$ to get

$$K(\mathbf{h}, u) = \mathrm{E}_{\mathbf{V}}\left[K_S(\mathbf{h} - \mathbf{V}u)\right] \tag{3.23}$$

Another related concept is the Taylor's hypothesis (Gneiting *et al.*, 2007b; Taylor, 1938), which considers a specific relationship between the purely spatial and purely temporal correlation functions corresponding to a stationary spatiotemporal correlation function. If $C(\mathbf{h}, u)$ satisfies Taylor's hypothesis, then there exists a velocity vector $\mathbf{v}$ (with the same dimension as $\mathbf{h}$) so that

$$K(\mathbf{0}, u) = K(\mathbf{v}u, 0) \tag{3.24}$$

Intuitively, we could visualize $\tilde{\mathbf{h}} = \mathbf{v}u$ as the distance propagated by the reference frame in time u.

3.4.4 Calculation of distance on a sphere

Before moving on to discuss some examples of spatiotemporal correlation models in the next section, we would like to clarify how we would calculate the distance between two spatial locations on a sphere. This issue is in fact one of the problems faced in the modeling of valid spatiotemporal correlations for environmental data, because in such case the data is usually observed at particular locations on the surface of the Earth. Since the Earth is roughly spherical, one should not simply use the Euclidean measure. A simple and intuitive alternative is to use the great circle distance to calculate the separation between two sites located at $\mathbf{s}_i$ and $\mathbf{s}_j$. Since we denote locations on the surface of the Earth by their latitudes and longitudes, we use the notation $\mathbf{s}_i = (s_i^{lat}, s_i^{lon})$ for all spatial locations in the following analysis[1]. The great circle distance is then obtained by (Stein, 2005)

$$G(\mathbf{s}_i, \mathbf{s}_j) = R_E \cos^{-1}\left[\sin s_i^{lat} \sin s_j^{lat} + \cos s_i^{lat} \cos s_j^{lat} \cos(s_i^{lon} - s_j^{lon})\right] \tag{3.25}$$

where R_E is the radius of the Earth. The distance in (3.25) has the advantage of giving the correct distance on Earth. However, unfortunately this may lead to an invalid correlation model if we simply substitute the Euclidean distance with the great circle distance (Jun & Stein, 2007; Stein, 2005) by substituting

$$\|\mathbf{s}_i - \mathbf{s}_j\| \to G(\mathbf{s}_i, \mathbf{s}_j) \tag{3.26}$$

To ensure that positive definiteness is retained, we consider the method of firstly projecting the 3-dimensional spherical surface onto a 2-dimensional plane, and then calculate the distance using the Euclidean metric. This would result in a valid correlation model, but an obvious disadvantage of this method is that the actual distances will be distorted due to the projection (Stein, 2005). This method would be more appropriate when the sites are relatively close together. In such case, we assume that the coordinates lie on the same latitude, and the projection of $\mathbf{s} = (s^{lat}, s^{lon})$ onto the plane is given by

$$\tilde{\mathbf{s}} = P(\mathbf{s}) = (s^{lat}, s^{lon} \cos \bar{L}) \tag{3.27}$$

[1]We follow Stein (2005) and put the degree of latitude as the first coordinate.

where $\bar{L}$ is the mean of the latitudes of all sites. The distance between two sites is then calculated by the Euclidean metric and we have

$$\|\tilde{\mathbf{h}}\| = \|P(\mathbf{s}_i) - P(\mathbf{s}_j)\| \tag{3.28}$$

A constant may then be multiplied to $\|\tilde{\mathbf{h}}\|$ to convert the unit from distance of one degree arc length on the Earth to kilometers, which is about 111.

As discussed in Section 3.4.3, we may construct an anisotropic correlation model by using the Lagrangian reference frame. In our case of wind data, we assume that the velocity of the weather front is $\mathbf{v} = (v^{lat}, v^{lon}) = (v \sin\theta, v \cos\theta)$, where θ is the angle measured in a clockwise direction from the east. Then, using (3.22), we have

$$\begin{aligned} K(\mathbf{h}, u) &= K_S(\|\tilde{\mathbf{h}} - \tilde{\mathbf{v}}u\|) \\ &= K_S(\|\tilde{\mathbf{s}}_i - (\tilde{\mathbf{s}}_j + \tilde{\mathbf{v}}u)\|) \\ &= K_S(\|P(\mathbf{s}_i) - P(\mathbf{s}_j + \mathbf{v}u)\|) \end{aligned} \tag{3.29}$$

i.e. we first project the original coordinates using P in (3.27) and then calculate the Euclidean distance using $\|\cdot\|$.

3.5 Spatiotemporal correlation models

In this section, we review some classes of spatiotemporal correlation models that have been considered in the literature. We will apply these models in some of the data analysis in Chapter 5 and Chapter 6, and compare the results with our spatiotemporal correlation model to be described in Section 3.5.2.

It should be emphasized that different choices of correlation models imply different assumptions regarding the underlying physical processes that generate the data. This is because of the relationship between stationary solutions of p^{th}-order stochastic differential equations and their spectral densities, which in term relates to the correlation functions through Fourier transforms (Doob, 1944; Rasmussen & Williams, 2006, Appendix B). For example, the Matérn correlation model, which will be discussed later, consists of a parameter ν that controls

the smoothness of the stochastic process. The process is smoother for larger values of ν. The exponential correlation model $C(t) = \exp(-\beta|t|)$ turns out to be equivalent to the Matérn correlation model with $\nu = 1/2$. This corresponds to the Ornstein-Uhlenbeck process that is not mean-square differentiable. If the first differences of the data is not white noise, then the data is at least once mean-square differentiable and the exponential correlation model may not be appropriate. Furthermore, for one-dimensional case, choosing a particular ν such that $\nu + 1/2 = p$ for some integer p gives rise to certain continuous-time AR(p) Gaussian process (Rasmussen & Williams, 2006, Chapter 4). Many physical flow processes are derived from second order SDE (i.e. $p = 2$) and a Matérn correlation model with $\nu = 3/2$ could be appropriate. It also implies that the process is only once mean-square differentiable. Nevertheless, it is common in the literature to keep using the exponential (i.e. $\nu = 1/2$) or squared exponential (i.e. $\nu = \infty$) correlation model instead of a Matérn correlation model because the latter involves evaluation of the Bessel functions and it is difficult to estimate ν.

3.5.1 Examples of spatiotemporal correlation models

In the following, we denote the correlation model by $K(\mathbf{h}, u)$, where $\mathbf{h} = \mathbf{s}_i - \mathbf{s}_j$ is the spatial lag and $u = t_i - t_j$ is the temporal lag. First, let us consider a relatively simple model proposed by Cressie & Huang (1999). They develop classes of spatiotemporal covariance functions by applying Bochner's Theorem (Bochner, 1955) and build the models using Fourier transforms. In the application of the modeling of wind speed over the tropical western Pacific Ocean, they propose an isotropic, nonseparable spatiotemporal correlation model

$$K(\mathbf{h}, u) = \exp(-a|u| - b^2\|\mathbf{h}\|^2 - c|u|\|\mathbf{h}\|^2) \tag{3.30}$$

where $a, b > 0$ and $c \geq 0$ are constant parameters. The correlation model in (3.30) becomes separable for $c = 0$. However, Gneiting (2002b) shows that (3.30) is not a valid correlation model because it is not positive definite, due to some erroneous assumptions in the behavior of the spectral densities. Nevertheless, Example 4 in Cressie & Huang (1999) is an alternative valid choice and we will

consider it here. Denoting $K_{ST}(\mathbf{h}, u)$ as the spatiotemporal component and $K_0(u)$ as the autocorrelation component, it can be expressed as

$$\begin{aligned} & & K(\mathbf{h},u) &= K_{ST}(\mathbf{h},u) + \delta_{\mathbf{h}=\mathbf{0}} K_0(u) \\ \text{where} & & K_{ST}(\mathbf{h},u) &= \frac{c_0(a|u|+1)}{[(a|u|+1)^2 + b^2\|\mathbf{h}\|^2]^{(d+1)/2}}, \\ & & K_0(u) &= (1-c_0)(a|u|+1)^{-d} \end{aligned} \tag{3.31}$$

In (3.31), $a, b > 0$ are the respective temporal and spatial scaling parameters, d is the dimension of $\mathbf{h}$ with $d = 2$ in this case, and $0 \leq c_0 \leq 1$ controls the magnitude of the spatial nugget effect K_0. We see that Cressie's model in (3.31) belongs to the Cauchy family (Gneiting, 2002a).

Next, we consider the class of spatiotemporal correlation models proposed by Gneiting (2002b). As opposed to the approach proposed by Cressie & Huang (1999), which considers the frequency domain and relies on closed form Fourier inversions, Gneiting (2002b) shows that a class of valid spatiotemporal correlation models can be constructed directly in the space-time domain by considering the combinations of certain positive functions with appropriate properties of monotonicity. Gneiting (2002b) also revisits the Irish wind data in Haslett & Raftery (1989) and suggests the isotropic but nonseparable spatiotemporal correlation model (Gneiting, 2002b, Equation (14))

$$\begin{aligned} & & K(\mathbf{h},u) &= K_{ST}(\mathbf{h},u) + \delta_{\mathbf{h}=\mathbf{0}} K_0(u) \\ \text{where} & & K_{ST}(\mathbf{h},u) &= \frac{c_0}{(a|u|^{2\alpha}+1)^{\tau}} \exp\left(\frac{-b\|\mathbf{h}\|^{2\gamma}}{(a|u|^{2\alpha}+1)^{\beta\gamma}}\right), \\ & & K_0(u) &= \frac{(1-c_0)}{(a|u|^{2\alpha}+1)^{\tau}} \end{aligned} \tag{3.32}$$

where $a > 0$ and $\alpha, \gamma \in [0, 1]$ and $\tau \geq \beta d/2$ to ensure positive definiteness. Again, d is the dimension of $\mathbf{h}$ with $d = 2$ in this case. The parameter $\beta \in [0, 1]$ governs the extent of space-time interaction, with $\beta = 0$ corresponding to a separable model. Again, $0 \leq c_0 \leq 1$ controls the magnitude of the spatial nugget term $K_0(u)$ in (3.32). In practice, one may like to fix the value of some parameters, and Gneiting (2002b) considers the special case with $\tau = 1$ and $\gamma = 1/2$. We will

adopt these specific parametric values as well in our spatiotemporal forecast to be discussed in Chapter 5. We see that Gneiting's model in (3.32) is built from the combination of the Cauchy family and the powered exponential family.

As a last example, we consider the class of Matérn correlation models (Matérn, 1986) which is advocated by Stein (2005). Although Matérn correlation models are more difficult to estimate due to the vast amount of computation involved in the evaluation of Bessel functions, they are attractive because they contain a specific parameter which controls the smoothness of the stochastic process. Matérn correlation models are expressed in terms of $\mathcal{K}_\nu$, i.e., the modified Bessel functions of the second kind of order ν. The order ν is a critical parameter which controls the smoothness of the stochastic process, and the process is smoother for larger values of ν. This is because the process is m times mean square differentiable if $\nu > m$ (Diggle & Ribeiro, 2007; Stein, 1999).

Stein (2005) also considers the application of Matérn correlation models on the Haslett & Raftery (1989) Irish wind data. The model is nonseparable, but is more sophisticated than those applied in Cressie & Huang (1999) and Gneiting (2002b) because Stein's model is anisotropic. Stein (2005) considers a transformation of coordinates that incorporates the information of a prevailing wind velocity vector in Ireland. Following Stein (2005), define

$$\mathcal{M}_\nu(r) = r^\nu \mathcal{K}_\nu(r) \tag{3.33}$$

where $\mathcal{K}_\nu$ is the modified Bessel functions of the second kind of order ν. The spatiotemporal correlation model can then be written as (Stein, 2005, Equation (16))

$$\begin{aligned} K(\mathbf{h}, u) &= K_{ST}(\mathbf{h}, u) + \delta_{\mathbf{h}=\mathbf{0}} K_0(u) \\ \text{where} \qquad K_{ST}(\mathbf{h}, u) &= \frac{\mathcal{M}_{\nu+\xi|u|^\gamma}(\alpha\|\tilde{\mathbf{h}}\|)}{2^{\nu+\xi|u|^\gamma}\Gamma(\nu+\xi|u|^\gamma+1)}, \\ K_0(u) &= \frac{\kappa}{\nu' + |u|^\gamma} \end{aligned} \tag{3.34}$$

where $\Gamma(\cdot)$ is the Gamma function $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$. The ranges of parameters are $\xi, \alpha \geq 0$, $\nu, \nu' > 0$ and $\gamma \in (0, 2]$ so as to ensure positive definiteness. The

spatial lag $\tilde{\mathbf{h}}$ is obtained as a transformation of the original spatial vector $\mathbf{h} = \mathbf{s}_i - \mathbf{s}_j$, and is given by

$$\tilde{\mathbf{h}} = \mathbf{h} - \mathbf{v}u \tag{3.35}$$

where $\mathbf{v} = (v \sin\theta, v\cos\theta)$ is the velocity of wind and u is the temporal lag. This is the Lagrangian approach as discussed in Section 3.4.3.

The value of κ in the spatial nugget term in (3.34) is not a free parameter. Instead, it is a function of ν and ν' so as to satisfy the condition that $K(\mathbf{0}, 0) = 1$. This can be seen from the fact that

$$\lim_{\mathbf{h}\to\mathbf{0}} K_{ST}(\mathbf{h}, u) = \frac{1}{2^{\nu+\xi|u|^\gamma}\Gamma(\nu + \xi|u|^\gamma + 1)} \tag{3.36}$$

and so

$$K(\mathbf{0}, u) = \frac{\kappa}{\nu' + |u|^\gamma} + \frac{1}{2^{\nu+\xi|u|^\gamma}\Gamma(\nu + \xi|u|^\gamma + 1)} \tag{3.37}$$

For $u = 0$, we have $K(\mathbf{0}, 0) = 1$ and so we have

$$\kappa = \nu' \left(1 - 2^{-\nu}\Gamma^{-1}(\nu + 1)\right) \tag{3.38}$$

3.5.2 Our anisotropic spatiotemporal correlation model

3.5.2.1 Idea from Lagrangian approach

After discussing various classes of correlation models, we consider building our spatiotemporal correlation model for a general wind power data set. To fit the empirical correlation of wind power with a physically meaningful correlation model, we consider a nonseparable, anisotropic correlation function which is stationary in both space and time, as non-stationarity in space will be captured in the variances instead. Due to the anisotropy of wind pattern, the correlation depends on the relative orientation of the wind farms as demonstrated in Section 3.3. As discussed in Section 3.4.3 and Section 3.4.4, one could consider the Lagrangian approach and construct a correlation model in the form

$$\begin{aligned} K(\mathbf{h}, u) &= K_{ST}(\mathbf{h}, u) + \delta_{\mathbf{h}=\mathbf{0}} K_0(u) \\ \text{where} \quad K_{ST}(\mathbf{h}, u) &= K_S(\|\mathbf{h} - \mathbf{v}u\|) \end{aligned} \tag{3.39}$$

with velocity $\mathbf{v}$ representing the main direction of the movement of the weather front, and K_S is a valid purely spatial isotropic correlation model. This has been applied in Stein (2005) on the wind speed data in Ireland, as discussed in Section 3.5.1. Note that in (3.39), one essentially reduces a anisotropic spatiotemporal correlation model $K_{ST}(\mathbf{h}, u)$ into a purely spatial isotropic correlation model $K_S(\|\tilde{\mathbf{h}}\|)$ by transforming the spatial lag by $\tilde{\mathbf{h}} = \mathbf{h} - \mathbf{v}u$.

On the other hand, following a similar approach in (3.39), one could consider transforming the temporal lag u and reducing a anisotropic spatiotemporal correlation model $K_{ST}(\mathbf{h}, u)$ into a purely temporal isotropic correlation model $K_T(|\tilde{u}|)$, where $\tilde{u}$ is some combination of $\mathbf{h}$ and u. As seen from the correlation plots below, this approach could be more intuitive and directly explains the shape of the anisotropic cross correlations.

3.5.2.2 Features of anisotropic correlations

To see this, we revisit the asymmetry of correlations between wind power generated at two wind farms as shown in Table 3.1. From the data, we clearly observe that when two wind farms lie roughly along the direction of wind propagation, the cross correlations of wind power at a negative temporal lag is significantly different from that at a positive temporal lag. Moreover, another feature that is not shown in Table 3.1 is the value of temporal lag at which the correlations attain maximum. Because of the movement of the weather front along a particular direction, the correlations between wind power generated at two wind farms that lie along the direction of propagation of weather front in general exhibit the following two features:

Feature 1 : For any temporal lag $u \neq 0$, the correlation $K(\mathbf{h}, u)$ is significantly different from $K(\mathbf{h}, -u)$. This corresponds to the anisotropy in time, which is induced by the direction of movement of weather front, i.e. flow of information.

Feature 2 : The correlation $K(\mathbf{h}, u)$ attains maximum at a larger magnitude of temporal lag $|u|$[1]. This differs from the case when two wind farms lie

[1]We only need to consider the absolute value $|u|$. It is because if the correlations between

roughly perpendicular to the movement of the weather front, where $K(\mathbf{h}, u)$ attains maximum when $u \approx 0$. This corresponds to the time needed for the flow of information, which is carried by the movement of the weather front.

Figures 3.7 and 3.8 show some typical correlation plots between wind power generated at two wind farms. They are chosen from wind farms as shown on Figure 3.5 so that one could have a better idea of their relative locations. Figure 3.7 concerns with wind farms number 1 and 55, which lie approximately along the direction of the moving weather front. On the other hand, Figure 3.8 concerns with wind farms number 42 and 55 which lie roughly along the perpendicular direction. On the Figures, Δ_0 represents Feature 1, and is the difference in correlations $|K(\mathbf{h}, T_0) - K(\mathbf{h}, -T_0)|$, where T_0 is the magnitude of temporal lag at which the correlations attain maximum. Also, T_0 in Figure 3.7 is much larger than that in Figure 3.8, and this corresponds to Feature 2. These two features are further demonstrated in Figures 3.9 and 3.10 respectively.

wind farms i and j attain maximum at u, then the correlations between wind farms j and i attain maximum at $-u$ due to symmetry of correlation functions.

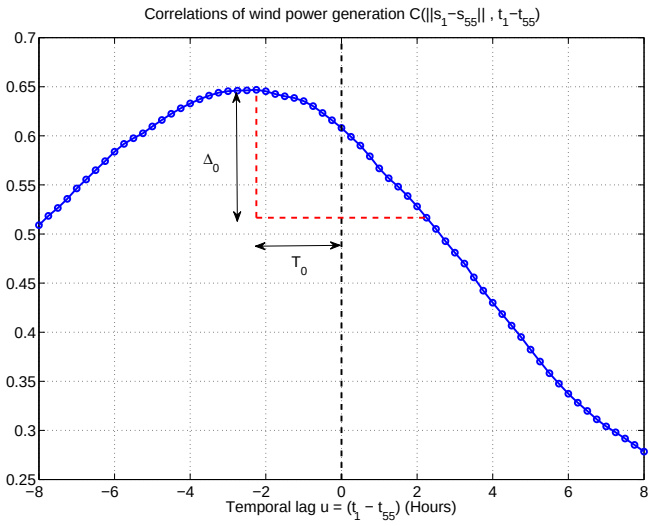


Figure 3.7: Wind power correlations between wind farms 1 and 55, where their locations are as shown in Figure 3.5. They lie approximately along the direction of the moving weather front. Δ_0 represents Feature 1, and is the difference in correlations $|K(\mathbf{h}, T_0) - K(\mathbf{h}, -T_0)|$, where T_0 represents Feature 2 and is the magnitude of temporal lag at which the correlations attain maximum.

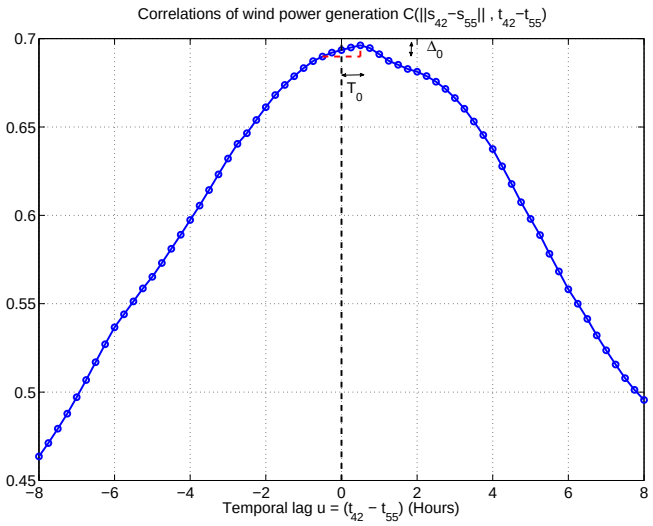


Figure 3.8: Wind power correlations between wind farms 42 and 55, where their locations are as shown in Figure 3.5. They lie approximately along the perpendicular direction to the moving weather front. Δ_0 represents Feature 1, and is the difference in correlations $|K(\mathbf{h}, T_0) - K(\mathbf{h}, -T_0)|$, where T_0 represents Feature 2 and is the magnitude of temporal lag at which the correlations attain maximum.

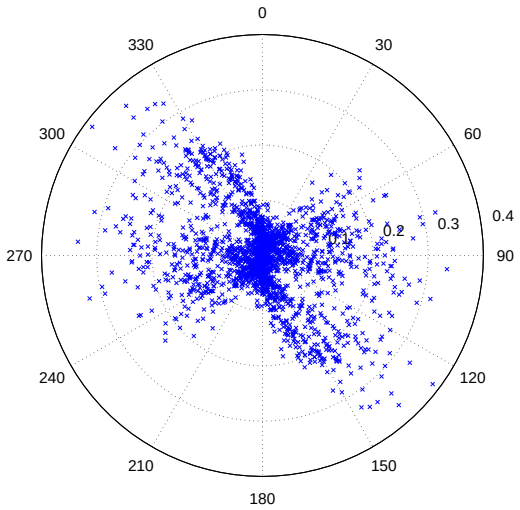


Figure 3.9: This is a polar plot showing the absolute difference between the correlations at positive temporal lag u and negative temporal lag $-u$, i.e. Δ_0 in Figures 3.7 and 3.8. The polar angles are the azimuth angles between each pair of wind farms. The radius show the absolute difference in correlations, i.e. $\Delta_0 = |K(\mathbf{h}, u) - K(\mathbf{h}, -u)|$. The figure shows that wind farms that lie along the north-westerly wind direction in general have a larger absolute difference of correlations at positive and negative temporal lags, where Δ_0 can be up to 0.4.

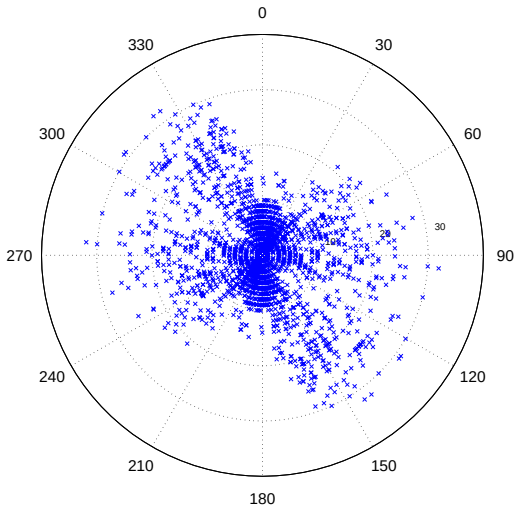


Figure 3.10: This is a polar plot showing the magnitude of temporal lag $|u|$ at which the correlations between two wind farms attain maximum, i.e. T_0 in Figures 3.7 and 3.8. The radius shows the value of T_0, and for instance $T_0 = 4$ stands for a temporal lag of an hour, as the frequency of our data is at 15 minutes. The figure shows that in general wind farms that lie along the north-westerly wind direction have maximum correlations in wind power generation at a larger magnitude of temporal lag T_0, which can be up to 7-8 hours.

3.5.2.3 Model specification

With these considerations in mind, we decide to construct a spatiotemporal correlation model which captures the shift of the temporal correlations, T_0, according to the orientations of the two wind farms, i.e. weather it is along or perpendicular to the direction of propagation of the weather front. We would like to describe this feature by a transformation in the temporal coordinates, which is basically a translation that shifts the point of maximum correlation away from zero. To describe the orientations of the two wind farms, we consider the azimuth angle $\theta(\mathbf{s}_i, \mathbf{s}_j) = \theta(\mathbf{h})$, which is the angle of $\mathbf{s}_j$ as measured from $\mathbf{s}_i$ in a clockwise direction from the north, as shown in Figure 3.11.

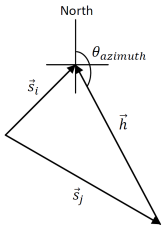


Figure 3.11: This figure illustrates the meaning of an azimuth angle between two wind farms located at $\mathbf{s}_i$ and $\mathbf{s}_j$. For $\mathbf{h} = \mathbf{s}_i - \mathbf{s}_j$, the azimuth angle $\theta(\mathbf{s}_i, \mathbf{s}_j) = \theta(\mathbf{h})$ is the angle of $\mathbf{s}_j$ as measured from $\mathbf{s}_i$ in a clockwise direction from the north.

In our model, we use the powered exponential functions as the building blocks, as they are simple isotropic correlation functions which are easy to calibrate and interpret. We build our spatiotemporal correlation model with a purely spatial and a purely temporal powered exponential functions $K_S(\mathbf{h})$ and $K_T(u)$ where

$$K_S(\mathbf{h}) = \exp(-(\alpha\|\mathbf{h}\|)^{2\gamma}), \qquad K_T(u) = \exp(-(\beta u)^{2\eta}) \tag{3.40}$$

and $\alpha, \beta > 0$, $0 < \gamma, \eta \leq 1$ are constant parameters. We then consider a spatiotemporal correlation model of the form $K_{ST}(\mathbf{h}, u) = K_S(\mathbf{h})K_T(u)$. We modify it to become nonseparable and anisotropic by transforming the temporal coordinates (but not the spatial ones) from u to $\tilde{u}$ such that

$$\tilde{u} = u + \frac{\mathbf{h} \cdot \mathbf{v}}{\|\mathbf{v}\|^2} \tag{3.41}$$

where $\mathbf{v}$ is a parameter representing the constant velocity of the weather front. As $K_T(u)$ is a valid correlation function, it follows that $K_T(\tilde{u})$ is also valid, i.e. positive definite (Appendix C).

Together with a purely spatial nugget effect $K_0(u)$ which accounts for the discontinuity of correlations at spatial lag $\mathbf{h} = \mathbf{0}$, our spatiotemporal correlation model is written as

$$\begin{aligned} K(\mathbf{h}, u) &= K_{ST}(\mathbf{h}, u) + \delta_{\mathbf{h}=\mathbf{0}} K_0(u) \\ \text{where} \qquad K_{ST}(\mathbf{h}, u) &= c_0 \exp\left(-(\alpha\|\mathbf{h}\|)^{2\gamma}\right) \exp\left(-(\beta\tilde{u})^{2\eta}\right), \\ K_0(u) &= \exp\left(-(\tilde{\beta}u)^{2\tilde{\eta}}\right) - c_0 \exp\left(-(\beta u)^{2\eta}\right) \\ \tilde{u} &= u + \frac{\mathbf{h}\cdot\mathbf{v}}{\|\mathbf{v}\|^2} \\ &= u + \frac{\|\mathbf{h}\|\cos(\theta - \theta_0)}{v} \end{aligned} \tag{3.42}$$

where $\alpha, \beta, \tilde{\beta} > 0$ are the scale parameters for space and time respectively, and $0 < \gamma, \eta, \tilde{\eta} \leq 1$ are the shape parameters with restrictions to ensure positive definiteness. The magnitude of the nugget effect is controlled by $c_0 \in [0, 1]$. $v > 0$ represents the speed of the weather front, and θ_0 is the azimuth angle describing the direction of its movement. Note that $\theta = \theta(\mathbf{s}_i, \mathbf{s}_j) = \theta(\mathbf{h})$ is the azimuth angle between two wind farms located at $\mathbf{s}_i$ and $\mathbf{s}_j$, as illustrated in Figure 3.11.

3.5.2.4 Discussions and remarks

In the following, we would like to provide several remarks concerning our spatiotemporal correlation model in (3.42). In addition, a summary of the properties of various correlation models, including those discussed previously in Section 3.5.1, is given in Table 3.3.

1. In contrast with the usual Lagrangian approach for building anisotropic correlations as discussed in Section 3.4.3, we do not consider transforming the spatial coordinates by $\tilde{\mathbf{h}} = \mathbf{h} - \mathbf{v}u$. Instead, as seen from the above observations in the correlation plots, we advocate a transformation of temporal coordinates by

$$\tilde{u} = u + \frac{\|\mathbf{h}\| \cos(\theta - \theta_0)}{v}$$

Results show that this model could give a much better fit compared with the usual Lagrangian model. Transforming the one-dimensional temporal lag could be a better approach and may induce less distortion.

2. The shifting term $\|\mathbf{h}\| \cos(\theta - \theta_0)/v$ represents the time needed for the weather front to propagate from $\mathbf{s}_i$ to $\mathbf{s}_j$.

3. A special case can be achieved when we set $\beta = \tilde{\beta}$ and $\eta = \tilde{\eta}$. However, to obtain a good fit, it is critical to allow them, especially the shape parameter η, to be different in K_{ST} and in K_0. The reason is that we observe a very different shape of correlation decay depending on whether $\mathbf{h} = \mathbf{0}$ or not. This difference is typical and is shown in Figure 3.12. This phenomenon, namely that the cross correlations are smoother than the autocorrelations, is true for any data sets due to the structure of the fourier transforms of valid correlation models, which could be derived mathematically (Gneiting *et al.*, 2010). As a result, we use a different parameter set $\tilde{\beta}, \tilde{\eta}$ to control the decay of autocorrelation. By rearranging the terms in (3.42), we see that the autocorrelation model is essentially given by

$$K(\mathbf{0}, u) = K_{ST}(\mathbf{0}, u) + K_0(u) = \exp\left(-(\tilde{\beta}u)^{2\tilde{\eta}}\right)$$

In fact, we define the spatial nugget effect as the difference between two exponential functions as given in (3.42) so as to obtain this particular autocorrelation model.

4. The spatial nugget effect $K_0(u)$, which is the difference between two exponential functions, must be positive definite in order for the model to be valid. We must include this constraint in our parameter estimation. Theoretically, one could impose a constraint on c_0 so as to ensure that the function is positive definite (Gregori *et al.*, 2008). However, due to the presence of η and $\tilde{\eta}$ in the functions, no closed form solutions for the permissible interval of c_0 could be found and thus numerical method is required. As we only consider a correlation matrix at fixed locations, we simply check if the matrix is positive definite with the chosen set of parameters.

5. Since the Earth is spherical, the azimuth angles in general do not satisfy the relation $|\theta(\mathbf{s}_i, \mathbf{s}_j) - \theta(\mathbf{s}_j, \mathbf{s}_i)| = 180°$. This results in

$$\begin{aligned} & \cos(\theta(\mathbf{s}_i, \mathbf{s}_j) - \theta_0) \quad \neq \quad -\cos(\theta(\mathbf{s}_j, \mathbf{s}_i) - \theta_0) \\ i.e. \quad & \cos(\theta(\mathbf{h}) - \theta_0) \quad \neq \quad -\cos(\theta(-\mathbf{h}) - \theta_0) \end{aligned}$$

and a lack of symmetry in the model. To ensure that the correlation model is symmetric, i.e. $K(\mathbf{h}, u) = K(-\mathbf{h}, -u)$, we slightly adjust all pairs of azimuth angles by taking averages such that $|\theta(\mathbf{s}_i, \mathbf{s}_j) - \theta(\mathbf{s}_j, \mathbf{s}_i)| = 180°$. The adjustment is very small in all cases and is of the order of $0.1°$.

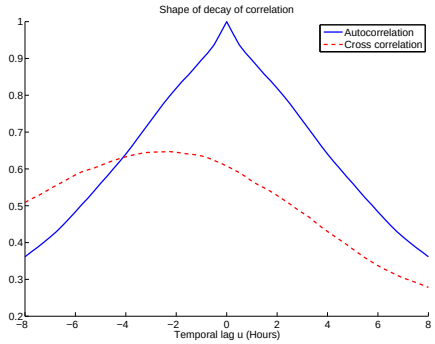


Figure 3.12: This figure shows the typically different shapes of decay between autocorrelations and cross correlations. In this example, the autocorrelation for wind power generated at wind farm number 1 is computed. It is clear that the decay is steeper and the function is not differentiable at zero. The cross correlation is computed between wind power generated at wind farms number 1 and 55. The decay of correlation is smoother than that for autocorrelations, and this feature is true for any data sets (Gneiting *et al.*, 2010).

Model	**Cressie**	**Gneiting**	**Stein**	**Lagrangian**	**Our Model**
Number of parameters	3	5	7	7	9
Building blocks	Cauchy	Cauchy, powered exponential	Matérn	Powered exponential	Powered exponential
Separable	No	No	No	No	No
Nugget	Yes	Yes	Yes	Yes	Yes
Isotropic	Yes	Yes	No	No	No

Table 3.3: A summary of the properties of various correlation models.

3.5.3 Estimation of correlation models

The estimation of correlation models can be achieved in various ways. Suppose that the empirically estimated correlations are given by $\hat{K}(\mathbf{h}, u)$, which we regard as the true correlations of the data. For a correlation model $K(\mathbf{h}, u|\boldsymbol{\Theta})$ with parameters denoted by $\boldsymbol{\Theta}$, we can fit the model by minimizing the sum of squared errors between $\hat{K}$ and K, i.e. minimize

$$\Phi = \sum_{t_i - t_j = 1,2,\ldots} \sum_{i,j=1}^{m} \left(\hat{K}(\mathbf{s}_i - \mathbf{s}_j; t_i - t_j) - K(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j|\boldsymbol{\Theta}) \right)^2 \tag{3.43}$$

However, Stein (1988) and Cressie & Huang (1999) show that obtaining a good fit near the origin, i.e. when K is large, is most critical. Thus, a better approach is to put more weight on the errors when K is large. The parameters $\boldsymbol{\Theta}$ are then obtained by minimizing the weighted squared errors

$$\Phi_w = \sum_{t_i - t_j = 1,2,\ldots} \sum_{i,j=1}^{m} \left(\frac{\hat{K}(\mathbf{s}_i - \mathbf{s}_j; t_i - t_j) - K(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j|\boldsymbol{\Theta})}{1 - K(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j|\boldsymbol{\Theta})} \right)^2 \tag{3.44}$$

where m is the total number of sites. We will adopt this weighted least square (WLS) approach to estimate the parameters in all the correlation models. In practice, one have to select a certain number of temporal lags, p_0, to be included in the estimation of (3.44) such that the infinite sum of $t_i - t_j = 1, 2, \ldots$ is truncated to $t_i - t_j = 1, 2, \ldots, p_0$. We will discuss this in more details in Section 3.6.

Note that one may also estimate the parameters by maximizing the likelihood of the correlation models. However, due to the vast amount of data, intensive computations will be required. Stein (2005) estimated the models using a restricted maximum likelihood estimator. Apart from its computational demand, an advantage of using a maximum likelihood estimator for the parameters $\boldsymbol{\Theta}$ is that the estimated model $K(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j|\hat{\boldsymbol{\Theta}})$ will then be guaranteed to be positive definite. It is also for this reason, Stein (2005) calculates the spatial lag $\|\mathbf{h}\|$ in his models using the formula for great circle distance (3.25) which does not give spatial distortions. As discussed in Section 3.4.4, using the great circle

distance instead of the Euclidean metric may lead to a violation of positive definiteness for the model, but if the model parameters are estimated by maximizing the restricted likelihood, one will get rid of this problem.

3.6 Generating spatiotemporal forecasts

Finally, equipped with the spatiotemporal correlation models together with the estimated parameters, we consider the practical problem of actually obtaining forecasts using the spatiotemporal kriging predictor, as introduced in (3.7) at the beginning of this Chapter. This involves a transformation of the data to meet with the Gaussian assumptions on kriging predictors, a choice of maximum temporal lag to be included in the correlation matrices, a choice between homoscedastic or heteroscedastic variances, and an inverse transformation of the data which ensures the forecasts lie in the permissible range of $[0, 1]$ for wind power generation.

3.6.1 Normalizing the wind power data

First, to calculate the spatiotemporal kriging predictors, recall that one has assumed that the data is Gaussian such that the kriging predictors minimize the variance of errors. Clearly, from Figure 1.5 (a), wind power generation is non-negative and significantly skewed, and so obviously non-Gaussian. Nevertheless, as discussed in Section 1.3.1, we could normalize the wind power data at each wind farm using the modified logit transformation in (1.1). We would apply this transformation in the following forecasts, and forecast performances are significantly enhanced by firstly normalizing the data.

3.6.2 Kriging predictor

Using the same notations in the description of the spatiotemporal kriging approach in Section 3.2, let us denote the normalized wind power data at wind farm k by $Z(\mathbf{s_k}, t)$. As described previously, under squared loss the optimal linear predictor for $Z(\mathbf{s_k}, t+h)$ depends on the covariances between $Z(\mathbf{s_k}, t+h)$ and $Z(\mathbf{s_j}, u)$ where $j = 1, \ldots, m$ and $u = 1, \ldots, t$, where m is the total number of

different sites. The problem boils down to determining the covariance matrices in (3.6).

For $h \geq 1$, consider the h-step ahead optimal linear predictor for $\mathbf{Z}(t+h)$. Further assume that the optimal linear predictor for $\mathbf{Z}(t+h)$ is determined solely by the observations at the $(p-1)$ latest temporal lags, that is, $t, t-1, \cdots, t-p+1$. Then, instead of calculating the covariance matrices of $\mathbf{Z}$ at every temporal lag $|u| < t+h$, it suffices to consider only the covariances at temporal lags $|u| \leq (p+h-1)$ below the cutoff value $p+h$. $\mathbf{\Sigma}$ in (3.4) is then written as

$$\mathbf{\Sigma}_p = \begin{pmatrix} \mathcal{C}(0) & \cdots & \mathcal{C}(-(p-1)) \\ \vdots & \ddots & \vdots \\ \mathcal{C}(p-1) & \cdots & \mathcal{C}(0) \end{pmatrix}_{(pm)\times(pm)} \tag{3.45}$$

where we use (3.6) to express the $m \times m$ matrices $\mathcal{C}(t_i, t_j) = \mathcal{C}(u)$ where $u = t_i - t_j$. The covariance matrices in $\mathbf{c}$ in (3.5) only concern with temporal lags from $u = -(p+h-1)$ to $u = -h$, and we have

$$\mathbf{c}_{p,h} = \begin{pmatrix} \mathcal{C}(-(p+h-1)) \\ \vdots \\ \mathcal{C}(-h) \end{pmatrix}_{(pm)\times m} \tag{3.46}$$

The h-step ahead spatiotemporal kriging predictor for $Z(\mathbf{s}, t+h)$ is then obtained by (3.7), with the corresponding kriging variance given by (3.8). Since we assume the optimal linear predictor $\mathbf{Z}(t+h)$ generated at time t only depends on $\mathbf{z}(t), \mathbf{z}(t-1), \ldots, \mathbf{z}(t-p+1)$, the kriging predictor is written as

$$p(\mathbf{Z}(t+h)) = \boldsymbol{\mu} + \mathbf{c}_{p,h}^T \mathbf{\Sigma}_p^{-1} \left[(\mathbf{z}(t-p+1) - \boldsymbol{\mu}), \ldots, (\mathbf{z}(t) - \boldsymbol{\mu})\right]^T \tag{3.47}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^T$ and $\mu_j = \sum_{i=1}^{p} z(\mathbf{s}_j, t-i+1)/p$.

For the wind power data in Ireland, we choose the value $p = 21$ in (3.45), i.e. the optimal linear predictor for $\mathbf{Z}(t+h)$ is determined by the observations at $t, t-1, \cdots, t-20$ only, observations in the past 5 hours. However, by observing the temporal lags in 3.46, note that we have to estimate the spatiotemporal correlation models up to a larger value of lag $p_0 > p$ if we want to generate multi-step ahead forecasts at horizons $h > 1$. In particular, we will consider 12−step

ahead (i.e. 3 hours) forecasts for the wind power generation, and the maximum temporal lags (both positive and negative) to be included in the weighted least square (WLS) approach in (3.44) would be $p_0 = p + h - 1 = 32$, i.e. $\pm$ 8 hours of autocorrelations and cross correlations.

This value of $p = 21$ in (3.45) is partly chosen by taking into the account of computational burden in the calculations of large covariance matrix when p is large. For instance, when $p = 21$, the dimension of the spatiotemporal correlation matrix, $\mathbf{\Sigma}_p$, is 1344×1344, and the calculation of the kriging predictor in (3.47) will involve the inversion of such a large matrix. Nevertheless, on the other hand, we do not want to choose a value of p which is too small and fails to capture the shape of the decay of correlations. From the correlation plots in Figures 3.7 and 3.8, we see that a temporal lag of $u = \pm 32$, i.e. $\pm$ 8 hours, is able to show the decay of the correlations satisfactorily and this is an appropriate number of temporal lags to be considered in the estimation of the models. Also, we test the robustness of the parameters when we minimize the WLS in (3.44) using different values of lags p_0 so that $t_i - t_j = 1, 2, \ldots, p_0$. Analysis shows that the estimated parameters remain stable when p_0 increases, as shown in an example in Figure 3.13. As a result, we choose $p_0 = 32$ in the objective function (3.44). With the maximum forecast horizon being $h = 12$, we consider $p = p_0 - h + 1 = 21$ in the calculation of kriging predictor in (3.45), (3.46) and (3.47). In other words, we generate the prediction of $\mathbf{Z}(t + h)$ using only observations at $t, t - 1, \cdots, t - 20$ only, i.e. observations in the past 5 hours.

3.6.3 Homoscedastic or heteroscedastic variances

In the kriging predictor in (3.47), we need to calculate the covariance matrices $\mathbf{\Sigma}_p$ and $\mathbf{c}_{p,h}$. As discussed previously, we decompose the covariances into a product of variances and correlations, as given in (3.13). With the correlation models $K(\mathbf{h}, u)$, it remains to decide a model for the variances. Here, we compare between two choices. The first one is the homoscedastic variance

$$\sigma^2_{\mathbf{s},t} = \sigma^2_{\mathbf{s}} \tag{3.48}$$

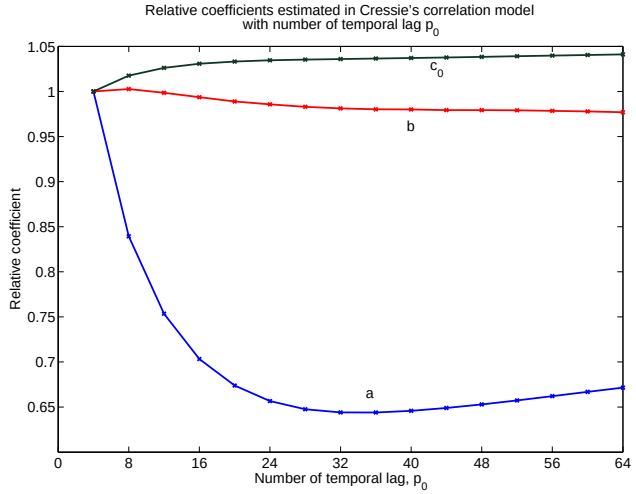


Figure 3.13: This plot shows an example of the robustness of the estimated parameters with different numbers of temporal lag p_0 included in the WLS estimation in (3.44). In this example, we consider the Irish wind data with correlations fitted to the Cressie's correlation model in (3.31) with three parameters a, b and c_0. Results show that fitting the empirical correlations up to $p_0 = 32$ temporal lags give a good enough estimation. This is also supported by the fact that it is most important to obtain good fit for large correlations near the origin (Cressie, 1993).

where we assume that the variances depends on spatial locations but are stationary through time. We then estimate the sample variances for each individual wind power using the training data, and obtain the covariance matrices $\mathbf{\Sigma}_p$ and $\mathbf{c}_{p,h}$ in the kriging predictor. This will be denoted as the homoscedastic spatiotemporal kriging.

The second choice is the heteroscedastic variance described by the GARCH(1,1) model

$$\sigma^2_{\mathbf{s},t} = \omega_0 + \alpha\varepsilon^2_{\mathbf{s},t-1} + \beta\sigma^2_{\mathbf{s},t-1} \tag{3.49}$$

where $\omega_0, \alpha, \beta > 0$ and $\alpha + \beta < 1$, and ε is the forecast error as usual. Note that we assume the variances to depend on spatial locations as well, and so the parameters ω_0, α and β are estimated for each wind power. The results will be multiplied with the correlations to give the covariance matrices. We call this the heteroscedastic spatiotemporal kriging.

3.6.4 Generating valid forecasts in $[0, 1]$

An additional problem that we face in wind power forecasts is to ensure that the forecasts generated by the models are valid, i.e. the values fall within the permissable domain $[0, 1]$ for normalized wind power generation. Even though we normalize the data y using the modified logit transformation in (1.1), the transformed data z is still bounded within an interval. For instance, if $\delta = 0.0312$, we have

$$y \in [0, 1] \qquad \rightarrow \qquad z \in [-3.5, 3.5] \tag{3.50}$$

However, when we calculate the spatiotemporal kriging predictor μ_z, we assume that z is Gaussian and so the predictor μ_z can be of any values which may lie outside the permissable region. To ensure that the forecasts are valid, we consider an approach similar to the one adopted in Chapter 2. We consider the use of truncated normal distribution to ensure that the forecasts are within the permissable region.

Once we obtain the kriging predictor μ_z and the corresponding kriging variance σ_z^2, we treat them as the location and scale parameters of a truncated normal distribution, with the left and right truncations at $z_{\min} = -3.5$ and $z_{\max} = 3.5$ respectively. In other words, we consider z to come from a distribution similar to that in (2.23) but with different truncations, i.e.

$$f(z; \mu_z, \sigma_z^2) = \frac{1}{\sigma_z} \frac{\varphi\left(\frac{z-\mu_z}{\sigma_z}\right)}{\Phi\left(\frac{z_{\max}-\mu_z}{\sigma_z}\right) - \Phi\left(\frac{z_{\min}-\mu_z}{\sigma_z}\right)}, \qquad z \in [z_{\min}, z_{\max}] \tag{3.51}$$

where φ and Φ are the standard normal density and distribution function respectively.

To generate forecasts for y, we use the method of simulations because through simulations we can easily obtain the density forecasts even for aggregated wind power generations[1]. For each forecast of μ_z, we consider drawing N random samples $\{z^i\}_{i=1}^N \in [z_{\min}, z_{\max}]$ from the corresponding truncated normal distribution in (3.51). Since we have transformed the data from y to z using the modified

[1]The aggregated wind power density forecasts could not be obtained if we consider an analytical approach, although it is still possible to write down the closed form of the densities for individual wind power.

logit transformation in (1.1), we apply the inverse and obtain N corresponding samples of $\{y^i\}_{i=1}^N \in [0,1]$ by calculating

$$y^i = \frac{(1+\delta)\exp(z^i) - \delta}{1+\exp(z^i)}, \qquad y^i \in [0,1] \tag{3.52}$$

From the samples of $\{y^i\}_{i=1}^N$, we can easily estimate the empirical individual density distributions of the forecasts. In addition, as we obtain a sample for each location of wind farm, we can aggregate the values to get a sample of total wind power generation. With all the samples we can then estimate the density distribution of aggregated wind power directly. In the application of the kriging approach in Chapters 5 and 6, we see that $N = 400$ is sufficient for the forecast densities. An example of the robustness of results is demonstrated in Figure 3.14. In this example, we consider one-step ahead forecasts generated by the spatiotemporal kriging approach with the empirical correlations. Forecast performances under RMSE and mean CRPS are stable with different numbers of samples from $N = 100$ to $N = 800$, showing that the forecast densities do not vary a lot with different numbers of samples.

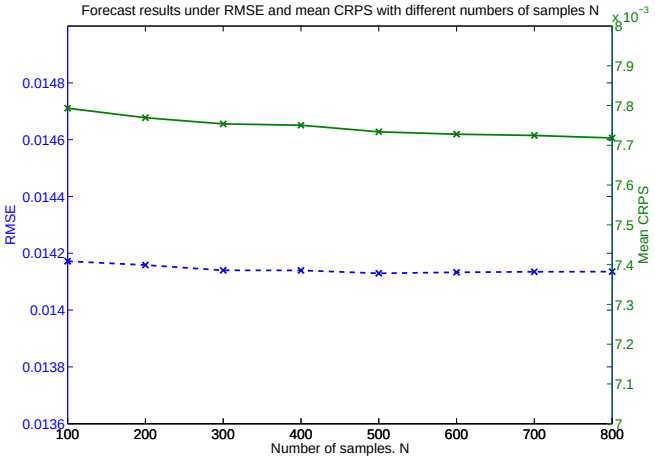


Figure 3.14: This figure shows the variations of forecast results under RMSE and mean CRPS with the number of samples drawn. In this example, we consider the Irish wind data and one-step ahead forecasts are generated by the spatiotemporal kriging approach with the empirical correlations. Forecast performances under RMSE and mean CRPS are stable with different numbers of samples from $N = 100$ to $N = 800$, showing that the forecast densities do not vary a lot with different numbers of samples.

Chapter 4

Two-stage Model with Latent Gaussian Processes

4.1 Introduction

In this chapter, we explore another approach which has been used widely to model spatial data. We consider the use of Gaussian processes, or Gaussian random fields, to describe the correlation structure of the data. Gaussian processes are relatively easy to implement and estimate since they are simply characterized by the mean and covariance functions. Inferences are also conveniently obtained in the case of Bayesian analysis due to their analytical and numerical tractability (Gibbs & MacKay, 1997; MacKay, 1998). A Gaussian process is an infinite dimensional stochastic process where any finite dimensional distributions are multivariate Gaussian distributions. In particular, if a spatiotemporal process $Z(\mathbf{s}, t)$ at location $\mathbf{s}$ and time t is a Gaussian process, then for any positive integer n and any $\{(\mathbf{s}_j, t_j)\}_{j=1}^{n}$, we have

$$(Z(\mathbf{s}_1, t_1), \ldots, Z(\mathrm{s}_n, t_n)) \sim N(\mu, \mathbf{\Sigma}) \tag{4.1}$$

where $\mu \in \mathbb{R}^{n\times 1}$ and $\mathbf{\Sigma} \in \mathbb{R}^{n\times n}$ are the mean and covariance matrices of the corresponding mean and covariance functions that characterize the Gaussian process. An excellent introduction on applications of Gaussian processes is given by Rasmussen & Williams (2006). Abrahamsen (1997) provides a detailed review on Gaussian processes.

4.1.1 Discrete-continuous model

Discrete-continuous models are models that explicitly describe the separate components in a data set, where each component can either be a discrete (e.g. count or indicator) or continuous. One of the most common uses is in the modeling of consumption decisions (Mansur *et al.*, 2005), where a discrete variable indicates the choice of consumption and a continuous variable models the distribution of demand or consumption level conditional on the particular choice of consumption. It is also widely used in health science and biostatistics such as the modeling of birth defects (Sammel *et al.*, 1997) and skin tumorigenesis (Dunson & Herring, 2005).

The idea of using a mixed discrete-continuous model is also seen for many environmental studies including rainfall or precipitation forecasts, since in such data sets there is always a point mass at zero which corresponds to the discrete component in the data set, i.e. rain or no rain. This point mass of zero can be also viewed as arising from the fact that rainfall observation cannot be negative. In other words, the amount of rainfall Y is a random variable bounded below by zero. More generally, many environmental and physical processes Y are bounded below by some constant c. In such case, a simple discrete-continuous model for Y could be formulated as

$$Y = \begin{cases} 0, & Z \leq c \\ f(Z), & Z > c \end{cases} \tag{4.2}$$

where Z is some latent random variable, most commonly being Gaussian. In that case, Stein (1992) calls this a truncated Gaussian process. The function f is a transformation chosen so that $f(Z)|Z > c$ follows some preferred distributions, which are mostly Gaussian but may also be lognormal (Bell, 1987).

The power truncated normal (PTN) model is then a special case of (4.2) where a power transformation is chosen, i.e., $f(Z) = Z^p$. It has been studied by Glasbey & Nevison (1997); Sanso & Guenni (1999, 2000) to model accumulated rainfall data. In the PTN model, one assumes that the non-negative variable of interest $Y(\mathbf{s}, t)$ at location $\mathbf{s}$ and time t is driven by a single latent spatiotemporal

Gaussian process $Z(\mathbf{s}, t)$ such that

$$Y = \begin{cases} 0, & Z \leq 0 \\ Z^p, & Z > 0 \end{cases} \tag{4.3}$$

The parameter p in the power transformation corresponds to the tail behavior of the data. As noted by Sanso & Guenni (2000), an advantage of (4.3) is that it is based on a Gaussian process and, thus, all traditional geostatistics and time series techniques can be applied. Also, if the parameter p in the power transformation is known or estimated, (4.3) simply means that $Y^{1/p} = Z^+$ where $Z^+ = Z$ when Z is positive and zero otherwise.

It turns out that a discrete-continuous model similar to the form of the truncated model in (4.3) is a nice candidate to model wind power data. This is largely due to the characteristic features of wind power distributions, which in some sense resemble those of rainfall distributions. The marginal distribution of wind power at a single wind farm is highly non-Gaussian. There are many zero values, and the distribution is typically right-skewed. In the case of aggregated wind power generation, this problem is less serious and one could still normalize the distribution approximately, for instance, using the logit transformation. However, if the focus is on wind power generation at a single, individual wind farm with only tens of wind turbines, one could almost be sure that there are chains of zeros in the data set and the distribution is much more skewed. We cannot obtain an approximate Gaussian distribution by applying a logit transformation because of the abundance of zeros and ones[1]. Since there is a discontinuity in the cumulative distribution of wind power generation, it is impossible to handle the probability mass at zero and one through a continuous transformation. As a result, we consider a discrete-continuous model to describe these features. In the next section, we describe how one could slightly modify (4.3) and apply a two-stage model for wind power generation.

[1]The upper bound of wind power is one as we normalize the data by dividing by the maximum capacity. Typically, the percentage of zero wind power can range from $4 - 20\%$, while that of maximum wind power can range from $0 - 5\%$ (However, in most cases it is smaller than 1%).

4.1.2 Two-stage model

A slightly more sophisticated discrete-continuous model is the two-stage model studied by Sloughter *et al.* (2007) and Berrocal *et al.* (2008), which was applied to describe accumulated precipitation. A recent paper by Berrocal *et al.* (2010) applies the model on ice formation as well. Two-stage model stands for the fact that we model the occurrence of an event (e.g. rainfall) in the first stage. Provided that the event has occurred, a second stage is applied to model the conditional distribution. Stern & Coe (1984) applies a two-stage approach to model rainfall data, where Markow chains are used to model the occurrence of rain and gamma distributions are used to model non-zero rainfall. Other papers using this idea include Bardossy & Plate (1992); Rappold *et al.* (2008). Since wind power generation and precipitation share similar features in the density distribution, namely, an abundant amount of zeros and a highly skewed distribution, we consider a similar two-stage model for wind power generation.

The main idea of the two-stage model is to first handle the occurrence of zero values, and then model the right-skewed probability density conditioned on observing non-zero values. As a result, the two-stage model is driven by two spatiotemporal Gaussian processes $W(\mathbf{s},t)$ and $Z(\mathbf{s},t)$, where W governs the probability of observing zero values and Z governs the density distribution of non-zero observations. In terms of the two-stage model, the wind power generation is a random process Y satisfying

$$Y = \begin{cases} 0, & W \leq 0 \\ f(Z), & W > 0 \end{cases} \tag{4.4}$$

In other words, the power truncated normal (PTN) model in (4.3) is a special case of the two-stage model (4.4) where $W = Z$ and $f(\cdot)$ is the power transformation. In this model, we assign a non-zero probability mass at $Y(\mathbf{s},t) = 0$, and for $Y(\mathbf{s},t) > 0$ we model the wind power using some suitable transformations $f(\cdot)$. One of the practical choices would be the truncated normal distributions, where we will describe more details about this later. We do not use a beta distribution as in Berrocal *et al.* (2008) since results show that truncated normal distributions

give good fits for the wind power data and the dynamics are easier to model.[1].

Now, suppose that Z is a standardized Gaussian process. Taking $f(\cdot)$, i.e. the distribution of Y conditioned on $Y > 0$, to be the truncated normal distribution, we can write

$$f = F^{-1} \circ \Phi \tag{4.5}$$

where F and Φ are the cumulative distribution function (cdf) of the truncated normal distribution and the standard normal distribution respectively.

In our study of wind power forecasting, we normalize the wind power generation of each wind farm by dividing the power by the maximum capacity of that wind farm. Thus, apart from having a lower bound at zero, our data also have an upper bound at one. To account for this feature, we slightly modify the two-stage model in (4.4) by including a third scenario indicated by $W \geq 1$. We consider wind power generation $Y(\mathbf{s}, t)$ being driven by two Gaussian processes $W(\mathbf{s}, t)$ and $Z(\mathbf{s}, t)$, so that at each location $\mathbf{s}$ and time t, we have

$$Y(\mathbf{s}, t) \quad = \quad \begin{cases} 0, & W(\mathbf{s}, t) \leq 0 \\ F^{-1} \circ \Phi(Z(\mathbf{s}, t)), & 0 < W(\mathbf{s}, t) < 1 \\ 1, & W(\mathbf{s}, t) \geq 1 \end{cases} \tag{4.6}$$

where F is the cdf of the truncated normal distribution. This would be a parametric distribution function governed by a location parameter ℓ and a scale parameter s^2. The evolution of $f = F^{-1} \circ \Phi$ could then be modeled in terms of the dynamics of ℓ and s^2.

[1]We try to fit the conditional wind power density to various parametric distributions, including the beta distribution, normal distribution and truncated normal distribution. The mean likelihoods obtained from using different number of wind power observations in the fit give consistent rankings of the goodness of fit of various distributions. For instance, using the Danish data, the mean log likelihoods of the beta distribution, normal distribution and truncated normal distribution are $93.24, 89.40$ and 92.75 respectively when 48 past observations are applied in the fit. In all cases, the beta distribution and the truncated normal distribution gives similar likelihoods, while the normal distribution gives a much lower likelihood as expected. As it is computationally more efficient to model the scale and location parameters in a truncated normal distribution, we decide to use that distribution instead of the beta distribution.

4.2 Two-stage model: Model specification

In this section, we describe in details about the model specification of the two-stage model in (4.6). This involves three major components:

1. Model for the first Gaussian process $W(\mathbf{s}, t)$ which governs the regime of wind power generation.

2. Model for the transformation $F(\cdot)$ which describes the continuous distribution of wind power generation between zero and one. Note that F in (4.6) is the cdf corresponding to f in (4.5). As mentioned above, we consider $f(\cdot)$ to be the density of a truncated normal variable.

3. Model for the second Gaussian process $Z(\mathbf{s}, t)$ which governs the magnitude of wind power generation conditional on the fact that it lies between zero and one.

4.2.1 Model for Gaussian process $W(\mathbf{s}, t)$

First we define a model for the latent Gaussian process $W(\mathbf{s}, t)$ that governs the occurrence of positive wind power generation through the relation that $Y(\mathbf{s}, t) > 0$ if and only if $W(\mathbf{s}, t) > 0$. We consider a probit model, which is often used for categorical variables that only consists of a certain possibilities of outcomes. To formulate this model, we define an indicator random variable $I_J(\mathbf{s}, t)$ by

$$I_J(\mathbf{s}, t) = \begin{cases} 1, & Y(\mathbf{s}, t) \in J \\ 0, & Y(\mathbf{s}, t) \notin J \end{cases} \tag{4.7}$$

where $Y(\mathbf{s}, t)$ is the wind power generation. We consider three indicator random variables $I_A(\mathbf{s}, t), I_B(\mathbf{s}, t)$ and $I_C(\mathbf{s}, t)$ where $A = \{0\}, B = (0, 1)$ and $C = \{1\}$. Then, we define $W(\mathbf{s}, t)$ to be the process

$$\begin{aligned} W(\mathbf{s}, t) = {} & \beta_0 + \beta_1 I_A(\mathbf{s}, t-1) + \beta_2 y(\mathbf{s}, t-1) I_B(\mathbf{s}, t-1) + \beta_3 I_C(\mathbf{s}, t-1) \\ & + \varepsilon(\mathbf{s}, t), \qquad \varepsilon(\mathbf{s}, t) \overset{i.i.d.}{\sim} N(0, \mathbf{\Sigma}_\varepsilon) \end{aligned} \tag{4.8}$$

At each time t, the spatial correlations among $W(\mathbf{s}, t)$ at different sites $\mathbf{s}$ are captured by the residuals $\varepsilon(\mathbf{s}, t)$ which have mean zero and covariance matrix $\mathbf{\Sigma}_\varepsilon$.

The distribution of $W(\mathbf{s},t)$ conditional on the past observation $y(\mathbf{s},t-1)$ is thus given by

$$W(\mathbf{s},t)|y(\mathbf{s},t-1) \sim N(\mu_W, \mathbf{\Sigma}_\varepsilon) \tag{4.9}$$

where $\mu_W = \beta_0+\beta_1 I_A(\mathbf{s},t-1)+\beta_2 Y(\mathbf{s},t-1)I_B(\mathbf{s},t-1)+\beta_3 I_C(\mathbf{s},t-1)$. Our target is to estimate the regression parameters $\beta_0, \beta_1, \beta_2$ and β_3 as well as to provide a parametric model to describe the covariance matrix $\mathbf{\Sigma}_\varepsilon$.

As $W(\mathbf{s},t)$ is latent, estimating the model parameters would be more tricky and requires a simulation approach, where the computation time hugely depends on the dimension of the simulation. For this reason, we assume that $W(\mathbf{s},t)$ are independent across time and so the correlation structure simply connects the spatial locations $\mathbf{s}$. Furthermore, we consider a simple and natural spatial correlation model characterized by an exponential decay with mutual distances. This is intuitive as one expects that at the same instant of time t, wind farms that are close together tend to be in similar regimes, i.e. zero, positive or maximum wind power generation. Of course, this model is not perfect. It cannot explain the situation when maximum wind power generation suddenly drops to zero due to shut down of turbines at extreme weather conditions. Nevertheless, this situation is very rare in real data, mainly due to the fact that we have wind data for the whole wind farm instead of for one single turbine. In fact, the number of maximum wind power observations (i.e. y = 1) for the whole wind farm is small, and is usually less than 0.25% at a single wind farm[1]. Conditional on observing a maximum wind power generation, the probability of observing zero wind power generation in the next 15 minutes is extremely low. For instance, in the Danish data, we only observe one such occasion across 49 wind farms throughout a 2 year data set.

As a result, we consider the spatial correlation model for $W(\mathbf{s},t)$, conditional on $Y(\mathbf{s},t-1)$, to be an exponential decay with mutual distances. This is captured by the correlations of the residuals, and can be written as

$$\mathrm{Cov}(\varepsilon(\mathbf{s}_i,t),\varepsilon(\mathbf{s}_j,t)) = \sigma_\varepsilon^2 \exp\left(-\frac{\|\mathbf{s}_i-\mathbf{s}_j\|}{\rho_\varepsilon}\right) \tag{4.10}$$

[1]This is as observed in our Irish and Danish data sets.

In such case, we have two parameters σ_ε and ρ_ε in the correlation model for W, which will be estimated as described in Section 4.3. Since $W(\mathbf{s}, t)$ is latent and unobserved, we shall first estimate $\beta_0, \beta_1, \beta_2$ and β_3 together with σ_ε^2 using an approximate maximum likelihood estimator. Then we shall estimate the correlation parameter ρ by the Expectation-Maximization (EM) algorithm with Gibbs sampling (Zeger & Karim, 1991).

4.2.2 Model for the transformation $F(\cdot)$

Next we deal with the dynamic evolution of the transformation $F(\cdot)$ that appears in the two-stage model (4.6). For the sake of efficient modeling, we assume that the distribution of wind power generation $Y(\mathbf{s}, t)$, conditional on $0 < W(\mathbf{s}, t) < 1$, is represented by an appropriate parametric distribution $f(y)$ truncated at zero and one. As discussed in Chapter 2, a nice and practical candidate could be the truncated normal distribution. This distribution is well suited for aggregated wind power, but it also provides a good approximation for individual wind power although the latter is in general more skewed.

The truncated normal distribution is governed by two parameters, namely, the location parameter ℓ and the scale parameter s^2. Roughly speaking, ℓ controls the mean and s^2 controls the variance of the distribution, although s^2 could affect the mean and ℓ could affect the variance as well. Denote $f(y|\ell, s^2)$ as the truncated normal distribution with location parameter ℓ and scale parameter s^2. Then, ℓ and s^2 are the respective mean and variance of the original normal distribution $N(\ell, s^2)$ without truncation. As ℓ and s^2 govern the dynamic evolution of the truncated distribution $f(y|\ell, s^2)$, modeling the dynamics of f boils down to the problem of modeling ℓ and s^2.

From the study of the modeling of aggregated wind power in Chapter 2, we find that the method of exponential smoothing is efficient, robust and provide competitive forecasts as compared to other ARIMA-GARCH models. More importantly, its advantages of being easy to estimate and implement are critical for practical applications, especially in the case of a portfolio of individual wind farms. In such situations, one could be dealing with 50-80 time series at one time, and it would be impractical to select a particular model for each time series. As a

result, we consider the method of exponential smoothing to forecast the location and scale parameters of $f(y|\ell, s^2)$ at each wind farm[1].

Note that we will not consider any spatial correlations in the exponential smoothing, since we aim at capturing this by the latent Gaussian process $Z(\mathbf{s}, t)$. As a result, in this section, we fix the location of each wind farm, $\mathbf{s}$, and drop the notation in all variables temporarily. The reason that we smooth and forecast the location and scale parameters is to describe the dynamic evolution of the truncated distribution which changes with time. As described in Chapter 2, for wind power y_t observed at a fixed wind farm, the smoothing and forecast equations are written as

$$\begin{aligned} S_t &= \alpha y_t + (1-\alpha) S_{t-1} \\ \hat{\ell}_{t+1|t} &= S_t + \phi_s (y_t - S_{t-1}) \\ \log V_t &= \gamma g(e_{t-1}) + (1-\gamma) \log V_{t-1} \\ \log \hat{s}^2_{\varepsilon;t+1|t} &= \log V_t + \phi_v \left[g(e_{t-1}) - \log V_{t-1} \right] \end{aligned} \tag{4.11}$$

where

$$\begin{aligned} g(e_t) &= \theta \left(|e_t| - \mathrm{E}[|e_t|] \right) \\ e_t &= \varepsilon_t / \sqrt{V_t} \\ \varepsilon_t &= y_t - \hat{\ell}_{t|t-1} \end{aligned} \tag{4.12}$$

In other words, S_t and V_t are the smoothed levels of the location and scale parameters respectively. e_t is the normalized residuals, θ is a parameter in the estimated variance $g(e_t)$, α and γ are the smoothing parameters in the updating equations, and ϕ_s and ϕ_v are used to account for autocorrelations in the forecasts.

The forecasts of location parameter ℓ and scale parameter s^2 will be used to drive the dynamics of the truncated normal distribution. For instance, we consider $F(y|\ell, s^2)$ to be the cdf of the truncated normal distribution. The evolution

[1]Note that we have tried some other methods such as modeling $f(y|\ell, s^2)$ with a non-parametric conditional distribution. However, the method is not very practical as it is computationally demanding, and we also verified that the out-of-sample forecasts are not as good as using exponential smoothing.

of $F(y|\ell, s^2)$ is then given by

$$F_{t+1|t}(y|\ell, s^2) = \frac{\Phi\left(\frac{y-\hat{\ell}_{t+1|t}}{\hat{s}_{t+1|t}}\right) - \Phi\left(\frac{-\hat{\ell}_{t+1|t}}{\hat{s}_{t+1|t}}\right)}{\Phi\left(\frac{1-\hat{\ell}_{t+1|t}}{\hat{s}_{t+1|t}}\right) - \Phi\left(\frac{-\hat{\ell}_{t+1|t}}{\hat{s}_{t+1|t}}\right)} \qquad ,0 \leq y \leq 1 \tag{4.13}$$

where Φ is the standard normal distribution function. The forecast in (4.13) could then be applied in (4.6) for the forecast of wind power generation $Y(\mathbf{s}, t+1|t)$.

4.2.3 Model for Gaussian process $Z(\mathbf{s}, t)$

With the model specification of the transformation F, we could now proceed to specify the Gaussian process $Z(\mathbf{s}, t)$ which governs the magnitude of positive wind power generation. As described above, we model F as the cdf of a truncated normal distribution, governed by location parameter ℓ and scale parameter s^2. According to the two-stage model in (4.6), we have

$$Y(\mathbf{s}, t)|W(\mathbf{s}, t) \in (0, 1) \sim F^{-1} \circ \Phi(Z(\mathbf{s}, t)) \tag{4.14}$$

where Φ is the distribution function of the standard normal distribution. Thus, the correlation structure of $Y(\mathbf{s}, t)$, conditional on $0 < W(\mathbf{s}, t) < 1$, could be described by that of $Z(\mathbf{s}, t)$. $Z(\mathbf{s}, t)$ is defined to be a standard Gaussian process with mean zero, and its marginal variance is chosen to be one so as to ensure identifiability (Berrocal *et al.*, 2010). In Berrocal *et al.* (2008) and Berrocal *et al.* (2010), $Z(\mathbf{s}, t)$ is assumed to have a pure spatial correlation structure only and is independent across time. They adopted an exponential model for the spatial correlation as in (4.10), that is,

$$\text{Cov}(Z(\mathbf{s}_i, t_i), Z(\mathbf{s}_j, t_j)) = \delta_{t_i,t_j} \exp\left(-\frac{\|\mathbf{s}_i - \mathbf{s}_j\|}{\rho_Z}\right) \tag{4.15}$$

However, in our application, there would be temporal correlations among $Z(\mathbf{s}, t)$, and we need to consider $Z(\mathbf{s}, t)$ to be a spatiotemporal Gaussian process instead of merely a spatial one. This could be observed from (4.14), where we have

$$Z(\mathbf{s}, t)|W(\mathbf{s}, t) \in (0, 1) \sim \Phi^{-1} \circ F(Y(\mathbf{s}, t)) \tag{4.16}$$

Although we suppress the dependence of location and scale parameters in F, one should keep in mind that F is dynamic since ℓ and s^2 evolves with time. Note that $Y(\mathbf{s},t)|Y(\mathbf{s},t) \in (0,1)$ in general have significant spatiotemporal correlations, although it is less correlated compared with the original data $Y(\mathbf{s},t)$ which include chains of zeros and ones. The transformed variable $F(Y(\mathbf{s},t))$ will still in general be spatiotemporally correlated, because the time evolving parameters ℓ and s^2 may not be able to account for all spatiotemporal effects. In addition, nonlinear transformations through F could also induce correlations too.

Now we have defined the Gaussian process $Z(\mathbf{s},t)$ to be

$$Z(\mathbf{s},t) \sim N(0, \mathbf{\Sigma}_Z(\mathbf{s},t)) \tag{4.17}$$

The next target is to determine an appropriate spatiotemporal correlation model $\mathbf{\Sigma}_Z(\mathbf{s},t)$ for $Z(\mathbf{s},t)$. The simplest candidate that may come to one's mind is the spatiotemporal version of (4.10), which simply describes the correlation as decaying exponentially with temporal and spatial lags. This could be written as

$$\text{Cov}(Z(\mathbf{s}_i,t_i), Z(\mathbf{s}_j,t_j)) = \exp\left(-\frac{\|\mathbf{s}_i-\mathbf{s}_j\|}{\rho_{Z,\mathbf{s}}}\right)\exp\left(-\frac{|t_i-t_j|}{\rho_{Z,t}}\right) \tag{4.18}$$

However, this simple model is isotropic and separable, which is not a realistic description of the correlation structure of wind power data that depends largely on movements of weather front and spatial locations. As a result, we will not apply the correlation model in (4.18) in our subsequent analysis. In fact, we have tried to apply this in the forecasts and results are very poor, which is expected.

Following Chapter 3, we consider some more sophisticated correlation models for $\mathbf{\Sigma}_Z$. These correlation models have been described in detail in Section 3.5, and we will not repeat the descriptions here. Among those models, we will focus on our non-isotropic and nonseparable spatiotemporal correlation model as constructed

in Section 3.5.2. Using the same notations as before, it is written as

$$\begin{aligned} K(\mathbf{h},u) &= K_{ST}(\mathbf{h},u)+\delta_{\mathbf{h}=\mathbf{0}}K_0(u) \\ \text{where} \qquad K_{ST}(\mathbf{h},u) &= c_0\exp\left(-(\alpha\|\mathbf{h}\|)^{2\gamma}\right)\exp\left(-(\beta\tilde{u})^{2\eta}\right), \\ K_0(u) &= \exp\left(-(\tilde{\beta}u)^{2\tilde{\eta}}\right)-c_0\exp\left(-(\beta u)^{2\eta}\right) \\ \tilde{u} &= u+\frac{\mathbf{h}\cdot\mathbf{v}}{\|\mathbf{v}\|^2} \\ &= u+\frac{\|\mathbf{h}\|\cos(\theta-\theta_0)}{v} \end{aligned} \tag{4.19}$$

where $\alpha,\beta,\tilde{\beta}>0$ are the scale parameters for space and time respectively, $0<\gamma,\eta,\tilde{\eta}\leq 1$ are the shape parameters, $0\leq c_0\leq 1$ controls the nugget effect, $v\geq 0$ is the speed of the weather front and θ_0 is the azimuth angle for the direction of the movement of the weather front, and $\theta=\theta(\mathbf{s}_i,\mathbf{s}_j)$ is the azimuth angle between $\mathbf{s}_i$ and $\mathbf{s}_j$.

Note that a small drawback for all the correlation models that we have discussed in Section 3.5 is that we cannot model negative spatiotemporal correlations. In contrast to the correlation structure of wind power $Y(\mathbf{s},t)$ which is significantly positive for all lags of interest, empirical results show that the correlations of $Z(\mathbf{s},t)$ could be negative, especially in some cases where the forecast horizon is short[1]. However, the problem is not severe since negative correlations mainly occur when the correlations are very weak and so it would not affect the forecast results to a large extent. One could consider some appropriate combinations of valid correlation models which allow negative correlations (Gregori *et al.*, 2008), but in our case, we believe that the risk of over-fitting the noise with such models is larger than the benefit because there is no clear physical explanations for the occurrence of negative correlations in our application.

[1] One may wonder why the correlations of $Z(\mathbf{s},t)$ could depend on forecast horizons. In fact, this dependency arises because we would like to draw samples of $Z(\mathbf{s},t)$ for forecasting purposes. To do so, one need to model how $Z(\mathbf{s},t)$ would be correlated in future times, which depends on the horizon. This also relates to the fact that $Z(\mathbf{s},t)$ is latent, and its value depends on the transformation F. The forecast for F depends on the horizon, thus inducing the dependency on $Z(\mathbf{s},t)$.

4.3 Two-stage model: Parameter estimation

After specifying the structure of the two-stage model, we describe the details of parameter estimation in this section. Recall that the two-stage model for wind power generation is given by (4.6), namely,

$$Y(\mathbf{s},t) \;=\; \begin{cases} 0, & W(\mathbf{s},t) \le 0 \\ F^{-1} \circ \Phi(Z(\mathbf{s},t)), & 0 < W(\mathbf{s},t) < 1 \\ 1, & W(\mathbf{s},t) \ge 1 \end{cases} \tag{4.20}$$

Note that $Z(\mathbf{s},t)$ plays a role only when $W(\mathbf{s},t)$ lies between $(0,1)$. As a result, the estimation of the parameters governing $Z(\mathbf{s},t)$ can only be inferred from part of the wind power data where $Y(\mathbf{s},t) \in (0,1)$.

4.3.1 Regression coefficients for the dynamics of $W(\mathbf{s},t)$

First, we estimate the parameters in the model for $W(\mathbf{s},t)$ in (4.8). The usual way of estimating parameters in a probit model is by maximum likelihood. We maximize the likelihood of $I = \{I_A(\mathbf{s},t), I_B(\mathbf{s},t), I_C(\mathbf{s},t)\}$ according to the probit model. To facilitate an efficient estimation, we have to temporarily assume that the correlation of W is weak and assume the errors are spatially independent. Robinson (1982) and Poirier & Ruud (1988) show that assuming independence when calculating the maximizing the likelihood gives a consistent and asymptotically normal estimator, although such estimation is inefficient. The spatial correlation parameter ρ_ε will be estimated afterwards using the Stochastic Expectation-Maximization (SEM) algorithm (Marschner, 2001).

Recall that in (4.8), we model $W(\mathbf{s},t)$ as

$$W(\mathbf{s},t) = \mu_W(\mathbf{s},t) + \varepsilon(\mathbf{s},t), \qquad \varepsilon(\mathbf{s},t) \overset{i.i.d.}{\sim} N(0, \mathbf{\Sigma}_\varepsilon(\mathbf{s})) \tag{4.21}$$

where

$$\mu_W(\mathbf{s},t) = \beta_0 + \beta_1 I_A(\mathbf{s},t-1) + \beta_2 Y(\mathbf{s},t-1) I_B(\mathbf{s},t-1) + \beta_3 I_C(\mathbf{s},t-1) \tag{4.22}$$

We have chosen the correlation model to be an exponential one such that $\mathbf{\Sigma}_\varepsilon = \sigma_\varepsilon^2 \exp(-\|\mathbf{s}_i - \mathbf{s}_j\|/\rho_\varepsilon)$. To estimate the parameters using maximum likelihood,

we have to temporarily assume spatial independence and let $\mathbf{\Sigma}_\varepsilon = \sigma_\varepsilon^2 \delta_{i,j}$ where i, j are location indices for different wind farms. Under this assumption, we have

$$\begin{aligned} P(Y(\mathbf{s},t) \in A) &= P(W(\mathbf{s},t) \leq 0) \\ &= P\left(\varepsilon(\mathbf{s},t) \leq -\mu_W\right) \\ &= 1 - \Phi\left(\frac{\mu_W}{\sigma_\varepsilon}\right) \end{aligned} \tag{4.23}$$

Similarly,

$$\begin{aligned} P(Y(\mathbf{s},t) \in C) &= P(W(\mathbf{s},t) \geq 1) \\ &= P\left(\varepsilon(\mathbf{s},t) \geq 1 - \mu_W\right) \\ &= 1 - \Phi\left(\frac{1-\mu_W}{\sigma_\varepsilon}\right) \end{aligned} \tag{4.24}$$

From (4.23) and (4.24), we have

$$P(Y(\mathbf{s},t) \in B) = \Phi\left(\frac{\mu_W}{\sigma_\varepsilon}\right) + \Phi\left(\frac{1-\mu_W}{\sigma_\varepsilon}\right) - 1 \tag{4.25}$$

Thus the log likelihood is written as

$$\begin{aligned} \ln L \;=\; & \sum_{Y(\mathbf{s}_i,t_n)\in A} \ln\left[1 - \Phi\left(\frac{\mu_W(\mathbf{s}_i,t_n)}{\sigma_\varepsilon}\right)\right] + \sum_{Y(\mathbf{s}_i,t_n)\in C} \ln\left[1 - \Phi\left(\frac{1-\mu_W(\mathbf{s}_i,t_n)}{\sigma_\varepsilon}\right)\right] \\ & + \sum_{Y(\mathbf{s}_i,t_n)\in B} \ln\left[\Phi\left(\frac{\mu_W(\mathbf{s}_i,t_n)}{\sigma_\varepsilon}\right) + \Phi\left(\frac{1-\mu_W(\mathbf{s}_i,t_n)}{\sigma_\varepsilon}\right) - 1\right] \end{aligned} \tag{4.26}$$

Note that when $Y = 0$ or $Y = 1$, we can only know that $W < 0$ or $W > 1$ respectively. The large range that W is allowed to lie within leads to the difficulty in estimating the coefficients β_1 and β_3 in (4.22), and the log likelihood $\ln L$ is quite insensitive to changes in β_1 and β_3. Nevertheless, results shows that the exact values of β_1 and β_3 do not affect forecast values very much. This is reasonable, as $W(\mathbf{s},t)$ only governs the regimes for $Y(\mathbf{s},t)$. For example, $W(\mathbf{s},t) = -1$ and $W(\mathbf{s},t) = -10$ both imply $Y(\mathbf{s},t) = 0$ in the two-stage model.

4.3.2 Parameters in correlation model for $W(\mathbf{s}, t)$

In Section 4.3.1, we have estimated the regression coefficients β's in the model for the Gaussian process $W(\mathbf{s}, t)$. However, recall that in the estimation of β's using maximum likelihood, we have to temporarily assume that there are no spatial correlations among $W(\mathbf{s}, t)$. Now, having estimated β's, we proceed to estimate the correlation model Σ_ε for the residuals in (4.21). We model this using an exponential decay across spatial lags $\|\mathbf{s}_i - \mathbf{s}_j\|$, which is given in (4.10) as

$$\Sigma_\varepsilon = \sigma_\varepsilon^2 \exp\left(-\frac{\|\mathbf{s}_i - \mathbf{s}_j\|}{\rho_\varepsilon}\right) \tag{4.27}$$

In fact, the parameter σ_ε, i.e. the variance of errors, has already been estimated together with the regression coefficients β's by maximizing the likelihood. Due to the fact that $W(\mathbf{s}, t)$ is unobservable, the remaining parameter ρ_ε is estimated using the Stochastic Expectation-Maximization (SEM) algorithm (Dempster *et al.*, 1977). SEM is shown to be asymptotically equivalent to the maximum likelihood estimator using the EM algorithm (Marschner, 2001).

To carry out the SEM algorithm, we have to generate a series of simulated values for the latent Gaussian process $W(\mathbf{s}, t)$. This is the stochastic component in the SEM algorithm, and would be done using the Gibbs sampler (Geman & Geman, 1984). With the simulated series of $W(\mathbf{s}, t)$, one could then calculate the conditional expectation of the log likelihood of $W(\mathbf{s}, t)$ given the observed wind power $y(\mathbf{s}, t)$. By maximizing this expected log likelihood, the parameter ρ_ε will be updated and the iteration repeats.

To describe this algorithm in more details, note that conditional on the past observations $y(\mathbf{s}, t-1)$ at time $t-1$, the density of $W(\mathbf{s}, t)$ is Gaussian and is given by (4.9). Writing $f(W(\mathbf{s}, t)|y(\mathbf{s}, t-1), \rho_\varepsilon) = f(W_t|y_{t-1}, \rho_\varepsilon)$ and suppressing the dependence on $\mathbf{s}$, we have

$$f(W_t|y_{t-1}, \rho_\varepsilon) \sim N(\mu_W, \mathbf{\Sigma}_\varepsilon) \tag{4.28}$$

where μ_W is simply the conditional mean of $W(\mathbf{s}, t)$ as defined in (4.22). However, when y_t is observed, the updated density distribution of $W(\mathbf{s}, t)$ is truncated,

depending on weather $y_t = 0$, $y_t \in (0,1)$ or $y_t = 1$. In other words, we have

$$f(W_t|y_t, y_{t-1}, \rho_\varepsilon) \sim N_R(\mu_W, \mathbf{\Sigma}_\varepsilon) \tag{4.29}$$

where R denotes the domain of truncation. We have $R = (-\infty, 0]$, $(0,1)$ or $[1,\infty)$ when $y_t = 0$, $y_t \in (0,1)$ or $y_t > 1$ respectively. With this truncated conditional distribution, we can simulate a series of $\tilde{w}(\mathbf{s},t)$ by Gibbs sampling. We then maximize the expected log likelihood, that is,

$$Q(\rho_\varepsilon) = \sum_t \left[\left(\sum_{\tilde{w}} \log\left[f(\tilde{w}_t|y_t, y_{t-1}, \rho_\varepsilon) \right] \right) f(\tilde{w}_t|y_t, y_{t-1}, \rho_\varepsilon) \right] \tag{4.30}$$

with respect to ρ_ε, where we assume that conditional on the observed values $y(\mathbf{s}, t-1)$, $\tilde{w}(\mathbf{s},t)$ are only spatially correlated and are independent across time. The updated parameter ρ_ε is used to draw a new series of $\tilde{w}(\mathbf{s},t)$ according to the conditional distribution (4.29). The expected log likelihood is maximized again, and this algorithm is repeated until the estimation for ρ_ε converges.

The Gibbs sampler of $\tilde{w}(\mathbf{s},t)$ is obtained using the algorithm provided by Geweke (1991), which deals with random sampling from truncated distributions. Rodriguez-Yam *et al.* (2004) also give an alternative approach, which applies a transformation so as to decorrelate the samples. Conditional on the observed values $y(\mathbf{s},t)$, we aim at drawing samples of $\tilde{w}(\mathbf{s},t)$ from a multivariate truncated normal distribution as in (4.29). Following Rodriguez-Yam *et al.* (2004), this is equivalent to drawing samples of $\tilde{w}(\mathbf{s},t)$ such that

$$\tilde{w}(\mathbf{s},t) \sim N_R(\mu_W, \mathbf{\Sigma}_\varepsilon(\mathbf{s})), \qquad R := \{\tilde{\mathbf{w}} \in \Re^N : \mathbf{c} \le \mathbf{B}\tilde{\mathbf{w}} \le \mathbf{d}\} \tag{4.31}$$

where N is the dimension of the distribution and R denotes the region of truncation. In our situation, we have $\mathbf{B} = \mathbf{I}$, $\mathbf{c} = (c_1, \ldots, c_N)'$ and $\mathbf{d} = (d_1, \ldots, d_N)'$ so that for $i = 1, \ldots, N$,

$$(c_i, d_i) \quad = \quad \begin{cases} (-\infty, 0), & y(\mathbf{s}_i, t_i) = 0 \\ (0,1), & 0 < y(\mathbf{s}_i, t_i) < 1 \\ (1, \infty), & y(\mathbf{s}_i, t_i) = 1 \end{cases} \tag{4.32}$$

The Gibbs sampler is then applied by updating the last sample $\tilde{\mathbf{w}} = (\tilde{w}_1^{(t)}, \ldots, \tilde{w}_N^{(t)})'$ at time t in the Gibbs path. For $j = 1, \ldots, N$, we draw $\tilde{w}_j^{(t+1)}$ from the truncated normal distribution $N_{(c_j,d_j)}(\mu_j, \sigma_j)$ where μ_j and σ_j are the conditional mean and conditional variance of $\tilde{w}_j^{(t+1)}$ respectively. In other words, the drawing of $\tilde{w}_j^{(t+1)}$ is conditional on having the $N-1$ values of $\{\tilde{w}_1^{(t+1)}, \ldots, \tilde{w}_{j-1}^{(t+1)}, \tilde{w}_{j+1}^{(t)}, \ldots, \tilde{w}_N^{(t)}\}$. To ensure that the samples converge to the required multivariate truncated normal distribution, we draw M_b samples in the burn-in period. Then we draw a further M_s samples of $\tilde{w}(\mathbf{s}, t)$, which are then used to calculate the expectation in (4.30). The number of M_b and M_s depends mainly on the dimension of the distribution.

4.3.3 Smoothing parameters for the dynamics of $F(y|\ell, s^2)$

4.3.3.1 Minimizing CRPS

Next we estimate the five parameters $\theta, \alpha, \gamma, \phi_s$ and ϕ_v in the smoothing and forecast equations (4.11) and (4.12), which govern the dynamics of the location parameter ℓ and scale parameter s^2 in the truncated distribution $F(y|\ell, s^2)$. For the reasons of parsimony and practicality, we assume that the smoothing parameters are the same for all wind farms[1].

The cdf of the truncated normal distribution $F(y|\ell, s^2)$ corresponds to the distribution of the wind power generation $Y(\mathbf{s}, t)$ conditional on $Y(\mathbf{s}, t) \in (0, 1)$. As we aim at generating superior density forecasts, a good calibration of $F(y|\ell, s^2)$ is important. Instead of minimizing the RMSE of point forecasts, we minimize the mean CRPS of the forecast densities $f_{t+1|t}(y_{t+1}|\hat{\ell}_{t+1|t}, \hat{s}^2_{t+1|t})$ (Gneiting *et al.*, 2006), where the mean is calculated from the observations of $Y(\mathbf{s}, t) \in (0, 1)$ across all wind farms. Note that we could not use the observations when wind power is zero or one. It is because in such cases, the transformation F does not play a role in the two-stage model 4.6.

Note that we expect most of the temporal correlations of wind power generated at a fixed wind farm to be captured by the dynamics of $F(y|\ell, s^2)$. However,

[1]We have estimated an individual set of smoothing parameters separately for each wind farm, but the out-of-sample forecast results are even worse. We also investigated if the smoothing parameters exhibit certain spatial patterns so that we could reduce the number of different sets of parameters. However, there are no patterns across the wind farms, and we conclude that the best approach is to estimate one optimal set of smoothing parameters for all wind farms.

since ℓ and s^2 for each wind farm is smoothed individually, spatial correlations of wind power generation across different wind farms cannot be captured. We rely on the Gaussian process $Z(\mathbf{s}, t)$ to capture spatial correlations. In fact, we extend the role of $Z(\mathbf{s}, t)$ to capture spatiotemporal correlations that remain after accounting for the dynamics through $F(y|\ell, s^2)$. Remaining spatiotemporal correlations are expected to be captured by the Gaussian process $Z(\mathbf{s}, t)$ with an appropriate spatiotemporal correlation model, and subsequently through the relationship between Z and Y in (4.16).

4.3.3.2 Rescaling the location and scale parameters

With the estimated smoothing parameters $\theta, \alpha, \gamma, \phi_s$ and ϕ_v, we obtain the distribution $F(y|\ell, s^2)$ at each location $\mathbf{s}$ and time t. According to (4.16), as long as we have determined the dynamics of $F(y|\ell, s^2)$, we could always recover the latent Gaussian process $Z(\mathbf{s}, t)$ when $0 < Y(\mathbf{s}, t) < 1$ by

$$Z(\mathbf{s}, t) = \Phi^{-1} \circ F(Y(\mathbf{s}, t)), \quad Y(\mathbf{s}, t) \in (0, 1) \tag{4.33}$$

For the two-stage model to be consistent, one would expect that the distribution of $Z(\mathbf{s}, t)$, as obtained in (4.33), should be approximately Gaussian. This is necessary and is a critical assumption in the subsequent forecast generations. However, it is found that according to the smoothing parameters estimated in F, $Z(\mathbf{s}, t)$ deviates from the Gaussian distribution and this violates our assumption. To solve this problem, we introduce two rescaling parameters δ_ℓ and k_s which tune the shape of the distribution of $Z(\mathbf{s}, t)$. They tune the distribution by modifying the location and scale parameters so that

$$\ell \mapsto \ell + \delta_\ell \quad , \quad s \mapsto k_s s \tag{4.34}$$

Since this modification changes the location and scale parameters, it affects the forecast distributions as well. As a result, we use an iterative procedure to ensure that the smoothing parameters together with the rescaling parameters minimize two objective functions, namely

1. The mean CRPS of the forecast densities

2. A distance measure between the distribution of $Z(\mathbf{s}, t)$ at each wind farm with that of the standard normal distribution $N(0, 1)$

Minimizing two objective functions simultaneously require techniques such as successive Pareto optimization and genetic algorithms (Deb *et al.*, 2000). Another approach is to simply construct an aggregate objective function (AOF) which combine all the objective functions under consideration, and minimize with respect to the AOF (Roy, 1971). In our case, we do not consider special techniques in multi-objective optimization because of the computational burden required, which is not practical in our forecasts of a large number of time series. Also, we do not consider an AOF because it is unclear how one should choose the function that combines the two objectives in an optimal way so that superior wind power forecasts could be obtained, which is our ultimate goal. As a result, we use the method of iteration to obtain a well-estimated set of smoothing parameters and rescaling parameters. Given the parametric distribution $F(y|\ell, s^2)$ and denote all the smoothing parameters by Θ, the iterative procedure is described as follows:

1. Estimate $\hat{\Theta}_{(1)}$ by minimizing the mean CRPS of the density forecasts $f_{t+1|t}(y_{t+1})$, so that $\hat{\ell}_{(1)} = \ell(\hat{\Theta}_{(1)})$ and $\hat{s}^2_{(1)} = s^2(\hat{\Theta}_{(1)})$

2. Estimate $\hat{\delta}_\ell^{(1)}, \hat{k}_s^{(1)}$ by minimizing the Kolmogorov-Smirnov (K-S) test statistic for the normality of $z(\mathbf{s}, t)$, where the values of $z(\mathbf{s}, t)$ are obtained using (4.33) and the distribution F is parameterized by ℓ and s^2 such that $\ell = \hat{\ell}_{(1)} + \delta_\ell$ and $s^2 = k_s \hat{s}^2_{(1)}$. The solutions are denoted by $\hat{\delta}_\ell^{(1)}$ and $\hat{k}_s^{(1)}$

3. For $n = 2, 3, \ldots,$

   (a) Estimate $\hat{\Theta}_{(n)}$ by minimizing the mean CRPS of the density forecasts so that $\hat{\ell}_{(n)} = \ell(\hat{\Theta}_{(n)}) + \hat{\delta}_\ell^{(n-1)}$ and $\hat{s}^2_{(n)} = \hat{k}_s^{(n-1)} s(\hat{\Theta}_{(n)})$

   (b) Estimate $\hat{\delta}_\ell^{(n)}, \hat{k}_s^{(n)}$ by minimizing the K-S test statistic for the normality of $z(\mathbf{s}, t)$, where $z(\mathbf{s}, t)$ are obtained using (4.33) and the distribution F is parameterized by ℓ and s^2 such that $\ell = \hat{\ell}_{(n)} + \delta_\ell$ and $s^2 = k_s \hat{s}^2_{(n)}$

(c) Repeat (3a) and (3b) until either one of the objective functions begins to increase. We use this as a criterion because this would be consistent with the concept of Pareto optimality (Zitzler & Thiele, 1999)

In our application, this procedure works quite well and the results are able to converge stably towards some neighborhood within the parameter space. In general, the number of iteration n is around 2-5, and usually it is stopped by an increase in the second objective function. It has been verified that both in-sample and out-of-sample forecast results in wind power generation improve significantly when we include these rescaling parameters in the transformation F, thus showing its importance in the model. For example, we consider generating one-step ahead probability forecasts for individual wind power in Ireland using the two-stage model with empirical correlations (with details of implementation to be included later in this chapter.). We evaluate out-of-sample probability forecasts using mean CRPS, and Figure 4.1 demonstrate the boxplot results at the 64 wind farms with and without rescaling. The boxplot on the left corresponds to simply estimating the smoothing parameters by minimizing mean CRPS, and no rescaling parameters is used. The boxplot on the right corresponds to the use of rescaling parameters, which are estimated together with the smoothing parameters as described above. Clearly, it is critical to include the rescaling parameters because otherwise the conditional wind power distribution will not be well-described by $F(y|\ell, s^2)$. This severely violates the meaning of $F(y|\ell, s^2)$ in the two-stage model (4.6), thus giving poor forecast results.

4.3.4 Parameters in correlation model for $Z(\mathbf{s}, t)$

To estimate the parameters in the correlation model for Σ_Z, we first need to calculate the empirical spatiotemporal correlations of $Z(\mathbf{s}, t)$. Note that the values of $Z(\mathbf{s}, t)$ are in fact dependent on the forecast horizon h, as described earlier in the footnote on page 94. Given the location and scale parameters, we obtain the distribution F for the wind power generation. As given in (4.33), for an observation of wind power $y(\mathbf{s}, t)$, we can calculate the corresponding $z(\mathbf{s}, t)$ by

$$z(\mathbf{s}, t) = \Phi^{-1} \circ F_t(y(\mathbf{s}, t)), \quad y(\mathbf{s}, t) \in (0, 1) \tag{4.35}$$

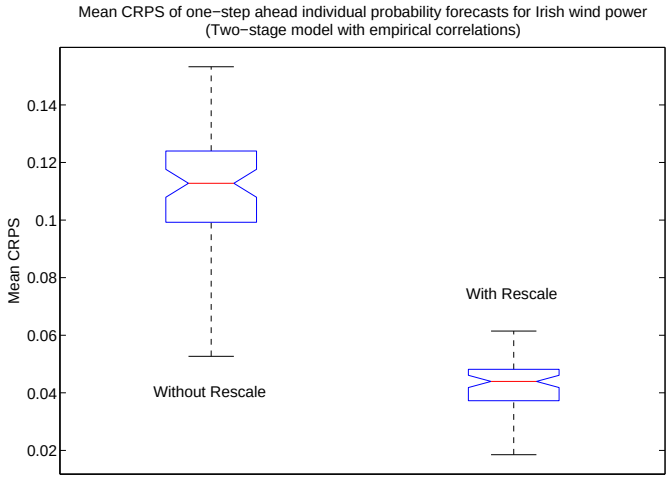


Figure 4.1: This figure shows the boxplot results of one-step ahead probability forecasts for 64 Irish wind power under mean CRPS, where we consider the two-stage model with empirical correlations (with details of implementation to be included later in this chapter.). The boxplot on the left corresponds to simply estimating the smoothing parameters by minimizing mean CRPS, and no rescaling parameters is used. Referring to the results in Table 5.7 in Chapter 5, we have $\theta = 236.82, \alpha = 0.19, \phi_s = 0.77, \gamma = 0.0003, \phi_v = 0.005$ and $\delta_\ell = 0, k_s = 1$, i.e. no rescaling. The boxplot on the right corresponds to the use of rescaling parameters, which are estimated together with the smoothing parameters as described above. In such case, we use $\theta = 21.12, \alpha = 1.05, \phi_s = 0.00, \gamma = 0.0092, \phi_v = 0.017$ and $\delta_\ell = -0.004, k_s = 0.77$ as shown in Table 5.7. Clearly, it is critical to include the rescaling parameters in the model.

where we denote $F_t = F(y|\ell_t, s_t^2)$. Clearly $Z(\mathbf{s}, t)$ depends on F_t, which evolves with forecast horizon[1]. In particular, for h-step ahead forecasts, the relationship in (4.35) becomes

$$z(\mathbf{s}, t+h|t) = \Phi^{-1} \circ F_{t+h|t}(y(\mathbf{s}, t+h)), \quad y(\mathbf{s}, t+h) \in (0,1) \tag{4.36}$$

As a result, we need to calculate the empirical correlations of $Z(\mathbf{s}, t+h|t)$ for each forecast horizon $h = 1, 2, \ldots$ and estimate the parameters in the correlation models accordingly. One would expect that the magnitude of correlations between $Z(\mathbf{s}, t+h|t)$ to increase for larger horizons h, as it would be more difficult to

[1]For instance, one expects that the density corresponding to F_t to be sharper for short forecast horizons, and more disperse for longer horizons.

account for the dynamics of $Y(\mathbf{s}, t+h|t)$ using $F_{t+h|t}$ and so the errors would be more correlated. Thus the role of the spatiotemporal models become more critical for h-step ahead forecasts when h is large.

Given the estimated parameters for the dynamics of ℓ and s^2, the estimated h-step ahead distributions $\hat{F}_{t+h|t}$ are obtained. We then calculate the h-step ahead estimated values of $z(\mathbf{s}, t+h|t)$ conditional on $y(\mathbf{s}, t+h) \in (0,1)$ using (4.36). There will be missing values in $z(\mathbf{s}, t+h|t)$ when $y(\mathbf{s}, t+h) = 0$ or 1. To handle this situation, one may calculate the spatiotemporal correlations of $z(\mathbf{s}, t+h|t)$ by considering only pairwise observations which are available, but then a serious problem may occur since the estimated correlation matrix may not be positive definite (Little & Rubin, 2002). We consider a simple method to handle this problem, which is to simply impute random variables from $N(0,1)$ to replace the missing values. Imputations are commonly applied in analysis with missing data, and results are usually quite robust (Henderson *et al.*, 2000; Rubin, 1987). It has also been checked that in our case, the resulting correlations are not largely altered when one impute different values for the missing $z(\mathbf{s}, t+h|t)$.

Now, denote the estimated values of $z(\mathbf{s}, t+h|t)$ by $\hat{Z}$ in general, where the dependence on h is suppressed. After obtaining these values of $\hat{Z}$, we could calculate the empirical spatiotemporal correlations $\mathbf{\Sigma}_{\hat{Z}}$. Note that we would need to determine a certain number of temporal lags to be included in the spatiotemporal correlations. This number of temporal lags will correspond to a time frame where wind power across different wind farms demonstrate interesting temporal interactions. Also, one would choose only a reasonable number of lags to avoid computing a spatiotemporal correlation matrix of a huge dimension.

On the other hand, we could also fit the empirical correlations to various models, in particular, those that we have discussed in Chapter 3. We estimate the parameters in the correlation models by minimizing the weighted least squares (WLS) as described in (3.44). The fitted correlations will then be applied to generate h-step ahead forecasts of $z(\mathbf{s}, t+h|t)$, and eventually the wind power generation $y(\mathbf{s}, t+h|t)$.

Some examples of the empirical correlations of $Z(\mathbf{s}, t+h|t)$ are shown in Figures 4.2 and 4.3. We consider autocorrelation at a fixed location as well as cross correlations at different locations. Two horizons at $h = 4$ (i.e. one hour)

and $h = 12$ (i.e. three hours) are chosen for comparison, and this shows that spatiotemporal correlations of $z(\mathbf{s}, t + h|t)$ become increasingly significant for larger forecast horizon h. It is observed that the cross correlations of $\hat{Z}$ at short forecast horizons are usually quite weak, and could sometimes be negative. In fact, we see that there are no obvious structures that exist in the cross correlations at short forecast horizons $h \leq 4$. Thus, in certain cases, we would not even model such correlations, and the use of empirical correlations will suffice. Forecast results are even worse if one applies a correlation model at such short horizons[1].

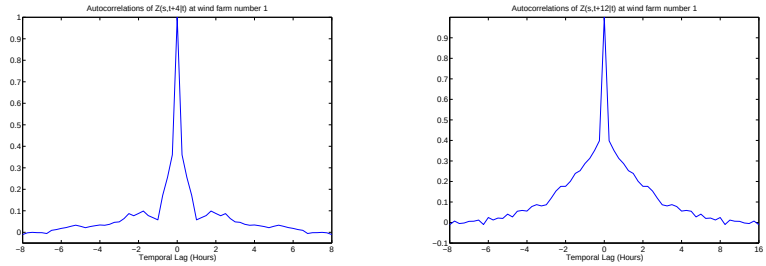


Figure 4.2: Decay of empirical autocorrelations of $z(\mathbf{s}, t + h|t)$ at $h = 4$ (Left) and $h = 12$ (Right).

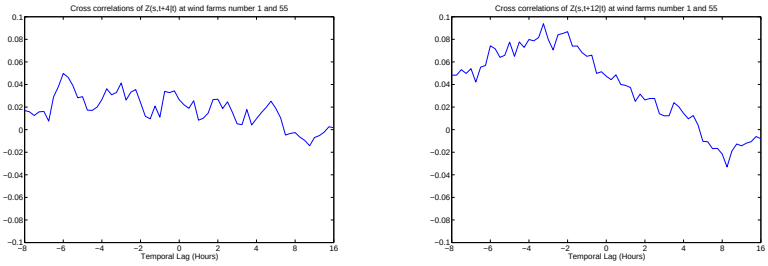


Figure 4.3: Decay of empirical spatiotemporal cross correlations of $z(\mathbf{s}, t+h|t)$ at $h = 4$ (Left) and $h = 12$ (Right). It is observed that the cross correlations of $\hat{Z}$ at short forecast horizons are usually quite weak, and could sometimes be negative. In fact, we see that there are no obvious structures that exist in the cross correlations at short forecast horizons $h \leq 4$. Thus, for practical forecasting purposes, we would not even model such correlations, and the use of empirical correlations will suffice. Forecast results are even worse if one applies a correlation model at such short horizons.

[1]The use of empirical correlations for short forecast horizons very much depends on the length of the training data. For long training sets, empirical estimations are superior and it could be better to use empirical estimates directly. For short training set, we see that modeling the correlations helps to improve forecasts. Thus, one should adapt to the available data and model accordingly.

4.3.5 Summary of two-stage model

There are many parameters involved in the description of the three major components in the two-stage model as stated in Section 4.2. To clarify all the procedures, Table 4.1 summarizes the various parameters in each of the components and how they are estimated.

Component	**1^{st} latent Gaussian process, $W(\mathbf{s},t)$**	**Model for transformation, $F(\cdot)$**	**2^{nd} latent Gaussian process, $Z(\mathbf{s},t)$**
Mean			
Model	Probit regression in (4.8)	Exponential smoothing in (4.11)	Zero mean by definition
Parameters	$\beta_0 - \beta_3, \sigma_\varepsilon$	$\alpha, \phi_s, \delta_\ell$	–
Estimation	Maximum likelihood	Minimize CRPS and $K-S$ distance	–
Variance/ Correlation			
Model	Exponential correlation in (4.10)	Exponential smoothing in (4.11)	Various spatiotemporal correlation models as discussed in Section 3.5
Parameters	ρ_ε	$\gamma, \phi_v, \theta, k_s$	Depend on correlation models (Refer to Section 3.5)
Estimation	Stochastic Expectation Maximization (SEM)	Minimize CRPS and $K-S$ distance	Weighted least squares (WLS)

Table 4.1: Summary of different components in the two-stage model, parameters involved and methods of estimations.

4.4 Generating forecasts through simulations

The two-stage model is now fully specified, and the details of parameter estimation have been described in the previous sections. In this section, we move on to describe the generation of wind power forecasts using the two-stage model. This boils down to the problem of drawing samples from the latent Gaussian processes $W(\mathbf{s},t)$ and $Z(\mathbf{s},t)$, conditional on the history of observations.

Using the exponential smoothing methods as described in Section 4.3.3, the dynamics of the location and scale parameters in the distribution $F(y|\ell, s^2)$ could be modeled. It follows that the h-step ahead forecasts of $F_{t+h|t} = F(y|\hat{\ell}_{t+h|t}, \hat{s}^2_{t+h|t})$

are obtained. We then have to draw samples of $w(\mathbf{s}, t+h|t)$ and $z(\mathbf{s}, t+h|t)$ from the appropriate conditional Gaussian distributions, and the resulting samples of h-step ahead forecasts of wind power generation $y(\mathbf{s}, t+h|t)$ are given by the two-stage model as in (4.6), that is,

$$y(\mathbf{s}, t+h|t) = \begin{cases} 0, & w(\mathbf{s}, t+h|t) \leq 0 \\ F^{-1}_{t+h|t} \circ \Phi(z(\mathbf{s}, t+h|t)), & 0 < w(\mathbf{s}, t+h|t) < 1 \\ 1, & w(\mathbf{s}, t+h|t) \geq 1 \end{cases} \quad (4.37)$$

where Φ is the univariate standard normal distribution function. This simulation approach has the advantage of generating probability forecasts easily, but is relatively time consuming. Expected values of wind power forecasts are then taken as the mean of the forecast densities. To draw $w(\mathbf{s}, t+h|t)$ and $z(\mathbf{s}, t+h|t)$ from the appropriate conditional Gaussian distributions, we need to calculate the conditional mean and conditional covariance of the respective Gaussian processes as described in the following sections.

4.4.1 Drawing samples of $W(\mathbf{s}, t)$

Compared with $z(\mathbf{s}, t+h|t)$, drawing samples of $w(\mathbf{s}, t+h|t)$ is relatively easy since we assume that there are only spatial correlations among $W(\mathbf{s}, t)$, and they are independent across time. Once we have estimated the correlation parameter ρ_ε using the SEM algorithm, the covariance matrix $\mathbf{\Sigma}_\varepsilon$ for $W(\mathbf{s}, t)$ is fixed and is given by (4.10). For one-step ahead forecasts, we could directly apply the model for $W(\mathbf{s}, t)$ in (4.8) and obtain

$$w(\mathbf{s}, t+1|t) = \beta_0 + \beta_1 I_A(\mathbf{s}, t) + \beta_2 y(\mathbf{s}, t) I_B(\mathbf{s}, t) + \beta_3 I_C(\mathbf{s}, t) \quad (4.38)$$

For $h > 1$, the generation of h-step ahead forecasts $w(\mathbf{s}, t+h|t)$ will require the estimation of $(h-1)$-step ahead forecasts $y(\mathbf{s}, t+h-1|t)$, which would have already been obtained because samples of $w(\mathbf{s}, t+h-1|t)$ and $z(\mathbf{s}, t+h-1|t)$ are generated in the previous steps already. In other words, h-step ahead forecasts

$w(\mathbf{s}, t+h|t)$ are simply obtained by iterating (4.38), i.e.

$$\begin{aligned} w(\mathbf{s}, t+h|t) &= \beta_0 + \beta_1 I_A(\mathbf{s}, t+h-1|t) + \beta_2 y(\mathbf{s}, t+h-1|t) I_B(\mathbf{s}, t+h-1|t) \\ &\quad + \beta_3 I_C(\mathbf{s}, t+h-1|t) \end{aligned} \tag{4.39}$$

where $y(\mathbf{s}, t+h-1|t)$ is the $(h-1)$-step ahead forecast for wind power, and the forecast values of the indicator variables I_A, I_B and I_C depend on weather the wind power forecast $y(\mathbf{s}, t+h-1|t)$ is equal to zero, lies within $(0, 1)$, or equal to one, respectively.

4.4.2 Drawing samples of $Z(\mathbf{s}, t)$

4.4.2.1 Spatiotemporal correlation matrices

Drawing samples of $z(\mathbf{s}, t+h|t)$ is slightly more complicated due to the spatiotemporal correlations, which leads to a correlation matrix with a higher dimension. As mentioned previously, we need to decide a maximum temporal lag in such case. This maximum temporal lag in the spatiotemporal correlation matrix will correspond to a time frame where wind power across different wind farms demonstrate interesting temporal interactions. Also, one would choose only a reasonable number of lags to avoid computing a spatiotemporal correlation matrix of a huge dimension, as in the forecast procedure one need to invert the correlation matrices.

Following the discussions on Page 79 in Chapter 3, we decide that a maximum temporal lag of 8 hours is reasonable and captures most of the structure in the spatiotemporal correlations. For instance, this period demonstrates clearly the shift of the position of maximum correlations between wind power generated at two wind farms lying along the direction of the movement of weather front. Now, as described in Section 4.3.4, we first obtain the estimated values $z(\mathbf{s}, t+h|t)$ in the learning data using (4.36). Empirical spatiotemporal correlations are calculated based on $z(\mathbf{s}, t+h|t)$, and results could be fitted to the spatiotemporal correlation models by minimizing the weighted least squares.

Now, let us denote the fitted (or empirical) spatial correlations at fixed temporal lags $\tau = t_i - t_j$ by $\hat{\mathfrak{C}}(t_i, t_j)$, so that $\hat{\mathfrak{C}}$'s are $N \times N$ correlation matrices and

N is the number of wind farms in the data set. In other words, $\hat{\mathfrak{C}}$'s are the spatial correlations between $Z(\mathbf{s}, t)$ at the N wind farms, which could be written as

$$\hat{\mathfrak{C}}(t_i, t_j) = \begin{pmatrix} \hat{C}_{11}(t_i, t_j) & \cdots & \hat{C}_{1N}(t_i, t_j) \\ \vdots & \ddots & \vdots \\ \hat{C}_{N1}(t_i, t_j) & \cdots & \hat{C}_{NN}(t_i, t_j) \end{pmatrix}_{N \times N} \tag{4.40}$$

and $\hat{C}_{pq}(t_i, t_j) = \text{Corr}(Z(\mathbf{s}_p, t_i), Z(\mathbf{s}_q, t_j))$. In fact, as our correlation model is stationary, we have $\hat{\mathfrak{C}}(t_i, t_j) = \hat{\mathfrak{C}}(t_i - t_j) = \hat{\mathfrak{C}}(\tau)$. To build the larger spatiotemporal correlation matrix, we could then simply stack up the spatial correlation matrices in (4.40) with different temporal lags. In particular, with the maximum temporal lag chosen to be 8 hours, i.e. 32 time steps, the fitted (or empirical) spatiotemporal correlation matrix for $Z(\mathbf{s}, t)$ is given by

$$\begin{aligned} \mathbf{\Sigma}_Z &= \begin{pmatrix} \hat{\mathfrak{C}}(t-p, t-p) & \cdots & \hat{\mathfrak{C}}(t-p, t) \\ \vdots & \ddots & \vdots \\ \hat{\mathfrak{C}}(t, t-p) & \cdots & \hat{\mathfrak{C}}(t, t) \end{pmatrix}_{(p+1)N \times (p+1)N} \\ &= \begin{pmatrix} \hat{\mathfrak{C}}(0) & \cdots & \hat{\mathfrak{C}}(-p) \\ \vdots & \ddots & \vdots \\ \hat{\mathfrak{C}}(p) & \cdots & \hat{\mathfrak{C}}(0) \end{pmatrix}_{(p+1)N \times (p+1)N} \end{aligned} \tag{4.41}$$

where $p = 32$ in this case.

4.4.2.2 Conditional mean and covariance

With the spatiotemporal correlation matrix $\mathbf{\Sigma}_Z$ in (4.41), consider the drawing of $z(\mathbf{s}, t + h|t)$ where $h < p$. Since $Z(\mathbf{s}, t + h|t)$ is a Gaussian process, it suffices to obtain the conditional mean and conditional covariance of $Z(\mathbf{s}, t + h|t)$. Let us stack the values of $Z(\mathbf{s}, t)$ at time t and different locations $\mathbf{s}_j$ by writing

$$\mathbf{Z}(t) = (Z(\mathbf{s}_1, t), \ldots, Z(\mathbf{s}_N, t))' \tag{4.42}$$

where N is the number of wind farms. Considering only a maximum temporal lag of $p = 32$ in the spatiotemporal correlation matrix (4.41) means that we regard $Z(\mathbf{s}, t)$ as a multivariate Gaussian random variable with finite dimension

of $(p+1)N$. For h-step ahead forecasts made at time t, we focus on the joint distribution of $\tilde{\mathbf{Z}}$ where

$$\tilde{\mathbf{Z}} = (\mathbf{Z}(t+h-p), \ldots, \mathbf{Z}(t+h))' \tag{4.43}$$

This would cover the required window of temporal lags that we are interested in. To discriminate the observed values in the past (up to present time t) with those to be predicted in the future, we further partition (4.43) in the form of

$$\begin{aligned} \tilde{\mathbf{Z}} &= (\mathbf{Z}(t+h-p), \ldots, \mathbf{Z}(t), |\mathbf{Z}(t+1), \ldots, \mathbf{Z}(t+h))' \\ &= (\tilde{\mathbf{Z}}_{past}, |\tilde{\mathbf{Z}}_{future})' \end{aligned} \tag{4.44}$$

In other words, our target is to generate samples of $\tilde{\mathbf{Z}}_{future} = (\mathbf{Z}(t+1), \ldots, \mathbf{Z}(t+h))'$ conditional on the observed values $\tilde{\mathbf{Z}}_{past} = (\mathbf{Z}(t+h-p), \ldots, \mathbf{Z}(t))'$. The correlation matrix $\tilde{\mathbf{Z}}$ could be partitioned accordingly into four sub-matrices such that

$$\boldsymbol{\Sigma}_{\tilde{\mathbf{Z}}} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \tag{4.45}$$

Using the notation in (4.41), the explicit expressions of the sub-matrices are simply given by

$$\begin{aligned} \boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_{\tilde{\mathbf{Z}}_{past}, \tilde{\mathbf{Z}}_{past}} &= \begin{pmatrix} \hat{\mathcal{C}}(0) & \cdots & \hat{\mathcal{C}}(h-p) \\ \vdots & \ddots & \vdots \\ \hat{\mathcal{C}}(p-h) & \cdots & \hat{\mathcal{C}}(0) \end{pmatrix}_{(p-h+1)N \times (p-h+1)N} \\ \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{\tilde{\mathbf{Z}}_{past}, \tilde{\mathbf{Z}}_{present}} &= \begin{pmatrix} \hat{\mathcal{C}}(h-p-1) & \cdots & \hat{\mathcal{C}}(-p) \\ \vdots & \ddots & \vdots \\ \hat{\mathcal{C}}(-1) & \cdots & \hat{\mathcal{C}}(-h) \end{pmatrix}_{(p-h+1) \times hN} \\ \boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{\tilde{\mathbf{Z}}_{present}, \tilde{\mathbf{Z}}_{past}} &= \begin{pmatrix} \hat{\mathcal{C}}(p-h+1) & \cdots & \hat{\mathcal{C}}(1) \\ \vdots & \ddots & \vdots \\ \hat{\mathcal{C}}(p) & \cdots & \hat{\mathcal{C}}(h) \end{pmatrix}_{hN \times (p-h+1)N} \\ \boldsymbol{\Sigma}_{22} = \boldsymbol{\Sigma}_{\tilde{\mathbf{Z}}_{present}, \tilde{\mathbf{Z}}_{present}} &= \begin{pmatrix} \hat{\mathcal{C}}(0) & \cdots & \hat{\mathcal{C}}(-h+1) \\ \vdots & \ddots & \vdots \\ \hat{\mathcal{C}}(h-1) & \cdots & \hat{\mathcal{C}}(0) \end{pmatrix}_{hN \times hN} \end{aligned} \tag{4.46}$$

The conditional mean of $\tilde{\mathbf{Z}}_{future}$ is then given by standard results on multivariate Gaussian distributions, which is written as

$$\mu_{\tilde{\mathbf{Z}}_{future}} = \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\tilde{\mathbf{Z}}_{past} \tag{4.47}$$

Similarly, the conditional covariance of $\tilde{\mathbf{Z}}_{future}$ is given by

$$\boldsymbol{\Sigma}_{\tilde{\mathbf{Z}}_{future}} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \tag{4.48}$$

Using (4.47) and (4.48), we can draw samples from the multivariate Gaussian distribution $N(\mu_{\tilde{\mathbf{Z}}_{future}}, \boldsymbol{\Sigma}_{\tilde{\mathbf{Z}}_{future}})$ and obtain values of $\tilde{\mathbf{Z}}_{future}$. Note that for the purpose of h-step ahead forecasts, we do not require all future values in $\tilde{\mathbf{Z}}_{future}$ from time $t+1$ to time $t+h$. We are in fact only interested in time $t+h$, i.e. the value of $\mathbf{Z}(t+h)$ in (4.44). Finally, we show a simple flow diagram of the random sampling of $w(\mathbf{s}, t+h|t)$ and $z(\mathbf{s}, t+h|t)$ in Figure 4.4, so as to clarify the procedure as described in this section.

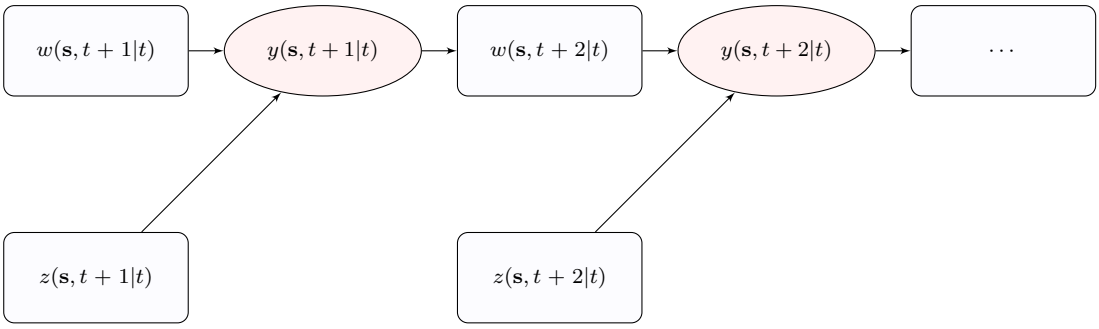


Figure 4.4: Flow diagram showing the sequence for drawing samples $w(\mathbf{s}, t+h|t)$ and $z(\mathbf{s}, t+h|t)$, conditional on observations up to time t. These samples can be applied in the two-stage model (4.37) to obtain the wind power forecasts $y(\mathbf{s}, t+h|t)$. As shown in (4.39), the drawing of $(h+1)$-step ahead forecasts $w(\mathbf{s}, t+(h+1)|t)$ requires the forecasts $y(\mathbf{s}, t+h|t)$, which would have already been obtained in the previous steps.

4.4.3 Samples of aggregated forecasts

The values of aggregated wind power forecasts is based on the individual forecast samples $y(\mathbf{s}, t+h|t)$. We obtain the normalized aggregated wind power forecast samples y_A by

$$y_A(t+h|t) = \frac{\sum_{j=1}^{N} \omega_j y(\mathbf{s}_j, t+h|t)}{\sum_{j=1}^{N} \omega_j} \tag{4.49}$$

where ω_j is the capacity of wind farm at location $\mathbf{s}_j$, and N is the total number of wind farms in the portfolio. The point forecast for the normalized aggregated wind power $\hat{Y}_A(t+h|t)$ is then taken as the mean of the samples $y_A(t+h|t)$, and the probabilistic forecast is the empirical density function that corresponds to the samples of $y_A(t+h|t)$, which is estimated using the build-in function in MATLAB 'ecdf'.

Chapter 5

Application to Irish Wind Data

5.1 Wind data

We have described various approaches for forecasting aggregated wind power in Chapter 2. We also build spatiotemporal models for forecasting individual wind power in Chapter 3 and Chapter 4. Now, in this chapter and the next, we are going to demonstrate how these models and approaches could be applied to data sets of wind power generation. We would compare the results of different models, analyze the results and evaluate the forecasts.

In this chapter, we consider wind power generation from a portfolio of 64 wind farms in Ireland, where the data is provided by Eirgrid Plc. This data set spans for approximately six months from 13-Jul-2007 to 01-Jan-2008, and the observations on wind power are recorded every 15 minutes. This comprises of 16512 observations at each of the 64 wind farms. The locations of the wind farms are shown in Figure 5.1, and they are mostly situated on the north-west coast where wind resources are most abundant. The details of this data set is summarized in Table 5.1. Note that Arklow Banks is the only offshore wind farm in the data set, which is located off the east coast of Ireland as shown in Figure 5.1.

No.	Name	Lat (N)	Lon (W)	Capacity (MW)	Mean	Variance
1	Bellacorick	54.14	9.54	6.45	.20	.07
2	Cark	54.89	7.89	15	.32	.08
3	Corrie Mountain / Spion Kop	54.13	8.15	6	.34	.10
4	Cronalaght	55.07	8.22	4.98	.39	.09
5	Crockahenny	55.15	7.27	5	.43	.11
6	Currabwee	51.68	9.11	4.62	.26	.08
7	Drumlough Hill	55.20	7.44	4.8	.35	.11
8	Golagh	54.70	7.93	15	.19	.04
9	Inverin	53.26	9.37	3.3	.28	.08
10	Kilronan	54.08	8.14	5	.27	.07
11	Culliagh	54.87	7.89	11.88	.32	.10
12	Milane Hill	51.69	9.21	5.94	.24	.08
13	Tursillagh	52.32	9.60	22	.29	.09
14	Beenageeha	52.33	9.57	3.96	.26	.09
15	Arklow Banks	52.89	5.92	25.2	.32	.11
16	Ballywater	52.54	6.25	31.5	.23	.08
17	Carnsore	52.18	6.37	11.9	.29	.08
18	Derrybrien	53.09	8.60	60	.20	.05
19	Kingsmountain	54.11	8.65	23.8	.21	.05
20	Meentycat	54.87	7.85	71	.26	.07
21	Glanlee	51.93	9.31	29.8	.21	.07
22	Cronelea	52.78	6.56	7.55	.37	.11
23	Coomagearlahy	51.94	9.32	42.5	.21	.06
24	Booltiagh	52.78	9.24	19.5	.17	.05
25	Kilbranish (Greenoge)	52.66	6.74	5	.39	.11
26	Ratrussan	54.00	7.10	70	.16	.03
27	Corneen	54.14	7.65	3	.35	.10
28	Black Banks I	54.06	8.20	3.4	.31	.08
29	Inis Mean	53.07	9.60	0.675	.29	.09
30	Anarget	54.74	8.16	3.1	.41	.13
31	Meenadreen	54.67	8.00	3.4	.24	.06
32	Burtonport	54.99	8.44	0.66	.41	.13
33	Meenanilta	54.85	7.83	5	.22	.06
34	Raheen Barr	53.91	9.34	18.7	.26	.07
35	Cuillalea (West of Kiltimagh)	53.85	9.07	3.4	.28	.07
36	Mienvee	52.65	8.15	0.7	.28	.09
37	Curraghgraigue	52.79	8.10	2.55	.28	.08
38	Sonnagh Old	53.13	8.64	7.65	.30	.09
39	Dundalk IT	54.00	6.39	0.5	.05	.03
40	Beale Hill	52.57	9.64	4.29	.30	.09
41	Largan Hill	53.95	8.60	5.94	.20	.06
42	Gartnaneane	53.95	6.92	15	.28	.07
43	Moanmore	52.67	9.49	12.6	.27	.09
44	Coomatallin	51.66	9.09	5.95	.36	.12
45	Mount Eagle	52.24	9.32	5.1	.30	.09
46	Altagowlan	54.12	8.14	7.65	.33	.08
47	Moneenatieve	54.14	8.15	3.96	.44	.13
48	Black Banks II	54.10	8.15	6.8	.28	.07
49	Gneeves	52.01	9.12	9.35	.30	.08
50	Ballinveny	52.83	7.94	2.55	.28	.08
51	Ballinlough	52.83	8.01	2.55	.32	.09
52	Geevagh	54.14	8.27	4.95	.19	.06
53	Taurbeg	52.25	9.14	26	.27	.08
54	Beam Hill	55.20	7.43	14	.31	.09
55	Carrig-Skehanagh	53.03	7.98	6.8	.26	.07
56	Kealkil (Curraglass)	51.81	9.32	8.5	.20	.07
57	Lahanaght Hill	51.65	9.22	4.25	.29	.09
58	Dunmore	53.79	6.53	1.7	.39	.10
59	Kilvinane	51.71	8.99	4.5	.25	.09
60	Sorne Hill	55.13	7.37	31.5	.31	.09
61	Richfield	52.22	6.59	27.1	.26	.09
62	Carrane Hill	54.14	8.23	3.4	.37	.10
63	Tournafulla	52.39	9.15	7.5	.32	.10
64	Lackan	54.25	9.07	6	.35	.10

Table 5.1: Summary of wind power data from 64 wind farms in Ireland

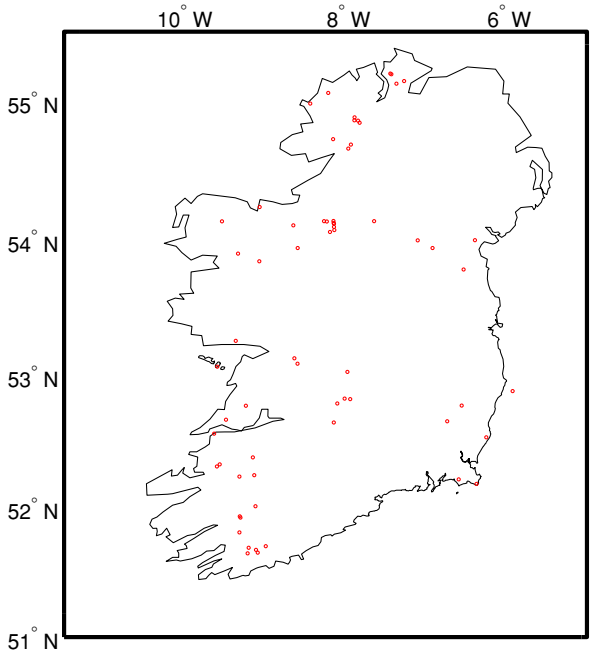


Figure 5.1: This figure shows the locations of 64 wind farms in Ireland. The wind farms are mostly situated on the north-west coast where wind resources are most abundant. Note that Arklow Banks, which is located off the east coast of Ireland, is the only offshore wind farm in the data set.

5.1.1 Research questions using this data set

With the wind power data in Ireland, one could investigate many interesting questions regarding with wind power forecasts. In particular, we want to analyze the results in certain particular aspects. These include:

1. In the forecast of aggregated wind power, how do the exponential smoothing methods compare with other ARIMA-GARCH models? Is it better to forecast aggregated wind power by directly modeling the univariate time series? Could we outperform the univariate models by considering spatiotemporal forecasts and summing up the results from each individual wind farm?

2. In Chapter 3, we describe the method of spatiotemporal kriging for forecasting individual wind power. In the spatiotemporal kriging models, one could use either a homoscedastic or heteroscedastic model for the variances at individual wind farms. How do these performances differ?

3. In the case of Ireland, weather fronts generally propagate from north-west to south-east (Fox *et al.*, 2007). Could our anisotropic correlation model capture this feature? How do the movement of weather fronts affect forecast results at different wind farm locations?

4. In the spatiotemporal kriging approach, the models cannot deal with the probability mass at zero wind power. The two-stage model introduced in Chapter 4 is constructed to handle this issue. How do the performances between the spatiotemporal kriging models and the two-stage models compare? Do the two-stage models perform better at forecasting extreme wind power?

These questions are, of course, only part of the many interesting problems concerning wind power forecasts. We aim at answering the above questions using the Irish data set in this chapter. More aspects on the robustness of the models as well as the forecast performances across different periods could be analyzed using a larger data set. This will be achieved using the Danish wind data described in the next chapter.

5.1.2 Training and testing data

To answer the above questions, we will need to fit the various models to the wind power data and evaluate the forecast performances. In all the subsequent analysis, we will dissect the data into a fixed training set for parameter estimation and a fixed testing set for out-of-sample forecast evaluation. Out of the 16512 observations from 13-Jul-2007 to 01-Jan-2008, we choose the first 11008 observations to be the training data. This comprises of approximately four months of data from July to October, 2007. The remaining 5504 observations will be used as the testing data for forecast evaluations, and this testing set has about two months of data in November to December, 2007. For the Irish data, we will not consider out-of-sample forecasts in terms of spatial locations, and so observations from all the 64 wind farms will be used in the training data. Instead, we will investigate this problem in Chapter 6 with a larger data set in Denmark.

In order to facilitate comparisons between data sets with different capacities, we normalize the individual wind power data by dividing by the capacity of the corresponding wind farm. Similarly, for aggregated wind power, we add up the actual wind power generation at all of the 64 wind farms, and divide the result by the total capacity, which is 792.355 MW in our case. By doing so, all wind power observations at an individual wind farm will be expressed as a value between $[0, 1]$, and observations of aggregated wind power across a portfolio of wind farms will be within $(0, 1)$ in practice. This could allow us to easily compare the forecasts across wind farms with different capacities. We could even compare forecast performances across different data sets, for instance, between the Irish and the Danish wind data.

In the following sections, we attempt to answer the questions posed in Section 5.1.1 by fitting the various models to the Irish wind data. First, we consider models for aggregated wind forecasts as described in Chapter 2. Then, we move on to fit the data to the spatiotemporal kriging models in Chapter 3 and study their forecast performances. Finally, we analyze the data using the two-stage model in Chapter 4.

5.2 Aggregated Forecasts

In this section, we consider forecasts for aggregated wind power. As described in Chapter 2, we analyze the following two different modeling approaches:

ARIMA-GARCH models: As in Section 2.1, we first normalize the aggregated data approximately using the logit transformation, and then model the transformed data by ARIMA-GARCH models with Gaussian innovations

Exponential smoothing: As in Section 2.2, we apply exponential smoothing on the original data, and the data are modeled using a truncated normal distribution with location and scale parameters obtained from exponential smoothing

5.2.1 Competing models

To analyze the role of heteroscedasticity in the forecast performances, for each of the above approaches, we consider one homoscedastic model and one heteroscedastic model. In other words, for the first approach using ARIMA-GARCH models, we will consider a plain ARIMA model and a full ARIMA-GARCH model. For the second approach using exponential smoothing, we will consider one with the smoothing of the location parameter only, and another with simultaneous smoothing of both the location and the scale parameters.

After transforming the data using the logit transformation in (1.4), the orders of the ARIMA-GARCH models are then selected by minimizing the BIC using these transformed data in the training set. For the plain ARIMA model, the optimal choice is the ARIMA(2,1,3) model. For the full ARIMA-GARCH model, results indicate that an ARIMA(4,1,3)-GARCH(1,1) model minimizes the in-sample BIC. Parameters are then estimated using the 'garchfit' toolbox in MATLAB, under the assumption of Gaussian innovations. For the exponential smoothing methods, we estimate the smoothing parameters by maximizing the truncated normal likelihood.

Using the notation introduced in Chapter 2, the four competing models that we consider for aggregated wind power forecasts in Ireland are listed below:

- The ARIMA(2,1,3) model [LT]
- The ARIMA(4,1,3)-GARCH(1,1) model [LT]
- The ETS$(A, N, N|EC)$ method [TN]
- The ETS$(A, N, N|EC)$-$(A, N, N|EC)$ method [TN]

where [LT] stands for logit transformation and [TN] stands for truncated normal distribution, so as to remind us how the densities are generated.

5.2.2 Benchmark models

To evaluate the forecast performances of the competing models, we compare the results with four benchmarks. The first two benchmarks are the persistence

(random walk) forecast and the constant (unconditional) forecast. The forecast distributions of these two benchmarks will be taken as truncated normal distributions in the form of (2.23). On the other hand, the third and fourth benchmarks are obtained by estimating empirical densities from the data. The third benchmark is the climatology forecast, and the fourth benchmark is an exponentially weighted moving average (EWMA) conditional probabilistic forecast. The details of their respective constructions are listed as follows:

Persistence forecast: We estimate the h-step ahead location parameter $\hat{\ell}_{t+h|t}$ and scale parameter $\hat{s}^2_{t+h|t}$ of the truncated normal distribution using the latest observations, that is,

$$\hat{\ell}_{t+h|t} = y_t, \qquad \hat{s}^2_{t+h|t} = \frac{\sum_{j=1}^{N_{r.v.}} (y_{t+1-j} - y_{t-j})^2}{N_{r.v.}} \tag{5.1}$$

for $t > N_{r.v.}$. We find that taking $N_{r.v.} = 48$, i.e., using data in the past 12 hours, gives an appropriate estimate for the realized variance $\hat{s}^2_{t+h|t}$. For $N_{r.v.}$ being too small, the estimated scale parameter will fluctuate greatly with time, while for $N_{r.v.}$ being too large, it would not be able to account for latest variations. Fixing $N_{r.v.} = 48$ may not be the optimal choice in all cases, but it is good enough to serve as a benchmark[1].

Constant forecast: We estimate the constant location parameter $\hat{\ell}_{t+h|t}$ and scale parameter $\hat{s}^2_{t+h|t}$ using data in the whole training set. They are given by the sample mean and the sample variance of the 11008 observations in the training set, so that

$$\hat{\ell}_{t+h|t} = \hat{\ell} = \frac{\sum_{j=1}^{11008} y_j}{11008}, \qquad \hat{s}^2_{t+h|t} = \hat{s}^2 = \frac{\sum_{j=1}^{11008} (y_j - \hat{\ell})^2}{11007} \tag{5.2}$$

Climatology forecast: An empirical unconditional density is fitted using data in the whole training set. The density has been shown in Figure 1.5 (b) previously. All h-step ahead forecasts would have the same unconditional density.

[1] In fact, we also consider optimizing $N_{r.v.}$ and obtain $N_{r.v.} = 12$. However, the in-sample mean CRPS is similar and in fact the out-of-sample forecasts are slightly worse than that of choosing $N_{r.v.} = 48$. For more details, please refer to Appendix E.4.

EWMA conditional probability forecast: To be in line with the use of exponential smoothing to estimate the location and scale parameters in Section 2.2, we consider an exponentially weighted moving average (EWMA) of a set of empirical conditional densities[1]. Due to computational efficiency as well as reliability of density estimations, at each time t we essentially consider the EWMA of 14 empirical conditional densities $g_{\text{emp}}(\{\Lambda_t^j\})$, where each of them is fitted using observations in the past j days with $j = 1, 2, \ldots, 14$ and $\{\Lambda_t^j\} = \{y_{t-96j+1}, y_{t-96j+2}, \ldots, y_t\}$ is the set of $(96 \times j)$ latest observations used to fit the empirical density. Up to an appropriate normalization constant, the h-step ahead EWMA empirical conditional probabilistic forecast is given by

$$f_{t+h|t}(y) \propto \sum_{j=1}^{14} (1-\lambda)\lambda^{j-1} g_{\text{emp}}(\{\Lambda_t^j\}) \tag{5.3}$$

so that for any fixed forecast origin t, the h-step ahead probability forecasts are identical for all $h > 1$. The smoothing parameter in (5.3) is estimated to be $\hat{\lambda} = 0.8012$, which is obtained by maximizing the log likelihood $\sum \log f_{t+1|t}(\lambda; y_{t+1})$. Figure 5.2 shows the exponential decrease of the weights being assigned to different empirical densities $g_{\text{emp}}(\{\Lambda_t^j\})$.

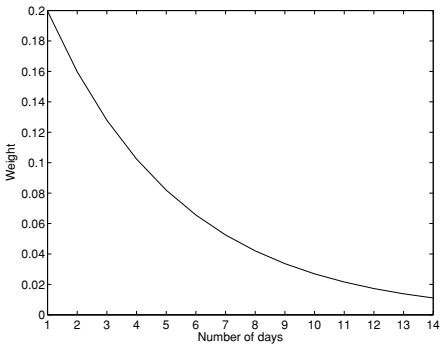


Figure 5.2: The exponential decrease of the weights $\lambda(1-\lambda)^{j-1}$ assigned to the empirical conditional densities $g_{\text{emp}}(\{\Lambda_t^j\})$ fitted with j days of latest observations, where $\hat{\lambda} = 0.8012$ is obtained by maximizing the likelihood using data in the training set. The EWMA empirical conditional probability forecasts are obtained as the weighted average of $g_{\text{emp}}(\{\Lambda_t^j\})$.

[1]We also consider a double kernel density benchmark but results show that the EWMA conditional probability forecast is good enough to serve as a decent benchmark. For more details on the double kernel density benchmark, please refer to Appendix E.4.

In summary, we consider four benchmarks and four competing models of generating multi-step probability forecasts, and compare their forecast performances from 15 minutes up to 24 hours ahead:

Model 1 : Persistence forecast [TN]

Model 2 : Constant forecast [TN]

Model 3 : Climatology forecast [Empirical density]

Model 4 : EWMA conditional probability forecast [Empirical density]

Model 5 : The ARIMA(2,1,3) model [LT]

Model 6 : The ARIMA(4,1,3)-GARCH(1,1) model [LT]

Model 7 : The ETS$(A, N, N|EC)$ method [TN]

Model 8 : The ETS$(A, N, N|EC)$-$(A, N, N|EC)$ method [TN]

5.2.3 Aggregated point forecasts

After explaining the four benchmarks and the four competing models in the previous sections, we generate 5504 out-of-sample h-step ahead forecasts for $1 \leq h \leq 96$, that is, from 15 minutes up to 24 hours ahead. For each forecast horizon h, we calculate the root mean squared error (RMSE) of the point forecasts, where the mean is taken over the 5504 h-step ahead forecasts in the testing set.

Figure 5.3 shows the results of point forecasts under RMSE. The two ARIMA-GARCH models outperform all other approaches for short forecast horizons within 12 hours, and are almost as good as the ETS$(A, N, N|EC)$-$(A, N, N|EC)$ method for horizons beyond 12 hours. Interestingly, the ARIMA(2,1,3) model is performing almost identically to the ARIMA(4,1,3)-GARCH(1,1) model. This phenomenon is in contrast with that for the ETS methods, where smoothing both the location and scale parameters do perform much better. It seems that including the dynamics of the conditional variance in the modeling of the logit transformed wind power z_t cannot improve the point forecasts under RMSE. These may be explained by Figure 1.6(b) which shows a significantly changing variance in the

original wind power y_t[1], and by Figure 2.1(a) which shows a fairly constant variance for z_t. We will further investigate into this issue in the evaluation of probability forecasts using the probability integral transform (PIT), where we see that the conditional variance models are indeed capturing the changes in volatility better and, thus, generate more reliable probability forecasts. A summary of the point forecast performances under RMSE is given in Table 5.2.

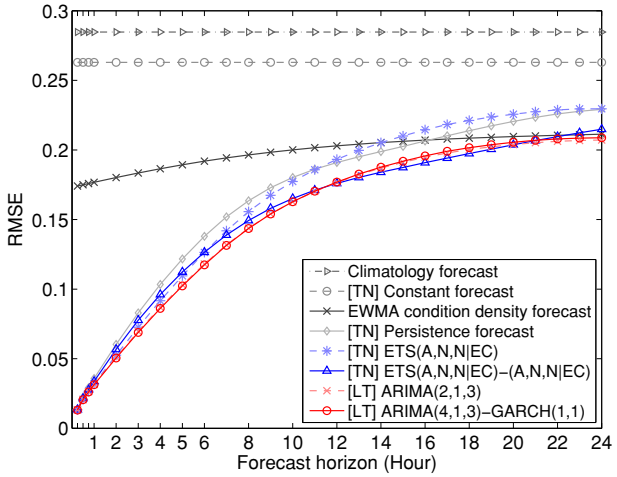


Figure 5.3: Root mean squared error (RMSE) of aggregated point forecasts generated by the eight models on Page 121 at forecast horizons from 15 minutes to 24 hours ahead.

	1 hour	6 hours	12 hours	24 hours
Persistence forecast	.036 (7.00)	.138 (5.50)	.191 (5.21)	.229 (2.78)
Constant forecast	.263 (21.02)	.263 (11.01)	.263 (5.57)	.263 (3.44)
Climatology forecast	.285 (21.21)	.285 (11.55)	.285 (6.37)	.285 (4.23)
EWMA conditional forecast	.177 (20.34)	.192 (9.12)	.203 (3.25)	.211 (1.11)
ARIMA(2,1,3)	.032 (2.98)	.118 (1.00)	.177 (0.29)	**.207**
ARIMA(4,1,3)-GARCH(1,1)	**.031**	**.117**	.177 (0.41)	.209 (0.56)
ETS$(A, N, N\|EC)$	.032 (4.81)	.126 (4.42)	.193 (4.73)	.230 (2.70)
ETS$(A, N, N\|EC)$-$(A, N, N\|EC)$	.033 (6.17)	.125 (3.28)	**.175**	.214 (0.68)

Table 5.2: Summary of aggregated point forecast performances of the four benchmarks and the four competing models, where the values are the RMSE of the 5504 out-of-sample forecasts at horizons of 1,6,12 and 24 hours. The bold values indicate the best model at that forecast horizon under RMSE. The bracketed values are the modified Diebold-Mariano statistics for the differences between that model with the best model with bolded values. Statistics with values greater than 1.96 correspond to significant differences at 95% confidence interval.

[1]This significantly changing variance may also due to the limited data employed.

5.2.4 Aggregated probability forecasts

For the evaluation of probability forecasts, we calculate the mean continuous ranked probability score (CRPS) of the probability forecasts as described in Appendix D, where the mean is taken over the 5504 h-step ahead forecasts in the testing set.

Figure 5.4 shows the performances of probability forecasts under mean CRPS. The rankings are similar to those under RMSE in point forecasts. The two ARIMA-GARCH models outperform all other models for all forecast horizons. Table 5.3 summarizes the main results. Again, the results of the ARIMA(2,1,3) model are very similar to those of the ARIMA(4,1,3)-GARCH(1,1) model. In contrast, the ETS$(A, N, N|EC)$-$(A, N, N|EC)$ method is significantly better than the ETS$(A, N, N|EC)$ method.

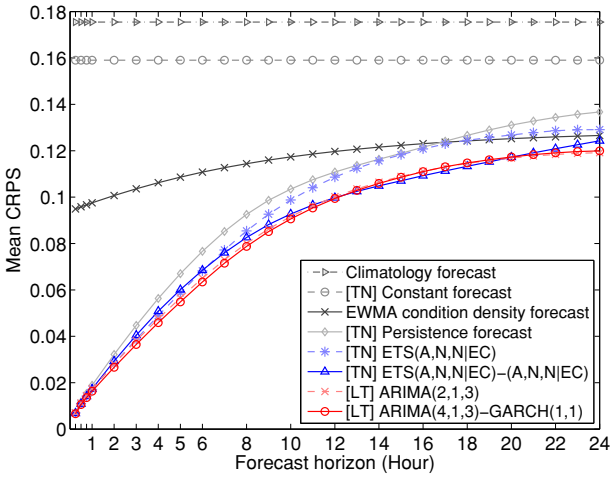


Figure 5.4: Mean CRPS of aggregated probability forecasts generated by the eight models on Page 121 at forecast horizons from 15 minutes to 24 hours ahead. Rankings of the performances are similar to those under RMSE in point forecasts.

To investigate the value of including the dynamics of conditional variances in the models, we consider the probability integral transform (PIT) as described in Appendix D. We evaluate the forecast calibrations of one-step ahead (i.e. 15-min) forecasts using the PIT diagrams, where the distribution of the PIT is uniform in the perfect case (Appendix D). For each model, we calculate the observed quantiles of the PIT values over all 5504 observations in the testing

	1 hour	6 hours	12 hours	24 hours
Persistence forecast	.019	.077	.111	.137
Constant forecast	.159	.159	.159	.159
Climatology forecast	.175	.175	.175	.175
EWMA conditional density	.098	.111	.120	.127
ARIMA(2,1,3)	.017	.065	.100	**.119**
ARIMA(4,1,3)-GARCH(1,1)	**.016**	**.063**	**.099**	.120
ETS($A, N, N\|EC$)	.017	.068	.109	.129
ETS($A, N, N\|EC$)-($A, N, N\|EC$)	.017	.069	.100	.124

Table 5.3: Summary of aggregated probability forecast performances of the four benchmarks and the four competing models, where the values are the mean CRPS of the 5504 out-of-sample forecasts at horizons of 1,6,12 and 24 hours. The bold values indicate the best model at that forecast horizon under mean CRPS.

data. In the ideal case, the quantile-quantile (QQ) plot should be a straight line along the diagonal. Figure 5.5 shows the QQ-plot for the PIT values of one-step ahead (i.e. 15-min) forecasts generated by the two ETS methods and the two ARIMA-GARCH models. We see that the ETS($A, N, N|EC$)-($A, N, N|EC$) method and the ARIMA(4,1,3)-GARCH(1,1) model indeed generate probability forecasts which are better calibrated. In particular, the overall calibration of the ETS($A, N, N|EC$)-($A, N, N|EC$) method is the best, indicating that it provides the most reliable descriptions of the changing volatility over time. Consistency bars of the QQ-plot can also be obtained by surrogate resampling that could account for serial correlations of the PIT values (Pinson *et al.*, 2010).

Figure 5.5 only reflects information on the marginal distributions of the PIT values. Stein (2009) suggests that it is also valuable to evaluate the distributions conditional on volatile periods. It is particularly important to capture the dynamics of the variance during times of large volatilities, since for most of the times one does not want to underestimate the risk by proposing an over-confident probabilistic forecast. Underestimating large risks usually leads to more disastrous outcome than overestimating small risks. Following Stein (2009), we compare the ability of the approaches in capturing volatility dynamics during the largest 10% of variance. To estimate the variance of the data in the testing set, we directly adopt the persistence forecast $\hat{s}^2_{\varepsilon;t+1|t}$ in (5.1), which essentially gives the 12-hour moving average of realized variance. Figure 5.6 shows the changing variance, where the largest values mostly occur in early December. The times

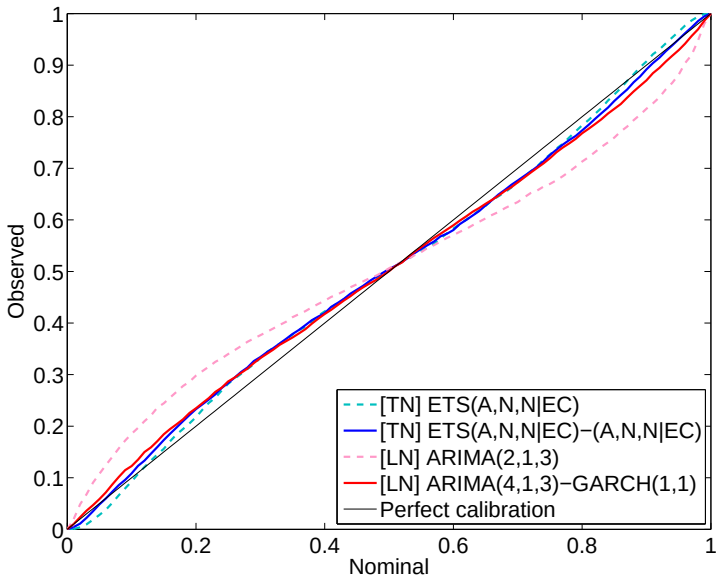


Figure 5.5: We show the QQ-plot for the PIT values of one-step ahead (i.e. 15-min) forecasts generated by the two ETS methods and the two ARIMA-GARCH models. For perfect calibration, the QQ-plot should be a straight line along the diagonal. We see that the ETS$(A, N, N|EC)$-$(A, N, N|EC)$ method and the ARIMA(4,1,3)-GARCH(1,1) model indeed generate probability forecasts which are better calibrated. In particular, the overall calibration of the ETS$(A, N, N|EC)$-$(A, N, N|EC)$ method is the best, indicating that it provides the most reliable descriptions of the changing volatility over time.

corresponding to the largest 10% of variance are selected and we compare the distribution of $z(y_{t+1})$ at those times. The PIT diagrams are shown in Figure 5.7. It demonstrates that the ARIMA-GARCH model indeed gives better calibrated one-step ahead probability forecasts than the ARIMA model during volatile periods. The differences between the two ETS methods are even more significant, where the ETS$(A, N, N|EC)$ method gives over-confident probability forecasts that underestimate the spread.

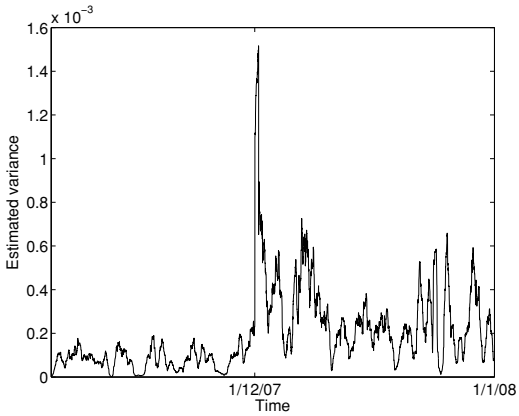


Figure 5.6: Estimated variance of data in the testing set using the persistence forecast $\hat{s}^2_{\varepsilon;t+1|t}$ in (5.1), which essentially gives the 12-hour moving average of realized variance. Clearly the variance changes with time and the largest values mostly occur in early December.

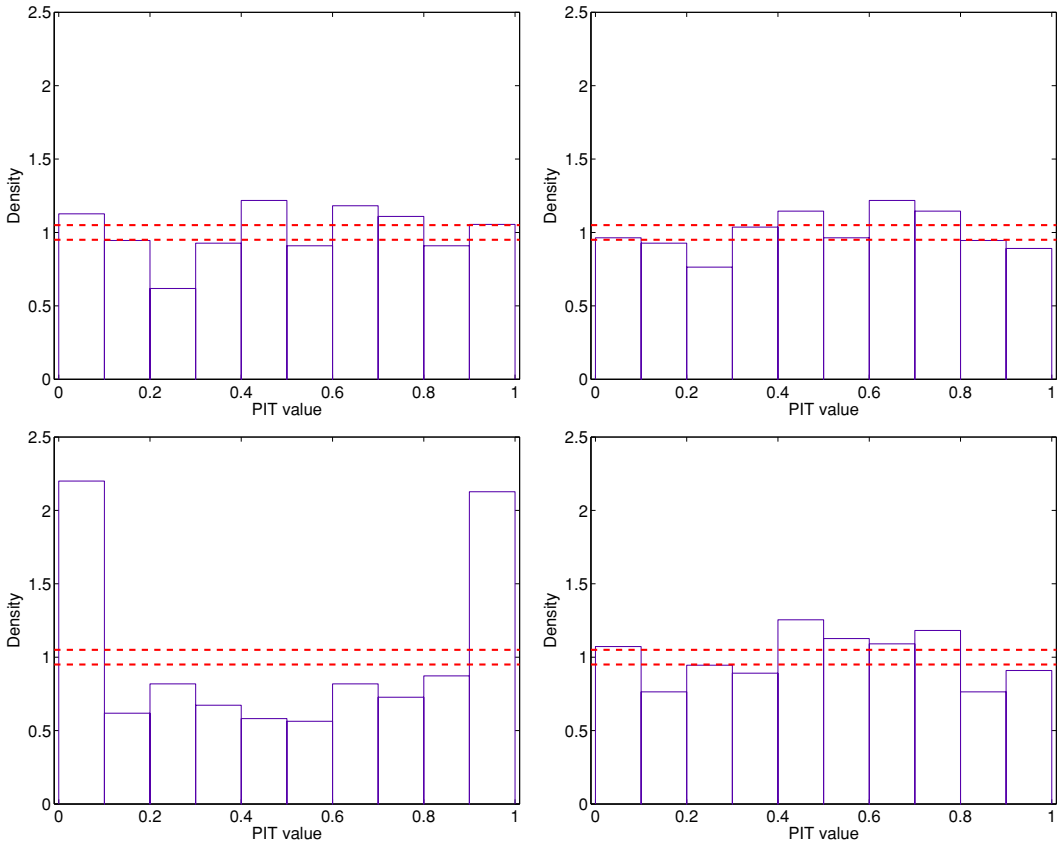


Figure 5.7: Histograms of PIT values conditioned on the largest 10% of estimated variance, where the one-step ahead probability forecasts are generated using the ARIMA(2,1,3) model (top left), the ARIMA(4,1,3)-GARHC(1,1) model (top right), the ETS$(A, N, N|EC)$ method (bottom left) and the ETS$(A, N, N|EC)$-$(A, N, N|EC)$ method (bottom right). The dotted lines correspond to the 95% consistency intervals.

5.3 Forecasts using spatiotemporal kriging

After discussing the results of aggregated forecasts, in this section, we apply the spatiotemporal kriging approach discussed in Chapter 3 to generate spatiotemporal forecasts for the Irish wind power. Using the methodology as detailed in Section 3.6, together with appropriate spatiotemporal correlation models as constructed in Section 3.5, we generate spatiotemporal wind power forecasts for each of the 64 wind farms in Ireland. We focus on short-term forecasts as this is an important horizon for optimal planning of power dispatch strategies. Thus, we consider multiple forecast horizons from 15 minutes (one-step ahead) up to 3 hours (12-steps ahead).

5.3.1 List of models

To clarify the various models that we consider in this section, we describe each of them in detail. First, we consider four benchmarks in Model 1 to Model 4. They are the ARIMA(4,1,3)-GARCH(1,1) model for aggregated wind power, the conditional density model, the ARIMA(1,1,1)-GARCH(1,1) model for individual wind power and the VAR(1) model respectively. Second, to perform spatiotemporal kriging, we consider six different correlation models in Model 5 to Model 10. These six models correspond to the homoscedastic version of spatiotemporal kriging, and in such case we assume the variances are stationary. Finally, we consider three heteroscedastic spatiotemporal kriging models in Model 11 to Model 13, where we model the changing variances with a GARCH(1,1) process. As in the homoscedastic models, the differences between the heteroscedastic models lie in the estimation of correlations. The details of the thirteen models are listed below:

Model 1 : We choose the first benchmark as the ARIMA(4,1,3)-GARCH(1,1) model for aggregated wind power, which has been denoted as Model 6 on page 121 when we analyze the aggregated wind power. We include this benchmark because it generates very competitive aggregated forecasts as shown in Figure 5.3, especially at short horizons below 3 hours. Note that this benchmark only serves for aggregated forecasts only.

Model 2 : We consider a conditional density forecast for individual wind power as another benchmark. The conditional densities are obtained by kernel smoothing with the built-in function 'ksdensity' in MATLAB, using the past N hours of data. By minimizing the in-sample mean square errors (MSE) across all wind farms, we find that $N = 5$ provides an optimal benchmark. The MSE of in-sample forecasts vary with N as shown in Figure 5.8.

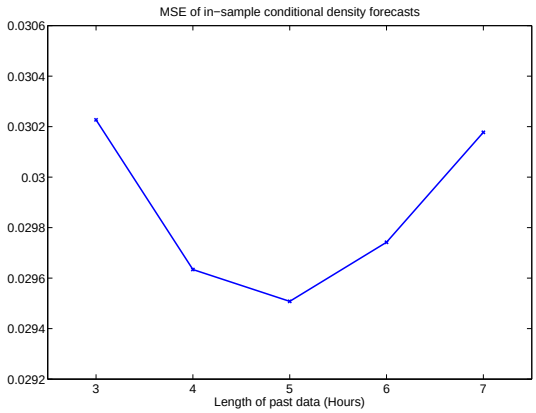


Figure 5.8: The figure shows the in-sample mean square error of the conditional density forecasts across all wind farms. It shows that the error is smallest when we use 5 hours of past data in estimating the conditional density.

Model 3 : We consider a univariate ARIMA(1,1,1)-GARCH(1,1) model for each individual wind power time series[1]. As in the benchmark Model 1, we firstly normalize the data using the modified logit transformation. To ensure the forecasts are within $[0, 1]$, we consider the mean and variance forecasts obtained from the ARIMA(1,1,1)-GARCH(1,1) model as the location and scale parameters for the corresponding truncated normal distribution, and draw 400 samples from the truncated normal distribution. We then inverse transform the samples and obtain both the density forecasts for individual and aggregated wind power.

[1]We have checked that using different number of lags in the ARIMA-GARCH model does not alter its performance as a benchmark significantly. In fact, the ARIMA(1,1,1)-GARCH(1,1) model generates more competitive out-of-sample forecasts. For more details, please refer to Appendix E.5.

Model 4 : We include a multivariate time series model as a benchmark, which is the VAR(1) model

$$\mathbf{Y}_t = \mathbf{c} + \mathbf{\Phi}\mathbf{Y}_{t-1} + \mathbf{e}$$

where $\mathbf{e}$ is the error with covariance matrix $\mathbf{\Sigma}$ and $\mathbf{Y}_t = (Y_{1,t}, \cdots, Y_{N,t})'$ is the wind power at the N wind farms at time t. We normalize the data using the modified logit transformation, and estimate the coefficient matrix $\mathbf{\Phi}$, the constant vector $\mathbf{c}$ and the covariance matrix $\mathbf{\Sigma}$ using the training data[1]. As in Model 3, we draw 400 samples from the truncated multivariate normal distribution, inverse transform the samples and obtain both the density forecasts for individual and aggregated wind power.

Model 5 : After defining four benchmarks as above, we now turn to the spatiotemporal kriging models. As described in Chapter 3, we firstly normalize the data using the modified logit transformation, and the spatiotemporal kriging predictor only depends on the covariance of the data. In this model, we consider the homoscedastic version and estimate the constant variances using the training data. We then calculate the empirical correlations and obtain the covariance matrix in the kriging predictor.

Model 6 : This is the spatiotemporal kriging model similar to Model 5, except that we consider fitting the empirical correlations with the correlation model in (3.31) proposed by Cressie & Huang (1999), i.e.

$$\begin{aligned} & & K(\mathbf{h}, u) &= K_{ST}(\mathbf{h}, u) + \delta_{\mathbf{h}=\mathbf{0}} K_0(u) \\ \text{where} & & K_{ST}(\mathbf{h}, u) &= \frac{c_0(a|u|+1)}{[(a|u|+1)^2 + b^2\|\mathbf{h}\|^2]^{3/2}}, \\ & & K_0(u) &= (1-c_0)(a|u|+1)^{-2} \end{aligned}$$

with parameters a, b and c_0[2].

[1] We estimate the coefficients using the 'vectorar' function in the UCSD GARCH Toolbox written by Kevin Sheppard, which could be freely downloaded from http://www.kevinsheppard.com/wiki/UCSD_GARCH.

[2] For the ranges of the parameters in different models, please refer to Section 3.5

Model 7 : This is the spatiotemporal kriging model similar to Model 5, except that we consider fitting the empirical correlations with the correlation model in (3.32) proposed by Gneiting (2002b) with chosen parameters $\tau = 1$ and $\gamma = 1/2$, i.e.

$$\begin{aligned} K(\mathbf{h}, u) &= K_{ST}(\mathbf{h}, u) + \delta_{\mathbf{h}=\mathbf{0}} K_0(u) \\ \text{where} \qquad K_{ST}(\mathbf{h}, u) &= \frac{c_0}{(a|u|^{2\alpha}+1)} \exp\left(\frac{-b\|\mathbf{h}\|}{(a|u|^{2\alpha}+1)^{\beta/2}}\right), \\ K_0(u) &= \frac{(1-c_0)}{(a|u|^{2\alpha}+1)} \end{aligned}$$

with parameters a, b, c_0, α and β.

Model 8 : This is the spatiotemporal kriging model similar to Model 5, except that we consider fitting the empirical correlations with the correlation model in (3.34) proposed by Stein (2005), i.e.

$$\begin{aligned} K(\mathbf{h}, u) &= K_{ST}(\mathbf{h}, u) + \delta_{\mathbf{h}=\mathbf{0}} K_0(u) \\ \text{where} \qquad K_{ST}(\mathbf{h}, u) &= \frac{\mathcal{M}_{\nu+\xi|u|^\gamma}(\alpha\|\mathbf{h}-\mathbf{v}u\|)}{2^{\nu+\xi|u|^\gamma}\Gamma(\nu+\xi|u|^\gamma+1)}, \\ K_0(u) &= \frac{\kappa}{\nu'+|u|^\gamma} \end{aligned}$$

and $\mathbf{v} = (v\sin\theta_0, v\cos\theta_0)$, with parameters $\nu, \nu', \xi, \gamma, \alpha, v$ and θ_0.

Model 9 : This is the spatiotemporal kriging model similar to Model 5, except that we consider fitting the empirical correlations with the Lagrangian correlation model described in (3.22), where we choose a powered exponential function for the purely spatial correlation function K_S and include an appropriate nugget effect, i.e.

$$\begin{aligned} K(\mathbf{h}, u) &= K_{ST}(\mathbf{h}, u) + \delta_{\mathbf{h}=\mathbf{0}} K_0(u) \\ \text{where} \qquad K_{ST}(\mathbf{h}, u) &= (1-c_0)\exp(-(\alpha\|\mathbf{h}-\mathbf{v}u\|)^{2\gamma}), \\ K_0(u) &= c_0\exp(-(\beta u)^{2\eta}) \end{aligned}$$

and $\mathbf{v} = (v\sin\theta_0, v\cos\theta_0)$, with parameters $\alpha, \beta, c_0, \gamma, \eta, v$ and θ_0.

Model 10 : This is the spatiotemporal kriging model similar to Model 5, except that we consider fitting the empirical correlations with our correlation model described in (3.42), i.e.

$$\begin{aligned} & & K(\mathbf{h},u) &= K_{ST}(\mathbf{h},u)+\delta_{\mathbf{h}=\mathbf{0}}K_0(u) \\ \text{where} & & K_{ST}(\mathbf{h},u) &= c_0\exp\left(-(\alpha\|\mathbf{h}\|)^{2\gamma}\right)\exp\left(-(\beta\tilde{u})^{2\eta}\right), \\ & & K_0(u) &= \exp\left(-(\tilde{\beta}u)^{2\tilde{\eta}}\right)-c_0\exp\left(-(\beta u)^{2\eta}\right) \\ & & \tilde{u} &= u+\frac{\mathbf{h}\cdot\mathbf{v}}{\|\mathbf{v}\|^2}=u+\frac{\|\mathbf{h}\|\cos(\theta-\theta_0)}{v} \end{aligned}$$

with parameters $\alpha, \beta, \tilde{\beta}, c_0, \gamma, \eta, \tilde{\eta}, v$ and θ_0.

Model 11 : This is the heteroscedastic version of Model 5, where we model the changing variances by a GARCH(1,1) process. With the estimated empirical correlations, we obtain the non-stationary covariance matrices for the kriging predictor.

Model 12 : This is the heteroscedastic version of Model 9, where we model the changing variances by a GARCH(1,1) process. With the fitted correlations using the Lagrangian model, we obtain the non-stationary covariance matrices for the kriging predictor.

Model 13 : This is the heteroscedastic version of Model 10, where we model the changing variances by a GARCH(1,1) process. With the fitted correlations using our correlation model, we obtain the non-stationary covariance matrices for the kriging predictor.

5.3.2 Parameter estimation

5.3.2.1 Modified logit transformation

Before moving on to analyze the forecast results obtained in the above models, we show some results of the estimated parameters, especially for the correlation models in Model 6 to Model 10. The first parameter that we need to estimate is the constant δ in the modified logit transformation. Applying the estimation procedure described in Section 1.3.1 on the training set of the Irish wind data,

we obtain $\delta_{\min} = 0.0312$. Figure 5.9 shows the result of the median Kolmogorov-Smirnov test statistic for different values of δ in (1.2). As a result, we transform the wind power generation y into $z = \log[(y + 0.0312)/(1 - y + 0.0312)]$. The continuous distribution of z is then approximately Gaussian at each wind farm. Figure 5.10 shows an example of a density distribution $f(z)$ of the transformed wind power generation using (1.1) with $\delta = \delta_{\min} = 0.0312$. Although the continuous part is approximately Gaussian, a probability mass at the lower bound still exists.

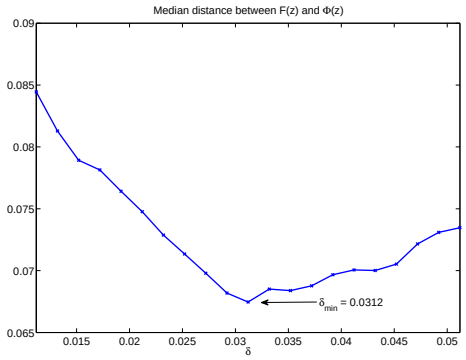


Figure 5.9: The figure shows the median distance between the distribution of the empirical wind power data at each wind farm j, $F^j(z)$, and the corresponding normal distribution $\Phi(z|\mu_j, \sigma_j)$, where z is the transformed wind power using (1.1), $\mu_j = \mathrm{E}[z_j]$ is the mean and $\sigma_j^2 = \mathrm{Var}[z_j]$ is the variance. The distance is calculated as $\max(|F(z) - \Phi(z))|$, which is essentially the Kolmogorov-Smirnov test statistic in (1.2). Using only the training data set, We obtain $\delta_{\min} = 0.0312$.

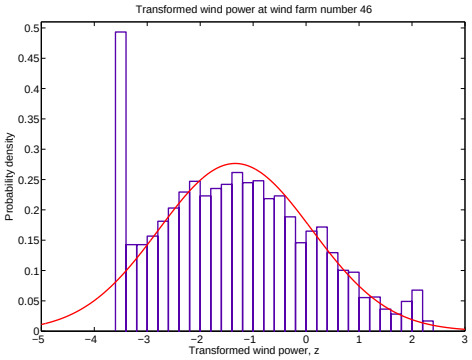


Figure 5.10: Density distribution $f(z)$ of the modified logit transformation of wind power generation at wind farm number 46 using (1.1), with $\delta = \delta_{\min} = 0.0312$. The solid line is the fitted Gaussian distribution $N(\mu, \sigma^2)$ with mean $\mu = \mathrm{E}[z]$ and variance $\sigma^2 = \mathrm{Var}[z]$. Only data in the training set is used.

5.3.2.2 Parameters in correlation models

Next, we fit the correlation models and estimate the parameters as described in Section 3.5.3. A summary of the fitted correlation models is given in Table 5.4. We give the goodness of fit in terms of mean square error (MSE), mean absolute error (MAE), Akaike's information criterion (AIC) and Bayesian information criterion (BIC). Note that all errors are weighted according to (3.44). As discussed in Raftery (1996) and Burnham (2004), only differences in AIC and BIC with their corresponding minimums matter, as they are free of scaling constants and can be interpreted as the likelihood ratio between two models. As a result, we also give values of AIC − AIC(Min) and BIC − BIC(Min). Furthermore, we show the normalized AIC and BIC values by dividing by the total number of observations (i.e. $(64 \times 64 \times 32) - 64$ in this case) in the fitting of the models. The parameter v is the estimated speed of the weather front, and θ_0 is the direction of movement in degrees as measured in a clockwise direction from the north. We also check if the models are able to describe the shape of decay of cross correlations correctly. We find that Cressie's and Gneiting's models, both with fewer parameters and contain the Cauchy class as a building block, give a fit for cross correlations which is too sharp at the peak and is regarded as too rough, as shown in Figure 5.11. Another interesting observation is that in the estimation of the exponent β in Gneiting's model, i.e. Model 7, we obtain $\beta \approx 1$ is many cases. As $\beta \in [0, 1]$ governs the extent of space-time interaction with $\beta = 0$ corresponding to a separable model (Gneiting, 2002b), it demonstrates that spatiotemporal wind power generation is highly non-separable.

Model	Cressie	Gneiting	Stein	Lagrangian	Our Model
Number of parameters	3	5	7	7	9
MSE	0.040	0.036	0.032	0.035	0.029
MAE	0.157	0.148	0.134	0.142	0.127
AIC	-853710	-873937	-918353	-889899	-944128
AIC - AIC(Min)	90418	70191	25775	54229	0
Normalized AIC	-6.5165	-6.6709	-7.0099	-6.7927	-7.2066
BIC	-853676	-873894	-918279	-889825	-944033
BIC - BIC(Min)	90357	70139	25754	54208	0
Normalized BIC	-6.5162	-6.6705	-7.0093	-6.7921	-7.2059
v (km/hr)	-	-	17.0	28.6	36.9
$\theta_0(^\circ)$	-	-	102.3	100.6	107.2
Fitted cross correlations	Rough	Rough	Smooth	Smooth	Smooth

Table 5.4: This table shows the summary of the fit of the various correlation models. We give the goodness of fit in terms of MSE, MAE, AIC and BIC. Large values of AIC − AIC(Min) and BIC − BIC(Min) means that our model, which gives the lowest AIC and BIC values, is significantly better than the other alternatives. v is the estimated speed of the weather front and θ_0 is the direction of movement in degrees from the north. We find that Cressie's and Gneiting's models give a fit for cross correlations which is too sharp at the peak and is regarded as too rough, as shown in Figure 5.11.

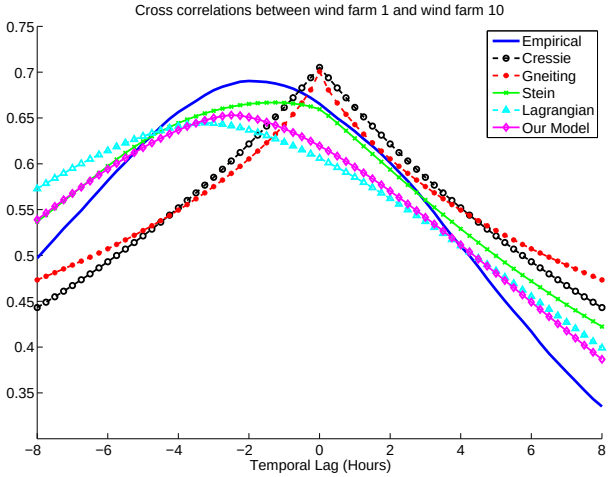


Figure 5.11: This figure shows the fit for the cross correlations between wind power generated at wind farm number 1 and 10. This is a typical fit, where Cressie's and Gneiting's models give a fit for cross correlations which is too sharp at the peak.

5.3.3 Individual forecast evaluations

Now, equipped with the calibrated models, we proceed to generate h-step ahead out-of-sample forecasts for each of the 64 wind farms in Ireland. In this section, we focus on individual forecasts, i.e. forecasts for wind power at individual wind farms. We analyze the forecast results from two aspects as follows:

Box-plots showing results from all wind farms at a fixed horizon: In the first aspect, we analyze the individual forecasts at all 64 wind farms and show the results in terms of box-plots. In the box plots, the central red line is the median value, and the edges of the box are the 25% and 75% quantiles respectively. The sizes of the notches indicate the uncertainties of the median. For instance, two medians are significantly different at 95% confidence interval if the notches do not overlap. A dotted horizontal line is also provided to clarify the location of minimum score among the models.

Compare results of two selected wind farms at all horizons: In the second aspect, we study the individual forecasts at two of the selected single wind farms so as to investigate the forecast performances with respect to the locations of the wind farms. One of the wind farms is located on the north-west coast, and the other is located on the south-east coast. An interesting result is that our correlation model is able to capture the useful information in the movement of the weather front, and forecasts better for wind farms located on the south-east coast.

5.3.3.1 Forecasts at all 64 wind farms

First, let us look into the forecasts for wind power generation at all 64 wind farms in Ireland. As there are a total of 64 evaluation scores for all wind farms, we use box plots to show the distribution of scores across different models. We focus on 12-step ahead forecasts (i.e. 3 hours) where the differences between various models are most significant. In the following analysis, the forecast horizon will be 3 hours unless otherwise specified. Figure 5.12 shows the RMSE of the point forecasts generated by the models. As expected, the performances of point forecasts are not affected by the inclusion of heteroscedastic effects. For instance, Model 10 and Model 13 perform similarly. The best point forecasts under RMSE

are generated by the spatiotemporal kriging models with Lagrangian correlation or our correlation model, where the difference between the two models are small. The conditional density benchmark is outperformed significantly.

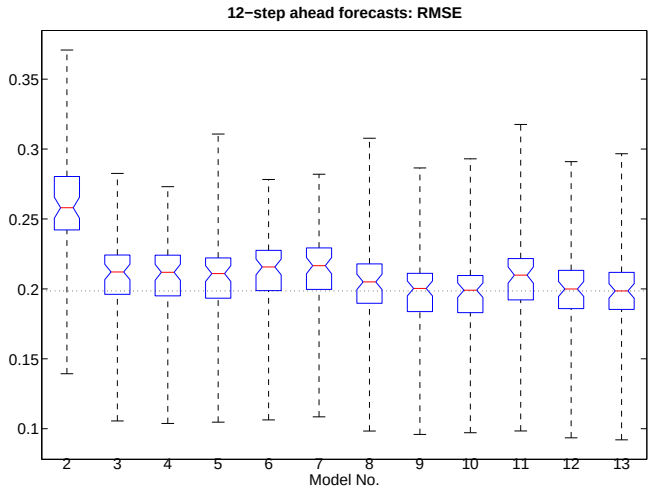


Figure 5.12: The figure shows the box-plot of RMSE for 12-step ahead (i.e. 3 hours) individual point forecasts generated by each model. The model numbers are labeled according to the list described in Section 5.3.1. As expected, the performances of point forecasts are not affected by the inclusion of heteroscedastic effects. For instance, Model 10 and Model 13 perform similarly. The best point forecasts under RMSE are generated by the spatiotemporal kriging models with Lagrangian correlation or our correlation model, where the difference between these models are small. The conditional density benchmark is outperformed significantly. *(2: Conditional density, 3: ARIMA(1,1,1)-GARCH(1,1), 4: VAR(1), 5: Empirical, 6: Cressie, 7: Gneiting, 8: Stein, 9: Lagrangian, 10: Our model, 11: Empirical-H, 12: Lagrangian-H, 13: Our model-H)*

Next, we consider the performances of quantile forecasts under MQE as described in Appendix D. Quantile forecasts are point forecasts at a particular quantile of the forecast density. We consider quantile values from 1% to 99%. To show the results clearly, for each model, we calculate the mean of the MQE across all 64 wind farms. We then plot this mean MQE versus different quantile values. Results are shown in Figure 5.13. It is clear that the conditional density benchmark is performing very poorly. An interesting observation is that the individual ARIMA(1,1,1)-GARCH(1,1) benchmark generates the best quantile forecasts at extreme quantile values of 1% and 99%. This could due to the robustness of the individual ARIMA-GARCH models. The heteroscedastic spatiotemporal kriging models with Lagrangian correlation (Model 12) or our correlation model (Model 13) give superior quantile forecasts at quantile values around 50% as shown in the zoom-in plot in Figure 5.14, although they are slightly outperformed by the ARIMA(1,1,1)-GARCH(1,1) benchmark at extreme quantiles.

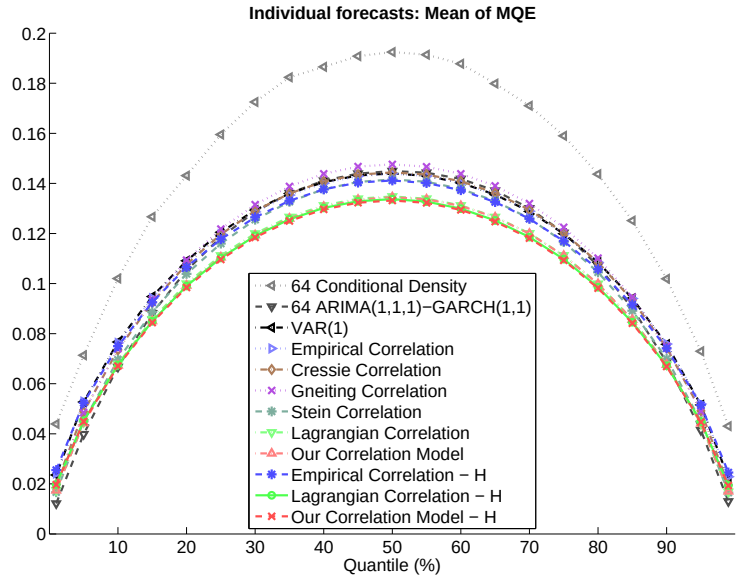


Figure 5.13: The figure shows the mean of the MQE across all 64 wind farms versus different quantile values. It is clear that the conditional density benchmark is performing very poorly. An interesting observation is that the individual ARIMA(1,1,1)-GARCH(1,1) benchmark generates the best quantile forecasts at extreme quantile values of 1% and 99%. This could due to the robustness of the individual ARIMA-GARCH models.

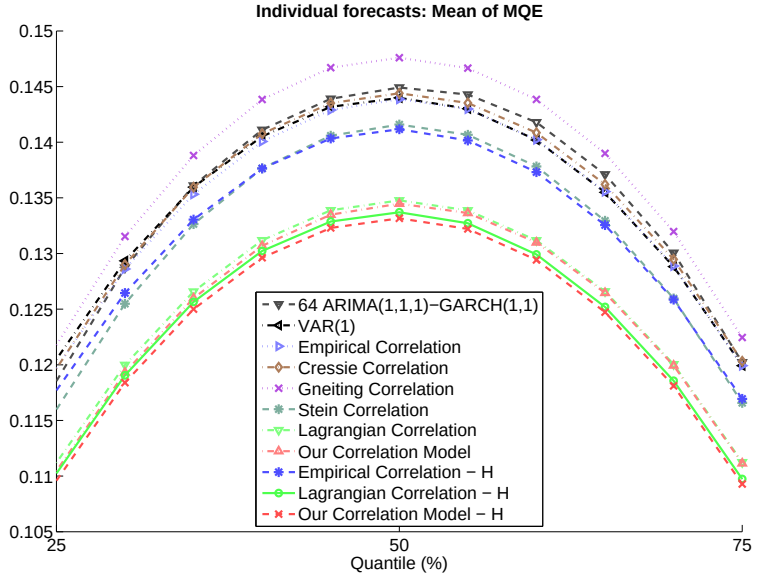


Figure 5.14: The figure shows the zoom-in plot of Figure 5.13 at quantile values between $25\% - 75\%$. The heteroscedastic spatiotemporal kriging models with Lagrangian correlation (Model 12) or our correlation model (Model 13) give superior quantile forecasts at quantile values around 50%. The conditional density benchmark is not shown here due to its poor performance.

Finally, we evaluate the probability forecasts produced by the models. We use the mean CRPS as the score for evaluation of probability forecasts, as described in Appendix D. We have also evaluate the calibration of the probability forecasts using the PIT histograms and QQ-plots, as for the aggregated probability forecasts in Section 5.2.4. However, due to limited space and vast number of wind farms in the individual forecasts, we do not include evaluations of calibrations here. Nevertheless, we will include the evaluations of calibrations for aggregated forecasts generated by spatiotemporal kriging models in Section 5.3.4, so as to provide an idea of their performances. Note that if the forecasts allow for probability masses at zeros and ones, as in the individual forecasts generated by the two-stage models, we should consider randomized PIT values (Czado *et al.*, 2009). However, this problem does not occur for aggregated forecasts as we obtain the probability density by simulations.

Figure 5.15 shows the boxplots of the mean CRPS for each model. The results are similar to those for point forecasts under RMSE. One may observe that the heteroscedastic spatiotemporal kriging models, i.e. Model 11 to Model 13, seem to perform similarly as their homoscedastic counterparts and are unable to improve probability forecasts. However, we will see in the following analysis that the improvements highly depend on the wind power variability of the particular wind farm. In other words, including the modeling of heteroscedastic volatility may or may not improve probability forecasts.

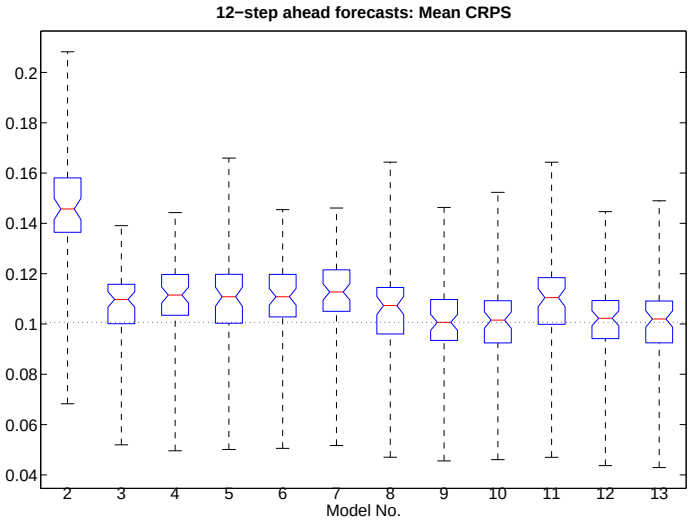


Figure 5.15: The figure shows the box-plot of mean CRPS for 12-step ahead (i.e. 3 hours) individual probability forecasts generated by each model. We see that the heteroscedastic spatiotemporal kriging models, i.e. Model 11 to Model 13, perform better in terms of probability forecasts when compared to their homoscedastic counterparts, i.e. Model 5, Model 9 and Model 10. However, the heteroscedastic models almost perform the same as their homoscedastic counterparts in terms of point forecasts. This clearly demonstrates the fact that modeling the changing variances help to improve density forecasts but not necessarily point forecasts. *(2: Conditional density, 3: ARIMA(1,1,1)-GARCH(1,1), 4: VAR(1), 5: Empirical, 6: Cressie, 7: Gneiting, 8: Stein, 9: Lagrangian, 10: Our model, 11: Empirical-H, 12: Lagrangian-H, 13: Our model-H)*

5.3.3.2 Forecasts at a single wind farm

Apart from evaluating the forecast performances at all 64 wind farms as a whole, we also investigated the forecast performances at a particular wind farm. In such case, we consider the forecasts at all 12 horizons. Among the 64 wind farms, we choose two of them for further analysis. The first one, Anarget, is a wind farm located on the north-west coast. The second one, Kilbranish, is situated on the south-east coast. It turns out that forecast performances of the models differ significantly at different wind farms. This could be explained by the movement of the weather front in Ireland, which in general propagates from the west towards the east. Thus it carries useful information that could be utilized to improve forecasts at wind farms situated on the south-east coasts.

Wind farm located in north-west: Anarget

Anarget is wind farm number 30 as in Table 5.1, which is located on the north-west coast. We also choose to study it in more detail because the wind power at Anarget has a relatively large variance, which is more challenging in terms of forecasts. Its wind power time series is shown in Figure 5.16.

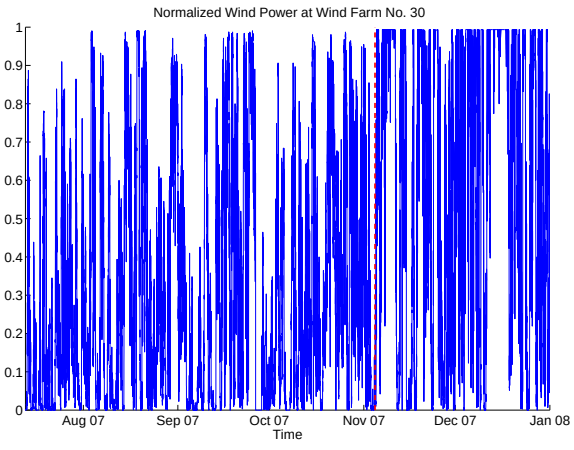


Figure 5.16: The figure shows the normalized wind power time series at wind farm number 30, Anarget, which is located at the north-west coast of Ireland. The red dotted line indicates the division of the training data and testing data used in our forecasts.

Figure 5.17 shows the results of point forecasts under RMSE. Note that as the conditional density benchmark (Model 2) is significantly out-performed by all other models, we will not show its results in Figure 5.17 as well as some of the following figures. The best model is the spatiotemporal kriging model with Lagrangian correlations. It slightly out-performs that of our correlation model. The heteroscedastic spatiotemporal kriging models in general give better forecasts than the homoscedastic counterparts at longer forecast horizons. This phenomenon is most obviously seen in the spatiotemporal kriging model with empirical correlations.

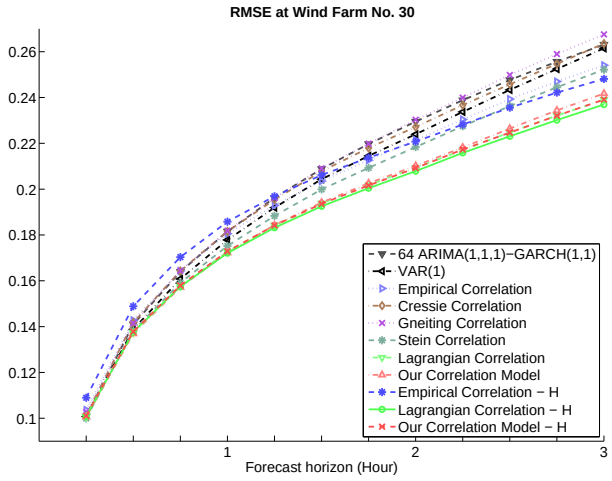


Figure 5.17: This figure plots the RMSE of the point forecasts for wind farm number 30 (Anarget) at a horizon from 15 minutes up to 3 hours ahead. The best model is the spatiotemporal kriging model with Lagrangian correlations. It slightly out-performs that of our correlation model. The heteroscedastic spatiotemporal kriging models in general give better forecasts than the homoscedastic counterparts at longer forecast horizons. This phenomenon is most obviously seen in the spatiotemporal kriging model with empirical correlations.

For the evaluation of quantile forecasts, we consider a single horizon at 3 hours ahead. Figure 5.18 shows the MQE at different quantile values. The rankings of the models are quite similar to those in Figure 5.13, with the heteroscedastic spatiotemporal kriging models giving the best quantile forecasts for most quantile values. Again, the ARIMA(1,1,1)-GARCH(1,1) benchmark performs the best at extreme quantiles.

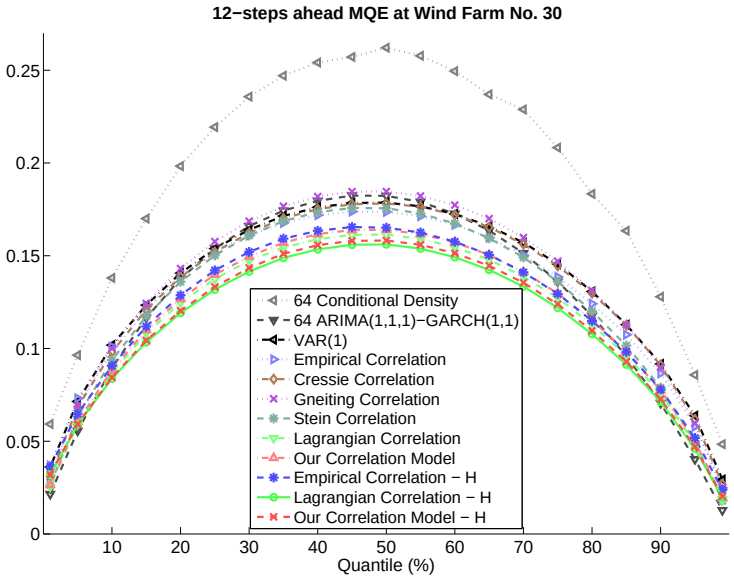


Figure 5.18: This figure plots the MQE of the quantile forecasts for wind farm number 30 (Anarget) at 12-step ahead forecast horizon, i.e. 3 hours ahead. The MQE are calculated at different quantile values. The rankings of the models are quite similar to those in Figure 5.13, with the heteroscedastic spatiotemporal kriging models giving the best quantile forecasts for most quantile values. Again, the ARIMA(1,1,1)-GARCH(1,1) benchmark performs the best at extreme quantiles.

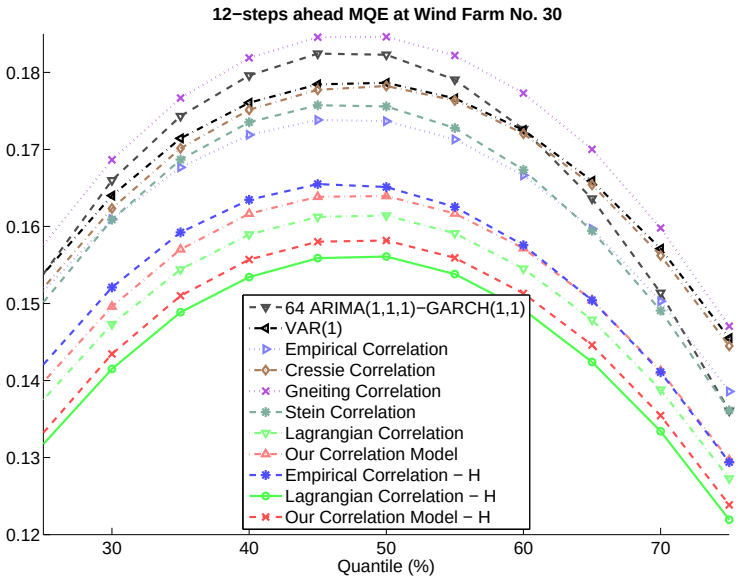


Figure 5.19: This figure shows the zoom-in plot of Figure 5.18 at quantile values between $25\% - 75\%$.

Finally, for probability forecasts, we show the advantages of using a heteroscedastic model over a homoscedastic one. Figure 5.20 plots the results of the mean CRPS. We observe a significant decrease in the mean CRPS for probability forecasts made by the heteroscedastic spatiotemporal kriging models, especially for longer forecast horizons. The Lagrangian model still slightly out-performs our correlation model, both in the homoscedastic and the heteroscedastic versions.

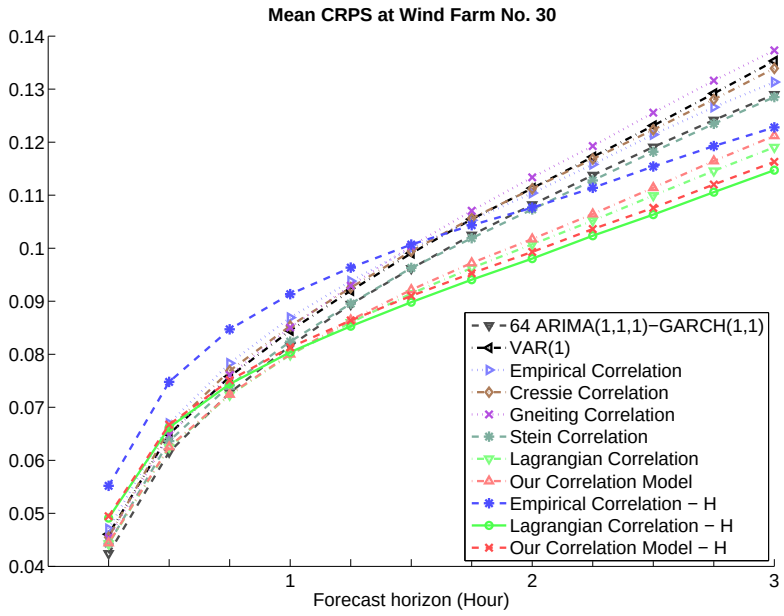


Figure 5.20: This figure plots the mean CRPS of the probability forecasts for wind farm number 30 (Anarget) at a horizon from 15 minutes up to 3 hours ahead. We observe a significant decrease in the mean CRPS for probability forecasts made by the heteroscedastic spatiotemporal kriging models, especially for longer forecast horizons. The Lagrangian model still slightly out-performs our correlation model, both in the homoscedastic and the heteroscedastic versions.

Wind farm located in south-east: Kilbranish

After looking into a wind farm that situates on the north-west coast of Ireland, we consider another one which locates on the south-east coast for comparison. We choose to study wind farm number 25, i.e. Kilbranish, and its location is shown in Figure 3.5. The time series of normalized wind power generation at Kilbranish is given in Figure 5.21.

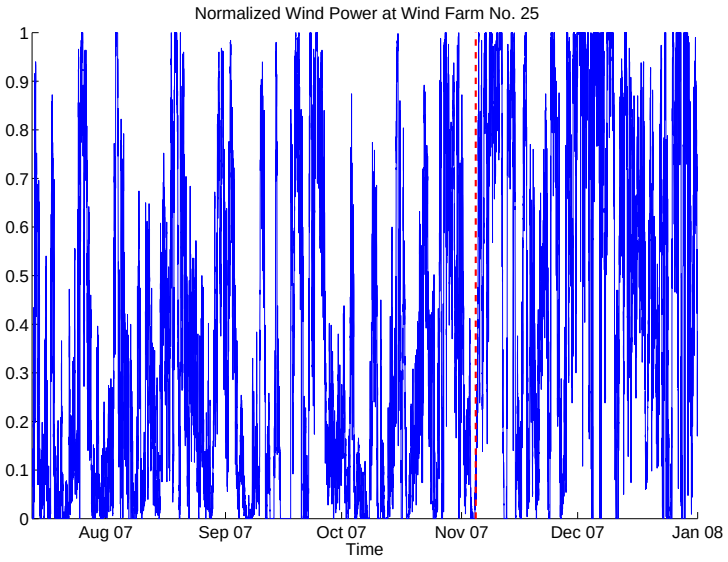


Figure 5.21: The figure shows the normalized wind power time series at wind farm number 25, Kilbranish, which is located at the south-east coast of Ireland. The red dotted line indicates the division of the training data and testing data used in our forecasts.

First, we show the RMSE of the point forecasts. The main differences of the results between Kilbranish and Anarget is that for Kilbranish, the spatiotemporal kriging approach with our correlation model gives much better results compared with the approach using Lagrangian correlations. The improvements using our correlation models become more significant for longer forecast horizons. Interestingly, the same spatiotemporal kriging approach with the empirical correlations perform poorly, especially for short forecast horizons. In this case, modeling the correlations give an important contribution in the improvement of point forecasts.

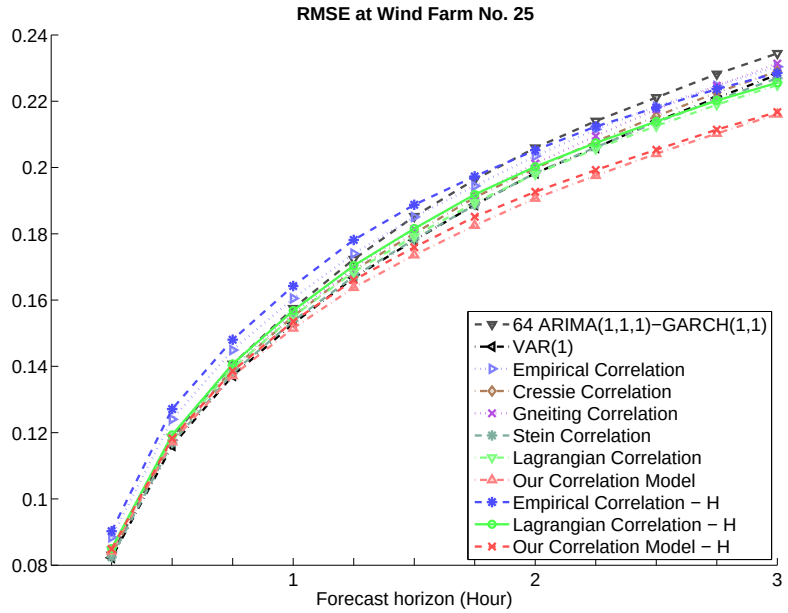


Figure 5.22: This figure shows the RMSE of the point forecasts for wind farm number 25 (Kilbranish) at a horizon from 15 minutes up to 3 hours ahead. The main differences of the results between Kilbranish and Anarget is that for Kilbranish, the spatiotemporal kriging approach with our correlation model give much better results compared with the approach using Lagrangian correlations. The improvements using our correlation models become more significant for longer forecast horizons. Interestingly, the same spatiotemporal kriging approach with the empirical correlations perform poorly, especially for short forecast horizons. In this case, modeling the correlations is important.

For the performances of quantile forecasts under MQE, we consider a single forecast horizon at 12-steps ahead (i.e. 3 hours) and show the results in Figure 5.23. There are two main differences with those obtained for Anarget in Figure 5.18. First, our correlation models perform much better than the Lagrangian model, as seen from the much lower MQE around 50% quantile values. This is shown more clearly in the zoom-in plot in Figure 5.24. Second, the conditional density benchmark, although still out-performed by all models at all quantile values, is performing relatively well compared with the results in Anarget. This could be explained by the lower variance (11%) of wind power as observed at Kilbranish as compared to that of Anarget (13%), which could result in better forecasts using conditional densities.

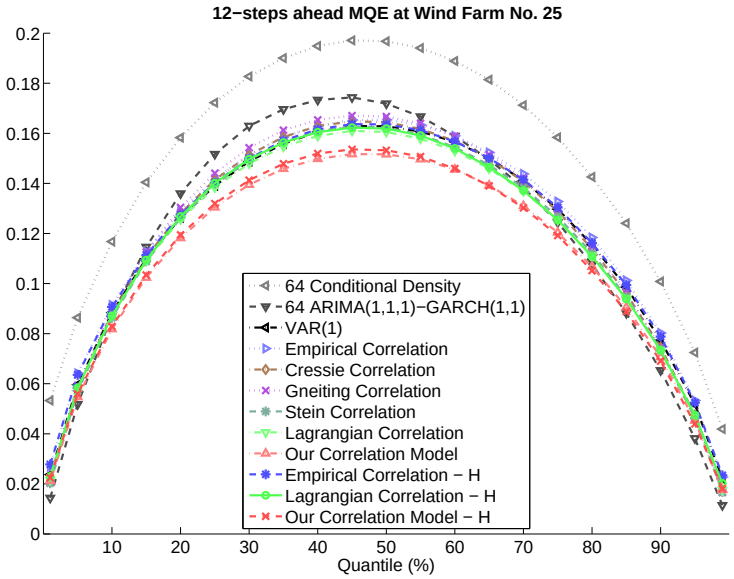


Figure 5.23: This figure shows the MQE of the quantile forecasts for wind farm number 25 (Kilbranish) at forecast horizon of 3 hours ahead. There are two main differences with those obtained for Anarget in Figure 5.18. First, our correlation models perform much better than the Lagrangian model, as seen from the much lower MQE around 50% quantile values. Second, the conditional density benchmark, although still out-performed by all models at all quantile values, is performing relatively well compared with the results in Anarget. This could be explained by the lower variance (11%) of wind power as observed at Kilbranish, which results in better forecasts using conditional densities.

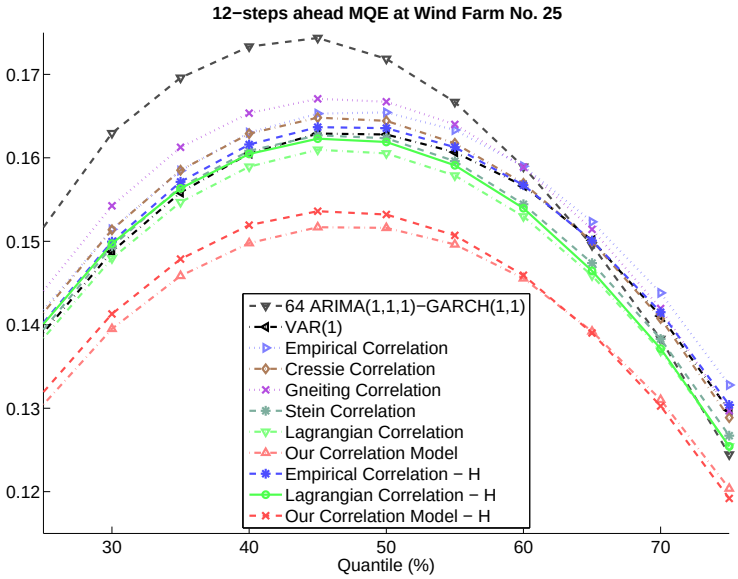


Figure 5.24: This figure shows the zoom-in plot of Figure 5.23 at quantile values between $25\% - 75\%$.

Finally, we investigate the performances of probability forecasts at Kilbranish. Results under mean CRPS are shown in Figure 5.25. It is very encouraging to see that probability forecasts using the spatiotemporal kriging approach with our correlation model perform very well. This is mainly due to the location of Kilbranish, which is at the south-east coast and further analysis on this aspect will be provided in the following section. However, in contrast to the case at Anarget, the heteroscedastic versions of the spatiotemporal kriging models do not improve the probability forecasts here.

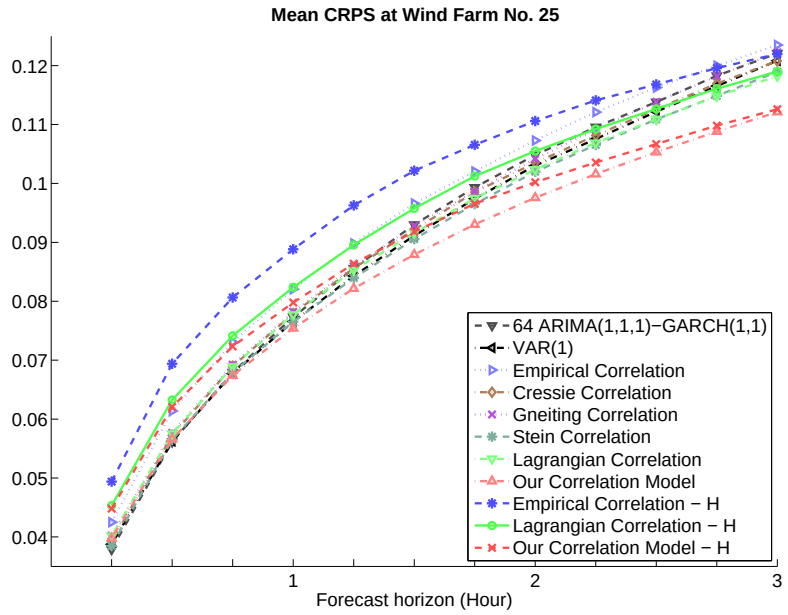


Figure 5.25: This figure shows the mean CRPS of the probability forecasts for wind farm number 25 (Kilbranish) at a horizon from 15 minutes up to 3 hours ahead. It is very encouraging to see that probability forecasts using the spatiotemporal kriging approach with our correlation model perform very well. This is mainly due to the location of Kilbranish, which is at the south-east coast and further analysis on this aspect will be provided in the following section. However, as contrast to the case at Anarget, the heteroscedastic versions of the spatiotemporal kriging models cannot improve probability forecasts here.

5.3.3.3 Comparison of forecast performances at different locations

From the above analysis of forecast performances at a single wind farm, we see that the locations of the wind farms may play an important role in the results obtained by various models. This is physically reasonable, as one expects the movement of the weather front in Ireland, which in general propagates towards the east, helps us to generate better forecasts at wind farms situated on the eastern part of Ireland. Clearly, historical observations of wind power at other sites, particularly those situated on the western part, could give very useful information on the future wind power generated at wind farms on the east coast as they are highly correlated. The spatiotemporal correlations essentially reflect this information, and thus it could be beneficial to consider forecasts from a spatiotemporal model in such cases. On the other hand, for wind farms that are located in the north or north-western part of Ireland, one may find that the spatiotemporal correlations are less useful in predicting the wind power generation. In these cases, a univariate approach may give competitive forecasts.

Note that the Lagrangian model produces better forecasts at Anarget, but our model produces better forecasts at Kilbranish. In general, we have an interesting result that our model gives better predictions for wind power generated at wind farms located in the south-east. As the weather front in Ireland propagates roughly from west to east, we expect that successful utilization of this propagation of information could greatly help us to predict wind power generated at wind farms in the east. We verify this effect and demonstrate how our correlation model, which better captures the correlation structure of the wind power generation, indeed produces superior forecasts in such case. Figure 5.26 shows the percentage differences between RMSE of forecasts generated by the Lagrangian model (Model 12) and those by our model (Model 13), where the forecasts are generated at horizon $h = 12$, i.e. 3 hours ahead. We standardize the percentage differences, reflecting the standard deviation of the percentage difference at a particular location with respect to all 64 wind farms. In the figure, the blue (red) dots indicate that compared with other wind farms, our correlation model performs relatively well (poor) at that location. The size and the deepness of colour of the dots are proportional to the magnitude of percentage differences. Clearly,

our model performs much better at wind farms located on the south-east coast of Ireland. The same analysis with mean CRPS generate very similar results and will not be shown here.

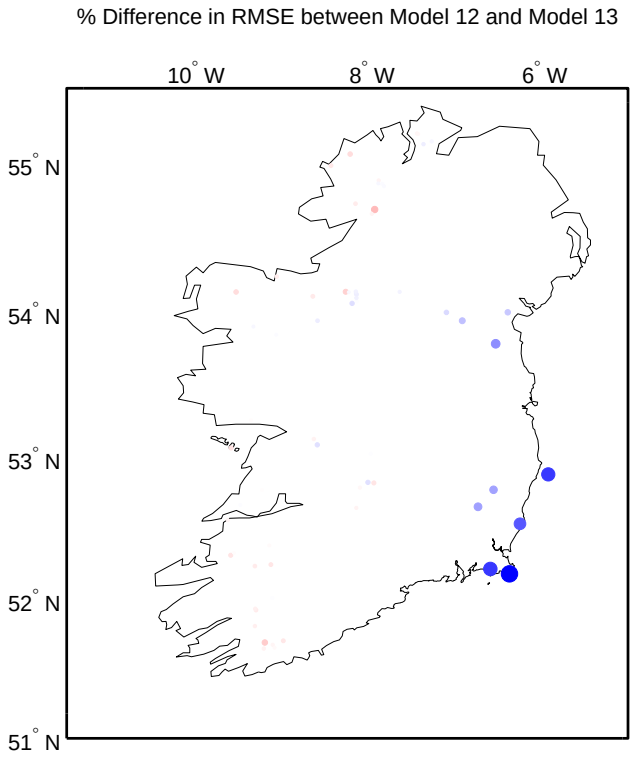


Figure 5.26: The figure shows the percentage differences of RMSE of point forecasts generated by the Lagrangian model (Model 12) and those by our model (Model 13), where the forecasts are generated at horizon $h = 12$, i.e. 3 hours ahead. We standardize the percentage differences, reflecting the standard deviation of the percentage difference at a particular location with respect to all 64 wind farms. The blue (red) dots indicate that compared with other wind farms, our correlation model performs relatively well (poor) at that location. The size and the deepness of colour of the dots are proportional to the magnitude of percentage differences. Clearly, our model performs much better at wind farms located on the south-east coast of Ireland. Our best forecasts are about 3% better than those obtained by the Lagrangian model.

5.3.4 Aggregated forecast evaluations

After evaluating the performances of individual wind power forecasts at a single wind farm, we turn to study the forecast performances for aggregated forecasts. Aggregated forecasts are obtained simply by adding the individual wind power forecasts at all 64 wind farms. Again, we normalize the aggregated wind power by dividing it by the total capacity of the 64 wind farms, which is 792.355 MW. Similar to the case for individual forecasts, we consider three aspects of performances, namely the point forecasts, the quantile forecasts and the probability forecasts. We evaluate the performances of aggregated forecasts at all horizons from 15 minutes to 3 hours ahead.

Results of point forecasts under RMSE are plotted in Figure 5.27. Again, as the conditional density benchmark is significantly out-performed by all other models, we do not include its results in the figure. Our correlation model produces the best point forecasts, no matter whether it is applied using the homoscedastic or heteroscedastic version of the spatiotemporal kriging model. The next best results are produced by the Lagrangian correlation model, and then the univariate ARIMA(4,1,3)-GARCH(1,1) benchmark. Note that this benchmark is applied directly to model the aggregated time series, and it generates very good point forecasts which are even better than the spatiotemporal kriging model using empirical correlations.

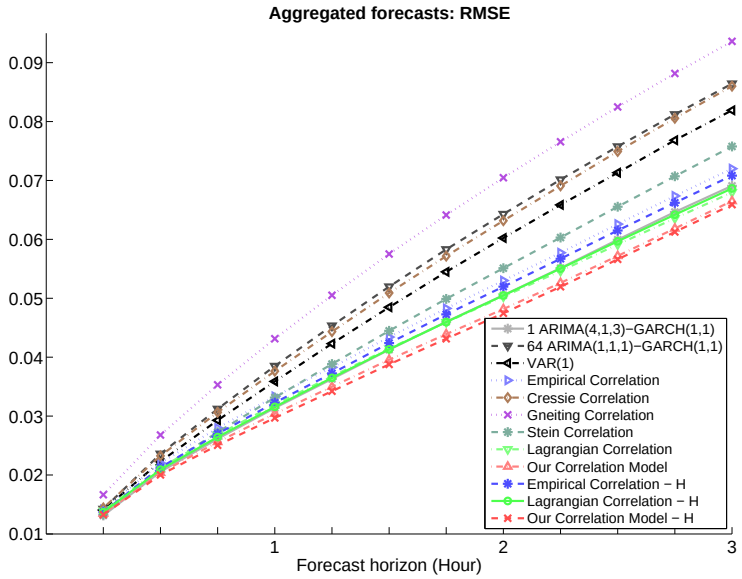


Figure 5.27: The figure shows the RMSE of aggregated point forecasts. Our correlation model produces the best point forecasts, no matter it is applied on the homoscedastic or heteroscedastic version of the spatiotemporal kriging model. Results are followed by the Lagrangian correlation model, and then the univariate ARIMA(4,1,3)-GARCH(1,1) benchmark. Note that this benchmark is applied directly to model the aggregated time series, and it generates very good point forecasts which are even better than the spatiotemporal kriging model using empirical correlations.

Next, we show the results of quantile forecasts for aggregated wind power in Figure 5.28. We plot the MQE of the 12-step ahead aggregated quantile forecasts at different quantiles. Again, we see that the univariate ARIMA(4,1,3)-GARCH(1,1) model is superior in generating well-calibrated quantile forecasts, especially at extreme quantile values. This is not surprising, as aggregated time series are smoother and easier to model. It also involves much fewer parameters, and so avoids the problem of over-fitting the noise in individual wind power time series. Our correlation model generates the best quantile forecasts around quantile values of 50%, as shown clearly on the zoom-in plot in Figure 5.29. In particular, the heteroscedastic version is slightly better than the homoscedastic counterpart, as in our correlation model, the Lagrangian correlation model and the empirical correlations.

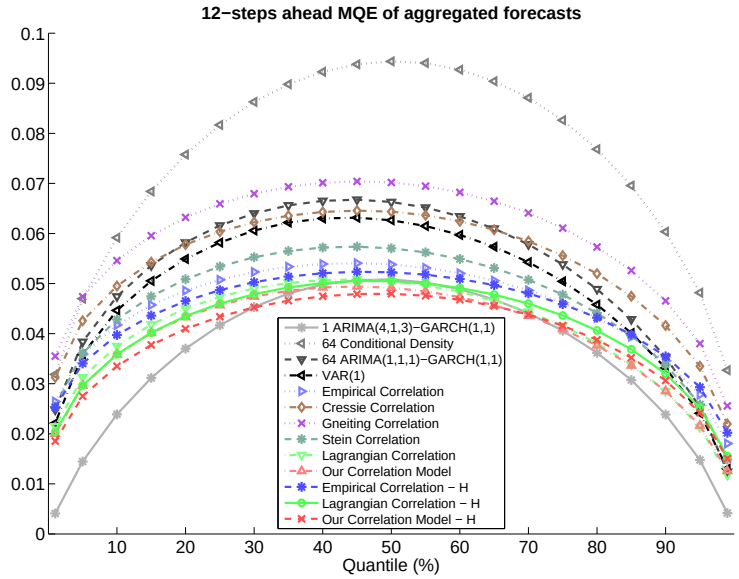


Figure 5.28: The figure shows the plot of MQE of aggregated quantile forecasts at different quantile values. Again, we see that the univariate ARIMA(4,1,3)-GARCH(1,1) model is superior in generating well-calibrated quantile forecasts, especially at extreme quantiles.

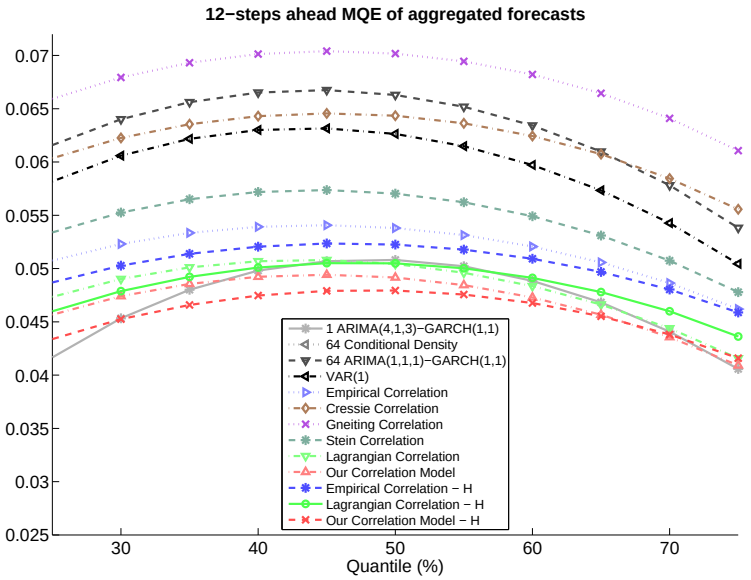


Figure 5.29: The figure shows the zoom-in plot of MQE of Figure 5.28 at quantile values between $25\% - 75\%$. Our correlation model generates the best quantile forecasts around quantile values of 50%. In particular, the heteroscedastic version is slightly better than the homoscedastic counterpart, as in our correlation model, the Lagrangian correlation model and the empirical correlations.

Finally, we look at the probability forecasts generated by the models. We use the mean CRPS to evaluate the forecasts, and results are shown in Figure 5.30. The results and the rankings of the models are very similar to those obtained for point forecasts in Figure 5.27. As mentioned earlier, we also show the evaluations of calibrations of forecast densities here, so as to give an idea of the goodness of forecast calibrations obtained by spatiotemporal models. Similar to the analysis in Section 5.2.4, we calculate the PIT values using the forecast densities[1] and the actual observations (Appendix D). For perfect calibration, the QQ-plot of the PIT values should lie on the straight line along the diagonal. Results are shown in Figure 5.31. We observe that the spatiotemporal kriging models in general generate relatively well-calibrated forecast densities, compared with their corresponding heteroscedastic counterparts. It should be emphasized that our correlation model produces better-calibrated forecast densities than the empirical correlations as well as the simple Cressie's correlation model. However, the heteroscedastic spatiotemporal kriging models produce over-confident forecast densities that are not disperse enough, which result in the shape that resembles the sigmoid function. Lastly, although forecasts using the two-stage model will only be discussed in the next section, we include the evaluations of the two-stage model with our correlation model in the figure for comparison. Interestingly, with the ability to model the probability masses at the boundaries of the forecast densities, the two-stage model performs significantly better than all other models at lower quantiles. However, its performance is not as good for larger quantiles. Similar to the heteroscedastic spatiotemporal kriging models, the two-stage model tends to produce forecast densities that are not disperse enough.

[1]For spatiotemporal forecasts, the forecast densities are obtained by simulations.

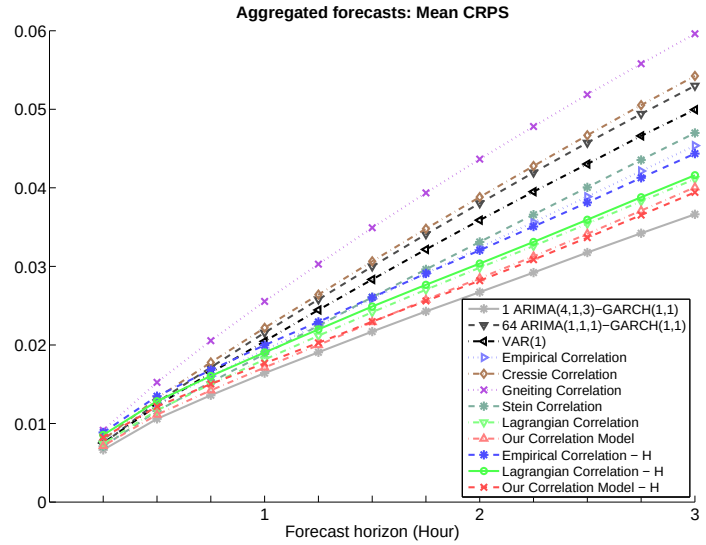


Figure 5.30: The figure shows the mean CRPS of aggregated probability forecasts. Our correlation model generates the best density forecasts, but the performance is only slightly better than the Lagrangian model.

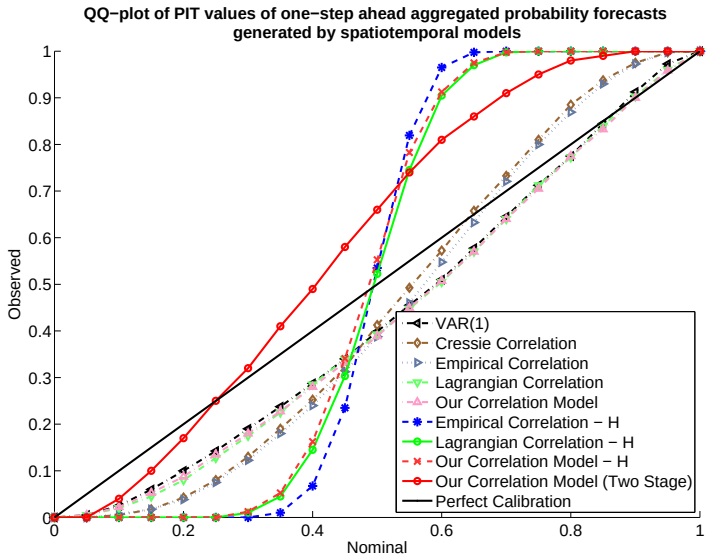


Figure 5.31: The figure shows the QQ-plot of PIT values for one-step ahead aggregated probability forecasts generated by various spatiotemporal models. We observe that the spatiotemporal kriging models in general generate relatively well-calibrated forecast densities, compared with their corresponding heteroscedastic counterparts. It should be noted that our correlation model produces better-calibrated forecast densities than the empirical correlations as well as the simple Cressie's correlation model. However, the heteroscedastic spatiotemporal kriging models produce over-confident forecast densities that are not disperse enough, which result in the shape that resembles the sigmoid function. We also include the evaluations of the two-stage model with our correlation model for comparison. Interestingly, with the ability to model the probability masses at the boundaries of the forecast densities, the two-stage model performs significantly better than all other models at lower quantiles. However, its performance is not as good for larger quantiles. Similar to the heteroscedastic spatiotemporal kriging models, the two-stage model tends to produce forecast densities that are not disperse enough.

5.3.4.1 Significance of forecast improvements

As the differences in forecast performances are quite small for different models, one may want to know if the differences are statistically significant or not. A possible way is to consider the Diebold-Mariano Test (Diebold & Mariano, 1995). For instance, suppose two models, $M1$ and $M2$, generate N h-step ahead forecasts $\{\hat{y}_t^{M1}\}, \{\hat{y}_t^{M2}\}$ respectively, where $t = 1, 2, \ldots$. The time series of forecast errors are then given by $\{e_t^{M1}\}, \{e_t^{M2}\}$, where $e_t = y_t - \hat{y}_t$. For a specific forecast evaluation metric $g(e_t)$, we would like to know if $g(e_t^{M1})$ is significantly different from $g(e_t^{M2})$. In other words, we would like to test if

$$d_t = g(e_t^{M1}) - g(e_t^{M2}) \tag{5.4}$$

is expected to be zero. The Diebold-Mariano test statistic is given by

$$DM = \frac{\bar{d}}{\sqrt{\hat{V}(\bar{d})}} \tag{5.5}$$

where $\bar{d} = \sum_{t=1}^{N} d_t/N$ is the mean of d_t, and $\hat{V}(\bar{d}$ is the estimated variance of $\bar{d}$, given approximately by

$$\hat{V}(\bar{d}) \approx N^{-1}\left(\gamma_0 + 2\sum_{k=1}^{h-1}\gamma_k\right) \tag{5.6}$$

where γ_k is the autocovariance of $\bar{d}$ at lag k. We include the autocovariances up to lag $h-1$ in (5.6) because it is well known that optimal h-step ahead forecast errors follow an MA$(h-1)$ process (see Section A.3.3). Under the null hypothesis that d_t has mean zero, the statistic in (5.5) converges to the standard normal distribution.

However, the Diebold-Mariano test using (5.5) is known to be seriously oversized, especially when the forecast horizon $h > 1$. For this reason, Harvey *et al.* (1997) propose a modified test statistic that applies an approximately unbiased

estimation for $\hat{V}(\bar{d})$, the variance of $\bar{d}$. The modified statistic is given by

$$\tilde{DM} = \left(\frac{N + 1 - 2h + N^{-1}h(h-1)}{N} \right)^{-1/2} DM \tag{5.7}$$

where N is the sample size, h is the forecast horizon and DM is the original statistic given by (5.5). As we consider multi-step ahead forecasts where $h = 1 - 12$, we apply the modified Diebold-Mariano test statistic $\tilde{DM}$ to test for the difference between the RMSE of aggregated forecasts generated by different models. In particular, we consider

$$d_t = (e_t^M)^2 - (e_t)^2 \tag{5.8}$$

where e_t^M are the forecast errors generated by Model 1 to Model 12, and e_t are those generated by heteroscedastic spatiotemporal kriging with our correlation model, i.e. Model 13. The test statistics are calculated between different models as well as for different horizons h. Our model generates the lowest RMSE than all other models (except Model 10) for horizons $h = 2 - 12$, but it is beaten by the univariate ARIMA(4,1,3)-GARCH(1,1) benchmark (i.e. Model 1) for $h = 1$. Results of test statistics are shown in Table 5.5. For horizon $h = 1$, Model 1 gives the lowest RMSE. Note that heteroscedastic spatiotemporal kriging with our correlation model (Model 13) could not beat its corresponding homoscedastic version (Model 10) significantly for horizons beyond $h = 5$.

Model No.	1	2	3	4	5	6	7	8	9	10	11	12
Horizon												
1	-1.56	35.56	8.65	8.11	6.52	12.01	20.81	0.00	9.32	3.04	3.83	9.03
2	3.19	20.38	10.99	9.70	8.12	10.94	14.19	5.69	7.19	2.61	6.45	6.57
3	4.19	15.82	10.65	9.95	7.91	10.24	12.05	7.47	6.76	2.68	6.62	6.03
4	4.43	13.45	10.12	9.66	7.36	9.48	10.74	7.91	6.32	2.26	6.58	5.65
5	4.30	11.95	9.60	9.22	6.65	8.87	9.83	7.84	5.91	1.86	6.03	5.36
6	4.23	10.91	9.12	8.83	6.46	8.41	9.17	7.81	5.41	1.50	5.85	5.12
7	4.05	10.16	8.78	8.50	6.20	8.08	8.70	7.75	5.12	1.24	5.62	5.05
8	3.75	9.58	8.47	8.21	5.88	7.81	8.36	7.57	4.71	1.01	5.36	5.00
9	3.46	9.12	8.17	7.97	5.49	7.57	8.06	7.34	4.26	0.76	5.13	4.94
10	3.17	8.75	7.88	7.75	5.14	7.38	7.81	7.17	3.72	0.63	4.86	4.84
11	2.97	8.45	7.62	7.52	4.77	7.21	7.60	6.97	3.24	0.54	4.56	4.71
12	2.64	8.20	7.40	7.30	4.36	7.05	7.41	6.83	2.71	0.54	4.21	4.51

Table 5.5: The table shows the modified Diebold-Mariano test statistic in (5.7), where we test for the differences of RMSE generated by Model 1 to Model 12 as compared with those generated by heteroscedastic spatiotemporal kriging with our correlation model, i.e. Model 13. Under the null hypothesis that the difference is zero, the distribution of test statistic converges to the standard normal distribution. Test statistics are calculated for different models as well as for different horizons h. Heteroscedastic spatiotemporal kriging with our correlation model (Model 13) generates significantly smaller RMSE than all other models (except Model 10) for horizons $h = 2 - 12$, but it is beaten by the aggregated ARIMA(4,1,3)-GARCH(1,1) benchmark (Model 1) for $h = 1$. For horizon $h = 1$, Model 1 gives the lowest RMSE. Note that Model 13 could not beat Model 10 significantly for horizons beyond $h = 5$. *(1: ARIMA(4,1,3)-GARCH(1,1), 2: Conditional density, 3: ARIMA(1,1,1)-GARCH(1,1), 4: VAR(1), 5: Empirical, 6: Cressie, 7: Gneiting, 8: Stein, 9: Lagrangian, 10: Our model, 11: Empirical-H, 12: Lagrangian-H)*

5.4 Forecasts using two-stage model

In the last section, we have focused on the forecasts of wind power using the spatiotemporal kriging approach only. In this section, we turn to the two-stage models as described in Chapter 4. The two-stage models have an advantage over the spatiotemporal approaches because they are capable to model the probability masses in the wind power distribution. We now move on to analyze the forecast results using the two-stage models, and we will compare the results with those obtained from the spatiotemporal kriging approaches. For the sake of clarity, we will list out the various models that we consider in the forecasts in this section. Before we generate the forecasts using the two-stage models, we will also describe some results regarding parameter estimation.

5.4.1 List of models

The list of models that we consider in this section will be of a similar structure as those described in Section 5.3.1, except that we do not consider the conditional density benchmark, Model 2 in Section 5.3.1, as its performance is poor. Apart from the two-stage models, we include the heteroscedastic spatiotemporal kriging approach with our correlation model, i.e. Model 13 in Section 5.3.1, so as to compare the performances between the spatiotemporal kriging approach and the two-stage models.

To clarify the models, we include a list of description below. Model 1 to Model 3 are benchmarks. Model 4 to Model 9 are the two-stage models, where the differences among them lie only on the structure of the spatiotemporal correlation models for the latent Gaussian process $Z(\mathbf{s}, t)$. For each two-stage model, we draw 100 samples in the simulation. Also note that for short forecast horizons of $h \leq 4$, we do not model any spatiotemporal correlations of $Z(\mathbf{s}, t)$ and will apply the empirical correlations directly, as this is found to give better results. Finally, Model 10 is the heteroscedastic spatiotemporal kriging approach with our correlation model.

Model 1 : The ARIMA(4,1,3)-GARCH(1,1) benchmark for aggregated wind power, i.e. Model 1 in Section 5.3.1.

Model 2 : The univariate ARIMA(1,1,1)-GARCH(1,1) benchmark for each individual wind power time series, i.e. Model 3 in Section 5.3.1.

Model 3 : The VAR(1) model as a multivariate benchmark, i.e. Model 4 in Section 5.3.1.

Model 4 : Two-stage model with empirical correlations of $Z(\mathbf{s}, t)$ calculated using the training data.

Model 5 : Two-stage model where the empirical correlations of $Z(\mathbf{s}, t)$ are fitted with the correlation model in (3.31) proposed by Cressie & Huang (1999),

i.e.

$$\begin{aligned} & & K(\mathbf{h},u) &= K_{ST}(\mathbf{h},u)+\delta_{\mathbf{h}=\mathbf{0}}K_0(u) \\ \text{where} & & K_{ST}(\mathbf{h},u) &= \frac{c_0(a|u|+1)}{[(a|u|+1)^2+b^2\|\mathbf{h}\|^2]^{3/2}}, \\ & & K_0(u) &= (1-c_0)(a|u|+1)^{-2} \end{aligned}$$

with parameters a, b and c_0.

Model 6 : Two-stage model where the empirical correlations of $Z(\mathbf{s},t)$ are fitted with the correlation model in (3.32) proposed by Gneiting (2002b) with chosen parameters $\tau = 1$ and $\gamma = 1/2$, i.e.

$$\begin{aligned} & & K(\mathbf{h},u) &= K_{ST}(\mathbf{h},u)+\delta_{\mathbf{h}=\mathbf{0}}K_0(u) \\ \text{where} & & K_{ST}(\mathbf{h},u) &= \frac{c_0}{(a|u|^{2\alpha}+1)}\exp\left(\frac{-b\|\mathbf{h}\|}{(a|u|^{2\alpha}+1)^{\beta/2}}\right), \\ & & K_0(u) &= \frac{(1-c_0)}{(a|u|^{2\alpha}+1)} \end{aligned}$$

with parameters a, b, c_0, α and β.

Model 7 : Two-stage model where the empirical correlations of $Z(\mathbf{s},t)$ are fitted with the correlation model in (3.34) proposed by Stein (2005), i.e.

$$\begin{aligned} & & K(\mathbf{h},u) &= K_{ST}(\mathbf{h},u)+\delta_{\mathbf{h}=\mathbf{0}}K_0(u) \\ \text{where} & & K_{ST}(\mathbf{h},u) &= \frac{\mathcal{M}_{\nu+\xi|u|^\gamma}(\alpha\|\mathbf{h}-\mathbf{v}u\|)}{2^{\nu+\xi|u|^\gamma}\Gamma(\nu+\xi|u|^\gamma+1)}, \\ & & K_0(u) &= \frac{\kappa}{\nu'+|u|^\gamma} \end{aligned}$$

and $\mathbf{v} = (v\sin\theta_0, v\cos\theta_0)$, with parameters $\nu, \nu', \xi, \gamma, \alpha, v$ and θ_0.

Model 8 : Two-stage model where the empirical correlations of $Z(\mathbf{s},t)$ are fitted with the Lagrangian correlation model in (3.22), where we choose a powered exponential function for the purely spatial correlation function K_S and

include an appropriate nugget effect, i.e.

$$\begin{aligned} & & K(\mathbf{h},u) &= K_{ST}(\mathbf{h},u) + \delta_{\mathbf{h}=\mathbf{0}} K_0(u) \\ \text{where} & & K_{ST}(\mathbf{h},u) &= (1-c_0)\exp(-(\alpha\|\mathbf{h}-\mathbf{v}u\|)^{2\gamma}), \\ & & K_0(u) &= c_0 \exp(-(\beta u)^{2\eta}) \end{aligned}$$

and $\mathbf{v} = (v\sin\theta_0, v\cos\theta_0)$, with parameters $\alpha, \beta, c_0, \gamma, \eta, v$ and θ_0.

Model 9 : Two-stage model where the empirical correlations of $Z(\mathbf{s},t)$ are fitted with our correlation model in (3.42), i.e.

$$\begin{aligned} & & K(\mathbf{h},u) &= K_{ST}(\mathbf{h},u) + \delta_{\mathbf{h}=\mathbf{0}} K_0(u) \\ \text{where} & & K_{ST}(\mathbf{h},u) &= c_0 \exp\left(-(\alpha\|\mathbf{h}\|)^{2\gamma}\right) \exp\left(-(\beta\tilde{u})^{2\eta}\right), \\ & & K_0(u) &= \exp\left(-(\tilde{\beta}u)^{2\tilde{\eta}}\right) - c_0 \exp\left(-(\beta u)^{2\eta}\right) \\ & & \tilde{u} &= u + \frac{\|\mathbf{h}\|\cos(\theta-\theta_0)}{v} \end{aligned}$$

with parameters $\alpha, \beta, \tilde{\beta}, c_0, \gamma, \eta, \tilde{\eta}, v$ and θ_0.

Model 10 : The heteroscedastic spatiotemporal kriging approach, where we model the changing variances by a GARCH(1,1) process. With the fitted correlations using our correlation model as stated above, we obtain the non-stationary covariance matrices for the kriging predictor. This model is labeled as Model 13 in Section 5.3.1.

5.4.2 Parameter estimation

5.4.2.1 Regression coefficients for $W(\mathbf{s},t)$

We first estimate the regression coefficients in the model for the latent Gaussian process $W(\mathbf{s},t)$ in (4.8) using maximum likelihood. The parameters are shown in Table 5.6. Note that β_1 is negative, because if the last observation of wind power $y(\mathbf{s},t-1)$ is zero, then the expected value of $W(\mathbf{s},t)$ is $\hat{W}(\mathbf{s},t) = \beta_0 + \beta_1$. Since it is likely that the next wind power observation $y(\mathbf{s},t)$ will be zero, we expect $\hat{W}(\mathbf{s},t) < 0$, which corresponds to $\hat{Y}(\mathbf{s},t) = 0$. For similar reasons, we

expect β_2 and β_3 to be positive. In Table 5.6, we also give the standard errors of the parameters, which are calculated using the jackknife approach (Efron, 1981; Farewell, 1978). We omit the data from one of the wind farms each time, repeat the procedure of parameter estimation and obtain the standard errors of the estimated parameters.

Parameters	β_0	β_1	β_2	β_3	σ_ε
Estimates	.126	-1.44	.705	3.96	.092
Standard errors	7.50×10^{-5}	.069	1.74×10^{-4}	.130	4.27×10^{-5}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

Table 5.6: The estimated regression coefficients in the model for $W(\mathbf{s}, t)$ in (4.8), and the standard errors of the estimated parameters using the jackknife approach. The values in brackets are the corresponding p-values.

5.4.2.2 Correlation model for $W(\mathbf{s}, t)$

Conditional on the estimated regression coefficients β's, we estimate ρ_ε in the correlation model for $W(\mathbf{s}, t)$ in (4.10), using the algorithm of SEM as described in Section 4.3. We estimate that $\hat{\rho}_\varepsilon = 35$. This means that the correlation is halved when the distance $\|\mathbf{s}_i - \mathbf{s}_j\|$ increases by $35 \ln 2 = 24.3$ km. The variation of the expected log likelihood of $W(\mathbf{s}, t)$ for different values of correlation parameter ρ_ε is shown in Figure 5.32.

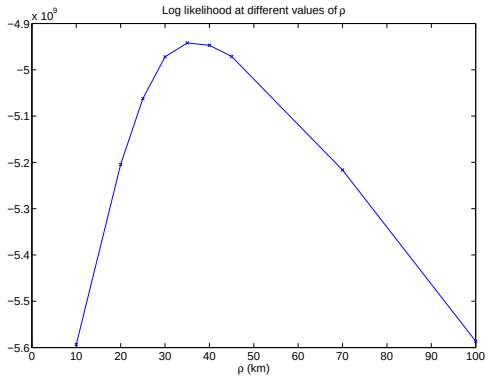


Figure 5.32: Expected log likelihood of $W(\mathbf{s}, t)$ for different values of correlation parameter ρ_ε in the exponential correlation model in (4.10). The estimated value is $\hat{\rho}_\varepsilon = 35$. This means that the correlation is halved when the distance $\|\mathbf{s}_i - \mathbf{s}_j\|$ increases by 24.3 km.

5.4.2.3 Smoothing parameters for $F(y|\ell, s^2)$

Next, we estimate the smoothing parameters in the distribution function $F(y|\ell, s^2)$, which generates the dynamics for the location parameter ℓ and the scale parameter s^2. To obtain a good calibration, we also need to simultaneously estimate the two rescaling parameters. Using the recursive algorithm as described in Section 4.3.3.2, we obtain the following estimates as listed in Table 5.7. In the estimation, we stop at the 3^{rd} iteration for the rescaling parameters and the subsequent 4^{th} iteration for the smoothing parameters, as highlighted in red in Table 5.7. This is because the objective function 2 begins to increase afterwards.

Parameter	θ	α	ϕ_s	γ	ϕ_v	***Obj. 1***		δ_ℓ	k_s	***Obj. 2***
Iteration										
1	236.82	0.19	0.77	0.0003	0.005	*0.031*		-0.005	1.24	*0.0233*
2	25.62	0.47	0.48	0.0008	0.018	*0.033*		-0.008	0.90	*0.0195*
3	19.94	0.86	0.17	0.0102	0.015	*0.031*		-0.004	0.77	*0.0192*
4	21.12	1.05	0.00	0.0092	0.017	*0.030*		-0.004	0.78	*0.0193*
5	21.29	1.01	0.03	0.0090	0.017	*0.030*		-0.004	0.79	*0.0193*

Table 5.7: This table shows the results of estimating the smoothing and rescaling parameters using the iteration method. *'Obj. 1'* and *'Obj. 2'* correspond respectively to the objective functions 1 and 2 on page 100. We stop at the 3^{rd} iteration for the rescaling parameters and the subsequent 4^{th} iteration for the smoothing parameters, as highlighted in red above. This is because the objective function 2 begins to increase afterwards.

5.4.2.4 Correlation model for $Z(\mathbf{s}, t)$

Finally, we need to estimate the parameters in the various correlation models for $Z(\mathbf{s}, t)$ by minimizing the weighted least squares (WLS) given in (3.44). For each correlation model, we need to estimate a set of parameters at each horizon, which vary slightly across horizons. In the case for Irish data, modeling the spatiotemporal correlations is useful in improving forecasts, because the training data is not too long. This is in contrast to the case for the Danish data, which we will see in Chapter 6 that for short forecast horizons $h \leq 4$, the structures of correlations are not significant and the use of empirical correlations actually performs better. The fitted results are similar to those in the spatiotemporal kriging models discussed in Section 5.3.2.2.

5.4.3 Individual forecast evaluations

In our simulations of forecasts, we draw 100 samples for each forecast at all wind farms and all time points within the testing set. We have tried to include more samples, but this does not alter the results significantly. For the sake of comparison with results obtained using spatiotemporal kriging in Section 5.3, we consider the same forecast horizons $h = 1 - 12$, i.e., from 15 minutes up to 3 hours ahead.

For individual forecasts, we use box plots to display the forecast performances at each wind farm. As the differences of forecast performances are relatively small for short horizons, we only consider three hours ahead forecasts, i.e. at forecast horizon $h = 12$. We compare the performances between Model 2 to Model 10, while Model 1 is a univariate model for aggregated wind power generation and so does not generate any individual forecasts.

Results of point forecasts under RMSE are shown in Figure 5.33. The two-stage models with anisotropic correlations, i.e. Model 1 to Model 9 produce superior forecasts over the best spatiotemporal kriging approach, i.e. i.e. Model 10. In the two-stage model, Stein's correlation model performs the best and gives the lowest RMSE. This is slightly different with the results using the spatiotemporal kriging approach as discussed in Section 5.3, where in that case our correlation model performs the best. A reason may due to the fact that correlations for the latent Gaussian process $Z(\mathbf{s}, t)$ is much weaker than those for the wind power generation $Y(\mathbf{s}, t)$. This happens especially for the cross correlations as shown in Figure 4.3. As a result, our correlation may have a problem of overfitting and thus the forecast results are slightly worse than those using Stein's model or the Lagrangian model, which have fewer parameters. Nevertheless, the results of Model 1 to Model 9 are quite similar and the differences in the median are not significant.

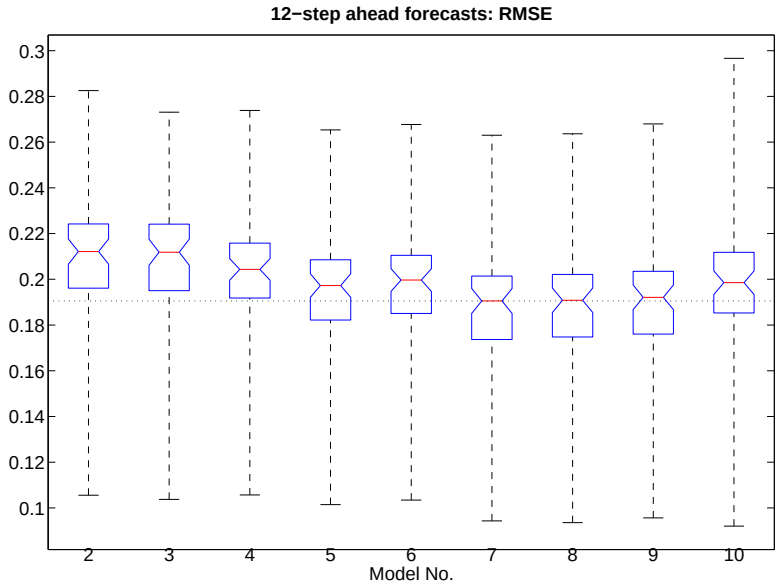


Figure 5.33: The figure shows the box-plots of RMSE of three hours ahead individual point forecasts. The two-stage models with anisotropic correlations, i.e. Model 1 to Model 9 produce superior forecasts over the best spatiotemporal kriging approach, i.e. i.e. Model 10. In the two-stage model, Stein's correlation model performs the best and gives the lowest RMSE. Nevertheless, the results of Model 1 to Model 9 are quite similar and the differences in the median are not significant. *(2: ARIMA(1,1,1)-GARCH(1,1), 3: VAR(1), 4: Empirical (Two-stage), 5: Cressie (Two-stage), 6: Gneiting (Two-stage), 7: Stein (Two-stage), 8: Lagrangian (Two-stage), 9: Our model (Two-stage) (Two-stage), 10: Our model-H)*

For quantile forecasts, we consider a fixed horizon of 12-steps ahead, i.e. 3 hours. We plot the mean of the MQE across the 64 wind farms versus different quantile values. Results are shown in Figure 5.34. Similar to the performances of point forecasts, the two-stage models with Stein's correlation, Lagrangian correlations and our correlation model give the best quantile forecasts in general. Only at extremely large quantile values does the ARIMA(1,1,1)-GARCH(1,1) benchmark perform the best, which is in line with previous discussions. A zoom-in plot for quantiles around 50% is shown in Figure 5.35.

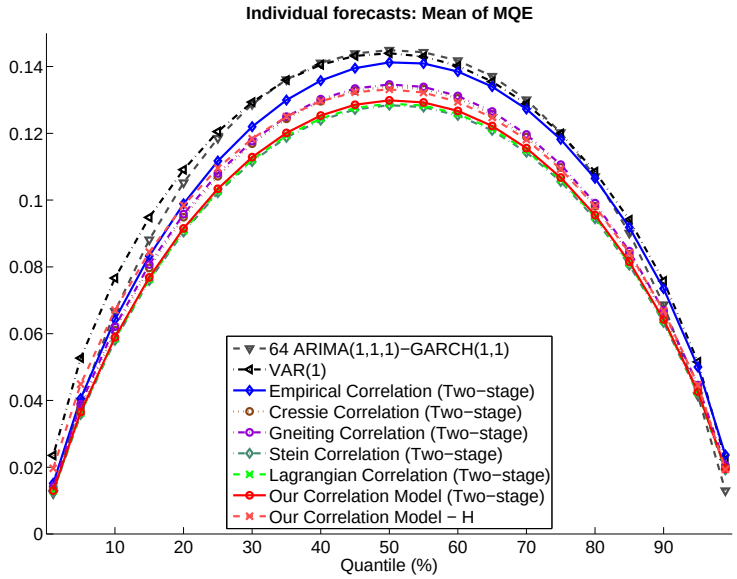


Figure 5.34: The figure shows the plot of MQE of individual quantile forecasts at different quantile values. The two-stage models with Stein's correlation, Lagrangian correlations and our correlation model give the best quantile forecasts in general. Only at extremely large quantile values does the ARIMA(1,1,1)-GARCH(1,1) benchmark perform the best, which is in line with previous discussions.

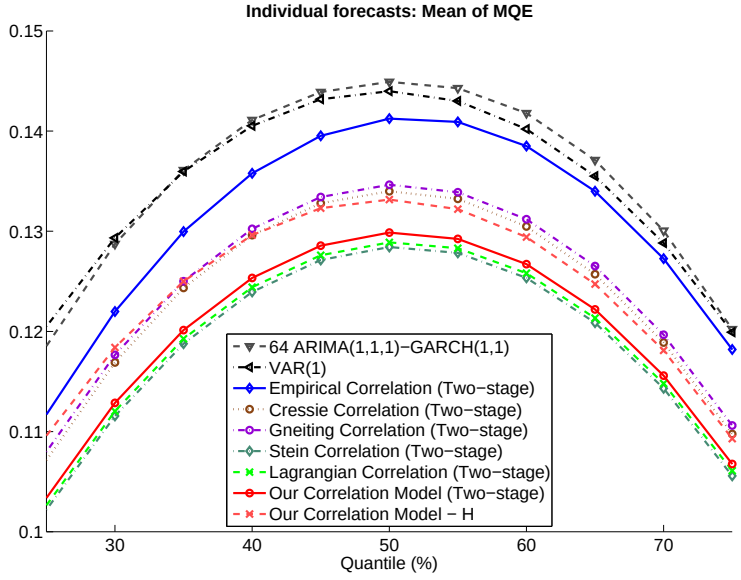


Figure 5.35: The figure shows the zoom-in plot of MQE of Figure 5.34 at quantiles between $25\% - 75\%$.

Finally, for probability forecasts, we consider the CRPS. Figure 5.36 shows the results in box-plots. As in the results of point forecasts under RMSE, the two-stage model with Stein's correlation, i.e. Model 7, performs the best and gives the lowest mean CRPS. The two-stage models with anisotropic correlations, i.e. Model 7 to Model 9 give superior probability forecasts compared with the best spatiotemporal kriging approach, i.e. Model 10. This clearly demonstrates the advantages of using the two-stage model over the spatiotemporal kriging approach.

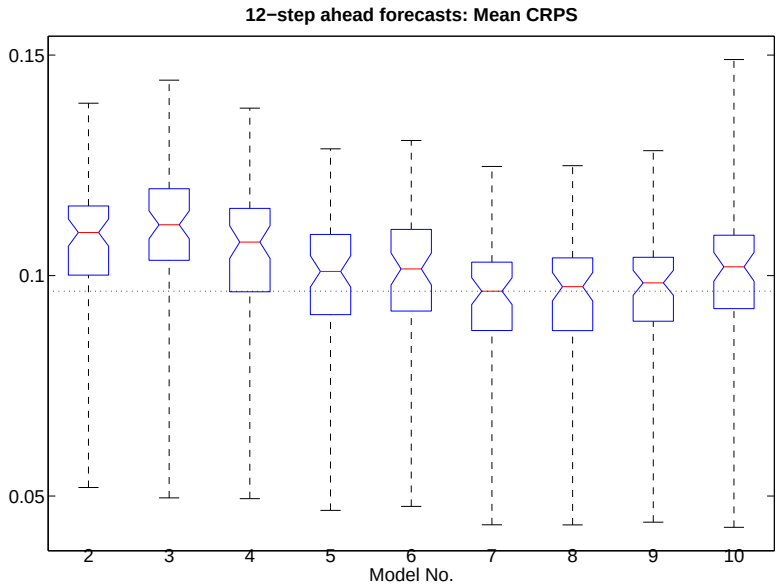


Figure 5.36: The figure shows the box-plots of mean CRPS of three hours ahead individual probability forecasts. The two-stage model with Stein's correlation, i.e. Model 7, performs the best and gives the lowest mean CRPS. The two-stage models with anisotropic correlations, i.e. Model 7 to Model 9 give superior probability forecasts compared with the best spatiotemporal kriging approach, i.e. Model 10. This clearly demonstrates the advantages of using the two-stage model over the spatiotemporal kriging approach. *(2: ARIMA(1,1,1)-GARCH(1,1), 3: VAR(1), 4: Empirical (Two-stage), 5: Cressie (Two-stage), 6: Gneiting (Two-stage), 7: Stein (Two-stage), 8: Lagrangian (Two-stage), 9: Our model (Two-stage) (Two-stage), 10: Our model-H)*

5.4.4 Aggregated forecast evaluations

After evaluating the point forecasts, in this section, we further consider aggregated forecasts for the total wind power generation at all of the 64 wind farms. Here we will include Model 1 in Section 5.4.1, which is the univariate benchmark applied directly on the aggregated wind power time series. We consider a range of forecast horizons at $h = 1 - 12$, i.e. from 15 minutes up to 3 hours ahead.

Performances of aggregated point forecasts under RMSE are shown in Figure 5.37. The best models are the two-stage approaches with Lagrangian correlation and our correlation model. The best spatiotemporal kriging approach with our correlation model also produces competitive forecasts, and even out-performs the two-stage approach with empirical correlations and some simpler isotropic correlation models. It should also be noticed that the univariate ARIMA(4,1,3)-GARCH(1,1) benchmark produces much better point forecasts compared with the spatiotemporal VAR(1) benchmark model.

Quantile forecast performances are shown in Figure 5.38, where we plot the MQE across different quantiles. The two-stage models with Stein's correlation, Lagrangian correlations and our correlation model give the best quantile forecasts in general. Only at extremely large quantile values does the univariate ARIMA(4,1,3)-GARCH(1,1) benchmark perform the best, which is in line with previous discussions. An interesting observation is on the poor quantile forecasts generated by the spatiotemporal kriging approach, especially at quantiles below 20% and over 80%. This clearly shows that by modeling the probability masses explicitly in the two-stage models, we could tremendously improve the forecasts when the normalized wind power observations are close to the extreme values of zero and one.

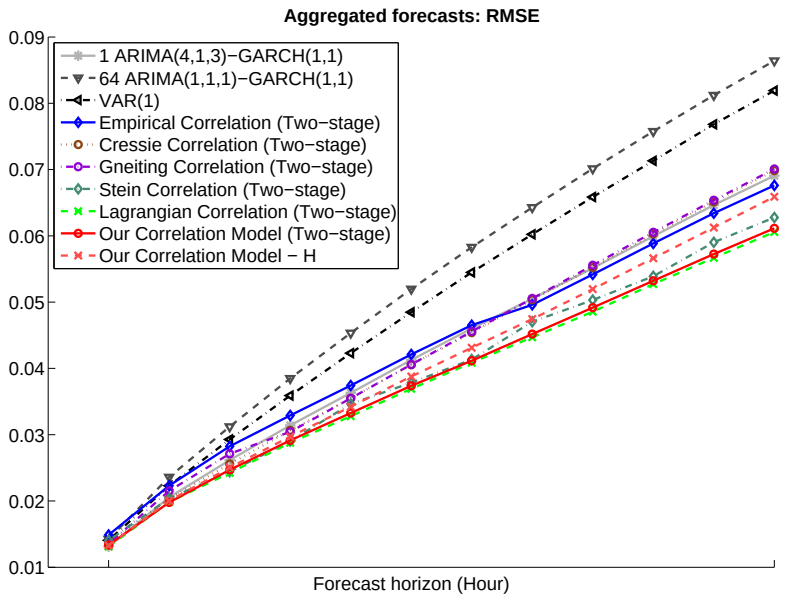


Figure 5.37: RMSE of aggregated point forecasts at forecast horizons 15 minutes up to 3 hours ahead. The best models are the two-stage approaches with Lagrangian correlation and our correlation model. The best spatiotemporal kriging approach with our correlation model also produces competitive forecasts, and even out-performs the two-stage approach with empirical correlations and some simpler isotropic correlation models. It should also be noticed that the univariate ARIMA(4,1,3)-GARCH(1,1) benchmark produces much better point forecasts compared with the spatiotemporal VAR(1) benchmark model.

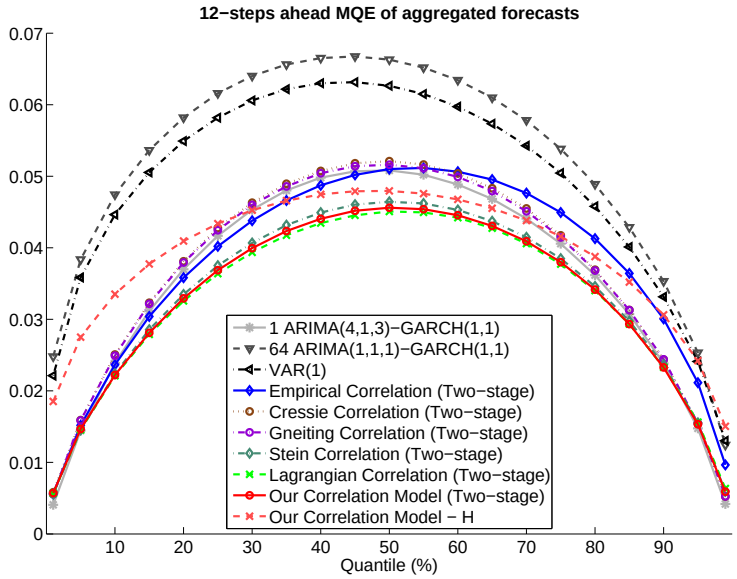


Figure 5.38: The figure shows the plot of MQE of aggregated quantile forecasts at different quantile values. The two-stage models with Stein's correlation, Lagrangian correlations and our correlation model give the best quantile forecasts in general. Only at extremely large quantile values does the univariate ARIMA(4,1,3)-GARCH(1,1) benchmark perform the best, which is in line with previous discussions. An interesting observation is on the poor quantile forecasts generated by the spatiotemporal kriging approach, especially at quantiles below 20% and over 80%. This clearly shows that by modeling the probability masses explicitly in the two-stage models, we could tremendously improve the forecasts when the normalized wind power observations are close to the extreme values of zero and one.

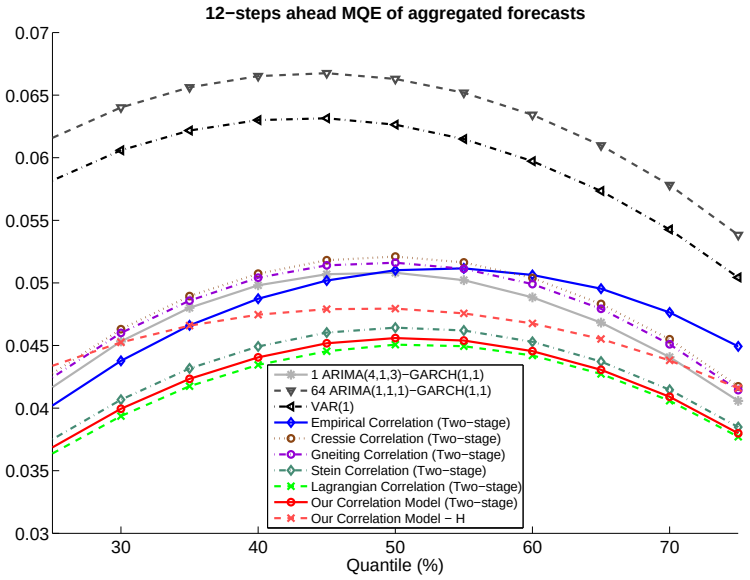


Figure 5.39: The figure shows the zoom-in plot of MQE of Figure 5.38 at quantiles between $25\% - 75\%$.

Finally, we show the mean CRPS of the probability forecasts in Figure 5.40. Again, the two-stage models with Stein's correlation, Lagrangian correlations and our correlation model give the best probability forecasts. An interesting difference between the results of probability forecasts and those of point forecasts in Figure 5.37 is on the performance of the univariate ARIMA(4,1,3)-GARCH(1,1) benchmark. For probability forecasts, this simple benchmark is very competitive indeed. Its performance under mean CRPS is even superior than the best spatiotemporal kriging model, as well as the two-stage models with simpler isotropic correlations. This demonstrates the difficulty of utilizing spatiotemporal information so as to generate superior forecasts for aggregated time series. In other words, if one's interest is solely on the forecasts of aggregated wind power generation, then it may sometimes be better to generate forecasts by directly modeling the aggregated time series.

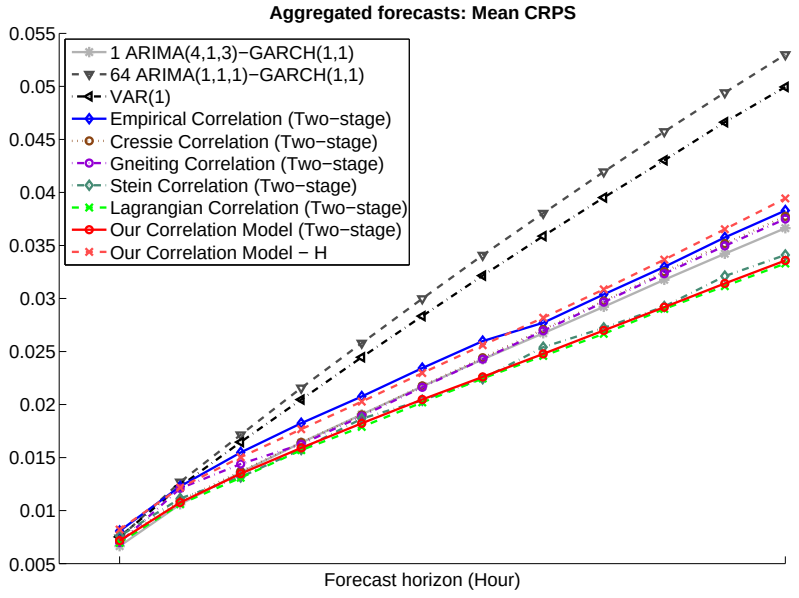


Figure 5.40: Mean CRPS of aggregated probability forecasts at forecast horizons 15 minutes up to 3 hours ahead. The two-stage models with Stein's correlation, Lagrangian correlations and our correlation model give the best probability forecasts. The ARIMA(4,1,3)-GARCH(1,1) benchmark is very competitive. Its performance under mean CRPS is even superior than the best spatiotemporal kriging model, as well as the two-stage models with isotropic correlations.

Chapter 6

Application to Danish Wind Data

6.1 Wind data

After formulating and investigating various approaches of individual and spatiotemporal forecasts in the previous chapters, where we have concentrated on Irish wind data, we now turn our focus to the application of these techniques to another data set to demonstrate the usefulness and robustness of our models. In this chapter, we consider wind power generation in Denmark. We obtain the data from the data provider, the Transmission System Operator in Denmark, through the SafeWind project[1].

The wind power data is recorded at every 15 minutes, which is at the same frequency as the Irish wind data obtained from Eirgrid. This is a nice match and we could compare the results more easily. This data set is cleaner than the Eirgrid data set for Ireland, and has longer periods of valid data for a larger number of wind farms. As a result, we consider a 2-year data set starting from 1-Jan-2009 to 31-Dec-2010, which consists of a total of 70076 observations at each wind farm.

To select a number of wind farms that we would like to consider in our analysis without introducing any survival or sample bias, we impose two reasonable and well-checked criteria on the wind power time series that we use. First, we want to avoid including time series that contain a huge amount zeros. Of course, we expect chains of zeros to occur in wind power data. However, sometimes the zeros

[1]The website of the SafeWind project can be found at http://www.safewind.eu/.

can occur due to the temporary shut down of the turbines for maintenance, and we do not want to include such events that may affect our analysis. Moreover, it is common that large amount of missing data are reflected from unreasonably long periods (e.g. several months) of zeros. As a result, we choose to analyze wind power time series that contain at most 10% of zero observations. This has been verified using the data that 10% of zeros is a reasonable level for a reliable wind power time series in Denmark. Using this filter, a number of time series with up to 40% of zeros are removed. Second, we also want to analyze wind power with more variation such that they are more difficult to forecast and give rise to a more challenging and practical problem. Thus we consider wind power that has a variance of at least 5%. We have checked that this bound of 5% does not lead to removal of many wind power time series. In fact, those time series that are removed are largely overlapped with those with huge amount of zeros, which are already removed using the first filter. In general, selecting time series using the above criteria helps us to avoid using wind data that are unreliable.

Using these two criteria, we select 49 wind farms across Denmark. Their locations are shown in Figure 6.1, with more details being given in Table 6.1. All of them are on-shore wind farms.

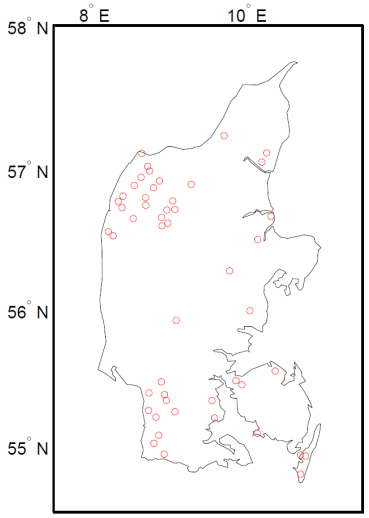


Figure 6.1: The locations of 49 wind farms in Denmark. All of them are on-shore wind farms.

No.	Wind Farm Index	Lat (N)	Lon (E)	Capacity (MW)	Mean	Variance
1	HHO_060	57.11	8.64	6.9	0.29	0.08
2	PAN_060	57.24	9.71	13.11	0.23	0.06
3	FLB_060	56.99	8.74	8.471	0.26	0.07
4	NOR_060	57.02	8.72	14.125	0.25	0.07
5	TRP_060	56.94	8.63	2.268	0.24	0.06
6	VIL_060	56.89	9.28	14.97	0.25	0.06
7	ULD_060	57.12	10.26	10.515	0.19	0.05
8	GAD_060	57.05	10.20	6.955	0.21	0.06
9	BLL_060	56.60	8.90	13.072	0.23	0.06
10	KRE_060	56.66	8.90	9.225	0.22	0.06
11	VND_060	56.62	8.98	11.7	0.22	0.05
12	RSM_060	56.78	9.04	6.925	0.25	0.06
13	STY_060	56.72	9.07	11.88	0.22	0.06
14	RSL_060	56.71	8.97	11.34	0.23	0.06
15	THO_060	56.28	9.78	5.25	0.22	0.05
16	HRL_060	56.50	10.14	9.4355	0.21	0.05
17	WUN_060	56.67	10.31	63	0.24	0.06
18	SVAV060	55.99	10.04	2.95	0.20	0.05
19	SKI_060	55.92	9.09	4.8	0.25	0.07
20	FLAV060	56.87	8.79	11.69	0.25	0.06
21	SIN_060	56.75	8.69	16.75	0.24	0.06
22	SEJ_060	56.92	8.87	5.3	0.27	0.07
23	DRS_060	56.80	8.69	13.151	0.23	0.06
24	HVIV060	56.65	8.53	10.3455	0.26	0.07
25	SNEV060	56.89	8.55	20.09	0.26	0.07
26	HUR_060	56.73	8.39	20.633	0.33	0.08
27	BED_060	56.81	8.40	22.286	0.25	0.06
28	VES_060	56.77	8.34	2.195	0.29	0.08
29	FJE_060	55.33	9.55	3.66	0.21	0.05
30	HEA_060	55.20	9.58	2.649	0.20	0.05
31	GRM_060	55.25	9.07	7.12	0.22	0.06
32	HVG_060	55.26	8.73	9.25	0.21	0.05
33	VIY_060	55.02	8.80	5.391	0.22	0.06
34	HJV_060	55.38	8.94	7.89	0.20	0.06
35	SHY_060	55.34	8.96	2.41	0.21	0.05
36	BBR_060	55.08	8.86	12.96	0.20	0.05
37	FFT_060	55.21	8.82	16.56	0.20	0.05
38	GBR_060	55.39	8.74	5.55	0.24	0.06
39	JEJ_060	54.94	8.93	4.5	0.20	0.05
40	HOSV060	55.47	8.89	2.22	0.23	0.06
41	LVI_060	56.53	8.27	2.95	0.26	0.07
42	KBYV060	56.56	8.21	2.722	0.22	0.06
43	MLM_060	54.92	10.76	11.355	0.24	0.07
44	RUDV060	54.93	10.70	1.705	0.25	0.07
45	TRY_060	54.79	10.70	16.65	0.27	0.07
46	BBY_060	55.55	10.37	11.67	0.21	0.05
47	NBY_060	55.48	9.86	5.55	0.21	0.05
48	EJB_060	55.45	9.94	3	0.22	0.05
49	HNE_060	55.09	10.14	2.165	0.23	0.06

Table 6.1: Summary of wind power data from 49 wind farms in Denmark. The wind farm index are provided according to the data provider, the Transmission System Operator in Denmark, from the SafeWind project (http://www.safewind.eu/).

The variance of wind power in Denmark is significantly lower than that in Ireland, due to that fact that Denmark is a smaller country with a flatter terrain. Figure 6.2 shows the distributions of wind power variance across the 49 wind farms during the 2 year period from 2009 to 2010, where we standardize the results to show the relative variances across different locations. The wind power has a larger (smaller) variance than average if the colour of the dot is red (blue). It is observed that the variance is relatively large at the most windy locations in the north-west coast of Denmark. This is associated with higher mean wind speeds from the north-west (Moller, 1992), and results in higher variability.

In addition, we show the variations of the mean and variance of wind power generation throughout the year. Figure 6.3 shows the corresponding variations for aggregated wind power. It is clear that an increase in wind power generation corresponds to larger variations of wind power output. There are also significant seasonal differences throughout the year, with larger wind power output in the winter (Oct - Jan) and lower in the summer (Jun - Aug). The mean variation among all 49 individual wind farms is very similar and not shown here. Note that although there is a clear seasonality throughout the year, which is expected, we do not consider models with such seasonalities in our following analysis. The reason is that our forecast horizons are from 15 minutes to 3 hours ahead, which are regarded as very short-term wind power forecasts. Unlike forecast at horizons of days or even weeks ahead, in our case, we see that seasonalities at low frequencies do not help to improve short term forecasts. Finally, two examples of Danish wind power time series are shown in Figures 6.4 and 6.5. Wind farm No. 1 is located on the north-west coast and has a relatively large variance of about 8%. Wind farm No. 46 is located on the south-east coast and has a relatively low variance of about 5%.

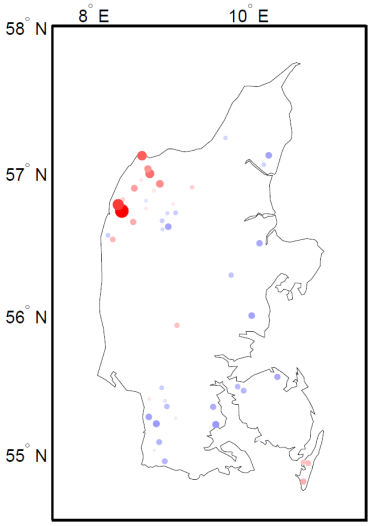


Figure 6.2: The figure shows the variances of wind power across the 49 wind farms in Denmark, where we standardize the results to show the relative variances across different locations. The size and the deepness of the colors of the dots are both proportional to the standardized variance. The wind power has a larger (smaller) variance than average if the colour of the dot is red (blue). As shown in Table 6.1, the range of the variances is quite small and all the variances are approximately between $5\% - 8\%$.

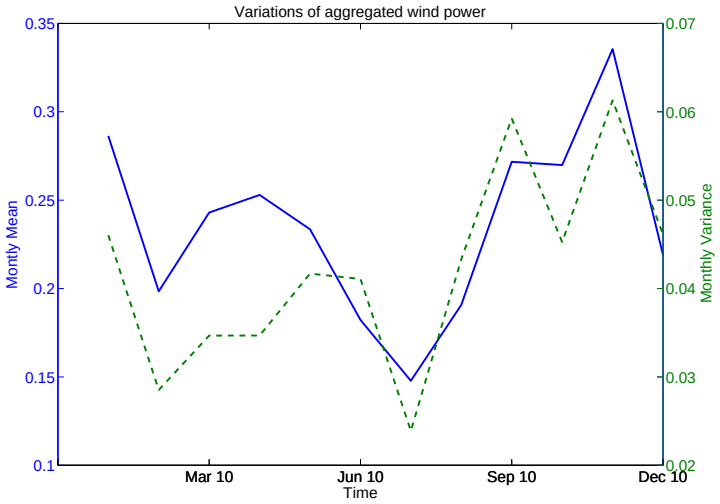


Figure 6.3: The figure shows the variations of the mean and variance of aggregated wind power generation throughout the year in 2010. It is clear that an increase in wind power generation corresponds to larger variations of wind power output. There are also significant seasonal differences throughout the year, with larger wind power output in the winter (Oct - Jan) and lower in the summer (Jun - Aug)

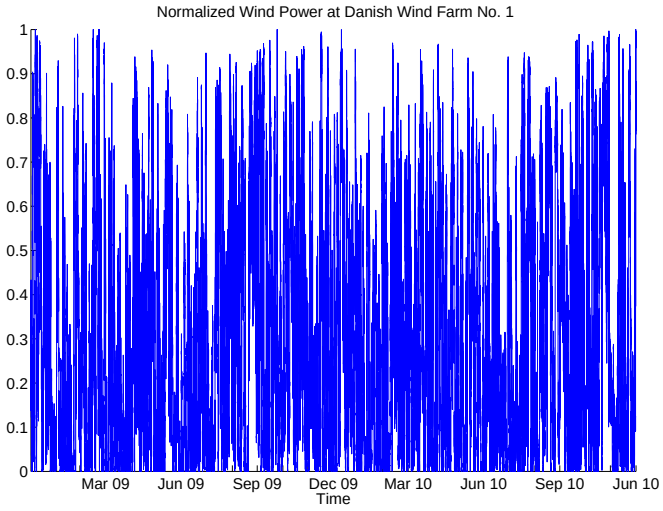


Figure 6.4: Normalized wind power time series of Danish wind farm No. 1, which is located on the north-west coast and has a relatively large variance of about 8%. Percentages of zeros and ones are 8.6% and 0.04% respectively. This time series is quite different from that of the Irish wind power which can have large percentage (e.g. 4%) of maximum wind power.

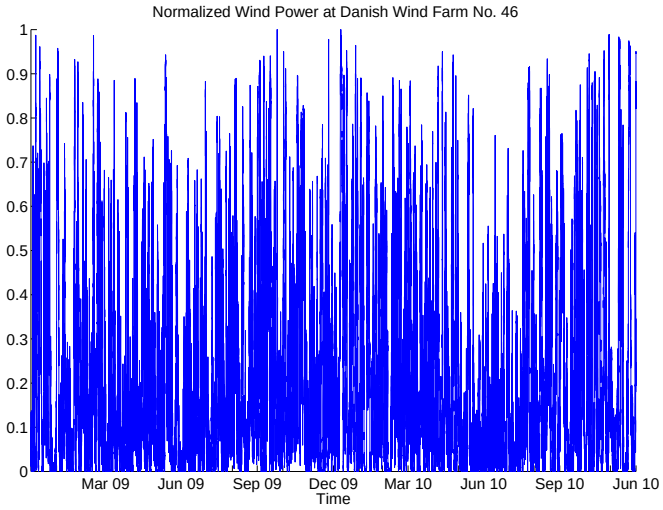


Figure 6.5: Normalized wind power time series of Danish wind farm No. 46, which is located on the south-east coast and has a relatively low variance of about 5%. Percentages of zeros and ones are 7.7% and 0.003% respectively. This time series is quite different from that of the Irish wind power which can have large percentage (e.g. 4%) of maximum wind power.

6.1.1 Training and testing data

To enhance comparison with the Irish data, we use the same forecast approach and models as in Chapter 5. We normalize each wind power time series by dividing by the corresponding capacity of that wind farm. Thus, the data is bounded within $[0, 1]$. The variance of the normalized wind power in Denmark is around 0.05, which is significantly lower than those in Ireland (about 0.08). In this chapter, we are going to forecast the Danish wind power using the following models:

1. Benchmark models, e.g. Persistence forecasts, Vector Autoregressive models, ARIMA-GARCH models, etc.

2. Spatiotemporal kriging models discussed in Chapter 3, with appropriate correlation modeling to account for the spatiotemporal information among wind farms

3. Two-stage model discussed in Chapter 4 to handle the probability masses at zero and one (i.e. maximum wind power generation)

As we have a much longer data set for the Danish wind power, we consider a training set of one year and perform out-of-sample forecasts in the testing set of one month following the training data. We carry out such analysis for each of the 12 months in 2010, in the sense of a rolling estimation and forecast. For instance, we estimate the models using training data from 1-Jan-2009 to 31-Dec-2010, and generate out-of-sample forecasts for observations from 1-Jan-2010 to 31-Jan-2010, and so on. As a result, we could compare forecasts at different times of a year, and analyze how the forecasts from different models behave under different conditions of wind power data. For example, we could examine if larger wind power variations lead to better forecasts in some models, or if the prevailing wind direction in certain months affect the fit of our correlation models and hence forecast results.

6.1.2 Research questions using this data set

With extra data from the Danish wind farms, we hope to investigate additional questions that have not been included in the analysis of Irish wind data. In

particular, due to the much longer length of data available from Danish wind farms, we are able to look into some questions which cannot be analyzed in the Irish case. The major questions that we would like to analyze in this chapter include:

1. How do the performances of various models differ across the year? Would some models perform better in certain months when the wind variability is relatively high/low? How do the ranks among different models change with the year?

2. In the case of Ireland, the weather front generally propagates from north-west to south-east, and so the forecast performances are better in the south-east where information from the north-west could be utilized. Is there a similar phenomenon for wind power generation in Denmark?

3. How robust are the results obtained by various models? What will be the expected performances if we forecast the wind power situated at a completely new location which is out of the original sample?

To answer the above questions, we have fitted the data with various models in a similar fashion as described in Chapters 3 and 4. Forecast performances are compared using metrics including RMSE, CRPS and MQE, with all being described earlier in Appendix D. Again, we consider a forecast horizon of $h = 1 - 12$-step ahead, i.e., from 15 minutes up to 3 hours.

6.2 Forecasts using spatiotemporal models

6.2.1 List of models

In this section, we consider models that will be applied on the spatiotemporal forecasts of the Danish wind data. We select only some of the best models that we have seen in the study of Irish wind data in Chapter 5, because this could allow us to further examine the strengths of those better models. The models that we will consider in this chapter are listed below, and for more details the reader is referred to Section 5.3.1 as all of them have been discussed earlier.

Model 1 : The ARIMA$(p, 1, q)$-GARCH(r, s) model for aggregated wind power, where we estimate the lags p, q, r, s by minimizing the BIC. This could be different for each learning data.

Model 2 : The conditional density forecasts for individual wind power. The conditional densities are obtained by kernel smoothing using the past N hours of data, where we estimate N by minimizing the in-sample mean square errors across all wind farms. The range of N for different training data varies from 9-12.

Model 3 : A univariate ARIMA(1,1,1)-GARCH(1,1) benchmark for each individual wind power time series.

Model 4 : The VAR(1) model as a competitive multivariate benchmark.

Model 5 : Spatiotemporal kriging approach with a volatility model of GARCH(1,1) to account for heteroscedasticity. The kriging predictors are calculated using empirical correlations.

Model 6 : Spatiotemporal kriging approach with a volatility model of GARCH(1,1) to account for heteroscedasticity. The kriging predictors are calculated using our correlation model.

Model 7 : Two-stage model with the correlations of the latent Gaussian process $Z(\mathbf{s}, t)$ being estimated empirically.

Model 8 : Two-stage model with the correlations of the latent Gaussian process $Z(\mathbf{s}, t)$ being estimated using our correlation model.

In the following forecast evaluations, Model 1 to Model 8 above will be compared with regard to aggregated forecasts. For forecasts at individual wind farms, Model 1 will be excluded.

6.2.2 Individual forecasts evaluations

First, we consider the individual forecasts at each of the 49 wind farms. Because of the length of the data set available, we place focus on the variations of monthly forecast performances using different models. To show the results clearly, we decide to fix the forecast horizon at 12-step ahead, i.e. 3 hours. At this horizon, forecast performances of the models differ more significantly and it will be easier to demonstrate the results.

In Figure 6.6, we show the mean of the point forecast performances of all wind farms under RMSE. The results are demonstrated across 12 months from Jan 2010 to Dec 2010. We see that in general the rankings of the models are quite consistent. In particular, the point forecasts obtained from the conditional density benchmark is significantly outperformed by others throughout the year. The VAR(1) benchmark is always better than the 49 individual ARIMA(1,1,1)-GARCH(1,1) benchmark, which is as expected because the former is a multivariate model but the latter is univariate. Due to the large amount of data used in the training set (1 year), the forecast performances using empirical correlations are always difficult to beat. In some cases, the simpler kriging approach with empirical correlations generates even better forecasts than the more complicated two-stage model. Nevertheless, the two-stage models in general still give the best forecasts.

In Figure 6.7, we show the mean of the probabilistic forecast performances of all wind farms under mean CRPS. The results are slightly different from those for point forecasts. Here, the conditional density benchmark can sometimes generate better density forecasts than other models due to its flexibility through non-parametric fitting. The two stage models are still outperforming all other candidates, and again, the models with empirical correlations perform better than our correlation model. In spite of this, modeling the correlations still have two advantages over using empirical correlations. First, a model for spatiotemporal correlations allow one to generate forecasts at locations that are not within the observation samples. Second, when the number of training data points is small, correlation models may do better than the empirical correlations and provide better forecasts.

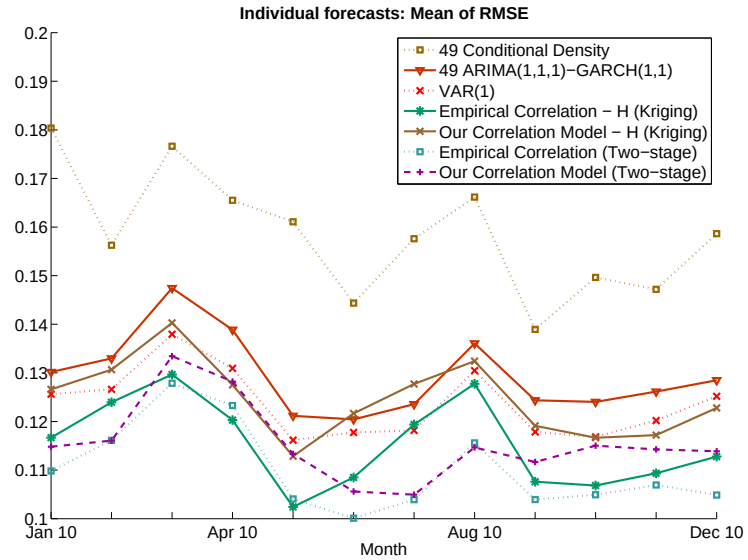


Figure 6.6: Monthly individual point forecasts at a horizon of 3 hours ahead. The results are the mean of RMSE across all 49 wind farms. The point forecasts obtained from the conditional density benchmark is significantly outperformed by others throughout the year. The VAR(1) benchmark is always better than the 49 individual ARIMA(1,1,1)-GARCH(1,1) benchmark, which is as expected because the former is a multivariate model but the latter is univariate. Due to the large amount of data used in the training set (1 year), the forecast performances using empirical correlations are always difficult to beat. In some cases, the simpler kriging approach with empirical correlations generates even better forecasts than the more complicated two-stage model. Nevertheless, the two-stage models in general still give the best forecasts.

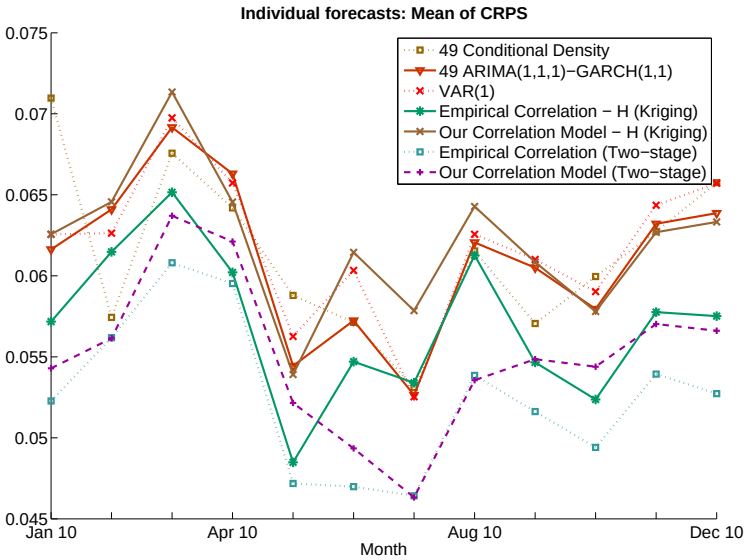


Figure 6.7: Monthly individual probability forecasts at a horizon of 3 hours ahead. The results are the mean of mean CRPS across all 49 wind farms. In contrast to point forecasts, the conditional density benchmark can sometimes generate better density forecasts than other models due to its flexibility through non-parametric fitting. The two stage models are still outperforming all other candidates, and again, the models with empirical correlations perform better than our correlation model.

Apart from showing the mean CRPS of the probability forecasts, we show some results of the evaluations of forecast calibrations. In particular, we consider the probability forecasts at Danish Farm No. 1 in January 2010. We calculate the PIT values of the probability forecasts using the observations, and generate a QQ-plot for the PIT distributions. Results are shown in Figure 6.8. Similar to the evaluations of forecast calibrations for aggregated forecasts in Ireland, we see that the heteroscedastic spatiotemporal kriging models do not produce well-calibrated probability forecasts. They tend to generate over-confident probability densities with small variances. Despite their relatively good performances in terms of RMSE and mean CRPS, it is worthwhile to investigate the issues of calibrations in future research. The conditional density benchmark also produces poorly-calibrated forecast densities. It is interesting to see that the two-stage models give much better forecast calibrations compared with the heteroscedastic spatiotemporal kriging models[1]. Finally, we observe that the univariate ARIMA(1,1,1)-GARCH(1,1) benchmark and the multivariate VAR(1) benchmark both generates relatively well-calibrated forecast densities that are competitive with the two-stage models, although these benchmarks are significantly outperformed in terms of RMSE and mean CRPS.

[1]The homoscedastic spatiotemporal kriging models perform much better than the heteroscedastic counterpart in terms of forecast calibrations, as shown previously in Figures 5.31.

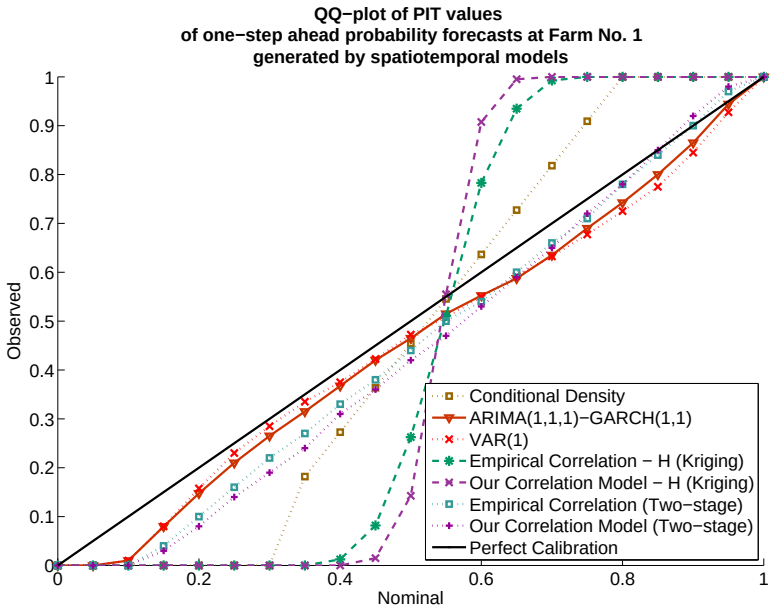


Figure 6.8: This figure shows the QQ-plot of PIT values for one-step ahead probability forecasts at Danish Farm No. 1 in January 2010. Similar to the evaluations of forecast calibrations for aggregated forecasts in Ireland, we see that the heteroscedastic spatiotemporal kriging models do not produce well-calibrated probability forecasts. They tend to generate over-confident probability densities with small variances. The conditional density benchmark also produces poorly-calibrated forecast densities. It is interesting to see that the two-stage models give much better forecast calibrations compared with the heteroscedastic spatiotemporal kriging models. Finally, we observe that the univariate ARIMA(1,1,1)-GARCH(1,1) benchmark and the multivariate VAR(1) benchmark both generates relatively well-calibrated forecast densities that are competitive with the two-stage models, although these benchmarks are significantly outperformed in terms of RMSE and mean CRPS.

6.2.3 Aggregated forecasts evaluations

Next, we consider the aggregated forecast performances, where the aggregated wind power are obtained by adding up all the power from the 49 wind farms, and normalized by the total capacity of 487.259 kW. Results of point forecasts under RMSE are shown in Figure 6.9. The main difference between the results for individual forecasts and those for aggregated forecasts is that the rankings of the models in the latter are more consistent throughout the year. Also, it is more clear that for aggregated forecasts, the forecast errors are in general larger during March, April, October and November. This pattern is less obvious for the individual forecasts.

Regarding the performances, it is clear that in general the benchmarks are significantly out-performed by other models. However, there is again one exception, where the aggregated ARIMA-GARCH model is doing very well. This is expected, as aggregation smoothes out the time series and thus a model fitting directly to the aggregated data could prevent the risk of fitting the noises. Only the two-stage models can consistently out-perform the aggregated benchmark.

For probability forecasts, performances under mean CRPS are shown in Figure 6.10. There is a major difference between results in RMSE and those in CRPS. For point forecasts, it has been seen that the 49 conditional density forecasts produce poor results under RMSE. However, this benchmark is much more competitive regarding probability forecasts under mean CRPS, with results even better than the spatiotemporal kriging approaches. The two-stage models still generate the most competitive probability forecasts. Similar to the results for individual forecasts, the empirical correlations perform very well and in many cases out-perform our correlation models. This is also expected as the length of learning data is very large.

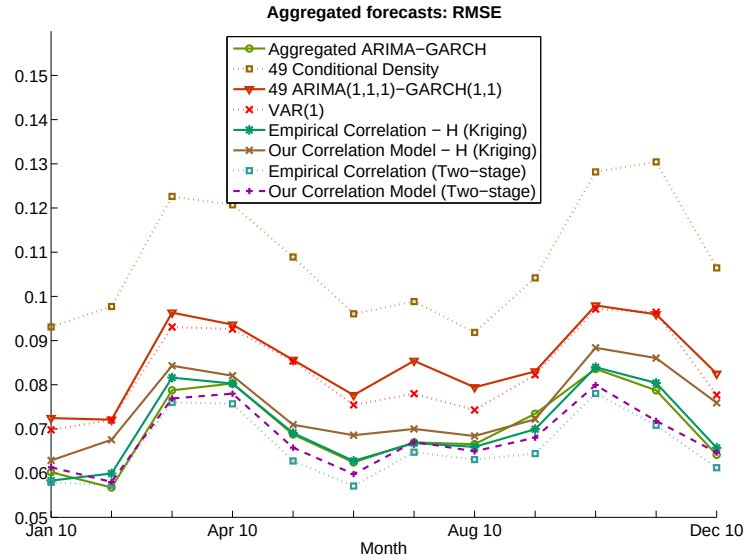


Figure 6.9: Monthly aggregated point forecasts at a horizon of 3 hours ahead. In general, the benchmarks are significantly out-performed by other models, except for the aggregated ARIMA-GARCH model which is doing very well. This is expected, as aggregation smoothes out the time series and thus a model fitting directly to the aggregated data could prevent the risk of fitting the noises. Only the two-stage models can consistently out-perform the aggregated benchmark.

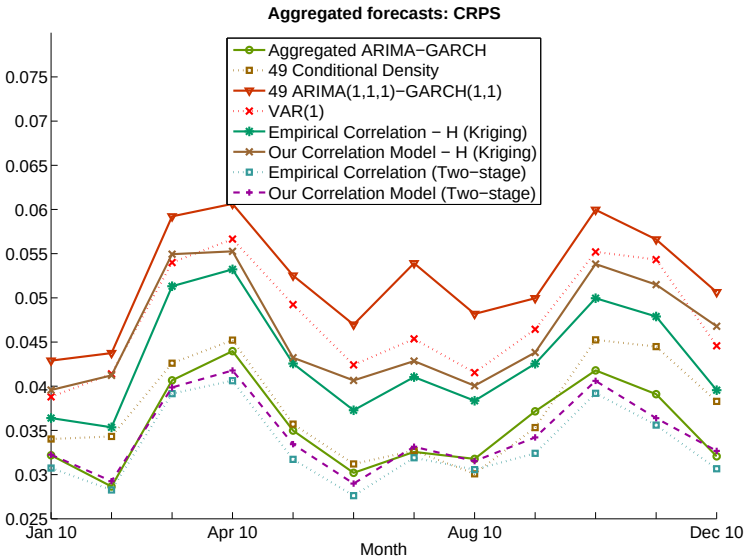


Figure 6.10: Monthly aggregated probability forecasts at a horizon of 3 hours ahead. The 49 conditional density forecasts produce poor results under RMSE, but is much more competitive regarding probability forecasts under mean CRPS. The two-stage models still generate the most competitive probability forecasts.

We also evaluate the forecast calibrations for aggregated forecasts. Again, we consider the probability forecasts in January 2010 so as to compare with results in Figure 6.8. We calculate the PIT values of the probability forecasts using the observations, and generate a QQ-plot for the PIT distributions. Results are shown in Figure 6.11. The forecast densities for the heteroscedastic spatiotemporal kriging models are poorly calibrated and are over-confident. Forecast calibrations for the two-stage models are much better, in particular for the two-stage model with empirical correlations.

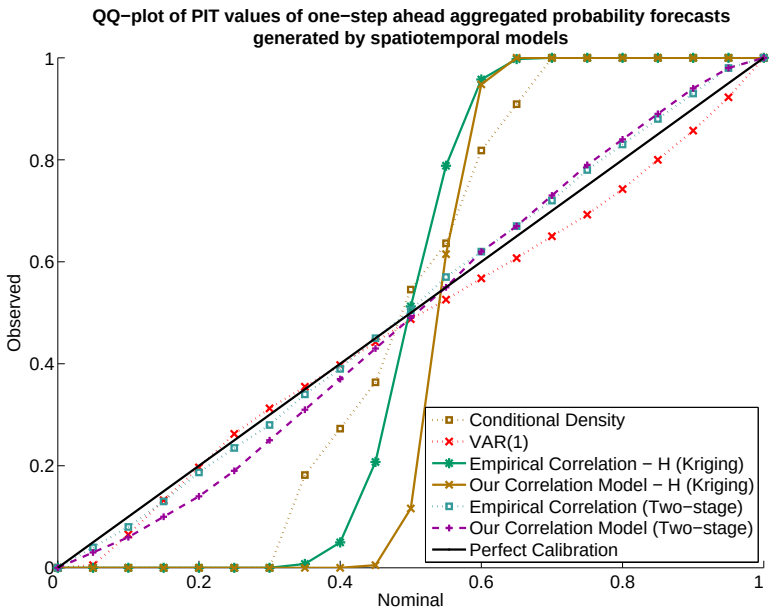


Figure 6.11: This figure shows the QQ-plot of PIT values for one-step ahead aggregated probability forecasts in January 2010. The forecast densities for the heteroscedastic spatiotemporal kriging models are poorly calibrated and are over-confident. Forecast calibrations for the two-stage models are much better, in particular for the two-stage model with empirical correlations.

6.3 Comparison of forecast performances at different locations

In this section, we investigate how the forecast performances could differ across wind farms in Denmark, and to understand if this could be due to the movement of weather front as in the case for the Irish wind data. We compare the forecast performances of our correlation model versus those using empirical correlations. In our correlation model, there is a parameter, θ_0, to account for the direction of the movement of the prevailing weather front. For the Danish data, we find that the azimuth angle of such movement is about $60 - 90°$, meaning that the wind in general blows towards the east. This matches with the prevailing south-westerly wind in Denmark (Barthelmie *et al.*, 1999), showing that our model captures this feature appropriately. To show how this feature helps to improve forecasts, we compare the RMSE of point forecasts between the two-stage approaches with empirical correlations and with our correlation model, i.e. Model 7 and Model 8 respectively. Again, we consider their performances for 3 hours ahead forecasts in Jan 2010. We calculate the percentage differences in RMSE at each wind farm as forecasted using Model 7 and Model 8. Of course, as discussed earlier, the empirical correlations perform exceptionally well due to the long training data used, which is expected. As a result, most percentage differences would be positive, as our correlation model (Model 8) is in general out-performed by the empirical approach (Model 7). Nevertheless, what we want to compare is their relative performances with respect to different locations across Denmark. For this reason, we standardize the percentage differences, reflecting the standard deviation of the percentage difference at a particular location with respect to all 49 wind farms. The results are shown in Figure 6.12. The blue (red) dots indicate that compared with other wind farms, our correlation model performs relatively well (poor). It shows that our correlation model gives better forecasts at locations where it could take advantage of the information obtained in wind farms located at their south-west. In particular, the forecasts using our correlation model is poorer in the south-west coast of Denmark.

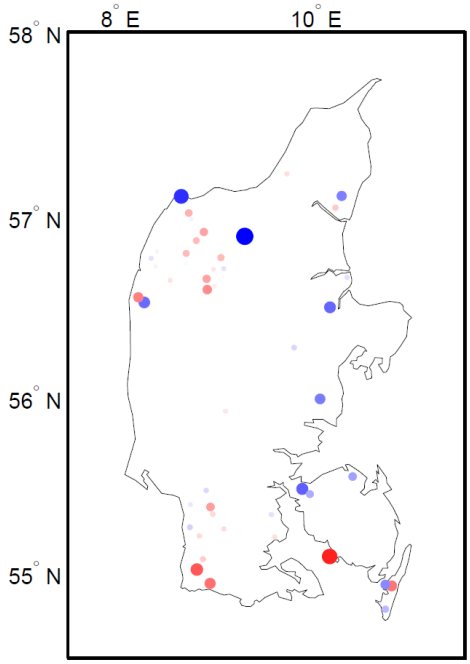


Figure 6.12: Comparison of the percentage differences in RMSE at each wind farm as forecasted using Model 7 and Model 8. We standardize the percentage differences, reflecting the standard deviation of the percentage difference at a particular location with respect to all 49 wind farms. The blue (red) dots indicate that compared with other wind farms, our correlation model performs relatively well (poor). It shows that our correlation model gives better forecasts at locations where it could take advantages from the information obtained in wind farms located at their south-west. In particular, the forecasts using our correlation model is poorer in the south-west coast of Denmark.

6.4 Robustness Analysis

All the previous forecasts considered are out-of-sample in terms of time. In this section, we analyze how our models perform when we consider out-of-sample forecasts in terms of spacial location, i.e. when we forecast wind power at a new location not included in the learning data. This problem is important as in many cases, wind farm investors would like to plan for the locations to install new wind turbines. Also, it could be slow and inefficient to estimate the models using a large learning set. Ideally, we hope the models could be as robust as possible such that parameters estimated using a small sample of wind farms could readily be used to

forecast a much larger data set including wind farms at new locations. Moreover, the test of robustness of our models is an important gauge of their performances at locations that no history of wind farm data are available. In such cases, we could not even compare their performances with benchmark forecasts or simple models that do not require parameter estimations (e.g. Empirical correlations, conditional density forecasts, etc). By comparing the robustness of their spatially and temporally out-of-sample performances at wind farms which we have history observations, we could deduce the expected forecast performances at a completely new location.

6.4.1 Methodology

To test for robustness, we apply a random sampling approach. First, we randomly draw 25 out of 49 wind farms for in-sample parameter estimation. The remaining 24 wind farms will then be used for out-of-sample forecasts, where the forecast performances are recorded under scores such as RMSE and CRPS. To reduce sampling biases and variances, we repeat this procedure 15 times, resulting in a total of 360 randomly chosen, spatially and temporally out-of-sample forecasts. Because of this vast amount of computation required, we decide to choose a shorter learning set of 2 months from 1-Oct-2010 to 30-Nov-2010, which consists of 5856 observations at each wind farm. The testing set for out-of-sample forecasts is the following one-month data from 1-Dec-2010 to 31 Dec-2010.

We select the following four models for the study of both spatially and temporally out-of-sample forecast robustness. The details are given as follows:

1. Spatiotemporal kriging approach with empirical correlations
2. Spatiotemporal kriging approach with our correlation model
3. Two-stage model with empirical correlations
4. Two-stage model with our correlation model

For the spatiotemporal kriging approaches, we estimate δ for the modified logit transformation using the in-sample learning data. As we cannot estimate

any parameters using data in the out-of-sample wind farms, we apply the homoscedastic volatility model for the spatiotemporal kriging approach. This is because we cannot estimate the parameters in the heteroscedastic model, say, GARCH(1,1), using the out-of-sample data. For the models that apply empirical correlations, we estimate the correlations using the history observations in the out-of-sample wind farms, i.e. data from 1-Oct-2010 to 30-Nov-2010 that are observed in the 24 out-of-sample wind farms. We then forecasts for the period from 1-Dec-2010 to 31-Dec-2010, as mentioned previously. For the approach using our correlation model, we estimate the parameters using truly in-sample history observations in the 25 wind farms in the learning set. As our correlation model is simply a function of the latitudes, longitudes and the azimuth angles of the 24 out-of-sample wind farms, we could then input the corresponding information into our calibrated model and obtain the out-of-sample forecasts. For the two-stage models, there are more parameters to be estimated, but all will be done in the same fashion as described above.

To compare the results, we also include the forecast performances using the competitive VAR(1) benchmark. However, note that in this case, we are actually estimating the parameters in the VAR(1) benchmark using the history observations at the 24 'out-of=sample' wind farms. As a result, the VAR(1) model only serves as a benchmark for the forecast performances at those 24 wind farms, and itself is not really an out-of-sample forecasts in terms of spatial locations.

6.4.2 Spatially out-of-sample forecast performances

Using the methodology as described, we obtain 360 out-of-sample forecasts using the 4 models, together with 360 in-sample (in terms of spatial location) forecasts using the VAR(1) benchmark. Each horizon from 15 minutes ahead up to 3 hours ahead will consist of 360 samples, where the 360 samples could be the forecasts for any individual wind farm among the total of 49 in out data set.

Figure 6.13 shows the out-of-sample point forecast performances under RMSE. They are the mean of the RMSE over the 360 randomly drawn samples. Standard deviations are not shown on the figure as they are too difficult to be seen from such scale. Nevertheless, it could be seen that the differences of mean RMSE between

the spatiotemporal kriging approaches and the two-stage models are significant at 95% confidence intervals for a forecast horizon of 12 hours. For instance, the standard deviation of the mean of RMSE at forecast horizon $h = 12$ is of the order of 10^{-4}. Clearly, the VAR(1) benchmark is out-performed by other models, although its parameters are estimated using the in-sample data (in terms of spatial locations). For shorter forecast horizons within 2 hours, the spatiotemporal kriging approach is better than the two-stage models. However, for longer horizons, the two-stage models start to out-perform the spatiotemporal kriging approach, and produce significantly better out-of-sample forecasts in general. Note that the performances of the two-stage models using empirical correlations and our correlation models are exactly the same for forecast horizons up to one hour ahead. This is as described in Chapter 4, as there are no observable structures in the correlations of the latent Gaussian process Z for small horizons, and so it is better to use the empirical correlations directly in such cases. Figure 6.14 shows the out-of-sample probability forecast performances under mean CRPS. They are the mean of the mean CRPS over the 360 randomly drawn samples.

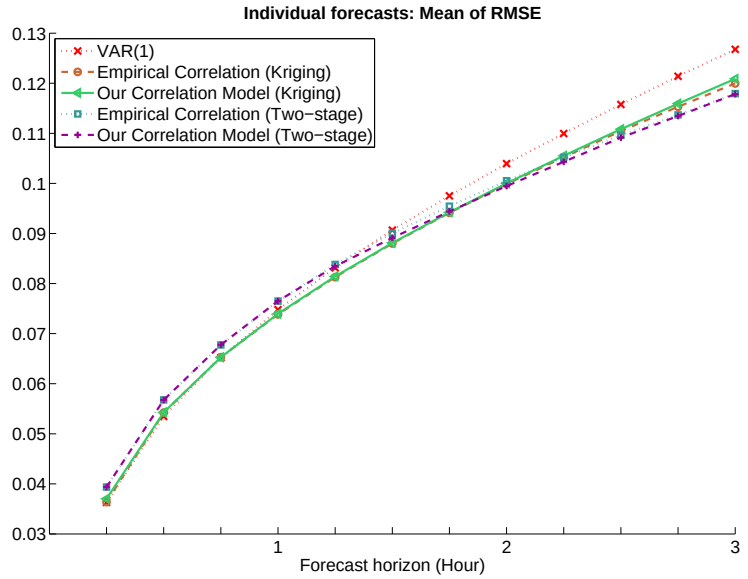


Figure 6.13: The figure shows the mean of RMSE of the out-of-sample point forecasts over the 360 randomly drawn samples. Clearly, the VAR(1) benchmark is out-performed by other models, although its parameters are estimated using the in-sample data (in terms of spatial locations). For shorter forecast horizons within 2 hours, the spatiotemporal kriging approach is better than the two-stage models. However, for longer horizons, the two-stage models start to out-perform the spatiotemporal kriging approach, and produce significantly better out-of-sample forecasts in general.

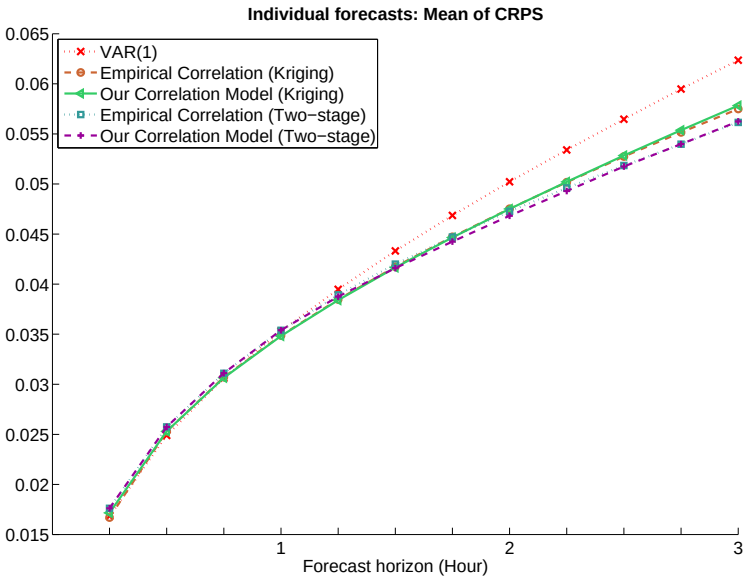


Figure 6.14: The figure shows the mean of CRPS of the out-of-sample probability forecasts over the 360 randomly drawn samples. Results are very similar to those obtained for point forecasts under RMSE.

6.5 Subsampling analysis

Apart from testing for model robustness, we would also like to compare how different subsampling approaches could affect out-of-sample forecast results of a model. As in the case for robustness analysis, we are interested in both spatially and temporally out-of-sample forecasts, as these give important statistics on how the models perform when one considers forecasting at a completely new location. The major concern here is the choice of subsampling approach. How do the spatially out-of-sample forecasts depend on the choice of the in-sample observations, which are used to calibrate our models? To answer this question, we consider four choices of subsampling approaches as follows.

6.5.1 Methodology

In this subsampling analysis, we consider four different combinations of in-sample data and out-of-sample data. To generate spatially out-of-sample forecasts, the in-sample data consists of observations at wind farms different from those in the out-of-sample data. In-sample data is used to calibrate the models, and forecasts are generated for the out-of-sample data. To obtain the statistics of the forecast results, we randomly select wind farms for the in-sample training set and out-of-sample testing set respectively, constrained on one of the four different combinations of these data sets. These combinations are listed as follows, and a summary is given in Table 6.2.

Uniformly random : The in-sample and out-of-sample wind farms are chosen randomly from all the 49 wind farms.

Similar geographical locations : We group the wind farms into four regions as indicated in Figure 6.15. The in-sample and out-of-sample wind farms are drawn from each of the four regions respectively, and so they both consist of wind farms of a similar geographical profile.

Different geographical locations 1 : As opposed to drawing from a similar geographical profiles, we draw wind farms located on the eastern part as the in-sample training set, and wind farms located on the western part as the

out-of-sample testing set. The division between eastern and western parts is as shown in Figure 6.16. We try to include roughly the same number of wind farms in both parts.

Different geographical locations 2 : We draw wind farms located on the western part as the in-sample training set, and wind farms located on the eastern part as the out-of-sample testing set.

Subsampling approach	In-sample farms	Out-of-sample farms
Uniformly random	Uniform Random	Uniform Random
Similar geographical locations	Composed of the 4 regions	Composed of the 4 regions
Different geographical locations 1	Eastern part	Western part
Different geographical locations 2	Western part	Eastern part

Table 6.2: Summary of the four different combinations of subsampling approaches. The 4 regions in the subsampling approach with similar geographical locations are as shown in Figure 6.15. The wind farms are divided into eastern and western parts as shown in Figure 6.16.

Because of the constrains in some of the subsamples, e.g. only wind farms located on the west coast are considered, we choose a smaller number of in-sample and out-of sample wind farms as compared with the robustness analysis in Section 6.4. We draw 15 wind farms in each of the in-sample training set and out-of-sample testing set. To obtain a larger number of sample results, for each of the model, we repeat this subsampling analysis for 24 times, i.e. a total number of $15 \times 24 = 360$ out-of-sample forecasts at individual wind farms are obtained. The training set is from 1-Oct-2010 to 30-Nov-2010, and the testing set is from 1-Dec-2010 to 31 Dec-2010, i.e. the same as that considered in the robustness analysis in Section 6.4.

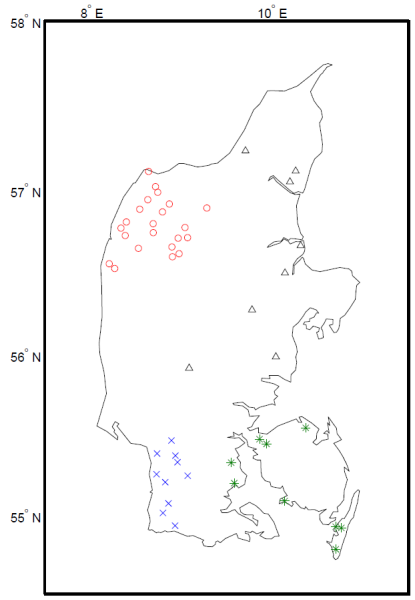


Figure 6.15: We group the 49 wind farms in Denmark according to their spatial locations. We define four regions, namely the north-west group as denoted by the red circles, the south-west group as denoted by the blue crosses, the south-east group as denoted by the green asterisks, and the remaining scatted group as denoted by the black triangles.

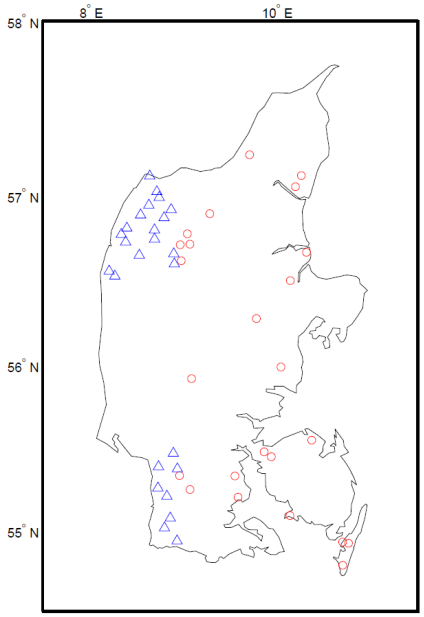


Figure 6.16: We divide the 49 wind farms in Denmark into 25 in the western part and 24 in the eastern part, as denoted by the blue triangles and the red circles respectively.

6.5.2 Results using different subsampling approaches

With the four different subsampling approaches as described above, we generate out-of-sample forecasts for each subsampling. In particular, we consider the results obtained by the homoscedastic spatiotemporal kriging approach with our correlation model. To account for the differences in the out-of-sample data sets considered in each subsampling approach, we use the conditional density benchmark as a reference for the performances, where this benchmark is simply Model 2 in the previous forecast analysis. In other words, we consider the scores obtained using the spatiotemporal kriging approach relative to that obtained using the conditional density benchmark.

In Figure 6.17, we show the results in terms of density forecasts under mean CRPS, where the mean is calculated over 2972 observations in the testing period as well as 360 samples of individual forecasts obtained from the 24 subsamples. The mean CRPS is the relative score obtained by our model as compared with the conditional density benchmark. The random sampling approach gives the worst forecasts. For shorter horizons within 2 hours, forecasts obtained for the western samples based on the training data in the eastern samples are superior than the opposite approach, i.e. forecasts for the eastern samples based on the training data in the western samples. This should due to the fact that the wind farms in the eastern part are more disperse and cover a larger region, thus providing more robust estimated parameters in the model. This results in better forecasts generated at the western wind farms. However, we see that better forecasts at the eastern wind farms are obtained when the forecast horizon is beyond 2 hours. Overall speaking, the forecasts generated by considering subsamplings of similar geographical profiles are the best, which is as expected.

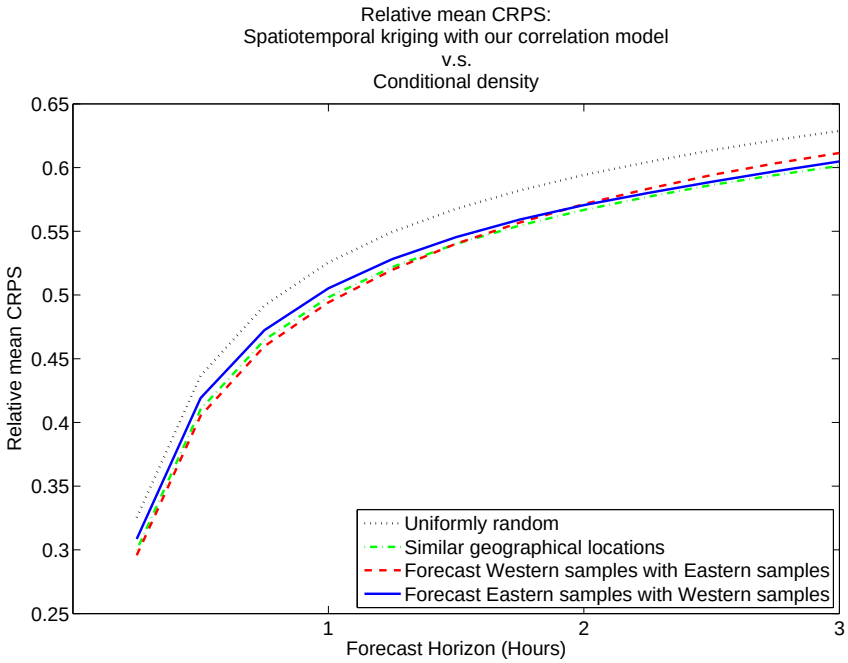


Figure 6.17: The figure shows the relative mean CRPS obtained by the homoscedastic spatiotemporal kriging approach with our correlation model, as compared with that obtained by the conditional density benchmark. The random sampling approach gives the worst forecasts. For shorter horizons within 2 hours, forecasts obtained for the western samples based on the training data in the eastern samples are superior than the opposite approach, i.e. forecasts for the eastern samples based on the training data in the western samples. This should due to the fact that the wind farms in the eastern part are more disperse and cover a larger region, thus providing more robust estimated parameters in the model. This results in better forecasts generated at the western wind farms. However, we see that better forecasts at the eastern wind farms are obtained when the forecast horizon is beyond 2 hours. Overall speaking, the forecasts generated by considering subsamplings of similar geographical profiles are the best, which is as expected.

Chapter 7

Conclusion and Discussion

In this thesis, we work on the problem of short-term wind power forecasting base on the motivation that such forecasts are critical for the successful integration of wind power into electricity grids. The long-term impact could be substantial in both economical and environmental aspects. To generate the forecasts, we need to study the data in details and construct models that could capture the various features of the data. For this reason, we analyze the general properties of portfolios of wind power data such as non-stationarities, distributional characteristics, occurrences of probability masses and anisotropic spatiotemporal correlations. We also need to ensure that the models are robust enough, such that they are useful across different wind power data sets. For this reason, we analyze our models using two large data sets, namely 64 wind farms in Ireland and 49 wind farms in Denmark.

Our work could in general be divided into two categories: aggregated forecasts and spatiotemporal forecasts. In the following, we summarize our major contributions in these two categories.

7.1 Summary of contributions

7.1.1 Aggregated wind power forecasts

In Chapter 2, we investigate two approaches of generating multi-step probability forecasts for aggregated wind power. In the first approach, we demonstrate that

the logit transformation is a good method to normalize aggregated wind power, which is otherwise highly non-Gaussian and nonstationary. We then fit ARIMA-GARCH models with Gaussian innovations for the logit transformed data and obtain competitive probability forecasts. In the second approach, we explore the robustness of exponential smoothing methods and directly apply them to the non-Gaussian wind data. To generate legitimate forecasts within the interval $[0, 1]$, we propose to use parametric distributions that depend on a location parameter and a scale parameter, where the parameters are obtained from the exponential smoothing methods. In particular, we investigate the truncated normal distribution and we demonstrate that such an approach generates robust forecasts with better quantile calibrations than the first approach, especially at extreme quantiles.

Although the approach using exponential smoothing methods with truncated normal distributions cannot fully beat the approach using ARIMA-GARCH models with logit transformed data, it is still a useful alternative to produce competitive probability forecasts due to the following reasons:

1. Forecast performances of the exponential smoothing methods are more robust under different lengths of training data, especially when the size of the training set is relatively small and the estimated parameters in the ARIMA-GARCH models may not be reliable for extrapolating into the testing set. This has been demonstrated using the Irish data, where we take only 40% (as opposed to 67% in Chapter 5) of the data as the training set and the remaining 60% as the testing set. In such case, the ETS$(A, N, N|EC)$-$(A, N, N|EC)$ method performs better than the ARIMA-GARCH models with logit transformed data.

2. In the first approach using ARIMA-GARCH models, we have to select the best model using BIC whenever we consider an updated training set. This is not necessary for the exponential smoothing methods. This advantage is important since in practice many forecasting problems require frequent online updating.

3. The second approach using exponential smoothing allows us to choose a parametric function $D(y|\ell, s^2)$ for the forecast densities, which gives us more flexibilities. One may also generate improved probability forecasts by testing various possible choices of $D(y|\ell, s^2)$. This advantage is particularly important when there are no obvious transformations to normalize the data, and when the data shows evidence that supports simple parametric forecast densities.

7.1.1.1 List of major results

In conclusion, we have developed two approaches of generating multi-step probability forecasts for aggregated wind power, and the major results could be summarized as follows:

- Demonstrate that the logit transformation is a good method to normalize aggregated wind power.
- Show the importance of modeling heteroscedasticity in aggregated wind power.
- Explore the exponential smoothing methods and propose an instantaneous smoothing in both the mean and variance levels, which results in improved forecasts.
- Develop an efficient and robust algorithm to generate multi-step ahead probability forecasts with exponential smoothing and truncated normal distributions, which could be generalized to other non-Gaussian data sets by choosing an appropriate parametric distribution.

7.1.2 Spatiotemporal wind power forecasts

In Chapters 3 and 4, we develop two different spatiotemporal models for the generation of spatiotemporal wind power forecasts. The first one is the kriging model, which is an extension of spatial kriging in geostatistics and relies on the assumption of Gaussian distributions. Here we apply the modified logit transformation to normalize the individual wind power and calculate the corresponding kriging predictor. The second one is the two-stage model with latent Gaussian processes, which is constructed to explicitly model the probability masses in individual wind power distributions. In this model, we draw samples from two latent Gaussian processes and obtain probability densities of the wind power. In this way, aggregated probability forecasts are obtained naturally together with the individual probability forecasts, which is not the case when we apply the spatiotemporal kriging approach.

7.1.2.1 Correlation modeling

In both models, a critical component is the modeling of spatiotemporal correlations. We analyze the spatiotemporal correlations of wind power carefully and explore the anisotropy in the correlation structure. These anisotropies, as shown in both the Irish and Danish data, are supported by the corresponding meteorological references that wind in general blows in the direction of the propagation of weather front. With the major feature in mind, we propose a nonseparable, anisotropic spatiotemporal correlation model for wind power data, which consists of parameters that explicitly explain the anisotropy in terms of the velocity of the weather front movements. Our correlation model, as compared to other simpler isotropic models, is more realistic and captures the main feature of wind power data. As a result, we obtain superior forecasts using our anisotropic correlation model.

7.1.2.2 Two-stage approach

From the analysis of the Irish and Danish wind data, we compare the relative performances of the spatiotemporal kriging models in Chapter 3 with the two-stage

models in Chapter 4. The two-stage model is more sophisticated and involve the direct modeling of the probability masses in individual wind power distributions. Because of the ability to model the occurrence of zero and maximum wind power, quantile forecasts generated by the two-stage models are superior than those generated by the spatiotemporal kriging models, especially at extreme quantiles below 25% and above 75%. It follows that probability forecasts generated by the two-stage models are superior as well.

7.1.2.3 List of major results

In conclusion, we have developed two spatiotemporal models for the generation of multi-step probability forecasts for wind power, and the major results could be summarized as follows:

- Develop a nonseparable, anisotropic spatiotemporal correlation model for wind power data, which consists of parameters that explicitly explain the anisotropy in terms of the velocity of the weather front movements.
- Spatiotemporal models, being able to extract information from the correlation structures of wind power data, are able to generate superior aggregated forecasts as compared to the univariate models that are directly applied on the aggregated data.
- However, it should be emphasized that those univariate models generate very competitive aggregated forecasts as it is easier to model the relatively smooth aggregated data. Only the best spatiotemporal models (e.g. the two-stage models) with realistic correlation structures (e.g. our correlation model) could beat the univariate aggregated forecasts.
- In the spatiotemporal kriging approach, we consider both the homoscedastic and the heteroscedastic version, in which we model the variances in the latter case. Results show that for wind power with larger variability, the heteroscedastic effect is important for capturing the dynamics of volatility and thus generate better probability forecasts as compared to the homoscedastic version. However, further research on the heteroscedastic spatiotemporal

kriging models should be made to investigate how to improve their forecast calibrations.

- Quantile forecasts generated by the two-stage models are superior than those generated by the spatiotemporal kriging models, especially at extreme quantiles below 25% and above 75%. It follows that probability forecasts generated by the two-stage models are superior as well.

7.2 Comparing the Irish and Danish results

Apart from the above general findings, we also gain interesting insights through analyzing two different data sets from Ireland and Denmark. Two observations are made as follows:

1. In our correlation model (3.42), there is a parameter θ_0 which describes the direction of movement of the weather front. Of course, the estimated value of θ_0 differs according to the data set concerned, as well as to whether we are using the spatiotemporal kriging approach or the two-stage model. In both the Irish and Danish data sets, we obtain reasonable values of θ_0 that match with references in meteorology. The value of θ_0 for the Irish data is around 105°, corresponding to a direction of movement towards the south-east. The value of θ_0 for the Danish data is around 60°, corresponding to a direction of movement towards the north-east.

2. The Danish wind data is a much larger data set compared with the Irish data. The training set for the Danish data consists of one year of observations, while the training set for the Irish data only consists of about four months of observations. This lead to the differences in the results obtained by models that incorporate empirical correlations. With a larger number of observations, the empirical correlations estimated in the Danish data are more accurate and are more likely to be able to outperform other correlation models. On the other hand, with fewer training data and higher wind variability, the models that incorporate empirical correlations are outperformed by the anisotropic correlation models in the case of Irish data.

7.3 Discussions and future developments

7.3.1 Aggregated wind power forecasts

The approaches of aggregated wind power forecasts that we develop in Chapter 2 is valuable and one may consider extending the methodology to handle individual wind power generated at a single wind farm. In such case, one could test if better parametric distributions other than the truncated normal distributions could be applied for individual wind power, because unlike aggregated power, individual power has distributions which are further away from Gaussian.

Due to the advances in computational power, it would also be interesting to develop non-parametric models for aggregated forecasts as well. In such case, the generation of multi-step probability forecasts would mostly rely on Monte Carlo simulations. An example is the work of Manzan & Zerom (2008), where forecast densities are produced by non-parametric bootstrapping.

Another approach of aggregated wind power forecast focus on the modeling of power curves that translate wind speed to wind power (Sanchez, 2006). As a result, studies on the modeling of the stochastic power curve would be important and it is worthwhile to investigate more into this area. For sophisticated wind turbines situated at flat terrains, the corresponding power curves could be relatively stable. There would be large potentials to generate superior forecasts through the modeling of both the wind speeds and the power curves.

7.3.2 Spatiotemporal wind power forecasts

In the wind literature, spatiotemporal modeling is becoming more and more important due to the increasing availability of large environmental data sets and the improvements in computational power. The main focus in these models is the correlation structure. One of the interesting approaches is the process-convolution method developed and studied by Higdon (1998), which has been applied to the modeling of ocean temperatures and ozone concentrations. In the process-convolution method, the correlation functions are parameterized via smoothing

kernels. It has the advantage of being able to build nonstationary correlations in a simple way that ensures positive definiteness.

Another challenging and powerful approach for spatiotemporal forecasting is to apply a dynamic state space framework in the spatiotemporal model. One could decompose the stochastic process into components which represent spatial and temporal variabilities at different scales. This approach has the advantage of being capable to explain the physical process in terms of different components, and it is flexible to accommodate complicated problems. Model estimation could be done by a Bayesian approach, which allows one to exploit the full probabilistic inference for the parameters as well as the forecasts (Stroud *et al.*, 2001). However, due to the size and dimensions of spatiotemporal data sets, it may be inconvenient to adopt a computational intensive Bayesian approach. As a result, non-Bayesian approaches such as the method of moments or maximization of the likelihood are common, while the method of moments is more practical since it is usually more efficient and stable. Interesting non-Bayesian approaches include the application of a spatiotemporal Kalman filter (Mardia *et al.*, 1998) to obtain optimal predictions. This spatiotemporal Kalman filter essentially combines spatial kriging with temporal smoothing. Literature devoted to this approach includes Huang & Cressie (1996), Wikle & Cressie (1999) and Huang & Hsu (2004).

7.4 Conclusion

In conclusion, this thesis is a significant contribution to wind power forecasting techniques, and we demonstrate how aggregated and spatiotemporal probability forecasts could be obtained through exponential smoothing methods, kriging predictors, correlation models and two-stage models with latent Gaussian processes. Through the extensive analysis of the Irish and Danish wind power data sets, valuable insights have been added to techniques in wind power modeling and forecasting.

Appendix A

ARIMA-GARCH Processes

A.1 ARMA process

Suppose that the time series $\{Y_t : t = 0, 1, 2, \cdots\}$ is covariance stationary. Furthermore, without loss of generality, assume that $\{Y_t\}$ is purely non-deterministic and $\mathrm{E}[Y_t] = 0$. According to Wold's representation theorem, we can express $\{Y_t\}$ as an infinite moving average (MA(∞)) process

$$Y_t = \varepsilon_t + \sum_{j=1}^{\infty} \psi_j \varepsilon_{t-j} \tag{A.1}$$

where $\sum_{j=1}^{\infty} |\psi_j| < \infty$ and $\{\varepsilon_t\}$ is a white noise process with distribution $D(0, \sigma^2)$. In other words, we have $\mathrm{E}[\varepsilon_t] = 0$ and $\mathrm{E}[\varepsilon_{t-j}\varepsilon_{t-k}] = \sigma^2 \delta_{jk}$. The linear process in (A.1) has an infinite number of parameters. In practice, we can approximate the process with a finite number of parameters by introducing autoregressive terms into the moving average process. In general, we consider an autoregressive moving average (ARMA) process

$$Y_t = \sum_{i=1}^{p} \phi_i Y_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \tag{A.2}$$

We can write this in a compact form

$$\Phi(B)Y_t = \Theta(B)\varepsilon_t \tag{A.3}$$

where $\Phi(B)$ and $\Psi(B)$ are polynomials of the backshift operator B of order p and q respectively, i.e. $\Phi(B) = 1 - \sum_{j=1}^{p} \phi_j B^j$, $\Theta(B) = 1 + \sum_{j=1}^{q} \theta_j B^j$ and $B^k Y_t = Y_{t-k}$. This is denoted as the ARMA(p, q) process. If the roots of $\Phi(z) = 0$ lie outside the unit circle, then $\{Y_t\}$ is stationary and it can be expressed as a linear sum of white noise. For the simplest case where $p = 1$, we require $|\phi_1| < 1$. If $|\phi_1| = 1$, we say that a unit root exists and we should take the first difference of the time series and consider $W_t = Y_t - Y_{t-1}$. If W_t follows an ARMA(p, q) process, then Y_t follows an autoregressive integrated moving average (ARIMA) process denoted by ARIMA(p, d, q) with $d = 1$. On the other hand, in order to express the white noise $\{\varepsilon_t\}$ in terms of the present and past values of y_t, we impose an invertible condition such that the roots of $\Theta(z) = 0$ lie outside the unit circle (Box *et al.*, 1994). Since $\{Y_t\}$ is covariance stationary, its unconditional mean $\mathrm{E}[Y_t] = 0$ is constant. However, its conditional mean $\mathrm{E}[Y_t|\mathcal{F}_{t-1}]$, where the information set is taken as $\mathcal{F}_{t-1} = \{y_k, \varepsilon_k\}_{k=-\infty}^{t-1}$, varies with time according to (A.2). On the other hand, the unconditional variance is equal to the conditional variance, and we have $\mathrm{E}[Y_t^2] = \mathrm{E}[Y_t^2|\mathcal{F}_{t-1}] = \sigma^2$.

A.2 GARCH process

Sometimes we would also like to model the variance of $\{Y_t\}$ since there may be evidences showing that the series $\{\varepsilon_t^2\}$ is time-varying. This phenomenon has long been observed in financial time series, where the volatility of the returns exhibits persistence and clustering in time. Engle (1982) proposed the ARCH(r) model for the white noise process $\{\varepsilon_t\}$, which is defined by

$$\begin{aligned} \varepsilon_t &= \sigma_t \nu_t, \qquad \nu_t \overset{i.i.d.}{\sim} D(0,1) \\ \sigma_t^2 &= \omega + \sum_{j=1}^{r} \alpha_j \varepsilon_{t-j}^2 \end{aligned} \tag{A.4}$$

where $D(0,1)$ denotes a distribution with mean zero and a unit variance. A white noise process $\{\varepsilon_t\}$ satisfying (A.4) is called an autoregressive conditional heteroscedastic (ARCH) process of order r. Note that $\{\varepsilon_t\}$ are uncorrelated but not i.i.d. random variables. (A.4) is analogous to the AR(p) process for the conditional mean, except that it is modeling the variance instead of the mean of a time series. Since the variance, unlike the mean, must be positive, we require that $\omega \geq 0$ and $\alpha_j > 0$ for all j. To ensure that the process is covariance stationary, we impose the condition that $\sum_{j=1}^{r} \alpha_j < 1$. This process allows the conditional variance to change with time according to (A.4), while the unconditional variance remains constant and is given by $\mathrm{E}[\varepsilon_t^2] = \omega / \left(1 - \sum_{j=1}^{r} \alpha_j\right)$ (Tsay, 2005).

In general, σ_t^2 may depend on an infinite number of lags in $\{\varepsilon_{t-j}^2\}_{j=1}^{\infty}$. Bollerslev (1986) proposed that we may approximate the ARCH(∞) process by introducing a finite number of lag $\{\sigma_{t-j}^2\}_{j=1}^{s}$ into the ARCH model. This is called the generalized autoregressive conditional heteroscedastic (GARCH) process. The GARCH(r, s) process is defined by

$$\begin{aligned} \varepsilon_t &= \sigma_t \nu_t, \qquad \nu_t \overset{i.i.d.}{\sim} D(0,1) \\ \sigma_t^2 &= \omega + \sum_{i=1}^{r} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2 \end{aligned} \tag{A.5}$$

This process is similar to the ARMA(p, q) process for the conditional mean. In fact, it is well known that if $\{\varepsilon_t\}$ follows a GARCH(r, s) process, then $\{\varepsilon_t^2\}$ follows an ARMA(p, r) process where $p = \max(r, s)$. To ensure that σ_t^2 are positive, we require $\alpha_i > 0$ and $\beta_j > 0$ for all i, j. To ensure that the unconditional variance of $\{\varepsilon_t\}$ is finite, we require that $\Sigma_{j=1}^{max(r,s)}(\alpha_j + \beta_j) < 1$ where $\alpha_j = 0$ for $j > r$ and $\beta_j = 0$ for $j > s$ Tsay (2005).

Combining the ARMA process for the mean and the GARCH process for the

variance, an ARMA(p,q)-GARCH(r,s) process for Y_t can be written as

$$\begin{aligned} Y_t &= \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \\ \varepsilon_t &= \sigma_t \nu_t, \qquad \nu_t \overset{i.i.d.}{\sim} D(0,1) \\ \sigma_t^2 &= \omega + \sum_{i=1}^{r} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2 \end{aligned} \tag{A.6}$$

where $\phi_i, \theta_j, \omega, \alpha_i, \beta_j$ are constant coefficients satisfying conditions as mentioned previously.

A.3 Optimal multi-step forecasts

A.3.1 Cost function

Equipped with an ARMA(p,q) or ARMA(p,q)-GARCH(r,s) process as the data generating process, the next task is to obtain optimal forecasts. In order to define optimality, we have to determine a criterion so that judgements on different forecasts could be objectively based on it. This criterion is called a cost function, which is a function of both the random variable and the forecast. For example, suppose that our forecast for Y_{t+h} is $\hat{y}_{t+h|t}$. Then the cost function is $C(Y_{t+h}, \hat{y}_{t+h|t})$. In general, we may assume that the cost function depends only on the forecast error, i.e. we consider $C(e_{t+h})$ where

$$e_{t+h} = Y_{t+h} - \hat{y}_{t+h|t} \tag{A.7}$$

In this case, we say that a forecast is optimal if the expected cost $\mathrm{E}[C(e_{t+h})|\mathcal{F}_t]$ is minimized, where $\mathcal{F}_t$ is the information set at time t and typically consists of all linear functions of past observations $\{y_k\}_{k=-\infty}^{t}$.

A common cost function being adopted is the quadratic function so that $C(e) = Ke^2$ for some positive constant K. It is well known that if the cost

function is symmetric, then the optimal forecast is given by the conditional mean of Y_{t+h}, i.e.

$$\hat{y}_{t+h|t} = \mathrm{E}[Y_{t+h}|\mathcal{F}_t] \tag{A.8}$$

Furthermore, if the distribution of the white noise $\{\varepsilon_t\}$ is Gaussian, the optimal forecast is a linear sum of the present and past values $\{y_{t-j}\}_{j=0}^{\infty}$. This optimal forecast is also known as the minimum mean square error (MMSE) forecast, since it minimizes the expected square error.

A.3.2 Optimal forecast for ARMA process

Now we consider the optimal forecast for an ARMA(p, q) model in (A.2). It would be convenient to express this model in the MA(∞) form since we can apply $\mathrm{E}[\varepsilon_{t+h}|\mathcal{F}_t] = \mathrm{E}[\varepsilon_{t+h}] = 0$ when we compute optimal forecasts using (A.8). We write

$$Y_t = \Psi(B)\varepsilon_t = \sum_{k=0}^{\infty} \psi_k \varepsilon_{t-k} \tag{A.9}$$

where $\psi_0 = 1$. According to (A.9), the optimal forecast is then easily obtained as

$$\begin{aligned} \hat{y}_{t+h|t} &= \mathrm{E}[Y_{t+h}|\mathcal{F}_t] \\ &= E\left[\varepsilon_{t+h} + \sum_{j=1}^{\infty} \psi_j \varepsilon_{t+h-j} \middle| \mathcal{F}_t\right] \\ &= \sum_{j=0}^{\infty} \psi_{h+j} \varepsilon_{t-j} \end{aligned} \tag{A.10}$$

since we have $\mathrm{E}[\varepsilon_{t-j}|\mathcal{F}_t] = \varepsilon_{t-j}$ for $j \geq 0$ and $\mathrm{E}[\varepsilon_{t-j}|\mathcal{F}_t] = 0$ for $j < 0$.

One may want to obtain the optimal forecast in terms of the original ARMA coefficients since ψ_j's may not be available explicitly. To do so, we compare (A.9) with (A.3) and obtain $\Phi(B)\Psi(B) = \Theta(B)$, which can be expressed explicitly as

$$\left(1 - \sum_{j=1}^{p} \phi_j B^j\right)\left(\sum_{k=0}^{\infty} \psi_k B^k\right) = \left(1 + \sum_{j=1}^{q} \theta_j B^j\right) \tag{A.11}$$

Equating the coefficients of B^k, we have (Granger & Newbold, 1986)

$$\psi_k - \sum_{j=1}^{p} \phi_j \psi_{k-j} = \theta_k, \qquad k = 1, 2, 3, \ldots \tag{A.12}$$

where $\psi_0 = 1$ and $\psi_k = 0$ for all $k < 0$. (A.10) together with the recurrence equation for ψ_k in (A.12) gives

$$\begin{aligned}
\hat{y}_{t+h|t} - \sum_{j=1}^{p} \phi_j \hat{y}_{t+h-j|t} &= \sum_{i=0}^{\infty} \psi_{h+i} \varepsilon_{t-i} - \sum_{j=1}^{p} \left(\phi_j \sum_{i=0}^{\infty} \psi_{h+i-j} \varepsilon_{t-i} \right) \\
&= \sum_{i=0}^{\infty} \left(\psi_{h+i} - \sum_{j=1}^{p} \phi_j \psi_{h+i-j} \right) \varepsilon_{t-i} \\
&= \sum_{i=0}^{\infty} \theta_{h+i} \varepsilon_{t-i} \\
\Rightarrow \qquad \hat{y}_{t+h|t} &= \sum_{j=1}^{p} \phi_j \hat{y}_{t+h-j|t} + \sum_{i=0}^{\infty} \theta_{h+i} \varepsilon_{t-i}
\end{aligned} \tag{A.13}$$

Taking $\hat{y}_{t-j|t} = y_{t-j}$ for $j \geq 0$, (A.13) provides a useful recursive formula for generating the optimal h-step ahead forecasts based on optimal forecasts within horizon $h - 1$, together with present and past one-step ahead forecast errors $\{\varepsilon_{t-i}\}_{i=0}^{\infty}$ where $\varepsilon_{t-i} = y_{t-i} - \hat{y}_{t-i|t-i-1}$.

A.3.3 Optimal forecast for ARIMA process

The optimal forecasts for an ARIMA model can be obtained from (A.13) in a straight forward way. For simplicity, suppose that the first differenced time series $W_t = Y_t - Y_{t-1}$ follows an ARMA(p, q) process, i.e. Y_t follows an ARIMA$(p, 1, q)$ process. We can obtain the optimal forecast $\hat{w}_{t+1|t}$ for W_{t+1} by (A.13). The optimal forecast for Y_{t+1} is then $\hat{y}_{t+1|t} = y_t + \hat{w}_{t+1|t}$, and the optimal forecast for

Y_{t+h} is

$$\begin{aligned}\hat{y}_{t+h|t} &= \hat{y}_{t+h-1|t} + \hat{w}_{t+h|t} \\ &= \hat{y}_{t+h-2|t} + \hat{w}_{t+h-1|t} + \hat{w}_{t+h|t} \\ &= \cdots \\ &= y_t + \sum_{j=1}^{h} \hat{w}_{t+j|t} \end{aligned} \tag{A.14}$$

As an example, let us compute the optimal forecast for an ARIMA(0,1,1) process. In this case $W_t = Y_t - Y_{t-1}$ follows an ARMA$(0,1)$ process, and with reference to the general ARMA(p,q) process in (A.2) we only have one non-zero coefficient θ_1. By (A.13), we have $\hat{w}_{t+1|t} = \theta_1 \varepsilon_t$ and $\hat{w}_{t+h|t} = 0$ for all $h > 1$. As a result, by (A.14), for $h = 1, 2, \ldots$ we have

$$\hat{y}_{t+h|t} = y_t + \theta_1 \varepsilon_t \tag{A.15}$$

As another example, let us also compute the optimal forecast for an ARIMA(1,1,1) process. Here we have two non-zero coefficients ϕ_1 and θ_1. By (A.13), we have $\hat{w}_{t+1|t} = \phi_1 w_t + \theta_1 \varepsilon_t$ and $\hat{w}_{t+j|t} = \phi_1 \hat{w}_{t+j-1|t}$ for $j > 1$. Recursively, for $j = 1, 2, \ldots$ we have

$$\hat{w}_{t+j|t} = \phi_1^{j-1}(\phi_1 w_t + \theta_1 \varepsilon_t) \tag{A.16}$$

Substituting (A.16) into (A.14) and writing $w_t = y_t - y_{t-1}$, for $h = 1, 2, \ldots$ we have

$$\begin{aligned}\hat{y}_{t+h|t} &= y_t + \sum_{j=1}^{h} [\phi_1^{j-1}(\phi_1 w_t + \theta_1 \varepsilon_t)] \\ &= \phi_1^h y_t + \left(\frac{1-\phi_1^h}{1-\phi_1}\right)(y_t - \phi_1 y_{t-1} + \theta_1 \varepsilon_t) \end{aligned} \tag{A.17}$$

Finally, we close this section by making a remark on some of the well known properties of optimal forecasts (Diebold, 2007). First, optimal forecasts must be unbiased, as seen from the expression as a conditional mean in (A.8). More

importantly, the series of h-step ahead forecasts errors $\{e_{t+h|t} : t \in T\}$ is correlated and is in general an MA$(h-1)$ process. As a result, this must be taken into account when we carry out forecast evaluation for h-step ahead forecasts.

A.4 Variance of forecast errors

A.4.1 Conditional homoscedastic white noise

Once we have obtained the optimal forecasts, we would also want to know about the variances of the forecast errors. Using (A.9) and (A.10), the forecast error of the optimal h-step ahead forecast is

$$\begin{aligned} e_{t+h} &= Y_{t+h} - \hat{y}_{t+h|t} \\ &= \sum_{j=0}^{h-1} \psi_j \varepsilon_{t+h-j} \end{aligned} \tag{A.18}$$

This indeed shows that the series of h-step ahead forecast errors $\{e_{t+h|t} : t \in T\}$ follows an MA$(h-1)$ process and is in general correlated. (A.18) implies that the variance of the forecast errors is

$$\begin{aligned} \mathrm{Var}[e_{t+h}|\mathcal{F}_t] &= \mathrm{E}[e_{t+h}^2|\mathcal{F}_t] - \mathrm{E}[e_{t+h}|\mathcal{F}_t]^2 \\ &= E\left[\left(\sum_{j=0}^{h-1} \psi_j \varepsilon_{t+h-j}\right)^2 \middle| \mathcal{F}_t\right] \\ &= \sigma^2 \sum_{j=0}^{h-1} \psi_j^2 \end{aligned} \tag{A.19}$$

since $\mathrm{E}[e_{t+h}|\mathcal{F}_t] = 0$ for optimal forecasts, and we assume $\mathrm{E}[\varepsilon_{t+j}\varepsilon_{t+k}] = \sigma^2\delta_{jk}$.

The variances of forecast errors of an ARIMA process can also be found in a similar way. Suppose that $W_t = Y_t - Y_{t-1}$ follows an ARMA(p,q) process. Using

(A.14) and (A.18), the forecast error of the optimal h-step ahead forecast is

$$\begin{aligned} e_{t+h} &= Y_{t+h} - \hat{y}_{t+h|t} \\ &= \left(y_t + \sum_{j=1}^{h} W_{t+j} \right) - \left(y_t + \sum_{j=1}^{h} \hat{w}_{t+j|t} \right) \\ &= \sum_{j=1}^{h} \eta_{t+j} \\ &= \sum_{j=1}^{h} \left(\sum_{k=0}^{j-1} \psi_k \varepsilon_{t+j-k} \right) \\ &= \sum_{j=1}^{h} \Omega_{h-j} \varepsilon_{t+j} \end{aligned} \tag{A.20}$$

where $\eta_{t+h} = (W_{t+h} - \hat{w}_{t+h|t})$ is the h-step ahead forecast error of W_{t+h} and $\Omega_k = \Sigma_{j=0}^{k} \psi_j$ so that ψ_j are the coefficients of ε_j when W_t is expressed as an MA(∞) process as in (A.9). The variance of forecast errors is then given by

$$\mathrm{Var}[e_{t+h}|\mathcal{F}_t] = \sigma^2 \sum_{j=1}^{h} \Omega_{h-j}^2 \tag{A.21}$$

where we assume $\mathrm{E}[\varepsilon_{t+j}\varepsilon_{t+k}] = \sigma^2 \delta_{jk}$.

As an example, let us consider the variance of forecast errors for the ARIMA(0,1,1) process. Here we only have one non-zero coefficient θ_1, thus we have $\psi_1 = \theta_1$. As a result, $\Omega_0 = 1$ and $\Omega_j = 1 + \theta_1$ for all $j \geq 1$. By (A.21), the variance of the h-step ahead forecast errors is

$$\mathrm{Var}[e_{t+h}|\mathcal{F}_t] = \sigma^2[1 + (h-1)(1+\theta_1)^2] \tag{A.22}$$

As another example, let us consider the variance of forecast errors for the ARIMA(1,1,1) process. Now we have two non-zero coefficients ϕ_1 and θ_1. Using (A.12), we obtain $\psi_j = \phi_1^{j-1}(\phi_1 + \theta_1)$ for $j \geq 1$. As a result, $\Omega_0 = 1$ and $\Omega_j = 1 + (\theta_1 + \phi_1)(1 - \phi_1^j)/(1 - \phi_1)$ for all $j \geq 1$. By (A.21), the variance of the

h-step ahead forecast errors is

$$\text{Var}[e_{t+h}|\mathcal{F}_t] = \sigma^2 \sum_{j=1}^{h} \left[1 + (\theta_1 + \phi_1) \left(\frac{1 - \phi_1^{h-j}}{1 - \phi_1} \right) \right]^2 \tag{A.23}$$

A.4.2 Conditional heteroscedastic white noise

In general, the conditional variances of the white noise need not be constant with time. Indeed, as described in Section A.2, the conditional variances could follow a GARCH process, although the unconditional variance remains constant. Since the GARCH process (A.5) for $\{\varepsilon_t\}$ is analogous to an ARMA process (A.3) for $\{\varepsilon_t^2\}$, we can obtain the optimal h-step ahead forecasts for the conditional variance of the white noise $\hat{\sigma}^2_{t+h|t}$ in a similar way as discussed in Section A.3. For simplicity, consider the GARCH(1,1) model in (A.5) where $r = s = 1$. The one-step ahead forecast is simply $\hat{\sigma}^2_{t+1|t} = \omega + \alpha_1 \varepsilon_t^2 + \beta_1 \sigma_t^2$. For forecast horizon $h > 1$, we may write $\varepsilon_t^2 = \sigma_t^2 \nu_t^2$ and so

$$\begin{aligned} \hat{\sigma}^2_{t+h|t} &= \text{E}[\sigma^2_{t+h}|\mathcal{F}_t] \\ &= \text{E}[\omega + \alpha_1 \varepsilon^2_{t+h-1} + \beta_1 \sigma^2_{t+h-1}|\mathcal{F}_t] \\ &= \text{E}[\omega + (\alpha_1 + \beta_1)\sigma^2_{t+h-1} + \alpha_1 \sigma^2_{t+h-1}(\nu^2_{t+h-1} - 1)|\mathcal{F}_t] \\ &= \omega + (\alpha_1 + \beta_1)\hat{\sigma}^2_{t+h-1|t} \end{aligned} \tag{A.24}$$

since $\text{E}[\nu^2_{t+h-1} - 1|\mathcal{F}_t] = 0$ for $h > 1$. This is the same as the recursive forecast formula (A.13) for the AR(1) process with $\omega = 0$ and $\phi_1 = \alpha_1 + \beta_1$. Applying (A.24) recursively, we have

$$\hat{\sigma}^2_{t+h|t} = \omega \left(\frac{1 - (\alpha_1 + \beta_1)^h}{1 - (\alpha_1 + \beta_1)} \right) + (\alpha_1 + \beta_1)^h \sigma_t^2 \tag{A.25}$$

which converges to the unconditional variance $\omega/[1 - (\alpha_1 + \beta_1)]$ as $h \to \infty$, provided that $\alpha_1 + \beta_1 < 1$.

Now consider $\{Y_t\}$ which follows an ARMA process with white noise following a GARCH process. To obtain the variance of forecast errors of Y_{t+h}, let us refer

to the expression in (A.18). Since the conditional variance of the white noise process is time-varying, we have

$$\begin{aligned} \mathrm{Var}[e_{t+h}|\mathcal{F}_t] &= E\left[\left(\sum_{j=0}^{h-1}\psi_j\varepsilon_{t+h-j}\right)^2\middle|\mathcal{F}_t\right] \\ &= \sum_{j=0}^{h-1}\psi_j^2\hat{\sigma}^2_{t+h-j|t} \end{aligned} \tag{A.26}$$

where $\mathrm{E}[\varepsilon^2_{t+h}|\mathcal{F}_t] = \hat{\sigma}^2_{t+h|t}$ and $\{\varepsilon_t\}$ are uncorrelated. This is analogous to (A.19). Similarly, if $\{Y_t\}$ follows an ARIMA process with white noise following a GARCH process, the variance for the forecast errors is analogous to (A.21) and is given by

$$\mathrm{Var}[e_{t+h}|\mathcal{F}_t] = \sum_{j=1}^{h}\Omega^2_{h-j}\hat{\sigma}^2_{t+j|t} \tag{A.27}$$

where $\Omega_k = \Sigma_{j=0}^{k}\psi_j$, and ψ_j are the MA coefficients of $W_t = Y_t - Y_{t-1}$ expressed as an MA(∞) process.

To illustrate, let us consider that $W_t = Y_t - Y_{t-1}$ follows the ARMA(0,1) process. Furthermore, the white noise follows the GARCH(1,1) process, and so the conditional variance of the white noise is described by (A.24). For this ARIMA(0,1,1)-GARCH(1,1) process, the variance of the forecast errors can be directly written down using (A.27). With the only non-zero coefficient $\psi_1 = \theta_1$, we have $\Omega_0 = 1$ and $\Omega_k = 1 + \theta_1$ for all $k > 1$, and so

$$\mathrm{Var}[e_{t+h}|\mathcal{F}_t] = \hat{\sigma}^2_{t+h|t} + (1+\theta_1)^2\sum_{j=1}^{h-1}\hat{\sigma}^2_{t+j|t} \tag{A.28}$$

where $\hat{\sigma}^2_{t+j|t}$ is obtained using (A.25). In general, for all cases where Y_t follows an ARIMA$(p,1,q)$-GARCH(r,s) process, the variance of forecast errors will be given by (A.27) with the substitution of the corresponding coefficients Ω^2_{h-j} and the j-step ahead forecasts of the conditional variances of white noise $\hat{\sigma}^2_{t+j|t}$.

Appendix B

Exponential Smoothing Methods

B.1 Model for conditional mean

In this section we describe how to use exponential smoothing methods to generate multi-step forecasts. Exponential smoothing methods are also called the exponentially weighted moving average (EWMA), since the one-step ahead forecast at time t is given by

$$\hat{y}_{t+1|t} = \alpha y_t + \alpha(1-\alpha)y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + \cdots \tag{B.1}$$

where α is a smoothing constant. If $|1-\alpha| < 1$, then the weights add up to unity and their magnitudes decrease exponentially with time in the past. Clearly, (B.1) can be written in a recursive form

$$\hat{y}_{t+1|t} = \alpha y_t + (1-\alpha)\hat{y}_{t|t-1} \tag{B.2}$$

We can then introduce a level index S_t, which is updated by the recursive formula

$$S_t = \alpha y_t + (1-\alpha)S_{t-1} \tag{B.3}$$

and the one-step ahead forecast is given by the value of the latest level index, i.e.

$$\hat{y}_{t+1|t} = S_t \tag{B.4}$$

The updating formula (B.3) and the forecast formula (B.4) constitute the simple exponential smoothing method. In general, we may also introduce indices for the trend and seasonality, and include those effects in the updating and forecast formulae. Exponential smoothing methods are simply recipes for generating point forecasts. It does not tell us anything about the uncertainty of a point forecast, not to mention about the whole density distribution of the forecast errors. So a question arises: is it possible to express an exponential smoothing method in terms of a stochastic model in a consistent way? By consistency we mean

1. The updating formulae are equivalent to the data generating process described by the stochastic model

2. The point forecast is optimal, i.e. it is the one-step ahead minimum mean square error (MMSE) forecast according to the stochastic model given an appropriate information set $\mathcal{F}_t$

This problem was first studied by Muth (1960), although his primary interest was not to produce density forecasts using exponential smoothing methods. Muth shows that the simple exponential smoothing method is optimal for the ARIMA(0,1,1) model. Other literature on the optimality of the exponential smoothing methods and their relations with the ARIMA models includes Cogger (1974), Ledolter & Box (1978) and Satchell & Timmermann (1994, 1995). In the form of the ARIMA(0,1,1) model, (B.3) and (B.4) can be expressed as

$$w_t = \varepsilon_t + (\alpha - 1)\varepsilon_{t-1}, \qquad \varepsilon_t \overset{i.i.d.}{\sim} D(0, \sigma^2) \tag{B.5}$$

where $w_t = y_t - y_{t-1}$, ε_t is the white noise with distribution D such that $\mathrm{E}[\varepsilon_t] = 0$ and $\mathrm{Var}[\varepsilon_t] = \sigma^2$, and α is the smoothing constant in (B.3). In order to be invertible, we require that $|\alpha - 1| < 1$, i.e. $0 < \alpha < 2$. Note that if $\alpha > 1$, the weights in the moving average are alternative in sign and this may not be intuitive. Most literature in exponential smoothing thus simply confines α to $(0, 1)$. However,

we allow $\alpha > 1$ here since we regard simple ES as an optimal method for the ARIMA(0,1,1) model. This is a linear model with additive errors. Following the taxonomy by Hyndman *et al.* (2008), we denote this exponential smoothing method by ETS(A, N, N). ETS is both an abbreviation for exponential smoothing as well as an acronym for error, trend and seasonality respectively. The A inside the bracket stands for additive errors, the first N stands for no trend, and the second N stands for no seasonality.

In some cases, the performance of simple exponential smoothing can be greatly improved by introducing an error correction term in the forecast formula (B.4) (Taylor, 2003). This error correction term accounts for the autocorrelations in the forecast errors by a correlation coefficient ϕ where $|\phi| < 1$. With the same updating formula (B.3), the one-step ahead forecast is now given by

$$\hat{y}_{t+1|t} = S_t + \phi(y_t - S_{t-1}) \tag{B.6}$$

It can be shown that the ARIMA(1,1,1) model is optimal for this simple exponential smoothing method with error correction. Expressed in terms of the ARIMA(1,1,1) model, we have

$$w_t = \phi w_{t-1} + \varepsilon_t + (\alpha - 1)\varepsilon_{t-1}, \qquad \varepsilon_t \overset{i.i.d.}{\sim} D(0, \sigma^2) \tag{B.7}$$

where $w_t = y_t - y_{t-1}$, $0 < \alpha < 2$, $|\phi| < 1$ and ε_t is the white noise as defined previously. We denote this method by ETS$(A, N, N|EC)$, where EC stands for error correction.

B.2 Model for conditional mean and variance

In the last section, we assume that the $\text{Var}[\varepsilon_t] = \sigma^2$ which is constant and does not vary with time. In general we may consider time-varying conditional variances. In such case, we have to describe the dynamics in both the mean and the variance, and we propose to apply exponential smoothing methods in a simultaneous way. In the simplest case, we have a level index S_t for the mean and a level index V_t

for the variance. Instead of updating S_t alone as in (2.7), we update both S_t and V_t according to the two updating formulae

$$\begin{aligned} S_t &= \alpha y_t + (1-\alpha) S_{t-1} \\ V_t &= \gamma (y_t - \hat{y}_{t|t-1})^2 + (1-\gamma) V_{t-1} \end{aligned} \tag{B.8}$$

where α, γ are smoothing constants. As discussed above, we take $0 < \alpha < 2$. On the other hand, since $V_t > 0$, we have $0 < \gamma < 1$. The one-step ahead forecasts are given by

$$\begin{aligned} \hat{y}_{t+1|t} &= S_t \\ \hat{\sigma}^2_{t+1|t} &= V_t \end{aligned} \tag{B.9}$$

From previous discussions, the exponential smoothing method for the mean is optimal for the ARIMA(0,1,1) model. Since the conditional variances of the white noise is time-varying, we write

$$w_t = \varepsilon_t + (\alpha - 1)\varepsilon_{t-1}, \qquad \varepsilon_t | \mathcal{F}_{t-1} \overset{i.i.d.}{\sim} D(0, \sigma_t^2) \tag{B.10}$$

where $w_t = y_t - y_{t-1}$. Note the difference between (B.5) and (B.10) for the description of the white noise.

On the other hand, the exponential smoothing method for the variance is optimal for the GARCH(1,1) model. This can be seen by applying the ARIMA(0,1,1) model for the conditional mean in (B.10). We can then express $\varepsilon_t = y_t - \hat{y}_{t|t-1}$ in the updating formula for V_t in (B.8). Combining both the models for the mean and the variance, we see that this exponential smoothing method in (B.9) and (B.10) is optimal for an ARIMA(0,1,1)-GARCH(1,1) model, which is expressed as

$$\begin{aligned} w_t &= \varepsilon_t + (\alpha - 1)\varepsilon_{t-1}, \qquad \varepsilon_t | \mathcal{F}_{t-1} \overset{i.i.d.}{\sim} D(0, \sigma_t^2) \\ \sigma_t^2 &= (1-\gamma)\sigma_{t-1}^2 + \gamma \varepsilon_{t-1}^2 \end{aligned} \tag{B.11}$$

where $w_t = y_t - y_{t-1}$. We denote this as the ETS(A, N, N)-(A, N, N) method, which means that we perform simple exponential smoothing for the mean and the variance simultaneously. Note that we have an integrated GARCH (IGARCH) model for the variance since the sum of the coefficients for the ARCH and GARCH terms is $(1 - \gamma) + \gamma = 1$. Thus the variance process is non-stationary. Shocks are persistent and the unconditional variance does not exist (Engle & Bollerslev, 1986).

As mentioned in Section B.1, it may be useful to add an error correction term in the forecast formula if the forecast errors are highly correlated. We can do this similarly for the exponential smoothing method in both mean and variance. The updating formulae in (B.8) will remain the same, but one or both of the forecasting formulae in (B.9) will have an additional term. For example, if we add an error correction term to both the forecast formulae for the mean and the variance, then (B.9) becomes

$$\begin{aligned} \hat{y}_{t+1|t} &= S_t + \phi_s(y_t - S_{t-1}) \\ \hat{\sigma}^2_{t+1|t} &= V_t + \phi_v \left[(y_t - \hat{y}_{t|t-1})^2 - V_{t-1}\right] \end{aligned} \tag{B.12}$$

where $|\phi_s| < 1$ and $|\phi_v| < 1$ are some coefficients accounting for the correlation of the respective forecast errors in the mean and the variance. Note that since the forecast formula for $\hat{y}_{t+1|t}$ includes an extra error correction term, the original updating formula for V_t in (B.8) is also affected, despite that the form of (B.8) remains invariant. Now it can be shown that the updating formulae (B.8) together with the forecast formulae (B.12) is optimal for the ARIMA(1,1,1)-GARCH(2,1) model, which is expressed as

$$\begin{aligned} w_t &= \phi_s w_{t-1} + \varepsilon_t + (\alpha - 1)\varepsilon_{t-1}, \qquad \varepsilon_t|\mathcal{F}_{t-1} \overset{i.i.d.}{\sim} D(0, \sigma_t^2) \\ \sigma_t^2 &= (1 - \gamma)\sigma^2_{t-1} + (\gamma + \phi_v)\varepsilon^2_{t-1} - \phi_v \varepsilon^2_{t-2} \end{aligned} \tag{B.13}$$

where $w_t = y_t - y_{t-1}$. We denote this as the ETS$(A, N, N|EC)$-$(A, N, N|EC)$ method, which means that we perform simple exponential smoothing with error correction term for the mean and the variance simultaneously. In this case, we

also have an IGARCH model for the variance since the sum of the coefficients for the ARCH and GARCH terms is $(1-\gamma)+(\gamma+\phi_v)-\phi_v = 1$. Thus the variance process is again non-stationary.

In general, we may construct different combinations of exponential smoothing methods in both the mean and the variance, and find a corresponding ARIMA-GARCH model in which the exponential smoothing method is optimal. For instance, we may construct the ETS$(A,N,N|EC)$-(A,N,N) method, which applies simple exponential smoothing with error correction term for the mean, and simple exponential smoothing for the variance. It can be shown that this exponential smoothing method is optimal for the ARIMA(1,1,1)-GARCH(1,1) model.

B.3 Optimal forecast and forecast variance

B.3.1 Conditional homoscedastic white noise

According to the ARIMA models corresponding to the exponential smoothing methods as described in Section B.1, we can directly write down the optimal h-step ahead forecasts $\hat{y}_{t+h|t}$ and variance of forecast errors of the exponential smoothing methods using the general formulae in sections A.3 and A.4. For instance, the ETS(A,N,N) method is optimal for the ARIMA(0,1,1) model and can be expressed as (B.5). According to (A.15) with $\theta_1 = \alpha - 1$, the optimal h-step ahead forecast is

$$\hat{y}_{t+h|t} = y_t + (\alpha-1)\varepsilon_t \tag{B.14}$$

The variance of forecast errors is given by (A.22), which is

$$\mathrm{Var}[e_{t+h}|\mathcal{F}_t] = \sigma^2[1+(h-1)\alpha^2] \tag{B.15}$$

As another example, the ETS$(A,N,N|EC)$ method is optimal for the ARIMA(1,1,1) model and can be expressed as (B.7). According to (A.17) with $\phi_1 = \phi$ and

$\theta_1 = \alpha - 1$, the optimal h-step ahead forecast is

$$\hat{y}_{t+h|t} = \phi^h y_t + \left(\frac{1-\phi^h}{1-\phi}\right)[y_t - \phi y_{t-1} + (\alpha - 1)\varepsilon_t] \tag{B.16}$$

Note that by (2.8), we can express the white noise as $\varepsilon_t = y_t - [S_{t-1} - \phi(y_{t-1} - S_{t-2})]$. Using the updating formula (2.7) and after some algebra, we can express (B.16) as

$$\hat{y}_{t+h|t} = S_t + \frac{\alpha\phi(1-\phi^{h-1})}{1-\phi}(y_t - S_{t-1}) + \phi^h(y_t - S_{t-1}) \tag{B.17}$$

We prefer to express the optimal h-step ahead forecast as (B.17) instead of (B.16) because we want to eliminate ε_t and compute $\hat{y}_{t+h|t}$ directly from the observation y_t and the level index S_t. On the other hand, we could also obtain (B.17) by computing $\hat{y}_{t+h|t}$ recursively using the forecast formula (2.8) for the ETS$(A, N, N|EC)$ method. The variance of the h-step ahead forecast errors is given by (A.23), which is

$$\begin{aligned} \text{Var}[e_{t+h}|\mathcal{F}_t] &= \sigma^2 \sum_{j=1}^{h} \left[1 + (\alpha + \phi - 1)\left(\frac{1-\phi^{h-j}}{1-\phi}\right)\right]^2 \\ &= \sigma^2 \sum_{j=1}^{h} \left[\phi^{h-j} + \frac{\alpha(1-\phi^{h-j})}{1-\phi}\right]^2 \end{aligned} \tag{B.18}$$

B.3.2 Conditional heteroscedastic white noise

Again, according to the GARCH models corresponding to the exponential smoothing methods as described in Section B.2, we can directly write down the optimal h-step ahead forecasts $\hat{y}_{t+h|t}$ and variance of forecast errors using the general formulae in sections A.3 and A.4. Consider the ETS(A, N, N)-(A, N, N) method which is optimal for the ARIMA(0,1,1)-GARCH(1,1) model. Using the result of optimal h-step ahead forecast for the ARIMA(0,1,1) model as given in (B.14), we have $\hat{y}_{t+h|t} = y_t + (\alpha - 1)\varepsilon_t$. For the variance of forecast errors, we apply (A.25) and (A.28) with $\theta_1 = \alpha - 1$, $\omega = 0, \alpha_1 = 1 - \gamma$ and $\beta_1 = \gamma$. Since $\hat{\sigma}^2_{t+h|t} = \sigma^2_t$ for

all h, we have

$$\text{Var}[e_{t+h}|\mathcal{F}_t] = \sigma_t^2[1 + (h-1)\alpha^2] \tag{B.19}$$

Note that (B.19) is analogous to (B.15), except that the conditional variance of the white noise σ_t^2 is now time-dependent. For other cases like the ETS$(A, N, N|EC)$-$(A, N, N|EC)$ method, which is optimal for the ARIMA(1,1,1)-GARCH(2,1) model, the optimal h-step ahead forecast and variance of forecast errors can be written down similarly.

Appendix C

Positive Definite Covariance Functions

Consider a spatiotemporal covariance function $C(\mathbf{s}_i, t_i; \mathbf{s}_j, t_j)$, where $\mathbf{s}$ is the spatial coordinate and t is the temporal coordinate. Further assume that it is stationary, so that we have

$$C(\mathbf{s}_i, t_i; \mathbf{s}_j, t_j) = C(\mathbf{h}, u) \tag{C.1}$$

where $\mathbf{h} = \mathbf{s}_i - \mathbf{s}_j$ is the spatial lag and $u = t_i - t_j$ is the temporal lag. $C(\mathbf{h}, u)$ is positive definite if for any positive integer N and any $a_1, a_2, \ldots, a_N \in \mathbb{R}$, $(\mathbf{s}_1, t_1), \ldots, (\mathbf{s}_N, t_N) \in \mathbb{R}^d \times \mathbb{R}$, we have

$$\sum_{i=1}^{N} \sum_{j=1}^{N} C(\mathbf{s}_i - \mathbf{s}_j; t_i - t_j) \geq 0 \tag{C.2}$$

Now, suppose that $C_0(\mathbf{h}, u)$ is a stationary covariance function, and consider transforming its temporal coordinates from u to $\tilde{u}$ such that

$$\tilde{u} = u + \mathbf{h} \cdot \mathbf{v} \tag{C.3}$$

where $\mathbf{v} \in \mathbb{R}^d$ is some constant vector. Then, as in Ma (2003), it can be shown that

$$\begin{aligned} C(\mathbf{h}, u) &= C_0(\tilde{u}) \\ &= C_0(\mathbf{h}, u + \mathbf{h} \cdot \mathbf{v}) \end{aligned} \tag{C.4}$$

is also a valid covariance function. It is easy to check that $C(\mathbf{h}, u)$ is positive definite by showing that

$$\begin{aligned} \sum_{i=1}^{N}\sum_{j=1}^{N} C(\mathbf{s}_i - \mathbf{s}_j; t_i - t_j) &= \sum_{i=1}^{N}\sum_{j=1}^{N} C_0(\mathbf{s}_i - \mathbf{s}_j; t_i - t_j + \mathbf{h} \cdot \mathbf{v}) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} C_0(\mathbf{s}_i - \mathbf{s}_j; (t_i + \mathbf{s}_i \cdot \mathbf{v}) - (t_j + \mathbf{s}_j \cdot \mathbf{v})) \\ &\geq 0 \end{aligned} \tag{C.5}$$

Appendix D

Techniques of Forecast Evaluations

Here we review some important ideas in evaluating forecasts (McSharry *et al.*, 2009). This will provide a necessary background for us to analyze the models and evaluate the results when we apply our approaches to the Irish data in Chapter 5 and the Danish data in Chapter 6.

One of the important issues in forecasting problems is to evaluate the forecast performances of different models. Depending on the particular application of the forecasts, interests can range from point forecasts, interval forecasts, quantile forecasts, or most generally, probability density forecasts. In decision theory, a forecast is judged by a corresponding loss function $L(Y, \hat{Y})$ (or sometimes called the cost function $C(Y, \hat{Y})$) that is defined by the user to quantify the loss. To be consistent, $\hat{Y}$ need to be the optimal forecast under this loss function $L(Y, \hat{Y})$ (Appendix A.3). The loss function is itself a random variable that depends on the outcomes of Y. To evaluate the forecast performance of a model, we obtain predictions $\hat{Y}$ from the model and calculate statistics of the loss function, in particular the mean values. As a result, forecast performances are ranked by the mean scores in this case.

Apart from using scores to evaluate forecasts, another approach is to rely on

visualizations of appropriate plots, sometimes called diagnostic tools. This technique is often applied to the evaluation of density forecasts such as the calibration and sharpness of a density, while it is difficult to reflect all information in a single score. The most common example is the probability integral transform (PIT) histograms (Diebold *et al.*, 1998). Other examples include the reliability diagrams (Pinson *et al.*, 2010). In the following, we describe some evaluation methods that will be applied later in the analyses of Irish and Danish wind data.

D.1 Scoring rules

To evaluate the forecasts generated by various models, we consider three different aspects of performances, namely the performances in 1). mean forecasts, 2). quantile forecasts, and 3). density forecasts. For the performances in mean and quantile forecasts, we use the corresponding appropriate loss function to calculate the errors, where the mean and the quantiles are obtained as optimal forecasts under that loss function (Gneiting, 2010). For probability forecasts, we consider the continuous ranked probability scores. Details of the evaluation of different scores are provided as follows:

Mean forecasts - Root mean square error (RMSE): It is shown that mean forecasts are the optimal forecasts under the squared loss function

$$L(Y, \hat{Y}) = (Y - \hat{Y})^2 \tag{D.1}$$

As a result, from the density forecasts generated from the 1000 samples, we calculate the mean value as the mean forecasts $\hat{Y}$ and measure the performance using root mean square errors (RMSE), i.e. we calculate

$$RMSE = \sqrt{\mathrm{E}[(Y - \hat{Y})^2]} \tag{D.2}$$

where the expectation is taken over all forecasts in the testing data.

Quantile forecasts - Mean quantile error (MQE): For quantile forecasts, one needs to consider asymmetric loss functions. For a fixed quantile value $\alpha \in [0, 1]$, let us consider the check loss function

$$L_\alpha(Y, \hat{Y}) = 2[\alpha - \mathbf{1}(Y < \hat{Y})] \cdot (Y - \hat{Y}) \tag{D.3}$$

where $\mathbf{1}(\cdot)$ is the indicator function which is equal to one when the argument is positive. This is an appropriate loss function as the optimal forecasts under this loss is exactly the $\alpha-$quantile. Note that for $\alpha = 0.5$, we have

$$L_{0.5}(Y, \hat{Y}) = |Y - \hat{Y}| \tag{D.4}$$

which reduces to the absolute error. As a result, we evaluate median forecasts by firstly calculating the median from the samples, take it as $\hat{Y}$ and calculate the mean absolute error (MAE), i.e.

$$MAE = \mathrm{E}|Y - \hat{Y}| \tag{D.5}$$

For quantiles at other values apart from $\alpha = 0.5$, we use the same approach and calculate the mean quantile error[1] (MQE), i.e.

$$MQE = \mathrm{E}\left[2[\alpha - \mathbf{1}(Y < \hat{Y})] \cdot (Y - \hat{Y})\right] \tag{D.6}$$

where $\hat{Y}$ is the $\alpha-$quantile forecast.

Probability forecasts - Continuous ranked probability score (CRPS): For probability density forecasts, we use the continuous ranked probability score (CRPS) (Gneiting & Raftery, 2007) as the loss function. CRPS is negatively oriented so that a lower score corresponds to a better probabilistic forecast. Gneiting & Raftery (2007) discussed the properties of CRPS

[1] Mean quantile error is also called mean quantile score, and both names are equivalent.

extensively, showing that it is a strictly proper score and a lower score always indicate a better probabilistic forecast. CRPS has become one of the popular tools for probabilistic forecast evaluations, especially for ensemble forecasts in meteorology[1]. For each h-step ahead probabilistic forecast $f_{t+h|t}$, let $F_{t+h|t}$ be the corresponding cumulative distribution function. The CRPS is computed as

$$\begin{aligned} CRPS &= \int_{-\infty}^{\infty} [F_{t+h|t}(y) - \mathbf{1}(y - y_{t+h})]^2 dy \\ &= \int_{0}^{1} [F_{t+h|t}(y) - \mathbf{1}(y - y_{t+h})]^2 dy \end{aligned} \tag{D.7}$$

where $\mathbf{1}(\cdot)$ is the indicator function which is equal to one when the argument is positive, F is the forecast distribution for y and y_0 is the observed value.

In fact, there is an interesting relation between the MQE and the mean CRPS. Laio & Tamea (2007) show that the CRPS in (D.7) can be expressed as an integral of the quantile scores (D.3) at all quantiles. By substituting the optimal quantile forecast $\hat{Y}$ into the equation for CRPS and expressing $F^{-1}(\alpha) = \hat{Y}$, we have

$$CRPS = \int_{0}^{1} L_\alpha(Y, \hat{Y}) d\alpha \tag{D.8}$$

One could even put different weights on the quantile scores and obtain a qunatile-weighted CRPS, which could be useful when one is particularly interested in a certain range of quantiles (Gneiting & Ranjan, 2010).

[1] Another common score for probability forecasts is the negative log likelihood (NLL) scores (Taylor *et al.*, 2009), but we will not apply it here. CRPS is more robust than the NLL scores, as the latter is always severely affected by a few extreme outliers (Gneiting *et al.*, 2005). One may need to calculate the trimmed mean of the NLL scores in order to resolve this problem (Weigend & Shi, 2000). Also, CRPS assesses both the calibration and the sharpness of the probability forecasts, while the NLL score assesses sharpness only (Gneiting *et al.*, 2007a).

D.2 Diagnostic diagrams

Another common diagnostic tool for the evaluation of density forecasts is the probability integral transform (PIT) histogram Clements & Smith (2000); Diebold *et al.* (1998). In this case, we calculate the PIT values $z(y_{t+1})$ for the one-step ahead probability forecasts $f_{t+1|t}$, which are given by

$$z(y_{t+1}) = \int_0^{y_{t+1}} f_{t+1|t}(y)dy \tag{D.9}$$

Diebold *et al.* (1998) show that the series of PIT values $z(y_{t+1})$ are i.i.d uniform if $f_{t+1|t}$ coincides with the true underlying density from which y_{t+1} is generated. As a result, a diagnostic diagram could be obtained by plotting the distribution of the PIT values and compare the histogram with that of a uniform distribution $U[0, 1]$. Different characteristics of the PIT histograms, such as being U-shaped or skewed, could then infer different imperfections in the forecast densities.

Appendix E

Remarks on Alternative Models

E.1 Diurnal seasonality and seasonal exponential smoothing

In Section 1.3.2, we have seen that aggregated wind power exhibits diurnal seasonalities as shown in Figure 1.8. To investigate weather we could improve the forecasts by accounting for the diurnal cycle, we deseasonalize the data by subtracting this diurnal seasonality. In the case of Irish wind power, we forecast the deseasonalized data by selecting the best ARIMA$(p, 1, q)$-GARCH(r, s) model based on minimizing the BIC. We select an ARIMA(2,1,3)-GARCH(1,1) model and perform an out-of-sample forecasts. Forecasts of wind power are then obtained by adding the diurnal cycle to the deseasonalized data. By comparing with the ARIMA(4,1,3)-GARCH(1,1) approach (with logit transformation) as discussed in Section 5.2, we obtain the results as shown in Figure E.1. In the figure, we compare point forecasts and consider the percentage differences in RMSE between the diurnal seasonality model and the ARIMA(4,1,3)-GARCH(1,1) benchmark, where positive differences mean that the diurnal seasonality model has a larger forecast error.

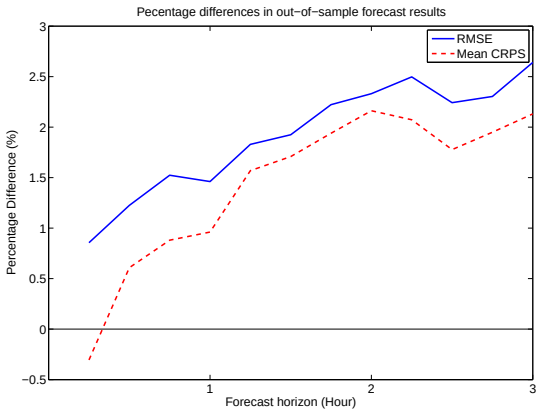


Figure E.1: This figure plots the relative performances of the simple diurnal seasonality model with the ARIMA(4,1,3)=GARCH(1,1) approach (with logit transformation) as discussed in Section 5.2. Positive percentage differences mean that the diurnal seasonality model has a larger forecast error.

Apart from simply analyzing deseasonalized data, we also try to apply a multiplicative seasonal exponential smoothing approach, which is also known as the Holt-Winter's method (Holt, 1957; Taylor, 2003). In this approach, we consider a seasonal period of length $L = 96$, i.e. one day, as we observe some evidences of diurnal seasonalities in Figure 1.8. We smooth through the mean level S_t and the seasonal cycle M_t by two smoothing parameters $0 < \alpha, \gamma < 1$, and the updating equations can be written as

$$\begin{aligned} S_t &= \alpha\left(\frac{y_t}{M_{t-L}}\right) + (1-\alpha)S_{t-1} \\ M_t &= \gamma\left(\frac{y_t}{S_t}\right) + (1-\alpha)M_{t-L} \end{aligned} \tag{E.1}$$

where y_t is the wind power data. The h-step ahead forecasting equation is

$$y_{t+h|t} = S_t * M_{t+h-L} \tag{E.2}$$

We initialize the updating equations by setting $S_j = \bar{y} = \sum_{j=1}^{L} y_j/L$ and $M_j = y_j/\bar{y}$ for $j = 1, \ldots, L$[1], and the updating equations in (E.1) are applied for $t > L$.

[1]Note that by definition, we have $\sum_{j=1}^{L} M_j = L$. However, smoothed values of M_t may not

As wind power forecasts $y_{t+h|t}$ must lie within $[0, 1]$, we apply the same methodology as discussed in Chapter 2 and consider $y_{t+h|t}$ as a proxy for the location parameter ℓ of a truncated normal distribution $f(y|\ell, s^2)$. We then estimate the smoothing parameters α, γ and the scale parameter s^2 by maximizing the truncated normal likelihood, as discussed on Section 2.2.3. Using the same training and testing data for aggregated Irish wind power as discussed in Chapter 5, we obtain forecast results that are worse than all models including the persistence benchmark in Section 5.2. This is probably due to the relative instability in the multiplicative seasonal components, and also short term seasonalities are more difficult to model as the signal-to-noise ratio could be small. Results comparing the point forecasts between the seasonal exponential smoothing method to the ARIMA(4,1,3)-GARCH(1,1) benchmark are shown in Figure E.2. Clearly, the RMSE of the point forecasts obtained by the seasonal exponential smoothing method is much larger. We only consider $1-4$-step ahead forecasts here because the performances of the seasonal exponential smoothing method deteriorate rapidly with forecast horizons.

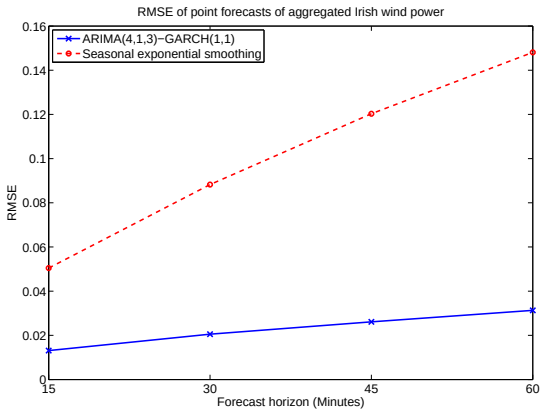


Figure E.2: This figure shows the RMSE of the point forecasts obtained by the seasonal exponential smoothing method and the ARIMA(4,1,3)-GARCH(1,1) benchmark as discussed in Section 5.2. Clearly, the RMSE of the point forecasts obtained by the seasonal exponential smoothing method is much larger. We only consider $1-4$-step ahead forecasts here because the performances of the seasonal exponential smoothing method deteriorate rapidly with forecast horizons.

satisfy this and we must normalize $M_{t-L+1}, \ldots, M_t$ so that $\sum_{j=1}^{L} M_{t-L+j} = L$.

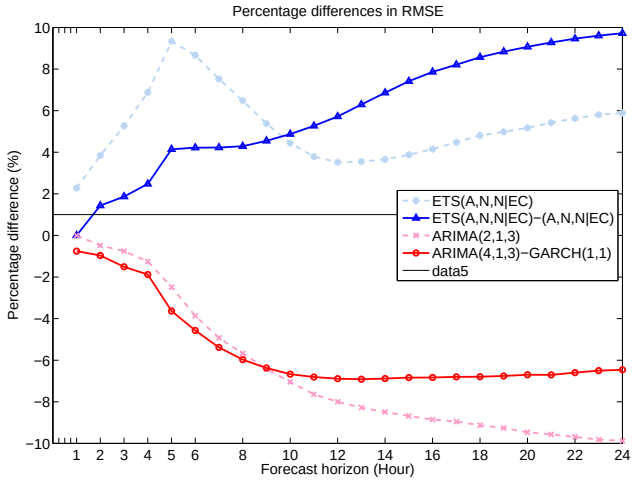
E.2 Exponential smoothing with logit transformation

In Section 2.2 we introduce an approach for forecasting aggregated wind power by exponential smoothing methods followed by the use of truncated normal distributions. One may wonder whether applying exponential smoothing methods on logit transformed data (similar to the descriptions in Section 2.1 would produce better results. By considering the aggregated wind power in Ireland, we demonstrate that the approach described in Section 2.2 is better. Exponential smoothing is very robust and it is better to apply it directly to the data without transformations.

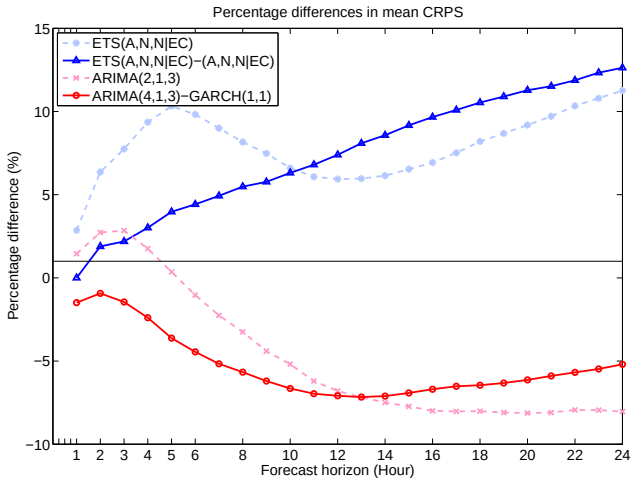
With the same training and testing data for the Irish wind power as described in Chapter 5, we investigate the performances of the following four models which appear in the analysis in Section 5.2:

- The ARIMA(2,1,3) model
- The ARIMA(4,1,3)-GARCH(1,1) model
- The ETS$(A, N, N|EC)$ method
- The ETS$(A, N, N|EC)$-$(A, N, N|EC)$ method

Note that we do not include the brackets [LT] or [TN] here, which have been used to denote the approach of logit transformation in Section 2.1 and truncated normal distribution in Section 2.2 respectively. Instead, we compare the two approaches with the above four models. In other words, for each of the above four models, we apply the corresponding [LT] or [TN] version to generate forecasts. To clarify the results, we show the percentage differences between model forecasts using the [LT] version and the [TN] version. Percentage differences in RMSE and mean CRPS for the testing data in the Irish wind power are shown in Figures E.3(a) and E.3(b) respectively. Results smaller (larger) than zero mean that the [LT] version is better (poorer) than the [TN] version. Results show that when one applies ARIMA-GARCH models, logit transformation [LT] is a better approach. On the other hand, when one applies exponential smoothing methods, truncated normal distribution [TN] is a better approach.



(a)Percentage differences in RMSE for the two approaches.



(b)Percentage differences in mean CRPS for the two approaches.

Figure E.3: Percentage differences in (a) RMSE, and (b) mean CRPS for the [LT] and [TN] approaches. We consider the same training and testing set in the Irish data in Chapter 5. The brackets [LT] and [TN] have been used to denote the approach of logit transformation in Section 2.1 and truncated normal distribution in Section 2.2 respectively. Results smaller (larger) than zero mean that the [LT] version is better (poorer) than the [TN] version. Results show that when one applies ARIMA-GARCH models, logit transformation [LT] is a better approach. On the other hand, when one applies exponential smoothing methods, truncated normal distribution [TN] is a better approach.

E.3 Exponential smoothing with sliding window

In the thesis, we mainly consider a single parameter estimation procedure using the training data, and then apply the fitted model for forecasting wind power in the testing data. One exception is in the spatiotemporal forecast of the Danish wind power, where we use a sliding window of one year as a training set.

In this section, we investigate briefly the use of a sliding window model (i.e. quasi-stationary) and see how the forecast results could be improved. In general, one expects that using a sliding window which is too short will induce too much uncertainties in the parameter estimation, thus leading to poor forecasts. On the other hand, one also expects that using a sliding window which is too long may not be optimal, because changing regimes could not be captured from a change of parameters. If one does not optimize the length of the sliding window, it is recommended that choosing a longer sliding window could be safer because in general the forecast performances deteriorate relatively slowly if one increase the length of sliding window beyond the optimal.

We did a simple analysis on estimating the parameters of the ETS$(A, N, N|EC)$-$(A, N, N|EC)$ method using the method of sliding window. We consider the normalized aggregated wind power in Ireland. As in the thesis, we consider 5504 observations in the testing data, i.e about 2 months. In the exponential smoothing method with sliding window, we consider sliding windows of lengths from one day up to 30 days for parameter estimation. Then we generate one-step-ahead out-of-sample forecasts, where the window of forecasts are of the same length as the sliding window used in parameter estimation. Evaluation of out-of-sample forecasts are based on the mean CRPS, where the forecast density is of the form of truncated normal distribution as discussed in 2.2.2.1. Results are shown in Figure E.4. Clearly, a short sliding window smaller than 5 days could give very poor forecasts. The optimal length of sliding window in this case is about 10 days, i.e. 960 observations. Increasing the length of the sliding window beyond 10 days deteriorate forecast performances, but at a relatively slow rate. Indeed, we find that using all training data (i.e. 11008 observations) for parameter estimation gives a mean CRPS of 0.0092, as shown by the dashed line in the figure.

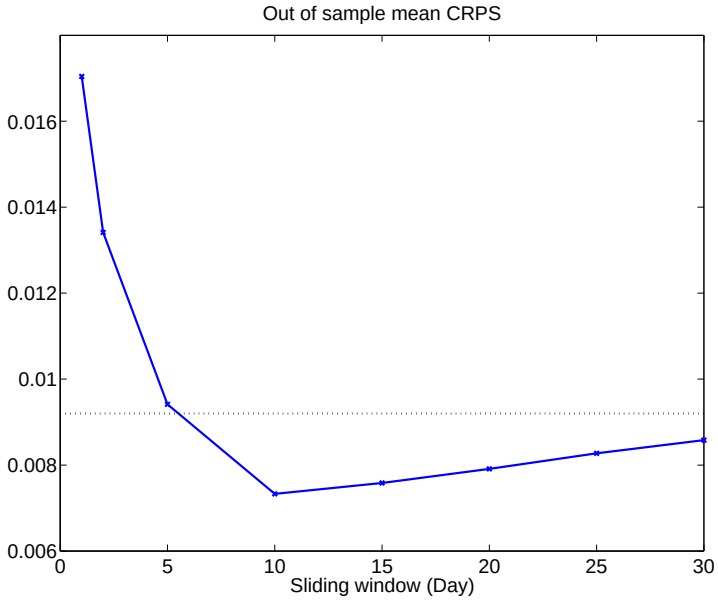


Figure E.4: This figure shows the mean CRPS of one-step-ahead out-of-sample forecasts of the normalized aggregated wind power in Ireland, with parameter estimation using different lengths of sliding window. Clearly, a short sliding window smaller than 5 days could give very poor forecasts. The optimal length of sliding window in this case is about 10 days, i.e. 960 observations. Increasing the length of the sliding window beyond 10 days deteriorate forecast performances, but at a relatively slow rate. Indeed, we find that using all training data (i.e. 11008 observations) for parameter estimation gives a mean CRPS of 0.0092, as shown by the dashed line in the figure.

E.4 Double kernel density benchmark

In Section 5.2 we study the aggregated wind power forecasts in Ireland. We introduce several benchmark models, with one of them being the EWMA conditional probability forecast. We have considered other similar types of benchmarks, for example, the double kernel density benchmark as described here. We show that the EWMA conditional probability forecast in Section 5.2 serves as a good enough benchmark.

The double kernel density benchmark has a similar idea as the work in Taylor (2008); Yu & Jones (1998). We apply exponential smoothing in time and consider a truncated normal kernel in wind power, and the density of wind power is obtained by

$$f(y_t) = \sum_{k=1}^{\infty} w_k K_{h,k}(y_t) \tag{E.3}$$

where $w_k = (1-\lambda)\lambda^{k-1}$ is the weight in exponential smoothing in time with $0 < \lambda < 1$ so that $\sum_{k=1}^{\infty} w_k = 1$ and thus $f(y_t)$ is a probability density. To account for the fact that wind power is a bounded variable, the kernel $K_{h,k}$ is a truncated normal density centered at y_{t-k} with bandwidth h (Taylor *et al.*, 2009), and with truncations at zero and one, i.e.

$$K_{h,k}(y_t) = \frac{1}{h} \frac{\varphi\left(\frac{y_t - y_{t-k}}{h}\right)}{\Phi\left(\frac{1-y_{t-k}}{h}\right) - \Phi\left(\frac{-y_{t-k}}{h}\right)} \tag{E.4}$$

where φ and Φ are the density and distribution of the standard normal respectively.

As (E.3) is essentially an exponential smoothing of the kernels with smoothing parameter λ, it could be written as an updating and forecasting algorithm with smoothed density $S(y_t)$, i.e.

$$\begin{aligned} S(y_t) &= \lambda K_{h,0}(y_t) + (1-\lambda) S(y_{t-1}) \\ \hat{f}(y_{t+1}) &= S(y_t) \end{aligned} \tag{E.5}$$

with $S(y_1) = K_{h,0}(y_1)$ and $K_{h,0}(y_t)$ is a truncated normal density centered at y_t with bandwidth h. As we aim at generating superior probability forecasts, we estimate the smoothing parameter λ and the bandwidth h by minimizing the in-sample mean CRPS. Using the same learning set in the Irish data as described in Chapter 5, we obtain $\hat{\lambda} = 0.999975$ and $\hat{h} = 8.8 \times 10^{-3}$. With such a large value of $\hat{\lambda}$, it means that essentially the best 15-minute ahead probability forecast $\hat{f}(y_{t+1})$ is obtained by a truncated normal density with location parameter being the last observation y_t and scale parameter being h. As a result, for short horizons, this double kernel density benchmark gives very similar forecasts as the persistence benchmark described on Page 119. However, it is significantly worse than the persistence benchmark at long horizons, as shown in Figure E.5. We see that it is good enough to include the persistence forecast and the EWMA conditional probability forecast as the two benchmarks in our analysis, as the double kernel density forecast is similar to the persistence forecast for short horizons, while its performance is relatively poor for horizons beyond 8 hours. Figure E.5 could be compared directly with Figure 5.4.

In addition, recall that for the persistence benchmark on Page 119, we estimate the realized variance with the latest $N_{r.v.}$ observations and simply choose $N_{r.v.} = 48$, i.e. 12 hours of past data. One could optimize the value of $N_{r.v.}$, and we did that by minimizing the in-sample mean CRPS. Result is shown in Figure E.6, and we obtain the optimal value of $N_{r.v.} = 12$. In Figure E.5, we include the performance of this benchmark, and we see that using $N_{r.v.} = 48$ is in fact better for out-of-sample forecasts in the case of Irish data.

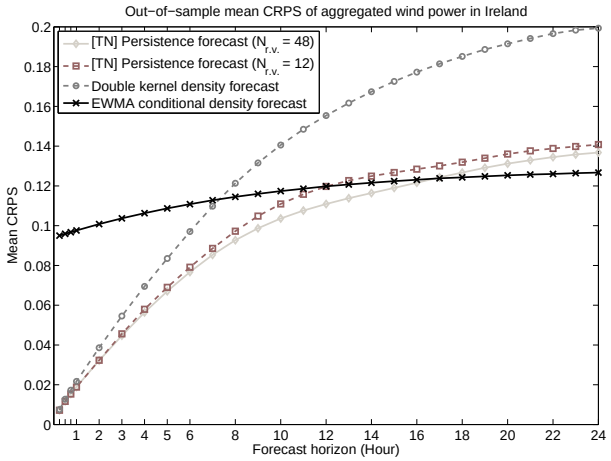


Figure E.5: This figure plots the mean CRPS of the probability forecasts for aggregated wind power in Ireland, with training and testing sets as described in Chapter 5. The persistence forecast and the EWMA conditional density forecast are described in detail in Section 5.2, with the former denoted by [TN] to remind us that it assumes a truncated normal distribution. We see that it is good enough to include the persistence forecast and the EWMA conditional probability forecast as the two benchmarks in our analysis, as the double kernel density forecast is similar to the persistence forecast for short horizons, while its performance is relatively poor for horizons beyond 8 hours. This figure could be compared directly with Figure 5.4.

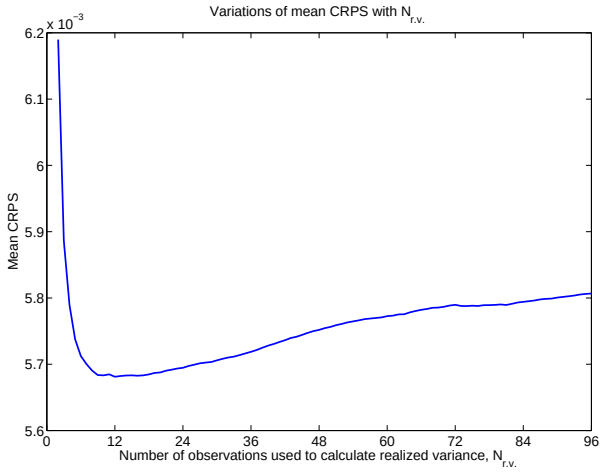


Figure E.6: This figure shows the variation of in-sample mean CRPS with $N_{r.v.}$, which is the number of past observations used in the estimation of realized variance $\hat{s}^2$ used in (5.1) for the persistence benchmark in aggregated wind power forecasts. In this case, we find that the optimal value is $N_{r.v.} = 12$.

E.5 ARIMA$(p,1,q)$-GARCH(r,s) benchmark

In Section 5.3.1, we introduce the univariate ARIMA(1,1,1)-GARCH(1,1) benchmark model for individual wind power forecasts. Such benchmark does not require any model selection. In this appendix, we compare the forecast results of such a robust ARIMA(1,1,1)-GARCH(1,1) benchmark with another similar version, where we select the best ARIMA$(p,1,q)$-GARCH(r,s) model for each individual wind power by minimizing the Bayesian Information Criteria (BIC). We show that simplicity and robustness are always the key criteria in generating superior out-of-sample forecasts in large data sets, and in this case, selecting the best ARIMA$(p,1,q)$-GARCH(r,s) model for each individual wind power could lead to over-fitting which gives poorer forecasts than the simple ARIMA(1,1,1)-GARCH(1,1) benchmark.

We test the results using the same Irish data with 64 individual wind power, and consider the same learning and testing data as mentioned in Chapter 5. To clarify the comparisons, we calculate the percentage differences in RMSE and mean CRPS for the out-of-sample aggregated wind power forecasts[1] between the ARIMA$(p,1,q)$-GARCH(r,s) model and the ARIMA(1,1,1)-GARCH(1,1) benchmark. Positive percentage differences mean that the ARIMA$(p,1,q)$-GARCH(r,s) model has a larger forecast error, and vice versa. Results are shown in Figure E.7, which clearly shows that the ARIMA$(p,1,q)$-GARCH(r,s) model is actually performing worst for out-of-sample forecasts, although it is likely that it has a better fit for in-sample data.

Other than selecting an ARIMA$(p,1,q)$-GARCH(r,s) model for each individual wind power, we could also select one best ARIMA$(p,1,q)$-GARCH(r,s) model jointly for all individual wind power by minimizing the total BIC. For the same Irish data as described above, the best model for all individual wind power turns out to be the ARIMA(3,1,1)-GARCH(1,1) model. However, results show that the out-of-sample forecast performances using the ARIMA(3,1,1)-GARCH(1,1) model is indeed worst than that of the ARIMA(1,1,1)-GARCH(1,1) benchmark.

[1]The aggregated forecasts are obtained by summing the forecasts for each individual wind power that are generated by the corresponding ARIMA$(p,1,q)$-GARCH(r,s) model with p,q,r,s being selected by minimizing the BIC.

A plot of the percentage differences in RMSE and mean CRPS for the out-of-sample aggregated wind power forecasts between the ARIMA(3,1,1)-GARCH(1,1) model and our ARIMA(1,1,1)-GARCH(1,1) benchmark is shown in Figure E.8.

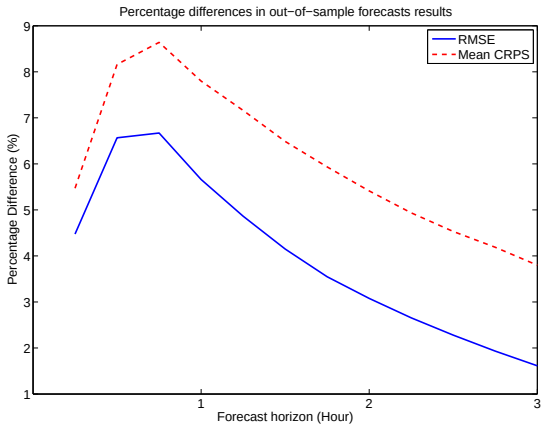


Figure E.7: This figure plots the relative performances of selecting the best ARIMA$(p, 1, q)$-GARCH(r, s) model for each individual wind power, as compared with the robust ARIMA(1,1,1)-GARCH(1,1) benchmark. We test the results using the same Irish data in Chapter 5. Positive percentage differences mean that the ARIMA$(p, 1, q)$-GARCH(r, s) model has a larger forecast error.

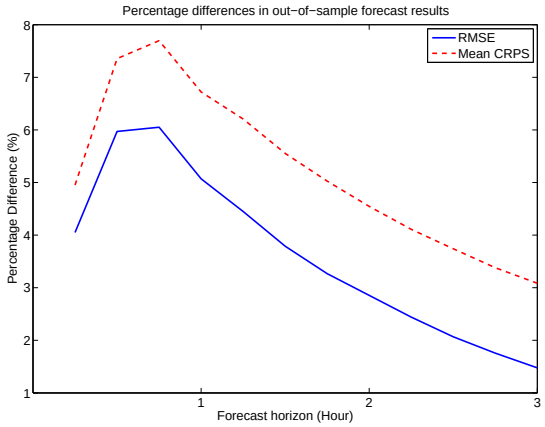


Figure E.8: This figure plots the relative performances of selecting one best ARIMA$(p, 1, q)$-GARCH(r, s) model jointly for all individual wind power by minimizing the total BIC, as compared with the robust ARIMA(1,1,1)-GARCH(1,1) benchmark. We test the results using the same Irish data in Chapter 5, and we select the ARIMA(3,1,1)-GARCH(1,1) model based on the BIC. Positive percentage differences mean that the ARIMA$(p, 1, q)$-GARCH(r, s) model has a larger forecast error.

Appendix F

Computational Remarks

F.1 Computational time: parameter estimation

In the two-stage model in Chapter 4, we need to estimate the scale parameter ρ_ε in the exponential correlation model for the latent Gaussian process $W(\mathbf{s}, t)$. This is done by Gibbs sampling. We firstly calculate the likelihood for several different values of ρ_ε and choose a best initial value for the optimization. Depending on the number of observations in the training data and the number of the wind farms, the computational time to obtain an optimal value of ρ_ε varies from 6-12 hours for our Irish and Danish data. This is being run on a typical personal computer with Intel Core 2 CPU each at 2.67 GHz.

F.2 Computational time: samplings

In the two-stage model in Chapter 4, we need to draw samples from the two latent Gaussian processes $W(\mathbf{s}, t)$ and $Z(\mathbf{s}, t)$. To obtain 100 samples for 5504 observations at 64 wind farms (i.e. each sample has a dimension of 5504×64), the computation time varies from 40 minutes for one-step ahead forecasts to 2 hours for 12-step ahead forecasts on a typical personal computer with Intel Core 2 CPU each at 2.67 GHz.

References

Abrahamsen, P. (1997). A review of Gaussian random fields and correlation functions, second edition. Technical Report 917, Norwegian Computing Centre. 83

Banerjee, S., Bradley & Gelfand, A.E. (2003). *Hierarchical Modeling and Analysis for Spatial Data (Monographs on Statistics and Applied Probability)*. Chapman & Hall/CRC. 6

Barbounis, T., Theocharis, J., Alexiadis, M. & Dokopoulos, P. (2006). Long-term wind speed and power forecasting using local recurrent neural network models. *Energy Conversion, IEEE Transactions on*, **21**, 273 – 284. 7

Bardossy, A. & Plate, E.J. (1992). Space-time model for daily rainfall using atmospheric circulation patterns. *Water Resources Research*, **28**, 1247–1259. 86

Barthelmie, R.J., Courtney, M.S., Lange, B., Nielsen, M., Sempreviva, A., Svenson, J., Olsen, F. & Christensen, T. (1999). Offshore wind resources at Danish measurement sites. In *1999 European Wind Energy Conference: wind energy for the next millennium*, 1101–1104, James & James (Science Publishers) Ltd. 188

Bell, T.L. (1987). A space-time stochastic model of rainfall for satellite remote-sensing studies. 84

Berrocal, V.J., Raftery, A.E. & Gneiting, T. (2008). Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. *The Annals of Applied Statistics*, **2**, 1170–1193. 24, 86, 92

Berrocal, V.J., Raftery, A.E., Gneiting, T. & Steed, R.C. (2010). Probabilistic weather forecasting for winter road maintenance. *Journal of the American Statistical Association*, 1–16. 86, 92

Bochner, S. (1955). *Harmonic analysis and the theory of probability*. University of California Press, Berkeley and Los Angeles. 56, 64

Bollerslev, T. (1986). Generalized autorregressive conditional heteroskedasticity. *Journal of Econometrics*, **31**, 301–327. 209

Box, G., Jenkins, G.M. & Reinsel, G. (1994). *Time Series Analysis: Forecasting & Control (3rd Edition)*. Prentice Hall. 208

Bremnes, J.B. (2004). Probabilistic wind power forecasts using local quantile regression. *Wind Energy*, **7**, 47–54. 22, 23

Bremnes, J.B. (2006). A comparison of a few statistical models for making quantile wind power forecasts. *Wind Energy*, **9**, 3–11. 18

Brown, B.G., Katz, R.W. & Murphy, A.H. (1984). Time series models to simulate and forecast wind speed and wind power. *Journal of Applied Meteorology*, **23**, 11841195. 2

Brown, R.G. & Meyer, R.F. (1961). The fundamental theorem of exponential smoothing. *Operations Research*, **9**, 673–685. 32

Burnham, K.P. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods Research*, **33**, 261–304. 133

Carlin, J. & Haslett, J. (1982). The probability distribution of wind power from a dispersed array of wind turbine generators. *Journal of Applied Meteorology*, **21**, 303–313. 9

Christoffersen, P.F. & Diebold, F.X. (1997). Optimal prediction under asymmetric loss. *Econometric Theory*, **13**, 808–817. 44

Clements, M.P. & Smith, J. (2000). Evaluating the forecast densities of linear and non-linear models: applications to output growth and unemployment. *Journal of Forecasting*, **19**, 255–276. 232

Cogger, K.O. (1974). The optimality of general-order exponential smoothing. *Operations Research*, **22**, 858–867. 219

Costa, A., Crespo, A., Navarro, J., Lizcano, G., Madsen, H. & Feitosa, E. (2008). A review on the young history of the wind power short-term prediction. *Renewable and Sustainable Energy Reviews*, **12**, 1725–1744. 3

Cressie, N. (1993). *Statistics for Spatial Data*. Wiley-Interscience. 6, 45, 47, 58, 59, 60, 80

Cressie, N. & Huang, H.C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, **94**, 1330–1340. 56, 64, 65, 66, 76, 129, 159

Czado, C., Gneiting, T. & Held, L. (2009). Predictive model assessment for count data. *Biometrics*, **65**, 1254–1261. 139

Davies, N., Pemberton, J. & Petruccelli, J.D. (1988). An automatic procedure for identification, estimation and forecasting univariate self exiting threshold autoregressive models. *Journal of the Royal Statistical Society. Series D (The Statistician)*, **37**, 199–204. 23

Deb, K., Agrawal, S., Pratap, A. & Meyarivan, T. (2000). A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii. In M. Schoenauer, K. Deb, G. Rudolph, X. Yao, E. Lutton, J. Merelo & H.P. Schwefel, eds., *Parallel Problem Solving from Nature PPSN VI*, vol. 1917 of *Lecture Notes in Computer Science*, 849–858, Springer Berlin / Heidelberg. 101

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**, 1–38. 97

Diebold, F.X. (2007). *Elements of Forecasting (4th Edition)*. Cincinnati: South-Western College Publishing. 213

Diebold, F.X. & Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, **13**, 253–63. 156

Diebold, F.X., Gunther, T.A. & Tay, A.S. (1998). Evaluating density forecasts: with applications to financial risk management. *International Economic Review*, **39**, 863–883. 229, 232

Diggle, P.J. & Ribeiro, P.J. (2007). *Model-Based Geostatistics*. Springer Series in Statistics. 6, 66

Diggle, P.J., Tawn, J.A. & Moyeed, R.A. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **47**, 299–350. 6

Doherty, R. & O'Malley, M. (2005). A new approach to quantify reserve demand in systems with significant installed wind capacity. *IEEE Transactions on Power Systems*, **20**, 587–595. 3

Doob, J. (1944). The elementary Gaussian processes. *The Annals of Mathematical Statistics*, **15**, 229–282. 63

Dunson, D.B. & Herring, A.H. (2005). Bayesian latent variable models for mixed discrete outcomes. *Biostatistics*, **6**, 11–25. 84

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika*, **68**, 589–599. 162

Engle, R.F. (1982). Autorregressive conditional heteroskedasticity with estimates of United Kingdom inflation. *Econometrica*, **50**, 987–1008. 208

Engle, R.F. & Bollerslev, T. (1986). Modeling the persistence of conditional variances. *Econometric Reviews*, **5**, 1–50. 222

Farewell, V.T. (1978). Jackknife estimation with structured data. *Biometrika*, **65**, pp. 444–447. 162

Fox, B., Flynn, D., Bryans, L., Jenkins, N., Milborrow, D., O'Malley, M., Watson, R. & Anaya-Lara, O. (2007). *Wind Power Integration: Connection and System Operational Aspects*. Institution of Engineering and Technology. 2, 3, 53, 116

Gardner, E.S. (2006). Exponential smoothing: The state of the art. *Journal of Forecasting*, **4**, 1–28. 32

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741. 97

Geweke, J. (1991). Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints and the evaluation of constraint probabilities. Computer Sciences and Statistics Proceedings of he 23^{rd} Symposium on the Interface. 98

Gibbs, M. & MacKay, D.J. (1997). Efficient implementation of Gaussian processes. Tech. rep., Cambridge University. 83

Giebel, G., Kariniotakis, G. & Brownsword, R. (2003). The state-of-the-art in short-term prediction of wind power – a literature overview. *EU project Anemos, Deliverable Report D1.1*. 3

Gipe, P. (2004). *Wind Power: Renewable Energy for Home, Farm and Business*. Chelsea Green Publishing Co. 7

Glasbey, C.A. & Nevison, I.M. (1997). Rainfall modelling using a latent Gaussian variable. In *Modelling Longitudinal and Spatially Correlated Data: Methods, Applications, and Future Directions*, 233–242, Springer. 84

GNEITING, T. (2002a). Compactly supported correlation functions. *Journal of Multivariate Analysis*, **83**, 493 – 508. 65

GNEITING, T. (2002b). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, **97**, 590–600. 56, 64, 65, 66, 130, 133, 160

GNEITING, T. (2010). Quantiles as optimal point forecasts. *International Journal of Forecasting*, **In Press, Corrected Proof**, –. 229

GNEITING, T. & RAFTERY, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359–378. 230

GNEITING, T. & RANJAN, R. (2010). Comparing density forecasts using threshold and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*. 231

GNEITING, T., RAFTERY, A.E., WESTVELD, A.H. & GOLDMAN, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, **133**, 1098–1118. 231

GNEITING, T., LARSON, K., WESTRICK, K., GENTON, M.G. & ALDRICH, E. (2006). Calibrated probabilistic forecasting at the Stateline Wind Energy Center: The regime-switching spacetime method. *Journal of the American Statistical Association*, **101**, 968–979. 23, 40, 99

GNEITING, T., BALABDAOUI, F. & RAFTERY, A.E. (2007a). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**, 243–268. 231

GNEITING, T., GENTON, M.G. & GUTTORP, P. (2007b). Geostatistical space-time models, stationarity, separability and full symmetry. In *Statistical Methods for Spatio-Temporal Systems*, 151–175, Chapman & Hall/CRC. 50, 57, 58, 60, 61

Gneiting, T., Kleiber, W. & Schlather, M. (2010). Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association*, **105**, 1167–1177. 74, 75

Granger, C.W.J. & Newbold, P. (1986). *Forecasting Economic Time Series (2nd Edition)*. Academic Press. 212

Gregori, P., Porcu, E., Mateu, J. & Sasvri, Z. (2008). On potentially negative space time covariances obtained as sum of products of marginal ones. *Annals of the Institute of Statistical Mathematics*, **60**, 865–882. 74, 94

Hamilton, J.D. (1994). *Time Series Analysis*. Princeton University Press. 36

Harvey, D., Leybourne, S. & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, **13**, 281–291. 156

Haslett, J. & Raftery, A.E. (1989). Space-time modelling with long-memory dependence: Assessing Ireland's wind power resource. *Applied Statistics*, **38**, 1–50. 8, 49, 50, 51, 59, 65, 66

Henderson, R., Diggle, P. & Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, **1**, 465–480. 104

Higdon, D.M. (1998). A process-convolution approach to modeling temperatures in the North Atlantic ocean. *Journal of Ecological and Environmental Statistics*, 173–190. 5, 205

Holt, C.E. (1957). Forecasting trends and seasonal by exponentially weighted averages. *ONR Memorandum no. 52*. 234

Holttinen, H. (2005). Optimal electricity market for wind power. *Energy Policy*, **33**, 2052 – 2063. 22

Huang, H.C. & Cressie, N. (1996). Spatio-temporal prediction of snow water equivalent using the Kalman filter. *Computational Statistics & Data Analysis*, **22**, 159–175. 6, 206

Huang, H.C. & Hsu, N.J. (2004). Modeling transport effects on ground-level ozone using a non-stationary space-time model. *Environmetrics*, **15**, 251–268. 206

Hyndman, R.J., Koehler, A.B., Ord, J.K. & Snyder, R.D. (2008). *Forecasting with Exponential Smoothing: The State Space Approach*. Springer Series in Statistics. 20, 32, 33, 220

Johnson, N.L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, **36**, 149–176. 18

Jorgenson, D.W. (1967). Seasonal adjustment of data for econometric analysis. *Journal of the American Statistical Association*, **62**, 137–140. 16

Juban, J., Siebert, N. & Kariniotakis, G. (2007). Probabilistic short-term wind power forecasting for the optimal management of wind generation. In *Power Tech, 2007 IEEE Lausanne*, 683 – 688. 22

Jun, M. & Stein, M.L. (2007). An approach to producing spacetime covariance functions on spheres. *Technometrics*, **49**, 468–479. 62

Kariniotakis, G., Stavrakakis, G. & Nogaret, E. (1996). Wind power forecasting using advanced neural networks models. *Energy Conversion, IEEE Transactions on*, **11**, 762 –767. 6

Krige, D.G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, **52**, 119–139. 45

Laio, F. & Tamea, S. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, **11**, 1267–1277. 231

Landberg, L. (1999). Short-term prediction of the power production from wind farms. *Journal of Wind Engineering and Industrial Aerodynamics*, **80**, 207 – 220. 6

Landberg, L., Giebel, G., Nielsen, H.A., Nielsen, T. & Madsen, H. (2003). Short-term prediction–an overview. *Wind Energy*, 273 – 280. 3

Lange, M. (2005). On the uncertainty of wind power predictions—analysis of the forecast accuracy and statistical distribution of errors. *Journal of Solar Energy Engineering*, **127**, 177–184. 41

Lau, A. & McSharry, P. (2010). Approaches for multi-step density forecasts with application to aggregated wind power. *Ann. Appl. Statist.*, **4**, 1311–1341. 24, 27

Ledolter, J. & Box, G.E.P. (1978). Conditions for the optimality of exponential smoothing forecast procedures. *Metrika*, **25**, 77–93. 34, 219

Lemos, R.T. & Sans, B. (2009). A spatio-temporal model for mean, anomaly, and trend fields of North Atlantic sea surface temperature. *Journal of the American Statistical Association*, **104**, 5–18. 5

Lilliefors, H.W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, **62**, pp. 399–402. 13

Little, M.A., McSharry, P.E. & Taylor, J.W. (2009). Generalized linear models for site-specific density forecasting of U.K. daily rainfall. *Monthly Weather Review*, **137**, 1029–1045. 8

Little, R.J.A. & Rubin, D.B. (2002). *Statistical Analysis with Missing Data, Second Edition*. Wiley-Interscience, 2nd edn. 104

Ma, C. (2003). Families of spatio-temporal stationary covariance models. *Journal of Statistical Planning and Inference*, **116**, 489–501. 60, 227

Ma, C. (2005). Linear combinations of space-time covariance functions and variograms. *Signal Processing, IEEE Transactions on*, **53**, 857–864. 60

Ma, L., Luan, S., Jiang, C., Liu, H. & Zhang, Y. (2009). A review on the forecasting of wind speed and generated power. *Renewable and Sustainable Energy Reviews*, **13**, 915 – 920. 7

MacKay, D.J. (1998). Introduction to Gaussian processes. Tech. rep., Cambridge University. 83

Madsen, H., Pinson, P., Kariniotakis, G., Nielsen, H.A. & Nielsen, T.S. (2005). Standardizing the performance evaluation of short-term wind prediction models. *Wind Engineering*, **29**, 475–489. 22

Mansur, E.T., Mendelsohn, R. & Morrison, W. (2005). A discrete-continuous choice model of climate change impacts on energy. http://www.earth.columbia.edu/cgsd/documents/mansur.pdf. 84

Manzan, S. & Zerom, D. (2008). A bootstrap-based non-parametric forecast density. *International Journal of Forecasting*, **24**, 535–550. 23, 205

Marcellino, M., Stock, J.H. & Watson, M.W. (2006). A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series. *Journal of Econometrics*, **135**, 499–526. 23

Mardia, K., Goodall, C., Redfern, E. & Alonso, F. (1998). The kriged Kalman filter. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, **7**, 217–282. 206

Marschner, I.C. (2001). On stochastic versions of the EM algorithm. *Biometrika*, **88**, 281–286. 95, 97

Massey, F.J.J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, **46**, pp. 68–78. 13

Matérn, B. (1986). *Spatial variation*. Lecture notes in statistics, Springer-Verlag. 66

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, **58**, 1246–1266. 45

May, D.R. & Julien, P.Y. (1998). Eulerian and Lagrangian correlation structures of convective rainstorms. *Water Resources Research*, **34**, 2671–2683. 60

McSharry, P., Pinson, P. & Gerard, R. (2009). Methodology for the evaluation of probabilistic forecasts. Safewind report, Oxford University. 228

Milligan, M., Schwartz, M. & Wan, Y. (2004). Statistical wind power forecasting models: Results for U.S. wind farms. The 17^{th} Conference on Probability and Statistics in the Atmospheric Sciences/2004 American Meteorological Society Annual Meeting, Seattle, Washington, January 11-15, 2004. 3, 22

Moeanaddin, R. & Tong, H. (1990). Numerical evaluation of distributions in nonlinear autoregression. *Journal of Time Series Analysis*, **11**, 33–48. 23

Moller, J.T. (1992). Balanced coastal protection on a Danish North Sea coast. *Journal of Coastal Research*, **8**, pp. 712–718. 175

Muth, J.F. (1960). Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association*, **55**, 299–306. 219

Myers, D.E. (1982). Matrix formulation of Co–Kriging. *Mathematical Geology*, **14**, 249–257. 47, 48

Nelson, D.B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, **59**, 347–70. 37

Olsen, R.J. (1980). Approximating a truncated normal regression with the method of moments. *Econometrica*, **48**, pp. 1099–1105. 41

Parsons, B., Milligan, M., Zavadil, B., Brooks, D., Kirby, B., Dragoon, K. & Caldwell, J. (2004). Grid impacts of wind power: a summary of recent studies in the United States. *Wind Energy*, **7**, 87–108. 2

Patton, A.J. & Timmermann, A. (2007). Properties of optimal forecasts under asymmetric loss and nonlinearity. *Journal of Econometrics*, **140**, 884 – 918. 44

Pinson, P. (2010). Very short-term probabilistic forecasting of wind power time series with generalized logit-normal distributions. *Journal of the Royal Statistical Society: Series C, (In press)*. 40

Pinson, P. & Madsen, H. (2009). Ensemble-based probabilistic forecasting at Horns Rev. *Wind Energy*, **12**, 137–155. 4

Pinson, P., Chevallier, C. & Kariniotakis, G. (2007a). Trading wind generation from short-term probabilistic forecasts of wind power. *Power Systems, IEEE Transactions on*, **22**, 1148 –1156. 22

Pinson, P., Chevallier, C. & Kariniotakis, G. (2007b). Trading wind generation with short-term probabilistic forecasts of wind power. *IEEE Transactions on Power Systems*, **22**, 1148–1156. 3

Pinson, P., Nielsen, H.A., Madsen, H. & Nielsen, T.S. (2008). Local linear regression with adaptive orthogonal fitting for the wind power application. *Statistics and Computing*, **18**, 59–71. 7

Pinson, P., McSharry, P. & Madsen, H. (2010). Reliability diagrams for non-parametric density forecasts of continuous variables: Accounting for serial correlation. *Quarterly Journal of the Royal Meteorological Society*, **136**, 77–90. 124, 229

Poirier, D.J. & Ruud, P.A. (1988). Probit with dependent observations. *Review of Economic Studies*, **55**, 593–614. 95

Potter, C. & Negnevitsky, M. (2006). Very short-term wind forecasting for Tasmanian power generation. *Power Systems, IEEE Transactions on*, **21**, 965 – 972. 6

Raftery, A.E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, **83**, 251–266. 133

Rappold, A.G., Gelfand, A.E. & Holland, D.M. (2008). Modelling mercury deposition through latent space-time processes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **57**, 187–205. 86

Rasmussen, C.E. & Williams, C.K.I. (2006). *Gaussian processes for machine learning*. MIT Press. 63, 64, 83

Robinson, P.M. (1982). On the asymptotic properties of estimators of models containing limited dependent variables. *Econometrica*, **50**, 27–41. 95

Rodriguez-Yam, G., Davis, R.A. & Scharf, L. (2004). Efficient Gibbs sampling of truncated multivariate normal with application to constrained linear regression. 98

Roy, B. (1971). Problems and methods with multiple objective functions. *Mathematical Programming*, **1**, 239–266, 10.1007/BF01584088. 101

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc. 104

Sammel, M.D., Ryan, L.M. & Legler, J.M. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **59**, 667–678. 84

Sanchez, I. (2006). Short-term prediction of wind energy production. *International Journal of Forecasting*, **22**, 43–56. 4, 205

Sanso, B. & Guenni, L. (1999). Venezuelan rainfall data analysed by using a Bayesian space-time model. *Applied Statistics*, **48**, 345–362. 5, 40, 84

Sanso, B. & Guenni, L. (2000). A nonstationary multisite model for rainfall. *Journal of the American Statistical Association*, **95**, 1089–1100. 84, 85

Satchell, S. & Timmermann, A. (1994). Optimal properties of exponentially weighted forecasts in the presence of different information sources. *Economics Letters*, **45**, 169–174. 219

Satchell, S. & Timmermann, A. (1995). On the optimality of adaptive expectations: Muth revisited. *International Journal of Forecasting*, **11**, 407 – 416. 219

Sideratos, G. & Hatziargyriou, N. (2007). An advanced statistical method for wind power forecasting. *Power Systems, IEEE Transactions on*, **22**, 258 –265. 7

Sloughter, J.M., Raftery, A.E., Gneiting, T. & Fraley, C. (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review*, **135**, 3209–3220. 86

Stein, M.L. (1988). Asymptotically efficient prediction of a random field with a misspecified covariance function. *The Annals of Statistics*, **16**, 55–63. 76

Stein, M.L. (1992). Prediction and inference for truncated spatial data. *Journal of Computational and Graphical Statistics*, **1**, 91–110. 84

Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series in Statistics. 66

Stein, M.L. (2005). Space-time covariance functions. *Journal of the American Statistical Association*, **100**, 310–321. 62, 66, 68, 76, 130, 160

Stein, M.L. (2009). Spatial interpolation of high-frequency monitoring data. *The Annals of Applied Statistics*, **3**, 272–291. 124

Stern, R.D. & Coe, R. (1984). A model fitting analysis of daily rainfall data. *Journal of the Royal Statistical Society. Series A (General)*, **147**, 1–34. 86

Stroud, J.R., Müller, P. & Sans, B. (2001). Dynamic models for spatiotemporal data. *Journal Of The Royal Statistical Society Series B*, **63**, 673–689. 6, 206

Tastu, J., Pinson, P., Kotwa, E., Madsen, H. & Nielsen, H.A. (2011). Spatio-temporal analysis and modeling of short-term wind power forecast errors. *Wind Energy*, **14**, 43–60. 2, 4, 23

Taylor, G.I. (1938). The spectrum of turbulence. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, **164**, 476–490. 61

Taylor, J.W. (2003). Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of the Operational Research Society*, **54**, 799–805. 32, 33, 220, 234

Taylor, J.W. (2004). Volatility forecasting with smooth transition exponential smoothing. *International Journal of Forecasting*, **20**, 273 – 286. 32

Taylor, J.W. (2008). Using exponentially weighted quantile regression to estimate value at risk and expected shortfall. *Journal of Financial Econometrics*, **6**, 382–406. 240

Taylor, J.W., McSharry, P.E. & Buizza, R. (2009). Wind power density forecasting using wind ensemble predictions and time series models. *IEEE Transactions on Energy Conversion*, **24**, 775–782. 4, 7, 18, 231, 240

Tsay, R.S. (2005). *Analysis of financial time series*. Wiley Series in Probability and Statistics, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2nd edn. 29, 30, 37, 209

van Wijk, A.J.M., Halberg, N. & Turkenburg, W.C. (1992). Capacity credit of wind power in the Netherlands. *Electric Power Systems Research*, **23**, 189 – 200. 2

Wagner, H.J. & Mathur, J. (2009). Physics of wind energy. In H.J. Kowalski & J. Mathur, eds., *Introduction to Wind Energy Systems*, Green Energy and Technology, 17–28, Springer Berlin Heidelberg. 7, 11

Weigend, A.S. & Shi, S. (2000). Predicting daily probability distributions of S&P500 returns. *Journal of Forecasting*, **19**, 375 – 392. 231

Weisser, D. & Foxon, T. (2003). Implications of seasonal and diurnal variations of wind velocity for power output estimation of a turbine: a case study of Grenada. *International Journal of Energy Research*, **27**, 1165–1179. 17

Wikle, C.K. & Cressie, N. (1999). A dimension-reduced approach to space-time Kalman filtering. *Biometrika*, **86**, 815–829. 206

Wikle, C.K., Berliner, L.M. & Cressie, N. (1998). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics*, **5**, 117–154. 6

Wikle, C.K., Milliff, R.F., Nychka, D. & Berliner, L.M. (2001). Spatiotemporal hierarchical Bayesian modeling: Tropical ocean surface winds. *Journal of the American Statistical Association*, **96**, 382–397. 5

Yu, K. & Jones, M.C. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, **93**, pp. 228–237. 240

Zeger, S.L. & Karim, M.R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, pp. 79–86. 90

Zitzler, E. & Thiele, L. (1999). Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. *Evolutionary Computation, IEEE Transactions on*, **3**, 257 –271. 102