Bayesian Spatio-temporal Modelling for Forecasting Ground Level Ozone Concentration Levels

Chun Yin Yip, B.Sc. (Hons)

Supervisor: Dr. S. K. Sahu

Advisor: Prof. J. J. Forster

A thesis submitted for the degree of Doctor of Philosophy at University of Southampton

May 2010

Abstract

Accurate, instantaneous and high resolution spatial air-quality information can better inform the public and regulatory agencies of the air pollution levels that could cause adverse health effects. The most direct way to obtain accurate air quality information is from measurements made at surface monitoring stations across a study region of interest. Typically, however, air monitoring sites are sparsely and irregularly spaced over large areas. That is why, it is now very important to develop space-time models for air pollution which can produce accurate spatial predictions and temporal forecasts.

This thesis focuses on developing spatio-temporal models for interpolating and forecasting ground level ozone concentration levels over a vast study region in the eastern United States. These models incorporate output from a computer simulation model known as the Community Multi-scale Air Quality (Eta-CMAQ) forecast model that can forecast up to 24 hours in advance. However, these forecasts are known to be biased. The models proposed here are shown to improve upon these forecasts for a two-week study period during August 2005.

The forecasting problems in both hourly and daily time units are investigated in detail. A fast method, based on Gaussian models is constructed for instantaneous interpolation and forecasts of hourly data. A more complex dynamic model, requiring the use of Markov chain Monte Carlo (MCMC) techniques, is developed for forecasting daily ozone concentration levels. A set of model validation analyses shows that the prediction maps that are generated by the aforementioned models are more accurate than the maps based solely on the Eta-CMAQ forecast data. A non-Gaussian measurement error model is also considered when forecasting the extreme levels of ozone concentration. All of the methods presented are based on Bayesian methods and MCMC sampling techniques are used in exploring posterior and predictive distributions.

Acknowledgements

I am greatly indebted to a lot of people during the course of my PhD studies and the preparation of the thesis. I would like to thank a few of them for their precious support.

First, I would like to thank my supervisor Dr. Sujit Sahu for his continuous support throughout my post-graduate studies. He provided me with deep insights in statistics and the opportunity to spend several months in Duke University and Environmental Protection Agency in the USA, which was proved to be invaluable and fruitful. Dr. Sahu has also helped improving my arguments in the thesis and presentation skills. I would never be able to finish this thesis without his guidance.

I would also like to take this opportunity to thank my advisor Prof. Jonathan Forster. His great enthusiasm and the invaluable suggestions he has given me on several occasions are truly inspirational. A special thank goes to Prof. Philip Prescott, who was my undergraduate tutor when I did my Mathematics BSc. degree in Southampton. He inspired me tremendously in various branches of mathematics and that laid the foundation of my current career.

I am particularly thankful to my roommates Saeed Al-Ghamdi, Sarah Carnaby, Shan Lin, Andrew Rose and Jeffrey Samuel for their humour, kindness and assistance. Besides studying together, we were good companions at lunchtimes and formed a good team in pub quizzes.

Finally, I owe my warmest thanks to my parents. With their love, patient and never-ending support throughout the last twenty-eight years, I can always combat any obstacles in my life.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

As we enter a new age where air pollution data can be accessed in real-time, new space-time models are needed to provide continuous, updated maps of current and future air pollution levels. The most direct way to obtain accurate air quality information is from measurements made at surface monitoring stations across particular study regions. Typically, however, air monitoring sites are sparsely and irregularly spaced over large areas. Thus, it is now important to develop computationally efficient models to combine sparsely observed air monitoring data and numerical model output available everywhere, in a coherent way for better prediction of air pollution over a short period of time.

This thesis is motivated by the need to:

1. capture spatio-temporal variation in air pollution,
2. improve biased forecasts from numerical model output,
3. fuse ground level observation data with deterministic computer model output,
4. quantify uncertainty in forecasts through the use of Bayesian methods,
5. produce high resolution maps of air pollution.

These core problems are considered in a holistic framework in this thesis. Tools and methods such as Kriging, time series analysis, data assimilation, Bayesian hierarchical modelling, and extreme value theory are used extensively towards solving the core problems in studying air pollution–ozone concentration levels in particular in the eastern United States (US).

The structure of the rest of this chapter is as follows: Section 1.1 provides a brief outline of the chemical properties, measurement and regulation laws relating to ozone concentration; Section 1.2 introduces the data set that is used throughout this thesis. Sections 1.3-1.7 address the five core scientific problems; Section 1.8 presents the outline of the thesis and contribution it can make to the subject area. Finally Section 1.9 summarises this chapter.

1.1 Ozone

Ozone is an odourless and colourless gas composed of three oxygen atoms, which can be found both in the earth's upper atmosphere and at ground level. At ground level, ozone can cause a number of respiratory and health problems like coughing, throat irritation, congestion, bronchitis, emphysema and asthma especially for people who are sensitive to high air pollution levels.

In the earth's troposphere, ozone is indirectly created by automobile engines, industrial boilers, power plants and refineries; these sources emit hydrocarbons and nitrogen oxides (NO_x) that react chemically in the presence of sunlight. Meteorological conditions such as sunlight intensity, temperature, wind direction and speed hugely influence the concentration and distribution of ozone. In the US, ground level ozone concentration levels are usually high during mid April to the end of September due to presence of sunlight and high temperature.

Ground level ozone concentration is usually measured by unattended photometers (Hedges, 1999). The measurements are affected by surrounding

conditions but the monitoring locations chosen are usually sparsely and irregularly spaced due to administrative reasons. The automatic instruments can malfunction and as a result may record inaccurate measurements continuously until rectified by human intervention. Often, this leads to a series of consecutive missing observations.

The US Environmental Protection Agency (USEPA) have developed the Air Quality Index (AQI) to indicate the levels of ozone and other common air pollutants at ground level. The index is based on the level of ozone concentration measured by a nationwide monitoring system from more than a thousand locations across the country. The AQI is measured against the air quality standards established by USEPA under the Clean Air Act to protect public health and the environment, as a useful safety indicator for the general public and also for the policy makers in the USEPA. For example, at ozone levels between 80 and 120 parts per billion (ppb), it is suggested that even moderate outdoor exertion for longer periods of time can increase one's risk in experiencing ozone-related effects.

Apart from observations, ground level ozone concentration levels can also be predicted by computer simulation models based on conservation laws and fluid dynamics. The predictions obtained from such models are known to be biased. More details regarding these predictions and ground level monitoring observations are discussed in the following sections.

1.2 Data

1.2.1 Observed Data

Ozone concentration data are obtained from $n = 390$ monitoring stations covering the region in the US between –84.70^{o}W and –68^{o}W from 2nd–15th August in 2005. Locations of sites are irregularly-spaced over the region, see Figure 1.1 for more details. Samples at each location are obtained for every

hour. Thus, 312 × 390 observations are obtained in total. There are about 20% of the data missing which we assume to have occurred completely at random. The concentration values vary from $0-192$ppb with a mean of about 34ppb. Data from 40 randomly chosen sites are set aside for model validation purposes and data from the remaining 350 sites are used for modelling and estimation.

1.2.2 Eta-CMAQ Computer Simulation Output

The National Oceanic and Atmospheric Administration (NOAA) in the United States designed the Community Multi-scale Air Quality (CMAQ) modelling system (http://www.epa.gov/amad/CMAQ/index.html) to forecast levels of various air pollutants such as ground level ozone concentration (Ching and Byun, 1999). The CMAQ forecasting model is not a statistical one but a deterministic differential equations model which takes several inputs based on emission, meteorology, transportation dynamics and ground characteristics that affect the level of air pollutants. It contains an interface processor which incorporates information from different modules such as meteorology, emissions and photolysis rates. The modules are mainly developed by first principles and requisite information for initial and boundary conditions is prepared as preprocessors. These modules are actually smaller computer programmes which provide information to the Chemical Transition Model (CTM) and also act as components in the system that can be replaced if they are not satisfactory enough. The CTM itself consists of six physical and chemical process components: (1) advection and diffusion, (2) gas phase chemistry, (3) plume-in-grid modelling, (4) particle modelling and visibility, (5) cloud processes, and (6) photolysis rates.

The conceptual structure of the CTM is shown in Figure 1.2. The CTM performs chemical transport modelling for multiple pollutants on multiple scales under certain physical assumptions such as incompressible atmosphere

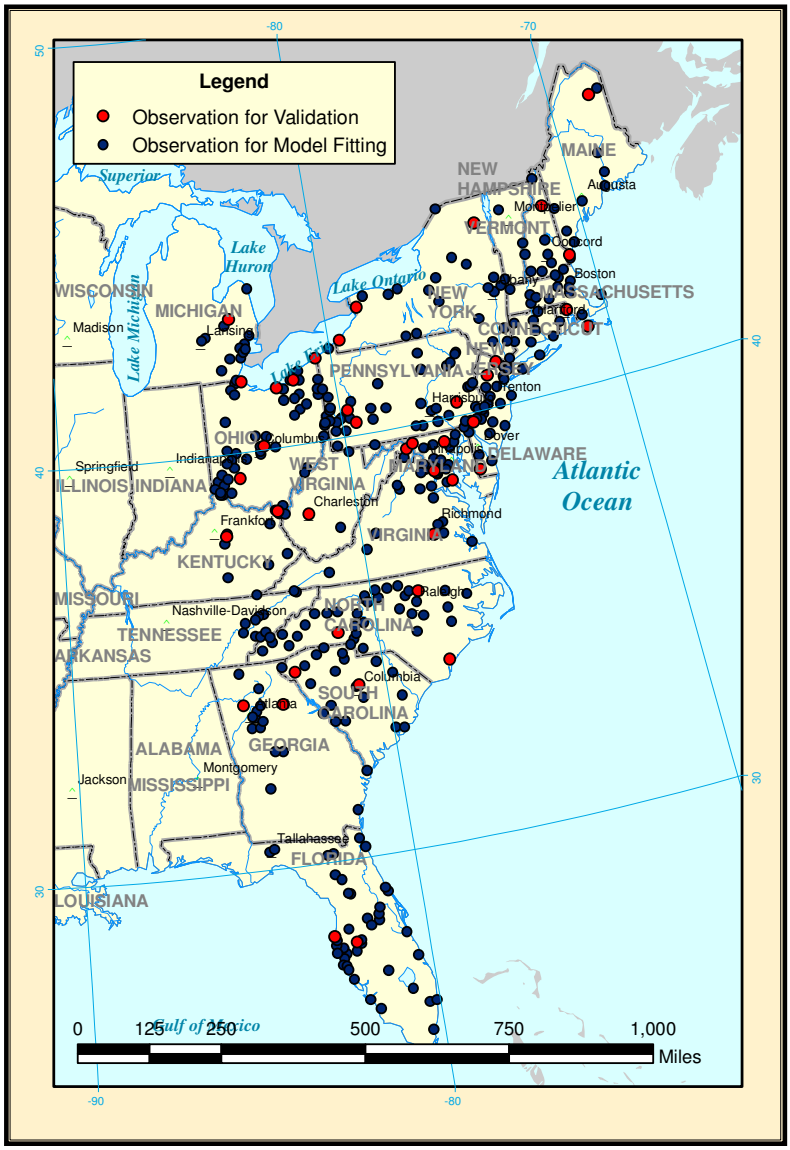


Figure 1.1: Plot of the 350 observation sites and 40 set-aside validation sites in the eastern US.

and non-divergent flow field assumptions. The philosophy behind this is to provide a modelling system in "one atmosphere" which makes the simulation as realistic as the real world. The Meteorology Modeling System version 5(MM5) mentioned in Figure 1.2 is a complex community model that includes proper grid definitions, physical models and a four-dimensional data assimilation scheme, and it produces the meteorological fields for the CTM. The CMAQ modelling systems also contain the following processors and interfaces:

- Meteorology-Chemistry Interface Processor (MCIP) translates and processes data generated from the MM5 for the CTM. MCIP interpolates the meteorological data needed, converts between coordinate systems, and computes the cloud, surface and planetary boundary layer parameters. It also imports information derived from the land use information from land use processor.

- Emission-Chemistry Interface Processor (ECIP) translates data from the Model-3 Emissions Processor and Projection System (MEPPS) for the CTM. ECIP generates hourly three-dimensional emission data for Eta-CMAQ. Meteorological data required for predicting emissions come from MCIP and MM5.

- Initial Conditions (ICON) and Boundary Conditions (BCON) provide concentration fields for chemicals for the initial simulation state and for the grids surrounding the modelling domains respectively. The ICON and BCON use raw data or previously modelled simulation data.

- Photolysis Processor (JPROC) deals with the temporally varying photolysis rates. JPROC uses ozone, temperature aerosol number density and earth's surface albedo (sunlight reflectivity) raw data to produce the initial photolysis rates and a table of photo-dissociation reaction rates for the CTM.

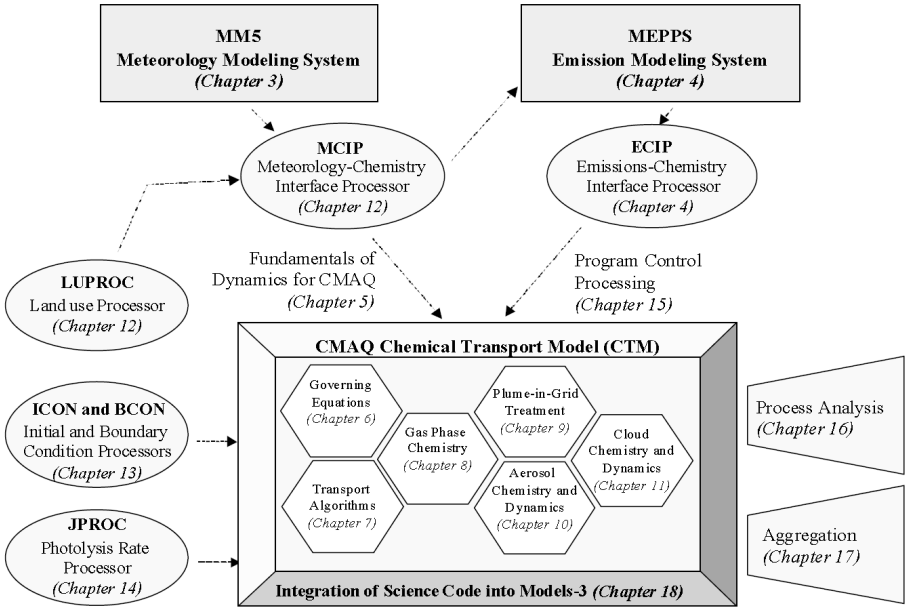


Figure 1.2: A graphical illustration of the Eta-CMAQ modelling systems structure (Ching and Byun, 1999).

The last stage of the CMAQ modelling systems is process analysis and aggregation. Process analysis is a pre-processor programme which aims to detect any error and uncertainties in a model through the parametrisation schemes and the input data. Aggregation is a statistical procedure which can derive the required seasonal and annual estimates without executing multiple model runs. This is not useful for many short-term predictions but the output can be an indicator of any short-term change.

There are many possible versions of CMAQ models depending on the particular choices of the component modules and initial conditions. In this thesis we shall use output from a particular version known as the Eta-CMAQ model. This model produces the forecast for each hour as an average concentration level for a 12 square-kilometre grid cell. There are 259×268 such grid cells covering much of the continental US. In our study region of the eastern US there are only 9119 such grid cells, see Figure 1.3.

Note that observed ozone concentration levels may have been fed into the Eta-CMAQ model through the ICON processor. However, the Eta-CMAQ model produces forecasts of ozone concentration levels up-to 48 hours in advance. Obviously, these forecasts do not use any observed data during this forecasting period of 48 hours. That is why it is reasonable to assume that the observed data are independent of the Eta-CMAQ forecasts. Moreover, the Eta-CMAQ forecasts are average hourly values for a grid cell while the observed data correspond to hourly values at a particular location referenced by a latitude-longitude pair. This gives rise to a spatial mis-alignment problem between the Eta-CMAQ forecasts and observed data. This problem is well-known in the literature, see Section 1.5.

In this thesis we shall use the Eta-CMAQ forecasts as a regressor for the ozone concentration levels since it is reasonable to expect that these two will be very similar. Indeed, see Figure 1.4 where we plot the hourly recorded data with the Eta-CMAQ forecast for the grid cell covering that location for the period 2nd–15th August, 2005 for two randomly chosen sites. It is noted that the Eta-CMAQ forecasts sometimes capture the measurement processes very well but may fail at other times due to various reasons. This is also supported by the fact that the average correlation between the hourly Eta-CMAQ forecasts for the period 2nd–15th August and the corresponding observed data for the 390 monitoring sites is 0.54.

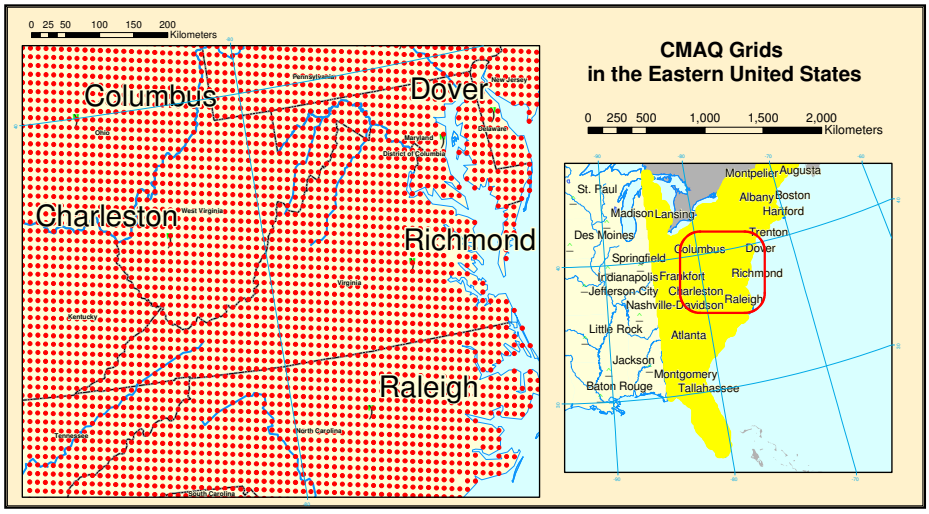


Figure 1.3: Left: an enlarged view of the centroids of the grids cells; Right: all 9119 grid cells in the eastern United States. Note that, for easy reference, centroids are shown instead of grid cells.

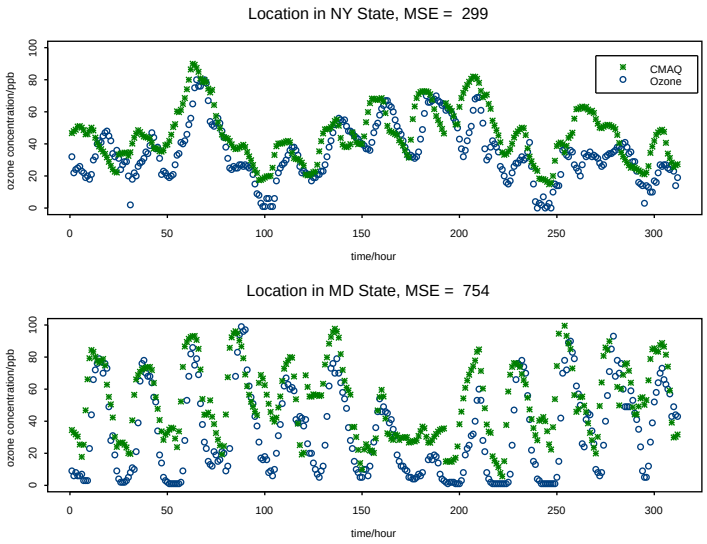


Figure 1.4: Time series plots of observed hourly ozone concentration data and the Eta-CMAQ forecasts for the grid cell which includes the data site for two randomly chosen sites in the Eastern US. The blue circles represent the observed data and the green asterisks are the Eta-CMAQ forecasts; the top panel is for a site in the state of New York and the bottom panel is for a site in Maryland. The MSE in each plot is the mean-square error between the observed data and the Eta-CMAQ forecasts.

1.3 Capturing Spatio-Temporal Variation

Geostatistics methods developed by Matheron (1971) give a good foundation on capturing spatial variation. Time series analysis (for example, Chatfield, 2004) gives a concrete foundation on the temporal evolution of the meteorological fields. However, those methods do not consider the space-time variation simultaneously. It is especially of interest to consider large space-time domain because data from such a domain usually gives a large variation. Long range and large scale spatio-temporal variations, in general, are easier to capture than smaller ones which we will describe mathematically in terms of notion of covariance functions in Chapter 3.

1.4 Improving Biased Forecasts

Forecasting is a key focus of this thesis. Le and Zidek (2006) point out that forecasting weather more than two or three days ahead would be difficult since tiny perturbations in initial conditions can propagate large changes in model output. Although computer model outputs are useful, a more reliable forecast will be the synthesis of both statistical and deterministic computer simulation model. The forecasts should update the computer model output in the light of new observations. We pursue forecasting in Chapters 4 and 5 of this thesis.

Lorenz (1963) studies an atmosphere analogy of a deterministic system of non-linear ordinary differential equations. The experimental results show that the solution of a weather system is unstable and non-periodic. This makes it very difficult to make long-range predictions. Unlike the studies in Bayarri *et al.* (2007) and Kennedy and O'Hagan (2001), this thesis is not going to analyse the relationship between the Eta-CMAQ forecasts and input meteorological parameters since for some unobserved parameters, the input parameters themselves are computer modules. Also, the initial conditions in

these models are unknown.

1.5 Fusing Observations with Computer Simulation Model Output

Often, probability forecasts are more informative than the deterministic point estimates. The probabilistic forecasts can be produced by combining observations and computer simulation models. Fusing observations with computer model data is called data assimilation (DA) in the meteorology community. However, DA can be easily interpreted as a problem in Bayesian statistics.

Kolmogorov (1941) is the first to consider the problem on fitting, interpolating and smoothing from different sources of data. Gandin (1963) further develops this idea and introduces *optimal interpolation* in minimising root mean square error sum of squares. This method is of more statistical interest than before. Matheron (1971) generalises Kriging in his seminal work but he considers it as an interpolation method rather than a data assimilation method. Lorenc (1986) points out that the methodology in optimal interpolation is very similar to Kriging.

Wikle and Berliner (2006), from a Bayesian settings, review the relationship to optimal/Kriging interpolation with a numerical example. Their model is not hierarchical and the structure is restricted to a relatively small class. But a conceptual example on hierarchical models has also been proposed. A Bayesian hierarchical model can be formed by a general three-stage factorisation consisting of model equations for data, processes and parameters, see Section 2.4 for further details. This approach allows us to adopt complex models which are not accessible in either with universal Kriging or with least square fitting and prediction.

In relation to the problem in fusing observation and computer model output, numerous Bayesian approaches have been proposed. Lorenc (1986) is

the first to consider the data fusion problem in a Bayesian setting. With the development of Markov chain Monte Carlo (MCMC) algorithms in the 1990s, many high dimensional model-based Bayesian approaches have been developed: 1) Jun and Stein (2004) compare the correlation structure of computer model and observations; 2) Fuentes and Raftery (2005) combine computer model output (CMAQ) and observation by joint multivariate normal distributions; 3) Zimmerman and Holland (2005) use different data sources with different measurement errors and biases. However, none of them deal with space-time forecasts at the same time as is done here.

1.6 Quantifying Uncertainty

As detailed in Section 1.2.2, the existing deterministic ozone concentration models are based on numerous physical first principles in Newtonian mechanics, laws of thermodynamics and fluid dynamics and so on. Philosophically, many scientists believe that if one possesses the exact knowledge on the current state of the universe, "clever" enough to understand and compute all the physical laws then the real world will exactly follow the physical model constructed. In 1814, Laplace gives a general comment on his view on determinism and probability in the introduction of *Essai philosophique sur le probabilités* (see details in Grattan-Guinness, 2005):

"If an intelligence, at a given instant, knew all the forces that animate nature and the position of each constituent being; if, moreover, this intelligence were sufficiently great to submit these data to analysis, it could embrace in the same formula the movements of the greatest bodies of the universe and those of the smallest atoms: to this intelligence nothing would be uncertain, and the future, as the past, would be present to its eyes."

and also,

"The regularity which astronomy shows us in the movements of the comets doubtless occurs in all phenomena. The curve described by a simple molecule of

air or water vapour is regulated in a manner just as certain as the orbits of the planets; the only difference between these is that introduced by our ignorance. Probability is relative in part to this ignorance, and in part to our knowledge."

Although there are many ways to address the uncertainty in physical dynamical systems, the above belief establishes the viewpoint that using probability to deal with uncertainty is perhaps the best method. Bernardo and Smith (1994) view uncertainty as an "incomplete knowledge in relation to a specified object". Under the Bayesian paradigm uncertainty is often evaluated by calculating the posterior distribution formed using the likelihood and the prior distribution, see Section 2.4. This method quantifies uncertainty via a probability distribution. For practical purposes, by assessing the posterior distributions, answering questions such as :"what is the chance that tomorrow's ozone concentration will be higher than a certain value?" and "what location has over 95% chance that the ozone concentration level will be lower than 70ppb?" will be possible. See Section 5.2 where we make this sort of inference.

1.7 Producing High Resolution Maps

There is a need for a high resolution visualisation of inferential atmospheric information. An accurate, colourful map is a good way to show the meteorological properties to the general public. These inferential characteristics are originally represented as properties of posterior predictive distributions. The posterior mean/median and standard deviation at various locations are of interest to the general public and policy makers. These statistics provide information on what is expected along with the degree of uncertainty. The linkage between the inference and the posterior predictive distributions will be discussed in Chapter 2 under a hierarchical Bayesian framework. A map representing the probability of the occurrence of a certain event can also be

of interest to the general public. A general review of important methods in spatial statistics will be detailed in the next chapter. See Section 2.7 where issues regarding mapping, map projections and distance calculation have been discussed.

1.8 Thesis Outline

The work in this thesis shows the implementation of statistical techniques for solving the core problems discussed in this chapter. Bayesian inference methods are used throughout this thesis. Chapter 2 provides a review of statistical modelling and important methods in spatial statistics. Chapter 3 investigates the possible covariance structures for atmospheric processes. Chapter 4 discusses a fast forecasting approach for hourly recorded ozone concentration levels. In Chapter 5 the spatio-temporal models and forecasting issues for the daily 8-hour maximum ozone concentration are considered. Chapter 6 discusses the non-Gaussian error models for extreme events, while Chapter 7 concludes the thesis and provides some possible future research directions. An appendix contains definitions and properties of the common statistical distributions used in the thesis.

1.9 Summary

In this chapter, the details of the observed data and Eta-CMAQ numerical model output are discussed. Five closely related scientific problems are also addressed in Sections 1.3-1.7. The motivations of the problems considered in this thesis have been introduced. A unified framework for putting these purposes together is needed. A Bayesian statistical framework appears to be a persuasive way to solve the problem. These form the foundations of the work in the rest of the thesis.

Chapter 2

Review of Statistical Modelling of Spatial Data

Spatial statistics, in a very wide sense, is the study of analysing spatially dependent data. Many approaches have been developed over the last 40 years in this field since Matheron's (1971) seminal work was published. The literature in this area has its own unique set of keywords such as *Kriging* and *variogram*. Methodologies in spatial statistics are especially needed in a wide range of applications in mining, air pollution modelling, property market pricing, epidemiology, assessing flood risk and so on.

In this chapter we first describe different types of spatially dependent data and then discuss the main issues in statistical modelling of such data under the Bayesian paradigm. We then review Bayesian methods and present a number of Bayesian model choice criteria. We discuss a number of important methods in spatial statistics including many variants of Kriging. Finally, we discuss general issues regarding mapping, map projections and distance calculations between two locations. The general discussions on modelling framework laid down in this chapter will be used in developing the modelling strategies adopted in the subsequent chapters.

2.1 Types of Spatially Dependent Data

Let $\boldsymbol{s}$ be a point in the d-dimensional Euclidean space $\mathbb{R}^d$. Suppose the attribute we observe at location index $\mathbf{s}$ is $Z(\boldsymbol{s})$. A spatial process in d dimensions can be formulated by the collection of random variables,

$$\left\{Z(\mathbf{s}) : \mathbf{s} \in S \subset \mathbb{R}^d\right\},$$

where S is a continuous geographically referenced region and is a subset of $\mathbb{R}^d$. If we observe the attribute at n spatial locations $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$ then we have an observation vector $\mathbf{Z} = (Z(\mathbf{s}_1), Z(\mathbf{s}_2), \ldots, Z(\mathbf{s}_n))'$. The notation $\mathbf{z}$ will denote the actual realisation of $\mathbf{Z}$.

2.1.1 Point-referenced Data

Following Cressie (1993, Chapter 1), we classify spatial data sets into three important types. The location index $\mathbf{s}$ varies *continuously* over S, where S is a fixed subset of $\mathbb{R}^d$. It means the observation $Z(\mathbf{s})$ can be taken at any point within S. The number of points in S is thus, theoretically infinite. The collection of $Z(\mathbf{s})$ is also called geo-referenced or usually named as geostatistical data. The USEPA ozone concentration data is an example, see Figure 1.1 for a map where these data were observed.

2.1.2 Areal Data

Areal data, sometimes called lattice data, are defined on a fixed and countable domain S. This means that the observation $Z(\boldsymbol{s})$ in S is taken from an area or a region rather than a point. Although the number of such regions can be infinite, in practice the number of regions in S is finite, for example the number of postcode districts in England is finite. The study region can be either regularly or irregularly-spaced. The standardised mortality ratios of lip cancer data in Scotland is an example of irregular areal data (see Figure 2.1,

Clayton and Kaldor, 2005); our computer simulation output data described in Section 1.2.2 are of the regular type.

2.1.3 Point Pattern Data

Assume that the domain of data collection points is *stochastic*; its index set will demonstrate the locations of random events in a spatial point pattern, see the example in Figure 2.2 (Mark and Esler, 1970). A more rigorous mathematical approach is to regard it as a random countable subset of the surface S. To make it computationally tractable, the realisations of the processes are locally finite subsets of S.

In this thesis we only concentrate on analysing the first two types of spatial data. Therefore, the collection of observation locations is assumed to be *fixed* rather than *stochastic*.

2.1.4 Spatio-temporal Data

Spatial data are often observed repeatedly in time. The temporal component of spatial data is always important in statistical analysis. In many spatio-temporal problems where there is no obvious trend over time, temporal observations are often regarded as replicates in the spatial domain. This kind of set-up, however, is not always realistic. Schabenberger and Gotway (2005) point out that the problem can be tackled by one of the followings:

1. separate spatial analysis for T time points;

2. separate temporal analysis for n locations;

3. spatio-temporal data analysis with methods for random fields in $\mathbb{R}^{d+1}$.

The first two approaches can be considered as conditional methods because the former is obviously T sets of pure spatial problems and the latter is merely

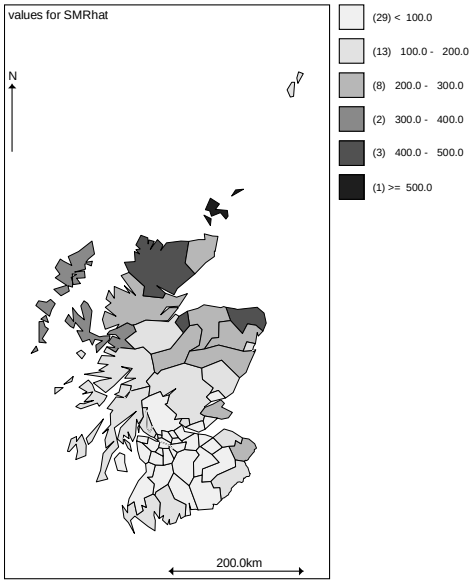


Figure 2.1: An areal representation of the standardised mortality ratios of lip cancer in Scotland.

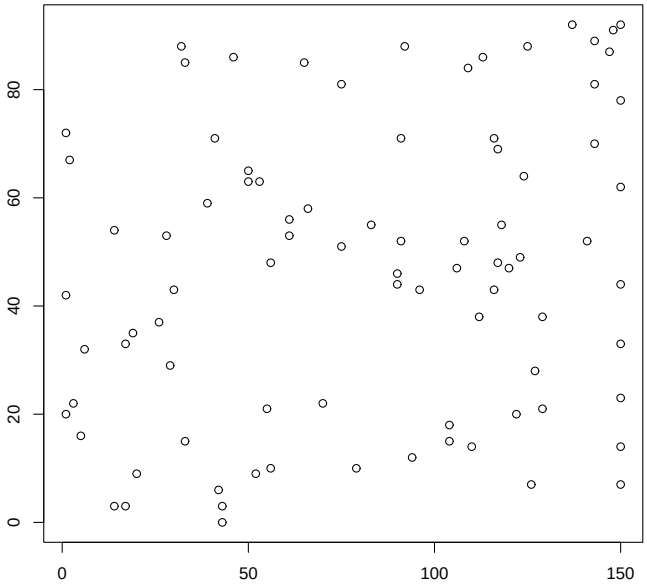


Figure 2.2: Point pattern showing locations of trees found in a forest in New Zealand: the black circle indicates the locations of trees in a plot. The diameter of each tree is also recorded.

n sets of time-series problems. These two approaches are not attractive to modellers and the main difficulties are as follows:

- The observations may contain some missing data and this will in turn complicate the two-stage methods. In some designs, the space-time coordinates may be irregularly distributed and this mis-alignment may result in the failure of the method.
- Separate analysis in space will allow predictions in space only. It is a difficult task to incorporate the space-time interactions.

The third approach is, often, the most preferable one. To express the space-time process, we should expand our previous formulation of a stochastic process in the following way,

$$\left\{Z\left(s,t\right):s\in S\subset\mathbb{R}^{d},t\in T\subset\mathbb{R}\right\}.$$

2.2 Modelling Principles

We first discuss the following important issues in practical modelling of spatially and temporally dependent data.

2.2.1 Modelling Purpose

We construct models in order to explain the reality. The word "explain" can have two separate meanings. In a narrow sense, explanatory power of a model means "explain what we have already known". Whereas, the predictive power of a model means "explain what we do not know yet", e.g., a time series forecasting problem where it is necessary to predict future events. Model calibration and testing scientific hypothesis are generally regarded as explanatory while forecasting and validation problems are always predictive. Prediction can be either interpolative or extrapolative. Some models may be weak in extrapolation but strong in interpolation. This usually happens in non-parametric spline models. Once we make the purpose clear, a good model which satisfies our objectives can be created. Our interest here is mainly on predictions but explanatory ability is also required for a better understanding of the real world.

2.2.2 Model Simplicity

Often, a simpler model is preferred to a more complex one. If, according to our model choice criterion, the performance of two models are about the same, then the simplest model must be chosen. A simpler model would be easier to implement and interpret. By removing unnecessary assumptions, a "simple" model can sometimes be obtained. For a covariance function, non-separability corresponds to the interaction between space and time while non-stationarity corresponds to the local differences in the dependency between different spatial locations.

2.2.3 Speed of Computation

Computation speed is often of paramount importance in many problems. For example, when predicting tomorrow's weather based on today's data it is imperative that the forecasts are issued as easily as possible. A simpler model may allow us to obtain the forecasts quite speedily but forecasts may be somewhat inaccurate.

2.3 Steps in Statistical Modelling

Statistical modelling is the process to generate realisations of real world systems. We will discuss the view of the process of statistical analysis with emphasis on problems motivated by the environment.

2.3.1 Deciding the Modelling Purposes

Le and Zidek (2006, Chapter 5) give a series of possible modelling purposes for environmental modelling such as prediction, hypothesis testing, impact assessment, data summary and knowledge representation. All of these influence our ways of constructing models and methods for model choice. A valid model provides an accurate representation of the phenomenon of interest in here. The word"accurate" is used to denote how well it matches with the modelling purposes. The model can be valid in several ways. Davis (1992) suggests that model validation can be divided into three different types: descriptive, predictive and structural validity.

Descriptive validity means that a model is able to explain the phenomenon. For regression models, measurements like the goodness of fit and the root mean square error may represent how well a model fits the data. In Bayesian context, we can use the deviance information criterion (DIC) (Spiegelhalter *et al.*, 2002) and other criteria based on posterior predictive loss to assess

descriptive validity.

Predictive validity means that a model can predict desired features of a complex system. The predictive performance is often assessed via numerical criteria such as the root mean square error and other measures.

Structural validity means that a model can accommodate the relationship between certain attributes or objects which appear in the real world. Many deterministic models are particularly strong due to their physical nature. For environmental modelling, this kind of model can incorporate the scientific knowledge from physical and chemical processes.

2.3.2 Exploratory Data Analysis (EDA)

EDA often gives us a preliminary idea of whether models are useful. There are well established graphical EDA tools in statistics for postulating suitable models, see for example, Tukey (1977). Tukey (1980) further asserts that EDA should play a major role in statistical modelling and not merely used as a bundle of descriptive statistical tools. A careful exploratory analysis helps to clarify the modelling purposes and identify the key features in the real world systems. The most obvious EDA tool is a map of the data. A map plotting build-in function can be usually found in many statistical programming languages. In addition, in spatial statistics, evaluating estimated variogram plot is an important exploratory work for assessing spatial dependency, see Section 3.1.3 for further details.

2.3.3 Model Specification

Once the scientific question is clear, a set of plausible models can be suggested. For the problems in meteorology, a plausible statistical model should capture some major features of the real world such as space-time dependency, measurement error and missing values of data. From a Bayesian

point of view, in this stage, we have to specify both the likelihood and prior distributions. However, prior knowledge for complex systems is sometimes too difficult to specify subjectively. A natural conjugate prior may be used in this case. Some further discussion on prior specification is provided in Chapter 4.

2.3.4 Model Choice and Validation

A suitable model choice criterion is often used for selecting models from a set of candidate models. Given a finite set of candidate models, a good model selection technique must strike a balance between goodness of fit and penalty on model complexity.

Model validation is a procedure to review the plausibility of the model describing the data. Using an example psychology, Gelman *et al.*, (2004) show that the fitted model captures a general pattern of the data but misses some key features. Thus, checking the goodness of fit and the model assumptions is an important step to judge if the model is adequate for the modelling purposes.

Hodges and Dewar (1992) point out that the standard of quality for a model should be based on its intended uses. The validity of a model is not decided by a yes-no question but by its degree of credibility. Section 2.5 below lists a number of Bayesian model choice criteria.

2.3.5 Inference and Remodelling

A modeller often chooses at least one model using the adopted model choice criteria. Statistical inference using the chosen model is a formal process to make conclusion using data. Usually, estimates of important features of the posterior distribution are provided. Since many models under hierarchical framework are quite complex, the posterior distribution is often approxi-

mated by MCMC methods. From the approximated posterior distribution, point estimations, credible intervals and predictions can be easily evaluated. See Section 2.4 for a brief review.

Often, it is a good idea to review the modelling purposes on the light of the inferences made using the adopted model. Further investigation may also be necessary if a single model does not emerge as the best one according to a multiplicity of model choice criteria. Both of these may suggest re-modelling of the data.

2.4 Bayesian Modelling

All the modelling and analysis work in this thesis are under the Bayesian paradigm. This paradigm is more natural than the traditional frequentist approach and lets us deal with the uncertainty in the model and its parameters. The total uncertainty can be interpreted by a probability distribution. For an environmental application, it is important to evaluate the uncertainty and to give a scientific interpretation using probability statements. For more detailed introduction on Bayesian modelling, see Bernardo and Smith (1994). They provide a theoretical introduction while Banerjee *et al.* (2004) describe a framework on the applications in spatial and spatio-temporal modelling.

2.4.1 Bayesian Inference

For a full Bayesian inference, we first construct a *likelihood* model $f(\boldsymbol{z}|\boldsymbol{\theta})$ for observed data $\boldsymbol{z} = (z_1, \ldots, z_n)$ given unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. For the specification of the unknown parameters, we add a *prior distribution* $\pi(\boldsymbol{\theta})$. Using the Bayes theorem, we obtain the *posterior distribution* as:

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{z}) = \frac{f(\mathbf{z}|\boldsymbol{\theta})\pi(\theta)}{\int f(\mathbf{z}|\boldsymbol{\theta})\pi(\theta)d\boldsymbol{\theta}} \tag{2.1}$$

where the integral in the denominator is over the whole parameter space for $\boldsymbol{\theta}$. The denominator,

$$\pi(\boldsymbol{z}) = \int f(\boldsymbol{z}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{2.2}$$

is known as the marginal likelihood or the prior predictive distribution of data, $\mathbf{z}$ and is a constant free of $\boldsymbol{\theta}$. That is why the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{z})$ is often written as proportional to the product of the likelihood and the prior distribution.

The Bayes theorem can be used to develop complex hierarchical models, see e.g. Wikle (2003). In this set up the likelihood of the data is written as a conditional distribution of data given underlying processes and parameters controlling the underlying process. In the second stage the conditional distribution of the processes is specified given the values of the parameters and in the third stage of the hierarchy a suitable prior distribution for the parameters is specified. The Bayes theorem can then be used to obtain the posterior distribution of the processes and parameters in the following way:

$$\begin{aligned} & \pi(\text{process}, \text{parameters} \mid \text{data}) \\ \propto \quad & f(\text{data} \mid \text{process}, \text{parameters}) \times \pi(\text{process} \mid \text{parameters}) \times \pi(\text{parameters}). \end{aligned}$$

In this hierarchical Bayesian setup, the ability to utilise scientific knowledge and to characterise uncertainty under the three-stage model building is very strong and useful. This can be extended to a fourth stage to account for uncertainties in the hyper-parameters present in the prior distribution of the parameters. In such a case a prior distribution on those hyper-parameters must be specified and included in the above hierarchical specification. The models we develop in the later chapters are all based on this setup of hierarchical Bayesian modelling.

These complex hierarchical models are often analytically intractable and are hard to fit. A numerical integration algorithm such as the MCMC algorithms is usually needed in practice to evaluate the posterior distribution for making inference.

2.4.2 Markov Chain Monte Carlo

Recent development in high speed computational facilities enables the possibilities of using more complex models in statistical data analysis. The MCMC methods provide Monte Carlo integration techniques for exploring posterior distributions in Bayesian analysis.

Gibbs Sampler

The Gibbs sampler introduced by Geman and Geman (1984) and developed by Gelfand and Smith (1990) for applied Bayesian statistical modelling enables dependent sampling from non-normalised high-dimensional posterior distributions. The samples are iteratively drawn from lower dimensional full conditional distributions, e.g., $\{\pi(\theta_i \mid \theta_{j\neq i}, \boldsymbol{z}), i = 1, \ldots, k\}$. Starting from an initial value $\boldsymbol{\theta}^{(0)}$, at iteration j, the Gibbs sampler draws:

$$\begin{aligned}
\theta_1^{(j)} &\sim \pi(\theta_1 \mid \theta_2^{(j-1)}, \ldots, \theta_k^{(j-1)}, \boldsymbol{z}) \\
\theta_2^{(j)} &\sim \pi(\theta_2 \mid \theta_1^{(j)}, \theta_3^{(j-1)}, \ldots, \theta_k^{(j-1)}, \boldsymbol{z}) \\
&\vdots \\
\theta_k^{(j)} &\sim \pi(\theta_k \mid \theta_1^{(j)}, \ldots, \theta_{k-1}^{(j)}, \boldsymbol{z}).
\end{aligned}$$

Note that always the most recent value of $\boldsymbol{\theta}$ is used in conditioning. Non-standard conditional distributions can be sampled using adaptive rejection Metropolis sampling proposed by Gilks and Wild (1992). Features of $\pi(\boldsymbol{\theta}|\mathbf{z})$ are estimated by forming suitable averages of $\boldsymbol{\theta}^{(j)}, j = 1, \ldots, L$ for a large value of L. This type of Monte Carlo integration strategy is used throughout the thesis.

Metropolis-Hastings Algorithm

The *raison d'être* of the Metropolis-Hastings algorithm (Metropolis *et al.*, 1953 and Hastings, 1970) is to draw samples from non-standard posterior distributions by rejecting samples obtained from a proposal distribution which

is much easier to sample from. Given a density $\pi(\boldsymbol{\theta}|\mathbf{z})$ that we wish to sample from, a proposal density $q(\boldsymbol{\theta'} \mid \boldsymbol{\theta})$ is chosen, where $\boldsymbol{\theta'}$ denotes the new sample values. The proposal density $q(\boldsymbol{\theta'} \mid \boldsymbol{\theta})$ usually is an easy-to-sample distribution conditional on the present value $\boldsymbol{\theta}$. The acceptance probability of the proposed value is

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta'}) = \min\left\{1, \frac{\pi(\boldsymbol{\theta'}|\mathbf{z})q(\boldsymbol{\theta} \mid \boldsymbol{\theta'})}{\pi(\boldsymbol{\theta}|\mathbf{z})q(\boldsymbol{\theta'} \mid \boldsymbol{\theta})}\right\}. \tag{2.3}$$

The sampling algorithm of every single sampling from a conditional distribution can be summarised as follows:

1. sample a candidate value $\boldsymbol{\theta'}$ from the proposal density $q(\boldsymbol{\theta'} \mid \boldsymbol{\theta})$,
2. obtain the acceptance probability $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}')$ in Equation (2.3),
3. sample a uniform distributed random variable U on $(0, 1)$.
4. if $U < \alpha(\boldsymbol{\theta}, \boldsymbol{\theta'})$ then accept the candidate value $\boldsymbol{\theta'}$ else, assign the present value $\boldsymbol{\theta}$ to the new value.

The Gibbs sampling is a special case of the algorithm which has a zero rejection rate. Therefore a Gibbs sampler can be easily embedded within a Metropolis routine. See for example Gilks *et al.* (1996) for an overview of the related topics.

2.5 Bayesian Model Choice Criteria

The Bayesian model choice criteria defined in this section are all based on the notion of predictive distributions. The prior predictive distribution has been defined as the marginal likelihood in Equation (2.2). The Bayes factor defined below compares the marginal likelihoods for two competing models. Many other Bayesian model choice criteria are based on the posterior predictive

distribution defined by:

$$\pi(\boldsymbol{z}_{\text{rep}}|\boldsymbol{z}) = \int f(\boldsymbol{z}_{\text{rep}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{z})d\boldsymbol{\theta} \tag{2.4}$$

where $\boldsymbol{z}_{\text{rep}}$ is a future replicate of the observed data.

Often, the Bayesian model choice criteria are approximated using samples obtained from the MCMC algorithms, see Appendix A. Let $\boldsymbol{z}_{\text{rep}}^{(j)}$ denote the jth sample from the posterior predictive distribution (2.4), $j = 1, \ldots, L$.

2.5.1 Bayes Factor

A pure Bayesian method for comparing models is to use the Bayes factor. The Bayes factor B_{12} for comparing models M_1 and M_2 with data $\mathbf{z}$ is given by,

$$B_{12} = \frac{\pi(\mathbf{z}|M_1)}{\pi(\mathbf{z}|M_2)} \tag{2.5}$$

where $\pi(\mathbf{z}|M_i)$ is the marginal likelihood for model $M_i, i = 1, 2$ defined by (2.2). The Bayes factor is interpreted by the rounded scale based on Jeffreys (1961), see also Raftery (1996).

B_{12}	$2\log B_{12}$	Evidence for M_1
< 1	< 0	Negative
$1 - 3$	$0 - 2.2$	Not worth more than bare mention
$3 - 20$	$2.2 - 6$	Positive
$20 - 150$	$6 - 10$	Strong
> 150	> 10	Very Strong

There are many methods available for approximating marginal likelihoods for calculating the Bayes factor, see for example, Newton and Raftery (1994),Chib (1995) and Meng and Wong (1996). The Bayes factor, however, is more difficult to compute for large dimensional problems and is not considered any further in this thesis. Instead we turn to the following model choice criteria which is most suitable when the Gaussian distribution is employed at the first stage of a hierarchical Bayesian model.

2.5.2 Predictive Model Choice Criterion

Gelfand and Ghosh (1998) proposed the following model choice criterion based on ideas discussed in Laud and Ibrahim (1995). The predictive model choice criterion (PMCC) is given by:

$$\text{PMCC} = \sum_{i=1}^{n} E(Z_{i,\text{rep}} - z_i)^2 + \sum_{i=1}^{n} Var(Z_{i,\text{rep}}). \tag{2.6}$$

The first term in the above is a goodness of fit term while the second is a penalty term for model complexity. The model with the smallest value of PMCC is selected among the competing models. Thus, to be selected a model must strike a good balance between goodness of fit and model complexity. In practice, PMCC is calculated using samples $\boldsymbol{z}_{i,\text{rep}}^{(j)}, j = 1, \ldots, L$ from the posterior predictive distribution (2.4).

2.5.3 Prediction Quality

For our modelling purpose, we are mostly concerned with the predictive validity of a model due to the forecasting aims of this thesis. Some criteria for assessing prediction quality (see for example, Atkinson and Lloyd, 1998, Moyeed and Papritz, 2002 and Stephenson, 2006) are given below:
Root Mean Square Error: $RMSE = \left\{\frac{1}{m}\sum_{i=1}^{m}\left(\hat{z}_i - z_i\right)^2\right\}^{\frac{1}{2}}$,
Mean Absolute Error: $MAE = \frac{1}{m}\sum_{i=1}^{m}|\hat{z}_i - z_i|$,
Relative Bias: $rBIAS = \frac{\sum_{i=1}^{m}(\hat{z}_i - z_i)}{m\bar{z}}$,
Relative Mean Separation: $rMSEP = \frac{\sum_{i=1}^{m}(\hat{z}_i - z_i)^2}{\sum_{i=1}^{m}(\bar{z}_p - z_i)^2}$,
where m is the total number of observations we want to validate, z_i is the observation value and $\hat{z}_i$ is the prediction value, $\bar{z}$ is the arithmetic mean of the observations, $\bar{z}_p$ is the arithmetic mean of the predictions.

The first two criteria represent the discrepancy between the model predictions and the measurements while the latter two represent the bias between the predictions and the real values which can be either positive or negative.

We can compare our prediction values with the naïve regression, Eta-CMAQ prediction and all the candidate models to see the difference between them. Using these criteria is advantageous since the Eta-CMAQ model is a deterministic model and we are unable to assess its uncertainty directly.

2.6 Methods of Geostatistics

Geostatistics often refers to the techniques proposed by the Ecole des Mines de Paris in Fontainebleau led by statistician Georges Matheron in the 1970s (Chilès and Delfiner, 1999, Preface). The materials we cover in this chapter are related to *geostatistics* and Matheron (1971) gives the following definition:

"Geostatistics are the application of the theory of the regionalised variables to the estimation of mineral deposits (with all that this implies). A regionalised variable $f(\mathbf{s})$ is a function which denotes the value at the spatial point $\mathbf{s}$."

The primary aim of *geostatistics* is to construct a statistical model to explain and predict spatial data. The geostatistical prediction methods are tools for predicting the regionalised variable at a new location from observations. The collection of methods is known as *Kriging*, a term coined by G. Matheron in honour of the South African mining engineer D. G. Krige who documented the technique on estimating ore-grade of gold mine in his masters thesis in 1951. These methods have been generalised as an essential part of *geostatistics*.

2.6.1 Inverse Distance Weighting

Spatial interpolation is essential to produce high resolution maps. In many interpolation methods, the response at a new location is usually assumed as a weighted sum of the known values. A popular but crude method for irregularly-spaced data is the Inverse Distance Weighting (IDW) (Shepard, 1968) method. Using the IDW method, the interpolated value at location $\mathbf{s}'$

is

$$Z(\mathbf{s}') = \frac{\sum_{i=1}^{n} \frac{1}{d_i^p} z(\mathbf{s}_i)}{\sum_{i=1}^{n} \frac{1}{d_i^p}}, \tag{2.7}$$

where p is an arbitrary positive real number, d_i is the distance between the interpolated point, s' and the location of the ith observation, s_i. This method assumes that there is no uncertainty in the observed data. It is obvious that more reliable values can be obtained for the locations near to the observation sites. However, this method becomes problematic when a prediction point is out of the convex polygon bounded by the observation locations.

2.6.2 Simple Kriging

Assume the following model structure with known parameters $\mu(\mathbf{s})$ and $\mathbf{\Sigma}$.

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \omega(\mathbf{s}), \tag{2.8}$$

where $\boldsymbol{\omega} = (\omega(\mathbf{s}_1), \ldots, \omega(\mathbf{s}_n))^T$ has mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$. Elements of this covariance matrix depend on the assumption of covariance function for the field $Z(\mathbf{s})$ and will discussed in detail in Chapter 3. Thus, $E[\boldsymbol{Z}(\mathbf{s})] = \boldsymbol{\mu}(\mathbf{s})$ and $Var[\boldsymbol{Z}(\mathbf{s})] = \mathbf{\Sigma}$. To minimise the mean square error of the prediction of $Z(\mathbf{s}')$, we assume a linear estimator $\hat{Z}(\mathbf{s}') = \lambda_0 + \boldsymbol{\lambda}'\mathbf{Z}(\mathbf{s})$. The mean square error (MSE) can be formulated as

$$E\left[\left(\hat{Z}(\mathbf{s}') - Z(\mathbf{s}')\right)^2\right] = Var[\boldsymbol{\lambda}'\mathbf{Z}(\mathbf{s}) - Z(\mathbf{s}')] + (\lambda_0 + \{\boldsymbol{\lambda}'\boldsymbol{\mu}(\mathbf{s}) - \mu(\mathbf{s}'))\}^2.$$

Obviously this becomes a simple optimisation problem and can be solved assuming $\mathbf{\Sigma}$ to be non-singular. Let $Var[Z(\mathbf{s}')] = \sigma_\epsilon^2$ and $Cov(\boldsymbol{Z}(s), \boldsymbol{Z}(\boldsymbol{s'})) = \boldsymbol{\rho}$. The best linear predictor under the squared-error loss corresponds to:

$$\begin{aligned} \lambda_0 &= \mu(\mathbf{s}') - \boldsymbol{\lambda}'\boldsymbol{\mu}(\mathbf{s}), \\ \boldsymbol{\lambda} &= \mathbf{\Sigma}^{-1}\boldsymbol{\rho}. \end{aligned}$$

The optimal predictor is given by

$$\hat{Z}(\mathbf{s}') = \mu(\mathbf{s}') + \boldsymbol{\rho}'\mathbf{\Sigma}^{-1}(\mathbf{Z}(\mathbf{s}) - \boldsymbol{\mu}(\mathbf{s})).$$

2.6.3 Ordinary Kriging

In ordinary Kriging it is assumed that $\mu(\boldsymbol{s}) = \mu$ for all $\boldsymbol{s}$ in (2.8) where μ is an unknown parameter. The prediction problem minimises:

$$Q = E\left[\left(\hat{Z}\left(\mathbf{s}'\right) - Z\left(\mathbf{s}'\right)\right)^2\right] - 2\kappa\left(\boldsymbol{\lambda}'\mathbf{1} - 1\right). \tag{2.9}$$

where κ is a Lagrange multiplier with respect to the constraint $\sum_{i=1}^{n}\lambda_i = 1$ and $\mathbf{1}$ is a vector with all elements equal to 1. The optimal predictor is given by

$$\hat{Z}\left(\mathbf{s}'\right) = \hat{\mu} + \boldsymbol{\rho}'\boldsymbol{\Sigma}^{-1}\left(\mathbf{Z}\left(\mathbf{s}\right) - \hat{\mu}\mathbf{1}\right),$$

where $\hat{\mu} = \left(\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}\right)^{-1}\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}\left(\mathbf{s}\right)$. Note that $\hat{\mu}$ is free of $\boldsymbol{s}'$.

2.6.4 Universal Kriging

In universal Kriging, we assume a model structure where $\boldsymbol{\Sigma}$ is known and $\boldsymbol{\mu}\left(\mathbf{s}\right)$ is unknown and allow it to vary over space in the linear form $\boldsymbol{\mu}\left(\mathbf{s}\right) = \mathbf{X}\boldsymbol{\beta}$. The model (2.8) is modified to

$$\mathbf{Z}\left(\mathbf{s}\right) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\omega}\left(\mathbf{s}\right). \tag{2.10}$$

Suppose that the best linear predictor is restricted to be in the form of $\hat{Z}\left(\mathbf{s}'\right) = \mathbf{a}'\mathbf{Z}\left(\mathbf{s}\right)$ for unknown values of $\boldsymbol{a}$. The mean-squared prediction error is given by:

$$\begin{aligned} E\left[\left(\mathbf{a}'\mathbf{Z}\left(\mathbf{s}\right) - Z\left(\mathbf{s}'\right)\right)^2\right] &= Var\left[\mathbf{a}'\mathbf{Z}\left(\mathbf{s}\right)\right] + Var\left[Z\left(\mathbf{s}'\right)\right] - 2Cov\left[\mathbf{a}'\mathbf{Z}\left(\mathbf{s}\right), Z\left(\mathbf{s}'\right)\right] \\ &= \mathbf{a}'\boldsymbol{\Sigma}\left(\mathbf{s}\right)\boldsymbol{a} + \sigma_\epsilon^2 - 2\boldsymbol{a}'\boldsymbol{\rho}. \end{aligned}$$

We obtain the optimal value of $\mathbf{a}$ as:

$$\mathbf{a} = \boldsymbol{\Sigma}_X^{-}\boldsymbol{\rho} + \boldsymbol{\Sigma}^{-1}\mathbf{X}\left(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\boldsymbol{x}\left(\mathbf{s}'\right),$$

where $\Sigma_X^{-} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{X}\left\{\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\right\}^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{x}(\mathbf{s}')$ is the vector of covariate values at $\mathbf{s}'$.

The optimal predictor is given by

$$\hat{Z}(\mathbf{s}') = \mathbf{a}'\mathbf{Z}(\mathbf{s}) = \boldsymbol{x}(\mathbf{s}')'\hat{\boldsymbol{\beta}}_{GLS} + \boldsymbol{\rho}'\boldsymbol{\Sigma}^{-1}\left(\mathbf{Z}(\mathbf{s}) - \mathbf{X}\hat{\boldsymbol{\beta}}_{GLS}\right),$$

where $\hat{\beta}_{GLS} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}\boldsymbol{\Sigma}^{-1}\mathbf{Z}(\mathbf{s})$ is the generalised least squares estimator of $\boldsymbol{\beta}$.

2.6.5 Hierarchical Bayesian Kriging

Bayesian version of Kriging has been developed over the years. By now there is already a substantial amount of literature written on this field, see for example, Le and Zidek (1992), Handcock and Stein (1993), Ecker and Gelfand (1997), Banerjee *et al* (2004) and the references therein. Bayesian hierarchical modelling provides a more powerful and flexible framework for both explanatory and predictive inference. First, we work with the simplest form of hierarchical spatial model under the framework of Banerjee *et al.*(2004, Chapter 5) and we call this *Gaussian random effects* (GRE) model. The model is the sum of three components given by:

$$Z(\mathbf{s}_i) = \mu(\mathbf{s}_i) + w(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), \tag{2.11}$$

where $Z(\mathbf{s}_i)$ is the observed data, $\mu(\mathbf{s}_i)$ is the mean function at location $\mathbf{s}_i$, $i = 1, \ldots, n$ and $\epsilon(\boldsymbol{s_i})$ is, from a regression point of view, the error term. From a geostatistical point of view, the residual $Z(\mathbf{s}_i) - \mu(\mathbf{s}_i)$ is partitioned into two pieces: the partial sill $w(\mathbf{s}_i)$ and the nugget effect $\epsilon(\mathbf{s}_i)$. The partial sill vector $\boldsymbol{w} = (w(s_1), \ldots, w(s_n))^T$ is assumed to be normally distributed with a zero mean and a covariance matrix $\sigma^2\boldsymbol{\Sigma}$ independent of the nugget which is independently normally distributed with a zero mean and a variance σ^2_ϵ. The sum $w(\mathbf{s}_i) + \epsilon(\mathbf{s}_i)$ represents the sill which is also distributed as a Gaussian random variable. The meaning of the terms *nugget* and *partial sill* is discussed in Chapter 3 in Section 3.1.3.

Let $\boldsymbol{\theta}$ denote the vector of all the model parameters. The posterior predictive distribution of the observation at an unobserved site $\boldsymbol{s}'$ is given by:

$$f(z(\mathbf{s}') \mid \boldsymbol{z}) = \int f(\mathbf{z}(\mathbf{s}') \mid w(\mathbf{s}'), \boldsymbol{\theta}) \pi(w(\mathbf{s}')|\mathbf{w}, \boldsymbol{\theta}) \pi(\mathbf{w}, \boldsymbol{\theta}|\mathbf{z}) dw(\mathbf{s}') d\mathbf{w} d\boldsymbol{\theta}, \tag{2.12}$$

where the $f(z(s')|w(\mathbf{s}'), \boldsymbol{\theta})$ is the probability density of the observation at an unobserved site given $w(\mathbf{s}')$, and $\boldsymbol{\theta}$; $\pi(w(\mathbf{s}')|\mathbf{w}, \boldsymbol{\theta})$ is the probability density of $w(\mathbf{s}')$ given $\mathbf{w}$, and $\boldsymbol{\theta}$; $\pi(\mathbf{w}, \boldsymbol{\theta}|\mathbf{z})$ is the joint posterior distribution of $\mathbf{w}$ and $\boldsymbol{\theta}$ given $\mathbf{z}$. In general, the distribution (2.12) is analytically intractable. Evaluation of this distribution requires numerical integration algorithms such as the MCMC techniques (Gilks *et al.*, 1996). The hierarchical modelling setup allows us to handle any missing data values that are often found in practical applications.

2.7 Issues in Mapping

Spatial data are often interpreted as geographically referenced and are presented as maps. Cartographer may represent those data on a map together with a valid coordinate system. The earth is three-dimensional. However, we seldom deal with all three dimensions in practice and usually work with the two-dimensional surface. To link the two-dimensional map and the actual surface of the earth, we need to construct a projection mapping between the reality and the imaginary topologies.

2.7.1 Cartography

The world is not a perfect sphere but is an irregular shape which makes it difficult to model precisely. The most typical projection is called the *Mercator projection* which projects a spherical surface to a plane map. This kind of projection distorts the surface very much although they may be visually

user-friendly. On the other hand, a perfect sphere gives a fairly good approximation of the earth. Despite the irregularity of the surface, Maling (1992) points out that it is appropriate to think of the earth as a sphere with its radius varying by about 10 km on either side, with an average value of 6371 km.

Distance is one of the most important measurements for a surface. Intuitively, a projection to a R^2 space would be a very good solution as we can calculate the Euclidean distance between any two spatial points instead of more complicated ones. However, this possibility is precluded by Gauss' Theorema Egregium (Remarkable Theorem in English) which shows the local isometry between R^2 and the reality is impossible. In the nineteenth century in Göttingen, Germany, Karl Friedrich Gauss was trying to draw such a map on a plain paper without any distortion in distance but failed to do so (see details in Montiel and Ros, 2005). The theorem ensures that local isometry is invariant through a constant Gauss curvature, since the Gauss curvatures over the plane and the surface of sphere are not identical, this kind of projection turns out to be impossible to work with.

2.7.2 Metric Space

In a very wide sense, the concept of distance does not only refer to the Euclidean distance. Euclidean distance is a popular approximation in the spatial statistics community due to its mathematical elegance. The distance sometimes measures the shortest path from one spatial location to another. However, such a measurement may not be realistic in some contexts. Constructing a suitable distance function is essential for modellers.

Let S be a non-empty set. A metric is a map $d : S \times S \to [0, \infty)$ which satisfies following properties, $\forall \mathbf{s}_i, \mathbf{s}_j, \mathbf{s}_k \in \mathbf{s}$,

1. $d(\mathbf{s}_i, \mathbf{s}_j) \geq 0$ and $d(\mathbf{s}_i, \mathbf{s}_j) = 0$ if and only if $\mathbf{s}_i = \mathbf{s}_j$.

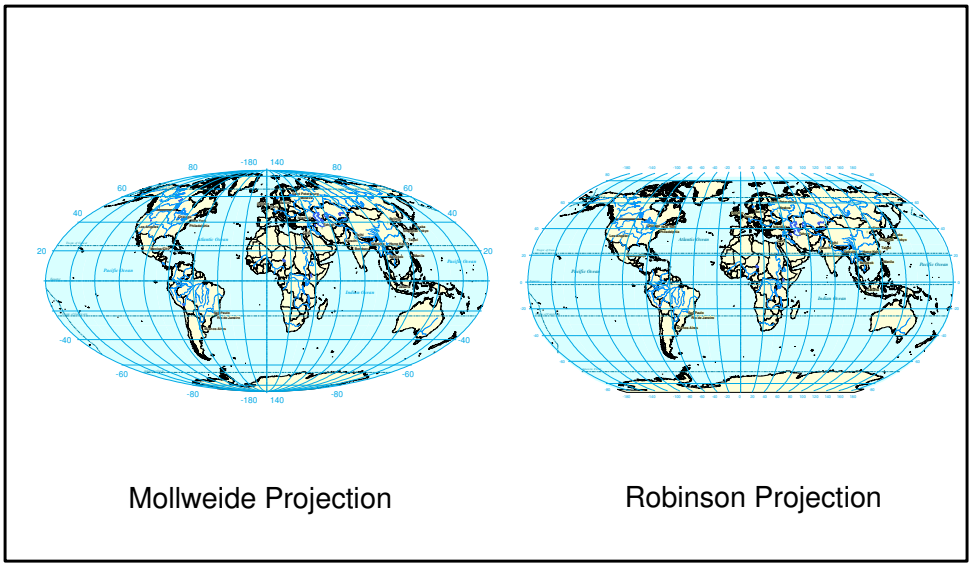


Figure 2.3: Projections may distort the area, distance and angle. (The actual area of south pole in two projections are different.)

2. $d(\mathbf{s}_i, \mathbf{s}_j) = d(\mathbf{s}_j, \mathbf{s}_i)$.

3. $d(\mathbf{s}_i, \mathbf{s}_k) \leq d(\mathbf{s}_i, \mathbf{s}_j) + d(\mathbf{s}_j, \mathbf{s}_k)$.

This gives the distance from location $\mathbf{s}_i$ to $\mathbf{s}_j$. Thus, all functions d satisfying these axioms could be understood as a realisation of the general concept of distance. There are some examples of useful metrics in $\mathbb{R}^2$.

$$\begin{aligned} \|h\|_1 &= |x_1 - x_2| + |y_1 - y_2| && \text{(City Block)} \\ \|h\|_2 &= \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} && \text{(Euclidean)} \quad \text{where } s_i = (x_i, y_i). \\ \|h\|_\infty &= \max(|x_1 - x_2|, |y_1 - y_2|) && \text{(Dominant)} \end{aligned}$$

Curriero (2006) shows that the above metrics are vector norms of $\mathbf{s}_i$ and $\mathbf{s}_j$ and all vector norms metrics are positive definite for the exponential covariance function defined in Section 3.1.6.

2.7.3 Calculating Geodesic Distances

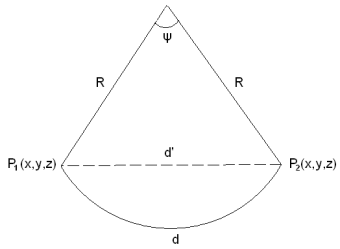


Figure 2.4: Chordal and Geodesic Distances.

Two of the most common coordinate systems used in the ellipsoid earth projection are spherical polar coordinate system and geographical coordinate system. These two systems can be transformed to each other with simple formula: $\lambda = \text{longitude} \times \pi \div 180$ and $\theta = \text{latitude} \times \pi \div 180$.

For any 2 points on the surface, the shortest distance can be formulated by the geodesic path which is a curve along the spherical surface between 2 points.

Using elementary trigonometry, we first consider the three-dimensional Cartesian coordinate system on Euclidean space, see Figure 2.4,

$$(x, y, z) = (R\cos\theta\cos\lambda, R\cos\theta\sin\lambda, R\sin\theta)\,.$$

Then d is the arc distance between two points P_1 and P_2,

$$d = R\cos^{-1}\left(\sin\theta_1\sin\theta_2 + \cos\theta_1\cos\theta_1\cos\left(\lambda_1 - \lambda_2\right)\right).$$

Although no analytical proof has been found to prove the validity of the metric d, the metric is positive definite since this is an approximation of the Euclidean norm in terms of spherical polar coordinate. See Banerjee (2005) for further justifications for using this metric.

2.8 Summary

This chapter has presented many important issues related to spatial data modelling. It also has reviewed a number of key concepts in Bayesian modelling and computation. A number of important methods in geostatistics such as Kriging have been discussed. Some issues in map projection have been discussed and the concept of distance in two and three dimensions has been reviewed. Subsequent chapters will use these concepts to demonstrate the problems that arises in practical environmental monitoring. As for the theoretical side, the hierarchical Bayesian Kriging methods will be further developed by adopting a more complex model structure and we shall remove many simple assumptions regarding the mean and covariance structure that have been made here.

Chapter 3

On Choosing a Space-time Covariance Function

In spatio-temporal modelling, covariance structure of data is often of particular interest. In environmental forecasting applications, capturing spatio-temporal covariance structure is one of the core parts of modelling. A model can be demonstrably wrong but fit for some purposes such as forecasting and interpolation. A number of currently used models for space time covariance function is presented in this chapter.

This chapter is organised as follows. In Section 3.1 we discuss the key concepts regarding covariance functions in space with many examples. We devote Section 3.2 to discuss a number of well known strategies for constructing non-stationary covariance functions. Section 3.3.1 develops covariance functions in space and time. In Section 3.4 we review the literature on hypothesis testing for covariance structure. Section 3.5 illustrates a problem in joint estimation of parameters describing the assumed covariance function and other model parameters for mean and variance. Section 3.6 experiments with a number of models for covariance function for the US EPA ozone concentration data example introduced in Chapter 1. This section chooses the

exponential covariance function that we use heavily in the later chapters. We conclude the chapter with a few summary remarks.

3.1 Covariance Functions in Space

A covariance function $C(\mathbf{s}_i, \mathbf{s}_j) \equiv \text{Cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j))$ describes the dependency of random variables. Sometimes we may write it as $C_s(\mathbf{s}_i, \mathbf{s}_j)$ or $C_s(h)$ where $(h = ||\mathbf{s}_i - \mathbf{s}_j||)$ which gives more emphasis on spatial dependency and the distance. A spatial process, $\boldsymbol{Z}$ is said to be Gaussian, if $\boldsymbol{Z}$ follows a multivariate normal distribution. To formulate the spatial response over a surface, it is usually assumed that the response values from sites which are closer would have larger values of correlation.

3.1.1 Stationarity

We borrow the idea of *stationarity* from the general theory of stochastic processes. The heuristic idea of a stationary spatial process means that the covariance of the responses at two different sites is translational invariant.

Definition 3.1. A process is said to be strictly stationary (also called strongly stationary) if, for any given $n \geq 1$, any set of n sites $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ and for any $\mathbf{h} \in \mathbb{R}^d$,

$$P(Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)) = P(Z(\mathbf{s}_1 + \mathbf{h}), \ldots, Z(\mathbf{s}_n + \mathbf{h}))$$

In practice, it is more useful to define a weaker form of stationarity.

Definition 3.2. A process $Z(\mathbf{s})$ is said to be weakly stationary(also called second-order stationarity) if, $Cov(Z((\mathbf{s}), Z((\mathbf{s} + \mathbf{h})) = C(\mathbf{s} + \mathbf{h}, \mathbf{h}) \equiv C(\mathbf{h})$ for all $\mathbf{h} \in \mathbb{R}^d$ such that $\mathbf{s}$ and $\mathbf{s} + \mathbf{h}$ both lie within S.

It is easy to show that strict stationarity implies weak stationarity. The converse is valid for Gaussian processes only. We adopt the concept of weak stationary in the thesis and the term *stationary* is used to mean this.

3.1.2 Isotropy

A process is said to be *isotropic* if the covariance function depends only on distance which is rotational and translational invariant. Otherwise, the process is called *anisotropic*. The advantage of using *isotropic* covariance function is that this kind of function is easy to formulate with just a set of parametric families of covariance functions of distance only. The representation of an isotropic covariance function can be simplified to

$$C(\mathbf{s}+\boldsymbol{h},\mathbf{s}) = C(h), \tag{3.1}$$

where h is the distance between sites $\mathbf{s}+\boldsymbol{h}$ and $\mathbf{s}$.

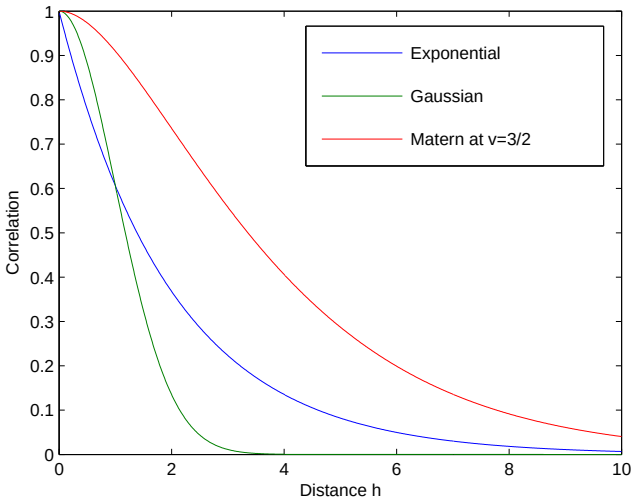


Figure 3.1: Some Covariance Functions.

3.1.3 Variogram

Exploratory Data Analysis (EDA) is usually implemented on the data before we actually formulate the model. Variogram is a useful add-on to those *EDA* tools like the Box plot, Scatter plot and histogram which have been introduced by John Tukey (1977). These tools together give us a better understanding of the behaviour of a spatially dependent stochastic process. The theoretical variogram $\gamma(\mathbf{s}_i, \mathbf{s}_j)$ is defined as:

$$2\gamma(\mathbf{s}_i, \mathbf{s}_j) = E\left[\{Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\}^2\right] \tag{3.2}$$

Sometimes, the function $\gamma(\mathbf{s}_i, \mathbf{s}_j)$ is also called a semi-variogram. For a stationary process, $Z(\mathbf{s})$ we have

$$\begin{aligned} 2\gamma(\mathbf{s}+\mathbf{h}, \mathbf{s}) &= Var(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})) \\ &= Var(Z(\mathbf{s}+\mathbf{h})) + Var(Z(\mathbf{s})) - 2Cov(Z(\mathbf{s}+\mathbf{h}), Z(\mathbf{s})) \\ &= C(\mathbf{0}) + C(\mathbf{0}) - 2C(\mathbf{h}) \\ &= 2[C(\mathbf{0}) - C(\mathbf{h})] \end{aligned}$$

However, the above definition does not include the white noise process. We further narrow down the process to isotropic making it more accessible, i.e.: $\gamma(h) = \gamma(\mathbf{s}+\boldsymbol{h}, \boldsymbol{h})$ where h is the distance between the locations $\mathbf{s}+\mathbf{h}$ and $\mathbf{s}$. The isotropic variogram is empirically estimated by:

$$\hat{\gamma}(h) = \frac{1}{2N_h} \sum_{(i,j)|h_{i,j} \cong h} (z(\mathbf{s}_i) - z(\mathbf{s}_j))^2 \tag{3.3}$$

where N_h represents the discrete number of point combinations that are roughly h distance apart, and $h_{i,j}$ is the distance between two locations $\mathbf{s}_i$ and $\mathbf{s}_j$. The covariance function is usually only of mathematical interest but the variogram offers a clearer picture on the illustrations of the *nugget*, *sill*, *partial sill* and *range*, see Figure 3.2 for an illustration. With $\lim_{h \to 0^+} \gamma(h) = \sigma_\epsilon^2$, the nuggets of Gaussian and exponential functions are about 0.6 unit while for

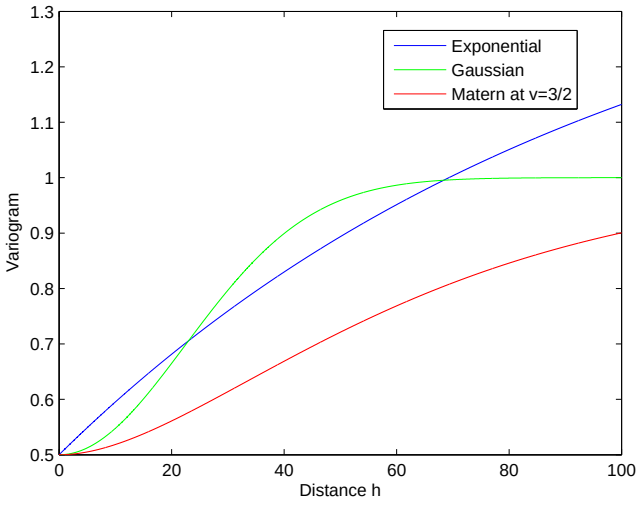


Figure 3.2: Some Variograms.

the Matèrn class example it is 0.5 unit in Figure 3.2. The nugget describes the asymptotic behaviour of the variogram when the distance is close to zero. The sill describes the saturated variogram which can be represented by $\lim_{h\to\infty}\gamma(h) = \sigma_\epsilon^2 + \sigma^2$. The partial sill σ_ϵ^2 is the sill minus the nugget. The range is defined as the minimum value of the distance at which $\gamma(h)$ first reaches the sill.

3.1.4 Positive Definiteness

A covariance function is valid if and only if the function is positive definite, i.e: for any a_i and $a_j \in R$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C\left(\mathbf{s}_i, \mathbf{s}_j\right) \geq 0.$$

Since every valid covariance function is positive definite, the inverse of a covariance matrix can be obtained by the Cholesky decomposition. This is advantageous compared to the method by matrix inversion using the first

principle. The speed of the algorithm is $O(n^3)$ and the accuracy of the results can be well-preserved by making some of the multiplication using double precision.

3.1.5 Spectral Representation and Bochner's Theorem

Sometimes, the covariance function can be represented in a spectral form. The Bochner's theorem (Bochner, 1960) states that, for a measure $F(\omega)$ and a random process in $\mathbb{R}$, the covariance function must be of the form:

$$C(h) = \int \cos(\omega h) dF(\omega). \tag{3.4}$$

This is obviously a Fourier transform. This deduces a corollary that a parametric covariance function must have a corresponding parametric form of spectral density. The spectral density of a covariance function can be further simplified via:

$$f(\omega) = \frac{1}{\pi} \int_0^\infty \cos(\omega h) C(h) dh. \tag{3.5}$$

Similarly,

$$C(h) = \frac{1}{\pi} \int_0^\infty \cos(\omega h) f(\omega) d\omega. \tag{3.6}$$

3.1.6 Some Parametric Covariance Functions

Below we list a set of popular parametric covariance functions often used in practice.

Exponential:

$$C(h) = \begin{cases} \sigma^2 \exp(-\phi h), & h > 0 \\ \sigma^2, & \text{otherwise} \end{cases}$$

Gaussian:

$$C(h) = \begin{cases} \sigma^2 \exp(-\phi h^2), & h > 0 \\ \sigma^2, & \text{otherwise} \end{cases}$$

Matèrn:

$$C(h) = \begin{cases} \sigma^2 \frac{1}{2^{v-1}\Gamma(v)} (\alpha h)^v K_v(\alpha h), & h > 0 \\ \sigma^2 & \text{otherwise} \end{cases}$$

where $\Gamma(.)$ is the usual gamma function and $K_v(.)$ is the modified Bessel function of order v (see, e.g., Abramowitz and Stegun, 1965, Chapter 9).

Spherical:

$$C(h) = \begin{cases} \sigma^2(1 - \frac{3}{2}\frac{h}{\alpha} + \frac{1}{2}(\frac{h}{\alpha})^3), & h \leq \alpha \\ \sigma^2 & \text{otherwise.} \end{cases}$$

Wave:

$$C(h) = \begin{cases} \sigma^2 \frac{\sin(\phi h)}{\phi h}, & h > 0 \\ \sigma^2 & \text{otherwise} \end{cases}$$

Model	Covariance Function $C(h)$	Spectral Density Representation $f(\omega)$
Exponential	$\exp(-\phi h)$	$\frac{\phi}{\pi(\phi^2+\omega^2)}$
Gaussian	$\exp(-\phi h^2)$	$\frac{1}{2\sqrt{\phi\pi}} \exp(-\omega^2/(4\phi))$
Matèrn	$\frac{1}{2^{v-1}\Gamma(v)} (\alpha h)^v K_v(\alpha h)$	$\frac{\alpha^{2v}\Gamma(v+1/2)}{\Gamma(v)\Gamma(1/2)} (\alpha^2+\omega^2)$
Spherical	$(1 - \frac{3}{2}\frac{h}{\alpha} + \frac{1}{2}(\frac{h}{\alpha})^3)$	$\frac{3}{2\pi\alpha^3\omega^4}(\alpha^2\omega^2 - 2\cos(\alpha\omega) - 2\alpha\omega\sin(\alpha\omega) + 2)$
Wave	$\sin(\phi h)/(\phi t)$	$1/(2\phi)$ for $\phi > \omega$

Table 3.1: Table of parametric family of covariance functions and their spectral densities.

In Table 3.1 we can see both the exponential and Gaussian covariance functions have simple parametric forms. The evaluation of spectral density is also straightforward. The Matèrn class model is more important since it includes the special cases of the exponential and Gaussian covariance functions (with $v = 1/2$ and $v \to \infty$ respectively). The spherical and wave covariance functions are not very practically useful and they have complicated spectral densities which are often more difficult to work with.

3.2 Non-stationary Covariance Functions

The covariance functions described in the previous section are all based on the assumption of stationarity. In this section we describe a number of methods for constructing non-stationary covariance functions.

Sampson-Guttorp Method

Sampson and Guttorp (1992) propose an elegant deformation method which projects a stationary imaginary plane (D-field) to the non-stationary real geographical plane (G-field). The approach is non-parametric, see the Bayesian implementation in Schmidt and O'Hagan (2003).

Parametric Non-stationary Covariance Functions

Hughes-Oliver *et al.* (1998) give a parametric point source model which considers the distance of a particular spatial point from the point source $\boldsymbol{c}$,

$$C_s(\mathbf{s}_i, \mathbf{s}_j) = \exp(-\delta_1 h \exp[\delta_2 \mid e_i - e_j \mid + \delta_3 \min(e_i, e_j)]) \tag{3.7}$$

where δ_1, δ_2 and δ_3 are parameters controlling the degree of non-stationarity and e_i is the distance between the point $\mathbf{s}_i$ and the point source $\boldsymbol{c}$. We regard the $\mid e_i - e_j \mid$ as a new metric. A similar approach can be adopted by mixing two covariance functions with two different metrics. However, the application of this covariance function is limited to some point-source pollution problems.

Convolution Method

Higdon *et al.* (1999) propose a convolution method to handle non-stationarity.

A valid covariance function can be obtained by using the convolution:

$$C(\mathbf{s}_i, \mathbf{s}_j; \theta) = \int_D K(\mathbf{s}_i - \mathbf{s}) K(\mathbf{s}_j - \mathbf{s}) C_{\theta(\mathbf{s})}(\mathbf{s}_i, \mathbf{s}_j) d\mathbf{s} \tag{3.8}$$

where $K(\cdot)$ is a kernel function. Epanechnikov (1969) kernel is usually used since this is an optimal kernel for estimation of non-stationarity. Fuentes and Raftery (2005) provide a practical example illustrating this method. However, the bandwidth parameter in $K(\cdot)$ is difficult to decide which is more important than the choice of a kernel (Wand and Jones, 1994).

Mixture of different metric models

Cressie *et al.* (2006) use the weighted mixture of two different covariance functions for Euclidean and stream distance metrics to obtain a non-stationary covariance function. This covariance function incorporates both geographical and hydrological information from the data.

Non-stationary adaptive spectrum

Pintore and Holmes (2004) give a general procedure to construct a non-stationary covariance function via tempering. A positive tempering process function $\eta(\mathbf{s})$ is used to weight the stationary spectrum at location $\mathbf{s}$. The new spectrum is a function of frequency ω and a spatial location $\mathbf{s}$,

$$f_{NS}(\omega, \mathbf{s}) = f_{NS}^{(\mathbf{s})}(\omega) \propto [f(\omega)]^{\eta(\mathbf{s})} \tag{3.9}$$

so that a valid spectrum for a valid covariance function can be formed:

$$f_{NS}(\omega, \mathbf{s}_i, \mathbf{s}_j) = f_{NS}^{(\mathbf{s}_i, \mathbf{s}_j)}(\omega) = f^{(\mathbf{s}_i)}(\omega)^{1/2} f^{(\mathbf{s}_j)}(\omega)^{1/2} \propto [f(\omega)]^{\frac{\eta(\mathbf{s}_i)+\eta(\mathbf{s}_j)}{2}}. \tag{3.10}$$

This formula gives us a possibility to generate a non-parametric covariance function. However, it would be very difficult to obtain Fourier transform integrals for the most common covariance functions by usual numerical integration methods. For example, consider an exponential covariance spectrum with $\phi = 0.01$,

$$\frac{0.01}{\pi(0.01^2 + \omega^2)}, \tag{3.11}$$

then the fourth derivative of an exponential covariance spectrum becomes,

$$f''''(\omega) = \frac{3.84 \times \omega^4}{\pi(0.01^2 + \omega^2)^5} + \frac{2.88 \times \omega^2}{\pi(0.01^2 + \omega^2)^4} + \frac{0.24 \times \omega^4}{\pi(0.01^2 + \omega^2)^3}.$$

Therefore, $\lim_{\omega \to 0} f''''(\omega) \approx 7.64 \times 10^{10}$, which is too large for making a reasonable precision in many numerical integration methods.

3.3 Covariance Functions in Space and Time

The space-time processes discussed in Section 2.1.4 is an example of multivariate spatial data. However, spatial data can also be multivariate without replication in time, for example, ozone and particulate matter concentrations in atmosphere are usually measured at the same monitoring station and highly correlated due to their physical and chemical properties (Le and Zidek, 2006, page 110). Furthermore, a monitoring network measuring multiple pollutants makes data multivariate.

3.3.1 Separable Covariance Functions

Let v_k denote the index for assigning fields at the same spatial location. A covariance function for multivariate data is called *separable* if

$$C\left(Z(\mathbf{s}_i, v_k), Z(\mathbf{s}_j, v_l)\right) = C_s\left(\mathbf{s}_i, \mathbf{s}_j\right) C_v\left(v_k, v_l\right), \tag{3.12}$$

where $Z(\mathbf{s}_i, v_k)$ is the response at $\mathbf{s}_i$ and v_k. The covariance matrix $\boldsymbol{\Sigma}$ of $Z(\mathbf{s}_1, v_1), \ldots, Z(\mathbf{s}_n, v_p)$ can also be represented in matrix form

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_s \otimes \boldsymbol{V}, \tag{3.13}$$

where $\boldsymbol{V}$ is a $p \times p$ matrix capturing inter-field covariance; $\boldsymbol{\Sigma}_s$ is a $n \times n$ matrix capturing spatial covariance, $\otimes$ is the Kronecker product operator. A set of spatio-temporal data, from a mathematical perspective, is also multivariate and makes no difference to a set of spatial data with an extra dimension.

From the definition in (3.12), a separable spatio-temporal covariance function is given by

$$C\left(\mathbf{s}_i, \mathbf{s}_j, t_k, t_l\right) = C_s\left(\mathbf{s}_i, \mathbf{s}_j\right) C_t\left(t_k, t_l\right).$$

This can be viewed as a special case of multivariate separable model in (3.12) and (3.13). The matrix representation is given by

$$\boldsymbol{\Sigma_{s,t}} = \boldsymbol{\Sigma}_s \otimes \boldsymbol{\Sigma}_t, \tag{3.14}$$

where $\boldsymbol{\Sigma}_{s,t}$ denotes the covariance matrix of $Z(\mathbf{s}_1, 1), \ldots, Z(\mathbf{s}_n, T)$; $(\boldsymbol{\Sigma}_s)_{i,j} = C_s\left(\mathbf{s}_i, \mathbf{s}_j\right)$; $(\boldsymbol{\Sigma}_t)_{i,j} = C_t\left(t_i, t_j\right)$. A separable model is advantageous for computation. For example, $\boldsymbol{\Sigma}_{s,t} = \boldsymbol{\Sigma}_s \otimes \boldsymbol{\Sigma}_t$ and $|\boldsymbol{\Sigma}_{s,t}| = |\boldsymbol{\Sigma}_s|^T |\boldsymbol{\Sigma}_t|^n$ can largely simplify the computation by these simple identities (for more details, see in Graham, 1981).

3.3.2 Non-separable Covariance Functions

The positive definiteness of a function is a necessary and sufficient condition for it to be a valid covariance function, see Section 3.1.4. Apart from the conventional family of isotropic separable covariance functions, there are many ways to construct valid models for the covariance function.

As we mentioned earlier in Equation (3.12) a separable covariance function can be formulated by assuming

$$Cov\left(Z\left(\mathbf{s}_i, t_k\right), Z\left(\mathbf{s}_j, t_l\right)\right) = C_s\left(\mathbf{s}_i, \mathbf{s}_j\right) C_t\left(t_k, t_l\right), \tag{3.15}$$

for all possible values of i, j, k and l. Assuming isotropy we can rewrite (3.15) as:

$$C\left(h, u\right) = C_s\left(h\right) C_t\left(u\right) \tag{3.16}$$

where h is the distance of two spatial locations and u is the difference between two time points.

A covariance model which does not obey the above property is called a *non-separable model*. Due to the epistemic uncertainty of the actual process,

separable covariance model is sometimes used even when we know that there exists a non-separable space-time interaction. In a practical sense, a separable covariance model can dramatically reduce the number of parameters and facilitate complicated mathematical representation much more easily. For example, Genton (2007) uses a separable approximation to a non-separable covariance function by Frobenius norm optimisation.

Cressie-Huang Models

Cressie and Huang (1999) have shown a method to construct non-separable stationary covariance functions. Under the conditions C1 and C2 below, the covariance function $C(h,u)$ is valid if

$$C(h,u) = \int e^{-ih\omega}\rho(\omega,u)k(\omega)d\omega \tag{3.17}$$

where

C1. For each $\omega \in \mathbb{R}$, $\rho(\omega,u)$ is a continuous autocorrelation function and $\int \rho(\omega,u)du < \infty$.

C2. $0 < \int k(\omega)d\omega < \infty$.

However, Cressie-Huang models limit the covariance functions into a relatively small class which can be easily covered by other models. We will not use the covariance function generated from this approach but a similar Fourier-transform based method will be used instead.

Gneiting Models

Gneiting (2002) proposed the following general class of valid non-separable, stationary covariance functions for random spatio-temporal processes. The class of covariance function is given by:

$$C\left(h,u\right) = \frac{\sigma^2}{\psi\left(u^2\right)^{\frac{d}{2}}}\varphi\left(\frac{h^2}{\psi\left(u^2\right)}\right) \tag{3.18}$$

where $\varphi\left(.\right)$ denotes a completely monotone function and ψ denotes a function with completely monotone derivative and $\sigma^2 > 0$. Tables 3.2 and 3.3 provide

a set of choices for the functions $\varphi(.)$ and $\psi(.)$ as suggested by Gneiting and Sasvári (1999) and Gneiting (2002).

Mixture Type Non-separable Models

The fact that the sum of two positive definite functions is positive definite allows us to form new valid covariance functions by finite mixtures. A mixture covariance function can define a non-separable covariance function in space and time. For example, Gilleland and Nychka (2006) use a mixture of exponential covariance functions with different decay parameters to capture short range and long range spatial dependencies. A mixture covariance model is a weighted sum of a set of covariance functions and is defined by:

$$C(h) = \sum_{i=1}^{m} w_i C_i(h) \tag{3.19}$$

where the weights $w_1, \ldots, w_m$ are non-negative and $\sum_{i=1}^{m} w_i = 1$; $C_i(h)$ is a valid spatial covariance function. Also, for a spatio-temporal process, a non-separable covariance function can be formed by combining a set of separable covariance functions.

$$C(h, u) = \sum_{i=1}^{m} w_i C_{s,i}(h) C_{t,i}(u) \tag{3.20}$$

where the weights $w_1, \ldots, w_m$ are non-negative and $\sum_{i=1}^{m} w_i = 1$; $C_{s,i}(h)$ is a valid spatial covariance function and $C_{t,i}(u)$ is a valid temporal covariance function.

Function	Parameters
$\varphi(t) = \exp(-ct^{\tau})$	$c > 0$, $0 < \tau \leq 1$
$\varphi(t) = (2^{v-1}\Gamma(v))^{-1}(ct^{1/2})^{v} K_v(ct^{1/2})$	$c > 0$, $v > 0$
$\varphi(t) = (1 + ct^{\tau})^{-v}$	$c > 0$, $0 < \tau \leq 1$, $v > 0$
$\varphi(t) = 2^{v}(\exp(ct^{1/2}) + \exp(-ct^{1/2}))^{-v}$	$c > 0$, $v > 0$

Table 3.2: Some completely monotone functions $\varphi(.)$, $t \geq 0$.

Function	Parameters
$\psi(t) = (at^{\alpha} + 1)^{\delta}$	$c > 0$, $a > 0$, $0 < \alpha \leq 1$, $0 \leq \delta \leq 1$
$\psi(t) = \log(at^{\alpha} + b)/\log(b)$	$a > 0$, $b > 1$, $0 < \alpha \leq 1$
$\psi(t) = (at^{\alpha} + b)/(b(at^{\alpha} + 1))$	$a > 0$, $0 < b \leq 1$, $0 < \alpha \leq 1$

Table 3.3: Some functions $\psi(.)$, $t \geq 0$ with completely monotone derivatives.

Note that the covariance function $C(h, u)$ defined by (3.20) can be non-separable in time and space. De Cesare, Myers and Posa (2001) introduce the following covariance function:

$$C(h, u) = k_1 C_s(h) C_t(u) + k_2 C_s(h) + k_3 C_t(u) \tag{3.21}$$

where k_1, k_2 and k_3 are positive real numbers.

3.4 Hypothesis Tests for Covariance Structure

The hierarchical Bayesian spatio-temporal modelling methods adopted in this thesis rely heavily on some suitable assumptions on space-time covariance functions. Assumptions such as stationarity and separability are widely accepted because of their mathematical simplicity and ability to interpret. To assess the validity of the models, conventional approaches usually formulate problems as a test of hypothesis. A number of such approaches have been suggested by Dutilleul (1999), and Mitchell *et al.* (2005). These methods need replicated space-time data. Fuentes and Raftery (2005) develop a spectral density approach to assess separability. Furthermore, Li *et al.* (2007a, 2007b) develop a hypothesis test for separability based on the asymptotics where no data replication is necessary. Guan *et al.* (2004) propose a non-parametric test for a more specific spatial isotropy. However, all the papers

we mentioned above assume simpler setup without any covariates adjustment.

A specific covariance structure is usually adopted when modelling the space-time variation of a process. However, choosing a suitable covariance structure is sometimes controversial for modellers. Gelman (2007) suggests that the purpose of model checking is not to reject a model (hypothesis) but rather to understand the ways in which it does not fit the data. It would be inappropriate to use a single scalar test statistic to look at the validity of a space-time structure since this is only a part of a full model. A more appropriate approach is to assess the whole model structure with the understanding of our original purposes. For example, rejecting the null hypothesis of separability in a space-time meteorological forecasting problem does not necessarily mean that the meteorological field is intrinsically non-separable, but it may also mean that a wrong model has been used.

3.5 Inconsistent Estimation for Covariance Parameters

Consider the model (2.11) $Z(\mathbf{s}_i) = \mu(\mathbf{s}_i) + w(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), i = 1, \ldots, n$ where the nugget effect, $\epsilon(\mathbf{s}_i) \sim N(0, \sigma_\epsilon^2)$ independent of the spatial random effects $w(\mathbf{s}_i), i = 1, \ldots, n$ which are assumed to follow the normal distributions with zero means and the Matèrn covariance function as given in Section 3.1.6, i.e. $\text{Cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j)) = \sigma^2 \frac{1}{2^{v-1}\Gamma(v)} (\alpha h)^v K_v(\alpha h)$ where h is the distance between $\mathbf{s}_i$ and $\mathbf{s}_j$; α, v, and σ^2 are unknown parameters. Also assume that $\mu(\mathbf{s}_i) = \mu$ for $i = 1, \ldots, n$. Now consider joint estimation of all the five parameters $\mu, \sigma_\epsilon^2, \sigma^2, \alpha$ and v.

Zhang (2004) shows that, regardless of the estimation method used, all five parameters cannot be estimated consistently from observed data. Moreover, Stein (1999) shows that spatial interpolation is sensitive to the product

$\sigma^2\alpha^{2v}$ but not to the parameters individually. In Bayesian inference settings the notion of inconsistent estimation is equivalent to weak identifiability of the parameters under non-informative prior specification, see, for example, Banerjee *et al.* (2008). In practical implementation using Gibbs sampling joint estimation is often poorly behaved due to this weak identifiability and extreme slow-mixing of the associated Markov chains under vague prior distributions for the parameters.

We illustrate the above result using the following simulation example. We simulate $n = 20$ data points from the model (2.11) with $\mu = 5.0, \sigma^2_\epsilon = 0.1, \sigma^2 = 0.5, \alpha = 0.015$ and $v = 0.5$.

We assume the prior distributions: $\mu \sim N(0, 10^4)$, $\sigma^2_\epsilon \sim IG(2, 1)$, $\sigma^2 \sim IG(2, 1)$, $\alpha \sim IG(3, 0.5)$, $v \sim IG(2, 1)$. The proper inverse gamma prior, $IG(2, 1)$ for the variance components avoids the possibility of having an improper posterior distribution and is used throughout this thesis. The 2.5%, 50% and 97.5% quantiles of the assumed inverse gamma distribution are 0.36, 1.66 and 5.50 respectively which covers a reasonable range of sill for the square-root of the ozone values. We have performed several sensitivity studies for the chosen values but the inferential conclusions never changed substantially and hence we do not report those in the thesis.

Figure 3.3 provides the trace plots of all five parameters for the first 10,000 iterations of the Metropolis-Hasting Algorithm. As expected, the plots show very slow mixing as individual parameters make long excursions away from their mean values. In addition, we also provide a scatter plot of the two variance components in Figure 3.4 which shows very high correlation between the sampled values.

Abt and Welch (1998) show that the asymptotic Fisher information matrix for the above five parameters is singular. These problems in estimation are explained in the literature using the notion of microergodicity, see Matheron (1989) who first defined this concept. The book by Stein (1999) gives

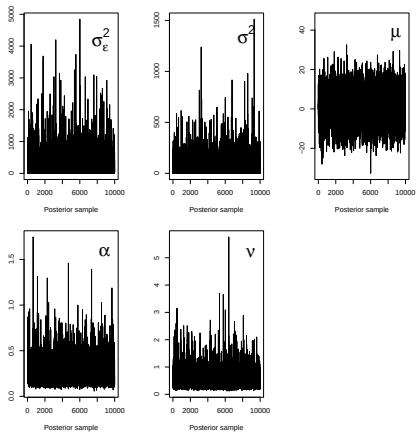


Figure 3.3: MCMC trace plots of all five parameters of a hierarchical spatial model with the Matèrn covariance function.

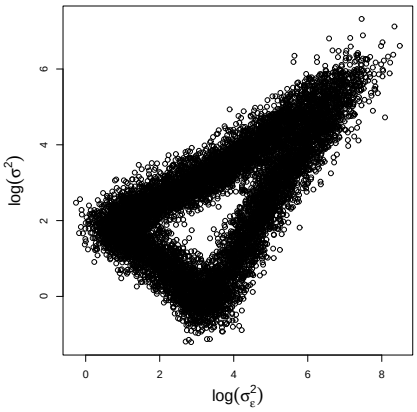


Figure 3.4: A scatter plot of the MCMC samples of the logarithm of two variance components σ^2 and σ^2_ϵ.

a precise mathematical definition of microergodicity for spatial processes. In this thesis, however, we do not investigate these theoretical problems any further; instead, we devote our attention to practical model building for space-time data.

While implementing the Gibbs sampler for practical Bayesian models, we often find that the full conditional distributions for the parameters describing the spatial dependence, for example, α and v above, are not conjugate and sampling those requires expensive likelihood evaluations in each iteration. These difficulties are exacerbated by the large number of locations-time point combinations we work with in this thesis as well as the desire to do spatial prediction over large spatial domains, for example the whole of eastern US in Chapters 4 and 5. For these reasons, we shall choose optimal values of these parameters using a validation mean square error criterion and estimate the variances conditional on those values. We note that this approach falls within the general empirical Bayes (EB) methodology. In the subsequent chapters the adopted EB methodology allows us to use a grid search technique for estimating the parameters describing the covariance function. Subsequent inference methods are conditional on the optimal values; for example, we estimate the variances which have conjugate full conditional distributions under our conjugate prior distribution assumptions.

3.6 US EPA Data Example

We return to the hourly ozone data set observed at 350 monitoring locations over 168 hours as described in Section 1.2. Fitting a spatio-temporal model with a non-separable covariance function will require storage and inversion of matrices of order 58,800 (=350 $\times$168). This is computationally prohibitive especially in our Bayesian setting with the use of MCMC computation algorithms. Therefore, we illustrate with a subset of the original data set to

reduce the computational burden. We take a subset of 20 locations from the full data set, see Figure 3.5. We also consider data from a reduced time window of 24 hours starting from 3pm on 2nd August, 2005. Our aim here is to compare the performances of various models, all based on Equation (3.22), but each with a different covariance function.

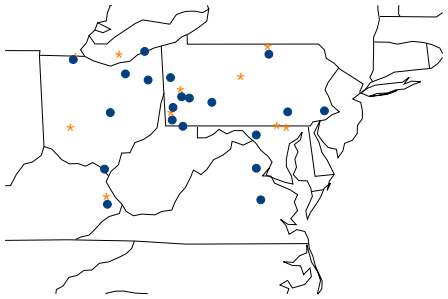


Figure 3.5: Blue circles: 20 sites for fitting; orange asterisks: 10 sites for validation.

3.6.1 Hierarchical Model

Let $z(s_i, t)$ denote the observed square root of the ozone concentration level at site $\mathbf{s}_i$ and at time t. Consider the following hierarchical Bayesian model,

$$Z(s_i, t) = \beta_0 + \beta_1 x(s_i, t) + \omega(s_i, t) + \epsilon(s_i, t) \tag{3.22}$$

where $x(s_i, t)$ is the square root of the Eta-CMAQ model ouput of the grid cell containing the site $\mathbf{s}_i$. The square root transformation is adapted throughout the thesis since it stabilises the variance and encourages normality, see Sahu *et al.* (2007) for further justification.

As in Section 3.5 we assume the $N(0, 10^4)$ distribution for β_0 and β_1 to have a flat prior for these mean parameters. Also following the ratio-

nale noted there we assume the inverse gamma distribution, $IG(2,1)$ for the nugget effect σ_ϵ^2 and σ^2, the variance of the spatio-temporal effect $w(\mathbf{s}_i, t)$.

Apart from the covariance functions mentioned in Table 3.1, we also consider a non-separable covariance function introduced by Gneiting (2002). In particular, we consider the following special case of the general space-time covariance function (3.18) given by:

$$C(h,u) = \frac{\sigma^2}{au^\tau + 1} \exp\left(\frac{-c\,h}{(au^\tau+1)^{\frac{\delta}{2}}}\right), \tag{3.23}$$

where h and u denote the spatial and temporal distance between two locations and two time points, respectively, and a, c, τ and δ are parameters. The parameter $\delta \in [0,1]$ controls the degree of space-time interaction. The special case $\delta = 0$ corresponds to a separable model. As δ increases from 0, the space-time interaction strengthens.

From our discussion in Section 3.5 we note that we cannot estimate all of these parameters together with all the parameters describing the mean function and variances consistently. That is why in our illustration below we assume that the parameters (a, c, τ, δ) are known.

3.6.2 Fitting Empirical Variogram

We use a distance-time auto-correlation plot to examine the covariance structure following Raftery, Haslett and McColl (1982). The idea is to fit a special case of separable model then to generalise it to a more general non-separable model. Haslett and Raftery (1989) used the graph of purely spatial correlation to fit the variograms while Gneiting (2002) fitted the variogram by considering space and time separately. Here we not only fit the variogram by using Equation (3.23) but also consider a number of other classes of variograms. Figure 3.6 shows the empirical spatial variogram and the correlogram of the temporal component from 409 sites and 7 days data. From the

empirical variogram, the appropriate range of the spatial variogram is about 450 kilometres. This choice of the range leads us to the following assumption for the purely spatial correlation function:

$$C(h, 0) = \exp(-0.0066h). \tag{3.24}$$

Thus this choice implies that we take $c = 0.0066$ in (3.23). Similarly, from the temporal correlogram we estimate that $a = 0.11$ and $\tau = 1.47$. Thus, we now have:

$$C(0, u) = \frac{1}{0.11u^{1.47} + 1}. \tag{3.25}$$

With these choices for a, c and τ, the covariance function (3.23) reduces to:

$$C(h, u) = \frac{\sigma^2}{0.11u^{1.47} + 1} \exp\left(\frac{-0.0066h}{(0.11u^{1.47} + 1)^{\frac{\delta}{2}}}\right). \tag{3.26}$$

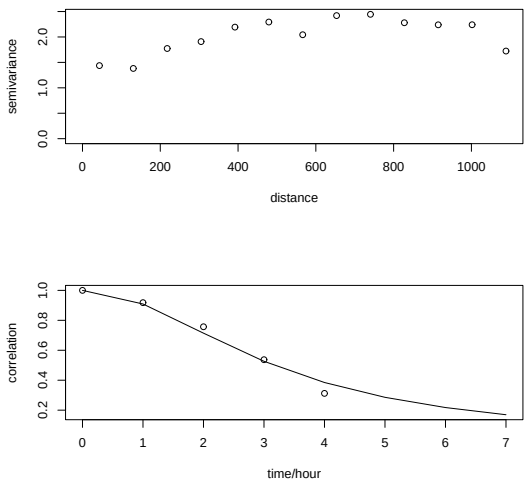


Figure 3.6: Empirical Variogram in spatial domain and Correlogram in time domain

3.6.3 Candidate Models for Space-Time Covariance Function

In this subsection we list a number of possible space-time covariance functions for the space-time process $w(\mathbf{s}_i, t)$ in the hierarchical model (3.22). We consider the covariance function in Equation (3.26) with $\delta = 0$ and $\delta = 1$. We also fit a separable covariance model with the exponential covariance function for both the spatial and temporal components. We call this Separable Model A. We also introduce four non-separable mixture models in Equation (3.20) with different weight parameters w_i. We now consider the following candidate models:

(1) Separable Gneiting model (3.26) with $\delta = 0$.

(2) Non-separable Gneiting model (3.26) with $\delta = 1$.

(3) A separable model: Kronecker product of an exponential covariance function, $e^{-\phi_s h}$ and a temporal covariance function, $e^{-\phi_t u}$ where $\phi_s = 0.0066$ and $\phi_t = 0.43$.

(4) Mixture model A: a mixture of a separable Gneiting model (1) and a separable exponential model (3) above with weights 0.7 and 0.3 respectively.

(5) Mixture model B: a mixture of a separable Gneiting model (1) and a separable exponential model (3) above with weights 0.9 and 0.1 respectively.

(6) Mixture model C: a mixture of a separable Gneiting model (1) and a separable exponential model (3) above with weights 0.5 and 0.5 respectively.

(7) Mixture model D: a mixture of a separable Gneiting model (1) and a separable exponential model (3) above with weights 0.3 and 0.7 respectively.

(8) A separable model: Kronecker product of a Gaussian spatial covariance function, $e^{-\phi_s h^2}$ and an exponential temporal covariance function, $e^{-\phi_t u}$ where $\phi_s = 1.5 \times 10^{-5}$ and $\phi_t = 0.43$.

3.6.4 Results

The model choice criteria for comparing the above models are provided in Table 3.4. The spatio-temporal models outperform the Eta-CMAQ and naïve linear regression predictions as expected. With more parameters, the mixture model *A* gives the best RMSE. The mixture models cannot be separated by multiplication and hence, they are non-separable. The exponential covariance model (3) gives a good prediction result which is slightly less accurate than the mixture model A. Gneiting models of covariance functions (see contour plot in Figure 3.7) and the models (1), (2), and (8) do not provide a better result than the simple separable exponential covariance function in descriptive or predictive senses. A possible explanation for this appears in Huang *et al.* (2007). Their empirical example shows a larger ratio of smoothing parameter σ^2 to nugget parameter σ_ϵ^2 which provides a better fit using a simpler model. A simulation study for checking the predictive performance of non-separable covariance models is also performed for this thesis but the results are found to be inconclusive. Henceforth, we do not use Gneiting's model and, instead set up a separable covariance structure for the regression models for the rest of the thesis instead.

Model	RMSE	MAE	rBIAS	rMSEP$\times 10^{-5}$	PMCC
CMAQ	25.133	18.878	16.610	1.311 $\times 10^5$	
Linear Regression	15.559	12.016	2.915	2.116	4164.5
(1) Gneiting Model, $\delta = 0$	14.797	11.046	-2.136	1.919	130.4
(2) Gneiting Model, $\delta = 1$	14.721	11.350	0.434	1.897	114.4
(3) Separable Model	12.164	9.507	-1.382	1.297	142.6
(4) Mixture Model A	11.948	9.243	-1.364	1.251	82.9
(5) Mixture Model B	12.176	9.200	-2.186	1.300	104.8
(6) Mixture Model C	12.211	9.462	-1.587	1.307	72.6
(7) Mixture Model D	12.497	9.826	0.243	1.367	68.4
(8) Gaussian Model	13.637	10.716	0.088	1.628	935.4

Table 3.4: Performance of all candidate models: the model (4) gives the lowest value of RMSE and MAE while the model (3) gives the second lowest value of RMSE and MAE.

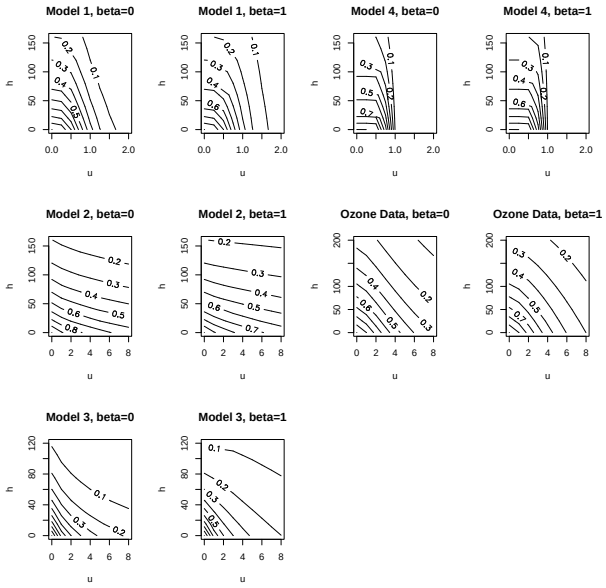


Figure 3.7: Contour plots of some simulated models with the degree of separability parameters $\delta = 0$ and $\delta = 1$. The contour lines denote the correlation. Some models show similar patterns regardless of the choice of the degree of separability parameters.

3.7 Summary

Identifying the modelling purposes is a crucial step in the statistical modelling. These influence the model choice criteria to be adopted for the particular modelling exercise. Covariance functions play a crucial role in space-time data modelling. The parametric covariance functions introduced in Chapter 2 are further generalised to non-separable and non-stationary cases. This chapter illustrates how to choose a space-time covariance function using Bayesian model choice criteria. The methods are illustrated with a real life data set on ozone concentration levels. In this example, the exponential covariance function performs very well when it is used within the GRE hierarchical model. The exploratory and the preliminary work here forms the basis of model construction and remodelling in the latter chapters.

Chapter 4

Interpolating and Forecasting Hourly Ozone Concentration Levels

4.1 Introduction

Real-time air quality information provides the general public a good insight to on the passive respiratory diseases and their prevention measures. Information is often presented in an illustrative rather than descriptive manner for example, using a multi-colour map. Producing an accurate, instantaneous and high resolution map is an easy way to visualise the information, however, we cannot produce such a map only with the measurements since they are often observed in sparse and spatially irregular monitoring networks. A natural choice is to extract the forecast information from high resolution numerical simulation output. However, the classical work by Lorenz (1963) shows that it is impossible to make detailed meteorological forecasts beyond a certain time limit due to extreme sensitivity to the initial conditions of the atmospheric dynamical systems. Thus, the computer simulation outputs

alone are not good enough for drawing the maps. Fusing information from the observed data and numerical simulation output appears to be a possible approach but suitable statistical methods linking two information sources are needed.

There are four possible scenarios in this prediction problem. The first one is a point-wise prediction problem which is methodologically same as the *Kriging* problem introduced in Section 2.6. The second scenario is an areal prediction problem which uses a set of areal data to obtain another set of areal predictions. The third scenario is still the areal prediction problem using observed point-referenced data often called up-scaling. As the fourth scenario a downscaling approach uses areal level data to predict point-referenced ones. The fourth scenario is usually more challenging because information lies in aggregated level data is used for making inference at a higher resolution. We usually fuse areal level data and point-referenced data with different support points to obtain a set of point-referenced prediction. This problem is often called *the change of support problem* in the spatial statistics literature. Several authors have addressed this problem, for example Gelfand *et al.* (2001) propose a unifying approach for prediction in above four scenarios. More recently, Gelfand and Sahu (2009) give a review on recent developments in data fusion.

For this type of applications, Fuentes and Raftery (2005) use Gaussian random fields to jointly model areal level computer model output and point level observed data. The Eta-CMAQ computer model output (see Section 1.2) is treated as linearly biased information for ground truth represented by the observed data. However, in this approach, the measurement errors in the observed data are not taken into account. A similar approach is developed by Zimmerman and Holland (2005). They modelled data from environmental monitoring networks on wet deposition as correlated variables by adopting the technique of *Co-Kriging*. Cowles and Zimmermann (2003)

propose a method to examine the temporal trend of the wet sulfate deposition data from two monitoring networks with different measurement errors, bias and variability. Information from data obtained from two networks have the effect of reducing uncertainty in the predictions.

The remainder of this chapter is organised as follows. In Section 4.2 we use a Gaussian random effects type model with exponential covariance function for modelling the hourly ozone concentration data. We present the models in a hierarchical setting and provide results for an example using a subset of the full data set. This method turns out to be very slow and as a result is not fit for the instantaneous prediction and forecasting problem for the hourly data. Section 4.3 develops an alternative Bayesian model which enables instantaneous forecasting of hourly ozone levels without resorting to MCMC. Section 4.3.3 presents the analytical results for achieving this. We then develop methods for forecasting the current hour's 8-hour average ozone concentration level. We illustrate with the data introduced in Chapter 1 and end the chapter with a few summary remarks.

4.2 A Model with Nugget Effect for Hourly Data

Let $Z_l(\mathbf{s}, t)$ denote the observed square-root ozone concentration, $\sqrt{Q_l(\mathbf{s}, t)}$, at location $\mathbf{s}$ and at hour t $(t = 0, \ldots, 23)$ of day l $(l = 1, \ldots, 7 = r)$ and $O_l(\mathbf{s}, t)$ denote the true value corresponding to $Z_l(\mathbf{s}, t)$. We develop models for data from n stations denoted by $\mathbf{s}_1, \ldots, \mathbf{s}_n$, for a running window of $r = 7$ days and 24 hours $(= T)$.

Further, let $x_l(\mathbf{s}, t)$ denote the square-root of the Eta-CMAQ ozone forecast value at the grid cell which includes location $\mathbf{s}$ and at hour t of day l. The work in Chapter 3 shows that $x_l(\mathbf{s}_i, t)$ can be a good predictor of $Z_l(\mathbf{s}_i, t)$. Figure 1.4 also shows heavy daily cycles in both ozone concentration and

their Eta-CMAQ forecasts.

We model these daily periodicities by including sine and cosine terms. The sine and cosine terms account for the seasonal (hourly) variations in the data. We define the mean function as follows:

$$\mu_l(\mathbf{s}_i, t) = \xi + \beta_1\, x_l(\mathbf{s}_i, t) + \sum_{k=1}^{6}\left[a_k \sin\left(\frac{2\pi tk}{24}\right) + b_k \cos\left(\frac{2\pi tk}{24}\right)\right], \quad (4.1)$$

for $i = 1, \ldots, n$, $t = 0, \ldots, 23$ and $l = 1, \ldots, 7$. Let $\boldsymbol{\beta}$ denote the collection of $p(= 14)$ un-known parameters ξ, β_1 and $a_k, b_k, k = 1, \ldots, 6$ to be estimated from data. Note that sine and cosine series describe the within day peaks and as a result are common for all the days.

We write the mean function (4.1) using the $\boldsymbol{\mu} = X\boldsymbol{\beta}$ where X, is the $nrT \times p$ design matrix and $\boldsymbol{\mu} = (\mu_1(s_1, 1), \mu_1(s_1, 2), \ldots, \mu_r(s_n, T))$ is the vector obtained by concatenating the mean function first by the $T = 24$ hours, then by the $r = 7$ days and then by the n sites.

4.2.1 Hierarchical Spatio-temporal Model

We adopt the hierarchical modelling framework (Section 2.6.5) for spatial and temporal data. The full hierarchical Bayesian modelling formulation includes the nugget term given by

$$Z_l(\mathbf{s}_i, t) = O_l(\mathbf{s}_i, t) + \epsilon_l(\mathbf{s}_i, t), \quad (4.2)$$

where

$$O_l(\mathbf{s}_i, t) = \mu_l(\mathbf{s}_i, t) + w_l(\mathbf{s}_i, t) \quad (4.3)$$

for $i = 1, \ldots, n, l = 1, \ldots, 7,\ t = 0, \ldots, 23$, the $\epsilon_l(\mathbf{s}_i, t)$ are assumed to be independently and identically distributed $N(0, \sigma^2_\epsilon)$ random variables, the space-time process $w_l(\mathbf{s}_i, t)$ is treated as a spatio-temporal random effect. The spatio-temporal process $w_l(\mathbf{s}_i, t)$ is assumed to be a zero-mean process with a separable covariance structure, given by:

$$\mathrm{Cov}\big\{w_l(\mathbf{s}_i, t),\ w_{l'}(\mathbf{s}_j, t')\big\} = \sigma^2_w\ \rho_s(\mathbf{s}_i - \mathbf{s}_j; \phi_s)\ \rho_t(d_{l,t,l',t'}; \phi_t), \quad (4.4)$$

where $d_{l,t,l',t'}$ is the number of hours between the tth hour on day l and t'th hour on day l'. For convenience, the two ρ's are taken to be exponential covariance functions, i.e., $\rho(d;\phi) = \exp(-\phi|d|)$. In Section 3.5 we remarked that the decay parameter ϕ is weakly identifiable. We would treat the ϕ_s and ϕ_t as unknown parameters and estimate them using the prediction quality criteria discussed in Section 2.5.3.

Let $\boldsymbol{w} = (\boldsymbol{w}_{*,1,0}, \ldots, \boldsymbol{w}_{*,7,23})$ and let $H(\boldsymbol{\phi})$ denote the correlation matrix of $\mathbf{w}$ using the covariance function (4.4). The spatio-temporal process $\boldsymbol{w}$ is now assumed to be:

$$\boldsymbol{w} \sim N\left(\mathbf{0}, \sigma_w^2 H(\boldsymbol{\phi})\right). \tag{4.5}$$

Note that this model reduces to the usual regression model with independent errors when we take $H(\phi) = I$. We compare this linear base model with the full spatio-temporal model in Section 4.2.4.

For convenience, we work with the precisions $\tau_\epsilon^2 = \frac{1}{\sigma_\epsilon^2}$ and $\tau_w^2 = \frac{1}{\sigma_w^2}$. The joint prior distribution of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \tau_\epsilon^2, \tau_w^2)$ is given by:

$$\pi(\boldsymbol{\beta}, \tau_\epsilon^2, \tau_w^2) = N\left(\boldsymbol{\beta}_0, \frac{V}{\tau_\epsilon^2}\right) G(a,b) G(a,b), \tag{4.6}$$

where $\boldsymbol{\beta}_0$ and V are suitable hyper-parameters and τ_ϵ^2 and τ_w^2 follow the gamma distribution with mean a/b, independently. We set the hyper-parameters, $\boldsymbol{\beta}_0 = \mathbf{0}$ and $V = 10^4 I$. As justified in Section 3.5 we take $a = 2$, $b = 1$, $A = 10^4$.

4.2.2 Computation Details

The log-likelihood is written as:

$$l(\boldsymbol{\theta}, \mathbf{w}; \mathbf{z}) \propto \frac{nrT}{2} \log(\tau_\epsilon^2) - \frac{\tau_\epsilon^2}{2} (\mathbf{z} - X\boldsymbol{\beta} - \mathbf{w})'(\mathbf{z} - X\boldsymbol{\beta} - \mathbf{w}),$$

where $\mathbf{z}$ denote the vector of all the data points. The log of the joint posterior distribution is given by

$$\begin{aligned}\log\left\{\pi(\boldsymbol{\theta},\mathbf{w},|\mathbf{z})\right\} &\propto \tfrac{nrT}{2}\log(\tau_\epsilon^2) - \tfrac{\tau_\epsilon^2}{2}(\mathbf{z}-X\boldsymbol{\beta}-\mathbf{w})'(\mathbf{z}-X\boldsymbol{\beta}-\mathbf{w})\\ &+ \tfrac{nrT}{2}\log(\tau_w^2) - \tfrac{\tau_w^2}{2}\,\mathbf{w}'H^{-1}(\boldsymbol{\phi})\mathbf{w}\\ &+ \tfrac{p}{2}\log(\tau_\epsilon^2) - \tfrac{\tau_\epsilon^2}{2}\,(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'V^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\\ &+ (a_\epsilon-1)\log(\tau_\epsilon^2) - b_\epsilon\tau_\epsilon^2 + (a_w-1)\log(\tau_w^2) - b_w\tau_w^2.\end{aligned}$$

This model is fitted using a Gibbs sampler. Development of the full conditional distributions with related discussion is provided below. Straightforward calculation yields the following full conditional distributions:

$$\begin{aligned}\tau_\epsilon^2|... &\sim G\left[\tfrac{nrT}{2}+\tfrac{p}{2}+a_\epsilon,\, b_\epsilon+\tfrac{1}{2}(\mathbf{z}-X\boldsymbol{\beta}-\mathbf{w})'(\mathbf{z}-X\boldsymbol{\beta}-\mathbf{w})+\right.\\ &\qquad \left.\tfrac{1}{2}\,(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'V^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right]\\ \tau_w^2|... &\sim G\left[\tfrac{nrT}{2}+a_w,\, b_w+\mathbf{w}'H^{-1}(\phi)\mathbf{w}\right],\\ \boldsymbol{\beta}|... &\sim N\left[V_1\,\left\{X'(\mathbf{z}-\mathbf{w}-\boldsymbol{\beta}_0)+V^{-1}\boldsymbol{\beta}_0\right\},\, V_1\right]\end{aligned}$$

where $V_1 = (V^{-1}+X'X)^{-1}$ and $|...$ is used to denote conditioning on the remaining parameters and observations.

The complete conditional distribution of $w_l(\mathbf{s}_i,t)$ is obtained in blocks as follows. Let $\mathbf{w}_{*j}$ denote the vector of $w_l(\mathbf{s}_j,t)$ for $l=1,\ldots,r, t=1,\ldots,T$. The prior complete conditional distribution of $\mathbf{w}_{*j}$ for $j=1,\ldots,n$ given all other columns $i\neq j, i=1,\ldots,n$ is normal with mean $\boldsymbol{\zeta}_j$ and covariance Λ_j where

$$\boldsymbol{\zeta}_j = -\sum_{i\neq j, i=1}^{n}\frac{(\Sigma_s)_{ij}^{-1}}{(\Sigma_s)_{jj}^{-1}}\mathbf{w}_{*i}, \quad \text{and } \Lambda_j = \sigma_w^2\,\frac{1}{(\Sigma_s)_{jj}^{-1}}\Sigma_t.$$

The likelihood contribution for $\mathbf{w}_{*j}$ is also normal with

$$\text{mean} = \boldsymbol{\xi}_j = \mathbf{z}_{*j} - \boldsymbol{\mu}_{*j}, \text{and covariance} = \sigma_\epsilon^2 I,$$

where I is the identity matrix of appropriate order. The posterior full conditional distribution given the remaining parameters and observations is now seen to be normal with mean

$$\boldsymbol{\chi}_j = \Omega_j\left(\frac{1}{\sigma_\epsilon^2}\boldsymbol{\xi}_j + \Lambda_j^{-1}\boldsymbol{\zeta}_j\right) \quad \text{and covariance } \Omega_j = \left(\frac{I}{\sigma_\epsilon^2}+\Lambda_j^{-1}\right)^{-1}.$$

These are sampled en-bloc in the Gibbs Sampler.

4.2.3 Spatial Interpolation and Forecasting

Using the above models we can interpolate the spatial surface at any time point t' which can be any time in the past or the future. For convenience, use the notation $Z(\mathbf{s}', t')$ to actually denote $Z_{l'}(\mathbf{s}', t')$. A consequence of this is that we shall use the index t to run from 1 to rT rather than $t = 1, \ldots, T$.

For a new location $\mathbf{s}'$ at time t', $Z(\mathbf{s}', t')$ is conditionally independent of $\mathbf{z}$ given $w(\mathbf{s}', t')$ with its distribution given by

$$Z(\mathbf{s}', t') \sim N\left(\mu(\mathbf{s}', t') + w(\mathbf{s}', t'),\ \sigma_\epsilon^2\right), \tag{4.7}$$

according to models (4.2) and (4.3). The posterior predictive distribution (see the general form in Equation (2.12)) of $Z(\mathbf{s}', t')$ is obtained by integrating over the unknown parameters with respect to the joint posterior distribution, that is:

$$\begin{aligned} \pi\left(Z(\mathbf{s}', t')|\mathbf{z}\right) &= \int \pi\left(Z(\mathbf{s}', t')|w(\mathbf{s}', t'), \boldsymbol{\theta}\right)\, \pi\left(w(\mathbf{s}', t')|\mathbf{w}, \boldsymbol{\theta}\right)\, \pi(\boldsymbol{\theta}, \mathbf{w}|\mathbf{z}) \\ &\quad dw(\mathbf{s}', t')\, d\mathbf{w}\, d\boldsymbol{\theta}. \end{aligned} \tag{4.8}$$

When using MCMC methods to draw samples from the posterior, the predictive distribution (4.8) is sampled by composition; draws from the posterior, $\pi(\boldsymbol{\theta}, \mathbf{w}|\mathbf{z})$ enable draws for $w(\mathbf{s}', t')$ (see below for the derivation of $\pi\left(w(\mathbf{s}', t')|\mathbf{w}, \sigma_w^2\right)$), and thus draws for $Z(\mathbf{s}', t')$. To report the predictions on the original scale, we simply work with the square of the predictive realisations drawn from (4.8).

To derive the distribution $\pi\left(w(\mathbf{s}', t')|\mathbf{w}, \sigma_w^2\right)$, note that $\begin{pmatrix} w(\mathbf{s}', t') \\ \mathbf{w} \end{pmatrix} \sim$

$$N\left[\begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix},\ \sigma_w^2 \begin{pmatrix} 1 & \Sigma_s'(\mathbf{s} - \mathbf{s}') \otimes \Sigma_t'(\mathbf{t} - t') \\ \Sigma_s(\mathbf{s} - \mathbf{s}') \otimes \Sigma_t(\mathbf{t} - t') & \Sigma_s \otimes \Sigma_t \end{pmatrix}\right]$$

where $\Sigma_s(\mathbf{s}-\mathbf{s}')$ is an $n \times 1$ column vector with the ith entry given by $\sigma(\mathbf{s}_i - \mathbf{s}') = \rho_s(\mathbf{s}_i - \mathbf{s}';\phi_s)$ and $\Sigma_t(\mathbf{t}-t')$ is a $rT \times 1$ column vector with entries obtained using the temporal correlation function $\rho_t(t-t';\phi_t)$. Hence,

$$w(\mathbf{s}',t')|\mathbf{w} \sim N\left(\sum_{j=1}^{n}\sum_{k=1}^{rT} b_{jk}(\mathbf{s}',t')w(\mathbf{s}_j,k),\ \sigma_w^2 C(\mathbf{s}',t')\right) \tag{4.9}$$

where

$$b_{jk}(\mathbf{s}',t') = \sum_{i=1}^{n}\sum_{m=1}^{rT}\sigma(\mathbf{s}_i-\mathbf{s}')\sigma(m-t')(\Sigma_s)_{ij}^{-1}(\Sigma_t)_{mk}^{-1} \tag{4.10}$$

$$C(\mathbf{s}',t') = 1 - \Sigma_{12}H^{-1}(\phi)\Sigma_{21} \tag{4.11}$$

where Σ_{12} is a $1 \times nT$ matrix with entry $\Sigma'_{12} = (\sigma(\mathbf{s}_1-\mathbf{s}')\sigma(1-t'),\ldots$ $,\sigma(\mathbf{s}_n-\mathbf{s}')\sigma(T-t'))$ and $\Sigma_{21} = \Sigma'_{12}$. The above expressions for $b_{jk}(\mathbf{s}',t')$ and $C(\mathbf{s}',t')$ simplify considerably under the separability assumption, see Section 4.3.3 for further details.

4.2.4 Results

We consider a smaller subset of our full data set to reduce the computational burden. The ozone concentration data are obtained from n = 116 sites in the states of Ohio, Pennsylvania, Maryland, Virginia, West Virginia, Kentucky and Washington D.C from 4th to 10th, August, 2005. Also, we set aside data from 43 sites for validation purposes; these sites are plotted in Figure 4.1.

In addition to addressing the instantaneous forecasting problem for a few hours ahead of the current hour, here we also investigate the capability of the models for forecasting for the next 27 hours so that we can forecast the next day's 8-hour maximum ozone concentration levels. This forecasting problem will be discussed fully in the next chapter where we formally define the daily 8-hour maximum ozone concentration levels. The aim here is to see if this model for hourly data can forecast for such a long lead time in advance.

As mentioned in Section 3.5 the decay parameters ϕ_s and ϕ_t in the model (4.4) cannot be jointly estimated consistently along with other parameters. By using empirical methods (as done in Section 3.6.2) we set the values ϕ_s and ϕ_t equal to 0.01 and 0.3, respectively. These correspond to a spatial range of 300 kilometers and a temporal range of 10hours.

The parameter estimates are presented in Table 4.1. We observe that most of the 95% credible intervals do not include zero; thus most of the parameters are significantly different from zero. We also see a significant positive effect of eta-CMAQ values. The 95% credible intervals for the Fourier series terms a_k and b_k do not both include zero until $k = 6$ implying that a_6 and b_6 can be dropped from the model. The estimates of the variance components show that the spatial random effects explain more variation than the nugget effect.

To compare the performance of the fitted hierarchical spatio-temporal model we also consider a simple linear regression sub-model with independent error distribution. From Table 4.2, we see that the PMCC of the spatio-temporal model is smaller than that for the linear base model. However the RMSE for the spatio-temporal model is only slightly better than that for the linear base model. However, in terms of the RMSE both the linear base model and the spatio-temporal model is better than the eta-CMAQ forecasting model. This fact is further evident in Figure 4.2 which also shows that the eta-CMAQ model produces upwardly biased forecasts.

The spatio-temporal model relies heavily on the Gibbs sampler which takes a long time, over 10 hours in an ordinary duo-processors personal computer, to run even for this small data set. However, this model is not able to improve the forecasts from a basic linear regression model significantly. Thus, an alternative model with less computational complexity and better forecasting ability is needed for obtaining instantaneous forecasts. We develop such a model by removing the nugget effect and improving the mean

structure in the next section.

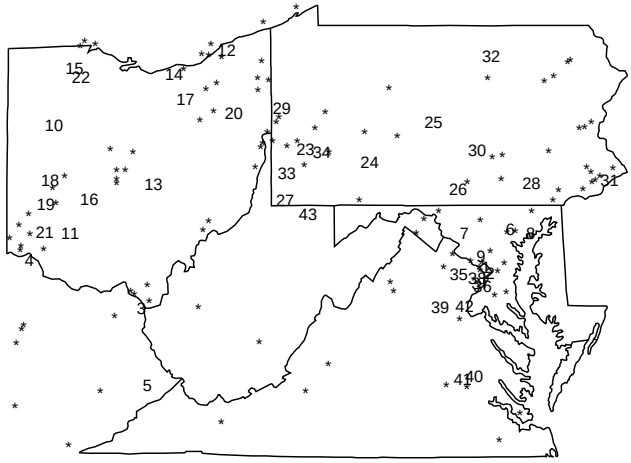


Figure 4.1: Plot of the 116 data sites and 43 validation sites in the study region.

Parameter	Mean	Standard Deviation	2.5 percentile	97.5 percentile
ξ	5.2722	0.1150	5.0717	5.5239
β_1	0.0349	0.0137	0.0052	0.0584
a_1	-0.1902	0.0664	-0.3031	-0.0805
b_1	-2.5279	0.0380	-2.6037	-2.4715
a_2	0.0885	0.0209	0.0379	0.1260
b_2	0.4096	0.0325	0.3436	0.4675
a_3	-0.3163	0.0317	-0.3895	-0.2635
b_3	-0.2285	0.0238	-0.2760	-0.1900
a_4	0.0458	0.0300	0.0023	0.1045
b_4	0.0339	0.0227	-0.0040	0.0821
a_5	-0.0452	0.0178	-0.0783	-0.0115
b_5	0.0452	0.0178	0.0102	0.0763
a_6	-0.0072	0.0165	-0.0426	0.0254
b_6	-0.0247	0.0190	-0.0608	0.0124
σ^2_ϵ	0.0490	0.0025	0.0443	0.0538
σ^2_ω	1.9944	0.0288	1.9370	2.0506

Table 4.1: Parameter estimates of the hierarchical spatio-temporal model in motivating example in Section 4.2.4. Most of the predictive intervals of the parameters do not include zero.

	RMSE	PMCC
Eta-CMAQ	27.53	$--$
Linear base Model	12.80	89051.93
Spatio-Temporal Model	12.67	6430.32

Table 4.2: Table for the comparison of the descriptive and predictive performance for various models.

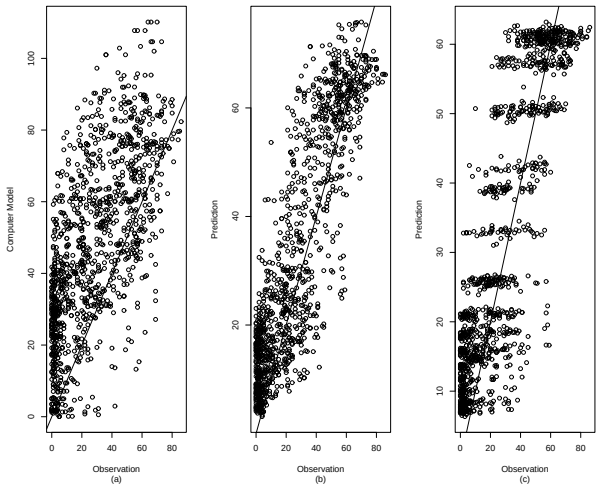


Figure 4.2: Validation plot for the comparison of prediction performance for (a) Eta-CMAQ Model (b) Linear Base Model, and (c) Spatio-Temporal Model. The $y = x$ line is superimposed in all the plots.

4.3 A Regression Model without Nugget Effect

In this section we develop a spatio-temporal model which can be fitted by exact methods. This is important since one of the objectives here is to obtain instantaneous hourly prediction surfaces. We re-formulate the models from Equation (4.2) and (4.3) by removing the nugget effect $\epsilon_l(\boldsymbol{s}, t)$ and developing a new mean function as follows. The full model is given by:

$$Z_l(\mathbf{s}_i, t) = \beta_0\, x_l(\mathbf{s}_i, t) + \xi(t) + w_l(\mathbf{s}_i, t), i = 1, \ldots, n,\ t = 1, \ldots, T, \quad (4.12)$$

where β_0 is an unknown regression co-efficient and $\xi(t)$ is the hourly intercept. We assume that the hourly intercept at any given hour remains constant for different days. The hourly intercept is defined by $\xi(t) = \beta_j$, where the hour $t(= 1, \ldots, T)$ corresponds to the jth hour of the day, $j = 1, \ldots, 24$. This mean structure provides much more direct interpretation using the hourly intercepts that the one given in Equation (4.1) we assumed previously in Section 4.2. This is also preferred by the USEPA and hence the reason for its adoption here. Note that we continue to model ozone concentration levels in the square-root scale.

Note that the model (4.12) is in the form: noisy data equal to the true mean level plus a random error where the true mean level is given by $\beta_0\, x_l(\mathbf{s}_i, t) + \xi(t)$ and the random error term, dependent in space and time, is given by $w_l(\mathbf{s}_i, t)$. Let $\boldsymbol{\beta}$ denote the unknown parameters $(\beta_0, \beta_1, \ldots, \beta_{24})$ and $p = 25$ denote the dimensionality of $\boldsymbol{\beta}$.

As before in Section 4.2 we assume that the spatio-temporal process $w_l(\mathbf{s}_i, t)$ is zero-mean Gaussian process with a separable covariance structure given by Equation (4.4). Thus we assume $\mathbf{w}$ to have the specification given by (4.5). We obtain optimal values of the decay parameters by a grid search as done previously in Section 4.2.1. The exact optimal values are

reported in the results Section 4.3.5.

For convenience, we work with the precision $\tau_w^2 = \frac{1}{\sigma_w^2}$. The joint prior distribution of $\boldsymbol{\beta}$ and τ_w^2 is given by:

$$\pi(\boldsymbol{\beta}, \tau_w^2) = N\left(\boldsymbol{\beta}_0, \frac{V}{\tau_w^2}\right) G(a_w, b_w), \tag{4.13}$$

where $\boldsymbol{\beta}_0$ and $\mathbf{V}$ are suitable hyper-parameters and τ_w^2 follows the gamma distribution with mean a/b. We use the same values of the hyper-parameters as in Section 4.2.1.

4.3.1 Posterior Distributions

The model in (4.12) can be written as

$$\mathbf{Z} \sim N\left(X\boldsymbol{\beta}, \sigma_w^2 H\left(\phi\right)\right).$$

The joint posterior distribution of $\boldsymbol{\beta}$ and τ_w^2, $\pi\left(\boldsymbol{\beta}, \tau_w^2|\mathbf{z}\right)$, is:

$$\begin{aligned}
&\propto (\tau_w^2)^{\frac{nrT+p}{2}+a_w-1} \exp\left[-\tfrac{\tau_w^2}{2}\left(\mathbf{z}-X\boldsymbol{\beta}\right)' H^{-1}\left(\phi\right)\left(\mathbf{z}-X\boldsymbol{\beta}\right)\right. \\
&\quad \left. -\tfrac{\tau_w^2}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'V^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - b_w\tau_w^2\right] \\
&\propto (\tau_w^2)^{\frac{nrT+p}{2}+a_w-1} \exp\left[-\tfrac{\tau_w^2}{2}\left\{\left(\mathbf{z}-X\boldsymbol{\beta}\right)' H^{-1}\left(\phi\right)\left(\mathbf{z}-X\boldsymbol{\beta}\right)\right.\right. \\
&\quad \left.\left. +(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'V^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + 2b_w\right\}\right].
\end{aligned}$$

Now we use the matrix identity:

$$\begin{aligned}
&(\mathbf{z}-X\boldsymbol{\beta})'H^{-1}(\phi)(\mathbf{z}-X\boldsymbol{\beta}) + (\boldsymbol{\beta}-\boldsymbol{\beta}_0)'V^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + 2b_w \\
= \quad & (\boldsymbol{\beta}-\boldsymbol{\beta}^*)'(V^*)^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}^*) + 2b_w^*
\end{aligned}$$

where

$$V^* = \left(V^{-1} + X'H^{-1}(\phi)X\right)^{-1} \tag{4.14}$$

$$\boldsymbol{\beta}^* = V^*\left(V^{-1}\boldsymbol{\beta}_0 + X'H^{-1}(\phi)\mathbf{z}\right) \tag{4.15}$$

$$b_w^* = b_w + \left\{\boldsymbol{\beta}_0'V^{-1}\boldsymbol{\beta}_0 + \mathbf{z}'H^{-1}(\phi)\mathbf{z} - (\boldsymbol{\beta}^*)'(V^*)^{-1}(\boldsymbol{\beta}^*)\right\}/2. \tag{4.16}$$

Hence the joint posterior distribution is given by

$$\pi\left(\boldsymbol{\beta}, \tau_w^2|\mathbf{z}\right) \propto \left(\tau_w^2\right)^{\frac{nrT+p}{2}+a_w-1} \exp\left[-\frac{\tau_w^2}{2}\left\{(\boldsymbol{\beta}-\boldsymbol{\beta}^*)'(V^*)^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)+2b_w^*\right\}\right].$$

Now we have the following results:

$$\begin{aligned}
\boldsymbol{\beta}|\mathbf{z}, \tau_w^2 &\sim N\left(\boldsymbol{\beta}^*, \sigma_w^2 V^*\right) && (4.17)\\
\tau_w^2|\mathbf{z}, \boldsymbol{\beta} &\sim G\left(\frac{nrT+p}{2}+a_w, \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)'(V^*)^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)+b_w^*\right) && (4.18)\\
\boldsymbol{\beta}|\mathbf{z} &\sim St\left(\boldsymbol{\beta}^*, 2b_w^*\frac{V^*}{nrT+2a_w}, nrT+2a_w\right) && (4.19)\\
\tau_w^2|\mathbf{z} &\sim G\left(nrT/2+a_w, b_w^*\right) && (4.20)
\end{aligned}$$

where $\mathbf{Y} \sim St(\boldsymbol{\mu}, \Sigma, \nu)$ has the density

$$f(\mathbf{y}|\boldsymbol{\mu}, \Sigma, \nu) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{p/2}}|\Sigma|^{-1/2}\left\{1+\frac{(\mathbf{y}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})}{\nu}\right\}^{-(\nu+p)/2}.$$

4.3.2 Predictive Distributions

Using the above models we can interpolate the spatial surface at any time point t' which can be any time in the past or the future. As before, for notational convenience, we drop the sub-script l' for day, i.e. we use the notation $Z(\mathbf{s}', t')$ to actually denote $Z_{l'}(\mathbf{s}', t')$. Let the regression vector at this new location-time combination be given by $\mathbf{x}_0$. We construct the joint distribution:

$$\begin{pmatrix} Z\left(s', t'\right) \\ \mathbf{Z} \end{pmatrix} \sim N\left\{\begin{pmatrix} \mathbf{x}_0'\boldsymbol{\beta} \\ X\boldsymbol{\beta} \end{pmatrix}, \sigma_w^2 \begin{pmatrix} 1 & \Sigma_{12} \\ \Sigma_{21} & H\left(\phi\right) \end{pmatrix}\right\},$$

where Σ_{12} and Σ_{21} are obtained appropriately using the covariance function (4.4). Now we obtain the conditional distribution

$$Z\left(s', t'\right)|\mathbf{z}, \boldsymbol{\beta}, \sigma_w^2 \sim N\left\{\mathbf{x}_0'\boldsymbol{\beta}+\Sigma_{12}H^{-1}\left(\phi\right)\left(\mathbf{z}-X\boldsymbol{\beta}\right), \sigma_w^2\left(1-\Sigma_{12}H^{-1}\left(\phi\right)\Sigma_{21}\right)\right\}.$$

Therefore,

$$\begin{aligned}\pi(Z(s',t')\,|\mathbf{z},\boldsymbol{\beta},\sigma_w^2) &\propto (\tau_w^2)^{1/2}\exp\Big[-\tfrac{\tau_w^2}{2m}\{z(s',t')\\ &\quad -\mathbf{x}_0'\boldsymbol{\beta}-\Sigma_{12}H^{-1}(\phi)(\mathbf{z}-X\boldsymbol{\beta})\}^2\Big]\\ &\propto (\tau_w^2)^{1/2}\exp\Big[-\tfrac{\tau_w^2}{2m}\{z(s',t')-\Sigma_{12}H^{-1}(\phi)\,\mathbf{z}-\mathbf{x}_0'\beta\\ &\quad +\Sigma_{12}H^{-1}(\phi)\,X\boldsymbol{\beta}\}^2\Big]\\ &\propto (\tau_w^2)^{1/2}\exp\Big[-\tfrac{\tau_w^2}{2m}\{z^*-(\mathbf{x}_0'-\Sigma_{12}H^{-1}(\phi)\,X)\boldsymbol{\beta}\}^2\Big]\\ &\propto (\tau_w^2)^{1/2}\exp\Big[-\tfrac{\tau_w^2}{2m}\{z^*-\mathbf{g}'\boldsymbol{\beta}\}^2\Big]\end{aligned}$$

where

$$\begin{aligned}m &= (1-\Sigma_{12}H^{-1}(\phi)\,\Sigma_{21})\\ z^* &= z(s',t')-\Sigma_{12}H^{-1}(\phi)\,\mathbf{z}\\ \mathbf{g}' &= \mathbf{x}_0'-\Sigma_{12}H^{-1}(\phi)\,X.\end{aligned}$$

This shows that

$$Z^*|\mathbf{z},\boldsymbol{\beta},\tau_w^2 \sim N\left(\mathbf{g}'\boldsymbol{\beta},\sigma_w^2\,m\right).$$

But we have already seen that

$$\boldsymbol{\beta}|\mathbf{z},\tau_w^2 \sim N\left(\boldsymbol{\beta}^*,\,\sigma_w^2\,V^*\right).$$

Hence by integrating out $\boldsymbol{\beta}$ we have

$$Z^*|\mathbf{z},\tau_w^2 \sim N\left(\mathbf{g}'\boldsymbol{\beta}^*,\sigma_w^2(m+\mathbf{g}'V^*\mathbf{g})\right).$$

Now since the posterior distribution of τ_w^2 is $G(nrT/2+a_w,b_w^*)$ the posterior predictive distribution of Z^* given $\mathbf{z}$ is:

$$Z^*|\mathbf{z} \sim St(\mathbf{g}'\boldsymbol{\beta}^*,2b_w^*\frac{1-\Sigma_{12}H^{-1}(\phi)\,\Sigma_{21}+\mathbf{g}'V^*\mathbf{g}}{nrT+2a_w},nrT+2a_w).$$

Now we obtain the posterior predictive distribution $Z(\mathbf{s}',t')|\mathbf{z}$ as follows:

$$\begin{aligned}Z(\mathbf{s}',t')|\mathbf{z} \sim St\Big(&\mathbf{x}_0'\boldsymbol{\beta}^*+\Sigma_{12}H^{-1}(\phi)\,(\mathbf{z}-X\boldsymbol{\beta}^*),\\ &2b_w^*\frac{1-\Sigma_{12}H^{-1}(\phi)\,\Sigma_{21}+\mathbf{g}'V^*\mathbf{g}}{nrT+2a_w},nrT+2a_w\Big).\end{aligned}\tag{4.21}$$

The posterior predictive distribution we are using here is on the square-root scale. We can predict on the original scale by evaluating below simple equality:

$$\begin{aligned} E(Z^2(\mathbf{s}',t') \mid \mathbf{z}) &= \left\{E(Z(\mathbf{s}',t') \mid \mathbf{z})\right\}^2 + Var\left\{E(Z(\mathbf{s}',t') \mid \mathbf{z})\right\} \\ &= \left\{\mathbf{x}_0'\boldsymbol{\beta}^* + \Sigma_{12}H^{-1}(\phi)(\mathbf{z} - X\boldsymbol{\beta}^*)\right\}^2 + \\ &\quad 2b_w^* \tfrac{1-\Sigma_{12}H^{-1}(\phi)\Sigma_{21}+\mathbf{g}'V^*\mathbf{g}}{nrT+2a_w-2} \end{aligned}$$

The variance of these predictions can be calculated using the fourth order of moment of the t-distribution. From Equation (4.21), let $\mu_{z*} = \mathbf{x}_0'\boldsymbol{\beta}^* + \Sigma_{12}H^{-1}(\phi)(\mathbf{z} - X\boldsymbol{\beta}^*)$, $\sigma_{z*}^2 = 2b_w^* \frac{1-\Sigma_{12}H^{-1}(\phi)\Sigma_{21}+\mathbf{g}'V^*\mathbf{g}}{nrT+2a_w}$ and $\nu_{z*} = nrT + 2a_w$. The variance of the response in original scale Z^2 is

$$Var(Z^2) = \frac{2\sigma_{z*}^2\nu_{z*}^2(\nu_{z*}-1)}{(\nu_{z*}-4)(\nu_{z*}-2)} + 8\sigma_{z*}^2\mu_{z*}\frac{\nu_{z*}}{\nu_{z*}-2}. \tag{4.22}$$

Although this can be used to assess uncertainty in the predictions, this exact formula for variance of Z^2 is not sufficient to construct prediction intervals since the distribution of the square of a non-central t is not a standard distribution. Besides, this type of exact formula cannot be derived if some other transformation, e.g. the logarithm had been used instead of the square-root used here. Hence there is a need for looking alternative methods for evaluating the uncertainty.

Apart from the above exact formula for variance of the predictions, there are three alternative methods of assessing uncertainty in the predictions using approximations. The first is a Monte Carlo method. We generate a large number of samples of $Z^{(j)^2}(\mathbf{s}',t'), j = 1,\ldots,1000$ from the posterior predictive distribution and use those to estimate the prediction variance and the prediction intervals.

This approach, however, will be slower than the second method based on the normal approximation for the square of the t-distribution (4.21) we adopt here. The approximation is justified by the fact that the degrees of freedom $nT + 2a_w$ is very large (more than 2500 in our application). The

approximate 95% prediction interval is given by

$$E(Z^2(\mathbf{s}', t')|\mathbf{z}) \pm 1.96 \times \sqrt{\mathrm{Var}(Z^2(\mathbf{s}', t')|\mathbf{z})}.$$

The third and final one is an approximation method using the well-known delta method (Oehlert, 1992), any continuous transformation $g(X)$'s variance can be approximated by $Var(X)g'(E(X))^2$. This method is quicker than the Monte Carlo method.

4.3.3 Simplifying the Computation

Spatio-temporal modelling often involves high dimension matrices. It takes unnecessary long time to compute those matrices by the first principles and also require storage of huge matrices. Analytical simplification of the expressions $\Sigma_{12}H^{-1}(\phi)$ and $\Sigma_{12}H^{-1}(\phi)\Sigma_{21}$ is possible because of the assumption of separability. As a result the computation becomes much faster.

Note that

$$\begin{pmatrix} 1 & \Sigma_{12} \\ \Sigma_{21} & H(\phi) \end{pmatrix} = \begin{pmatrix} 1 & \Sigma_s'(\mathbf{s}-\mathbf{s}') \otimes \Sigma_t'(\mathbf{t}-t') \\ \Sigma_s(\mathbf{s}-\mathbf{s}') \otimes \Sigma_t(\mathbf{t}-t') & \Sigma_s \otimes \Sigma_t \end{pmatrix}$$

where $\Sigma_s(\mathbf{s}-\mathbf{s}')$ is an $n \times 1$ column vector with the ith entry given by $\sigma(\mathbf{s}_i - \mathbf{s}') = \rho_s(\mathbf{s}_i - \mathbf{s}'; \phi_s)$ and $\Sigma_t(\mathbf{t}-t')$ is a $rT \times 1$ column vector with entries obtained using the temporal correlation function $\rho_t(t-t'; \phi_t)$.

Here $H^{-1}(\phi) = \Sigma_s^{-1} \otimes \Sigma_t^{-1}$. Hence the $1 \times nrT$ vector $\Sigma_{12}H^{-1}(\phi)$ will have elements (for $j = 1, \ldots, n$ and $k = 1, \ldots, rT$)

$$\begin{aligned} b_{jk}(\mathbf{s}', t') &= \sum_{i=1}^{n}\sum_{m=1}^{rT} \sigma(\mathbf{s}_i - \mathbf{s}')\sigma(m-t')(\Sigma_s)_{ij}^{-1}(\Sigma_t)_{mk}^{-1} \\ &= \sum_{i=1}^{n} \sigma(\mathbf{s}_i - \mathbf{s}')(\Sigma_s)_{ij}^{-1} \sum_{m=1}^{rT} \sigma(m-t')(\Sigma_t)_{mk}^{-1} \\ &= b_s(j, \mathbf{s}')\; b_t(k, t'), \end{aligned} \tag{4.23}$$

where

$$b_s(j,\mathbf{s}') = \sum_{i=1}^{n} \sigma(\mathbf{s}_i - \mathbf{s}')(\Sigma_s)^{-1}_{ij}, \text{ and } b_t(k,t') = \sum_{m=1}^{rT} \sigma(m-t')(\Sigma_t)^{-1}_{mk}.$$

The quantity $b_t(k,t')$ can be simplified considerably by noting that it resembles the inner product of a multiple of a particular column of Σ_t and a particular row of Σ_t^{-1}. First, consider the case $t' \leq rT$. In this case $b_t(k,t')$ is the inner product of the t'th column of Σ_t and kth row of Σ_t^{-1}. Hence $b_t(k,t')$ will be 1 if $t' = k$ and 0 otherwise. Now consider the case $t' > rT$. Suppose that we can write

$$\sigma(m-t') = \sigma(t'-rT)\sigma(rT-m)$$

for $m = 1,\ldots,rT$, thus $b_t(k,t')$ will be $\sigma(t'-rT)$ times the inner product of the rTth column of Σ_t and kth row of Σ_t^{-1}. (The above identity holds for the adopted exponential covariance function.) Thus we have proved the following results:

$$b_t(k,t') = \begin{cases} \delta_{k,t'}, & \text{if } t' \leq rT \\ \delta_{k,rT} \exp\left[-\phi_t(t'-rT)\right], & \text{if } t' > rT \end{cases}$$

where $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise.

Now we simplify the expression for the conditional variance. For the exponential covariance function, this provided a very convenient simplified form for $b_l(k,k')$; there is no need to perform any summation at all. Let

$$\begin{aligned} C(\mathbf{s}',t') &= 1 - \Sigma_{12} H^{-1}(\phi)\,\Sigma_{21} \\ &= 1 - \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{m=1}^{rT}\sum_{k=1}^{rT} \sigma(\mathbf{s}_i - \mathbf{s}')\sigma(m-t') \qquad (4.24) \\ &\quad (\Sigma_s)^{-1}_{ij}(\Sigma_t)^{-1}_{mk}\sigma(\mathbf{s}_j - \mathbf{s}')\sigma(k-t') \\ &= 1 - a_s(\mathbf{s}')\; a_t(t'), \qquad (4.25) \end{aligned}$$

where

$$a_s(\mathbf{s}') = \sum_{i=1}^{n}\sum_{j=1}^{n} \sigma(\mathbf{s}_i - \mathbf{s}')\left(\Sigma_s^{-1}\right)_{ij}$$

and

$$a_t(t') = \sum_{m=1}^{rT}\sum_{k=1}^{rT} \sigma\left(m - t'\right)\left(\Sigma_t^{-1}\right)_{mk}\sigma\left(k - t'\right) = \sum_{k=1}^{rT} b_t(k, t')\sigma\left(k - t'\right).$$

By substituting the values of $b_t(k, t')$ we obtain:

$$a_t\left(t'\right) = \begin{cases} 1, & \text{if } t' \leq rT \\ \exp\left[-2\phi_t\left(t' - rT\right)\right], & \text{if } t' > rT. \end{cases}$$

Now we have the following results:

$$C(\mathbf{s}', t') = \begin{cases} 1 - a_s(\mathbf{s}'), & \text{if } t' \leq rT \\ 1 - a_s(\mathbf{s}')\,\exp\left[-2\phi_t\left(t' - rT\right)\right], & \text{if } t' > rT. \end{cases}$$

This greatly simplifies the expression $C(\mathbf{s}', t')$ involving four different summations. Now a sum over only two indices is all that is required to evaluate $C(\mathbf{s}', t')$.

Apart from giving an equation for each spatial and time point, the expression $\Sigma_{12}H^{-1}\left(\phi\right)$ can also be represented in a matrix form. For some statistical programming languages, the calculation could be done via matrix multiplication. Consider the following identity (see Graham, 1981):

$$(ABC)^S = (C^T \otimes A)B^S \tag{4.26}$$

where S is a stack operator for a $m \times n$ matrix with comprising m dimension $n \times 1$ vectors. If A and C are positive definite, we have the following:

$$\begin{aligned} & (ABC)^S &&= (C^T \otimes A)B^S \\ \Rightarrow\ & ((ABC)^S)^T &&= (B^S)^T(C^T \otimes A)^T \\ \Rightarrow\ & ((ABC)^S)^T &&= (B^S)^T(C \otimes A^T) \end{aligned}$$

The right hand side of the identity is in the form of $\Sigma_{12}H^{-1}\left(\phi\right)$, so the expression $\Sigma_{12}H^{-1}\left(\phi\right)$ can be written as:

$$\Sigma_{12}H^{-1}\left(\phi\right) = ((\Sigma_t^{-1}(\Sigma_{12}^{-S})_{t\times s}\Sigma_s^{-1})^S)^T \tag{4.27}$$

where $-S$ is an inverse of stack operator that forms a $1 \times mn$ matrix back into a $n \times m$ matrix. The running time is originally $O(n^2m^2)$ and now is reduced to $O(n^2m)$ for $n > m$. For the spatio-temporal models we fitted, the Big O is reduced by $rT = 168$ times.

4.3.4 Predicting the Eight-hour Map

One of the useful applications of the hourly modelling is the ability to predict the 8-hour average ozone concentration at the current hour. In the EPA AIRNow environment air quality standard, the 8-hour average ozone concentration at the current hour t is the simple average of the 8-hourly concentrations at the current hour t, four past hours $(t-1, t-2, t-3, t-4)$, and the three future hours $(t+1, t+2$, and $t+3)$. The 8-hour ozone at time t, location $\mathbf{s}'$ is given by:

$$Q_8(\mathbf{s}, t) = \frac{1}{8} \sum_{k=-4}^{3} Z^2(\mathbf{s}, t+k). \tag{4.28}$$

The 8-hour averages are centred at the middle of 8 hours, for example, the 8-hour average at 4 pm is the average value obtained from the eight one-hourly measurements observed from 12 pm to 7 pm.

Here we use Z^2 since ozone is modelled in the square-root scale. Note that at any un-observed site $\mathbf{s}'$, $Z^2(\mathbf{s}', t)$ for any t is the square of the non-central t-distribution with parameters as given in (4.21). The posterior predictive distribution of $O_8(\mathbf{s}', t)$, defined as the sum of the non-central F-distributed random variables, is not available in closed form. As a result, we use Monte Carlo simulation to find the mean and standard deviation of the posterior predictive distribution of $O_8(\mathbf{s}', t)$ given the observed data $\mathbf{z}$ as follows. We generate a large number L of independent random variables, $Z^{(j)}(\mathbf{s}', t+k), j = 1, \ldots, B$ for each $k = -4, -3, \ldots, 3$ at each hour t at the given location $\mathbf{s}'$. Now we obtain $O_8^{(j)}(\mathbf{s}', t) = \frac{1}{8} \sum_{k=-4}^{3} Z^{(j)^2}(\mathbf{s}', t+k)$ for each $j = 1, \ldots, B$. The 8-hour

average is estimated by the sample mean, $\bar{O}_8(\mathbf{s}', t) = L^{-1} \sum_{j=1}^{L} O_8^{(j)}(\mathbf{s}', t)$ and the sample standard deviation of $O_8^{(j)}(\mathbf{s}', t)$ is used as an uncertainty estimate of the posterior predictive distribution. In the Monte Carlo simulation, we use independent samples $Z^{(j)^2}(\mathbf{s}', t+k)$, $k = -4, \ldots, 3$ for each j. In effect, we perform marginal predictions of $Z^2(\mathbf{s}', t+k)$ for each k, just as we do marginal predictions at all the different locations $\mathbf{s}'$ in the predictive grid of 3000 sites. Joint predictions and forecasting is computationally prohibitive in the instantaneous prediction problem of this chapter and are not pursued here.

4.3.5 Results

We use $n = 350$ sites real-time hourly ozone concentrations data covering the eastern U.S for a two-week period, 2nd-14th August, 2005. We set aside 40 additional for model validation. We only include data for 168 hours in our model: a simple linear regression result shows that the more distant past data do not improve the prediction. Using data hourly data from 2nd-15th August we form 30 data sets each having a start day between 2nd-7th August (giving six possible start days) and a start hour between 2PM-6PM (giving five possible start hours) to examine the prediction performance for repeated data sets.

Apart from validation, we also predict high resolution map for visualising the inferential output. We use the Eta-CMAQ forecast for 3000 randomly sampled grid cells out of available 9119 grid cells. As discussed in Chapter 3, over 90% of high ozone concentrations occurs between 2pm and 6pm. Therefore, our scientific interests would only concentrate on that period.

We will first examine the 3-hours-ahead forecasts. All the model based forecasts outperform the Eta-CMAQ forecasts. Validation plots in Figure 4.3 and 4.4 give the best and the worst prediction in 30 sets of the validation data.

The optimal/decay parameters are found out using a grid search. The grid search procedure is to select a set of parameters from choosing the one with the smallest RMSE. The optimal decay parameters ϕ_s and ϕ_t vary over the eight hours under the RMSE criterion. Thus, it is not worthwhile to use the same set of parameters for all hours. Alternatively, we can use the optimal decay parameters under the RMSE criterion to generate eight-hour ozone level, see Table 4.5 for an illustration. Most of the parameter estimates of ϕ_s and ϕ_t from 1-to-3-hours-ahead forecasts are the same. For the interpolation side, for those 5-hours, ϕ_s tend to have a larger value (shorter range). We use Monte Carlo simulation to find the eight-hour ozone level estimator $\hat{Q}_8(\mathbf{s}', t)$ at time t and location $\mathbf{s}'$. The predictive quality can be checked by the validation mean-square error at time t which is given by,

$$\text{RMSE(t)} = \left[\frac{1}{n_{v'}} \sum_{i=1}^{40} \left(Q_8(\mathbf{s}'_i, t) - \hat{Q}_8(\mathbf{s}'_i, t) \right)^2 I(Q_8(\mathbf{s}'_i, t)) \right]^{\frac{1}{2}} \tag{4.29}$$

where $I(Q(\mathbf{s}'_i, t)) = 1$ if $O_8(\mathbf{s}'_i, t)$ is available, and 0 otherwise, and $n_{v'} = \sum_{i=1}^{40} I(Q(\mathbf{s}'_i, t))$ is the total number of available observations at the 40 validation sites. In terms of the eight-hour scale RMSE, the model based method outperforms both the Eta-CMAQ forecast and the linear base model as well. For example, at 3pm on August 11, the RMSE for Eta-CMAQ and linear model is 24.61 and 8.99 respectively but only 4.82 for our model, see Figure 4.6 for the validation plot.

The prediction map in both Figures 4.8 and 4.9 also show an agreement to the superimposed observed data. From Figure 4.5, we observe that there is not much discrepancy between the Monte Carlo and delta method computations for the standard deviations.

Note that the 8-hour average predictions have lower uncertainties than the 3-hour ahead forecasts, as expected. A usual linear regression model is also examined which has no spatial correlation term. In both Figures 4.7 and 4.10, the linear regression model fails to exhibit the spatial variation in the

map and the predictive values do not show any spatial variation.

We also note that RMSE tends to increase as the length of the forecast period is increased. Table 4.6 shows uniform reductions in mean-square error that result in using the proposed Bayesian spatial-temporal model in relate to the regression model.

4.4 Summary

This Chapter has demonstrated that, as a data assimilation problem, probability forecasts of instantaneous short-term air quality information can be obtained from our model based method. In the example in Section 4.2.4, the hierarchical Bayesian spatio-temporal model shows a better descriptive power but fails to provide a better predictive performance than that of a simple linear regression model.

High resolution prediction map can be produced using Bayesian methods without an extensive MCMC computation. For scientific interpretation, probability statements from the forecast output is easier to address than using physical numerical model. The validation analyses in Section 4.3.5 show that the model based approaches can consistently outperform the computer simulation model.

One disadvantage for using the hierarchical models of this chapter is their weakness in forecasting far ahead in time. Hence we abandon the hope of predicting daily ozone concentration levels using a model for hourly data. In the next Chapter, we incorporate an auto-regressive space-time model for daily data for forecasting the next day's 8-hour maximum ozone concentration levels.

	RMSE	MAE	rBIAS	rMSEP
CMAQ	13.11	10.13	0.05855	0.6168
Linear	12.37	10.05	0.06989	0.5417
Bayes	10.98	9.16	0.06588	0.4289

Table 4.3: Predictive quality indicators for 3-hours ahead forecast at 2pm on 9th August.

	RMSE	MAE	rBIAS	rMSEP
CMAQ	15.46	11.73	0.05211	0.6417
Linear	19.36	16.73	-0.1919	0.7961
Bayes	13.19	10.48	-0.03318	0.4776

Table 4.4: Predictive quality indicators for 3-hours ahead forecast at 2pm on 12th August.

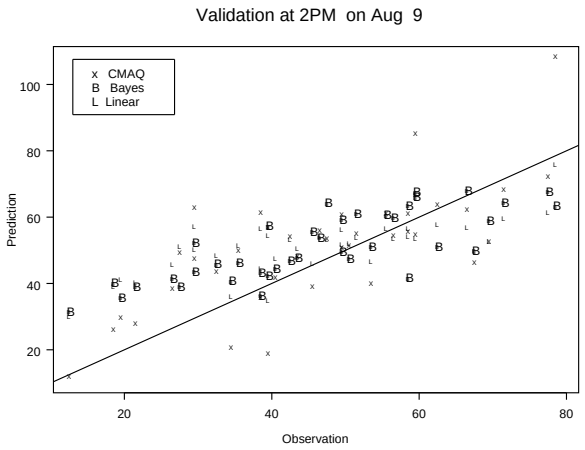


Figure 4.3: Validation plot for 2PM on 9th August when the proposed model performs the best.

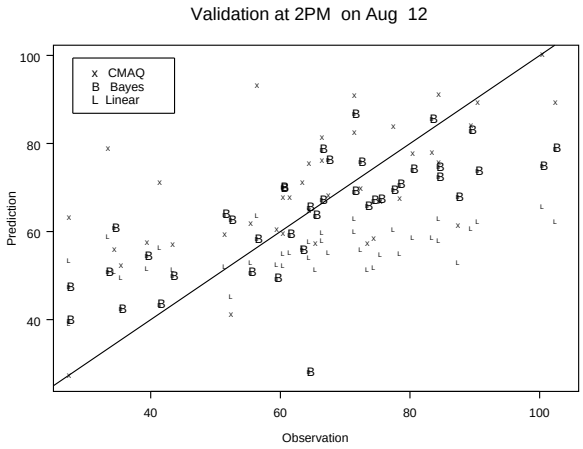


Figure 4.4: Validation plot for 2PM on 12th August when the proposed model performs the worst.

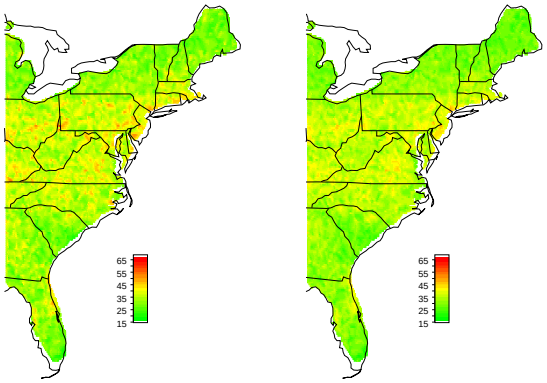


Figure 4.5: Left panel is the standard deviation map produced by Monte Carlo method for model-based 3-hours forecasts at 3pm on 11th August and right panel is for the same standard deviation map generated by delta method.

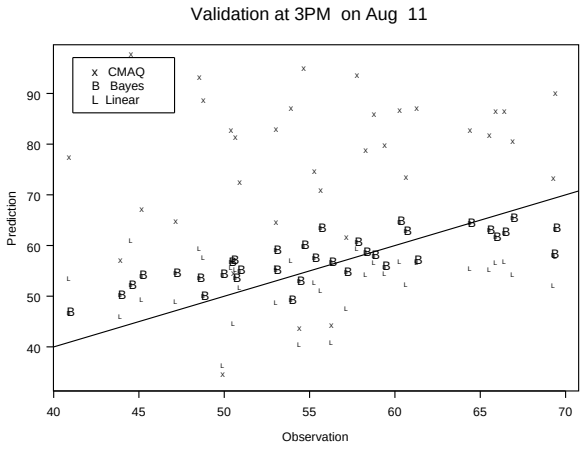


Figure 4.6: Validation plot of 8-hour average at 3pm on 11th August

.

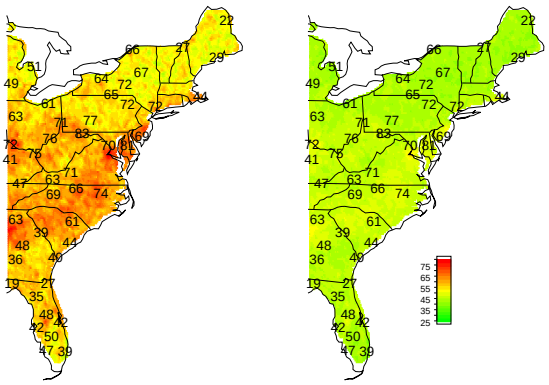


Figure 4.7: Left panel is the 8-hour average Eta-CMAQ map at 3pm on 11th August and right panel is the same map using an independent error regression model. Observed values from some selected sites are superimposed. (For visual clarity we present only a subset of the monitoring data).

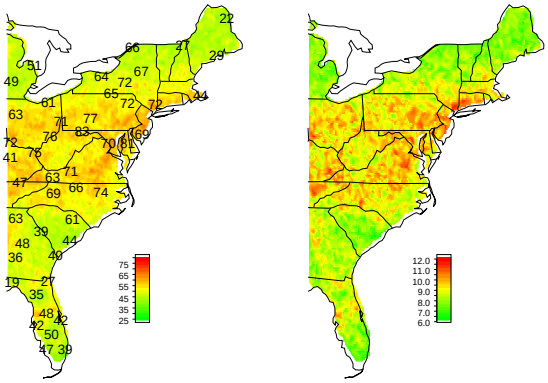


Figure 4.8: Left panel is the 8-hour average model based map at 3pm on 11th August and right panel is the standard deviation map. Observed values from some selected sites are superimposed. (For visual clarity we present only a subset of the monitoring data).

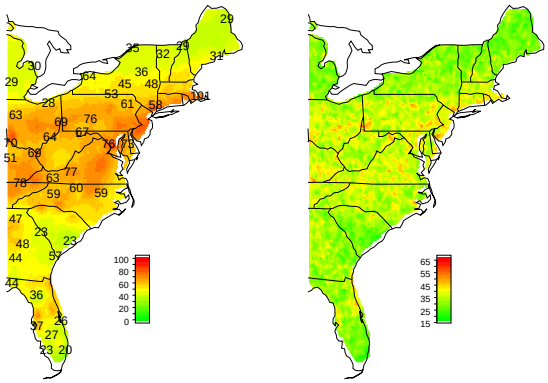


Figure 4.9: Left panel is the 3-hours ahead model based map at 3pm on 11th August and right panel is the standard deviation map. Observed values from some selected sites are superimposed. (For visual clarity we present only a subset of the monitoring data).

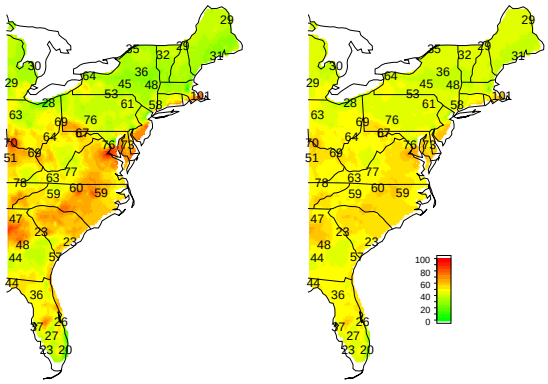


Figure 4.10: Left panel is for the 3-hours ahead Eta-CMAQ map at 3pm on 11th August and right panel is for the same map using an independent error regression model. Observed values from some selected sites are superimposed. (For visual clarity we present only a subset of the monitoring data).

Hour	11am	12pm	1pm	2pm	3pm	4pm	5pm	6pm
ϕ_s	0.06	0.012	0.06	0.06	0.012	0.006	0.006	0.006
ϕ_t	0.13	0.25	1.00	1.00	0.5	0.13	0.13	0.13

Table 4.5: Optimal decay parameters for 8-hour average at 3pm on August 11th.

Day	Start Hour				
	2pm	3pm	4pm	5pm	6pm
Last hour interpolation					
Aug 2	124.7	124.8	141.0	103.0	84.3
Aug 3	98.2	104.9	115.6	83.4	105.5
Aug 4	94.1	99.6	118.6	141.4	192.4
Aug 5	35.5	83.8	48.6	92.0	194.8
Aug 6	117.4	82.1	65.1	74.2	148.0
Aug 7	80.9	104.3	131.2	114.8	141.2
One hour ahead forecasts					
Aug 2	80.0	76.9	63.7	86.4	137.6
Aug 3	43.4	50.6	53.9	81.5	129.4
Aug 4	117.5	144.4	206.1	282.3	327.8
Aug 5	170.5	172.4	190.1	231.4	257.7
Aug 6	118.5	125.0	150.6	160.4	150.7
Aug 7	74.1	81.5	103.6	115.6	126.7
Two hours ahead forecasts					
Aug 2	88.1	71.3	40.4	82.7	148.3
Aug 3	58.9	58.2	53.5	63.8	138.6
Aug 4	119.9	137.1	199.1	298.7	375.1
Aug 5	167.8	180.5	227.3	235.2	267.0
Aug 6	104.4	117.0	171.3	190.3	141.6
Aug 7	80.2	79.1	96.6	143.2	133.4
Three hours ahead forecasts					
Aug 2	44.5	40.3	72.0	125.1	168.1
Aug 3	10.4	20.7	38.2	102.1	157.3
Aug 4	109.2	164.6	267.0	339.3	277.6
Aug 5	161.7	205.3	191.4	233.9	240.3
Aug 6	117.4	172.5	182.2	127.9	114.1
Aug 7	58.8	68.9	135.3	114.7	87.8

Table 4.6: Differences in mean square errors

Chapter 5

Bayesian Fusion for Daily 8-hour Maximum Ozone Concentration Levels

5.1 Introduction

The last Chapter developed methodologies for predicting a real-time 8-hour ozone concentration map at every hour. In this Chapter, we turn to the direct problem of modelling the daily maximum 8-hour average ozone concentration levels since it is necessary to have the forecasts for these one-day in advance. The definition of the daily 8-hour maximum in the last Chapter given by

$$Q_{8,\max}(\mathbf{s}) = \max_{t=0}^{23} Q_8(\mathbf{s}, t)$$

where $Q_8(\mathbf{s}, t)$ has been defined in Equation (4.28) as the average of the 8-successive hours' ozone concentration levels.

In this chapter we develop models for $Q_{8,\max}(\mathbf{s})$ for a number of different sites and days. For convenience, we shall use the $Z(\mathbf{s}_i, t)$ notation to denote the square-root of the daily 8-hour maximum ozone concentration level, $Q_{8,\max}(\mathbf{s})$ for a particular day, t. Similarly, we define $x(\mathbf{s}_i, t)$ to be the daily

8-hour maximum eta-CMAQ forecast for a grid cell containing the site $\mathbf{s}_i$ on day t.

Modelling of the daily data instead of the hourly data greatly reduces the computation problem. The daily model can be developed to be a more complex realistic model. We can also use spatially non-stationary time series models such as an auto-regressive model (ARM) for capturing both spatial variation and temporal evolution.

The plan for the remainder of this chapter is follows. In Section 5.2 we re-postulate the hierarchical Bayesian model from Section 4.2 for daily data. Section 5.3 develops the ARM and compares its performance with the previous model. The chapter ends with a few summary remarks.

5.2 A Gaussian Random Effect Model

The Bayesian hierarchical GRE models take the spatio-temporal random effects into account and assume a hierarchical structure

$$Z(\mathbf{s}_i, t) = \mu(\mathbf{s}_i, t) + \epsilon(\mathbf{s}_i, t), \tag{5.1}$$

$$\mu(\mathbf{s}_i, t) = \beta_0 + \beta_1 x(\mathbf{s}_i, t) + w(\mathbf{s}_i, t), \tag{5.2}$$

The time scale used here is $T = 1, \ldots, 7$ days. The nugget effect $\epsilon(\mathbf{s}_i, t)$ is assumed to follow the normal distribution with zero mean and variance σ_ϵ^2 and $\mathbf{w} \sim N(0, \sigma_w^2 \Sigma)$. We assume Σ to be a separable correlation matrix given by $\Sigma = \Sigma_s \otimes \Sigma_t$, where $\Sigma_{s(ij)} = \rho(d(\mathbf{s}_i, \mathbf{s}_j); \phi_s)$, $\Sigma_{t_k,t_l} = \rho(|t_k - t_l|; \phi_t)$, where $\rho(d; \phi) = \exp(-\phi d)$. We assume the following prior distribution:

$$\sigma_\epsilon^2 \sim IG(a, b)\ ,\ \sigma_w^2 \sim IG(a, b)\ ,\ \beta \sim N(0, A^2),$$

where $IG(a, b)$ denotes the inverse gamma distribution. We take $a = 2, b = 1$ and $A^2 = 10^4$ following the justification provided in Section 3.5.

The data model (5.1) represents the measurement error process. Equation (5.2) represents the process model which is obtained by a linear adjustment of computer model output that varies according to space and time.

5.2.1 Results

The original data set, described in Section 1.2, is transformed into a daily scale by calculating daily 8-hour maximum observed concentration levels and eta-CMAQ forecasts for each site. We model data from $n = 350$ sites and set aside data from 40 randomly selected sites for validation purposes.

We model the daily 8-hour maximum ozone concentration levels for a running window of $T = 7$ days and forecast one-day-ahead. The daily 8-hour maximum ozone concentration levels during 10th-13th August can be forecasted from the original data. In particular, we can forecast the levels for August 10th by modelling 7-days' data from August 3rd-9th and so on. We do this to repeatedly test out the methodologies to be adopted for the real problem.

For producing high resolution forecasting map, we obtain the Eta-CMAQ forecasts for 3000 randomly sampled grid points out of the total 9119 grid points in the eastern US. However, all the available data should be used in order to obtain a more accurate and a higher resolution result.

As discussed in Section 3.5, we fix the decay parameters $\boldsymbol{\phi}$ at optimal values by searching among candidate values in a grid. The candidate values correspond to range values of 50km, 250km, 500km in space and 1-day, 3-days, and 7-days in time. The values corresponding to 250km and 3-days provide the smallest RMSE.

From Table 5.1, all the one-day ahead forecasts derived from the hierarchical GRE models outperform their Eta-CMAQ counterparts in terms of RMSE. For the model bias, all rBIAS in Bayesian hierarchical model is smaller except for the one on the 10th which is slightly worse than Eta-

CMAQ forecasts. Overall, the Bayesian hierarchical model for the period 10th-13th August provides better prediction.

Figure 5.1 shows the validation plot for the forecasts based on Eta-CMAQ and GRE model. It shows that the Eta-CMAQ forecasts are upwardly biased while the GRE model forecasts are downwardly biased. The RMSE of Bayes GRE forecasts is 17.49ppb when the observed values are larger than 80ppb while the RMSE is only 9ppb otherwise. This validation analysis shows that the Bayesian hierarchical GRE model forecasts reduce the bias in numerical model output through assimilation. However, in the example, the model cannot capture the upper tail of the distribution very well.

The standard deviation map in Figure 5.2 shows that the forecasts for higher values are associated with larger standard deviations.

The probability maps (panels (c) and (d)) in Figure 5.2) show that many areas in Pennsylvania, West Virginia, New York and Connecticut will have their daily 8-hour maximum ozone concentration for August much larger than 80ppb with a very high probability.

Table 5.1: RMSEs and rBIAS

	RMSE		rBIAS	
Validation Days	Eta-CMAQ	Bayes GRE	Eta-CMAQ	Bayes GRE
Aug 10	18.66	7.82	0.2917	-0.05075
Aug 11	14.83	13.94	0.1204	-0.1235
Aug 12	12.70	9.72	0.1388	-0.08586
Aug 13	14.78	7.97	0.1867	-0.03447
Aug 10-13	15.41	10.22	0.1807	-0.07480

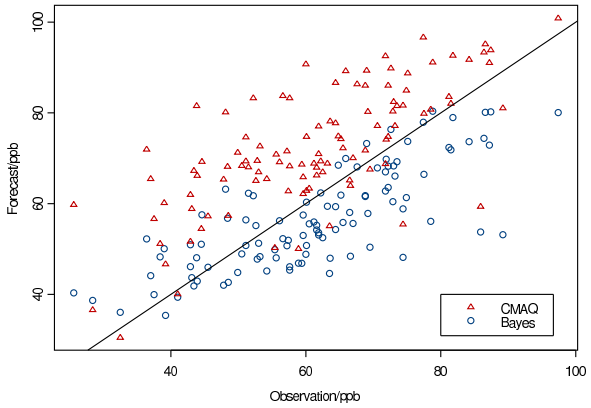


Figure 5.1: Validation plot for the period 10th-13th August of the Eta-CMAQ forecasts and the forecasts under the GRE model.

5.3 An Auto-regressive Model with Spatially Varying Slope

Following Sahu, Gelfand and Holland (2006, 2007), the pollutant process is modelled as a high-resolution space-time process. Although their work does not deal with data assimilation, this approach can be used to combine computer model output and observed data. A conceptual graph in Figure 5.3 shows the dependent structure of the modelling mechanism.

Corresponding to $Z(\mathbf{s}_i, t)$, let $O(\mathbf{s}_i, t)$ denote the true square-root ozone concentration level at $\mathbf{s}_i$ and at time t. We first assume:

$$Z(\mathbf{s}, t) = O(\mathbf{s}_i, t) + \epsilon(\mathbf{s}, t), \tag{5.3}$$

where $\epsilon(\mathbf{s}_i, t)$ is a white noise process, assumed to follow $N(0, \sigma_\epsilon^2)$ independently. Thus σ_ϵ^2, taken to be homogeneous in space and time, is the so called nugget effect.

Now we consider modelling $O(\mathbf{s}_i, t)$ as a function of previous values of

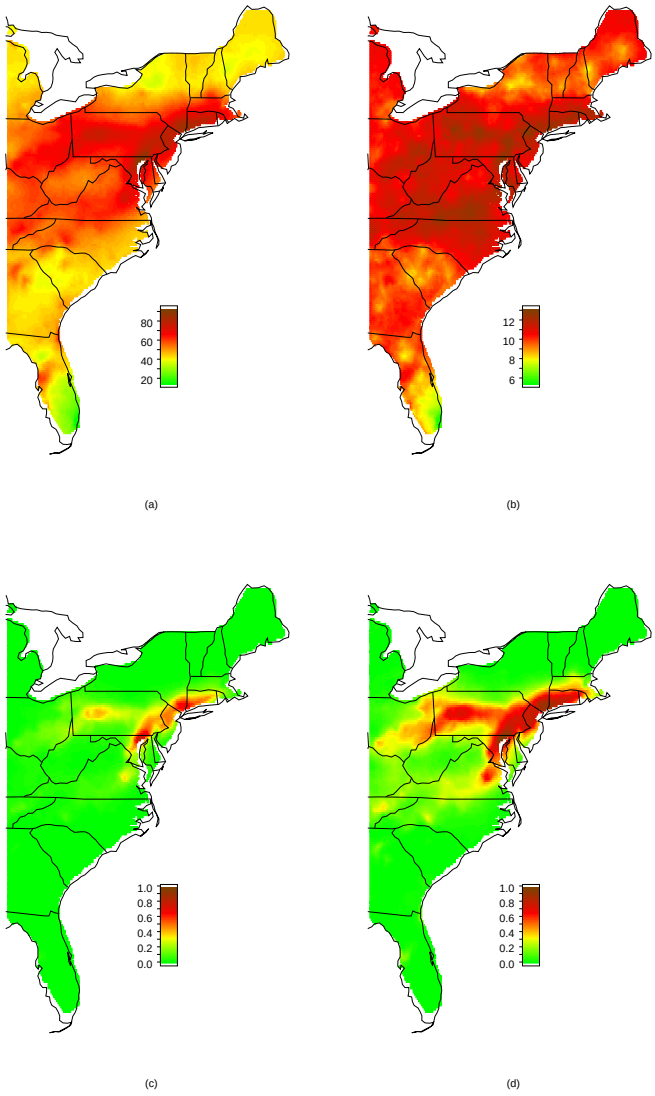


Figure 5.2: GRE model forecasts: (a) one-day-ahead forecasts of ozone level on 10th Aug. (b) standard deviation of one-day-ahead forecasts of ozone level on August 10 (c) probability of one-day-ahead forecasts exceeding 80ppb on August 10 (d) probability of one-day-ahead forecasts exceeding 70ppb on August 10.

ozone levels and the square root of the eta-CMAQ output $x(\mathbf{s}_i, t)$ for the grid-cell containing the site $\mathbf{s}_i$. The qualitative information of the atmospheric science tells us the ozone concentration level at a particular space-time point evolves from the successive state (day). In addition, from the EDA in Chapter 3, there is an auto-correlation between the successive day's ozone concentration measurement. The Eta-CMAQ forecast is also highly relevant for prediction of ozone levels. We add a spatially varying regression term ("slope") and the square root of the original Eta-CMAQ forecast as a predictor. Thus, we assume that,

$$O(\mathbf{s}_i, t) = \xi_t + \rho\, O(\mathbf{s}_i, t-1) + (\beta_0 + \beta(\mathbf{s}_i))\, x(\mathbf{s}_i, t) + \eta(\mathbf{s}_i, t), \tag{5.4}$$

for $i = 1, \ldots, n,\ t = 1, \ldots, T$ where ξ_t is a random walk process $\xi_t \sim N(\xi_{t-1}, \sigma_\xi^2)$, $\rho\, O(\mathbf{s}_i, t-1)$ is the auto-regressive term with $0 < \rho < 1$, $(\beta_0 + \beta(\mathbf{s}_i))\, x(\mathbf{s}_i, t)$ is the spatially varying regression term and $\eta(\mathbf{s}_i, t)$ is a spatially correlated but temporally independent error term. The term $\beta(\mathbf{s}_i)$ is particularly attractive because it is a Gaussian process centred at 0 allowing a non-stationary structure. The inclusion of the spatially-varying term β_0 and $\beta(\mathbf{s}_i)$ leads to a non-stationary model. A stationary sub-model with $\beta(\mathbf{s}_i) = 0$ is a special case for this model.

Another sub-model can be obtained by reducing the random walk process ξ_t to a single constant term. Then T parameters $\xi_1, \ldots, \xi_T$ would reduce to one. We expect that the full model to give a better fit than the sub-model but extra variability is introduced due to additional parameters. The systematic equation requires a corresponding initial condition for $O(\mathbf{s}_i, 0)$ which we choose to be grand mean of the data.

For computational convenience, we shall use the following vector notations: $\mathbf{Z}_t = (Z(\mathbf{s}_1, t), \ldots, Z(\mathbf{s}_n, t))'$, $\mathbf{O}_t = (O(\mathbf{s}_1, t), \ldots, O(\mathbf{s}_n, t))'$, $\mathbf{x}_t = (x(\mathbf{s}_1, t), \ldots, x(\mathbf{s}_n, t))'$, where X_t is a diagonal matrix whose ith diagonal entry is $x(\mathbf{s}_i, t)$, and $\boldsymbol{\epsilon}_t = (\epsilon(\mathbf{s}_1, t), \cdots, \epsilon(\mathbf{s}_n, t))'$. Now we write the above

models using vectors and matrices to facilitate computation. The first model equation will be the same as the equation (5.1) in GRE models:

$$\mathbf{Z}_t = \mathbf{O}_t + \boldsymbol{\epsilon}_t, \tag{5.5}$$

for $t = 1, \ldots, T$, where $\boldsymbol{\epsilon}_t = (\epsilon(\mathbf{s}_1, t), \cdots, \epsilon(\mathbf{s}_n, t))'$. Let $\mathbf{1}$ be the vector of dimension n with all elements unity and $\boldsymbol{\beta} = (\beta(\mathbf{s}_1), \ldots, \beta(\mathbf{s}_n))'$. From (5.4) we have:

$$\mathbf{O}_t = \xi_t \mathbf{1} + \rho\, \mathbf{O}_{t-1} + \beta_0 \mathbf{x}_t + X_t \boldsymbol{\beta} + \boldsymbol{\eta}_t, \tag{5.6}$$

for $t = 1, \ldots, T$, where $\boldsymbol{\eta}_t = (\eta(\mathbf{s}_1, t), \ldots, \eta(\mathbf{s}_n, t))'$.

For the measurement error in (5.5) we assume that $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \sigma_\epsilon^2 I_n)$, $t = 1, \ldots, T$, independently, where $\mathbf{0}$ is the vector with all elements zero and I_n is the identity matrix of order n. For the term ξ_t, we define it as a random walk process that $\xi_t \sim N(\xi_{t-1}, \sigma_\xi^2)$. For the spatially correlated error we assume that $\boldsymbol{\eta}_t \sim N(\mathbf{0}, \Sigma_\eta)$, $t = 1, \ldots, T$ where Σ_η has elements $\sigma_\eta(i, j) = \sigma_\eta^2 \rho_\eta(\mathbf{s}_i - \mathbf{s}_j; \phi_\eta)$. We take $\rho_\eta(\mathbf{s}_i - \mathbf{s}_j; \phi_\eta) = \exp(-\phi_\eta d(\mathbf{s}_i, \mathbf{s}_j))$ where $d(\mathbf{s}_i, \mathbf{s}_j)$ is the distance between sites $\mathbf{s}_i$ and $\mathbf{s}_j$, $i, j = 1, \ldots, n$. We choose the exponential covariance structure due to its mathematical elegance.

The spatially varying coefficients $\boldsymbol{\beta} \sim N(\mathbf{0}, \Sigma_\beta)$ where Σ_β has elements $\sigma_\beta(i, j) = \sigma_\beta^2 \rho(\mathbf{s}_i - \mathbf{s}_j; \phi_\beta)$. The parameters ϕ_η and ϕ_β are determined using grid-search methods similar to the one described in Section 4.2.4. For future reference we define S_η and S_β by the relations:

$$\Sigma_\eta = \sigma_\eta^2 S_\eta, \quad \Sigma_\beta = \sigma_\beta^2 S_\beta.$$

Let $\boldsymbol{\vartheta}_t = \xi_t \mathbf{1} + \rho\, \mathbf{O}_{t-1} + \beta_0 \mathbf{x}_t + X_t \boldsymbol{\beta}$, for $t = 1, \ldots T$. Further, let $\boldsymbol{\theta}$ denote all the parameters, β_0, $\boldsymbol{\beta}$, ρ, σ_ϵ^2, σ_η^2, σ_β^2, σ_ξ^2 and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_T)^T$. Let $\mathbf{w}$ denote all the augmented data, $\mathbf{o}_t$ and the missing data, denoted by $z^*(\mathbf{s}_i, t)$, for $i = 1, \ldots, n$, $t = 1, \ldots, T$, and $\mathbf{z}$ denote all the non-missing data $z(\mathbf{s}_i, t)$, for $i = 1, \ldots, n$, $t = 1, \ldots, T$. The log of the posterior distribution, denoted

by $\log \pi(\boldsymbol{\theta}, \mathbf{w}|\mathbf{z})$, can be written as

$$
\begin{aligned}
&-\tfrac{nT}{2}\log(\sigma_\epsilon^2) - \tfrac{1}{2\sigma_\epsilon^2}\textstyle\sum_{t=1}^{T}(\mathbf{Z}_t - \mathbf{O}_t)'(\mathbf{Z}_t - \mathbf{O}_t)\\
&-\tfrac{nT}{2}\log(\sigma_\eta^2) - \tfrac{1}{2\sigma_\eta^2}\textstyle\sum_{t=1}^{T}(\mathbf{O}_t - \boldsymbol{\vartheta}_t)'S_\eta^{-1}(\mathbf{O}_t - \boldsymbol{\vartheta}_t)\\
&-\tfrac{T}{2}\log(\sigma_\xi^2) - \tfrac{1}{2\sigma_\xi^2}\textstyle\sum_{t=2}^{T}(\xi_t - \xi_{t-1})^2\\
&-\tfrac{n}{2}\log(\sigma_\beta^2) - \tfrac{1}{2\sigma_\beta^2}\boldsymbol{\beta}'S_\beta^{-1}\boldsymbol{\beta} + \log\left(\pi(\boldsymbol{\xi}, \beta_0, \rho, \sigma_\epsilon^2, \sigma_\eta^2, \sigma_\beta^2, \sigma_\xi^2)\right)
\end{aligned}
$$

where $\pi(\boldsymbol{\xi}, \rho, \beta_0, \sigma_\epsilon^2, \sigma_\eta^2, \sigma_\beta^2, \sigma_\xi^2)$ is the prior distribution. We assume that a-priori β_0 is independently normally distributed with mean 0 and variance 10^4. The auto-regressive coefficient ρ is specified the $N(0, 10^4)$ distribution but restricted to the interval $I(0 < \rho < 1)$. The inverse of the variance components, $\frac{1}{\sigma_\epsilon^2}$, $\frac{1}{\sigma_\eta^2}$, $\frac{1}{\sigma_\beta^2}$, $\frac{1}{\sigma_\xi^2}$ are assumed to follow $G(a, b)$ independently, where the distribution $G(a, b)$ has the mean a/b. For the constant ξ sub-model, ξ is normally distributed with mean 0 and variance 10^4 a-priori . In our implementation we take $a = 2$ and $b = 1$ to have a proper prior specification for each of these variance components, see Section 3.5 for further justifications.

We label the three sub-models as follows:

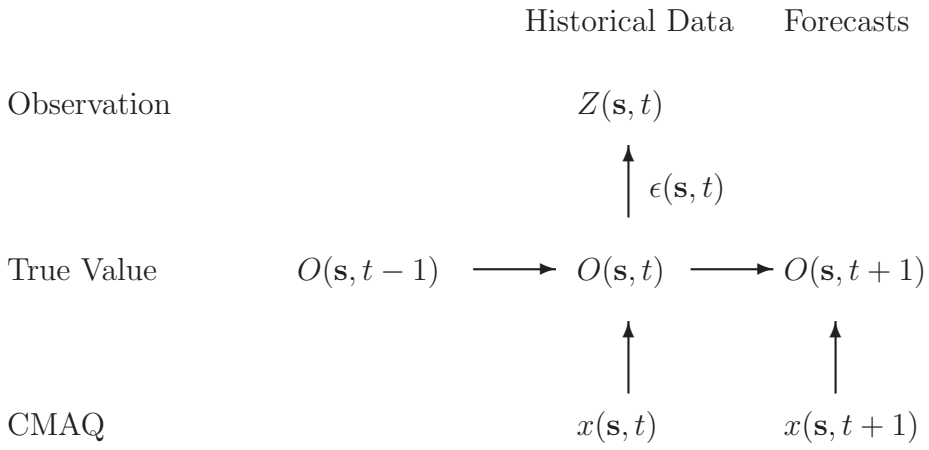
1. ARM(1): auto-regressive model with constant ξ and no spatially varying regression term.

2. ARM(2): auto-regressive model with varying ξ and no spatially varying regression term.

3. ARM(3): auto-regressive model with constant ξ and a spatially varying regression term.

5.3.1 Conditional Distributions for Gibbs Sampling

Conditional Distributions for: σ_ϵ^2, σ_η^2, σ_β^2, σ_ξ^2, $\mathbf{O}_t$, ρ, β_0 and $\boldsymbol{\beta}$



Measurement Equation: $Z(\mathbf{s},t) = O(\mathbf{s},t) + \epsilon(\mathbf{s},t)$

System Equation: $O(\mathbf{s}_i,t) = \xi_t + \rho\, O(\mathbf{s}_i,t-1) + (\beta_0 + \beta(\mathbf{s}_i))\, x(\mathbf{s}_i,t) + \eta(\mathbf{s}_i,t)$

Figure 5.3: Conceptual graph of the ARM framework.

Any missing value, $Z(\mathbf{s},t)$ is to be sampled from $N(O(\mathbf{s},t),\sigma_\epsilon^2)$, $t=1,\ldots,T$. Straightforward calculation yields the following complete conditional distributions:

$$\begin{aligned}
&\tfrac{1}{\sigma_\epsilon^2} \sim G\left(\tfrac{nT}{2}+a, b+\tfrac{1}{2}\textstyle\sum_{t=1}^{T}(\mathbf{Z}_t-\mathbf{O}_t)'(\mathbf{Z}_t-\mathbf{O}_t)\right),\\
&\tfrac{1}{\sigma_\eta^2} \sim G\left(\tfrac{nT}{2}+a, b+\tfrac{1}{2}\textstyle\sum_{t=1}^{T}(\mathbf{O}_t-\boldsymbol{\vartheta}_t)'S_\eta^{-1}(\mathbf{O}_t-\boldsymbol{\vartheta}_t)\right),\\
&\tfrac{1}{\sigma_\beta^2} \sim G\left(\tfrac{n}{2}+a, b+\tfrac{1}{2}\boldsymbol{\beta}'S_\beta^{-1}\boldsymbol{\beta}\right),\\
&\tfrac{1}{\sigma_\xi^2} \sim G\left(\tfrac{T}{2}+a, b+\tfrac{1}{2}\textstyle\sum_{t=2}^{T}(\xi_t-\xi_{t-1})^2\right).
\end{aligned}$$

Let $Q_\eta=\Sigma_\eta^{-1}$. The full conditional distribution of $\mathbf{O}_t$ is $N(\Lambda_t\boldsymbol{\chi}_t,\ \Lambda_t)$ where
Case 1: For $1\le t<T-1$:

$$\Lambda_t^{-1}=\frac{I_n}{\sigma_\epsilon^2}+(1+\rho^2)Q_\eta,$$

$$\boldsymbol{\chi}_t=\frac{\mathbf{Z}_t}{\sigma_\epsilon^2}+Q_\eta\left\{\xi\mathbf{1}+\rho\mathbf{O}_{t-1}+\beta_0\mathbf{x}_t+X_t\boldsymbol{\beta}+\rho\left(\mathbf{O}_{t+1}-\xi\mathbf{1}-\beta_0\mathbf{x}_{t+1}-X_{t+1}\boldsymbol{\beta}\right)\right\}.$$

Case 2: For $t=T$

$$\Lambda_t^{-1}=\frac{I_n}{\sigma_\epsilon^2}+Q_\eta,$$

$$\boldsymbol{\chi}_t=\frac{\mathbf{Z}_t}{\sigma_\epsilon^2}+Q_\eta\left\{\xi\mathbf{1}+\rho\mathbf{O}_{t-1}+\beta_0\mathbf{x}_t+X_t\boldsymbol{\beta}\right\}.$$

The full conditional distribution of ρ is $N(\Lambda\chi,\ \Lambda)$ where

$$\Lambda^{-1}=\sum_{t=1}^{T}\mathbf{O}_{t-1}'Q_\eta\mathbf{O}_{t-1}+10^{-4},\ \chi=\sum_{t=1}^{T}\mathbf{O}_{t-1}'Q_\eta(\mathbf{O}_t-\xi\mathbf{1}-\beta_0\mathbf{x}_t-X_t\boldsymbol{\beta}),$$

restricted in the interval (0, 1).

The full conditional distribution of β_0 is $N(\Lambda\chi,\ \Lambda)$ where

$$\Lambda^{-1}=\sum_{t=1}^{T}\mathbf{x}_t'Q_\eta\mathbf{x}_t+10^{-4},\ \chi=\sum_{t=1}^{T}\mathbf{x}_t'Q_\eta(\mathbf{O}_t-\xi\mathbf{1}-\rho\mathbf{O}_{t-1}-X_t\boldsymbol{\beta}),$$

The full conditional distribution of $\boldsymbol{\beta}$ is $N(\Lambda\boldsymbol{\xi},\ \Lambda)$ where

$$\Lambda^{-1}=\sum_{t=1}^{T}X_t'Q_\eta X_t+\Sigma_\beta^{-1},\ \text{and}$$

$$\boldsymbol{\xi} = \sum_{t=1}^{T} X_t' Q_\eta (\mathbf{O}_t - \xi \mathbf{1} - \rho \mathbf{O}_{t-1} - \beta_0 \mathbf{x}_t).$$

The full conditional distribution for random walk ξ_t is $N(\Lambda_t \boldsymbol{\chi}_t,\ \Lambda_t)$ where

Case 1: For $1 \leq t < T-1$:

$$\Lambda^{-1} = \frac{2}{\sigma_\xi^2} + \mathbf{1}' Q_\eta \mathbf{1}, \quad \chi = \mathbf{1}' Q_\eta \mathbf{a}_t + \frac{\xi_{t-1} + \xi_{t+1}}{\sigma_\xi^2}.$$

Case 2: For $t = T$

$$\Lambda^{-1} = \frac{1}{\sigma_\xi^2} + \mathbf{1}' Q_\eta \mathbf{1}, \quad \chi = \mathbf{1}' Q_\eta \mathbf{a}_t + \frac{\xi_{t-1}}{\sigma_\xi^2}.$$

For constant ξ sub-model, the full conditional distribution of ξ is $N(\Lambda\chi,\ \Lambda)$ where

$$\Lambda^{-1} = \frac{1}{\sigma_\xi^2} + T \mathbf{1}' Q_\eta \mathbf{1}, \quad \chi = \mathbf{1}' Q_\eta \sum_{t=1}^{T} \mathbf{a}_t$$

where $\mathbf{a}_t = \mathbf{O}_t - \rho \mathbf{O}_{t-1} - \beta_0 \mathbf{x}_t - X_t \boldsymbol{\beta}$.

5.3.2 Prediction Details

Prediction methods used for interpolation and forecast of a space-time point are slightly different. We first develop the methods of spatial interpolation of ozone levels at a new location $\mathbf{s}' \in S$ and any time t, $t = 1, \ldots, T$. We then look at the one step-ahead forecasting at time $t = T+1$ for any new location within the region S.

Spatial interpolation at location $\mathbf{s}'$ and time t is based upon the predictive distribution of $Z(\mathbf{s}', t)$ given in the model equations (5.4) and (5.5). According to Figure 5.3, $Z(\mathbf{s}', t)$, has the distribution:

$$Z(\mathbf{s}', t) \sim N\left(O(\mathbf{s}', t),\ \sigma_\epsilon^2\right) \tag{5.7}$$

and

$$O(\mathbf{s}', t) = \xi_t + \rho\, O(\mathbf{s}', t-1) + \left(\beta_0 + \beta(\mathbf{s}')\right) x(\mathbf{s}', t) + \eta(\mathbf{s}', t). \tag{5.8}$$

From the system equations in (5.8), the auto-regressive $O(\mathbf{s}', t)$ can only be sequentially determined from all the previous values. Hence, we introduce the notation $\mathbf{O}(\mathbf{s}, [t])$ to denote the vector $(O(\mathbf{s}, 1), \ldots, O(\mathbf{s}, t))'$ for $t \geq 1$.
The posterior predictive distribution of $Z(\mathbf{s}', t)$ is obtained by integrating over the unknown quantities in (5.7) with respect to the joint posterior distribution, i.e.,

$$\begin{aligned} \pi\left(Z(\mathbf{s}', t)|\mathbf{z}\right) &= \int \pi\left(Z(\mathbf{s}', t)|O(\mathbf{s}', [t]), \sigma_\epsilon^2\right) \pi\left(O(\mathbf{s}', [t])|\beta(\mathbf{s}'), \boldsymbol{\theta}, \mathbf{w}\right) \\ &\quad \pi\left(\beta(\mathbf{s}')|\boldsymbol{\theta}\right) \, dO(\mathbf{s}', [t]) \, d\beta(\mathbf{s}') \, d\boldsymbol{\theta} \, d\mathbf{w}. \end{aligned} \tag{5.9}$$

When using MCMC methods to draw samples from the posterior, the predictive distribution (5.9) is sampled by composition. Draws from the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{z}, \mathbf{w})$, and the conditional distributions $\pi\left(\beta(\mathbf{s}')|\boldsymbol{\theta}\right)$ facilitate evaluation of the above integral, details provided below.
To sample $\beta(\mathbf{s}')$ we have

$$\begin{pmatrix} \beta(\mathbf{s}') \\ \boldsymbol{\beta} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix}, \sigma_\beta^2 \begin{pmatrix} 1 & S_{\beta,12} \\ S_{\beta,21} & S_\beta \end{pmatrix}\right],$$

where $S_{\beta,12}$ is $1 \times n$ with the ith entry given by $\exp(-\phi_\beta d(\mathbf{s}_i, \mathbf{s}'))$ and $S_{\beta,21} = S'_{\beta,12}$. Therefore,

$$\beta(\mathbf{s}')|\boldsymbol{\theta} \sim N\left(S_{\beta,12} S_\gamma^{-1} \boldsymbol{\beta}, \sigma_\beta^2 \left(1 - S_{\beta,12} S_\beta^{-1} S_{\beta,21}\right)\right). \tag{5.10}$$

We draw $O(\mathbf{s}', t)$ from its conditional distribution given $\boldsymbol{\theta}, \mathbf{w}$ and $O(\mathbf{s}', [t-1])$. Analogous to (5.6), we obtain for $t \geq 0$

$$\begin{pmatrix} O(\mathbf{s}', t) \\ \mathbf{O}_t \end{pmatrix} \sim N\left[\begin{pmatrix} \xi + \rho\, O(\mathbf{s}', t-1) + (\beta_0 + \beta(\mathbf{s}'))\, x(\mathbf{s}', t) \\ \xi \mathbf{1} + \rho\, \mathbf{O}_{t-1} + \beta_0 \mathbf{x}_t + X_t \boldsymbol{\beta} \end{pmatrix}, \; \sigma_\eta^2 \begin{pmatrix} 1 & S_{\eta,12} \\ S_{\eta,21} & S_\eta \end{pmatrix}\right]$$

where $S_{\eta,12}$ is $1 \times n$ with the ith entry given by $\exp(-\phi_\eta d(\mathbf{s}_i, \mathbf{s}'))$ and $S_{\eta,21} = S'_{\eta,12}$. Hence,

$$O(\mathbf{s}', t)|\beta(\mathbf{s}'), \mathbf{O}_t, \boldsymbol{\theta}, \mathbf{w} \sim N(\chi, \; \Lambda) \tag{5.11}$$

where $\Lambda = \sigma_\eta^2 \left(1 - S_{\eta,12} S_\eta^{-1} S_{\eta,21}\right)$ and

$$\chi = \xi_t + \rho\, O(\mathbf{s}', t-1) + (\beta_0 + \beta(\mathbf{s}'))\, x(\mathbf{s}', t) + S_{\eta,12} S_\eta^{-1} \left(\mathbf{O}_t - \xi_t \mathbf{1} - \rho\, \mathbf{O}_{t-1} - \beta_0 \mathbf{x}_t - X_t \boldsymbol{\beta}\right).$$

In summary, we implement the following algorithm to predict $Z(\mathbf{s}', t), t = 1, \ldots, T$.

1. Draw a sample $\boldsymbol{\theta}^{(j)}, \mathbf{w}^{(j)}, j \geq 1$ from the posterior distribution.
2. Draw $\beta^{(j)}(\mathbf{s}')$ using (5.10).
3. Draw $\mathbf{O}^{(j)}(\mathbf{s}', [t])$ sequentially using (5.11). Note that the initial value value $O^{(j)}(\mathbf{s}', 0)$ is a constant for all $\mathbf{s}'$.
4. Finally draw $Z^{(j)}(\mathbf{s}', t)$ from $N\left(O^{(j)}(\mathbf{s}', t),\ \sigma_\epsilon^{2(j)}\right)$.

The ozone concentration in the original scale is the square of $Z^{(j)}(\mathbf{s}', t)$. If we want the predictions of the smooth ozone concentration process without the nugget term we would simply omit the last step in the above algorithm and square the realisations $\mathbf{O}^{(j)}(\mathbf{s}, t)$. We use the median of the MCMC samples and the lengths of the 95% intervals to summarise the predictions. The median as a summary measure preserves the one-to-one relationships between summaries for O and Z, and for O^2 and Z^2.

The one-step ahead Bayesian forecast at a location $\mathbf{s}'$ is given by the posterior predictive distribution of $Z(\mathbf{s}', T+1)$ which is determined by $O(\mathbf{s}', T+1)$. Note that using (5.11) we already have the conditional distribution of $O(\mathbf{s}', T)$ given $\beta(\mathbf{s}'), \mathbf{O}_t, \boldsymbol{\theta}$, and $\mathbf{w}$. We use model equation (5.4) to advance this conditional distribution one unit at a time in the future. The mean of the one step-ahead forecast distribution is given by $\xi_t + \rho\, O(\mathbf{s}', T) + (\beta_0 + \beta(\mathbf{s}'))\, x(\mathbf{s}', T)$, according to (5.4), and $O(\mathbf{s}', T+1)$ should be equal to this if we are interested in forecasting the mean. However, if we want to forecast an observation at location $\mathbf{s}'$ we simulate $O(\mathbf{s}', T+1)$ from the marginal distribution which has the above mean and variance σ_η^2. We work with this marginal distribution

rather than the conditional distribution since conditioning with respect to the observed information (i.e. Kriging) up to time T at the observation locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ has already been done in (5.11), and at the future time $T+1$ there is no new available information to condition on except for using the Eta-CMAQ output as regressor values. Then we follow the above algorithm and the MCMC output symmetrisation methods to evaluate the forecasts.

5.3.3 Results

We again use the same data set as in the previous example in Section 5.2.1. Models ARM(1), ARM(2) and ARM(3) are used in this analysis. Tables 5.2 to 5.13 provide the validation results for these models for validating on August 10, 11, 12 and 13. We can see that the performances of ARM(1) and ARM(2) are very close to each other. Although ARM(3) has a more complex structure than ARM(1), it does not outperform either models according to most of the criteria.

Figure 5.6 provides a plot of forecasts against observations for August 11 under the Eta-CMAQ model, and ARM(1) and ARM(3) models. Both these last two models perform better than the former, Eta-CMAQ model. At a higher ozone concentration level, the summary of prediction seems to be weaker than that for a lower value. However, these forecasts are much better than the forecasts based on the GRE model analysed previously in Section 5.2.

Figure 5.7 plots the 95% credible intervals for $\beta(\mathbf{s}_i), i = 1, \ldots, 350$. Most of these intervals include zero which implies that a spatially varying slope parameter is not significant. Hence, a spatially non-stationary model is not required for these data; henceforth we use the ARM(1) for analysis.

In Figure 5.4 we provide a map of the one-day ahead forecast surfaces on 11th Aug for ARM(1) and Eta-CMAQ. The observed ozone concentration levels are also superimposed on the graph. The maps show that the Bayesian

forecasts from ARM(1) are more accurate prediction than those from Eta-CMAQ model. Also note that the average ozone concentration in the model based map is significantly lower than that from Eta-CMAQ forecast map.

Figure 5.5 provides a map of the Bayesian forecasts from ARM(1) (left panel) for August 11 along with its uncertainty estimates (right panel). These give us a rough idea on what locations are forecasted to have high concentration levels with their associated uncertainties.

In Table 5.14, a comparison of hit and error rate of Eta-CMAQ and ARM(1) forecasts is shown. Here, hit is defined as the event where both the validation observation and the forecast for it were either both greater or less than 80ppb. The error, on the other hand, is defined as the event where the actual observation is less than 80ppb but the forecast is greater than 80ppb. The hit rate for the Bayes model is generally 10% higher than that for the Eta-CMAQ forecasts.

5.4 Summary

This Chapter provides a framework in modelling the order statistics (8-hour maximum ozone concentration levels here) of hourly recorded spatio-temporal ozone concentration data. Since the dimension is enormously reduced due to a lower temporal resolution, it allows extra computational power to adopt a more complex model. A spatially varying auto-regressive model enables inclusion of regional variation directly in the model. This chapter also demonstrates that it is possible to forecast at a future time point with the ARM by combining information from data and computer model output. Although the ARM improves the forecasting accuracy, it is still unable to predict the high ozone concentration values very accurately. In the next Chapter, the Gaussian assumption is removed by incorporating an additional hierarchy to address non-normality of the data.

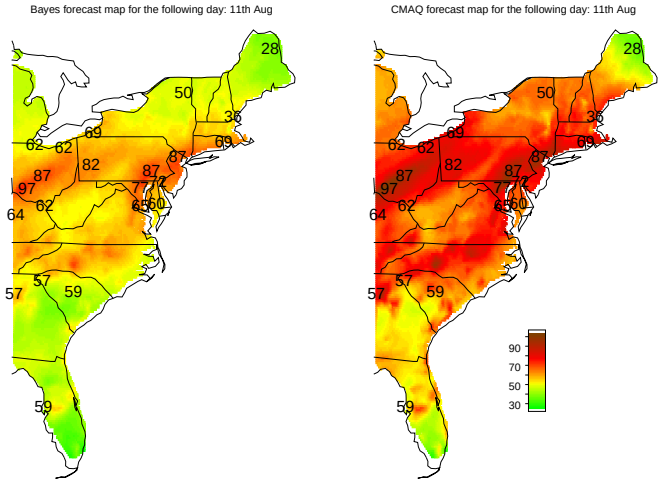


Figure 5.4: The one-day ahead forecast surfaces on August 11 for ARM(1) (left panel) and Eta-CMAQ (right panel). The observed ozone concentration level are superimposed on the graph.

Table 5.2: RMSEs, MAEs, rBIAS and rMSEP of ARM(1) on Aug 10: The forecasts rows represent the one-day-ahead ($1 \times 40 = 40$) set-aside validation sites. The interpolation rows represent the previous seven days ($7 \times 40 = 280$) set-aside validation sites. The two rows corresponding to Total represent all the validation data used ($8 \times 40 = 320$).

		RMSE	MAE	rBIAS	rMSEP
Forecasts	Eta-CMAQ	18.66	16.62	0.2917	0.8294
	Bayes	11.08	9.302	-0.1051	0.6101
Interpolation	Eta-CMAQ	15.15	12.02	0.1886	0.5293
	Bayes	6.996	5.367	-0.006381	0.1514
Total	Eta-CMAQ	15.7	12.67	0.2029	0.5748
	Bayes	7.705	5.923	-0.02001	0.1962

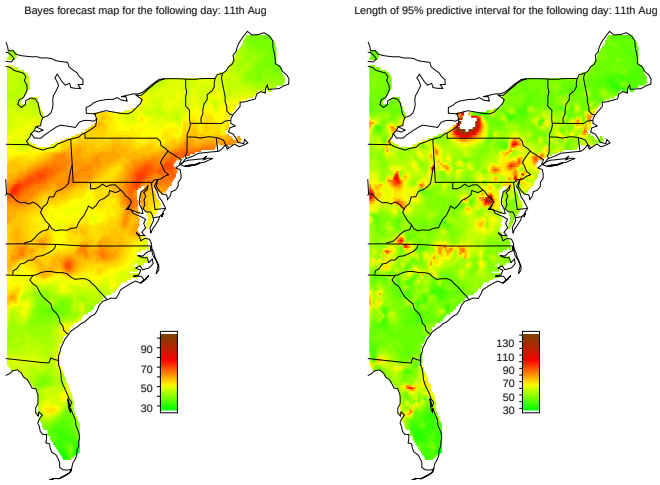


Figure 5.5: The one-day ahead forecast surface on August 11 for ARM(1) (left panel) and the length of 95% forecast intervals (right panel).

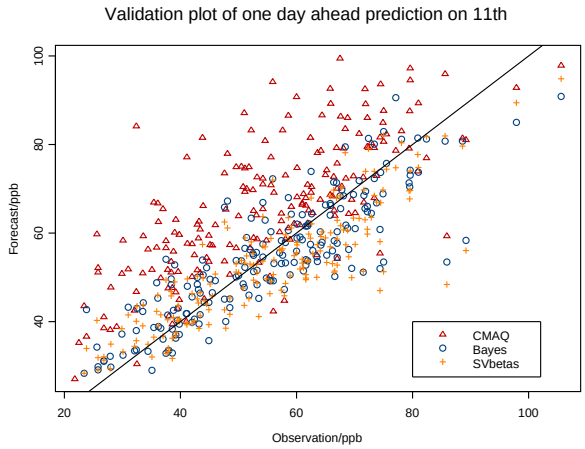


Figure 5.6: Validation plot for one-day ahead forecast on August 11 for Eta-CMAQ, ARM(1) (Bayes) and ARM(3) (SVbetas). The line $y = x$ is superimposed.

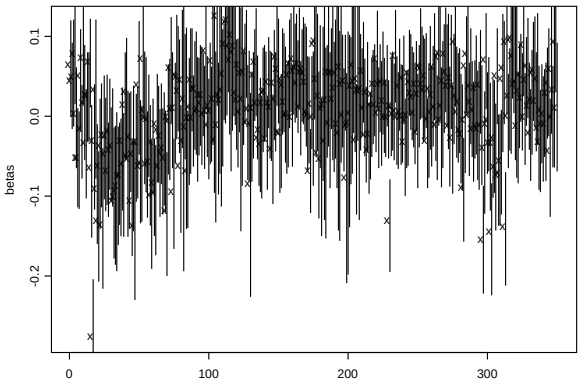


Figure 5.7: 95% creidible intervals for the 350 parameters, $\beta(\mathbf{s}_i), i = 1, \ldots, 350$ under the ARM(3) model when the data used in the fitting are for days August 4-10, 2005.

Table 5.3: RMSEs, MAEs, rBIAS and rMSEP of ARM(1) on Aug 11: The forecasts rows represent the one-day-ahead ($1 \times 40 = 40$) set-aside validation sites. The interpolation rows represent the previous seven days ($7 \times 40 = 280$) set-aside validation sites. The two rows corresponding to Total represent all the validation data used ($8 \times 40 = 320$).

		RMSE	MAE	rBIAS	rMSEP
Forecasts	Eta-CMAQ	14.83	11.94	0.1204	0.9529
	Bayes	13.48	11.18	-0.1061	0.8283
Interpolation	Eta-CMAQ	16.34	13.39	0.225	0.6324
	Bayes	6.688	5.172	-0.0129	0.1627
Total	Eta-CMAQ	16.14	13.19	0.2094	0.6585
	Bayes	7.981	6.004	-0.02681	0.2391

Table 5.4: RMSEs, MAEs, rBIAS and rMSEP of ARM(1) on Aug 12: The forecasts rows represent the one-day-ahead ($1 \times 40 = 40$) set-aside validation sites. The interpolation rows represent the previous seven days ($7 \times 40 = 280$) set-aside validation sites. The two rows corresponding to Total represent all the validation data used ($8 \times 40 = 320$).

		RMSE	MAE	rBIAS	rMSEP
Forecasts	Eta-CMAQ	12.7	10.01	0.1388	0.5212
	Bayes	12.48	10.71	-0.129	0.5235
Interpolation	Eta-CMAQ	16.37	13.31	0.2232	0.7724
	Bayes	6.384	5.017	-0.00982	0.1938
Total	Eta-CMAQ	15.92	12.86	0.2089	0.689
	Bayes	7.518	5.798	-0.03003	0.2346

Table 5.5: RMSEs, MAEs, rBIAS and rMSEP of ARM(1) on Aug 13: The forecasts rows represent the one-day-ahead ($1 \times 40 = 40$) set-aside validation sites. The interpolation rows represent the previous seven days ($7 \times 40 = 280$) set-aside validation sites. The two rows corresponding to Total represent all the validation data used ($8 \times 40 = 320$).

		RMSE	MAE	rBIAS	rMSEP
Forecasts	Eta-CMAQ	14.78	12.29	0.1867	0.6879
	Bayes	7.926	6.41	-0.02694	0.358
Interpolation	Eta-CMAQ	15.61	12.85	0.2119	0.6447
	Bayes	6.807	5.315	-0.0133	0.1877
Total	Eta-CMAQ	15.51	12.79	0.2084	0.6299
	Bayes	6.952	5.447	-0.01522	0.193

Table 5.6: RMSEs, MAEs, rBIAS and rMSEP of ARM(2) on Aug 10: The forecasts rows represent the one-day-ahead ($1 \times 40 = 40$) set-aside validation sites. The interpolation rows represent the previous seven days ($7 \times 40 = 280$) set-aside validation sites. The two rows corresponding to Total represent all the validation data used ($8 \times 40 = 320$).

		RMSE	MAE	rBIAS	rMSEP
Forecasts	Eta-CMAQ	18.66	16.62	0.2917	0.8294
	Bayes	11.21	9.522	-0.105	0.6246
Interpolation	Eta-CMAQ	15.15	12.02	0.1886	0.5293
	Bayes	6.815	5.172	-0.01005	0.1436
Total	Eta-CMAQ	15.7	12.67	0.2029	0.5748
	Bayes	7.592	5.787	-0.02317	0.1902

Table 5.7: RMSEs, MAEs, rBIAS and rMSEP of ARM(2) on Aug 11: The forecasts rows represent the one-day-ahead ($1 \times 40 = 40$) set-aside validation sites. The interpolation rows represent the previous seven days ($7 \times 40 = 280$) set-aside validation sites. The two rows corresponding to Total represent all the validation data used ($8 \times 40 = 320$).

		RMSE	MAE	rBIAS	rMSEP
Forecasts	Eta-CMAQ	14.83	11.94	0.1204	0.9529
	Bayes	12.56	10.8	-0.03709	0.8518
Interpolation	Eta-CMAQ	16.34	13.39	0.225	0.6324
	Bayes	6.996	5.308	-0.007965	0.1783
Total	Eta-CMAQ	16.14	13.19	0.2094	0.6585
	Bayes	8.000	6.069	-0.01231	0.2417

Table 5.8: RMSEs, MAEs, rBIAS and rMSEP of ARM(2) on Aug 12: The forecasts rows represent the one-day-ahead ($1 \times 40 = 40$) set-aside validation sites. The interpolation rows represent the previous seven days ($7 \times 40 = 280$) set-aside validation sites. The two rows corresponding to Total represent all the validation data used ($8 \times 40 = 320$).

		RMSE	MAE	rBIAS	rMSEP
Forecasts	Eta-CMAQ	12.7	10.01	0.1388	0.5212
	Bayes	10.62	8.353	-0.05676	0.4765
Interpolation	Eta-CMAQ	16.37	13.31	0.2232	0.7724
	Bayes	6.804	5.221	-0.01247	0.2199
Total	Eta-CMAQ	15.92	12.86	0.2089	0.689
	Bayes	7.443	5.651	-0.01998	0.2313

Table 5.9: RMSEs, MAEs, rBIAS and rMSEP of ARM(2) on Aug 13: The forecasts rows represent the one-day-ahead ($1 \times 40 = 40$) set-aside validation sites. The interpolation rows represent the previous seven days ($7 \times 40 = 280$) set-aside validation sites. The two rows corresponding to Total represent all the validation data used ($8 \times 40 = 320$).

		RMSE	MAE	rBIAS	rMSEP
Forecasts	Eta-CMAQ	14.78	12.29	0.1867	0.6879
	Bayes	7.965	6.334	0.005847	0.3675
Interpolation	Eta-CMAQ	15.61	12.85	0.2119	0.6447
	Bayes	6.677	5.274	-0.008957	0.1808
Total	Eta-CMAQ	15.51	12.79	0.2084	0.6299
	Bayes	6.845	5.402	-0.006878	0.1882

Table 5.10: RMSEs, MAEs, rBIAS and rMSEP of ARM(3) on Aug 10: The forecasts rows represent the one-day-ahead ($1 \times 40 = 40$) set-aside validation sites. The interpolation rows represent the previous seven days ($7 \times 40 = 280$) set-aside validation sites. The two rows corresponding to Total represent all the validation data used ($8 \times 40 = 320$).

		RMSE	MAE	rBIAS	rMSEP
Forecasts	Eta-CMAQ	18.66	16.62	0.2917	0.8294
	Bayes	8.91	6.944	-0.005075	0.4705
Interpolation	Eta-CMAQ	15.15	12.02	0.1886	0.5293
	Bayes	6.751	5.363	-0.01391	0.1408
Total	Eta-CMAQ	15.7	12.67	0.2029	0.5748
	Bayes	7.096	5.586	-0.01269	0.1667

Table 5.11: RMSEs, MAEs, rBIAS and rMSEP of ARM(3) on Aug 11: The forecasts rows represent the one-day-ahead ($1 \times 40 = 40$) set-aside validation sites. The interpolation rows represent the previous seven days ($7 \times 40 = 280$) set-aside validation sites. The two rows corresponding to Total represent all the validation data used ($8 \times 40 = 320$).

		RMSE	MAE	rBIAS	rMSEP
Forecasts	Eta-CMAQ	14.83	11.94	0.1204	0.9529
	Bayes	14.5	10.98	-0.1345	0.864
Interpolation	Eta-CMAQ	16.34	13.39	0.225	0.6324
	Bayes	7.005	5.665	-0.02627	0.1776
Total	Eta-CMAQ	16.14	13.19	0.2094	0.6585
	Bayes	8.449	6.402	-0.04244	0.2647

Table 5.12: RMSEs, MAEs, rBIAS and rMSEP of ARM(3) on Aug 12: The forecasts rows represent the one-day-ahead ($1 \times 40 = 40$) set-aside validation sites. The interpolation rows represent the previous seven days ($7 \times 40 = 280$) set-aside validation sites. The two rows corresponding to Total represent all the validation data used ($8 \times 40 = 320$).

		RMSE	MAE	rBIAS	rMSEP
Forecasts	Eta-CMAQ	12.7	10.01	0.1388	0.5212
	Bayes	17.16	15.05	-0.1912	0.7588
Interpolation	Eta-CMAQ	16.37	13.31	0.2232	0.7724
	Bayes	6.599	5.277	-0.01703	0.2065
Total	Eta-CMAQ	15.92	12.86	0.2089	0.689
	Bayes	8.828	6.617	-0.04656	0.318

Table 5.13: RMSEs, MAEs, rBIAS and rMSEP of ARM(3) on Aug 13: The forecasts rows represent the one-day-ahead ($1 \times 40 = 40$) set-aside validation sites. The interpolation rows represent the previous seven days ($7 \times 40 = 280$) set-aside validation sites. The two rows corresponding to Total represent all the validation data used ($8 \times 40 = 320$).

		RMSE	MAE	rBIAS	rMSEP
Forecasts	Eta-CMAQ	14.78	12.29	0.1867	0.6879
	Bayes	13.13	10.93	-0.1298	0.7107
Interpolation	Eta-CMAQ	15.61	12.85	0.2119	0.6447
	Bayes	6.525	5.352	-0.004788	0.1727
Total	Eta-CMAQ	15.51	12.79	0.2084	0.6299
	Bayes	7.63	6.025	-0.02234	0.2325

Period	Eta-CMAQ Hit	Error	Bayes Hit	Error
Aug 2-9	84.76	15.24	95.12	4.88
Aug 3-10	82.20	17.80	94.24	5.76
Aug 4-11	82.05	17.95	94.36	5.64
Aug 5-12	84.78	15.22	94.92	5.08
Aug 6-13	83.92	16.08	93.97	6.03

Table 5.14: Hit and error percentages for O_3 exceeding 80ppb. Here, hit is defined as the event where both the validation observation and the forecast for it were either both greater or less than 80ppb. The error, on the other hand, is defined as the event where the actual observation is less than 80ppb but the forecast is greater than 80ppb.

Chapter 6

Non-Gaussian Measurement Error Models

6.1 Introduction

The models we have considered so far in this thesis are all based on Gaussian distribution. Often in many environmental data modelling problems it is of interest to model the extreme values. For example, in the previous chapter we have modelled the daily maximum 8-hour average ozone concentration levels. Although these are averages of 8-successive hourly values, the end statistic is the maximum of the 24 8-hour averages for a day. The Gaussian distribution, though used somewhat successfully in the previous chapter, is not an appropriate model for the largest order statistic.

This chapter will develop an extreme value distribution (Section 6.2) for the daily maximum 8-hour average ozone concentration levels as an alternative to the Gaussian distribution assumed in the previous chapter. After a brief introduction to the spatial extreme values (Section 6.3) we describe the models and develop MCMC computation algorithms. We also provide details for prediction at unobserved location and also for future time points.

We illustrate the methods with a simulation and a real data example in Section 6.4. A few summary remarks are provided at the end.

6.2 Extremes and Spatial Extremes

It is often of interest to study the extremes of atmospheric fields. Many meteorological indicators can be regarded as fields of extremes. These kinds of fields can be understood as order statistics of fields of independent and identically-distributed or short-range dependent random variables. In general, there is no analytical form for most of the distributions but asymptotic forms can usually be obtained. The extreme value theory gives us a starting point on the asymptotic distribution of such a field. Gumbel (1958) and Galambos (1978) provide overviews on the theoretical foundation. A recent work by Coles (2001) gives a more practical review on the subject. Most of the classical theories treat the issue as a univariate statistical problem. However, many atmospheric science problems are indeed multivariate especially the ones spatio-temporally referenced. It is important to consider spatial extremes as a fusion of extreme value theory and multivariate statistics. For spatially-referenced extremes, Coles and Casson (1999) propose a spatial regression for extremes based on the Generalised Extreme Value distribution.

6.2.1 Generalised Extreme Value (GEV) Distribution

The problem is originally addressed by Fisher and Tippett (1928) for assessing the limiting behaviour of normalised random variables. Assuming that $X_1, X_2, \ldots, X_n$ are independently and identically-distributed random variables with distribution F, the maximum $M_n = \max\{X_1, X_2, \ldots, X_n\}$ converges to the cumulative distribution $H(M_n)$ as $n \to \infty$, the cumulative

GEV distribution function is given by,

$$H\left(M_n\right) = \begin{cases} \exp\left[-\left\{1+\xi\left(\frac{M_n-\mu}{\sigma}\right)\right\}^{-1/\xi}\right], & 1+\xi\left(M_n-\mu\right)/\sigma > 0, \xi \neq 0 \\ \exp\left\{-\exp\left[-\frac{(M_n-\mu)}{\sigma}\right]\right\}, & \xi = 0 \end{cases}$$

The model has three parameters: a location parameter, μ; a scale parameter, σ; and a shape parameter, ξ. The representation $M_n \sim GEV(\mu, \xi, \sigma)$ often refers to M_n following a GEV distribution with the above cumulative distribution function.

Gumbel (1958) provides a theorem that the distribution of the maximum M_n is either asymptotically Fréchet, Gumbel or Weibull distributed which can be re-parametrised to a GEV distribution. The GEV distribution with $\xi = 0$ is a special case that leads to the *Gumbel* distribution. This refers to a light-tailed distribution with finite higher order moments. With the Gumbel distribution, the shape parameter can be removed to reduce the dimension of the parameter space. If ξ is negative, the distribution belongs to the *Fréchet* distribution and bounded from above. If ξ is positive, the distribution belongs to the *Weibull* distribution which is heavy-tailed with infinite higher order moments. The support of the Gumbel distribution is on the real number line but it is not the same for the other two distributions. The support for these depend on the parameters. The likelihood function of the parameters under the full model is analytically intractable and can behave very badly unlike the ones based on standard distributions such as the normal, gamma and beta distributions. Modelling with the parameter $\xi(\neq 0)$ is much more challenging than that with the Gumbel sub-model corresponding to $\xi = 0$.

Prescott and Walden (1980, 1983) propose the maximum likelihood estimation methods for fitting the GEV distribution. Consider the log-likelihood function for the GEV parameters with m independent and identically dis-

tributed random variables $z_1, \ldots, z_m$, for $\xi \neq 0$ but $1 + \xi \frac{z_i - \mu}{\sigma} > 0$,

$$l(\mu, \sigma, \xi) = -m \log \sigma - (1+1/\xi) \sum_{i=1}^{m} \log \left[1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right] - \sum_{i=1}^{m} \left[1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right]^{-1/\xi}, \tag{6.1}$$

when for $\xi = 0$,

$$l(\mu, \sigma) = -m \log \sigma - \sum_{i=1}^{m} \left(\frac{z_i - \mu}{\sigma}\right) - \sum_{i=1}^{m} \exp\left\{-\left(\frac{z_i - \mu}{\sigma}\right)\right\}. \tag{6.2}$$

The maximum likelihood estimates cannot be found analytically. Some iterative numerical solution methods are usually required. Such numerical algorithms behave much better, i.e., show more stability when applied to the $\xi = 0$ case. Apart from the maximum likelihood approach, Hosking, Wallis and Wood (1985) use a weighted moment based technique. More recently, Coles and Powell (1996) give a general review in a Bayesian context by using MCMC methods.

Generating GEV Random Variables

Generating GEV random variables is straightforward. Given the parameters, GEV random number z_p can be always generated from the inverse of the cumulative distribution function:

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left\{1 - [-\log(U[0,1])^{-\xi}]\right\}, & 1 + \xi\,(x - \mu)\,/\sigma > 0, \xi \neq 0 \\ \mu - \sigma \log(-\log(U[0,1])), & \xi = 0 \end{cases}$$

where $U[0,1]$ is a pseudo-random sample from a uniform distribution bounded by 0 and 1. The intuitive interpretation of random sample z_p is the level exceeded by the annual maximum in any particular year with probability $U[0,1]$.

6.3 Spatio-temporal Models for Extreme Value

Location parameter μ, scale parameter σ and shape parameter ν can be replaced by the spatio-temporally-varying $\mu(\mathbf{s},t)$, $\sigma_g(\mathbf{s},t)$ and $\nu(\mathbf{s},t)$ to model the spatio-temporal data as follows:

$$H\left(z(\mathbf{s},t)\right) = \begin{cases} \exp\left[-\left\{1+\nu(\mathbf{s},t)\left(\frac{z(\mathbf{s},t)-\mu(\mathbf{s},t)}{\sigma_g(\mathbf{s},t)}\right)\right\}^{-1/\nu(\mathbf{s},t)}\right], & 1+\nu\frac{(z(\mathbf{s},t)-\mu(\mathbf{s},t))}{\sigma_g(\mathbf{s},t)} \\ & > 0, \\ & \nu(\mathbf{s},t)\neq 0 \\ \exp\left\{-\exp\left[-\frac{(z(\mathbf{s},t)-\mu(\mathbf{s},t))}{\sigma_g(\mathbf{s},t)}\right]\right\}, & \nu(\mathbf{s},t)=0. \end{cases} \tag{6.3}$$

However, this general model is seldom applied in real situation due to its high complexity. The parameter ν is often physically regarded as fixed over the spatio-temporal space which represents the consistency of the tail behaviour.

6.3.1 An Extreme Value Theory Extension to the ARM

Following Huerta and Sansò (2007) and Sang and Gelfand (2007), we incorporate the generalised extreme value distribution for the response $Z(\mathbf{s},t)^2 = Z_o(\mathbf{s},t)$ in the original scale. The measurement equation can be replaced by $Z_o(\mathbf{s},t) \sim GEV(\mu(\mathbf{s},t),\sigma_g,\nu)$ with the cumulative distribution function:

$$H\left(Z_o(\mathbf{s},t)\right) = \begin{cases} \exp\left[-\left\{1+\nu\left(\frac{z_o(\mathbf{s},t)-\mu(\mathbf{s},t)}{\sigma_g}\right)\right\}^{-1/\nu}\right], & 1+\nu\frac{(Z_o(\mathbf{s},t)-\mu(\mathbf{s},t))}{\sigma_g} > 0, \\ & \nu\neq 0 \\ \exp\left\{-\exp\left[-\frac{(z_o(\mathbf{s},t)-\mu(s,t))}{\sigma_g}\right]\right\}, & \nu=0 \end{cases} \tag{6.4}$$

and the location parameter $\boldsymbol{\mu}_t$ is time-varying which forms a ARM model with the same transition equation as in the last chapter:

$$\mu(\mathbf{s}_i,t) = O(\mathbf{s}_i,t) + \epsilon(\mathbf{s}_i,t), \tag{6.5}$$

In a vector representation, we write:

$$\boldsymbol{\mu}_t = \mathbf{O}_t + \boldsymbol{\epsilon}_t, \tag{6.6}$$

where $\boldsymbol{\mu}_t = (\mu(\mathbf{s}_1, t), \ldots, \mu(\mathbf{s}_n, t))'$. We also recall the transition equation:

$$O(\mathbf{s}_i, t) = \xi_t + \rho\, O(\mathbf{s}_i, t-1) + (\beta_0 + \beta(\mathbf{s}_i))\, x(\mathbf{s}_i, t) + \eta(\mathbf{s}_i, t). \tag{6.7}$$

The addition of the random walk ξ_t makes the forecast depend more on the recent than past data. The Gaussian process ARM treats the error term σ_ϵ^2 as the measurement error. In this setting, the measurement error of the predictor variable is not taken into account. The term $\boldsymbol{\epsilon}_t$ in Equation (6.6) is no longer defined as a measurement error, the term is simply the nugget effect describing small scale spatial variability. We assume that there is no difference between the response $Z_o(\mathbf{s}, t)$ and the latent true value except for the measurement error which is presumably epistemologically different.

6.3.2 Posterior Distribution and Gibbs Sampling

The posterior distribution is derived similarly as in the last chapter. In the derivation below we keep the same notation as much as possible. Let $\boldsymbol{\theta}$ denote all the parameters, β_0, $\boldsymbol{\beta}$, ρ, σ_ϵ^2, σ_η^2, σ_β^2, σ_ξ^2, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_T)$, ν, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_T)$ and σ_g. Let $\mathbf{w}$ denote all the augmented data, $\mathbf{o}_t$ and the missing data, denoted by $z_o^*(\mathbf{s}_i, t)$, for $i = 1, \ldots, n$, $t = 1, \ldots, T$, and $\mathbf{z}$ denote all the non-missing data $Z_o(\mathbf{s}_i, t)$, for $i = 1, \ldots, n$, $t = 1, \ldots, T$. The log of the posterior distribution, denoted by $\log \pi(\boldsymbol{\theta}, \mathbf{w}|\mathbf{z})$, can be written as

$$\begin{aligned}
&-\textstyle\sum_{i=1}^n \sum_{t=1}^T \left\{1 + \nu(\frac{Z_o(s_i,t) - \mu(s_i,t)}{\sigma_g})\right\}^{1/\nu} \\
&-\tfrac{nT}{2}\log(\sigma_\epsilon^2) - \tfrac{1}{2\sigma_\epsilon^2}\textstyle\sum_{t=1}^T (\boldsymbol{\mu}_t - \mathbf{O}_t)'(\boldsymbol{\mu}_t - \mathbf{O}_t) \\
&-\tfrac{nT}{2}\log(\sigma_\eta^2) - \tfrac{1}{2\sigma_\eta^2}\textstyle\sum_{t=1}^T (\mathbf{O}_t - \boldsymbol{\vartheta}_t)' S_\eta^{-1} (\mathbf{O}_t - \boldsymbol{\vartheta}_t) \\
&-\tfrac{T}{2}\log(\sigma_\xi^2) - \tfrac{1}{2\sigma_\xi^2}\textstyle\sum_{t=2}^T (\xi_t - \xi_{t-1})^2 \\
&-\tfrac{n}{2}\log(\sigma_\beta^2) - \tfrac{1}{2\sigma_\beta^2}\boldsymbol{\beta}' S_\beta^{-1}\boldsymbol{\beta} + \log\left(\pi(\xi, \beta_0, \rho, \sigma_\epsilon^2, \sigma_\eta^2, \sigma_\beta^2, \sigma_\xi^2, \nu, \boldsymbol{\mu}, \sigma_g)\right)
\end{aligned}$$

where $\pi(\boldsymbol{\xi}, \rho, \beta_0, \sigma_\epsilon^2, \sigma_\eta^2, \sigma_\beta^2, \sigma_\xi^2, \nu, \boldsymbol{\mu}, \sigma_g)$ is the prior distribution. The full conditional distributions needed for Gibbs sampling can be derived similarly as

has been done in the last chapter. Here we only need to derive the full conditional distributions for the three new parameters μ, ξ and σ (which are non-standard) and then discuss sampling methods using the Metropolis-Hastings algorithm. The distributions of all μ, ξ and σ are:

$$\begin{aligned} f(\mu(s_i,t) \mid \boldsymbol{\theta}_-) &\propto g_{GEV}(z_o(\mathbf{s}_i,t);\mu(\mathbf{s}_i,t),\nu,\sigma_g) N_{\mu(\mathbf{s}_i,t)}(\mathbf{0},10^4) N_{O(s_i,t)}(\mu(\mathbf{s}_i,t),\sigma_\epsilon^2) \\ f(\nu \mid \boldsymbol{\theta}_-) &\propto \prod_{i=1}^n \prod_{t=1}^T g_{GEV}(z_o(\mathbf{s}_i,t);\mu(\mathbf{s}_i,t),\nu,\sigma_g) N_\nu(\mathbf{0},10^4) \\ f(\sigma_g \mid \boldsymbol{\theta}_-) &\propto \prod_{i=1}^n \prod_{t=1}^T g_{GEV}(\mathbf{z}_o(s_i,t);\mu(\mathbf{s}_i,t),\nu,\sigma_g) IG_{\sigma_g}(a,b), \end{aligned}$$

where $g_{GEV}(\cdot,\cdot)$ is the probability density function of the generalised extreme value distribution, notation $\boldsymbol{\theta}_-$ denotes the set of all parameters in $\boldsymbol{\theta}$ except the one going to be sampled; $IG_x(\cdot)$ is the density function of the inverse gamma distribution. These conditional distributions are non-standard, and as a result we use the random-walk Metropolis-Hastings algorithm to sample from them. However, note that $\boldsymbol{\mu}$ and $1/\sigma_g$ can be sampled by adaptive rejection sampling (Gilks and Wild, 1992) since their distributions are log-concave.

6.3.3 Prediction Details

We adopt the Bayesian prediction techniques of the last chapter to predict under the spatial GEV model. The posterior predictive distribution of $Z_o(\mathbf{s}',t)$ is obtained by integrating over the unknown quantities in Equation (5.7) with respect to the joint posterior distribution, i.e.,

$$\begin{aligned} \pi\left(Z_o(\mathbf{s}',t)|\mathbf{z}\right) = \int & \pi\left(Z_o(\mathbf{s}',t)|\mu(\mathbf{s}',[t])\right) \pi\left(\mu(\mathbf{s}',t)|O(\mathbf{s}',[t]),\sigma_\epsilon^2\right) \\ & \pi\left(O(\mathbf{s}',[t])|\beta(\mathbf{s}'),\boldsymbol{\theta},\mathbf{w}\right) \pi\left(\beta(\mathbf{s}')|\boldsymbol{\theta}\right) \\ & dO(\mathbf{s}',[t])\, d\beta(\mathbf{s}')\, d\boldsymbol{\theta}\, d\mathbf{w}. \end{aligned} \tag{6.8}$$

When using MCMC methods to draw samples from the posterior, the predictive distribution (6.8) is sampled by composition. In summary, we implement the following algorithm to predict $Z(\mathbf{s}',t), t = 1,\ldots,T$.

1. Draw a sample $\boldsymbol{\theta}^{(j)}$ from the full conditional distributions.
2. Draw $\beta^{(j)}(\mathbf{s}')$ using (5.10).
3. For $1 \leq t \leq T$, draw ξ_t sequentially from the posterior distribution.
4. Draw $\mathbf{O}^{(j)}(\mathbf{s}', [t])$ sequentially using (5.11). Note that the initial value value $O^{(j)}(\mathbf{s}', 0)$ is a constant for all $\mathbf{s}'$.
5. Finally draw $Z^{(j)}(\mathbf{s}', t)$ from $GEV\left(\mu^{(j)}(\mathbf{s}', t),\ \nu^{(j)},\ {\sigma_g}^{(j)}\right)$.

The posterior predictive distribution of ozone concentration is directly obtained in original scale. We use the median of the MCMC samples and the length of the 95% intervals to summarise the predictions.

The Bayesian forecast at a location $\mathbf{s}'$ and time point $T+1$ is given by the posterior predictive distribution of $Z_o(\mathbf{s}', T+1)$. We employ a similar algorithm to the one in Section 5.3.2 as follows:

1. draw ξ_{T+1} from $N(\xi_T, \sigma_\xi^2)$,
2. evaluate $O^{(j)}(\mathbf{s}', T+1) = \xi_{T+1} + \rho\, O^{(j)}(\mathbf{s}', T) + (\beta_0 + \beta(\mathbf{s}'))\, x(\mathbf{s}', T)$,
3. draw $\mu^{(j)}(\mathbf{s}', T+1)$ from $N\left(O^{(j)}(\mathbf{s}', T),\ {\sigma_\epsilon^2}^{(j)}\right)$,
4. finally draw $Z^{(j)}(\mathbf{s}', T+1)$ from $GEV\left(\mu^{(j)}(\mathbf{s}', T+1),\ \nu^{(j)},\ {\sigma_g}^{(j)}\right)$.

6.4 Examples

6.4.1 A Simulation Study

The proposed full model based on the GEV distribution is very complicated due to its constraint support. Theoretically, a multivariate prior distribution for all the location, scale and shape parameters can be assumed. However, such a prior distribution is difficult to specify and the analysis gets much

more difficult, see e.g. Coles *et al.* (2003). Henceforth, we work with the Gumbel sub-model of the full GEV based model. For simplicity, we assume $\beta = 0$ in the simulation model. Parameter estimates, provided in Table 6.1, are very close to the simulation parameters. We now proceed to modelling real ozone concentration data using the Gumbel sub-model.

6.4.2 Analysis of Ozone Concentration Data

As mentioned before, we work with the Gumbel sub-model corresponding to $\nu = 0$. Following the previous chapter we take $\beta = 0$ since non-zero β did not lead to model improvement. We compare the model with time-varying $\xi(t)$ with the fixed ξ model.

We run the MCMC algorithm with a Metropolis-Hastings within Gibbs sampling for 5000 iterations after 1000 burn-in iterations. By trial and error the parameters of the transition kernel for updating $\boldsymbol{\mu}$ and σ_g have been tuned to produce acceptance rates between 25% and 30% - this is close to the theoretical optimal acceptance rate of 0.234 derived by Roberts, Gelman and Gilks (1997).

The predictions outperform the Eta-CMAQ forecasts but are not as good as those from ARM(1) and ARM(3) in terms of the RMSE, MAE, rBIAS and rMSEP, see Tables 6.3, 6.4, 6.5, and 6.6. However, the models here demonstrate interesting results. Forecasts for August 13 show that the GEV based model (EVTARM) is better in the upper tails. Table 6.2 shows that EVTARM has smaller RMSE than that from the ARM(1) when the observed values are high. The validation plot in Figure 6.1 shows that the EVTARM forecasts are more accurate in the upper tail although those are globally inferior. Therefore, this method is still promising but further exploration needs to be done.

Table 6.1: Simulation parameters and their estimates for the Gumbel submodel ($\nu = 0$).

True Parameter	Posterior Median	2.5%	97.5%
$\rho = 0.7$	0.6872	0.5742	0.8007
$\xi = 10.0$	13.1187	5.2083	21.2047
$\sigma_\epsilon^2 = 10.0$	0.7843	0.1932	2.3623
$\sigma_\omega^2 = 400.0$	476.6424	359.3349	624.5331
$\sigma_g = 8.0$	8.5442	7.2130	10.1263

6.5 Summary

In this Chapter, an extreme value distribution model has been developed for modelling ozone concentration data. The model is shown to improve the forecasts corresponding to the high observations in the tail. The model, however, does not globally outperform the Gaussian ARM of the previous chapter due to the presence of some low values. Often, the environmental impact is assessed using the upper tail of the distributions, and hence the GEV model proposed here is of considerable value to the modelling community. This and other non-Gaussian models using transformations will form the basis of future work.

Table 6.2: RMSE of the upper tail on Aug 13th

Observed Value	ARM(1)	EVTARM
All	6.94	7.37
> 50	7.26	7.30
> 60	7.64	7.61
> 70	8.59	8.28
> 80	10.53	9.45

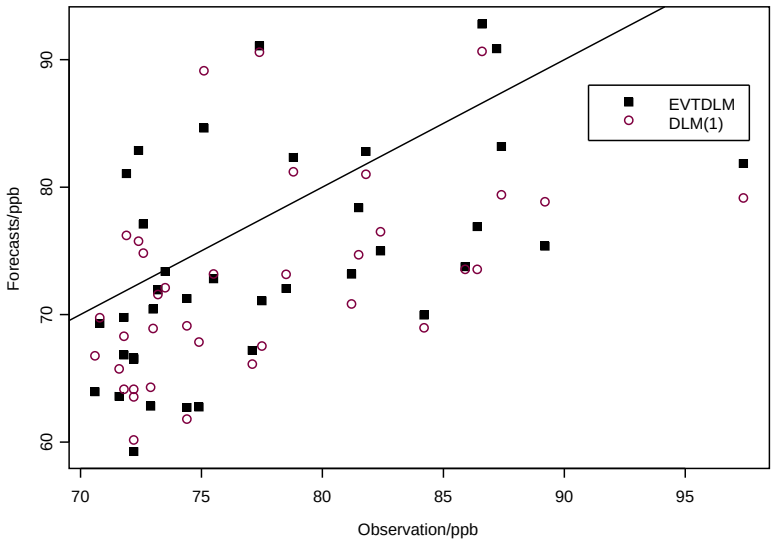


Figure 6.1: A validation plot of the upper tail on Aug 13th for the Gumbel Sub-model and ARM(1).

Table 6.3: RMSEs, MAEs, rBIAS and rMSEP of EVTARM on Aug 10: The forecasts rows represent the one-day-ahead ($1 \times 40 = 40$) set-aside validation sites. The interpolation rows represent the previous seven days ($7 \times 40 = 280$) set-aside validation sites. The two rows corresponding to Total represent all the validation data used ($8 \times 40 = 320$).

		RMSE	MAE	rBIAS	rMSEP
Forecasts	Eta-CMAQ	18.66	16.62	0.2917	0.8294
	Bayes	8.623	6.688	0.006629	0.4407
Interpolation	Eta-CMAQ	15.15	12.02	0.1886	0.5293
	Bayes	8.91	7.413	0.05506	0.2388
Total	Eta-CMAQ	15.7	12.67	0.2029	0.5748
	Bayes	8.87	7.311	0.04837	0.2549

Table 6.4: RMSEs, MAEs, rBIAS and rMSEP of EVTARM on Aug 11: The forecasts rows represent the one-day-ahead ($1 \times 40 = 40$) set-aside validation sites. The interpolation rows represent the previous seven days ($7 \times 40 = 280$) set-aside validation sites. The two rows corresponding to Total represent all the validation data used ($8 \times 40 = 320$).

		RMSE	MAE	rBIAS	rMSEP
Forecasts	Eta-CMAQ	14.83	11.94	0.1204	0.9529
	Bayes	15.22	12.55	-0.1648	0.843
Interpolation	Eta-CMAQ	16.34	13.39	0.225	0.6324
	Bayes	8.867	6.388	-0.05929	0.2762
Total	Eta-CMAQ	16.14	13.19	0.2094	0.6585
	Bayes	9.99	7.241	-0.07504	0.355

Table 6.5: RMSEs, MAEs, rBIAS and rMSEP of EVTARM on Aug 12: The forecasts rows represent the one-day-ahead ($1 \times 40 = 40$) set-aside validation sites. The interpolation rows represent the previous seven days ($7 \times 40 = 280$) set-aside validation sites. The two rows corresponding to Total represent all the validation data used ($8 \times 40 = 320$).

		RMSE	MAE	rBIAS	rMSEP
Forecasts	Eta-CMAQ	12.7	10.01	0.1388	0.5212
	Bayes	10.72	8.908	-0.07087	0.4698
Interpolation	Eta-CMAQ	16.37	13.31	0.2232	0.7724
	Bayes	8.009	6.531	0.03444	0.3006
Total	Eta-CMAQ	15.92	12.86	0.2089	0.689
	Bayes	8.432	6.857	0.01658	0.2974

Table 6.6: RMSEs, MAEs, rBIAS and rMSEP of EVTARM on Aug 13: The forecasts rows represent the one-day-ahead ($1 \times 40 = 40$) set-aside validation sites. The interpolation rows represent the previous seven days ($7 \times 40 = 280$) set-aside validation sites. The two rows corresponding to Total represent all the validation data used ($8 \times 40 = 320$).

		RMSE	MAE	rBIAS	rMSEP
Forecasts	Eta-CMAQ	14.78	12.29	0.1867	0.6879
	Bayes	9.563	8.18	-0.08053	0.4585
Interpolation	Eta-CMAQ	15.61	12.85	0.2119	0.6447
	Bayes	8.477	6.249	-0.07033	0.2754
Total	Eta-CMAQ	15.51	12.79	0.2084	0.6299
	Bayes	8.615	6.482	-0.07176	0.2804

Chapter 7

Conclusion and Future Work

7.1 Conclusion

This thesis has been motivated by the forecasting problem of environmental data. The aim of this investigation has been to produce more accurate forecasts of ozone concentration levels along with their uncertainty measures. A number of models with varying degree of forecasting ability has been proposed and experimented with. The major contributions of this thesis are:

- **Hierarchical Gaussian random effects models for producing probabilistic forecasts for hourly data**
  A fast and analytical GRE model has been shown to predict the 8-hour average ozone concentration of the current hour and the two hours ahead. The Eta-CMAQ model outputs are assimilated to the observed data as further information contributing to reconstructing the ground truth. A novel dimension reduction method is also developed. An alternative way based on delta method in approximating spatially-varying variance is also highlighted. High resolution model based Bayesian predictive maps are obtained and displayed for dissemination purposes. A portion of the work, written as Sahu, Yip and Holland (2010), presented

here has been published. Simulation and real data examples show that the spatio-temporal GRE models outperform the non-spatial models in terms of both predictions and model fitting. The use of geostatistical and exploratory data analysis tools are also emphasised.

- **Dynamic linear models for data assimilation of the daily 8-hour maximum ozone concentration data**

  Again the Eta-CMAQ output acts as further information to the modelling but this time also for the daily 8-hour maximum concentration data. An auto-regressive model for the mean response along with the Eta-CMAQ output as a co-variate has been shown to work well for the purposes of forecasting the next day's ozone level. A paper based on the modelling developments of this chapter has appeared as Sahu, Yip and Holland (2009).

- **Dynamic linear models with generalised extreme value distribution**

  The top level Gaussianity assumption of the previous chapter is replaced by the GEV distribution to accommodate non-Gaussianity of the spatial extremes. The complexity of the support constraints has forced us to work with the Gumbel sub-model of the GEV distribution. Model fitting has been done by using the Metropolis-Hastings algorithm within the Gibbs sampler. A simulation and a real data example show better forecasts corresponding to the observation values which are higher. This shows that the GEV models are promising for modelling spatial extremes, though further investigation is necessary to understand the full capability of the models.

7.2 Future Work

7.2.1 Non-Gaussian Measurement Error Models Based on Skew-Normal Distribution

The non-Gaussian measurement error models can be further improved by using other family of distributions. A random variable X has a skew-normal distribution (see e.g. Sahu *et al.*, 2003) with location parameter μ, scale parameter σ and skewness parameter α, denoted $X \sim SN(\mu, \sigma, \alpha)$ and it has the density function,

$$f(x) = 2\phi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\alpha\frac{x-\mu}{\sigma}\right).$$

Its mean and variance are

$$E(x) = \mu + \sigma\sqrt{\frac{2}{\pi}}\delta \text{ and } \text{Var}(x) = \sigma^2(1 - \frac{2}{\pi}\delta)$$

where ϕ and Φ are the density and cumulative distribution function of the standard normal distribution. The parameters α and δ have the relationship that $\alpha(\delta) = \delta/\sqrt{1-\delta^2}$ and $\delta(\alpha) = \alpha/\sqrt{1+\alpha^2}$.

There is a proposition (Dalla Valle, 2004) that ensures the skew-normal random variable can be generated from two normal i.i.d. random variables. If Y and W are independent $N \sim (0, 1)$ random variables, and Z is set equal to Y conditionally on $\alpha Y > W$, then $Z \sim SN(0, 1, \alpha)$. It is sufficient to generate Y and W i.i.d. $N \sim (0, 1)$, and put $Z = Y$ if $\alpha Y > W$ and $Z = -Y$ if $\alpha Y \leq W$. Therefore the random variable $X \sim SN(\mu, \sigma, \alpha)$ can be generated by $X = \mu + \sigma Z$.

From a well-known theorem (see example in Dalla Valle, 2004) that the maximum of two independent normally distributed random variables is a skewed-normal variable. Intuitively this can be viewed as an approximation to the daily 8-hour maximum ozone concentration levels since the proportion of all the daily 8-hour maxima in our data attributed to 4pm and 5pm is

greater than 60%. The suggested model is similar to the extreme-value-theory-based model in terms of its hierarchical structure.

7.2.2 Two-stage Joint Modelling Approach

Sahu, Gelfand and Holland (2010) develop a framework for joint modelling of point and grid referenced spatio-temporal data for wet deposition. This method incorporates at least two hierarchical structures. In the framework, the Eta-CMAQ output is no longer counted as a covariate but given a separable hierarchical structure. The observed deposition and the precipitation are highly related. In the model, biased Eta-CMAQ output serve as extra information for the observed data. This methodology is close to the work of Fuentes and Raftery (2005). The same approach can be applied to the forecasting problems here, but it is noted that slower MCMC methods will be required to obtain the results.

7.2.3 Heteroscedastic Models

In many applications, the variance also varies over region and changes with time due to the change in meteorological conditions and economic activities. Usually, in many financial applications, heteroscedastic models such as auto-regressive conditionally heteroscedastic (or ARCH) models and generalised auto-regressive conditionally heteroscedastic (or GARCH) models are used to capture the stochastic change of variance (Chatfield, 2004). The technique here uses a local conditional variance. A typical example is *stochastic volatility model*. For example, a stochastic volatility model for the nugget effect can be defined as a random walk process with another "white noise" component,

$$\log\left(\sigma_\epsilon^2(t)\right) \sim N\left(\log(\sigma_\epsilon^2(t-1)), \sigma_E^2\right) \tag{7.1}$$

where σ_E^2 is a "white noise" term for the above process. As the spatial effect is multivariate, the stochastic volatility (variance) is spatially dependent. The technique of discounted variance learning method (West and Harrison, 1997 and Pitt and Shephard, 1999) uses the inverse Wishart distribution to capture the spatial variation. This method is analogous to an AR process for the covariance matrix but in a multivariate context. Although this model complicates the computation a lot, it will be useful if the number of parameters of the matrix can be reduced.

Appendix A

Common Statistical Distributions

Here we list the definitions, first moments, variances and other important issues of the probability distributions that were used in this thesis.

A.1 Univariate Normal Distribution

X has the normal distribution with mean μ and variance σ^2, denoted $X \sim N(\mu, \sigma^2)$ if it has the probability density function:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right], \ -\infty < x < \infty.$$

A.2 Student's t Distribution

X has a Student's t distribution on ν degrees of freedom, denoted $X \sim t_\nu$ and it has a density function,

$$f(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi\nu}\Gamma(\nu/2)} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}, \ -\infty < x < \infty.$$

The mean and variance are

$$E(X) = 0 \text{ and } \mathrm{Var}(X) = \nu/(\nu - 2), \quad \text{when } \nu > 2.$$

A.3 Gamma Distribution

X has a gamma distribution on two parameters α and β, denoted $X \sim \Gamma(\alpha, \beta)$ and it has a density function,

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) \text{ for } 0 < x < \infty.$$

Its mean and variance are

$$E(X) = \alpha/\beta \text{ and } \mathrm{Var}(X) = \alpha/\beta^2.$$

A.4 Inverse Gamma Distribution

X has an inverse gamma distribution on two parameters α and β, denoted $X \sim\ IG(\alpha, \beta)$. If $1/X \sim \Gamma(\alpha, \beta)$, then $X \sim IG(\alpha, \beta)$. Its mean and variance are

$$E(X) = \frac{\beta}{\alpha - 1} \text{ and } \mathrm{Var}(X) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}, \text{ when } \alpha > 2.$$

A.5 Multivariate Normal Distribution

Multivariate normal distribution is a multivariate generalisation of univariate normal distribution. A vector $\mathbf{X}$ has a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix Σ, denoted $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$. A real random variable $\boldsymbol{X}$ is said to be normally distributed with mean vector $\boldsymbol{\mu}$ and covariance Σ if and only if

$$f(\boldsymbol{x}) = (2\pi\Sigma)^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \mu)'\mathbf{\Sigma^{-1}}\mathbf{x} - \mu)\right], \quad -\infty < \boldsymbol{x} < \infty.$$

A.6 Multivariate Student's t Distribution

A k-dimensional random vector $\mathbf{X}$ has a multivariate student t distribution, denoted $\mathbf{X} \sim St(\mu, \mathbf{\Sigma}, \alpha)$ and it has a density function,

$$f(\mathbf{x}) = c\left[1 + \frac{1}{\alpha}(\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]^{-(\alpha+k)/2}$$

where $c = \frac{\Gamma(\frac{1}{2}(\alpha+k))}{\Gamma(\frac{1}{2}\alpha)(\alpha\pi)^{k/2}} \mid \Sigma \mid^{-1/2}$. Its mean and variance are

$$E(\mathbf{X}) = \boldsymbol{\mu} \text{ and } \mathrm{Var}(\mathbf{X}) = \frac{\alpha}{\alpha - 2}\Sigma, \quad \text{when } \alpha > 2.$$

References

Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions. New York, Dover.*

Abt, M. and Welch, W. (1998). Fisher information and maximum likelihood estimation of covariance parameters in Gaussian stochastic processes. *Canadian Journal of Statistics,* **26 (1)**, 127-137.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* **19 (6)**, 716-723.

Atkinson, M. and Lloyd, D. (1998). Mapping precipitation in Switzerland with ordinary and indicator kriging. *Journal of Geographic Information and Decision Analysis,* **2(2)**, 65-76.

Banerjee, S. (2005). On geodetic distance computations in spatial modeling. *Biometrics,* **61 (2)**, 617-625.

Banerjee, S., Carlin, B.P. and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC.

Banerjee, S., Gelfand, A.E., Finley, A.O. and Sang H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society, Ser. B,* **70**, 825-848.

Bayarri, M.J., Berger, J.O., Paulo, R., Sacks, J., Cafeo, J.A., Cavendish,

J., Lin, Chin-Hsu and Tu, Jian (2007). A framework for validation of computer models. *Technometrics,* **49 (2)**, 138-154.

Berger, J.O., De Oliveria, V., and Sansò, B. (2001). Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association,* **96**, 1361-1374.

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian theory.* Wiley.

Berrocal, V.J., Gelfand, A.E., and Holland, D.M. (2008). A spatio-temporal downscaler for output from numerical models. *Submitted.*

Bochner, S. (1960). *Harmonic analysis and the theory of probability.* University of California Press.

Brown, P. J., Le, N. D., Zidek, J. V. (1994). Multivariate spatial interpolation and exposure to air pollutants. *Canadian Journal of Statistics,* **22**, 489-510.

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician,* **39**, 83-87.

Carroll, R. J., Chen, R., George, E. I., Li, T.H., Newton, H.J., Schmiediche, H. and Wang, N. (1997). Ozone exposure and population density in Harris County, Texas. *Journal of the American Statistical Association,* **92**, 392-404.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science,* **10 (3)**, 273-304.

Chandler, R.E. (2005). On the use of generalized linear models for interpreting climate variability. *Environmetrics,* **16**, 699-715.

Chang, H., Fu, A. Q., Le, N. D. and Zidek, J. V. (2007). Designing Environmental Monitoring Networks for Measuring Extremes. *Environmental and Ecological Statistics,* **14(3)**, 301-321.

Chatfield, C. (2004). *The analysis of time series: an introduction.* Chapman & Hall/CRC.

Chen, C-F. (1985). On asymptotic normality of limiting density functions with Bayesian Implications., *Journal of the Royal Statistical Society, Ser. B,* **47 (3)**, 540-546.

Chilès, J-P. and Delfiner, P. (1999). *Geostatistics, modeling spatial uncertainty.* Wiley.

Ching, J., and D. Byun (Eds.) (1999). Science algorithms of the EPA models-3 community multi-scale air quality (CMAQ) modeling system, *Rep. EPA/ 600/R-99/030, Natl. Exposure Res. Lab., Research Triangle Park, N.C.*

Chib, S. (1995). Marginal likelihood from the Gibbs ouput. *Journal of the American Statistical Association,* **90**, 1313-1321.

Clayton, D. and Kaldor, J. (2005). Empirical bayes estimates of age-standardized relative risks for use in disease mapping *Biometrics,* **43 (3)**, 671-681.

Coles, S. (2001). *An introduction to statistical modeling of extreme values.* Springer.

Coles, S. and Casson, E. (1999). Spatial regression models for extremes. *Extremes,* **1**, 449-468.

Coles, S., Pericchi, L. R. and Sisson, S (2003). A fully probabilistic approach to extreme rainfall modeling. *Journal of Hydrology,* **273**, 35-50.

Coles, S. and Powell, E. (1996). Bayesian methods in extreme value modelling: A review and new developments. *International Statistical Review,* **64**, 119-136.

Cooley, D., Naveau, P., Jomelli, V., Rabatel, A. and Grancher, D. (2006). A Bayesian hierarchical extreme value model for lichenometry. *Environmetrics,* **17**, 555-574.

Cowles, M. K., and Zimmerman, D.L (2003). A Bayesian space-time analysis of acid deposition data combined from two monitoring networks *Journal of Geophysical Research,* **108**, 9006, doi:10.1029/2003JD004001.

Cressie, N.A.C. (1993). *Statistics for Spatial Data. Revised ed..* John Wiles & Son Inc.

Cressie, N.A.C. and Huang, H. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association,* **94**, 1330-1340.

Cressie, N., Frey, J., Harch, B., and Smith, M. (2006). Spatial prediction on a river network. *Journal of Agricultural, Biological, and Environmental Statistics,* **11**, 127-150.

Curriero, F.C. (2006). On the use of non-Euclidean distance measures in geostatistics. *Mathematical Geology,* **38**, **8**, 907-926.

Daley, R. (1993). *Atmospheric Data Analysis.* Cambridge University Press.

Dalla Valle, A. (2004). The skew-normal distribution, *Skew-elliptical distributions and their applications: a journey beyond normality (Edited by M. Genton), Chapter 1*, Chapman & Hall/CRC.

Davis, P.K. (1992). *Generalizing concepts and methods of verification, validation, and accreditation (VV&A) for military simulations*, RAND, R-4249-ACQ.

De Cesare, L., Myers, D., and Posa, D. (2001). Estimating and modeling space-time correlation structures. *Statistics & Probability Letters,* **51**, 9-14.

De Iaco, S., Myers, D., and Posa, D. (2002). Nonseparable space-time covariance models: some parametric families. *Mathematical Geology,* **34**, **1**, 23-42.

DiCiccio, T. J., Kass, R. E., Raftery A. and Wasserman L. (1997). Computing bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association,* **92**, 903-915.

Dutilleul, P. (1999). The MLE algorithm for the matrix normal distribution. *J. Statist. Comput. Simulation,* **64**, 105-123.

Ecker, M. D. and Gelfand, A. E. (1997). Bayesian variogram modeling for an isotropic spatial process. *Journal of Agricultural, Biological and Environmental Statistics,* **2**, 347-369.

Eder, B., Kang, D., Mahur, R., Yu, S., and Schere, K. (2006) An operational evaluation of the Eta-CMAQ air quality forecast model. *Atmospheric Environment*, **40**, 4894–4905.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability and its Applications,* **14**, 153-158.

EPA (1999). *Smog – who does it hurt? - What you need to know about ozone and your health*, the United States Environmental Protection Agency.

Fisher, R and Tippett, L. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society* **24**, 180-190.

Fuentes, M. (2001). A high frequency kriging approach for non-stationary environmental processes. *Environmetrics,* **12**, 469-483.

Fuentes, M. and Raftery, A. (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models *Biometrics,* **61 (1)**, 36-45.

Galambos, J. (1978). *The asymptotic theory of extreme order statistics*. John Wiles & Son Inc.

Gandin, L. S. (1963). *Objective analysis of meteorological fields*. Gidrometeorizdar, (Translated from Russian by Israeli Program for Scientific Translations in 1965.)

Gelfand, A.E. and Ghosh, S.K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika,* **85**, 1-11.

Gelfand, A.E. and Sahu, S.K (2009). Combining monitoring data and computer model output in assessing environmental exposure. To appear in *Handbook of Applied Bayesian Analysis, edited by O'Hagan A. and West, M.*

Gelfand, A.E., Zhu, L. and Carlin, B.P. (2001). On the change of support problem for spatio-temporal data. *Biostatistics,* **2**, 31-45.

Gelman, A., Carlin, J., Stern, H. and Rubin D.(2004). *Bayesian data analysis*. Chapman & Hall/CRC.

Gelman, A. (2007). Comment: Bayesian Checking of the Second Levels of Hierarchical Models. *Statistical Science,* **22 (3)**, 349-352.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **6(6)**, 721-741.

Genton, M. (2007). Separable approximations of space-time covariance matrices. *Environmetrics,* **18**, 681-695.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo Methods in Practice*. Chapman & Hall/CRC.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics,* **41(2)**, 337-348.

Gilleland E. and Nychka, D. (2006). Statistical models for monitoring and regulating ground-level ozone. *Environmetrics,* **16**, 535-546.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association,* **97**, 590-600.

Gneiting, T. and Sasvári, Z. (1999). The characterization problem for isotropic covariance functions. *Mathematical Geology,* **31**, 105-111.

Graham A. (1981). *Kronecker products and matrix calculus with applications*. John Wiley & Sons, New York.

Grattan-Guinnes I. (2005). *Landmark writings in western mathematics 1640-1940*. Elsevier.

Guan, Y. , Sherman, M. and Calvin, J.A. (2004). A nonparametric test for spatial isotropy using subsampling . *Journal of the American Statistical Association,* **99**, 810-821.

Gumbel, E.J. (1958). *Statistics of extremes*. New York, Columbia University Press.

Guttorp, P., Meiring, W. and Sampson, P. D. (1994). A space-time analysis of ground level ozone data. *Environmetrics,* **5**, 241-254.

Handcock, M. S. and Stein, M. L. (1993). A Bayesian analysis of kriging. *Technometrics*, **35**, 403–410.

Haslett, J. and Raftery A.E. (1989). Space-time modelling with long-memory dependence: Assessing Ireland's Wind Power Resource. *Applied Statistics,* **38, 1**, 1–50.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications *Biometrika,* **57 (1)**, 97-109.

Hedges, S. (1999). Ozone Monitoring, Mapping, and Public Outreach: Delivering Real-Time Ozone Information to Your Community. *Technical Report EPA/625/R-99/007, United States Environmental Protection Agency.*

Higdon, D., Swall, J., Kern, J. (1999). Non-stationary spatial modelling. Bayesian Statistics 6, (Eds) JM Bernardo *et al.*. Oxford: Oxford University Press, 761-768.

Hodges, J.S. and Dewar, J.A. (1992), *Is it your or your model talking? A framework for model validation*, RAND, R-4114-AF/A/05D

Hosking, J.R.M., Wallis, J.R. and Wood E.F. (1985), Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics,* **27**, 251-261.

Huerta, G. and Sansò, B. (2007). Time-varying models for extreme values. *Environmental and Ecological Statistics,* **14, 1**, 285–299.

Hughes-Oliver, J.M., Gonzalez-Farias, G., Lu, J.-C., and Chen, D. (1998), Parametric nonstationary correlation models. *Statistics & Probability Letters,* **40**, 267-278.

Huang, H.-C., Martinez, F., Mateu, J., and Montes, F. (2007), Model comparison and selection for stationary space-time models. *Computational Statistics & Data Analysis,* **51**, 4577-4596.

Jeffreys H. (1961). *Thoery of probability.* 3rd Ed. OUP.

Jones, C. (1997). *Geographical information systems and computer cartography.* Prentice Hall.

Jun, M. and Stein, M. L. (2004). Statistical Comparison of Observed and CMAQ Modeled Daily Sulfate Levels. *Atmospheric Environment,* **38**, 4427-4436.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors, *Journal of the American Statistical Association,* **90***, 773-795.*

Kennedy, M. and O'Hagan, A. (2001). Bayesian calibration of computer models (with discussion)., *Journal of the Royal Statistical Society, Ser. B,* **63***, 425-464.*

Kibria, B. M. G., Sun, L., Zidek, J. V., and Le, N. D. (2002). Bayesian spatial prediction of random space-time fields with application to mapping PM2.5 exposure. *Journal of the American Statistical Association,* **97**, 112-124.

Kolmogorov, A. N. (1941). Interpolation and extrapolation of stationary random sequences, *Izv. Akad. Nauk SSSR Ser. Mat.*, **5**, 3–14

Krige, D.G. (1951). A statistical approach to some mine valuations and allied problems at the Witwatersrand. *Master's thesis of the University of Witwatersrand*

Laud, P.W. and Ibrahim, J.G. (1995). Predictive model selection. *Journal of the Royal Statistical Society. B,* **57**, 247-262.

Lawson, A.B. and Denison, D.G.T. (2002). *Spatial cluster modelling.* Chapman & Hall/CRC, Boca Raton, FL.

Le, N. D., Sun, W., Zidek, J. V. (1997). Bayesian multivariate spatial interpolation with data missing by design. *Journal of the Royal Statistical Society, Ser. B,* **59**, 501-510.

Le, N. D. and Zidek, J. Z. (1992). Interpolation with uncertain spatial covariances: a Bayesian alternative to kriging. *Journal Multivariate Analysis,* **43** , 351-374.

Le, N.D. and Zidek, J.Z. (2006). *Statistical analysis of environmental space-time processes.* Springer.

Li, B., Eriksson, M. , Srinivasan, R. and Sherman, M. (2007). A geostatistical method for Texas NexRad data calibration. *Environmetrics,* **19**, 1-19.

Li, B., Genton, M.G. and Sherman, M. (2007). A nonparametric assesment of properties of space-time covariance functions. *Journal of the American Statistical Association,* **102**, 736-744.

Li, B., Genton, M.G. and Sherman, M. (2008). On the asymptotic joint distribution of sample space-time covariance estimators. *Bernoulli,* **14**, 229-248.

Lorenc, A. C. (1986). Analysis methods for numerical weather prediction. *The Quarterly Journal of the Royal Meteorological Society*, **112**, 1177-1194.

Lorenz, E. N. (1963). Deterministic non-periodic flow. *Journal of the Atmospheric Sciences*, **20**, 130-141.

Maling, D.H. (1992). *Coordinate systems and map projections.* Oxford: Pergamon

Mark, A.F. and Esler, A.E. (1970). An assessment of the point-centred quarter method of plotless sampling in some New Zealand forests., *Proceedings of the New Zealand Ecological Society,* **17**, 106-110

Matheron, G. (1971). The theory of regionalized variables and its application., *Les Cahiers du Centre de Morphologie Mathematique de Fontainebleau, No. 5. Ecole Nationale Superieure des Mines de Paris.*

Matheron, G (1989). *Estimating and choosing.* trans. A. M. Hasofer. Springer-Verlag.

McMillan, N., Bortnick, S. M., Irwin, M. E. and Berliner, M. (2005). A hierarchical Bayesian model to estimate and forecast ozone through space and time. *Atmospheric Environment*, **39**, 1373–1382.

Meng, X. L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica,* **6**, 831-860.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics,* **21 (6)**, 1087-1092.

Mezić, I. and Runolfsson, T. (2008). Uncertainty propagation in dynamical systems. *Automatica,* **44(12)**, *3003-3013.*

Møller, J. and Waagepetersen, R. P. (2004). *Statistical inference and simulation for spatial point processes.* Chapman & Hall/CRC.

Montiel, S. and Ros, A. (2005). *Curves and surfaces.* AMS & RSME.

Moyeed, R. and Papritz, A. (2002). An empirical comparison of kriging methods for nonlinear spatial point prediction. *Mathematical Geology,* **34(4)***, 365-386.*

Newton, M. A. and Raftery, A.E. (1994). Approximate bayesian inference by the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society, Ser. B,* **56**, 3-48.

Oberkampf, W. L., DeLand, S. M., Rutherford, B. M., Diegert, K. V. and Alvin, K. F. (2002). Error and uncertainty in modeling and simulation. *Reliability Engineering and Systems Safety,* **75**, 333-357.

Oehlert, G. W. (1992). A note on Delta method. *The American Statistician,* **46**, 27-29.

O'Hagan A. and Forster, J. (2004). *Kandall's advanced theory of statistics: Volumae 2B: Bayesian Inference. Hodder Arnold*

Pintore, A. and Holmes, C.C. (2004). Non-stationary covariance functions via spatially adaptive spectra. *Journal of American Statistical Association, Accepted.*

Pitt, M. K. and Shephard, N. (1999). Time varying covariances: a factor stochastic volatility approach (with discussion). *Bayesian Statistics 6. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.),* **46**, 547-570. Oxford.

Preparata, F.P. and Shamos, M.I. (1985). *Computational geometry: an introduction. Springer-Verlag.*

Prescott, P. and Walden, A. (1980). Maximum likelihood estimations of the parameters of the generalized extreme value distribution. *Biometrika,* **67**, 723-724.

Prescott, P. and Walden, A. (1983). Maximum likelihood estimations of the three-parameter generalized extreme value distribution from censored samples. *Journal of Statistical Computing and Simulation,* **16**, 241-250.

Raftery, A.E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika,* **83**, 251-266.

Raftery, A. E., Haslett, J. and McColl, E. (1982). Wind power: a space time process? *Time series analysis: theory and practice*, Springer-Verlag.

Rappold, A. G., Gelfand, A. E. and Holland, D. M. (2008). Modeling mercury deposition through latent space-time processes. *Journal of the Royal Statistical Society, Ser. C,* **57***, 187-205.*

Roberts, G.O., Gelman, A. and Gilks, W.R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithm. *Annual of Applied Probability* **7**, 110-120.

Sahu S. K., Dey, D. K. and Branco, M. (2003). A new class of multivariate skew distributions with applications to Bayesian regression models. *The Canadian Journal of Statistics,* **11**, 61-86.

Sahu S. K., Gelfand, A. E. and Holland, D. M. (2006). Spatio-temporal modeling of fine particulate matter. *Journal of Agricultural, Biological, and Environmental Statistics,* **11**, 61-86.

Sahu, S. K., Gelfand, A. E. and Holland, D. M. (2007). High resolution space-time ozone modeling for assessing trends. *Journal of American Statistical Association,* **102**, 1221-1234.

Sahu, S. K., Gelfand, A. E. and Holland, D. M. (2010). Fusing point and areal level space-time data with application to wet deposition. To appear in *Applied Statistics: Journal of the Royal Statistical Society.*

Sahu, S. K. and Mardia, K. V. (2005). A Bayesian Kriged-Kalman model for short-term forecasting of air pollution levels. *Journal of the Royal Statistical Society, Ser. C,* **54**, 223-244

Sahu, S. K., Yip, S. and Holland, D. M. (2009). Improved space-time forecasting of next day ozone concentrations in the eastern US. *Atmospheric Environment,* **43**, *494–501.*

Sahu, S. K., Yip, S. and Holland, D. M. (2010). A fast Bayesian method for updating and forecasting hourly ozone levels. Environmental and Ecological Statistics, DOI 10.1007/s10651-009-0127-y.

Sampson. P.D. and Guttorp, P.(1992) Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association,* **87**, *108-119.*

Schabenberger, O. and Gotway, C.A. (2004). *Statistical methods For spatial data analysis*. Chapman & Hall/CRC.

Schmidt, A. and O'Hagan, A. (2003). Bayesian inference for nonstationary spatial covariance structures via spatial deformations. *Journal of the Royal Statistical Society, Ser. B,* **65 (3)**, *743-758*

Shaddick, G. and Wakefield, J. (2002). Modelling daily multivariate pollutant data at multiple sites. *Journal of the Royal Statistical Society, Ser. C,* **51**, *351-372*

Smith, R. L., Kolenikov, S. and Cox, L. H. (2003). Spatio-Temporal modelling of PM2:5 data with missing values. *Journal of Geophysical Research-Atmospheres,* **108**, *D249004, doi:10.1029/2002JD002914.*

Spiegelhalter D.J., Best N.G., Carlin B.P. and van der Linde A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Ser. B,* **64**, *583-640*

Stein, M.L. (1999). *Statistical interpolation of spatial data: some theory for Kriging*. Springer.

Stephenson J. (2006). *Non-stationary spatial statistics in the geosciences*, Ph.D thesis, Department of Eath Science and Engineering, Imperial College of Science, Technology and Medicine.

Sun L., Zidek, J. V., Le, N. D. and Ozkaynak, H. (2000). Interpolating Vancouver's daily ambient PM10 field. *Environmetrics,* **11**, *651-663.*

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Tukey, J. W. (1980). Exploratory data analysis is an attitude, a flexibility, and a reliance on display, NOT a bundle of techniques, and should be so taught. *The American Statistician,* **34**, *23-25.*

Walker, A.M. (1969). On the asymptotic behaviour of posterior distributions., *Journal of the Royal Statistical Society, Ser. B,* **31 (1)**, *80-88.*

Wand, M. P. and Jones M. C. (1994). *Kernel Smoothing*. Chapman & Hall, London.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer.

Wikle, C. K. (2003). Hierarchical models in environmental science. *Int. Stat. Rev.,* **71**, *181-199.*

Wikle, C. K. and Berliner L. M. (2006). A Bayesian tutorial for data assimilation Physica D, **230**, *1–16.*

Yaglom, A.M (1962). *An introduction to the theory of stationary random functions*. Prentice-Hall Inc.

Yip, S. and Sahu S. K. (2009). Bayesian hierarchical modelling for data assimilation of past observations and numerical model forecasts, *Technical Report, University of Southampton.*

Yu, S., Mathur, R., Schere, K., Kang, D., Pleim, J. and Otte, T. (2007). A detailed evaluation of the Eta-CMAQ forecast model performance for O3, its related precursors, and meteorological parameters during the 2004 ICARTT study. Journal of Geophysical Research, **112***, D12S14, doi:10.1029/2006JD007715.*

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of American Statistical Association,* **99***, 250–261.*

Zhang, H. and Zimmerman, D.L. (2005). Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika,* **92 (4)***, 921-936.*

Zhu Z. and Zhang, H. (2006). Spatial sampling design under the infill asymptotic framework. *Environmetrics,* **17***, 323–337.*

Zidek, J. V., Sun, L., Le, N. and Ozkaynak, H.(2002). Contending with space-time interaction in the spatial prediction of pollution: Vancouver's hourly ambient PM10 field. *Environmetrics,* **13***, 595-613.*

Zimmerman, D. L. and Holland, D. M. (2004). Complementary co-kriging: spatial prediction using data combined from several environmental monitoring networks. *Environmetrics,* **16***, 219-234.*