University of Southampton

Faculty of Social and Human Sciences

School of Mathematics

Bayesian Analysis of Daily Maximum Ozone Levels

by

Khandoker Shuvo Bakar

M.S. & B.Sc. (Hons.)

Thesis for the degree of Doctor of Philosophy

January 2012

Abstract

Ground level ozone is one of the six criteria primary pollutants that is monitored by the United States Environmental Protection Agency. Statistical methods are increasingly being used to model ground level ozone concentration data. This thesis is motivated by the need to perform practical data analysis, and to develop methods for modelling of ozone concentration data observed over a vast study region in the eastern United States (US).

For the purposes of analysis, we use two space-time modelling strategies: the dynamic linear models (DLM) and the auto-regressive (AR) models and obtain predictions and forecasts for set aside validation data. These methods are developed under the Bayesian paradigm and MCMC sampling techniques are used to explore the posterior and predictive distributions. Particularly, for analysis, we use a subset data set from the state of New York to illustrate the methods. Both the DLM and AR modelling approaches are compared in detail using the predictive and forecast distributions induced by them. The comparisons are facilitated by a number of theoretical results. These show better properties for the AR models under some conditions, which have been shown to hold for the real life example that we considered.

To address the challenge of modelling large dimensional spatio-temporal ozone concentration data, we adopt Gaussian predictive processes (GPP) technique and propose a rich hierarchical spatio-temporal AR model. The important utility of this method lies in the ability to predict the primary ozone standard at any given location for the modelled period from 1997-2006 in the eastern US. Different sensitivity analyses are performed, and, in addition, hold-out data sets are used for model validation. Specifically, this new modelling approach has been illustrated for evaluating meteorologically adjusted trends in the primary ozone standard in the eastern US over the 10 year period. This helps in understanding spatial patterns and trends in ozone levels, which in turn will help in evaluating emission reduction policies that directly affect many industries.

Forecasting of ozone levels is also an important problem in air pollution monitoring. We compare different spatio-temporal models for their forecasting abilities. The GPP based models provide the best forecast for set aside validation data.

In addition, in this thesis we use computer simulation model output as an explanatory variable for modelling the observed ozone data. Thus, the proposed methods can also be seen as a spatio-temporal downscaler model for incorporating output from numerical models, where the grid-level output from numerical models is used as a covariate in the point level model for observed data. This type of space and time varying covariate information enriches the regression settings like the methods used in this thesis.

Currently there is no package available that can fit space-time environmental data using Bayesian hierarchical spatio-temporal models. In this thesis we, therefore, develop a software package named `spTimer` in `R`. The `spTimer` package with its ability to fit, predict and forecast using a number of Bayesian hierarchical space-time models can be used for modelling a wide variety of large space-time environmental data. This package is built in `C` language to be computationally efficient. However, this `C`-code is hidden from the user and the methods can be implemented by anyone familiar with the `R` language.

This thesis can be extended in several ways for example, for multivariate data, for non-Gaussian first stage data, and for data observed in environmental monitoring of stream networks.

Acknowledgements

I would like to thank my supervisor Dr. Sujit K. Sahu for his great influence developing my knowledge with deep insights in statistics and research. He helped me to improve my ability to explain numbers in writings. I would have never been able to finish this thesis without his guidance and unlimited times he offered me.

I am very grateful to Professor Jonathan Forster for his time, advice and appreciation to me. It was a great working and learning environment at the University of Southampton which I will always be missing.

I am deeply indebted to Dr. Dave Woods who was always very much supportive to me.

Special thanks to Bernard for proof reading over the weekend. Many thanks to Shidah, Antony, Sean, Kieran and Natalie for the time and knowledge we shared together during my study period. Thanks are also due to Garry and Kulvir in the office of the School of Mathematics for their cordial response to me.

Finally, never enough thanks to Asha for her continuous support and my parents for their inspiration.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis considers analysis and modelling of daily maximum eight-hour average ground level ozone (O_3) concentrations. Specifically, we use Bayesian modelling approaches to predict and forecast ozone levels for spatio-temporal fields. These are important problems since ground level ozone concentrations have harmful effects on human health. In addition, it is important to find better models and faster computational techniques to obtain predictions and forecasts. These are the primary motivations of this thesis. Before going further on modelling approaches, we describe what is ozone and why we need to analyse ozone concentrations in the next section.

1.1 What is Ozone?

Ozone is a colourless and odourless reactive gas that occurs naturally in the atmosphere. It occurs in two layers of the earth's surface, namely the troposphere and the stratosphere. The troposphere is the lowest portion of earth's atmosphere and it ranges between 4 to 11 miles above ground depending on the latitude of the location (Mohanakumar, 2008; Chapter 1). The stratosphere, the second layer of the earth's atmosphere, is stratified in temperature, with warmer layers higher up and cooler layers farther down, see Figure 1.1. The stratosphere is situated between about 10 miles to 31 miles altitude above the surface at moderate latitudes. Ozone found in the troposphere has detrimental health effects while ozone in the stratosphere protects the earth's inhabitant from the sun's ultra violate (UV) rays.

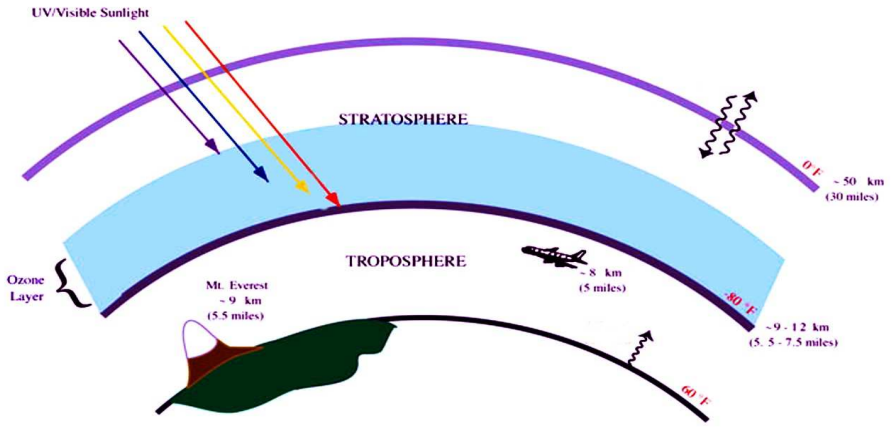


Figure 1.1: Ozone layer in the earth's stratosphere (picture source: NOAA).

1.1.1 Ozone Formation

Chemically ozone is a molecule composed of three atoms of oxygen and is created in the stratosphere when highly energetic solar radiation strikes oxygen molecules (O_2) and causes the two oxygen atoms to split apart. If a freed atom joins into another oxygen molecule then it forms ozone. This process is also known as photolysis (McGarth and Norrish, 1957). However, in the troposphere, ozone is created by chemical reactions between oxides of nitrogen (NO_x) and volatile organic compounds (VOC) in the presence of sunlight. The major source of NO_x and VOC are the emissions from industrial facilities and electric utilities, motor vehicle exhaust, gasoline vapours and the chemical solutions of solids or gases in a liquid[1].

1.1.2 Good Effects of Ozone

Ozone in the stratosphere is good, because in this level ozone helps to protect life on earth by absorbing UV radiation from the sun. UV radiation can cause skin cancer, cataracts (a clouding that develops in the crystalline lens of the human eye) and can harm immune system of human beings. UV can also damage sensitive crops (e.g., soybeans), and destroy some types of marine life (e.g.,*phytoplankton*)[2]. Thus, the increase of ozone in the stratosphere protects earth's life from the UV radiation of the sun, see Figure 1.1.

[1]For more details see: http://www.epa.gov/glo/
[2]http://www.epa.gov/air/ozonepollution/

1.1.3 Harmful Effects of Ozone

Ground level ozone (i.e., in the troposphere) is a pollutant that has direct and indirect bad effects on human health. It is a secondary pollutant and formed by a slow and complicated series of reactions from primary pollutants (see Subsection 1.1.1).

One of the main concerns regarding ground level ozone is its harmful effects on human lung functions, specifically, the damage it causes to the lung tissues, and subsequent respiratory functions. Studies have shown that susceptible persons suffering from bronchitis and asthma are specially likely to be affected most by high levels of ozone (see, WHO, 1979; 1987). Ozone is also responsible in reducing the immune system's ability to fight off bacterial infections in the respiratory system. Also, it is a main ingredient of urban smog (USEPA, 1999a).

Tropospheric ozone also damages vegetation and ecosystem, since high ozone concentration levels can affect the ability of plants to produce and store food. Effects on long-living species such as trees may accumulate over the years, resulting in damage to entire forests and the related ecosystems (Ashmore, 2005; Sitch *et al.*, 2007).

1.1.4 Ozone's Effect on Climate

Ozone occurring throughout the troposphere acts as a greenhouse gas (GHG), which traps heat from the sun and warms the earth's surface. Ozone's impact on climate consists of changes in temperature. Most of the atmospheric warming from tropospheric ozone comes from absorption of infrared energy radiated back towards space from the earth's ground surface. Hence, it may have an effect on global climate change (IPCC, 2007a, Chapter 7). A study evaluating the effects of changing global climate on regional ozone levels in 15 cities in the United States (US) finds, for instance, that average summer time daily maximum ozone concentrations could increase by 2.7 parts per billion (ppb) for a 5-year span in the 2020s and 4.2 ppb for a 5-year span in the 2050s. As a result, more people, specially the young and the elderly, might be forced to restrict outdoor activities (NRDC, 2004) when the ozone levels are high. Again, according to the Intergovernmental Panel on Climate Change (IPCC), tropospheric ozone is the third most important GHG after carbon dioxide (CO_2) and methane (CH_4),

that absorb heat radiation coming from the surface of the earth and trap this heat in the troposphere (IPCC, 2007b), see Figure 1.2.

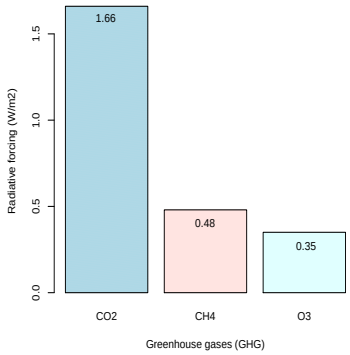


Figure 1.2: Three important green house gases (GHGs) that have effect on climate change.

1.2 Ozone Monitoring

This section describes how we can measure ground level ozone concentrations. Particularly, we provide information based on the ozone measurement approaches taken by the United States Environmental Protection Agency (USEPA). We also define how the hourly and daily ozone levels are obtained from hourly readings.

1.2.1 Ozone Monitoring Stations

In the eastern US, there is a large number of monitoring stations recording hourly ozone levels. Some of these stations are located in urban areas while some others are found in rural areas and the remaining sites are located near pollutant sources such as power stations. The monitoring sites in the urban areas, particularly around big cities, are known as National Air Monitoring Stations/State and Local Air Monitoring Stations (NAMS/SLAMS)[3], whereas the Clean Air Status and Trends Network (CASTNET)[4] sites operate in mostly non-urban areas, see for example, Figure 1.3 for a map of these sites in Ohio. The number of monitoring sites in CASTNET is relatively smaller than that in the NAMS/SLAMS.

[3]http://www.epa.gov/cludygxb/programs/namslam.html
[4]http://www.epa.gov/castnet/

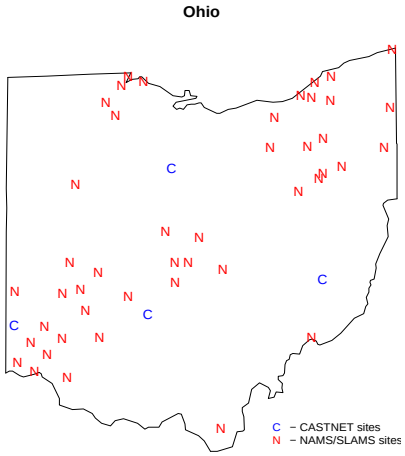


Figure 1.3: A map showing ozone monitoring sites in Ohio.

1.2.2 Hourly and Daily Ozone Concentrations

Initially the observed ozone concentrations are measured hourly. From these hourly readings, the USEPA calculates the daily maximum one hour and eight-hour average ozone concentrations. The daily maximum one hour average ozone concentrations focus on short time exposure at a high level and the eight-hour average provides greater protection against longer time exposure at a moderate level (USEPA, 1996). Our main interest will be the daily maximum eight-hour average ozone levels instead of the hourly ozone concentrations, because current air pollution regulations of the US are based on this. We now define a few key summary statistics of ozone concentration levels that are used in this thesis.

Daily Maximum Eight-Hour Average Ozone Concentrations

The daily maximum eight-hour average ozone level is the maximum of averages of the eight successive hourly ozone concentrations in a day. The procedure for obtaining the daily maximum eight-hour average ozone levels is as follows:

The eight-hour average ozone concentration at the current hour t is the simple average of the eight-hourly concentrations at the current hour t, four past hours ($t-1$, $t-2$, $t-3$, and $t-4$), and the three future hours ($t+1$, $t+2$, and $t+3$). For example, the eight-hour average at 2 P.M. will be the simple average of the

hourly ozone readings from 10 A.M. to 5 P.M.

The maximum of these eight-hour averages for a day is the daily maximum eight-hour ozone levels for that day. Note that the daily maximum for a particular calendar day will depend on the hourly ozone levels for 8-11 P.M. of the previous day and the hourly readings for midnight, 1 A.M. and 2 A.M. of the next day.

Annual 4th Highest Measurements

The annual 4th highest daily maximum eight-hour average is straightforwardly obtained as the 4th highest value of the daily maximum eight-hour averages for that year. Note that this is site specific, i.e., each location will have its annual 4th highest measurement (see details in USEPA, 1998).

Three-year Rolling Average Measurements

The three-year rolling average of the annual 4th highest daily maximum eight-hour average ozone concentration is obtained by averaging the annual 4th highest eight-hour daily maximum concentration levels over three successive years and assigning the average to the final year of averaging.

1.3 Ozone Standards

1.3.1 Air Quality Index (AQI)

A number of air quality indicators have been developed and used to understated and measure the air pollution exposure, for example, Ott (1978), Khanna (2000), Cogliani (2001), Chan and Yao (2008), Dingenen *et al.* (2009), Lee *et al.* (2011). USEPA uses the Air Quality Index (AQI), which is a uniform index for reporting and forecasting daily air quality for the US (USEPA, 1999b). It is used to report the five most common ambient air pollutants that are regulated under the Clean Air Act[5]: ground-level ozone, particulate matter, carbon monoxide, sulfur dioxide, and nitrogen dioxide. The AQI tells how clean or polluted the air is and how to avoid potential health effects.

The AQI uses a normalised scale from 0 to 500. Since levels rarely exceed a value of 200 in the US, in most cases only the range from 0 to 300 is shown.

[5]see http://www.epa.gov/air/caa/

The higher the AQI value, the greater the level of pollution and the greater the effects in health. The AQI is divided into six categories that correspond to different levels of health concern. For ozone, the breakpoints between these categories were selected based on a review of the health effects evidence. This evidence included concentration-response functions derived from a series of controlled human exposure studies (e.g., Folinsbee *et al.*, 1988; Horstmann *et al.*, 1990; McDonnell *et al.*, 1991). Table 1.1 provides the cut points for the ground level ozone concentrations[6].

Index Values	Levels of Health Concern	Cautionary Statements
0-50	Good	None.
51-100	Moderate	Unusually sensitive people should consider reducing prolonged or heavy exertion outdoors.
101-150	Unhealthy for Sensitive Groups	Active children and adults, and people with lung disease, such as asthma, should reduce prolonged or heavy exertion outdoors.
151-200	Unhealthy	Active children and adults, and people with lung disease, such as asthma, should avoid prolonged or heavy exertion outdoors. Everyone else, should reduce prolonged or heavy exertion outdoors.
201-300	Very Unhealthy	Active children and adults, and people with lung disease, such as asthma, should avoid all outdoor exertion. Everyone else, should avoid prolonged or heavy exertion outdoors.
301-500	Hazardous	Everyone should avoid all physical activity outdoors.

Table 1.1: The Air Quality Index guide including the cautionary statements and actions people can take to reduce their risk from exposure to air pollution at different levels of health concern.

1.3.2 Primary Ambient Air Quality Standard for Ozone

The Clean Air Act established two types of national air quality standards for ground level ozone: (i) Primary standards and (ii) Secondary standards. The primary standards set limits to protect public health, including the health of sensitive populations such as asthmatics, children, and the elderly.

The primary standards promulgated in 1997 was set at 80 ppb for averaged

[6]reference: http://www.airnow.gov/

over an eight-hour period. Allowing for rounding, USEPA considered areas with readings as high as 85 ppb to have attained the standard. The review completed in 2008 found evidence of health effects, at levels of exposure below the 80 ppb standard. As a result, both USEPA and the Clean Air Scientific Advisory Committee (CASAC) recommended strengthening the standard to 75 ppb in 2008.

The primary ozone standard is met if the three year average of the annual 4th highest daily maximum eight-hour average is less than 75 parts per billion (ppb). In this thesis the value 85 ppb will be used since we only analyse data until 2006 (see Chapter 4). A site is designated as a non-attainment site if the primary standard is not met at that site.

Note that the primary standard is site specific. However, ozone concentrations are only monitored in few fixed monitoring sites in each state, see e.g., Figure 1.3. That is why it is important to spatially model and predict ozone concentration levels. Section 1.5 provides a review of the modelling strategies for ozone levels.

1.4 Ozone Forecasting using Computer Models

1.4.1 Computer Simulation Models

To forecast ground level ozone concentration in a very fine spatial resolution, the National Oceanic and Atmospheric Administration (NOAA) in the US designed the Community Multi-scale Air Quality (CMAQ) modelling system[7]. The National Centers for Environmental Prediction (NCEP)[8] developed an Eta model (Black, 1994; Rogers *et al.*, 1996), and later, Otte *et al.* (2005) described the linkage between the Eta and the CMAQ model and proposed the Eta-CMAQ model. In the Eta-CMAQ, the Eta modelling approach is used to prepare the meteorological fields for input to the CMAQ system. The NCEP product generator software is used to perform bilinear interpolations and nearest-neighbour mapping of the Eta Post-processor output from Eta forecasting domain to the CMAQ forecast domain. The processing of the emission data for various pollutant sources has been adapted from the Sparse Matrix Operator Kernel Emissions (SMOKE) modelling system (Houyoux *et al.*, 2000) on the basis of the USEPA national emission inventory.

[7] http://www.epa.gov/amad/CMAQ/index.html
[8] http://www.ncep.noaa.gov/

1.4.2 CMAQ System

The CMAQ model is a deterministic differential equations model which takes several inputs based on meteorology, transportation dynamics, emission 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. 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: advection and diffusion, gas phase chemistry, plume-in-grid modelling, particle modelling and visibility, cloud processes, and photolysis rates. The CMAQ modelling systems also contain the following processors and interfaces:

(i) Meteorology-Chemistry Interface Processor (MCIP) interpolates the meteorological data needed and computes the cloud, surface and planetary boundary parameters.

(ii) Emission-Chemistry Interface Processor (ECIP) generates hourly emission data for the CMAQ.

(iii) Initial Conditions (ICON) provide concentration fields for chemicals for the initial simulation state.

(iv) For the grids surrounding the modelling domains Boundary Conditions (BCON) provides concentration fields for chemicals.

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

Process analysis and aggregation is the final stage of the CMAQ modelling system. Here, the process analysis detects errors and uncertainties in a model through data analysis and many parametrisation schemes. For further details on CMAQ systems, see Ching and Byun (1999).

The CMAQ model provided by the NOAA uses emission inventories, meteorological information and land use to estimate average pollution levels of ozone concentrations for the gridded cells over successive time periods[9]. The Eta-CMAQ model produces forecasts of ozone concentration levels up to 48 hours in advance and these forecasts do not use any observed data during this forecast period of 48 hours. The forecast values of ozone levels are obtained over 12 kilometre grids. There are 259×268 grid cells that cover much of the continental US.

The CMAQ models provide spatio-temporal coverage in a large area, however it has limitations. For example, as this is a computer simulation model it is not an exact representation of real situations. Moreover, the model output is biased due to errors in emission inventories.

1.4.3 Data Assimilation

The CMAQ model is not a statistical model but a deterministic differential equation model. Often, probability forecasts are more informative than the deterministic estimates, and the probabilistic forecasts can be produced by combining observations and computer simulation models. This type of approach is called *data assimilation* (DA). Moreover, the CMAQ forecasts are obtained for a grid cell while the observed daily ozone data are obtained at a particular location referenced by a longitude-latitude pair. This leads to the spatial misalignment problem between the CMAQ output and the observed ozone data. This problem is well known in literature (see for example, Lorence, 1986; Jun and Stein, 2004; Fuentes and Raftery, 2005). Recently, several new modelling techniques such as the downscaler models have been suggested. See for example, Berrocal *et al.*, 2010a, 2010b; Zidek *et al.*, 2011 and references therein.

In Section 1.4.2 we explained that output of the CMAQ models do not use any observed ozone data. Henceforth, we can assume that observed data are independent of the CMAQ output. In this thesis we use the daily CMAQ output as a covariate (Sahu *et al.*, 2009) to model observed ozone concentration levels, since it is reasonable to expect that these two will be very similar (see Figure 1.4). It can be observed that CMAQ output sometimes capture the actual measurement process very well, however, in other times it fails due to its deterministic

[9] http://www.epa.gov/asmdnerl/CMAQ/

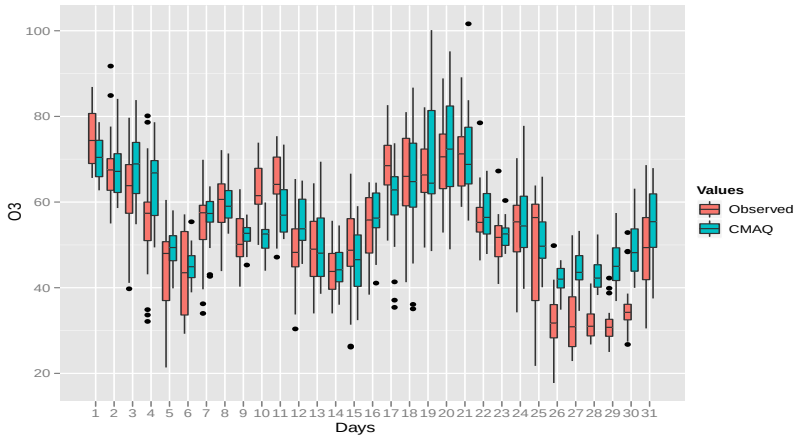


Figure 1.4: Time series plot of the observed daily ozone concentrations and CMAQ output for the grid cell that includes the data site in the state of Alabama for the month of July, 2006.

approach.

1.5 Review of Modelling Strategies for Ozone Concentrations

Recently, there has been a surge of interest in modelling ozone concentration levels. A growing diversity of literature on statistical methodologies are available in this context. We review some key modelling approaches that have been used to model ozone concentrations. Following the review, we provide a brief discussion on the need of scale transformation for modelling ozone levels.

1.5.1 Regression Based Approaches

Most of the approaches for modelling ozone levels are based on linear and non-linear regression methodology. There is a huge literature discussing regression approaches (e.g., Cassmassi and Bassett, 1991; Feister and Balzer, 1991; Fiore *et al.*, 1998; Fioletov *et al.*, 2002) where the ozone levels are modelled as functions of key meteorological variables, for example, temperature, wind speed, humidity, air pressure and cloud cover. Feister and Balzer (1991) used 313 meteorological parameters from three main sources: (a) geopotential at 18 grid points in Central Europe; (b) synoptic meteorological data; and (c) aerological data from a few

European stations. They provide a long-term trend of changes of surface ozone from 1972-1987. They conclude that cloudiness are probably not the main cause of the long-term changes in surface ozone, but changes in ozone trend are effected by the circulation and concentration of ozone precursors (e.g., NO_x, VOC). Fiore *et al.* (1998) also support the result of Feister and Balzer (1991) for the long-term trends in median and 90th percentile ozone concentrations at 549 sites across the US for the 1980-1995 period.

There are also papers discussing time series modelling of ozone levels, where the residuals are considered to have autocorrelation. For example, Galbally *et al.* (1986) develop a linear regression model to analyse the daily maximum one hour average ozone levels. Due to the high autocorrelation both in the raw data and the residuals this method considers the lag one autocorrelation of the errors. Temperature, wind speed and some other meteorological variables are also used, however they do not estimate any trend in their analysis.

Korsog and Wolff (1991) provide a robust regression methodology to analyse the urban daily maximum one hour ozone levels of eight major population centres in the north-eastern US. Their study examined the trends in ozone levels from 1973 to 1983. They found that the 75th percentile ozone concentrations are a good statistic for determining trends. In their analysis the surface temperature and upper air temperature variables were found to be the best predictors of ozone levels.

Bloomfield *et al.* (1996) discuss that statistical linear models have difficulty to capture the complex relationships between the meteorological variables and ozone. To overcome this difficulty they develop a parametric non-linear model. They used twelve meteorological variables as covariates in the model and estimated the trend in ozone levels from 1981 to 1991 in Chicago.

Huang and Smith (1999) extended the non-linear approach to classification and regression trees (Breiman *et al.*, 1984), where the meteorological influence is treated non-linearly through a regression tree. A particular advantage of this approach is that it allows to estimate different trends within the clusters produced by the regression tree analysis. They use ozone concentration data from Chicago to analyse their model and provide ozone-trend for 10 years (1981-1991).

Cox and Chu (1993) formulated a predictive model to analyse daily ozone concentrations using generalised linear models (GLM), assuming a conditional

Weibull distribution for the ozone concentrations given meteorology. They applied their modelling strategy to the annual distribution of ground-level ozone in 43 urban areas throughout the US. Their model includes a trend component that adjusts the annual rate of change in ozone for concurrent impacts of meteorological conditions, e.g., surface temperature and wind speed. Their results suggest that meteorologically adjusted upper percentiles of the distribution of daily maximum one hour ozone levels are decreasing in most urban areas over the period from 1981 to 1991. They also show that without meteorological components the assessed trends underestimate the rate of reduction in ozone.

Following the lines of Cox and Chu (1993) and Huang and Smith (1999), Cocchi *et al.* (2005) model the daily ozone levels under the Bayesian paradigm. They analysed the series of daily maxima of ozone concentrations over the metropolitan area of Bologna, in North of Italy for the period 1994 to 2002. Their analysis highlights the need for standardising the meteorological variables when assessing long-term trend in ozone concentrations. They also found that the trend obtained from the standardised meteorological variables behave differently compared to the yearly median of ozone observations.

Davis *et al.* (1998) analysed ozone concentrations using singular value decomposition and clustering to select the meteorological variables and used generalised additive models (GAM) to develop functional relationship between ozone and meteorological variables. They used one hour average ozone concentration levels from several sites in Houston, Texas, and did not estimate any trend in ozone levels. GAM are also used by Davis and Speckman (1999) to make next-day predictions of ozone levels in the Houston area and used the daily maximum eight-hour average ozone concentration data for analysis.

Camalier *et al.* (2007) also used GAM for modelling the daily maximum eight-hour average ozone concentrations in 39 of the 53 metropolitan areas that have been used in USEPA report (USEPA, 2004) for the period 1998-2004. Their approach also describes the statistical methodology for meteorologically adjusted ozone trends and characterises the relationship between meteorological variables and ozone. They use separate models for each urban area and do not consider the spatial correlation.

The approach of Dynamic linear models (DLM) described in West and Harrison (1997) are used by Zheng *et al.* (2007) to analyse ozone concentrations.

However, they do not consider any spatial correlations between the observations in different sites in the models. Here they compare this approach with the GAM to estimate trends in ozone concentration levels in the eastern US for the period 1997-2004. They also compared the results from both models for four monitoring locations chosen through principal components analysis (PCA) to represent regional patterns in ozone concentrations. After adjusting for the meteorological influence by the PCA, they found that the overall ozone trend showed a downward pattern for all four locations.

Quantile regression approach is used by Sousa *et al.* (2009), where they analysed the influence of the meteorological variables (e.g., temperature, solar radiation, wind direction and relative humidity) on hourly ozone concentration levels. In their study, hourly ozone data is used for the months June, July and August in 2003, that are obtained from urban location in Oporto, Northern Portugal. They forecast next day hourly ozone levels but did not obtain trends. They concluded that the quantile regression approach is useful to evidence the heterogeneity of the influence of the meteorological variables on different ozone levels.

1.5.2 Spatio-temporal Approaches

Spatial and spatio-temporal modelling are also popular for analysing ozone concentration levels. Guttorp *et al.* (1994) examine hourly ozone concentration data obtained from 17 sites concentrated around the Sacramento area of the San Joaquin Valley of California. They apply a spatio-temporal analysis, which indicated a relatively simple spatial covariance structure at night-time, and a more complex one during the afternoon. A simple separable space-time covariance model is used to analyse these data.

Carroll *et al.* (1997) develop another spatio-temporal model with an exponential space-time covariance function and applied it to hourly ozone concentration levels obtained from twelve monitoring sites in Harris County, Texas. The model they proposed for the ozone prediction consists of decomposing the ozone data into a trend part and an irregular part. Along with building the model, they develop a fast model-fitting method that can cope with the massive amounts of available data and the substantial number of missing observations.

Huerta *et al.* (2004) use the spatio-temporal version of the DLM (developed

by Stroud *et al.*, 2001), and apply it to hourly ozone concentration data obtained from 19 monitoring sites in Mexico city. This DLM is a state-space model and incorporates spatial covariance structure for the ozone levels and model parameters. They use seasonal variation and temperature as covariate effects in their model and analyse using Bayesian methods. Their methods provide short-term forecasts and spatial interpolations for the ozone concentration levels, however they do not estimate trends in ozone levels.

McMillan *et al.* (2005) proposed a hierarchical Bayesian model that describes the spatio–temporal behaviour of daily ozone levels within a domain covering Lake Michigan. Their model incorporates linkages between ozone and meteorology and estimates ozone levels over the entire modelling domain based upon unevenly distributed monitoring data. They provide prediction on spatial fields of ozone concentrations considering effects of the meteorological variables, such as temperature, humidity, pressure, and wind speed and direction. Trend analysis for ozone levels are not discussed in their study.

Sahu *et al.* (2007) propose another spatio-temporal method to analyse daily ozone levels based on autoregressive (AR) modelling. They use daily maximum eight-hour average ozone levels obtained from 53 monitoring sites in Ohio. Their model incorporates meteorological variables: maximum temperature, average relative humidity and wind speed in the morning and in the afternoon, observed at a collection of ozone monitoring sites as well as at several weather stations where ozone levels have not been observed. They handle this misalignment through spatial modelling. Their model is hierarchical in nature and specified within a Bayesian framework. They analyse 8 years of data from 1997-2004 and provide predictions at the validation sites. Long-term trends in ozone concentration levels are also analysed. In addition, they provide annual summaries of the ozone levels.

Using similar types of models, Sahu *et al.* (2009) provide next day forecast of the daily maximum eight-hour ozone concentration levels in the eastern US using 390 monitoring locations. They use forecast data obtained from a computer simulated model (see Section 1.4) as a predictor for the observed ozone levels in the eastern US.

Dou *et al.* (2010) compare the Bayesian spatial predictor (BSP) method (Le and Zidek, 1992; 2006) with the DLM for analysing hourly ozone concentration

data from several sites in Illinois, Missouri and Kentucky. The BSP has been proposed as an alternative to kriging (see Section 2.5 for kriging). Dou *et al.* (2010) provide spatial interpolations at the validation sites for both methods and have concluded that the BSP performs as well as the DLM and in some cases of missing observations, the BSP performs better for prediction. Following this Dou *et al.* (2011) also provide temporal forecast of hourly ozone concentrations using the BSP.

Another method, Bayesian melding (Fuentes and Raftery, 2005), is used by Liu *et al.* (2011) for predicting ozone concentrations in the unmonitored locations. They used data from the deterministic Air Quality Model (AQM) and the MAQSIP (Multi-scale Air Quality Simulation Platform) model. The melding methodology is applied and compared with kriging to predict and map spatial fields. However, they do not obtain any trends in ozone levels.

Bayesian spatial quantile regression modelling is proposed by Reich *et al.* (2011) to analyse daily maximum eight-hour average ozone concentrations in the eastern US. Different meteorological variables (e.g., average temperature, maximum wind speed and average cloud cover) are used as covariate on modelling ozone levels. They conclude that meteorological variables are strongly associated with ozone levels and the effects are stronger in the right tail than the centre of the distribution.

1.5.3 Scale Transformation of Ozone Concentrations

There are many modelling approaches where the original scale of ozone levels are used (e.g., Feister and Balzer, 1991; Fiore *et al.*, 1998; Davis *et al.*, 1998). However, in original scale ozone concentrations are unstable and as a result different variance stabilising transformations have been proposed in the literature. For example, the logarithmic transformation has been applied by Bloomfield *et al.* (1996) and Korsog and Wolff (1991). However, the log scale introduces negative skewness (Sahu *et al.*, 2007). A popular approach is the square root transformation, see e.g., Galbally *et al.* (1986), Cox and Chu (1993), Sahu *et al.* (2007). The square root transformation is adopted in this thesis because it encourages symmetry and stabilises the variance of the data (Carroll *et al.*, 1997; Sahu *et al.*, 2007; 2009).

1.6 Literature Review for the *big-n Problem*

Statistical modelling is often infeasible for ozone concentration data obtained from a large number of monitoring sites, because of the limitations in existing computational ability. This problem is also known as *big-n problem* in literature (see e.g., Banerjee *et al.*, 2004, page-387; Xia and Gelfand, 2006; Shekhar and Xiong, 2008).

In the *big-n problem*, exact likelihood based inference becomes unstable and infeasible since it involves computing quadratic forms and determinants associated with a large variance-covariance matrix (Stein, 2008). The large n dimensional variance-covariance matrix decomposition involves $O(n^3)$ computational complexity in time and $O(n^2)$ in storage that increases with the increase of spatial locations n (Cressie and Johannesson 2008). This problem, also arises in evaluation of the joint or conditional distributions in Gaussian processes models under hierarchical Bayesian setup (Banerjee *et al.* 2004), particularly, for example, in iterative algorithms.

The early approaches to solve the *big-n problem* are based on *ad-hoc*, for example, local kriging (Cressie, 1993), and sub-sampling from a large spatial dimension by a moving window approach see for example, Hass (1995); Pardo-Iguzquiza and Dowd (1997). However, these easy methods ignore a moderate amount of observed data in analysis.

The gradual methodological improvement leads to different approximation techniques of kriging equations, for example, low rank kriging (Nychka *et al.* 1996) where the reduced spatial points are obtained using space filling algorithm (Johnson *et al.* 1990). Kammann and Wand (2003) use this approach and account for non-linear covariate effects by the geoadditive models. Spectral domain approach is also used to reduce the dimension where likelihood approximation of the kriging equation is used (Stein, 1999; Paciorek, 2007), and has the limitations for analysing multivariate processes with non-stationary covariance functions. Xia and Gelfand (2006) use the moving average technique to approximate spatial random process as a linear combination of smaller random variables, thus reducing the large spatial dimension. However, their method is also only applicable for stationary spatial processes. Spatial prediction based on low rank smoothing splines is also used for massive spatial data sets (Hastie 1996,

Johannesson and Cressie 2004). Another technique, the Gaussian Markov random fields (Rue and Held, 2006) approximation with sparse matrix algorithms, is used to solve the problem for large spatial datasets (Hartman and Hossjer, 2008). However, their method is more suitable for areal data rather than point referenced spatial data sets. To overcome this problem, recently Lindgren and Rue (2011) used an explicit link between the Gaussian Markov random fields and Gaussian fields using a stochastic partial differential equation approach to tackle the gap. Reich *et al.* (2011) used a spatial quantile approach considering non-Gaussian processes in the models. They use quantile parameters as the reduced dimension of the data. However, their approach is sensitive to the choice of the number of quantiles.

Furthermore, multi-resolution spatial models (Huang, *et al.*, 2002, Johannesson and Cressie, 2004, Johannesson, *et al.*, 2007) can capture the non-stationarity of the data and provide fast optimal estimates. However, these methods cannot capture the heterogeneity across large spatial regions. Approaches like low rank and moderate rank matrix (Stein 2007, 2008) are also used to reduce the dimension of the data.

Fixed rank kriging (Cressie and Johannesson, 2008), can also handle the modelling of massive spatial data. These approaches capture the non-stationarity and heterogeneity in the data, but choice of smoothing parameters (e.g., basis functions) and knots sometimes may increase the complexity of the models. Moreover, their approach is not based on the likelihood function, because of the difficulties in maximisation with large number of parameters.

To avoid approaches related to the basis functions and to utilise the likelihood based approach, Banerjee *et al.* (2008) proposed the Gaussian predictive processes, that can analyse the heterogeneity of the massive spatial data and can easily handle the problem of the smoothness parameter with a solution of the *big-N problem*. However, the non-spatial error term of this approach induces positive bias and later Finley *et al.* (2009) proposed a modified process to address this problem. In addition, Guhaniyogi *et al.* (2011) introduced an adaptive technique for choosing knot sizes, where stochastic modelling of knots is used instead of fixing them in the predictive process models.

In this thesis, we adopt the Gaussian predictive processes methodology and propose a spatio-temporal model that can analyse ozone concentration levels ob-

tained from the vast region of the eastern Unites States (US), details are provided in later Chapter 6.

1.7 Thesis Organisation

The remainder of this thesis is organised as follows:

Chapter 2 reviews the statistical techniques that are used in spatial analysis. We also discuss the fundamental geo-statistical methodologies to analyse spatial and spatio-temporal data, particularly the approaches we have used in this thesis.

A review of the Bayesian paradigm is discussed in Chapter 3. Different Bayesian modelling strategies, criteria of model choice are explained in this chapter. We also discuss the Spatio-temporal Bayesian Gaussian processes models that have been adopted in this thesis to analyse ground level ozone concentrations.

Chapter 4 gives a description of available data that we use in this thesis. We describe data preparation, editing and cleaning which is necessary for the raw ozone data we obtained from the United States environmental protection agency (USEPA) for the whole eastern US. The computer simulation model output values are also discussed in this chapter. Different types of meteorological data are provided by the National Climatic Data Center (NCDC)[10], however in this thesis we only consider the variables that have significant effects on ozone. We also present summary statistics for all data sets after getting into analysable form.

In Chapter 5, we experiment with two different modelling strategies: the DLM (Huerta *et al.*, 2004) and the AR models (Sahu *et al.*, 2007). Theoretical properties of the models are discussed and compared to find out similarities and dissimilarities. To compare the model performances we provide simulation examples together with a real life application on daily maximum ozone concentrations observed in several sites in the state of New York for the months of July and August, 2006. We use CMAQ output as a covariate in the models. We conclude that the AR models perform better compared to the DLM both in theories and in practical example.

To analyse and model large dimensional data obtained from the whole eastern US, we propose a new spatio-temporal modelling strategy in Chapter 6. The

[10]http://www.ncdc.noaa.gov/oa/ncdc.html

proposed model is based on the Gaussian predictive processes (GPP) approximation and also considers the temporal dependency through the random effects. The methodology of tackling spatial misalignment between ozone monitoring and meteorological monitoring locations are also discussed in this chapter. Initially we use a smaller part of the full data set consisting of four states for comparing the proposed modelling approach with the AR models used in Chapter 5. We find better predictive performance of the GPP based models over the AR models. Here, we also obtain long term meteorology adjusted and unadjusted trends in ozone levels from 1997 to 2006 and discuss on the non-attainments of the primary ozone standards.

Chapter 7 describes the forecasting methodology of the models discussed in the earlier chapters. Three weeks of data from the eastern US are used in this chapter for analysis. Similar to Chapter 6, we compare the GPP based models with the GP, DLM, and AR models using a smaller set of data. We conclude that the model based on GPP approximation is the best among other modelling strategies. Finally we obtain one day ahead forecast for 7 days at the CMAQ grid locations spread around the eastern US.

In Chapter 8 we discuss the software package `spTimer` that have developed as a part of this thesis. This package is built using low-level language `C` that is hidden from the user and is designed for the open-source popular statistical software `R`. Currently, this package can fit, predict and forecast data using three types of Gaussian process spatio-temporal models. To validate the code we use several simulation studies and re-estimate the true model parameters. We also provide validation results for the simulated data sets.

Some concluding remarks are presented in Chapter 9. Here we discuss the summary of the thesis and introduce some idea that can be extended for future work.

1.8 Summary

The primary aim of this study is to model and analyse the daily maximum eight-hour average ground level ozone concentrations obtained from a large number of sites in the eastern US. Our interest is to predict and forecast ozone levels at unmonitored locations and also at future time points along with their associated

uncertainties through rigorous statistical models. In this chapter we provide a brief description of ozone concentration levels and its effects on human health, plant and vegetations. We also discuss a review of the available spatial and non-spatial modelling approaches to analyse the ozone concentration levels. Review of analysing large spatial data sets is also discussed to solve the well known *big-n problem*.

Chapter 2

Review of Geostatistics

2.1 Introduction

The statistical methods in analysing spatial data date back to Matheron (1963), where he proposed the term *geostatistics* and developed a range of estimation techniques in mining using spatial statistical approaches. Geostatistical analysis is important for modelling and understanding the spatial variability of a quantity that may vary in space, for example, ozone concentration levels, rainfall, and soil structure.

Spatio-temporal modelling is more recent than the spatial analysis methods and a large literature has evolved in the last two decades (see for example, Banerjee *et al.*, 2004; Le and Zidek, 2006; Finkenstadt *et al.*, 2007; Gelfand *et al.*, 2010 and references therein). Throughout the thesis, we will use the basic approaches and assumptions of the spatio-temporal models stated in this chapter.

This chapter reviews the geostatistical methods together with their properties and assumptions. The plan of this chapter is as follows: In Section 2.2 we describe different types of spatial data. Section 2.3 discusses the spatial and spatio-temporal processes. In Section 2.4 we provide the basic characteristics of the spatio-temporal covariance functions. We discuss different types of spatial interpolation techniques, also known as kriging in Section 2.5. In Section 2.6, we provide some brief description on cartography and geodetic distances. Finally, Section 2.7 ends this chapter with summary remarks.

2.2 Types of Spatial Data

Spatially dependent data are often classified into three major types see e.g., Banerjee *et al.* (2004, Chapter 1). These are: (i) point-referenced data (ii) point pattern data, and (iii) areal data. Below we discuss these three types of data.

2.2.1 Point-referenced Data

In *point-referenced data* (also known as *geostatistical data*) the random observation $Z(\mathbf{s})$ is measured at a location $\mathbf{s} \in \mathcal{S} \subset \Re^d$, and $\mathbf{s}$ varies continuously over the study region $\mathcal{S}$. Theoretically the number of locations in $\mathcal{S}$ is infinite. For example see Figure 1.3 where the ozone concentration levels are monitored in several sites in the state of Ohio.

2.2.2 Point Pattern Data

The second type of spatial data is known as *point pattern data*, where the study domain $\mathcal{S}$ is random and its index set gives the locations of random events that describe the observed spatial point patterns. An example of point pattern data is given in Figure 2.1, where the points represent locations of 3605 trees of the species *Beilschmiedia pendula* (Lauraceae) in a 1000 by 500 meter rectangular sampling region in the tropical rain forest of Barro Colorado Island (Condit, 1998).

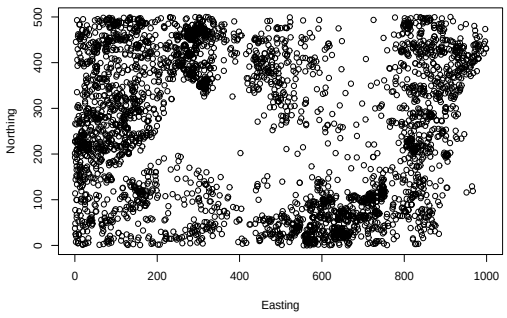


Figure 2.1: Example of point pattern data showing locations of trees in the rain forest of Barro Colorado Island.

2.2.3 Areal Data

The third and final type of spatial data is known as *areal data*, where the study domain $\mathcal{S}$ is a fixed subset with regular or irregular shape, but partitioned into a finite number of areal units with well-defined boundaries. For example, Figure 2.2 shows the average 4th highest ozone concentration levels for the 33 states in the eastern US in 1997.

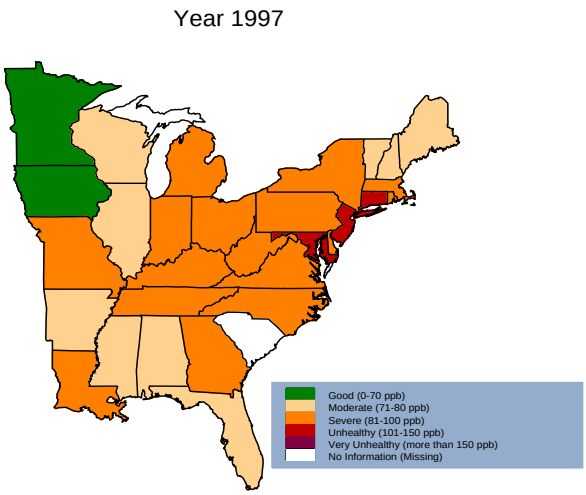


Figure 2.2: A choropleth map of the statewise average 4th highest ozone concentration levels in 1997.

Henceforth, we only describe modelling strategies for analysing point-referenced data since the main objective of this thesis is to model daily ozone concentration levels observed in many fixed monitoring stations in the eastern US study region.

2.3 Spatial and Spatio-temporal Processes

Let $\mathbf{s}$ be any spatial location within the study region $\mathcal{S}$. We write the spatial random process as:

$$Z(\mathbf{s}) : \mathbf{s} \in \mathcal{S} \subset \Re^d$$

where, $Z(\mathbf{s})$ is the measurement of the attribute of interest at location $\mathbf{s}$. Notationally, for n different locations, the measurements can be written as, $\mathbf{Z}(\mathbf{s}) = (Z(\mathbf{s}_1), ..., Z(\mathbf{s}_n))'$. To express the spatio-temporal process we need to include

temporal identity on the spatial process.

The spatial process $Z(\mathbf{s})$ can be extended to a spatio-temporal process with the help of an additional index, t for time. Thus we can use the notation:

$$Z(\mathbf{s},t) : \mathbf{s} \in \mathcal{S} \subset \Re^d, t \in \Re$$

to denote the spatio-temporal process of interest. When we have observed the $Z(\mathbf{s},t)$ process at n spatial locations $\mathbf{s}_1, ..., \mathbf{s}_n$ at t different time points, we may write the spatio-temporal process $Z(\mathbf{s},t) = (Z(\mathbf{s}_1,t), ..., Z(\mathbf{s}_n,t))'$, $1 \leq t \leq T$.

However, from a mathematical perspective, we can also represent the spatio-temporal process as a multivariate spatial process with dimension $d+1$, see e.g., Le and Zidek, (2006). They argue that every time point of the spatio-temporal process can be regarded as a separate spatial random field.

In this study, we will consider models for $Z(\mathbf{s},t)$ to be fully parametrised by the set of p (say) parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)'$. Our aim is to model the spatio-temporal dependence present in the observations, with a view to making a prediction $Z(s_0,t)$ at a new position s_0, or to provide a forecast at a future time point.

2.4 Characteristics of Space-time Covariance Functions

2.4.1 Stationarity

Before going further on model discussions, we define the terms *stationarity* and *isotropy*. The idea of stationarity comes from the general theory of stochastic processes. Consider two spatial locations, $\mathbf{s}$ and $\mathbf{s}+\mathbf{h}$, where $\mathbf{h} \in \Re^d$. A spatial process is called *strictly stationary* if, for any given $n \geq 1$, any set of n sites $\{\mathbf{s}_1, ..., \mathbf{s}_n\}$ and for any $\mathbf{h} \in \Re^d$, the joint distributions of $Z(\mathbf{s})$ and $Z(\mathbf{s}+\mathbf{h})$ are same, i.e.,

$$\pi(Z(\mathbf{s}_1), ..., Z(\mathbf{s}_n)) = \pi(Z(\mathbf{s}_1+\mathbf{h}), ..., Z(\mathbf{s}_n+\mathbf{h})).$$

Assume that the process has a valid covariance function $\text{Cov}(Z(\mathbf{s}), Z(\mathbf{s}+\mathbf{h}))$. The process $Z(\mathbf{s})$ is known as *second-order stationary* (also known as *weak sta-*

tionary), if

$$\text{Cov}(Z(\mathbf{s}), Z(\mathbf{s}+\mathbf{h})) = C(\mathbf{h}), \quad \forall \quad \mathbf{s} \in \mathcal{S}, \quad \mathbf{h} \in \Re^d,$$

where, $C(\mathbf{h})$ is a function that depends on the difference in the spatial locations, $\mathbf{h}$. For *non-stationary* spatial process either or both the above type of stationarity do not hold.

The function $C(\mathbf{h})$, that we have defined earlier is known as *covariogram*. The intrinsic stationary defines only the first and second moments of the differences $Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})$ but not anything about their distributions.

Similarly, spatio-temporal process say, $Z(\mathbf{s}, t)$ is considered to be *mean stationary* within its spatio-temporal domain $\mathcal{S} \times \mathcal{T}$, if its mean process is constant within spatio-temporal domain (Bruno *et al.*, 2009). For *weak stationary* of the spatio-temporal process $Z(\mathbf{s}, t)$ the mean function is assumed to be constant and the covariance function is assumed to depend on spatial and temporal covariances. We can observe that mean stationary only implies weak stationary if the first two moments i.e., mean and variance exist, whereas weak stationary only implies strict stationary if the spatio-temporal random process $Z(\mathbf{s}, t)$ is a Gaussian process, details of Gaussian process is given in Section 3.4.

2.4.2 Isotropy

A spatial process $Z(\mathbf{s})$ is termed as *isotropic* if its covariance function $C(\mathbf{h})$ depends only on the distance $|\mathbf{h}|$ between the two locations $\mathbf{s}$ and $\mathbf{s}+\mathbf{h}$. A process which is not isotropic is called *anisotropic*. Covariance functions of anisotropic processes exhibit different behaviour in different directions. Isotropic processes are popular because of their simplicity, and easy interpretability.

There are common parametric isotropic models available in spatial analysis. These models are in simple parametric form and are available as candidates for the semivariogram $\gamma(\mathbf{h})$. Figure 2.3 shows some of the covariance functions based on different isotropic models.

2.4.3 Separable and Nonseparable Covariance Functions

The separability of models refers to formation of the spatio-temporal covariance function as a product of the spatial and temporal covariance functions (see for

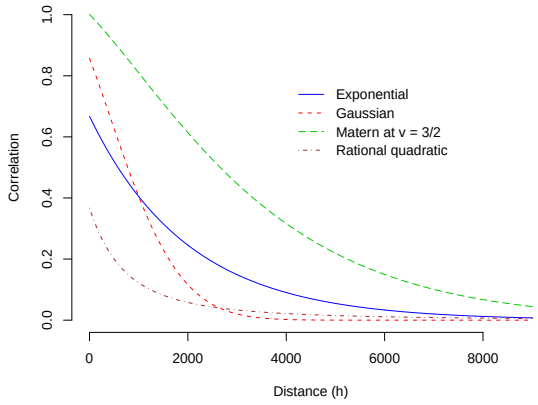


Figure 2.3: Some illustrations of covariance functions based on parametric isotropic models.

example Rouhani and Mayers 1990, Le and Zidek 2006, Diggle and Ribeiro 2007). A *separable* spatio-temporal covariance function is defined as:

$$C(Z(\mathbf{s},t);Z(\mathbf{s}',t')) = C_S(\mathbf{s},\mathbf{s}')C_T(t,t')$$

where, $\mathbf{s}$ and $\mathbf{s}'$ are the spatial locations and t and t' are the temporal points, and the terms $C_S(\mathbf{s},\mathbf{s}')$ and $C_T(t,t')$ represent the spatial and temporal covariance functions respectively. A space-time covariance function is called *nonseparable* if it cannot be represented as the product of spatial and temporal functions. For a separable process, the space time covariance function can be modelled separately.

The main advantage of assuming separability is the computational convenience, since the spatio-temporal covariance matrix can be written as the Kronecker product of two smaller dimensional matrices. However, there are many nonseparable models available. Cressie and Huang (1999) introduced several classes of nonseparable stationary covariance functions to model spatio-temporal data. They used Fourier transforms in their approach, and used the Bochner's theorem (1955) to guarantee positive definiteness for the covariance function. There are other approaches for constructing non-stationary covariance function see, for example Gneiting (2002), Stein (2005), Bruno *et al.* (2009) and the references therein.

2.4.4 Some Parametric Covariance Functions

In this section we discuss some parametric covariance functions (see Figure 2.3). We can write the covariance function $C(\mathbf{h})$ as:

$$C(\mathbf{h}) = \sigma^2 \kappa(\mathbf{s}_i, \mathbf{s}_j; \Phi)$$

where, σ^2 is the common variance term and $\kappa(\mathbf{s}_i, \mathbf{s}_j; \Phi)$ is the spatial correlation between locations $\mathbf{s}_i$ and $\mathbf{s}_j$ with smoothness and decay parameters Φ. In this thesis we frequently use the spatial exponential correlation function defined as:

$$\kappa(\mathbf{s}_i, \mathbf{s}_j; \phi) = \exp(-\phi||\mathbf{s}_i - \mathbf{s}_j||), \quad \phi > 0,$$

where, $||\mathbf{s}_i - \mathbf{s}_j||$ is the distance between sites $\mathbf{s}_i$ and $\mathbf{s}_j$, and ϕ is the spatial decay parameter. We also use Gaussian and spherical correlation function. The Gaussian correlation function is defined as:

$$\kappa(\mathbf{s}_i, \mathbf{s}_j; \phi) = \exp(-\phi||\mathbf{s}_i - \mathbf{s}_j||^2), \quad \phi > 0,$$

and we can define the spherical correlation function as:

$$\kappa(\mathbf{s}_i, \mathbf{s}_j; \phi) = 1 - \frac{3}{2}\phi||\mathbf{s}_i - \mathbf{s}_j|| + \frac{1}{2}(\phi||\mathbf{s}_i - \mathbf{s}_j||)^3), \quad 0 < ||\mathbf{s}_i - \mathbf{s}_j|| < 1/\phi,$$

The Matérn correlation function (Matérn 1986) that we used in this thesis is defined as:

$$\kappa(\mathbf{s}_i, \mathbf{s}_j; \phi, \nu) = \frac{1}{2^{\nu-1}\Gamma(\nu)}(2\sqrt{\nu}||\mathbf{s}_i - \mathbf{s}_j||\phi)^\nu K_\nu(2\sqrt{\nu}||\mathbf{s}_i - \mathbf{s}_j||\phi), \quad \phi > 0, \nu > 0,$$

where, K_ν is the modified Bessel function of the second kind with order ν. The special cases of this correlation function are also available in exponential and Gaussian form by replacing $\nu = 1/2$ and $\nu \to \infty$ respectively. For convenience, we also use $\nu = 2/3$ to obtain a close form of the Matérn correlation function, for more details see Cressie (1993) and Banerjee *et al.* (2004).

2.5 Kriging

In this section we discuss the classical approaches to spatial interpolation. In 1951, D.G. Krige, a South African mining engineer developed a method that was able to perform a spatial prediction for small amount of data. Later, Matheron (1963) formalised that method and termed it Kriging. Several enhancements of this method have been developed to deal with particular applications (see for example, Cressie, 1993 Chapter 3 and Stein, 1999). Some of the enhancements with mathematical settings are described below.

2.5.1 Simple Kriging

Let the stochastic response $\eta(\mathbf{s})$ (point-referenced data) at site $\mathbf{s}$ be strictly stationary, so that it is written as:

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \eta(\mathbf{s}) \tag{2.1}$$

where, $\mu(\mathbf{s})$ is a known function and $\eta(\mathbf{s})$ is the spatial error process and assumed it to be Gaussian with mean zero and covariance matrix $\mathbf{\Sigma}$ (say). Let $\mathbf{Z}(\mathbf{s}) = (z(\mathbf{s}_1), ..., z(\mathbf{s}_n))'$. The mean and variance of the process is written as, $E(\mathbf{Z}(\mathbf{s})) = \mu(\mathbf{s})$ and $Var(\mathbf{Z}(\mathbf{s})) = \mathbf{\Sigma}$. To obtain prediction at unknown site $\mathbf{s}_0$, we can estimate the optimal prediction $Z(\mathbf{s}_0)$ as:

$$\hat{Z}(\mathbf{s}_0) = \hat{\mu}(\mathbf{s}_0) + C'\mathbf{\Sigma}^{-1}(Z(\mathbf{s}) - \mu(\mathbf{s})).$$

where, $C' = \text{Cov}(Z(\mathbf{s}), Z(\mathbf{s}_0))$. This type of kriging is known as *simple kriging*.

2.5.2 Ordinary Kriging

Assume that the mean process $\mu(\mathbf{s}) = \mu$ is known and does not vary with spatial locations $\mathbf{s}$, hence the model in equation (2.1) is written as:

$$Z(\mathbf{s}) = \mu + \eta(\mathbf{s}).$$

The estimated optimal prediction $Z(\mathbf{s}_0)$ at site $\mathbf{s}_0$ is known as the *ordinary kriging*, and is written as:

$$\hat{Z}(\mathbf{s}_0) = \hat{\mu} + C'\boldsymbol{\Sigma}^{-1}(Z(\mathbf{s}) - \hat{\mu}),$$

where, $\hat{\mu} = (\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1})^{-1}\mathbf{1}'\boldsymbol{\Sigma}^{-1}Z(\mathbf{s})$, and $\mathbf{1}$ is a vector with all elements equal to 1.

2.5.3 Universal Kriging

Assume the mean process $\mu(\mathbf{s})$ is unknown and it varies over space in the linear regression form $\mu(\mathbf{s}) = \mathbf{X}(\mathbf{s})'\boldsymbol{\beta}$, and the covariance function $\boldsymbol{\Sigma}$ is known as in the model (2.1). The model is written as:

$$\mathbf{Z}(\mathbf{s}) = \mathbf{X}^T(\mathbf{s})\boldsymbol{\beta} + \boldsymbol{\eta}(\mathbf{s}),$$

where, $\boldsymbol{\eta}(\mathbf{s}) = (\eta(\mathbf{s}_1), ..., \eta(\mathbf{s}_n))'$ and $\boldsymbol{\eta}(\mathbf{s}) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, $\mathbf{Z}(\mathbf{s}) = (Z(\mathbf{s}_1), ..., Z(\mathbf{s}_n))'$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ is the p (say) parameters and $\mathbf{X}^T(\mathbf{s})$ is the $p \times n$ covariate matrix. Hence, we can estimate the optimal prediction at site $\mathbf{s}_0$ as:

$$\hat{Z}(\mathbf{s}_0) = \mathbf{X}^T(s_0)\hat{\boldsymbol{\beta}} + C'\boldsymbol{\Sigma}^{-1}(\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\hat{\boldsymbol{\beta}}).$$

where, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T(\mathbf{s})\boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{s}))^{-1}\mathbf{X}^T(\mathbf{s})\boldsymbol{\Sigma}^{-1}\mathbf{Z}(\mathbf{s})$. This type of kriging is known as the *universal kriging*.

Classical methods of kriging lack the ability to incorporate uncertainty associated with parameter estimation, and they are based on the assumption of an isotropic covariance function that is sometimes unrealistic in environmental applications. To overcome these problems the Bayesian approaches to kriging have been developed. We briefly discuss this in the next Chapter 3.

2.6 Cartography

Cartography is the study of making maps in ways that represents the spatial information. In maps, spatial data are presented with a valid coordinate system. In cartography and spatial analysis, one of the important questions is how to measure distance of the earth's surface. The earth has irregular spherical shape, this makes it difficult to obtain actual measurement of the surface of the earth.

In this section we briefly describe how spatial statisticians and geographers determine the distance between two locations in the earth's surface.

2.6.1 Geodetic Distances

In spatial statistics we model the spatial dependence between two random variables as a function of the distance between the two sites where they were observed. The ordinary Euclidean distance can be used to measure the distance for data sets covering relatively small spatial domains. However, for large spatial regions, e.g., north America, we need to consider the curvature of the earth to calculate such distances, see Figure 2.4.

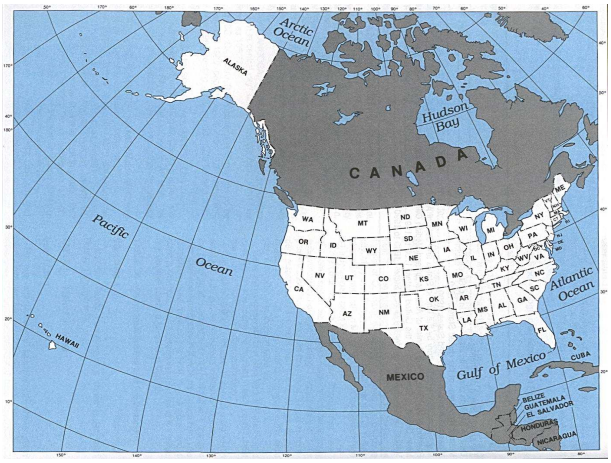


Figure 2.4: A map of north America illustrating the curvature pattern of the earth.

In geostatistics it is preferable to obtain distance between two observation sites using *geodetic* (or *geodesic*) distances but not using Euclidean distances (for further readings see Banerjee *et al.*, 2004, Chapter 1).

Geodetic distance is based on earth's longitude and latitude positions. Most common approach to measure the geodetic distance is known as the spherical law of Cosines. For example, for two points $P_1 = (\lambda_1, \theta_1)$ and $P_2 = (\lambda_2, \theta_2)$ on earth's surface, the geodetic distance d_{12} can be obtained as:

$$d_{12} = R \cos^{-1}(\sin\theta_1 \sin\theta_2 + \cos\theta_1 \cos\theta_2 \cos(\lambda_1 - \lambda_2))$$

where, R is the radius of the earth. Throughout this thesis we use the above formula for calculation of the geodetic distance between two spatial locations.

2.7 Summary

In this chapter we have reviewed the geostatistical methods and related concepts. We briefly discuss different types of spatial data and spatio-temporal processes. We also discuss some important characteristics of the space-time covariance functions. The kriging approaches are described to predict at unmonitored spatial locations. Some topics related to cartography are also discussed in this chapter.

Chapter 3

Review of Bayesian Modelling

3.1 Introduction

The Bayesian paradigm is used for making inference throughout this thesis. In this chapter we provide an overview of the key concepts in Bayesian modelling. A Bayesian approach is more natural than the traditional frequentist approaches since it lets us deal with the uncertainty in the model and its parameters. In Bayesian analysis the *prior* distribution has influence on the uncertainty of the model and the total uncertainty can be represented by a probability distribution. Particularly in environmental applications, it is important to evaluate the uncertainty and to give a scientific interpretation using probability statements. For more detailed introduction to Bayesian modelling, see Gelman *et al.* (2004), Bernardo and Smith (1994). For applications in spatial and spatio-temporal modelling, Banerjee *et al.* (2004) provide an overview of Bayesian modelling.

In Section 3.2 of this chapter, we describe the fundamental elements of the Bayesian paradigm. Section 3.3 provides a brief explanation of the Bayesian model choice criteria. Section 3.4 describes some space-time Bayesian modelling strategies in details. Finally in Section 3.5 we provide few summary remarks.

3.2 Bayesian Modelling and Computation

3.2.1 Bayesian Framework

In the Bayesian framework we update the prior knowledge using Bayes theorem to the *posterior* distribution. Let, $f(\mathbf{z}|\boldsymbol{\theta})$ be the *likelihood function* of parameters

$\boldsymbol{\theta} = (\theta_1, ..., \theta_p)'$ based on the observed data $\mathbf{z} = (z_1, ..., z_n)'$. If we have a prior distribution $\pi(\boldsymbol{\theta})$ for the parameters, then using the Bayes theorem we can obtain the posterior distribution as:

$$\pi(\boldsymbol{\theta}|\mathbf{z}) = \frac{f(\mathbf{z}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_\Theta f(\mathbf{z}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}, \tag{3.1}$$

where, the denominator of the above equation (3.1) is the integral over the parameter space Θ. This integral is also known as the *marginal likelihood* of the data $\mathbf{z}$ and it is free of the parameters $\boldsymbol{\theta}$, hence can be treated as a constant. That is why the posterior distribution is often written as proportional to the product of the likelihood and the prior distribution, i.e.,

$$\pi(\boldsymbol{\theta}|\mathbf{z}) \propto f(\mathbf{z}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

Bayesian methodology typically proceeds in the following steps: (i) write the likelihood of the parameters for a given set of data, (ii) assign prior distributions to the unknown parameters, (iii) calculate the posterior distribution and (iv) make inference based on the updated information in the posterior distribution.

3.2.2 Prior Choices

The choice of the prior distributions is a very important step in any Bayesian analysis. The most attractive choice of prior distributions should be the one that best takes into account any previous knowledge. These types of priors are known as *informative priors*. However, there is often no clear choice of prior distributions for unknown parameters.

For various likelihood functions, there exists prior distributions that lead to a posterior distribution, which comes from the same distribution family as the prior. These types of prior distributions are known as *conjugate priors*. For example, the prior distribution conjugate to a Bernoulli likelihood is a Beta distribution, and for a Poisson likelihood the conjugate prior distribution is the Gamma distribution. Choice of this type of prior distribution is attractive because of its straightforward computation. However, for different posterior and prior distribution families this type of prior might not be available.

To overcome this *conjugacy* problem, prior ignorance is used with a proper

prior specification with large variability. For example, the inverse-gamma distribution is used for the non-negative variance parameters in the models.

For the *noninformative prior* (also known as *vague prior*), the *uniform* distribution is commonly used. Besides in many situations *Jeffrey's rule* is applied to obtain the noninformative prior distribution (Gelman *et al.*, 2004, page, 62), where the noninformative prior distribution is taken as proportional to the square root of the determinant of the *Fisher's information matrix*.

3.2.3 Markov Chain Monte Carlo (MCMC)

In Bayesian analysis, complex hierarchical models are often analytically intractable and are hard to fit. Therefore *Markov chain Monte Carlo* (MCMC) methods are now popular for evaluating features of posterior distributions needed for making inference. In MCMC we generate a sequence of samples from the joint probability distribution of random variables. The purpose of such a sequence is to approximate the joint distribution, or to compute an integral (such as an expected value). For example, let $\mathbf{z}$, be the vector of the observed data and $\boldsymbol{\theta}$ the parameter vector. The MCMC algorithm generates a Markov chain, $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{n}$, from the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{z})$, where n is the number of MCMC samples. Then, the samples are used to estimate integrals using Monte Carlo methods. Thus, we get,

$$\hat{E}(h(\boldsymbol{\theta})) = \int_{\Theta} h(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{z})d\boldsymbol{\theta},$$

where, $h(\boldsymbol{\theta})$ is a function of $\boldsymbol{\theta}$.

There are number of different MCMC simulation techniques available, for example, Metropolis-Hastings algorithm, and Gibbs sampling. Details of these techniques are found in Gelman *et al.* (2004), and Chen *et al.* (2000).

Convergence of the MCMC algorithm is often hard to detect. Several convergence diagnostic methods exist, see for example, Gelman *et al.* (2004), Gilks *et al.* (1996). However, the time series plots of the MCMC iterates usually indicate the convergence properties of the MCMC algorithms. Ideally, the autocorrelation between successive iterates should be low and higher order autocorrelation should die down rapidly.

3.2.4 Metropolis-Hastings Algorithm

The Metropolis-Hastings (MH) algorithm (Metropolis *et al.*, 1953 and Hastings, 1970) is an MCMC method for obtaining random samples. Here, we draw samples from a non-standard posterior distribution by rejecting samples obtained from a proposal distribution in an appropriate fashion.

The MH algorithms are based on a Markov chain that depends on samples drawn from a proposal distribution and an acceptance-rejection mechanism. The proposal suggests an arbitrary next step in the chain and the acceptance-rejection step makes sure the appropriate limiting direction is maintained by rejecting unwanted moves of the chain. For example, let, $\pi(\boldsymbol{\theta}|\mathbf{z})$ be the density from which we want to sample. We chose a proposal density $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ denotes the current point. We write the MH algorithm as follows:

(i) Sample a candidate value $\boldsymbol{\theta}'$ from the proposal density $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$.

(ii) Calculate the acceptance probability $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min\left\{1, \frac{\pi(\boldsymbol{\theta}'|\mathbf{z})q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}|\mathbf{z})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\}$.

(iii) Sample a uniformly distributed random variable U on (0, 1).

(iv) If, $U < \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}')$ then accept the candidate value else assign the present value to the new value.

We use the MH algorithm in particular, if the posterior distribution and the conditional distributions are not standard distributions. There are some special cases of the MH algorithm as discussed below:

Metropolis Algorithm

The *Metropolis* algorithm is a especial case of the MH algorithm, where $q(.|\boldsymbol{\theta}) = q(\boldsymbol{\theta}|.)$. This yields the acceptance probability as:

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

Here, the acceptance ratio only depends on the ratio of the values of the target density $\frac{\pi(\boldsymbol{\theta}'|\mathbf{z})}{\pi(\boldsymbol{\theta}|\mathbf{z})}$.

Random-walk Metropolis

In *random-walk* (RW) Metropolis algorithm, we draw candidate from the following RW model,

$$\boldsymbol{\theta}' = \boldsymbol{\theta} + \tau\boldsymbol{\epsilon},$$

where, $\boldsymbol{\epsilon}$ is an independent error term with mean zero, and τ is a scaling factor, which we call the tuning parameter in this thesis. For a symmetric RW model, we get $q(.|\boldsymbol{\theta}) = q(\boldsymbol{\theta}|.)$, hence, the RW Metropolis algorithm is equivalent to the Metropolis algorithm.

3.2.5 Acceptance Rates

The tuning parameter determines the rate of acceptance in the Metropolis-Hastings algorithm. Large values allow bigger moves around the sample space with more rejections and small values yield a small rejection. A number of other factors effects the desired rate of acceptance, for example, choice of proposal distribution and the initial value of the chain. It is suggested that for Gaussian random walk proposals the desired acceptance rate is around 20-40% (see Gelman *et al.*, 1997, 2004, and references therein) when the parameter is one-dimensional.

3.2.6 Gibbs Sampler

The Gibbs sampler, introduced by Geman and Geman (1984), has been developed by Gelfand and Smith (1990). The Gibbs sampler simulates from multidimensional posterior distributions by iteratively sampling from the lower-dimensional conditional posterior distributions. Unlike the previous MH algorithms, the Gibbs sampler updates the chain one component at a time, instead of updating the entire vector. For example, starting from an initial value $\theta^{(0)}$, at iteration j, the Gibbs sampler draws:

$$\begin{aligned}
\theta_1^{(j)} &\sim \pi(\theta_1|\theta_2^{(j-1)}, \theta_3^{(j-1)}, ..., \theta_p^{(j-1)}, \mathbf{z}) \\
\theta_2^{(j)} &\sim \pi(\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)}, ..., \theta_p^{(j-1)}, \mathbf{z}) \\
... &\quad ... \\
\theta_p^{(j)} &\sim \pi(\theta_k|\theta_1^{(j)}, \theta_2^{(j)}, ..., \theta_{p-1}^{(j)}, \mathbf{z}).
\end{aligned}$$

The densities on the right hand sides of the above are called the *complete conditional* distributions or *full conditional* distributions. Throughout the thesis we will use the *Gibbs Sampling* approach to analyse and fit the Bayesian models.

3.3 Bayesian Model Choice Criteria

3.3.1 Bayes Factor

Suppose we have two models M_1 and M_2 with data $\mathbf{z}$ and the corresponding marginal likelihoods are $\pi(\mathbf{z}|M_1)$ and $\pi(\mathbf{z}|M_2)$. The Bayesian model choice criterion "Bayes factor" for these two models is given by,

$$BF = \frac{\pi(\mathbf{z}|M_1)}{\pi(\mathbf{z}|M_2)}$$

There are many methods available for approximating the 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 use the model choice criteria discussed in Section 3.3.3, which is most suitable when the Gaussian distribution is employed at the first stage of a hierarchical Bayesian model.

3.3.2 Deviance Information Criteria

Spiegelhalter *et al.* (2002) introduce the deviance information criteria (DIC) based on the posterior mean of the model parameters and the averages of the deviances (Dempster 1974) using a sample from the posterior distribution. Let $\mathbf{z}$ be the observed data with unknown quantities $\boldsymbol{\theta}$ and $\pi(\mathbf{z}, \boldsymbol{\theta})$ be the joint posterior distribution. We can write the Bayesian deviance as:

$$D(\boldsymbol{\theta}) = -2L(\boldsymbol{\theta}|\mathbf{z})$$

where, $L(\boldsymbol{\theta}|\mathbf{z})$ is the log-likelihood of the model. Now the goodness of fit of a model is obtained as $\bar{D} = E_{\boldsymbol{\theta}|\mathbf{z}}(D)$ and the model complexity is written as:

$$p_D = E_{\boldsymbol{\theta}|\mathbf{z}}(D) - D[E_{\boldsymbol{\theta}|\mathbf{z}}(\boldsymbol{\theta})]$$

Thus finally the DIC is obtained as:

$$DIC = E_{\boldsymbol{\theta}|\mathbf{z}}(D) + p_D = D[E_{\boldsymbol{\theta}|\mathbf{z}}(\boldsymbol{\theta})] + 2p_D$$

The model with lower DIC value indicates a better fitting model.

3.3.3 Predictive Model Choice Criteria

The predictive model choice criterion (PMCC), see e.g. Gelfand and Ghosh (1998), is suitable for comparing models with normally distributed error and is given by:

$$\text{PMCC} = \sum_{i=1}^{n} E(Z_{i,\text{rep}} - z_i)^2 + \sum_{i=1}^{n} \text{Var}(Z_{i,\text{rep}}), \tag{3.2}$$

where $Z_{i,\text{rep}}$ denotes a future replicate of the data z_i. 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. Throughout the thesis we use PMCC as a Bayesian model choice criteria.

3.3.4 Criteria for Validations

To compare the quality of predictions and forecasts obtained from the fitted models, in this thesis we use some validation criteria (see for example, Atkinson and Lloyd 1998, Moyeed and Papritz 2002, Stephenson 2006, and Yip 2009). We use the root mean squared error (RMSE), mean absolute error (MAE), relative bias (rBIAS), and relative mean separation (rMSEP). These validation criteria are defined as:

$$\begin{aligned}
RMSE &= \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\widehat{z}_i - z_i)^2} \\
MAE &= \frac{1}{m}\sum_{i=1}^{m}|\widehat{z}_i - z_i| \\
rBIAS &= \frac{1}{m\bar{z}}\sum_{i=1}^{m}(\widehat{z}_i - z_i) \\
rMSEP &= \sum_{i=1}^{m}(\widehat{z}_i - z_i)^2 / \sum_{i=1}^{m}(\bar{z}_p - z_i)^2,
\end{aligned}$$

where, m is the total number of observations we want to validate, z_i is the data indexed by i, $\widehat{z}_i$ is the prediction value, $\bar{z}$ and $\bar{z}_p$ are the arithmetic mean of the observations and predictions respectively.

3.4 Gaussian Process Models

A Gaussian process is a collection of random variables, any finite number of which have a Gaussian distribution with valid mean and variance. There is a huge literature on modelling spatio-temporal data based on Gaussian processes (see details in Section 1.5). In this section we describe some hierarchical Bayesian models, that are used in this thesis to analyse the daily maximum eight-hour average ozone levels.

3.4.1 Bayesian Linear Regression Models

Let $\mathbf{Z}$ be the $n \times 1$ response vector and $\mathbf{X}$ be an $n \times p$ design matrix. We write the linear model as:

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

where, $\boldsymbol{\beta}$ is the $p \times 1$ vector of parameters, σ^2 is the variance parameter of the error process and $\mathbf{I}$ is the $n \times n$ identity matrix.

Under Bayesian approach, we specify the prior distributions for the unknown model parameters. For example flat prior distributions as: $\pi(\boldsymbol{\beta}) \propto 1$, and for variance parameter as: $\pi(\sigma^2) \sim IG(a, b)$. Hence, we obtain the full conditional posterior distribution for β as:

$$\pi(\boldsymbol{\beta}|\sigma^2, \mathbf{z}) \sim N((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

and for σ^2 as:

$$\pi(\sigma^2|\boldsymbol{\beta}, \mathbf{z}) \sim IG\left(a + \frac{n}{2}, b + \frac{1}{2}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})\right)$$

We use Gibbs sampler to draw samples from the conditional distributions.

Suppose that we have observed new predictor values $\tilde{\mathbf{X}}$ and we want to predict

the outcome $\tilde{\mathbf{Z}}$. Thus, we obtain the posterior predictive distribution as:

$$\pi(\tilde{\mathbf{Z}}|\mathbf{z}) = \int \pi(\tilde{\mathbf{Z}}|\mathbf{z}, \boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta}, \sigma^2|\mathbf{z})d\boldsymbol{\beta}d\sigma^2$$

where, $\pi(\tilde{\mathbf{Z}}|\mathbf{z}, \boldsymbol{\beta}, \sigma^2) \sim N(\tilde{\mathbf{X}}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, and we can use MCMC methods to evaluate this integral.

3.4.2 Bayesian Kriging

Kitanidis (1986) provided one of the earliest articles using the Bayesian approach in spatial interpolation. He developed a theoretical framework for deriving the predictive distribution of a spatially dependent random variable with a covariance field assumed to be known. He derived the kriging estimator and its variance as special cases of the posterior mean and variance respectively under the Bayesian paradigm. However, the assumption of a known covariance function makes it difficult for application to wider settings. So, Handcock and Stein (1993) advanced the Kitanidis theory by assuming the covariance function as a functional form of the parameters. Their approach is extended by De Oliveria *et al.* (1997) where the random fields are non-linearly transformed to Gaussian distributions and uncertainty associated with such transformation is also considered. By now there is a substantial literature on this, see for example, Ecker and Gelfand (1997), Banerjee *et al.* (2004), Le and Zidek (2006) and the references therein.

The basic Bayesian model with Gaussian random effects can be written as:

$$\mathbf{Z}(\mathbf{s}) = \boldsymbol{\mu}(\mathbf{s}) + \boldsymbol{\eta}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s}),$$

where $\mathbf{Z}(\mathbf{s}) = (Z(\mathbf{s}_1), ..., Z(\mathbf{s}_n))'$ is the observed data, $\boldsymbol{\mu}(\mathbf{s})$ is the mean function at location $\mathbf{s}$, defined as $\boldsymbol{\mu}(\mathbf{s}) = \mathbf{X}(\mathbf{s})'\boldsymbol{\beta}$. The residuals are partitioned where $\boldsymbol{\eta}(\mathbf{s})$ is the spatially correlated error and distributed as zero mean stationary Gaussian spatial process with covariance $\sigma_\eta^2\Sigma$, where Σ is a correlation matrix with $\Sigma_{ij} = \kappa(\mathbf{s}_i - \mathbf{s}_j; \phi)$, $i, j = 1, ..., n$ and $\kappa(.)$ is a valid isotropic correlation function, ϕ is a spatial decay parameter. The second part $\boldsymbol{\epsilon}(\mathbf{s})$ is the non-spatial uncorrelated pure error also distributed normally with mean zero and variance $\sigma_\epsilon^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Thus, we can write,

$$\mathbf{Z}(\mathbf{s})|\boldsymbol{\theta} \sim N(\mathbf{X}'(\mathbf{s})\boldsymbol{\beta}, \sigma_\eta^2\Sigma + \sigma_\epsilon^2\mathbf{I}).$$

Let the collection of model parameters be given by $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_\eta^2, \sigma_\epsilon^2, \phi)'$, $\boldsymbol{\theta} \in \Theta$ and let $\pi(\boldsymbol{\theta})$ denote the prior distribution for $\boldsymbol{\theta}$. The posterior distribution is obtained as:

$$\pi(\boldsymbol{\theta}|\mathbf{z}(\mathbf{s})) \propto f(\mathbf{z}(\mathbf{s})|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

where, $f(\mathbf{z}(\mathbf{s})|\boldsymbol{\theta})$ is the likelihood function of the parameters. The posterior predictive distribution $Z(\mathbf{s}'|.)$ at an unobserved site $\mathbf{s}'$ is given by:

$$\pi(Z(\mathbf{s}')|\mathbf{z}(\mathbf{s})) = \int_\Theta \pi(Z(\mathbf{s}')|\boldsymbol{\theta}, \mathbf{z}(\mathbf{s}))\pi(\boldsymbol{\theta}|\mathbf{z}(\mathbf{s}))d\boldsymbol{\theta}$$

where, $\pi(Z(\mathbf{s}')|\boldsymbol{\theta}, \mathbf{z}(\mathbf{s}))$ is the probability density function of $Z(\mathbf{s}')$ at an unobserved site given $\boldsymbol{\theta}$ and $\mathbf{z}(\mathbf{s})$, and $\pi(\boldsymbol{\theta}|\mathbf{z}(\mathbf{s}))$ is the posterior distribution of $\boldsymbol{\theta}$. MCMC methods (see details in Section 3.2) can be used to evaluate the above integral.

3.4.3 Bayesian Spatio-temporal Gaussian Process (GP) Models

Let $Z(\mathbf{s}_i, t)$ denote the observed point-referenced data and $O(\mathbf{s}_i, t)$ is the true value corresponding to $Z(\mathbf{s}_i, t)$ at site $\mathbf{s}_i$, at time t, $i = 1, ..., n$ and $t = 1, ..., T$. In vector notation, $\mathbf{Z}_t = (Z(\mathbf{s}_1, t), ..., Z(\mathbf{s}_n, t))'$ and $\mathbf{O}_t = (O(\mathbf{s}_1, t), ..., O(\mathbf{s}_n, t))'$, we write the spatio-temporal linear regression models as:

$$\mathbf{Z}_t = \mathbf{O}_t + \boldsymbol{\epsilon}_t, \tag{3.3}$$

$$\mathbf{O}_t = \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\eta}_t \tag{3.4}$$

where, $\mathbf{X}_t$ is the $n \times p$ design matrix of covariate effects and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ is the $p \times 1$ vector of parameters respectively. The term $\boldsymbol{\epsilon}_t = (\epsilon(\mathbf{s}_1, t), ..., \epsilon(\mathbf{s}_n, t))' \sim N(\mathbf{0}, \sigma_\epsilon^2\mathbf{I}_n)$ is the independently distributed white noise error with variance σ_ϵ^2 also known as the nugget effect, and $\mathbf{I}_n$ is the $n \times n$ identity matrix. The term $\boldsymbol{\eta}_t = (\eta(\mathbf{s}_1, t), ..., \eta(\mathbf{s}_n, t))' \sim N(\mathbf{0}, \Sigma_\eta)$ is the spatially correlated error, with $n \times n$ variance-covariance matrix $\Sigma_\eta = \sigma_\eta^2 S_\eta = \sigma_\eta^2 \kappa(\mathbf{s}_i, \mathbf{s}_j; \phi, \nu)$, $i, j = 1, ..., n$; σ_η^2 is the site invariant common variance and $\kappa(.; \phi, \nu)$ is the spatial correlation matrix with spatial decay ϕ and smoothness ν parameters. The errors $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\eta}_t$ are assumed to be independent of each other.

Suppose we want to predict at location $\mathbf{s}'$ at time t. The posterior predictive distribution for $Z(\mathbf{s}', t)$ is obtained by integrating over the parameters with

respect to the joint posterior distribution as:

$$\pi(Z(\mathbf{s}',t)|\mathbf{z}) = \int \pi(Z(\mathbf{s}',t)|O_l(\mathbf{s}',t),\sigma_\epsilon^2,\mathbf{z})\pi(O(\mathbf{s}',t)|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{z})dO(\mathbf{s}',t)d\boldsymbol{\theta} \quad (3.5)$$

where, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_\eta^2, \phi, \nu)'$. According to (3.3) we obtain,

$$Z(\mathbf{s}',t) \sim N(O(\mathbf{s}',t), \sigma_\epsilon^2), \quad (3.6)$$

where, $O(\mathbf{s}',t)$ is the true prediction values at site $\mathbf{s}'$, and we obtain the samples for $O(\mathbf{s}',t)$ from:

$$\begin{pmatrix} O(\mathbf{s}',t) \\ \mathbf{O}_t \end{pmatrix} \sim N\left[\begin{pmatrix} X(\mathbf{s}',t)\boldsymbol{\beta} \\ \mathbf{X}_t\boldsymbol{\beta} \end{pmatrix}, \sigma_\eta^2 \begin{pmatrix} 1 & S_{\eta,12} \\ S_{\eta,21} & S_\eta \end{pmatrix}\right], \quad (3.7)$$

where, $S_{\eta,12}$ is $1 \times n$ with ith entry given by, $\kappa(\mathbf{s}_i, \mathbf{s}'; \phi, .)$, $i = 1, ..., n$ and $S_{\eta,12} = S'_{\eta,21}$. In summary, we draw sample $\boldsymbol{\theta}^{(j)}$, $j \geq 1$, from the full conditional posterior distributions and then draw $O^{(j)}(\mathbf{s}',t)$ from (3.7) and finally draw $Z^{(j)}(\mathbf{s}',t)$ from (3.6).

3.4.4 Bayesian Spatio-temporal Dynamic Linear Models (DLM)

The DLM, developed as a result of the popularity of Kalman filtering (Kalman 1960) methods, provide a dynamical state-space system that is thought to evolve from a pair of state and observation equations. The DLM is a state-space model and its Bayesian version is introduced by West and Harrison (1997), where in each time point the model parameter changes because the model is assumed it is locally appropriate in time. This sequential and dynamic approach is used in the spatio-temporal modelling by Stroud *et al.* (2001). Huerta *et al.* (2004) elaborate the DLM for temporal non-stationarities in the data. In general, the DLM is written as:

$$\mathbf{Z}_t = \mathbf{F}_t\boldsymbol{\theta}_t + \boldsymbol{\nu}_t, \quad t \geq 1, \quad (3.8)$$

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad t \geq 1, \quad (3.9)$$

where, the first equation (3.8) is known as the *observation equation* and the second equation (3.9) is known as the *system equation*. In the DLM equations, $\mathbf{Z}_t =$

$(Z(s_1,t),...,Z(s_n,t))'$ denote the observation vector for any $1 \leq t \leq T$, and T is the maximum number of time points in the data. Here $\boldsymbol{\nu}_t = (\nu(s_1,t),...,\nu(s_n,t))'$ is the spatially correlated error and is assumed to follow the $N(\mathbf{0}, \Sigma_\nu)$ distribution. The term $\mathbf{F}_t$ is the matrix of the covariate effects. We also assume that $\boldsymbol{\omega}_t \sim N(\mathbf{0}, \sigma^2_\omega \mathbf{I})$, where $\mathbf{I}$ denotes the identity matrix of appropriate order, and the initial state $\boldsymbol{\theta}_0$ is assumed to follow $N(\boldsymbol{\mu}, \sigma^2_\theta \mathbf{I})$ distribution for suitable values of the hyper-parameters $\boldsymbol{\mu}$ and σ^2_θ. The observations are spatially correlated, hence a spatially correlated covariance matrix must be assumed for Σ_ν. For convenience, in this thesis we assume the exponential covariance function to model spatial dependence and let

$$\Sigma_\nu = \sigma^2_\nu S_\nu = \sigma^2_\nu \exp(-\phi_\nu D)$$

where $\phi > 0$ is a spatial correlation decay parameter assumed to be known, and the $n \times n$ distance matrix D has elements d_{ij}, the distance between $\mathbf{s}_i$ and $\mathbf{s}_j$, $i, j = 1, ..., n$.

Huerta *et al.* (2004) considered the above DLM structure and used temperature as a covariate effect on ozone. They and similarly Duo *et al.* (2010) used a seasonal component $S_{kt}(\mathbf{a}_k)$, $k = 1, 2$, at time t, that consists of *sine* and *cosine* terms to describe the seasonal pattern of the ozone concentrations (see Section 1.5). Unlike these authors we do not include any seasonal term in the models as the seasonal terms are more relevant for modelling the diurnal cyclic components often present in the hourly ozone data.

To accommodate the covariate effects $\mathbf{X}_t$ and intercept at time t, we can assume that $\mathbf{F}_t = (1, \mathbf{X}_t)$; consequently $\boldsymbol{\theta}_t = (\alpha_t, \beta_t)'$. The error $\boldsymbol{\omega}_t$ in the DLM system equation, hence is written as, $\boldsymbol{\omega}_t = (\omega^\alpha_t, \omega^\beta_t)'$, where $\omega^\alpha_t \sim N(0, \sigma^2_\omega)$ and $\omega^\beta_t \sim N(0, \sigma^2_{\omega\beta})$.

In Bayesian structure, the function of the joint posterior distribution of the DLM based on Huerta *et al.* (2004) is $\pi(\boldsymbol{\theta}, \sigma^2_\omega, \sigma^2_{\omega\beta}, \sigma^2_\theta, \phi|\mathbf{z})$. So, the spatial interpolation at new site $\mathbf{s}'$ and time t is obtained from the posterior predictive distribution,

$$\pi(Z(\mathbf{s}', t)|\boldsymbol{\theta}_t, \sigma^2_\omega, \sigma^2_{\omega\beta}, \sigma^2_\nu, \sigma^2_\theta, \phi)$$

The MCMC algorithm can be applied to obtain the samples form the posterior distributions.

3.4.5 Bayesian Spatio-temporal Auto-regressive (AR) Models

We consider the AR model as proposed by Sahu *et al.* (2007), to analyse spatio-temporal ozone concentration data. Let $\mathbf{Z}_{lt} = (Z_l(\mathbf{s}_1, t), ..., Z_l(\mathbf{s}_n, t))'$ be the vector of observed and $\mathbf{O}_{lt} = (O_l(\mathbf{s}_1, t), ..., O_l(\mathbf{s}_n, t))'$ be the true square-root ozone concentration levels in day t and year l, $t = 1, ..., T$, $l = 1, ..., r$ at sites $\mathbf{s}$. In matrix notation the AR model is as follows:

$$\begin{aligned}
\mathbf{Z}_{lt} &= \mathbf{O}_{lt} + \boldsymbol{\epsilon}_{lt}, \qquad \boldsymbol{\epsilon}_{lt} \sim N(\mathbf{0}, \sigma^2_\epsilon I), \\
\mathbf{O}_{l1} &= \mu_l \mathbf{1} + \boldsymbol{\gamma}_l, \qquad \boldsymbol{\gamma}_l \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\gamma), \\
\mathbf{O}_{lt} &= \xi_l \mathbf{1} + \rho \mathbf{O}_{l(t-1)} + \mathbf{X}_{lt}\boldsymbol{\beta} + \boldsymbol{\eta}_{lt}, \qquad \boldsymbol{\eta}_{lt} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\eta),
\end{aligned}$$

where, $\boldsymbol{\epsilon}_{lt} = (\epsilon_l(\mathbf{s}_1, t), ..., \epsilon_l(s_n, t))$ is a white-noise process with σ^2_ϵ as the nugget effect. The term $\boldsymbol{\gamma}_l = (\gamma_l(\mathbf{s}_1, 1), ..., \gamma_l(\mathbf{s}_1, 1))'$ is the regional effect in year l at site $\mathbf{s}$ over a global level μ_l. The term $\boldsymbol{\eta}_{lt} = (\eta_l(\mathbf{s}_1, t), ..., \eta_l(\mathbf{s}_n, t))'$ is the spatially correlated error, ρ is the autoregressive process parameter with $0 < \rho < 1$, ξ_l is the global annual intercept, and $\mathbf{X}_{lt}$ is the covariate effects on true ozone levels. The covariance functions $\boldsymbol{\Sigma}_\eta = \sigma^2_\eta S_\eta$ and $\boldsymbol{\Sigma}_\gamma = \sigma^2_\gamma S_\gamma$ have elements $\sigma^2_\eta \exp(-\phi_\eta d_{ij})$ and $\sigma^2_\gamma \exp(-\phi_\gamma d_{ij})$. The term $\mathbf{X}_{lt}\boldsymbol{\beta}$ is written as:

$$\mathbf{X}_{lt}\boldsymbol{\beta} = \begin{pmatrix} \mathbf{x}'_l(s_1, t) \\ ... \\ \mathbf{x}'_l(s_n, t) \end{pmatrix} \begin{pmatrix} \beta_1 \\ ... \\ \beta_p \end{pmatrix} = \begin{pmatrix} x_{l1}(s_1, t) & x_{l2}(s_1, t) & ... & x_{lp}(s_1, t) \\ ... & ... & ... & ... \\ x_{l1}(s_n, t) & x_{l2}(s_n, t) & ... & x_{lp}(s_n, t) \end{pmatrix} \begin{pmatrix} \beta_1 \\ ... \\ \beta_p \end{pmatrix}$$

We obtain the conditional mean and variance of $\mathbf{Z}_{lt}$ as $E(\mathbf{Z}_{lt}|\mathbf{O}_{lt}) = \mathbf{O}_{lt}$ and $Var(\mathbf{Z}_{lt}|\mathbf{O}_{lt}) = \sigma^2_\epsilon \mathbf{I}$ respectively. Hence, $\mathbf{Z}_{lt}|\mathbf{O}_{lt} \sim N(\mathbf{O}_{lt}, \sigma^2_\epsilon \mathbf{I})$. Again, the conditional mean and variance of $\mathbf{O}_{lt}$ is obtained as: $E(\mathbf{O}_{lt}|\mathbf{O}_{l(t-1)}) = \boldsymbol{\vartheta}_{lt}$ and $Var(\mathbf{O}_{lt}|\mathbf{O}_{l(t-1)}) = Var(\boldsymbol{\eta}_{lt}) = \boldsymbol{\Sigma}_\eta$, where, $\boldsymbol{\vartheta}_{lt} = \xi_l \mathbf{1} + \rho \mathbf{O}_{l(t-1)}$. Hence, $\mathbf{O}_{lt}|\mathbf{O}_{l(t-1)} \sim N(\boldsymbol{\vartheta}_{lt}, \boldsymbol{\Sigma}_\eta)$, for $t = 2, ..., T$. Let $\boldsymbol{\theta}$ represent all parameters in the models and written as $\boldsymbol{\theta} = (\mu_l, \xi_l, \rho, \boldsymbol{\beta}, \sigma^2_\epsilon, \sigma^2_\gamma, \sigma^2_\eta)$.

A simpler version of the AR model is also available (Sahu, 2011), where the true values are modelled for time t, starting from day one, i.e., $t = 1, ..., T$ and $l = 1, ..., r$ using the equation as:

$$\mathbf{O}_{lt} = \rho \mathbf{O}_{lt-1} + \mathbf{X}_{lt}\beta + \boldsymbol{\eta}_{lt} \tag{3.10}$$

where, the initial value for $\mathbf{O}_{l0}$ is assigned a prior distribution with mean μ_l and covariance matrix $\Sigma_l = \sigma_l^2 S_0$, where S_0 has elements $\exp(-\phi_0 d_{ij})$, $i, j = 1, ..., n$.

The predictive distribution of the observation $Z_l(\mathbf{s}', t')$ at location $\mathbf{s}'$ and at time t is written as:

$$Z_l(\mathbf{s}', t) \sim N(O(\mathbf{s}', t), \sigma_\epsilon^2)$$

Hence, the posterior predictive distribution can be obtained as:

$$\begin{aligned} \pi(Z_l(\mathbf{s}', t)|\mathbf{z}) &= \int \pi(Z_l(\mathbf{s}', t)|O(\mathbf{s}', t), \sigma_\epsilon^2) \times \pi(O_l(\mathbf{s}', t)|\boldsymbol{\theta}, \mathbf{z}^*) \\ &\quad \times \pi(\boldsymbol{\theta}, \mathbf{z}^*|\mathbf{z}) \times \pi(\boldsymbol{\theta}|\mathbf{z}) dO_l(\mathbf{s}', t) d\mathbf{z}^* d\boldsymbol{\theta}, \end{aligned}$$

where, $\mathbf{z}^*$ denote the missing data, $\mathbf{z}$ denote the all non-missing data, see details in Sahu *et al.* (2007). Similar to all other Bayesian approaches we can use MCMC methods to draw samples from the posterior distribution.

3.5 Summary

In this chapter we provide a short review of Bayesian modelling. We discuss the Bayesian framework, choices for different types prior distributions and the MCMC algorithms. We also provide several model choice criteria for choosing different Bayesian models but only use the PMCC further in the later chapters of this thesis. Some Gaussian process spatio-temporal models are also discussed that are used in this thesis for analysing daily ozone concentration levels.

Chapter 4

Data Description

4.1 Introduction

We have already mentioned in Chapter 1 that our main interest is to model and analyse the daily maximum eight-hour average ozone concentration levels in a study region in the eastern US. There are large number of ozone monitoring sites in the eastern US, and there are problems associated with data collection methods in many of these sites. A lot of missing data arises due to this. In addition, there are instances of extreme observations at some of these sites. In this chapter we provide details regarding the procedures we have adopted for data cleaning and editing so that the data can be readily used for modelling purposes.

The eastern US ozone data set that we obtain after cleaning has daily ozone concentration levels for 153 days in the ozone season (May to September) for years 1997 to 2006 from 691 locations, see Figure 4.1. In this chapter we also provide many summary statistics and graphical displays to describe this large data set. We also obtain the summary statistics used to monitor the primary ozone standard as defined in Section 1.3.2. These summary statistics will be used in model based analysis in the later chapters.

Apart from the observed values, we have also obtained the CMAQ output for the daily maximum eight-hour ozone concentration levels, see Section 1.4 for details on CMAQ. As expected there is no missing data in these computer output. We present summaries of these forecasts and compare with those of the observed ozone concentration data.

It is well known that ground level ozone is affected by meteorological vari-

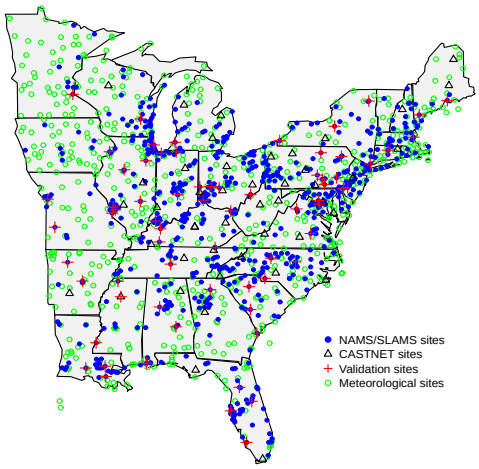


Figure 4.1: A plot of the 691 ozone monitoring locations in the eastern US, among them 646 are from NAMS/SLAMS and 45 are from CASTNETS. Hold-out sites for model validation are superimposed in the map together with the 746 meteorological monitoring sites in the eastern US.

ables such as, temperature, relative humidity and wind speed (see Section 1.5). Hence, in our study we also include these meteorological variables obtained from the National Climatic Data Center (NCDC)[1] of US Department of Commerce. We provide the data processing and summary statistics of these meteorological variables.

4.2 Daily Ozone Data

For 691 monitoring sites in the eastern US, with 10 years of data for 153 days in each year we get 1,057,230 observations for ozone concentrations. However, among them 110,363 (= 10.44%) observations are missing, and each site has more than 50% data of ozone levels. We also observe among these sites, 646 are from NAMS/SLAMS and rest of the site (45 sites) are from CASTNET (see Figure 4.1). The information regarding the NAMS/SLAMS and the CASTNET sites are given in Section 1.2.

[1]see, http://www.ncdc.noaa.gov/oa/ncdc.html

4.2.1 Data Preparation, Editing and Cleaning

The USEPA collects daily ozone concentration levels from about 1700 monitoring sites covering the 50 states of US. We consider a part of the eastern US as our study region (see Figure 4.1), where we finally have data from 691 sites.

Sites with more than 50% missing observations are discarded from the 1700 monitoring sites. We also remove ozone monitoring sites on the offshore area as we want to model the ozone concentration levels in the inland areas only.

During the 10 years of data collection the original locations of about 20 sites (out of 691) moved to a new location, which is a short distance away. For convenience, we treat the two locations to be the same and reference the combined site by the longitude latitude combination of the most recent site. For example, a site in the state of Alabama has longitude -87.005 for the period 1997 to 1999, but the same site has different longitude value -87.004 for the period 2000 to 2006. Hence, we replace the site longitude position in 1997-1999 by the longitude position in 2000-2006.

Additionally, there are 15 pairs of sites in our data set, which are less than a kilometre apart. For convenience and to reduce the number of sites for modelling we combine these pairs of sites as follows:

- We divide 15 pairs of sites into two categories, according to the nature of the observed data. The first category contains the 13 pairs for which there are relatively small differences between the ozone concentration levels. Figure 4.2 provides a typical example of the difference between the ozone concentration levels. We combine the observed ozone concentrations from these types of pairs by taking simple average of the available data from the two sites and we also refer the combined site by one of the two sites that is selected arbitrarily.

  In this category, six pairs of sites have missing values in only one or both of the two sites in the pair. For these pairs we simply combine the two sites by replacing the missing observation by the available observation and if both sites in the pair are missing then that observation in the combined site is treated as missing. Here, we refer the combined site by the site that has less missing observations in the pair.

- In the second category, the observation from one site contained many high

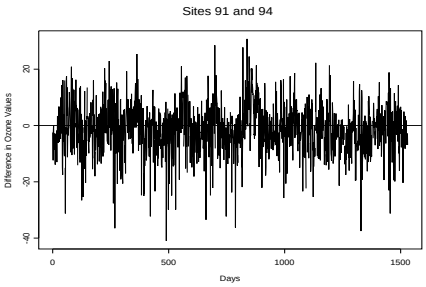


Figure 4.2: Time series plot of relatively small differences in ozone levels for a pair of sites.

out of range values, such as 500, see for example Figure 4.3. For this pair we replace such outlying observations by the observations from the other site in the pair, and we refer the combined site by the site that has no outlier.

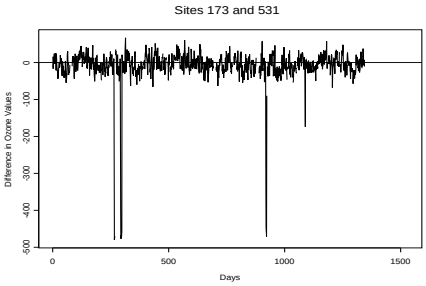


Figure 4.3: Time series plot of differences in ozone levels for a pair of sites that represent the second category: extreme observation.

4.2.2 Descriptive Statistics

In this section we discuss some summary statistics and graphical displays of the data set we prepared in the previous section. Recall that we have daily data from 691 monitoring sites for 153 ($= T$) days in a year (from May 1 to September 30) for 10 ($= r$) years. Out of these 1,057,230 ($= nrT$) possible observations, 110,363 (i.e., 10.44%) are missing. The first three years contain a higher percentage ($15\% - 25\%$) of missing observations compared to the later years. This is possibly because of the improvement in data collection methods after the millenium. However, the percentage of missingness increased to a double

digit number after 2004.

From Table 4.1, we can see that the available ozone concentration values range from 0.22 ppb to 246.22 ppb with mean 50.41 ppb and median 49.37 ppb.

Minimum	Mean	Median	Maximum
0.22	50.41	49.37	246.22

Table 4.1: Summary statistics for daily maximum eight-hour average ozone concentration levels in parts per billion (ppb).

The box-plot of ozone concentration values by year are given in Figure 4.4. Here, we can observe that the overall level goes up in the year 1998, comes down to the lowest level in 2000, and then rises again and comes down in the year 2004. After that the level again rises in 2005 and comes down in 2006.

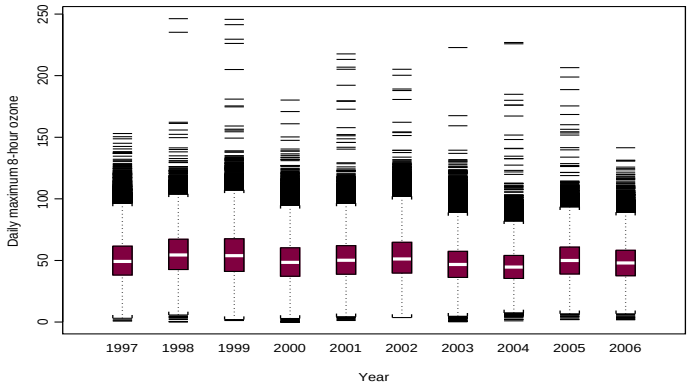


Figure 4.4: Box-plot of daily maximum eight-hour ozone concentration levels by years.

Figure 4.5 shows the levels of ozone concentration for different months. Here, we can observe that on average the ozone levels are highest in July and August and lowest in September. The levels in May and June are similar.

Figure 4.6 represents the box-plot of the ozone concentration levels of the different states within our study region of the eastern US. Some states, e.g., Maryland and Tennessee have much higher ozone levels than some others e.g., Maine and Vermont. The average ozone levels in most states fall between these two extremes and the levels in a state like New York, seems to represent a typical state. To illustrate, we shall analyse the ozone levels observed in New York in much more detail in Chapter 5.

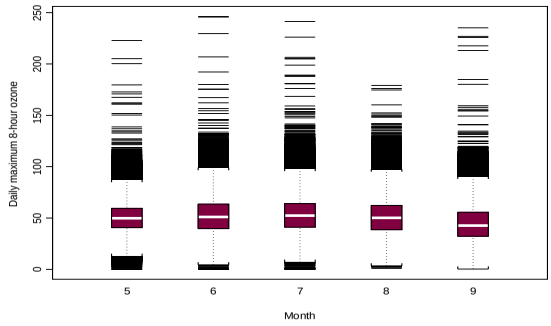


Figure 4.5: Box-plot of daily maximum eight-hour ozone concentration levels by months.

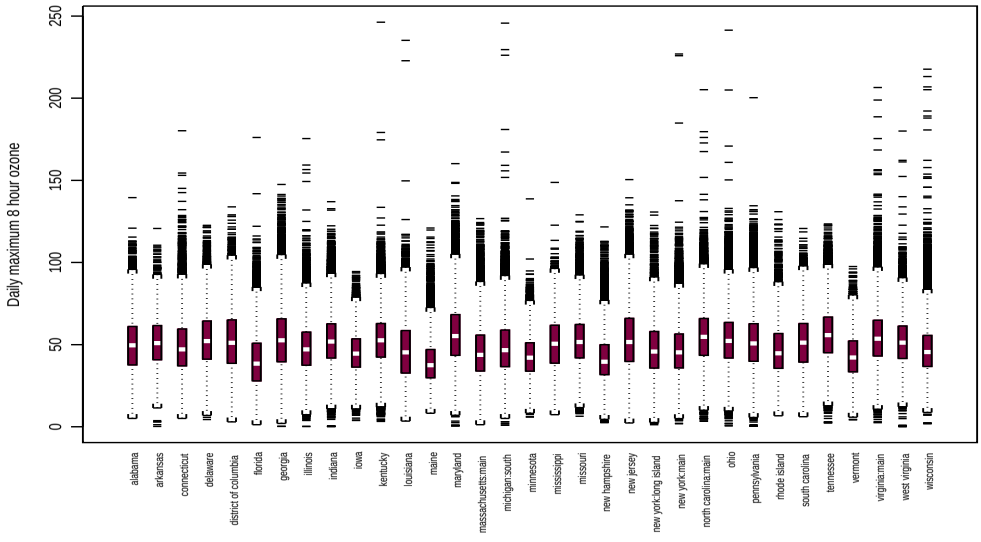


Figure 4.6: Box-plot of daily maximum eight-hour ozone concentration levels by states.

4.2.3 Annual 4th Highest Maximum Ozone Concentrations

Figure 4.7 provides a time series plot of the annual 4th highest maximum ozone concentrations for 691 sites in the eastern US. The figure shows the presence of some outlying monitoring sites for which there were some unusually high level of ozone concentration values. Most of the sites had their annual 4th highest concentration values greater than 85 ppb, which is the standard used in this thesis, see Section 1.3.

Most of the sites show a regular pattern, which yields an increase in 1998 and 1999, then decrease in 2000, again increase in 2002, decrease in 2004 and finally after a little increase in 2005 it decreased in 2006. This pattern of 4th highest maximum ozone concentrations is approximately similar to the pattern showed by the box-plot of the daily maximum eight-hour ozone concentration levels provided in Figure 4.4.

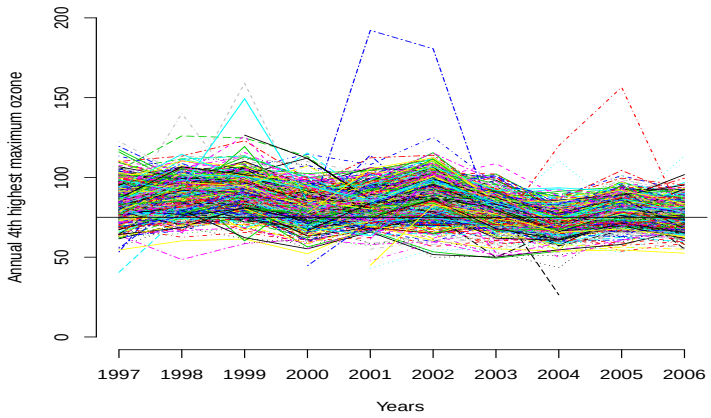


Figure 4.7: Time series plot of the 4th highest maximum ozone concentrations for 691 sites in the eastern US.

4.2.4 Three-Year Rolling Averages

The three-year rolling average of the annual 4th highest daily maximum ozone levels is obtained by taking the average of three years and aligning that with the last year of averaging, see Section 1.2.2.

For our eastern US data set we calculate the three year rolling averages for the years 1999-2006. The time-series plot of the three-year rolling averages shows a

downward slope in time (see Figure 4.8). Except for a few outliers the three-year averages range from 60-110 ppb.

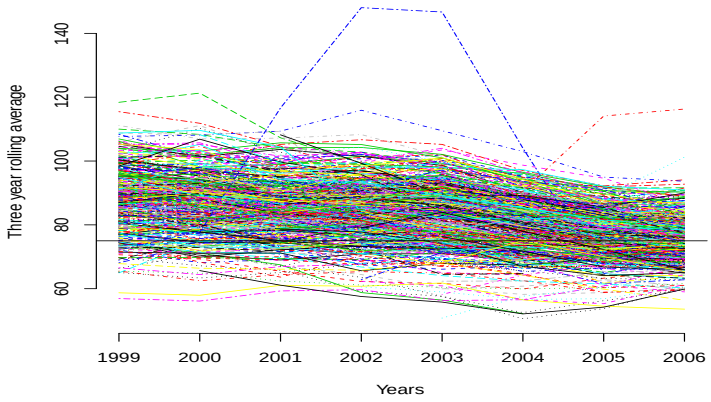


Figure 4.8: Time series plot of three-year rolling average of the 4th highest maximum ozone concentrations for 691 sites in the eastern US.

We discuss these outliers in Section 4.5.

4.3 CMAQ Output

In this thesis we use the CMAQ output as a covariate in the modelling (for details of CMAQ, see Section 1.4). These data are obtained for 9119 grid cells that covers the eastern US, see Figure 4.9. However, to use the data in modelling ozone concentration levels, we need to find the appropriate number of CMAQ grid cell points that match with the number of ozone monitoring sites. We have CMAQ output for 153 days in 2006 and 21 days (June 24 to July 15) in 2010. In this thesis, we use a part of the first set of CMAQ output as a covariate to compare different modelling strategies for ozone levels (see Chapter 5) and the second set of data is used to obtain next day forecasts for ozone levels using a novel methodology (see Chapter 7).

4.3.1 Data Preparation for CMAQ Output

We have already mentioned that the total number of observed ozone monitoring sites is 691, hence we need to locate possibly 691 grid cells for the CMAQ data.

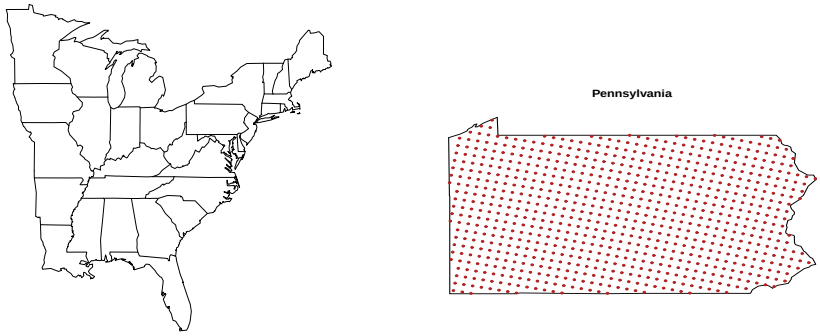


Figure 4.9: Panel (a) shows the 9119 CMAQ grid cells covering our study region in the eastern US. Panel (b) represents the CMAQ grid cells for the state of Pennsylvania.

Usually, in each grid cell we obtain one ozone monitoring site. However there are some grid cells where we can get more than one monitoring sites, for example, in the District of Columbia we have seen five ozone monitoring sites that are in one CMAQ grid cell. Therefore, for modelling purposes, the CMAQ values of that grid cell is used for modelling ozone levels of all five ozone monitoring sites in the District of Columbia.

4.3.2 Descriptive Statistics for CMAQ Output

The CMAQ output provides broadly similar patterns as the observed ozone values. However, there exists dissimilarities between them. For example, Table 4.2 shows that the CMAQ output varies from 2.70 ppb to 131.00 ppb where the observed ozone varies from 1.88 ppb to 141.50 ppb in 2006. The average and median of the forecast values and observed ozone levels also differ slightly.

	Minimum	Mean	Median	Maximum
Observed O_3	1.88	48.19	48.00	141.50
CMAQ output	2.70	51.25	51.02	131.00

Table 4.2: Summary statistics for daily maximum eight-hour average ozone concentration levels and CMAQ forecast values in ppb in year 2006.

4.4 Meteorological Data

The meteorological data for the variables: maximum temperature, dew points and average wind speed are obtained from the NCDC[2]. Figure 4.1 shows 746 monitoring sites of the meteorological variables together with the ozone monitoring sites in the eastern US. We have a total of 3,424,140 observations for 153 days in 10 years from 746 meteorological monitoring sites for the three variables. Among them 682,351 (i.e., 19.93%) are missing.

4.4.1 Data Preparation for Meteorological Variables

For convenience, we use the temperature on the $^{\circ}C$ (degree Celsius) scale rather than the $^{\circ}F$ (degree Fahrenheit) scale as used in the US. We also work with relative percentage humidity obtained using[3]:

$$\frac{T_d \times a}{T_d + b} = \frac{T \times a}{T + b} + \ln\left[\frac{RH}{100}\right] \tag{4.1}$$

where, $a = 17.271$ and $b = 237.7^{\circ}C$ are fixed constants, T_d is the dew point and T is the temperature in $^{\circ}C$, and RH is the percentage relative humidity.

We can see from Figure 4.1 that there are some sites where both meteorological and ozone concentration data are observed and there are also some other sites where only one type of data are observed. This misalignment in the data are handled using the spatial kriging (see Section 2.5) method. This approach of handling misalignment using kriging is applied in different literature (for example see, Reich *et al.*, 2011). Thus, after kriging we obtain a final non-missing data set of the meteorological variables for 691 ozone monitoring sites in the eastern US. Details of managing misalignment are discussed in Section 6.6.4 of this thesis.

4.4.2 Descriptive Statistics for Meteorological Variables

In this section we describe some summary statistics related to the maximum temperature, relative humidity and average wind speed in the eastern US. We can observe from Table 4.3 that maximum temperature varies from $-1.30^{\circ}C$ to $42.28^{\circ}C$. The mean and median RH are 5.74% and 5.66% respectively. The

[2]National Climatic Data Center (NCDC), website: http://www.ncdc.noaa.gov/oa/ncdc.html
[3]provided by NOAA, see http://www.hpc.ncep.noaa.gov/html/dewrh.shtml

average wind speed is measured in nautical miles per hour, varies from 0.39 knots to 40.06 knots.

Meteorological variables	Minimum	Mean	Median	Maximum
Maximum Temp. ($^{\circ}C$)	-1.30	27.61	28.30	42.28
Relative humidity (%)	2.11	5.74	5.66	13.29
Average wind speed	0.39	5.51	5.18	40.06

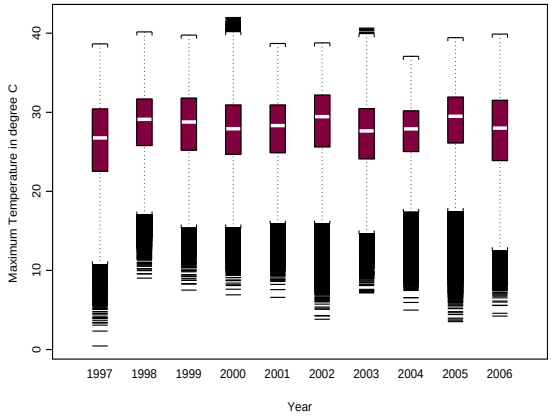
Table 4.3: Summary statistics for daily maximum temperature in $^{\circ}C$, percentage relative humidity and average wind speed in nautical miles in the eastern US.

Table 4.4 shows the correlation between observed ozone levels with the significant meteorological variables used in the modelling. We observe maximum temperature has a positive correlation with the observed maximum 8 hour ozone levels, whereas wind speed and relative humidity show negative correlation with ozone levels.

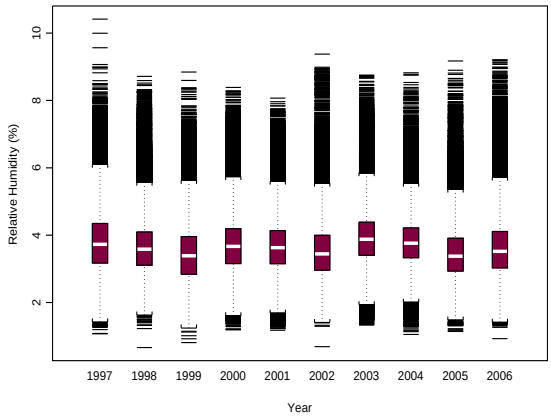
	o8hrmax	TEMP	WDSP	RH
o8hrmax	1.00	0.40	-0.23	-0.54
TEMP	0.40	1.00	-0.27	-0.41
WDSP	-0.23	-0.27	1.00	0.17
RH	-0.54	-0.41	0.17	1.00

Table 4.4: Correlation matrix of daily maximum 8 hour ozone levels and meteorological variables. Here, TEMP is maximum temperature in $^{\circ}C$, WDSP is the average wind speed in nautical miles and RH is the percentage relative humidity in the eastern US.

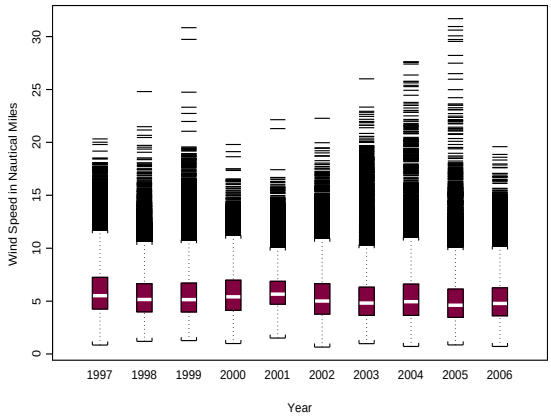
Figure 4.10(a) shows box-plot of the daily maximum temperature trend in the eastern US. Remarkably, the trend in maximum temperature matches closely with the overall trend in ozone levels in Figure 4.4. From Figure 4.10(b) and (c), we observe that the overall trend in average wind speed shows approximately the same pattern compared to the RH. However, there are some dissimilarities too, for example, in year 2003 the wind speed is lower than the years 2002 and 2004. We also observe the variability for the average wind speed is relatively higher compared to temperature and RH.



(a)



(b)



(c)

Figure 4.10: Box-plot of the three meteorological variables by years, (a) daily maximum temperature levels ($^{\circ}C$) (b) relative humidity in percentage and (c) daily average wind speed in nautical miles per hour in the eastern US.

4.5 Some Outliers in the Observed Ozone Concentration Levels

We observe some outliers from the annual 4th highest maximum and 3-year rolling average plots (see Figure 4.7 and 4.8). The monitoring sites for those outlier observations are identified, where all of them are NAMS/SLAMS sites (for NAMS/SLAMS sites see details in Section 1.2). Figure 4.11 represents the 5 monitoring sites superimposed in the eastern US map.

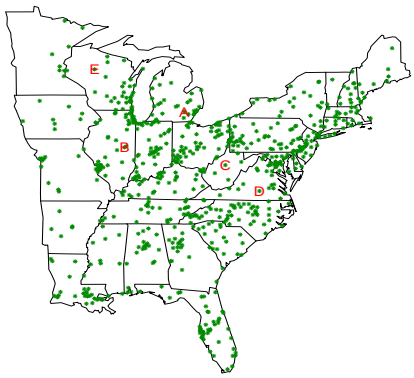


Figure 4.11: Map of the eastern US for the ozone monitoring sites with superimposed outlier observation locations A to E.

There are several factors that can increase ozone levels dramatically, for example effects of meteorological variables, sudden increase of the ozone emission sources etc. We observe major increase in ozone levels in the mid summer seasons, i.e., in the months of June and July. For example, Figure 4.12 shows the time-series plots of ozone levels for months June and July in 2002. We can also observe from the meteorological variables that the standard deviations for maximum temperature, relative humidity and wind speed are 3.0, 0.7, and 1.8 in the month of July 2002, which is relatively higher compared to the average monthly standard deviations 2.3, 0.4, and 1.2 respectively.

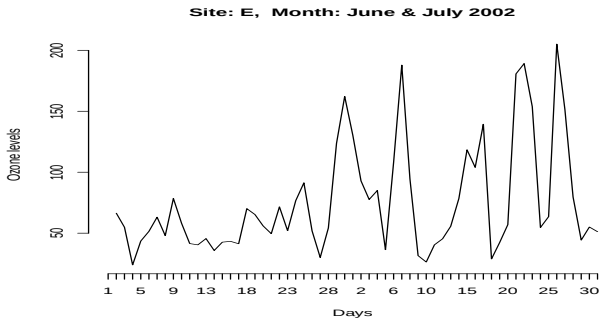


Figure 4.12: Time-series plot of ozone levels for location E, for the months of June and July in 2002.

4.6 Ozone Data for Forecast Models

Alongside the 10 years daily ozone concentration data in the eastern US, we have another set of daily ozone data for three weeks starting from 23 June to 14 July, 2010. We use this data set for forecasting of 7 days ahead that are analysed in Chapter 7. CMAQ output are also available for this time period that are used as covariate in the models.

Figure 4.13 represents the map of the eastern US, where ozone monitoring and CMAQ grid locations are superimposed. In the following sections we will provide some descriptive statistics related to the ozone and CMAQ output.

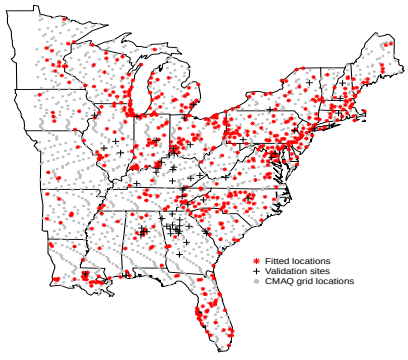


Figure 4.13: Plot of the 639 ozone monitoring sites in the eastern US in 2010. 62 hold-out sites, 577 sites for fitting forecast models, and 1451 CMAQ grid locations are superimposed.

	Minimum	Mean	Median	Maximum
Ozone levels	0.00	50.62	50.99	113.00
CMAQ output	16.50	59.19	60.36	145.50

Table 4.5: Summary statistics for daily ozone levels and CMAQ output in the eastern US.

4.6.1 Descriptive Statistics

In this data set we have 13,419 ozone observations for 21 days in 639 ozone monitoring locations, among them 299 ($\approx 2.23\%$) are missing. As expected we do not have any missing observations for the CMAQ grid output. Table 4.5 provides the summary statistics for ozone levels and CMAQ data, where we observe average level ozone measurement is littile bit higher for the CMAQ output. We use this data set in Chapter 7 to obtain forecast in the future time and for forecast validations.

4.7 Summary

In this chapter we present different data editing techniques and summary statistics of the daily ozone concentration levels. In addition, we describe data obtained from the computer simulated models known as CMAQ. We also describe meteorological data with their summary statistics. In the rest of this thesis we will use the data described in this chapter.

Chapter 5

Model Comparisons

5.1 Introduction

Several approaches have been proposed to model daily ozone concentration levels. Section 1.5 provides a review of these. In this chapter we compare two such approaches: the dynamic linear models (DLM) (Stroud *et al.*, 2001; Huerta *et al.*, 2004) and the hierarchical auto-regressive (AR) models (Sahu *et al.*, 2007).

There are few articles that compare dynamic linear approach with other models for ozone concentration levels. Zheng *et al.* (2007) used DLM and generalised additive models (GAM) to explain trend in ozone levels, however they do not consider the spatial correlations and applied principal component analysis (PCA) to represent regional patterns of ozone concentrations. Their results indicate both methods can easily estimate trends in ozone levels and provide good predictions. They conclude that additive models are attractive when estimates are needed quickly or when many similar but separate site specific analyses are required. In addition, dynamic models are much more flexible, readily addressing such issues as autocorrelation, the presence of missing values, and estimation of long-term trends or cyclical patterns.

Dou *et al.* (2010) compare the space-time version of the DLM with another estimation method, the Bayesian spatial predictor (BSP) (see details in Le and Zidek, 2006) to analyse hourly ozone concentrations. They conclude that BSP works at least as well as the DLM, and requires much less computational power, see Section 1.5.

In this chapter we compare the DLM with the AR modelling strategies by

providing some theoretical results regarding the predictive and forecasting distributions obtained using simplified versions of these models. The simplified models do not consider the covariate effects, but they represent the underlying basic spatio-temporal characteristics of the models.

We also simulate four replicated datasets from both models and compare their performances. As expected, the fitted model performs best when it is also the true simulation model. The models are also applied to a real-life ozone concentration data set obtained from New York for the months of July and August, 2006 (see Section 4.2 for data description).

The remainder of this chapter is organised as follows: in Section 5.2, we briefly describe both models and their simplified versions. Section 5.3 discusses the properties and theoretical results that we have obtained for the models. Section 5.4 provides a simulation study and the real data example on daily maximum eight-hour ozone concentration levels in New York. Finally, a few summary remarks are given in Section 5.5.

5.2 Model Specifications

The DLM and the hierarchical AR models are discussed in Section 3.4. The simplified versions of both the models are described below.

5.2.1 Simplified DLM

Following Dou *et al.* (2010), we consider a simplified version of the DLM where we assume that there are no covariate effects, i.e. $F_t = 1$, in equation (3.8), which corresponds to the model that has a site invariant mean. Consequently, ω_t turns out to be a scalar, and we assume $\omega_t \sim N(0, \sigma_\omega^2)$. The simplified model is given by:

$$Z(s_i, t) = \theta_t + \nu(s_i, t) \tag{5.1}$$

$$\theta_t = \theta_{t-1} + \omega_t \tag{5.2}$$

we assume the initial condition $\theta_0 \sim N(\mu, \sigma_\theta^2)$.

5.2.2 Simplified AR Models

To simplify the AR models, we again consider no effect of the meteorological variables on true ozone levels $O(s_i, t)$. We also assume no global intercepts to further simplify the model. The simplified AR model is written as:

$$Z(s_i,t) = O(s_i,t) + \epsilon(s_i,t), \quad t = 1,2,...,T \tag{5.3}$$

$$O(s_i,t) = \rho O(s_i,t-1) + \eta(s_i,t), \quad t = 1,2,...,T \tag{5.4}$$

with initial condition $O(s_i, 0) \sim N(\mu, \Sigma_0)$, where Σ_0 has elements $\sigma_0^2 \exp(-\phi_0 d_{ij})$, with ϕ_0 as a decay parameter.

5.3 Theoretical Results

For comparison of two modelling strategies: the DLM and the AR models, we consider their simplified versions stated in equations (5.1)-(5.2) and (5.3)-(5.4). We assume that the required components $\sigma_\nu^2, \sigma_\theta^2, \sigma_\omega^2$ and $\sigma_\epsilon^2, \sigma_0^2, \sigma_\eta^2$ of the respective models are known constants, and the autoregressive parameter ρ is also known for the AR model. These are assumed for the purpose of obtaining the theoretical results. In Appendix A we provide the proofs of the theories and required calculations.

5.3.1 Some Properties of the DLM and the AR Models

We provide some properties of the AR models and compare those with the properties of the DLM obtained by Dou *et al.* (2010). Now, the variance-covariance structure (see Dou *et al.*, 2010) of $Z(s_i, t)$ and $Z(s_j, t+k)$ for the DLM is written as:

$$\text{Cov}[Z(s_i,t), Z(s_j,t+k)] = \sigma_\theta^2 + t\sigma_\omega^2 + \sigma_\nu^2 \exp(-\phi d_{ij}) 1(k=0). \tag{5.5}$$

where, $k \geq 0$ is an integer and $1(k = 0)$ is the indicator function for the variance (σ_ν^2) of the spatially correlated error term of the model.

We now obtain a similar result for the AR models as follows. For any positive

integer t the AR models imply that:

$$Z(s_i,t) = \epsilon(s_i,t) + \eta(s_i,t) + \rho\,\eta(s_i,t-1) + ... + \rho^{t-1}\,\eta(s_i,1) + \rho^t\,O(s_i,0),$$

and for any integer $k > 0$:

$$\begin{aligned} Z(s_j,t+k) &= \epsilon(s_j,t+k) + \eta(s_j,t+k) + \rho\,\eta(s_j,t+k-1) + ... + \rho^{k-1}\,\eta(s_j,t+1) \\ &+ \rho^k\,\eta(s_j,t) + \rho^{k+1}\,\eta(s_j,t-1) + ... + \rho^{t+k-1}\,\eta(s_j,1) + \rho^{t+k}\,O(s_j,0). \end{aligned}$$

The assumptions of the AR models state that the spatial errors $\eta(s_i,t)$ and $\eta(s_j,t+k)$ are independent if $k > 0$ and the hierarchical error $\epsilon(s_i,t)$ is independent of the spatial error $\eta(s_j,t)$, and the initial random variable $O(s_i,0)$ is independent of both $\eta(s_i,t)$ and $\epsilon(s_i,t)$. Hence, we have

$$\begin{aligned} \mathrm{Cov}(Z(s_i,t),Z(s_j,t+k)) &= \mathrm{Cov}(\epsilon(s_i,t),\epsilon(s_j,t+k)) + \rho^{2t+k}\,\mathrm{Cov}(O(s_i,0),O(s_j,0)) \\ &+ \rho^k\,\mathrm{Cov}(\eta(s_i,t),\eta(s_j,t)) + \rho^{k+2}\,\mathrm{Cov}(\eta(s_i,t-1),\eta(s_j,t-1)) \\ &+ ... + \rho^{k+2t-2}\,\mathrm{Cov}(\eta(s_i,1),\eta(s_j,1)) \\ &= \rho^{2t+k}\sigma_0^2\exp(-\phi_0 d_{ij}) + \rho^k\tfrac{1-\rho^{2t}}{1-\rho^2}\sigma_\eta^2\exp(-\phi_\eta d_{ij}). \end{aligned}$$

Thus we arrive at the following general covariance function of the observations $Z(s_i,t)$ and $Z(s_j,t+k)$ at locations s_i, s_j, at time t and $t+k$ as:

$$\begin{aligned} \mathrm{Cov}[Z(s_i,t),Z(s_j,t+k)] &= \rho^{2t+k}\sigma_0^2\exp(-\phi_0 d_{ij}) + \rho^k\left[\frac{1-\rho^{2t}}{1-\rho^2}\right]\sigma_\eta^2\exp(-\phi_\eta d_{ij}) \\ &\quad +\sigma_\epsilon^2 1(k=0). \end{aligned} \tag{5.6}$$

where, $k \geq 0$ is an integer and $1(k=0)$ is the indicator function for the nugget effect (σ_ϵ^2) of the model.

These two general covariance functions given in equations (5.5) and (5.6) enable us to study many properties of the two models as discussed in the following two sub-sections. Details of the calculation are given in Appendix A.

5.3.2 Comparison of Correlation Structures

Using the expression for the general covariance function in equation (5.5) Dou *et al.* (2010) obtained the following results:

(i) $\mathrm{Cor}(Z(s_i,t),Z(s_j,t+k))$ for $i \neq j$ attains its maximum at $k = 0$ and

decreases as k increases. This can be a reasonable property since the correlation between observations at different locations can be expected to be the maximum at the current time because both of those locations may be influenced similarly by the prevailing meteorological and other conditions, e.g., power station emission volumes, affecting ozone production. The correlation should decrease at different times due to possible mismatches in the meteorological conditions at different times.

(ii) $\text{Cor}(Z(s_i,t), Z(s_j,t)) \to 1$ as $t \to \infty$ for $i \neq j$. This seems to be an unreasonable property. The correlation between any two fixed monitors should not increase with time.

(iii) $\text{Cor}(Z(s_i,t), Z(s_j,t)) \to 1$ as $d_{ij} \to 0$ for $i \neq j$. This is a reasonable property since the observations at two locations close to each other should be very similar.

(iv) $\text{Cor}(Z(s_i,t), Z(s_j,t)) \to \frac{\sigma_\theta^2 + t\sigma_\omega^2}{\sigma_\theta^2 + t\sigma_\omega^2 + \sigma_\nu^2}$ as $d_{ij} \to \infty$ for $i \neq j$. Ideally, this limit should be close to 0 since the observations at two far away locations should tend to be independent of each other. In order to achieve this ideal limit, Dou *et al.* (2010) suggested replacing σ_ω^2 by σ_ω^2/T and taking σ_θ^2 much smaller than σ_ν^2.

Similar properties of the AR models can be derived using the general covariance function obtained in equation (5.6).

(i) As in the DLM case, $\text{Cor}(Z(s_i,t), Z(s_j,t+k))$ for $i \neq j$ decreases as k increases.

(ii) $\text{Cor}(Z(s_i,t), Z(s_j,t)) \to \frac{\sigma_\eta^2 \exp(-\phi d_{ij})}{\sigma_\epsilon^2(1-\rho^2) + \sigma_\eta^2}$ as $t \to \infty$ for $i \neq j$, where ϕ. Unlike the case for the DLM, this correlation does not approach 1.

(iii) $\text{Cor}(Z(s_i,t), Z(s_j,t)) \to 1$ as $d_{ij} \to 0$ for $i \neq j$. This is a reasonable property as in the case for the DLM.

(iv) $\text{Cor}(Z(s_i,t), Z(s_j,t)) \to 0$ as $d_{ij} \to \infty$ for $i \neq j$. Unlike the case for the DLM, here the ideal limit is reached without any further condition or model adjustments.

5.3.3 Comparison of Variance Inequalities for Predictions

The differences in covariance structure imply very different behaviour in model based predictions and forecasting. In this section we investigate the prediction variances by examining five important inequalities capturing various possibilities for predictions. We compare the results for the AR models with those for the DLM obtained by Dou *et al.* (2010).

For simplicity we consider prediction at an unmonitored site s_0 given the observations at a monitored site s_1. We assume that all the parameters, ρ, ϕ, σ_θ^2, σ_ν^2, σ_ω^2, σ_0^2, σ_η^2, σ_ϵ^2 are known. Hence the conditional variance of $Z(s_0,t)$ given $Z(s_1,t')$ for any t and t' will be the predictive variance in the Bayesian setting since there is no need to integrate over any unknown parameters to obtain the predictive distributions. The comparisons performed in the simulation study and the real data example in the next section do not make these assumptions.

For the simplified versions of the DLM and the AR models in equations (5.1)-(5.2) and (5.3)-(5.4) respectively, with $n = 1$ and $t = 1, 2$, the joint distribution of observations $(z(s_0,1), z(s_0,2), z(s_1,1), z(s_2,2))'$ can be written as: $N(\mathbf{0}, \Sigma)$, where Σ is the variance-covariance matrix, obtained from equation (5.5) for the DLM and from equation (5.6) for the AR models.

By simple calculations we obtain the following conditional variances for the AR models:

$$\begin{aligned}
\mathrm{Var}(Z(s_0,1)|Z(s_1,1)) &= \sigma_\epsilon^2 + \rho^2\sigma_0^2 + \sigma_\eta^2 - \zeta^2 \tfrac{(\sigma_0^2\rho^2+\sigma_\eta^2)^2}{\sigma_\epsilon^2+\rho^2\sigma_0^2+\sigma_\eta^2} \\
\mathrm{Var}(Z(s_0,2)|Z(s_1,2)) &= \sigma_\epsilon^2 + \rho^4\sigma_0^2 + (1+\rho^2)\sigma_\eta^2 - \zeta^2 \tfrac{\{\rho^4\sigma_0^2+(1+\rho^2)\sigma_\eta^2\}^2}{\sigma_\epsilon^2+\rho^4\sigma_0^2+(1+\rho^2)\sigma_\eta^2},
\end{aligned}$$

where $\zeta = \exp(-\phi d_{01})$ denotes the spatial correlation between the observations at the two sites at any given time. The general covariance function (5.6) also allows us to calculate the conditional variances $\mathrm{Var}(Z(s_0,1)|Z(s_1,1), Z(s_1,2))$ and $\mathrm{Var}(Z(s_0,2)|Z(s_1,1), Z(s_1,2))$; the expressions for these are long and given in Appendix A.

- Now, we obtain the following results involving the conditional variances for the AR models:

$$\begin{aligned}
\mathrm{Var}(Z(s_0,1)|Z(s_1,1)) - \mathrm{Var}(Z(s_0,1)|Z(s_1,1), Z(s_1,2)) &= \tfrac{N_1}{\Delta_1(\sigma_\epsilon^2+\rho^2\sigma_0^2+\sigma_\eta^2)} \\
\mathrm{Var}(Z(s_0,2)|Z(s_1,2)) - \mathrm{Var}(Z(s_0,1)|Z(s_1,1), Z(s_1,2)) &= \tfrac{N_1}{\Delta_1(\sigma_\epsilon^2+\rho^4\sigma_0^2+(1+\rho^2)\sigma_\eta^2)},
\end{aligned}$$

where

$$N_1 = \zeta^2\rho^2\sigma_\epsilon^4(\rho^2\sigma_0^2 + \sigma_\eta^2)^2$$

and

$$\Delta_1 = \sigma_\epsilon^4 + \sigma_\eta^2(\rho^2\sigma_0^2 + \sigma_\eta^2) + \sigma_\epsilon^2\{\rho^2(1+\rho^2)\sigma_0^2 + (2+\rho^2)\sigma_\eta^2\}.$$

Thus the above two differences in variances are always non-negative. These two variance inequalities ascertain that the variance of the spatial prediction at site s_0 using data from both time points will always be smaller than that when the spatial prediction is done using data from only one time point. Dou *et al.* (2010) prove the exact same results for the DLM as:

$$\begin{aligned}
&\mathrm{Var}(Z(s_0,1)|Z(s_1,1)) - \mathrm{Var}(Z(s_0,1)|Z(s_1,1),Z(s_1,2)) = \tfrac{N_2}{\Delta_2(\sigma_\theta^2+\sigma_\omega^2+\sigma_\nu^2)} \\
&\mathrm{Var}(Z(s_0,2)|Z(s_1,2)) - \mathrm{Var}(Z(s_0,1)|Z(s_1,1),Z(s_1,2)) = \tfrac{N_2}{\Delta_2(\sigma_\theta^2+2\sigma_\omega^2+\sigma_\nu^2)},
\end{aligned}$$

where

$$N_2 = \sigma_\nu^4(\sigma_\theta^2 + \sigma_\omega^2)^2(1-\zeta)^2$$

and

$$\Delta_2 = (\sigma_\theta^2 + \sigma_\omega^2 + \sigma_\nu^2)(\sigma_\theta^2 + 2\sigma_\omega^2 + \sigma_\nu^2) - (\sigma_\theta^2 + \sigma_\omega^2)^2.$$

A striking difference between the two models lies in the expression for the factor in the numerator. Observe that both differences have a factor ζ^2 in the numerator which implies that the differences increase as the spatial correlation ζ increases. Intuitively, this is a very desirable property since spatial prediction should become more accurate as the spatial correlation increases. However, the same conclusion cannot be reached for the DLM since the variance differences involves the spatial correlation ζ only through a factor $(1-\zeta)^2$ in the numerator. This seems to be an undesirable property of the DLM.

- Dou *et al.* (2010) prove that, for the DLM, conditioned on the same amount of data, the predictive variance of $Z(s_0,1)$ would be no greater than that of $Z(s_0,2)$, that is,

$$\mathrm{Var}(Z(s_0,1)|Z(s_1,1),Z(s_1,2)) \leq \mathrm{Var}(Z(s_0,2)|Z(s_1,1),Z(s_1,2)).$$

The same inequality holds for the AR models only under the condition

$$\kappa \equiv \frac{\sigma_\eta^2}{\sigma_0^2} \geq 1 - \rho^2. \tag{5.7}$$

Note that this always holds if we set $\rho = 1$ as in the DLM case. For other values of ρ, this condition implies that the ratio of the process and the initial variance, κ must be bounded below by $1 - \rho^2$. This condition holds if we set σ_0^2 to be the limiting variance of η_t given by $\sigma_\eta^2/(1-\rho^2)$ as $t > \infty$.

However, this is a troublesome property as the conditional variance increases by time. Hence, under the condition in equation (5.7), the AR model can perform better than the DLM.

- All four conditional variances discussed so far for both the models can be proved to be monotonically decreasing function of spatial correlation ζ, or equivalently, increasing function of the distance, d_{01} between the data site, s_1 and the predictions site, s_0.

- In the time series modelling framework, it is worthwhile to investigate whether or not it is possible to make more accurate spatial prediction by conditioning on additional temporal data. That is, whether inequalities such as

$$\text{Var}(Z(s_0, 2)|Z(s_1, 2)) > \text{Var}(Z(s_0, 2)|Z(s_1, 1), Z(s_1, 2)), \tag{5.8}$$

can be expected to hold. The above inequality, however, is always true due to the fact that the conditional variance decreases as the number of conditioning random variables increases in a nested fashion.

A slight re-formulation of the above question is often more useful in practical modelling. Would the inequality (5.8) hold if for the prediction problem in the left hand side we ignore the data at time $t = 1$ completely and apply the model at time $t = 2$ for the first time? In this case, $\text{Var}(Z(s_0, 2)|Z(s_1, 2))$ when the model is applied for the first time at $t = 2$ will be exactly the same as $\text{Var}(Z(s_0, 1)|Z(s_1, 1))$. Hence, we need to in-

vestigate what conditions will guarantee the inequality

$$\mathrm{Var}(Z(s_0,1)|Z(s_1,1)) - \mathrm{Var}(Z(s_0,2)|Z(s_1,1),Z(s_1,2)) > 0. \tag{5.9}$$

For the DLM, Dou *et al.* (2010) show that (5.9) holds if and only if

$$\frac{\sigma_w^2}{\sigma_\theta^2} < \frac{\kappa+1}{\kappa}, \tag{5.10}$$

where $\kappa = \frac{\sigma_\eta^2}{\sigma_0^2} \equiv \frac{\sigma_\nu^2}{\sigma_\theta^2}$ under the DLM. Note that this condition (5.10) is free of the spatial correlation parameter ζ. We now investigate the conditions under which (5.9) holds for the AR models.

The analysis for the AR models is more complicated due to the presence of the extra temporal correlation parameter ρ. We consider the following special and limiting cases. Straightforward calculations yield that the variance difference in equation (5.9) is negative if $\sigma_0^2 = 0$. In addition it goes to ∞ as $\sigma_0^2 \to \infty$; hence, large values of σ_0^2 will guarantee that (5.9) holds.

Now it is interesting to investigate what happens if σ_0^2 takes any other value. We can prove that equation (5.9) holds if

$$\frac{\sigma_\epsilon^2}{\sigma_0^2} < \frac{\kappa+\rho^2}{\kappa-(1-\rho^2)},$$

when ζ approaches 1 (i.e. for large spatial correlation). Observe that for $\rho = 1$ the above condition reduces to the one for the DLM case (5.10) if $\sigma_\theta^2 \equiv \sigma_0^2$ and $\sigma_w^2 \equiv \sigma_\epsilon^2$. Note that $\frac{\kappa+\rho^2}{\kappa-(1-\rho^2)} \geq \frac{\kappa+1}{\kappa}$ always for any value of $0 < \rho^2 < 1$. This implies that the inequality (5.9) holds for a wider range of parameter values under the AR models than the DLM. We can also prove that, when $\zeta \to 0$ the inequality (5.9) holds if in addition we have $\sigma_0^2 > \sigma_\eta^2/(1-\rho^2)$.

5.3.4 Comparison of Variance Inequalities for Forecasts

In this section we provide some properties of the models for conditional variances for forecasting at an unmonitored site s_0. Similar to the previous section we assume all parameters of the models are known. For the simplified version of the DLM and the AR models in equations (5.1)-(5.2) and (5.3)-(5.4) respectively,

with $n = 1$ and $t = 1, 2$ and forecast at time $t = 3$, the joint distribution of observations $(z(s_0, 3), z(s_0, 2), z(s_1, 1), z(s_1, 2), z(s_2, 1), z(s_2, 2))'$ can be written as: $N(\mathbf{0}, \Sigma)$, where Σ is the variance-covariance matrix, obtained from equation (5.5) for the DLM, and from equation (5.6) for the AR models.

- The conditional variances of forecast for the DLM can be obtained as:

$$\text{Var}(Z(s_0, 3)|Z(s_1, 2)) = \sigma_\theta^2 + \sigma_\nu^2 + 3\sigma_\omega^2 - \frac{(\sigma_\theta^2 + 2\sigma_\omega^2)^2}{\sigma_\theta^2 + \sigma_\nu^2 + 2\sigma_\omega^2}.$$

  For the AR models we can write the same conditional variances as:

$$\text{Var}(Z(s_0, 3)|Z(s_1, 2)) = \rho^6\sigma_0^2 + (1+\rho^2+\rho^4)\sigma_\eta^2 + \sigma_\epsilon^2 - \zeta^2 \frac{(\rho^5\sigma_0^2 + \rho\sigma_\eta^2 + \rho^3\sigma_\eta^2)^2}{\rho^4\sigma_0^2 + \sigma_\eta^2 + \rho^2\sigma_\eta^2 + \sigma_\epsilon^2}.$$

  From the forecast variances of the DLM we can see that it does not depend on the spatial correlation between sites s_0 and s_1. However, for the AR models, the increase in spatial correlation (i.e., $\zeta \to 1$) yields less forecast variability, a desirable property.

- Similar to the prediction inequalities in equations (5.8) and (5.9), we can write the forecast conditional variance at time $t = 3$ given data at time $t = 2$ is equal to the conditional variance for forecasting based on only one time point, i.e., at time $t = 2$ given data at site s_1 as:

$$\text{Var}(Z(s_0, 3)|Z(s_1, 2)) = \text{Var}(Z(s_0, 2)|Z(s_1, 1)).$$

  We can obtain the following inequality similar to equation (5.9):

$$\text{Var}(Z(s_0, 2)|Z(s_1, 1)) - \text{Var}(Z(s_0, 3)|Z(s_1, 1), Z(s_1, 2)) > 0. \qquad (5.11)$$

  For the AR models the inequality in equation (5.11) holds as $\sigma_0^2 \to \infty$, henceforth large values of σ_0^2 will guarantee the inequality holds.

  Other results for the forecast conditional variances are also similar to the conditional variances of predictions discussed in Section 5.3.3.

- For both models the forecast variances with less amount of temporal observations is greater than the forecast variance with more temporal obser-

vations. Hence, for both models we can write the inequality:

$$\text{Var}(Z(s_0,3)|Z(s_1,1),Z(s_1,2)) \leq \text{Var}(Z(s_0,3)|Z(s_1,2)). \tag{5.12}$$

Again for more spatial observations, we can write the inequality:

$$\text{Var}(Z(s_0,3)|Z(s_1,1),Z(s_2,2)) \leq \text{Var}(Z(s_0,3)|Z(s_1,2)). \tag{5.13}$$

Hence, the conditional variance of forecasts decreases for the increase in both temporal and spatial observations.

In summary, the AR models are likely to have better properties if the initial variance σ_0^2 is large compared to the process variance σ_η^2. In practical examples where the models are more complex and parameters are unknown, we will not be able to verify the conditions required for the theoretical results, and we must, therefore, rely on empirical evidence. This is where various Bayesian and non-Bayesian model choice criteria can be used to perform model choice. The following section discusses this with several simulation and a real data example.

5.4 Examples

In this section we compare the DLM and AR models in practical data modelling situations where these are often implemented. We use the DLM and the AR models that are stated in Section 3.4. Unlike Hureta *et al.* (2004) and Duo *et al.* (2010), we do not include any seasonal term in the models for daily ozone data, because seasonal terms are more relevant for modelling the diurnal cyclic components often present in the hourly ozone data.

We consider modelling daily eight-hour maximum ozone concentration data from the 29 ozone monitoring sites in the state of New York for 62 days in the months of July and August in 2006. We shall use data from 25 sites for model fitting and the data from the remaining 4 sites will be used for model validation purposes. The state of New York is considered as the spatial domain because the ozone monitoring network in this state represents typical practical situations: a cluster of few sites in and around a big city (the city of New York here) and a moderate number of other sites, situated large distances apart, covering a vast

region; see Figure 5.1 for a map of New York and the location of the monitoring sites. The data from 62 ($= T$) days in July and August are modelled since these are in the high ozone season in the USA. The spatio-temporal domain considered here represents a moderate computational problem where we can implement the models and obtain results using a reasonable amount of computing time and effort.

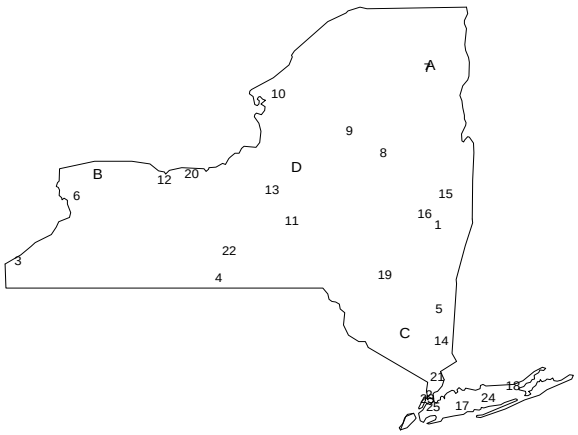


Figure 5.1: A map of the 29 ozone monitoring sites in the state of New York. Four randomly chosen sites labelled A,B,C and D are used for validation purposes and the remaining 25 sites (numbered 1 to 25) are used for modelling.

In the practical modelling of this section, following Sahu *et al.* (2009), as the single covariate we include the output of a computer simulation model known as the CMAQ model. Details of CMAQ modelling are given in Section 1.4 of Chapter 1. In both the DLM and the AR models we use the daily maximum eight-hour CMAQ ozone concentration output for the grid cell covering the monitoring site as the covariate. The spatial predictions at the unmonitored sites are performed using the CMAQ output at the corresponding grid cells. In our models we have also included other meteorological covariates (see Section 4.4) such as the daily maximum temperature, but none of those turn out to be significant in the presence of the CMAQ output. Figure 5.2 shows a strong linear relationship between ozone concentration values and the corresponding CMAQ output.

The full Bayesian model is completed by specifying prior distributions for all

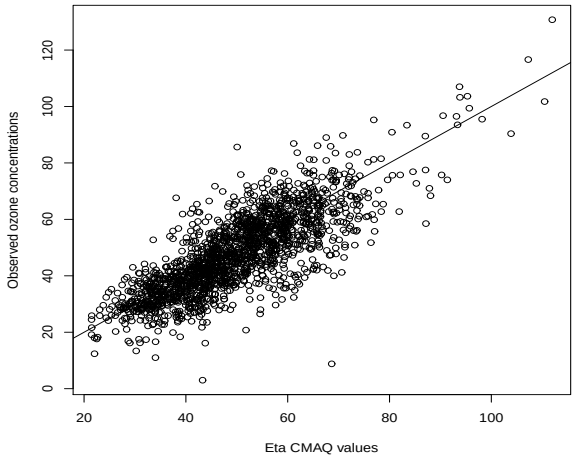


Figure 5.2: A scatter plot of daily maximum eight-hour average ozone concentration levels (ppb) against the CMAQ output (ppb) for the grid cells covering that monitoring sites from 25 sites in New York for 62 days in July and August 2006.

the unknown parameters. We work with the inverse of the variance components $\sigma^2_\epsilon, \sigma^2_\eta$, σ^2_0, σ^2_ν, σ^2_θ and σ^2_ω and assume an independent gamma prior distribution with parameters a and b having mean a/b for each of $1/\sigma^2_\epsilon$, $1/\sigma^2_\eta$, $1/\sigma^2_0$, $1/\sigma^2_\nu$, $1/\sigma^2_\theta$ and $1/\sigma^2_\omega$. In our implementation we take $a = 2$ and $b = 1$ implying that these variance components have prior mean 1 and infinite variance. We assign a flat prior $N(0, 10^4)$ for the regression co-efficient β. Following Sahu *et al.* (2009) we use an empirical Bayes method to estimate the value of the spatial correlation decay parameters ϕ_ν, ϕ_η and ϕ_0 since these parameters are often difficult to estimate from a joint Bayesian model, see Sahu *et al.* (2009) for more in this regard. We also use random-walk metropolis sampling scheme for the spatial decay parameter considering common ϕ. Both methods are used to obtain results as detailed below.

The fully specified Bayesian DLM and AR models cannot be compared using exact analytic methods as done in Section 5.3. Hence we use the PMCC model selection criteria (see Section 3.3) to compare the models. To assess the quality of the predictions we use validation criterion discussed in Section 3.3.4. The conclusions regarding the model choice and comparison turned to be the same as the ones reported below using PMCC and RMSE. For model fitting and predictions, we use the `R` package `spTimer`, that is developed as a part of this research (for

further details see Chapter 8).

5.4.1 A Simulation Example

We first provide a simulation example where we test out the two model choice criteria and the MCMC code we developed for fitting the two sets of models. We simulate four data sets from each of the DLM and AR models. Each data set consists of observations from 29 monitoring sites and 62 days in July and August, 2006. Note that the simulation model includes the CMAQ output as the single covariate. As mentioned above, data from 25 sites will be used for model fitting and the data from the remaining 4 sites will be used for model validation purposes. For both models we set the common value of ϕ at 0.01 for both simulation and fitting. The choice of the simulation model parameters is guided by the practical example provided in Section 5.4.3. For the AR simulation models we set $\rho = 0.2$, $\sigma^2_\epsilon = 0.04$, $\sigma^2_\eta = 0.6$, $\sigma^2_0 = 0.2$, $\mu = 8.0$, $\xi = 1.0$, and $\beta = 0.6$. For the simulation from the DLM we assume: $\sigma^2_\nu = 0.5$, $\Sigma_\omega = 0.06I$, $\Sigma_\theta = 0.2I$, and $\boldsymbol{\mu} = (1.0, 0.6)'$.

We implement the Gibbs sampler for each of the DLM and AR models where we keep the value of ϕ fixed at the simulation value; see the real data example below on how to choose this in practice. We note that the MCMC chains converge rapidly for both the models. 15000 iterates are used for making inference after discarding the first 5000 iterations. We also use multiple parallel runs and calculated the Gelman and Rubin statistics (Gelman and Rubin 1992), which we found to be satisfactory.

5.4.2 Results for the Simulation Example

Table 5.1 presents the values of the PMCC and RMSE for the two models fitted to four replicated simulation data sets from each of the two models. As expected, we see that both the model choice criteria pick the true simulation model. Note also that when data are simulated from the DLM the performance of the incorrectly fitted AR models is not too far away from the DLM. However, when the data are simulated from the AR models the performance of the incorrectly fitted DLM is some distance away from the AR models. Thus the AR models provide reasonably good performance even when data are simulated from the DLM.

	Simulation Model							
	AR				DLM			
	Fitted Model							
	AR		DLM		AR		DLM	
Data Set	PMCC	RMSE	PMCC	RMSE	PMCC	RMSE	PMCC	RMSE
1	831.35	3.36	1223.93	4.03	855.05	3.47	784.87	3.30
2	824.47	3.29	1201.52	4.00	894.21	3.62	797.61	3.52
3	865.15	3.47	1325.71	4.30	847.44	3.40	751.02	3.17
4	852.91	3.42	1311.24	4.23	841.67	3.38	745.42	3.14

Table 5.1: PMCC & RMSE for the DLM and AR models where each model has been fitted to four replicated simulation data sets.

Some other validation criteria are given in Table 5.2, where all those criteria pick the correct simulation model in each case. Figures 5.3 and 5.4 provide the prediction plots for both models for a validation site. It is observed that the AR model has smaller range of 95% prediction intervals compared to the DLM in both cases. We now proceed to the real data example.

AR data	Methods	MAE	rBIAS	rMSEP
1	AR	2.71	-0.024	0.14
	DLM	2.93	-0.005	0.18
2	AR	2.56	-0.009	0.08
	DLM	3.05	-0.005	0.09
3	AR	2.95	0.012	0.11
	DLM	3.21	-0.021	0.14
4	AR	2.89	0.009	0.15
	DLM	3.12	-0.001	0.17
DLM data				
1	AR	2.81	0.022	0.18
	DLM	2.32	-0.011	0.04
2	AR	2.92	0.003	0.23
	DLM	2.44	-0.003	0.04
3	AR	2.75	0.022	0.23
	DLM	2.37	-0.011	0.06
4	AR	2.31	0.013	0.21
	DLM	2.04	-0.009	0.05

Table 5.2: MAE, rBIAS & rMSEP for the DLM and AR models where each model has been fitted to four replicated simulation data sets.

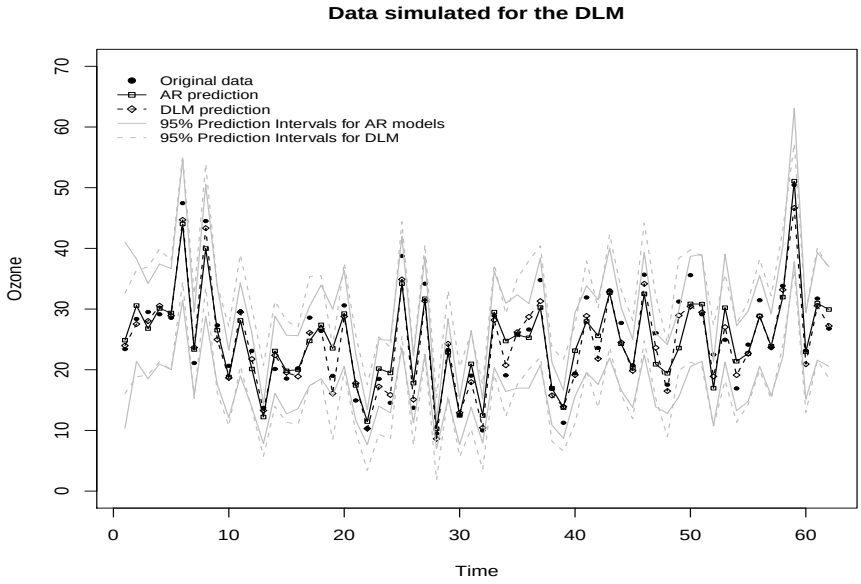


Figure 5.3: DLM and AR predictions at a site for the dataset generated from the DLM. The 95% prediction intervals obtained from both models are also superimposed.

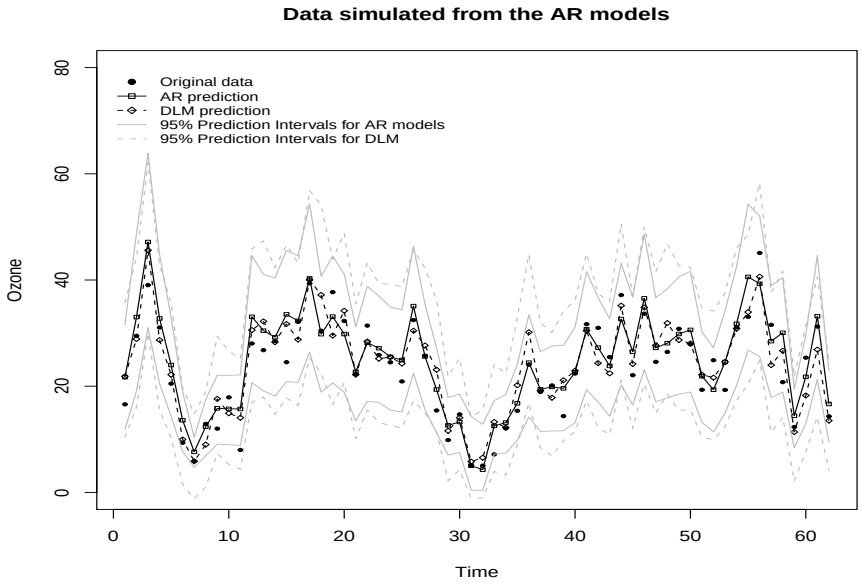


Figure 5.4: DLM and AR predictions at a site for the dataset generated from the AR models. The 95% prediction intervals obtained from both models are also superimposed.

5.4.3 The New York Data Example

We analyse the New York data set obtained from 29 monitoring sites for 62 days in July and August in 2006. Out of these 1798 observations 80 (4.45%) were found to be missing which we assume to be at random. In our Bayesian inference setup using MCMC we simply treat these missing values as unknown parameters and simulate from their full conditional distributions at each MCMC iteration.

As mentioned previously, we use data from 25 sites for model fitting and the data from the remaining four sites (labelled A-D in Figure 5.1) are used for validation. For covariate effect we use the output obtained from the CMAQ models. We consider those observations of CMAQ grid locations that are closest to the sites of the ozone observation (for details see Section 1.4).

Boxplot of the data from the 25 monitoring sites are provided in Figure 5.5. The plot shows moderately high level (more than 50 ppb) of ozone concentration values for most days. There is no apparent strong overall trend, although it seems that there is a slight decreasing trend during the last two weeks in August.

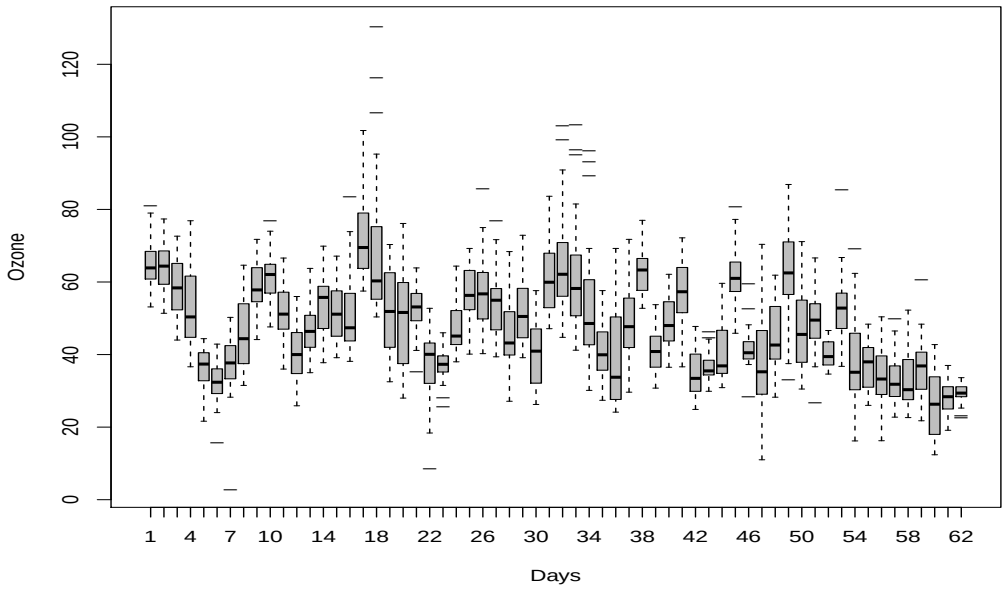


Figure 5.5: Boxplot of the daily maximum 8-hour average ozone concentration levels from 25 monitoring sites in New York for 62 Days in July and August 2006.

5.4.4 Sensitivity of the Prior Distributions

In this section we check the sensitivity of the prior distributions for the variance parameters. For both models we consider the variance parameters (i.e., σ^2_ϵ, σ^2_η and σ^2_0 for the AR models; σ^2_ν, σ^2_ω and σ^2_θ for the DLM) to follow inverse Gamma distribution with hyper-parameters a and b. To obtain a proper prior we used $a = 2$ and $b = 1$, and furthermore we fix ϕ_η, ϕ_0 and ϕ_ν at 0.01 based on the results in Section 5.4.5. Here, to find the sensitivity of the prior distributions we change the values of the parameters a and b. We use the RMSE as the indicator to compare between different prior distributions.

Table 5.3 provides some RMSE results for choosing different prior specifications. We can see for small changes in the hyper-prior parameters yields no change in the RMSE. However, for a higher value of the hyper-parameter change the validation result a lot. In the later case prior is very informative and RMSE is very large due to the very tight constraints implied by the informative prior.

Changes in a			Changes in b		
Hyper-prior	AR	DLM	Hyper-prior pair	AR	DLM
(2, 1)	6.92	8.62	(2, 2)	6.92	8.63
(3, 1)	6.93	8.63	(2, 3)	6.93	8.63
(4, 1)	6.93	8.65	(2, 4)	6.93	8.64
(1000, 1)	7.26	9.05	(2, 1000)	7.28	9.13

Table 5.3: RMSE for the DLM and AR models under different hyper-prior specifications.

5.4.5 Empirical Bayes Method for Choice of the Spatial Decay

We first use the validation data set to choose the spatial decay parameters ϕ_ν for the DLM and ϕ_η and ϕ_0 for the AR models. For each of these we consider the set of possible values: 0.05, 0.01, 0.005, 0.001 and choose the combination which provides the least value of the RMSE for the New York Data set. Tables 5.4 and 5.5 provide the RMSE values for different values of the decay parameters for the DLM and AR models, respectively. For the DLM, using Table 5.4 we choose ϕ_ν to be 0.01 and for the AR models using Table 5.5 shows that the optimal value of both the decay parameters ϕ_η and ϕ_0 is 0.01. Note that this value of the spatial decay parameter corresponds to a spatial range of about 300 kilometres,

i.e. spatial correlation becomes negligible after 300 kilometres.

ϕ_ν	0.050	0.010	0.005	0.001
RMSE	8.86	**8.57**	8.93	8.84

Table 5.4: RMSE values for the DLM for different values of ϕ_ν.

		ϕ_η			
		0.050	0.010	0.005	0.001
ϕ_0	0.050	7.08	6.98	7.69	8.75
	0.010	7.03	**6.92**	7.61	8.66
	0.005	7.05	6.96	7.62	8.67
	0.001	7.02	6.95	7.61	8.64

Table 5.5: RMSE values for the AR for different values of ϕ_η and ϕ_0.

The optimal values of the RMSE for the selected DLM and AR models are 8.57 and 6.92, respectively. This shows that the AR models perform much better in model validation than the DLM. In fact, these overall RMSE's are averages of the RMSE for each of the four validation sites, see Table 5.6. The RMSE for site D is highest since this is the farthest validation site from its nearest data site, see Figure 5.1. This table shows that the AR models outperform the DLM in all four validation sites. Moreover, the values of the PMCC criterion for the selected DLM and AR models are 847.9 and 1360.6, respectively. This also confirms that the AR models are better suited for this particular data set.

	A	B	C	D	Overall
AR	6.19	6.67	7.01	7.81	6.92
DLM	7.15	8.17	9.22	9.73	8.57

Table 5.6: RMSE values for the selected DLM and AR models for the overall and four validation sites.

Figure 5.6 shows the trace plots for DLM parameters σ^2_ν, σ^2_ω and σ^2_θ, and Table 5.7 provides the posterior summary statistics of the parameters σ^2_ν, σ^2_ω and σ^2_θ.

Table 5.8 provides the parameter estimates for the adopted AR models. It shows that the CMAQ output is a significant predictor since β is significant. The temporal correlation parameter ρ is also estimated to be significant. The estimate of the variance components show that the initial variance, σ^2_0 is much larger than

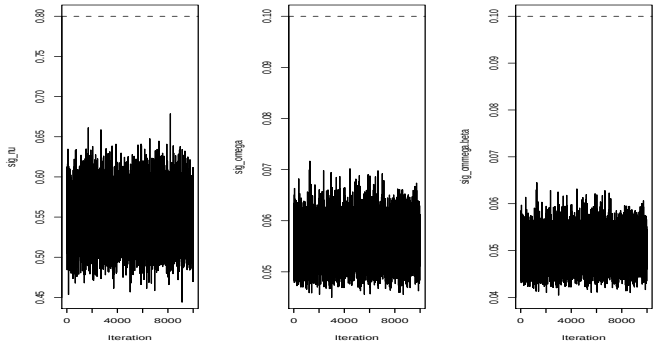


Figure 5.6: MCMC trace plots for the parameters σ_ν^2, σ_ω^2 and σ_θ^2 of the DLM for the New York data. The dashed line represents the initial values for the corresponding parameter.

Measurements	σ_ν^2	σ_ω^2	σ_θ^2
2.5%	0.495	0.049	0.044
Mean	0.558	0.056	0.050
Median	0.547	0.056	0.050
97.5%	0.607	0.063	0.057

Table 5.7: Summary statistics of the posterior distributions for the parameters σ_ν^2, σ_ω^2 and σ_θ^2.

the other two variance components. Thus the theoretical results discussed in Section 5.3 that required a large value of σ_0^2 are likely to hold here.

Model	Measurements	ρ	ξ	β	μ	σ_ϵ^2	σ_η^2	σ_0^2
AR	2.5%	0.18	0.52	0.55	7.91	0.036	0.5194	1.03
	Mean	0.23	1.07	0.62	8.04	0.039	0.5600	1.75
	Median	0.23	1.07	0.62	8.04	0.038	0.5600	1.68
	97.5%	0.27	1.58	0.68	8.16	0.041	0.6042	2.94

Table 5.8: Summary statistics of the posterior distributions for the parameters ρ, ξ, β, μ, σ_ϵ^2, σ_η^2 and σ_0^2 for the AR models.

The MCMC trace plots of AR model parameters σ_ϵ^2, σ_η^2, μ, ρ, ξ and β are given in Figure 5.7 and these indicate quick convergence.

5.4.6 Metropolis-Hastings Sampling for the Spatial Decay

In Section 5.4.5 we use empirical Bayes approach to estimate the spatial decay parameter for the models. In this section we consider a common spatial decay parameter, i.e., ϕ for the models. Henceforth, we use the Metropolis sampling algorithm to draw samples for ϕ. With appropriate tuning we obtain the acceptance

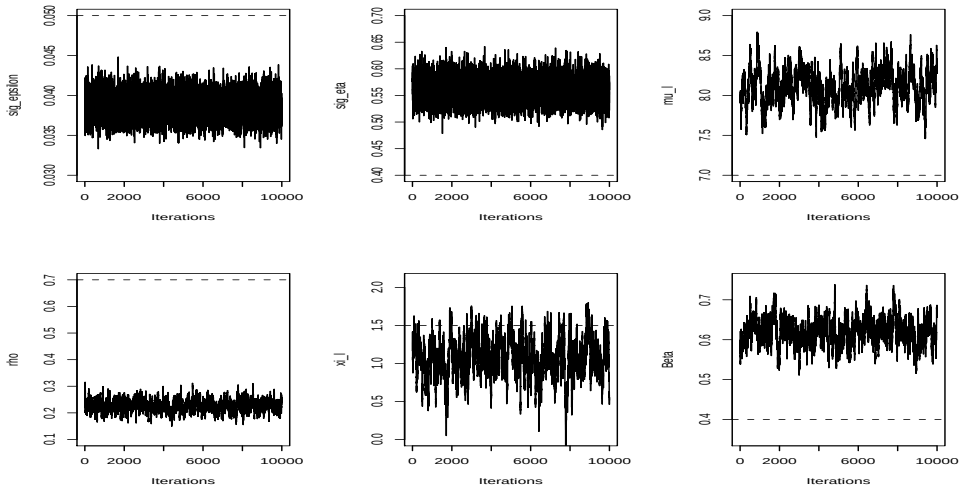


Figure 5.7: MCMC trace plots for the parameters σ_ϵ^2, σ_η^2, μ, ρ, ξ and β of the AR models fitted to the New York data set. The dashed line represents the initial values for the corresponding parameter.

rate for ϕ as 30.07% and 34.2% for the DLM and the AR models respectively. The estimates of ϕ are 0.011 for the DLM and 0.012 for the AR models, and are statistically significant. We also observe the PMCC value is 735.22 for the AR models that is lower than the PMCC value 1066.04 of the DLM.

Table 5.9 provides the parameter estimates for the AR model adopted using Metropolis algorithm of the ϕ parameter. It shows that the CMAQ output is a significant predictor since β is significant. The temporal correlation parameter ρ is also estimated to be significant. The spatial decay parameter is estimated to be 0.012. The estimates of the variance components show that on average, the initial variance, σ_0^2, is much larger than the process variance, σ_η^2; hence the theoretical results which required a large initial variance will hold.

	Mean	95% interval
μ	8.431	(7.582, 8.991)
ξ	1.226	(0.793, 1.811)
ρ	0.198	(0.157, 0.235)
β	0.669	(0.581, 0.734)
σ_ϵ^2	0.048	(0.037, 0.065)
σ_η^2	0.255	(0.198, 0.377)
σ_0^2	0.689	(0.592, 0.768)
ϕ	0.012	(0.009, 0.016)

Table 5.9: Parameter estimates of the selected AR model.

The RMSEs are given in Table 5.10, and we observe the AR model performs better compared to the DLM. We also get better results for the random walk Metropolis sampling than the empirical Bayes approach (see Table 5.6). Henceforth, in the following Chapters we use the Metropolis sampling scheme to sample ϕ parameter.

	A	B	C	D	Overall
AR	5.86	6.96	6.93	7.34	6.77
DLM	7.16	7.72	7.56	8.10	7.64

Table 5.10: RMSE values for the selected DLM and AR models for the overall and the four validation sites using the random walk Metropolis sampling.

5.4.7 Forecasts

We have also performed one step ahead forecasts using the DLM, and the AR models. Table 5.11 provides the RMSEs for the one step ahead forecast. The RMSE values obtained for the CMAQ observations are also presented. It is clearly observed that the AR models give smaller MSE compared to the other methods. We discuss forecasting further in Chapter 7, where different modelling strategies are compared.

DLM	AR models	CMAQ
9.61	8.65	11.85

Table 5.11: RMSE for seven day forecast using the DLM, the AR models, and the CMAQ values for the New York data set.

All these provide additional justifications for choosing the AR models for modelling the daily ozone data considered here.

5.5 Conclusions

In this Chapter, we compare the DLM and the AR modelling approaches to analyse the ozone concentration levels observed in New York in July-August, 2006. Here, we also provide some important properties and theoretical results for both the models. Theoretical results for simple versions of the two sets of models show better properties for the AR models under some conditions which have been shown to hold for the practical data example considered in this thesis. We have

followed this investigation by a simulation study for a more practical version of the models. As expected, the simulation study shows better performance of the DLM when the data are simulated from it. Similarly, the AR models are seen to be better when the data are simulated from it. Finally, we have compared the models by fitting them to a real data set for daily maximum eight-hour average ozone concentration levels in the state of New York for 62 days in July and August, 2006. A predictive Bayesian model choice criterion as well as set aside validation data show that the fitted AR model performs much better than the fitted DLM. These results show that the AR models can be much better than the DLM in practical ozone data modelling situations.

Chapter 6

Trend in Ozone Levels using Models based on Predictive Processes Approximations

6.1 Introduction

Results obtained in Chapter 5 show that the hierarchical autoregressive (AR) models provide better model fits and have superior predictive performances than the DLM. The hierarchical AR models, however, are not suitable for analysing large data sets observed over vast study regions such as the eastern United States (US). The problem here lies in inverting high dimensional spatial covariance matrices repeatedly in iterative model fitting algorithms. This is known as the *big-n problem* in literature (see details in Section 1.6).

Motivated by the need to model large data sets, this Chapter extends the Gaussian predictive processes (GPP) approximation technique of Banerjee *et al.* (2008) to include auto-regressive terms of the latent underlying space-time process. This auto-regressive process is defined at a set of a smaller number of knot locations within the study region and then spatial interpolation, i.e. kriging, is used to approximate the original space-time process. The model is fully specified within a hierarchical Bayesian setup and is implemented using Markov chain Monte Carlo (MCMC) techniques.

This study assesses that the proposed approximation modelling method offers a reliable solution to analyse large and non-stationary spatio-temporal ground

level ozone observations. Here, we aim to predict spatial patterns of the ozone levels in the eastern US and detect their long-term trends after adjusting for the effects of meteorological variables. We use a smaller data set to illustrate and compare the hierarchical AR and the GPP approximation models; and then use 10 years eastern US data to fit the models and obtain trends in ozone levels.

In this chapter we implement the Gibbs sampler for each of the GPP based approximation model where the Metropolis algorithm is used for sampling the ϕ parameter. We note that the MCMC chains converge rapidly for the models. 15000 iterates are used for making inference after discarding the first 5000 iterations. We also used multiple parallel runs and calculated the Gelman and Rubin statistics (Gelman and Rubin 1992), which we found to be satisfactory.

The rest of this chapter is organised as follows: Section 6.2 introduces the modified version of the hierarchical AR models discussed in Chapters 3 and 5. The following Section 6.3 represents the modelling strategy for large dimensional data using modified AR models based on GPP approximations. Section 6.4 describes the joint posterior details for the proposed models. The prediction details are then discussed in Section 6.5. In Section 6.6 we illustrate the proposed modelling approach for a smaller data set consisting of four states of the eastern US. The proposed model is compared with the hierarchical AR models in Section 6.7 using the four-state data example. The following Section 6.8 represents the analysis and prediction of the full eastern US data using the proposed model. Finally, we present some concluding remarks in Section 6.9.

6.2 Modified AR Models

The hierarchical AR models described in Section 3.4.5 assume the AR model for the true values of the modelled response $\mathbf{O}_{lt}$. Following Papamichael (2011) we modify this model so that the modified version does not assume a true level $O_l(\mathbf{s}_i, t)$ for each $Z_l(\mathbf{s}_i, t)$ but instead assumes a space-time random-effect denoted by $\eta_l(\mathbf{s}_i, t)$. It then assumes an AR model for these space-time random effects.

The top level general space-time random effect model is assumed to be:

$$\mathbf{Z}_{lt} = \mathbf{X}_{lt}\boldsymbol{\beta} + \boldsymbol{\eta}_{lt} + \boldsymbol{\epsilon}_{lt}, l = 1, ..., r,\ t = 1, ..., T. \tag{6.1}$$

where $\boldsymbol{\epsilon}_{lt} \sim N(\mathbf{0}, \sigma^2_\epsilon I)$ where I is the identity matrix. In the next stage of the modelling hierarchy the AR model is assumed as:

$$\boldsymbol{\eta}_{lt} = \rho\, \boldsymbol{\eta}_{lt-1} + \boldsymbol{\delta}_{lt}, \tag{6.2}$$

where $\boldsymbol{\delta}_{lt} \sim N(0, \sigma^2\kappa(d;\phi))$. Here $\kappa(d;\phi)$ denotes the correlation function which we take to be the exponential correlation function $\kappa(d;\phi) = \exp(-d\phi)$ in our illustration, although other choices can be adopted. Finally, the initial condition is assumed to be:

$$\eta_l(\mathbf{s}_i, 0) \sim N(0, \sigma_0^2\kappa(d;\phi_0)) \tag{6.3}$$

where σ_0^2 and ϕ_0 are unknown parameters.

Note that the marginal mean of the random effects $\boldsymbol{\eta}_{lt}$ is zero, but the conditional mean given $\boldsymbol{\eta}_{lt-1}$ is no longer zero due to the auto-regressive specification (6.2). This specification also implies a non-stationary marginal covariance function for $\boldsymbol{\eta}_{lt}$ that does not need to be explicitly derived nor is it required since model fitting proceeds through the conditional specification (6.2).

6.3 Models Based on GPP Approximations

The auto-regressive models specified in Section 6.2 create a random effect $\eta_l(\mathbf{s}_i, t)$ in (6.1) corresponding to each data point $Z_l(\mathbf{s}_i, t)$. This will lead to *the big-n problem*, as discussed in Section 1.6 when n is large. To overcome this problem we propose a dimension reduction technique through a kriging approximation following Banerjee *et al.* (2008).

The main idea here is to define the random effects $\eta_l(\mathbf{s}_i, t)$ at a smaller number of locations, called the knots, and then use kriging to predict those random effects at the data locations. The auto-regressive model is only assumed for the random effects at the knot locations and not for all the random effects at the observation location. The method proceeds as follows:

At the top level we continue to assume the model (6.1), but we do not specify $\boldsymbol{\eta}_{lt}$ directly through the auto-regressive model (6.2). Instead, we select $m << n$ knot locations, denoted by $\mathbf{s}^*_1, ..., \mathbf{s}^*_m$ within the study region and let the spatial random effects at these locations at time l and t be denoted by $\mathbf{w}_{lt} = (w_l(\mathbf{s}^*_1, t), \ldots, w_l(\mathbf{s}^*_m, t))'$. Discussion regarding the choice of these loca-

tions is given below. Assuming an underlying Gaussian process independently at each time point l and t, Banerjee *et al.* (2008) show that the process $\boldsymbol{\eta}_{lt}$ can be approximated by

$$\tilde{\boldsymbol{\eta}}_{lt} = A\mathbf{w}_{lt} \tag{6.4}$$

with $A = CS_w^{-1}$ where C denotes the n by m cross-correlation matrix between $\boldsymbol{\eta}_{lt}$ and $\mathbf{w}_{lt}$, and S_w is the correlation matrix of w_{lt}. Note that the common spatial variance parameter does not affect the above since it cancels in the product CS_w^{-1}. Also, there is no contribution of the means of either $\boldsymbol{\eta}_{lt}$ or $\mathbf{w}_{lt}$ in the above since those means are assumed to be 0.

The proposal here is to use the GPP approximation $\tilde{\boldsymbol{\eta}}_{lt}$ instead of $\boldsymbol{\eta}_{lt}$ in the top level model (6.1), thus we assume that:

$$\mathbf{Z}_{lt} = \mathbf{X}_{lt}\boldsymbol{\beta} + \tilde{\boldsymbol{\eta}}_{lt} + \boldsymbol{\epsilon}_{lt}, l = 1, \ldots r,\ t = 1, \ldots, T, \tag{6.5}$$

where $\tilde{\boldsymbol{\eta}}_{lt}$ is as given in (6.4). Analogous to (6.2), we specify $\mathbf{w}_{lt}$ at the knots conditionally given $\mathbf{w}_{lt-1}$ as:

$$\mathbf{w}_{lt} = \rho\, \mathbf{w}_{lt-1} + \xi_{lt}, \tag{6.6}$$

where $\boldsymbol{\xi}_{lt} \sim N(\mathbf{0}, \sigma_w^2 \kappa(d; \phi_w))$ independently. Again we assume that $\mathbf{w}_{l0} \sim N(\mathbf{0}, \sigma_l^2 S_0)$ independently for each $l = 1, \ldots, r$, where the elements of the covariance matrix S_0 are obtained using the correlation function, $\kappa(d; \phi_0)$, i.e. the same correlation function as previously but with a different variance component for each year and also possibly with a different decay parameter ϕ_0 in the correlation function.

The above modelling specifications are justified using the usual hierarchical modelling philosophies in the sense that the top level model is a mixed model with mean zero random effects and these random effects have structured correlations as implied by the spatial auto-regressive model at the second stage (6.6). These two model equations, together with the initial condition, however, are neither intended to, nor will ever imply the auto-regressive model (6.2) for the original random effects $\boldsymbol{\eta}_{lt}$ except for trivial cases such as the one where $m = n$ and all the knot locations coincide with the data locations. In general such a property can never be expected to hold without further conditions.

6.4 Joint Posterior Details

Define $N = nrT$ and let $\boldsymbol{\theta}$ denote all the parameters $\boldsymbol{\beta}$, ρ, σ^2_ϵ, σ^2_w, ϕ, ϕ_0, σ^2_l, $l = 1, ..., r$. Further, let $\mathbf{z}^*$ denote the missing data and $\mathbf{z}$ denote all the non-missing data. The log of the joint posterior distribution for the models in equations (6.5) and (6.6), denoted by $\log \pi(\boldsymbol{\theta}, \mathbf{z}^*|\mathbf{z})$ is written as:

$$
\begin{aligned}
&-\frac{N}{2}\log\sigma^2_\epsilon - \frac{1}{2\sigma^2_\epsilon}\sum_{l=1}^{r}\sum_{t=1}^{T}(\mathbf{Z}_{lt} - \mathbf{X}_{lt}\boldsymbol{\beta} - A\mathbf{w}_{lt})'(\mathbf{Z}_{lt} - \mathbf{X}_{lt}\boldsymbol{\beta} - A\mathbf{w}_{lt}) \\
&-\frac{mrT}{2}\log\sigma^2_w - \frac{rT}{2}\log|\mathbf{S}_w| - \frac{1}{2\sigma^2_w}\sum_{l=1}^{r}\sum_{t=1}^{T}(\mathbf{w}_{lt} - \rho\mathbf{w}_{lt-1})'\mathbf{S}_w^{-1}(\mathbf{w}_{lt} - \rho\mathbf{w}_{lt-1}) \\
&-\frac{m}{2}\sum_{l=1}^{r}\log\sigma^2_l - \frac{r}{2}\log|\mathbf{S}_0| - \frac{1}{2}\sum_{l=1}^{r}\frac{1}{\sigma^2_l}\mathbf{w}_{l0}\mathbf{S}_0^{-1}\mathbf{w}_{l0} + \log\pi(\boldsymbol{\theta})
\end{aligned}
\tag{6.7}
$$

where, $\log\pi(\boldsymbol{\theta})$ is the log of the prior distribution for the parameter $\boldsymbol{\theta}$. We assume the prior distributions $\boldsymbol{\beta} \sim N(\mathbf{0}, 10^4)$, $\rho \sim N(0, 10^4)I(0 < \rho < 1)$. Further, the prior distributions for the variance parameters are: $1/\sigma^2_\epsilon \sim G(a, b)$, $1/\sigma^2_w \sim G(a, b)$, where the Gamma distribution has mean a/b. We shall choose the values of a and b in such a way that guarantees a proper prior distribution for these variance components, see Chapter 3 for more on prior distributions.

6.4.1 Full Conditional Distribution for Covariate Coefficients

From the kernel of the joint posterior distribution (6.7), we obtain the full conditional distribution of the covariate coefficient, $\boldsymbol{\beta}$ as $N(\Delta\chi, \Delta)$ where,

$$\Delta^{-1} = \frac{1}{\sigma^2_\epsilon}\sum_{l=1}^{r}\sum_{t=1}^{T}X'_{lt}X_{lt} + 10^{-4}I,$$

$$\chi = \frac{1}{\sigma^2_\epsilon}\sum_{l=1}^{r}\sum_{t=1}^{T}X'_{lt}(\mathbf{Z}_{lt} - A\mathbf{w}_{lt}).$$

6.4.2 Full Conditional Distribution for Autoregressive Parameter

The full conditional distribution of the auto-regressive parameter ρ is $N(\Delta\chi, \Delta)I(0 < \rho < 1)$ where,

$$\Delta^{-1} = \sum_{l=1}^{r}\sum_{t=1}^{T}\mathbf{w}'_{lt-1}Q_w\mathbf{w}_{lt-1} + 10^{-4}$$

$$\chi = \sum_{l=1}^{r}\sum_{t=1}^{T} \mathbf{w}'_{lt-1} Q_w \mathbf{w}_{lt}$$

where $Q_w = \Sigma_w^{-1}$.

6.4.3 Full Conditional Distribution for Variance Parameters

We also obtain the full conditional distributions for the variance parameters of the models from the kernel of the posterior distribution in (6.7). The full conditional distribution of $\frac{1}{\sigma_\epsilon^2}$ is given by:

$$G\left(\frac{N}{2} + a, b + \frac{1}{2}\sum_{l=1}^{r}\sum_{t=1}^{T}(\mathbf{Z}_{lt} - X_{lt}\boldsymbol{\beta} - A\mathbf{w}_{lt})'(\mathbf{Z}_{lt} - X_{lt}\boldsymbol{\beta} - A\mathbf{w}_{lt})\right)$$

Similarly, the full conditional distribution for $\frac{1}{\sigma_w^2}$ is written as:

$$G\left(\frac{mrT}{2} + a, b + \frac{1}{2}\sum_{l=1}^{r}\sum_{t=1}^{T}(\mathbf{w}_{lt} - \rho\mathbf{w}_{lt-1})' Q_w (\mathbf{w}_{lt} - \rho\mathbf{w}_{lt-1})\right)$$

The full conditional distribution of σ_l^2 for $l = 1, \ldots, r$ is given by:

$$G\left(\frac{m}{2} + a, b + \frac{1}{2}\mathbf{w}_{l0}\mathbf{S}_0^{-1}\mathbf{w}_{l0}\right).$$

6.4.4 Full Conditional Distribution for Spatial Error Processes

The full conditional distribution for $\mathbf{w}_{lt}$ is given by: $N(\Delta\chi, \Delta)$ where

$$\Delta^{-1} = \frac{1}{\sigma_\epsilon^2}A'A + Q_w + \rho^2 Q_w$$

$$\chi = \frac{1}{\sigma_\epsilon^2}A'(\mathbf{Z}_{lt} - X_{lt}\boldsymbol{\beta}) + Q_w\mathbf{w}_{lt-1} + Q_w\mathbf{w}_{lt+1},$$

for $1 \leq t < T$. For $t = T$, we have

$$\Delta^{-1} = \frac{1}{\sigma_\epsilon^2}A'A + Q_w$$

$$\chi = \frac{1}{\sigma_\epsilon^2}A'(\mathbf{Z}_{lt} - X_{lt}\boldsymbol{\beta}) + Q_w\mathbf{w}_{lt-1}.$$

The full conditional distribution of $\mathbf{w}_{l0}$ is given by $N(\Delta\chi, \Delta)$ where,

$$\Delta^{-1} = \rho^2 Q_w + Q_0^{-1}$$

$$\chi = \rho Q_w \mathbf{w}_{l1} + \mu_l \Sigma_0^{-1} \mathbf{1}_m,$$

where $Q_0 = \Sigma_0^{-1}$.

6.4.5 Sampling the Spatial-Decay Parameter

We observe from the posterior distribution in (6.7) that the full conditional distribution of ϕ_w is not available in closed form. The log of the conditional posterior density (up to an additive constant) is given by:

$$\log \pi(\phi_w|...) = \log \pi(\phi_w) - \frac{rT}{2} \log |S_w| - \frac{1}{2} \sum_{l=1}^{r} \sum_{t=1}^{T} (\mathbf{w}_{lt} - \rho \mathbf{w}_{lt-1})' Q_w (\mathbf{w}_{lt} - \rho \mathbf{w}_{lt-1})$$

Similarly, the log of the conditional posterior density of ϕ_0 (up to an additive constant) is given by:

$$\log \pi(\phi_0|.) = \log \pi(\phi_0) - \frac{r}{2} \log(|S_0|) - \frac{1}{2} \sum_{l=1}^{r} \frac{1}{\sigma_l^2} \mathbf{w}_{l0} \mathbf{S}_0^{-1} \mathbf{w}_{l0}.$$

6.4.6 Sampling the Missing Observations

Using the nugget effect we obtain the conditional distribution for missing observations ($\mathbf{z}^*$) as:

$$\pi(\mathbf{z}_{lt}^*|.) \sim N\left(\boldsymbol{\mu}_{lt}^*, \sigma_\epsilon^2\right)$$

where, $\boldsymbol{\mu}_{lt}^*$ is the $q \times 1$ mean vector at time $t = 1, ..., T$ and $l = 1, ..., r$, for the q missing values that is equal to $\mathbf{X}_{lt}^* \boldsymbol{\beta} + \tilde{\boldsymbol{\eta}}_{lt}$, where $\mathbf{X}_{lt}^*$ is the corresponding covariates related to the q independent missing observations.

We use Markov chain Monte Carlo (MCMC) techniques (see details in Chapter 3) to obtain the estimates of the parameters. Further, in this chapter for simplicity we assume the spatial decay parameters for $\mathbf{w}_{l0}$ and $\mathbf{w}_{lt}$ are same, i.e., $\phi_0 = \phi_w = \phi$, however one can treat them differently. Hence, we sample and estimate the spatial decay parameter ϕ, instead of ϕ_0 and ϕ_w, from the posterior density $\pi(\phi|\rho, \mathbf{w}, \mathbf{z})$ that do not have any closed form of the distribution. Henceforth, we use *Metropolis-Hasting* (MH) algorithm (see details in Chapter 3) to

sample ϕ, given data.

6.5 Prediction Details

We want to predict the response $Z_l(\mathbf{s}', t)$ at a new site $\mathbf{s}'$ and time l and t. According to the top-level model (6.1), we obtain:

$$Z_l(\mathbf{s}', t) = \mathbf{x}_l(\mathbf{s}', t)'\boldsymbol{\beta} + \tilde{\eta}_l(\mathbf{s}', t) + \epsilon_l(\mathbf{s}', t),\ l = 1, \ldots r,\ t = 1, \ldots, T, \tag{6.8}$$

where $\mathbf{x}_l(\mathbf{s}', t)$ denotes the covariate value at the new location at time l and t, and the scalar $\tilde{\eta}_l(\mathbf{s}', t)$ is obtained using the following equation, obtained analogously as (6.4),

$$\tilde{\eta}_l(\mathbf{s}', t) = \mathbf{c}'(\mathbf{s}')S_w^{-1}\mathbf{w}_{lt} \tag{6.9}$$

where the kth element of the m by 1 vector $\mathbf{c}(\mathbf{s}')$ is given by $\psi(d; \phi)$ where d is the distance between the sites $\mathbf{s}_k^*$ and $\mathbf{s}'$.

Prediction is straightforward under any MCMC sampling scheme. At each iteration, j say, first one obtains the approximation $\tilde{\eta}_l^{(j)}(\mathbf{s}', t)$ calculated using the current parameter iterates $\boldsymbol{\theta}^{(j)}$ and $\mathbf{w}_{lt}^{(j)}$. The next step is to generate a new $Z_l^{(j)}(\mathbf{s}', t)$ using the model (6.8) and plugging in $\boldsymbol{\theta}^{(j)}$.

6.6 Illustration of the GPP based Models for the Four States Example

In this section, we illustrate the proposed GPP approximation models for the four states data. We use data obtained from the four states, e.g., Illinois, Indiana, Ohio, and Kentucky for the period of 10 years starting from 1997 to 2006 (see Chapter 4). We also use three meteorological variables, i.e., maximum temperature in $^\circ C$, average wind speed in nautical miles, and average relative humidity in percentage (see Chapter 4 for more details). Figure 6.1 represents the map of four states, and the ozone and meteorological monitoring sites. We fit data from 148 sites and 10%$\approx$ 16 of the total 164 sites are set aside for validation purpose. Rest of this section will elaborate the GPP approximate model fitting and the sensitivity analysis.

From Figure 6.2 we observe the yearly trend in daily ozone levels in the four

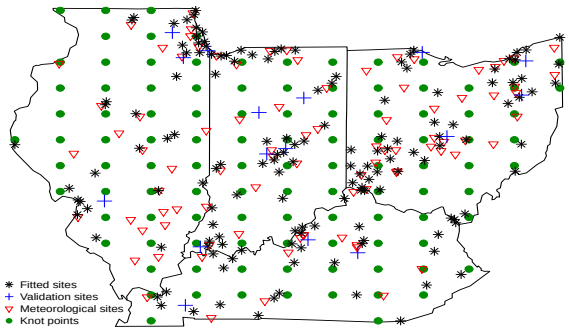


Figure 6.1: A map of the four states, Ohio, Indiana, Illinois and Kentucky. Total 164 ozone monitoring locations (of which 148 are used for model fitting and 16 are for validation), 88 meteorological sites and 107 grid knot points are superimposed.

states. The box-plot by years for all states have similar patterns. For example, the overall ozone level goes up in years 1998 and 1999 and comes down in 2000 and again goes up in 2002 and so on. Some extreme outlying daily ozone levels are recorded in the Ohio and Illinois states in 1999. Table 6.1 represents the summary statistics for ozone and meteorological variables used in this chapter. Ozone levels vary a great deal between 1.00 ppb to 241.40 ppb, we also observe high variability for wind speed.

	Minimum	Mean	Median	Maximum
Ozone	1.00	51.86	51.12	241.40
Max. Temp.	7.95	27.15	27.82	40.01
RH	0.78	3.65	3.56	9.09
WDSP	0.00	5.76	5.46	21.26

Table 6.1: Summary statistics for ozone levels (in ppb), maximum temperature (Max. Temp.) in degree C, percentage relative humidity (RH) and average wind speed (WDSP) in nautical miles per hour in the four states for years 1997-2006.

6.6.1 Sensitivity of Knot Sizes

We define five different sets of regular grid locations randomly starting from 6×6, 8×8, 10×10, 12×12 and 14×14 over the four states. From these regular grids, we choose the points inside the boundary of the four states as knot locations to fit the GPP based models, and finally consider knot sizes 26, 40, 60, 107 and 138

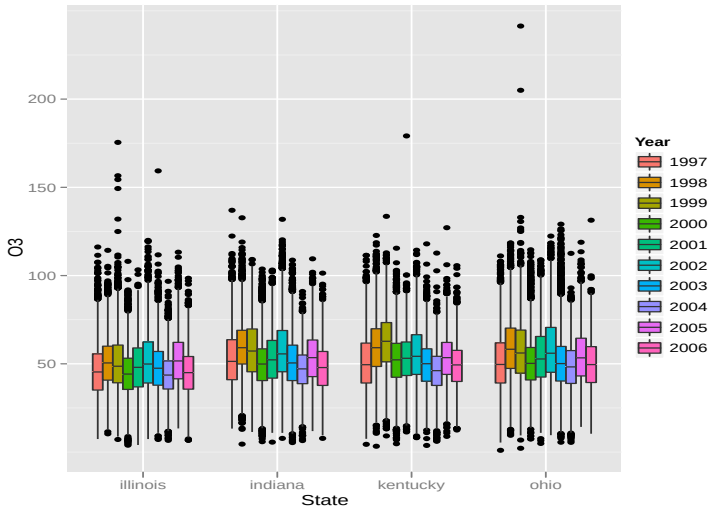


Figure 6.2: Box-plot of ozone levels observed in Illinois, Indiana, Ohio, and Kentucky by years.

respectively. Figure 6.1 shows an example of the knot location points (here, 107 knots) that we consider to model the ozone concentration data of the four states.

From Table 6.2 we obtain the model performance and validation criteria (see details in Section 3.3) for different set of knots. As expected, both validation and model choice criteria show better results as the number of knots increases. However, we observe the difference between the results for knot sizes 107 and 138 are very small. This result gives us an idea that after a particular choice of knot size the predictive performance of the models are approximately same. Hence, in this thesis we choose knot size 107 to analyse further the four states data.

	Validation Criteria				Model Choice Criterion
Knots	RMSE	MAE	rBIAS	rMSEP	GoF + P = PMCC
26	6.31	4.58	-0.006	0.15	35424.70+44128.77=79553.47
40	6.19	4.48	-0.009	0.15	32595.85+42797.55=75393.40
60	6.17	4.44	-0.006	0.15	30977.74+42128.34=73106.08
107	6.07	4.37	-0.009	0.14	28143.46+40893.51=69036.97
138	6.06	4.36	-0.009	0.14	27690.84+40503.23=68194.07

Table 6.2: Values of the model choice and validation criteria for different knot sizes for the four states example.

6.6.2 Sensitivity of Prior Selection

A sensitivity study has been conducted for different hyper-parameter values of the Gamma distribution. Knot size 107 is considered here as discussed in Section 6.6.1. Table 6.3 shows validation and model choice results for different combinations of the hyper-parameters. We observe prior specification with $a = 2$ and $b = 1$ gives the best validation and model choice results compared to other combinations of a and b.

	Gamma Prior Distribution		
	Validation Criteria		Model Choice Criterion
Gamma(a,b)	RMSE	MAE	GoF + P = PMCC
(a=2,b=1)	6.07	4.37	28143.46+40893.51=69036.97
(a=1,b=1)	6.10	4.41	29370.61+40273.76=69644.37
(a=2,b=2)	6.10	4.40	29383.05+40373.42=69756.47
(a=10,b=10)	6.10	4.41	29432.03+40188.96=69620.99

Table 6.3: Values of the model choice and validation criteria for different hyper-parameters for the four states example.

6.6.3 Choice for Sampling Spatial Decay Parameter

The full conditional distribution of the spatial decay (ϕ) parameter does not have any closed form. Hence, we can use different types of sampling scheme for choosing ϕ parameter. We can also use the empirical Bayes approach that has been discussed in Section 5.4.5, however omitted for brevity.

In this section we discuss only discrete and random-walk sampling strategies and see their performance based on the predictive and model choice criteria. For random-walk we use Metropolis-Hastings approach and use a suitable tuning parameter to obtain acceptance rate between 20% to 40% (see Gelman *et al.* 1997). Appropriate tuning yields 32.4% acceptance rate for the ϕ parameter. For discrete sampling of ϕ we choose points that are defined from 0.001 to 0.1 with 50 equal segments. Table 6.4 represents the results based on both sampling scheme, where better performance of random-walk Metropolis approach is observed over the discrete sampling.

	Sensitivity for ϕ sampling		
	Validation Criteria		Model Choice Criterion
	RMSE	MAE	GoF + P = PMCC
Discrete	6.16	4.46	29493.75+40235.93=69729.68
Random-walk	6.07	4.37	28143.46+40893.51=69036.97

Table 6.4: Values of the model choice and validation criteria for different sampling of ϕ for the four states example.

6.6.4 Adjustment of the Spatial Misalignment

From Figure 6.1 we observe that there are misalignments between ozone and meteorological monitoring locations. To adjust the misalignment we use Kriging method discussed in Section 2.5. In this section we use two techniques for Kriging, first one is the *single kriging* (SK) and second one is the *multiple kriging* (MK).

In the SK approach we krige the meteorological variables into the ozone monitoring sites only once and use these kriged values as covariates in the models. For MK, we sample the meteorological variables in each iteration from normal distribution with mean and variance defined by the kriged values and corresponding kriged variances respectively. The former case, does not take into account the variation that occurred due to kriging in the MCMC iteration algorithm. The MK approach includes the effect of kriging variance in the iteration, however yields computational burden.

Table 6.5 provides model choice and validation results for the GPP based model. Here we use both SK and MK adjustment techniques of spatial misalignment to measure the performance. We observe, the RMSE is larger for the MK approach, in addition the penalty of the model choice criteria is very large compared to the SK approach. We also observe the 95% prediction coverage for the SK is smaller than the MK, however is much closer than the coverage obtained from MK approach. Henceforth, the rest of the analysis is done considering of SK approach of adjusting spatial mis-alignment.

6.6.5 Results for Different Sets of Hold-Out Sites

The consistency in the model prediction are observed through the data of different sets of hold-out sites. We randomly choose 7 hold-out data sets, each of them consists of ozone observations from 16 monitoring sites. These 16 hold-out sites

	Validation Criteria		Model Choice Criterion	Coverage (%)
	RMSE	MAE	GoF + P = PMCC	95%
SK	6.07	4.37	28143.46+40893.51=69036.97	92.88
MK	8.05	6.14	42931.68+113752.95=156684.63	100.00

Table 6.5: Values of the model choice and validation criteria for single kriging (SK) and multiple kriging (MK) approach for imputing missing meteorological data.

are randomly chosen from the 164 ozone monitoring sites in the four states. Figure 6.3 represents the four states map, where data sets of different validation sites are given in numbers.

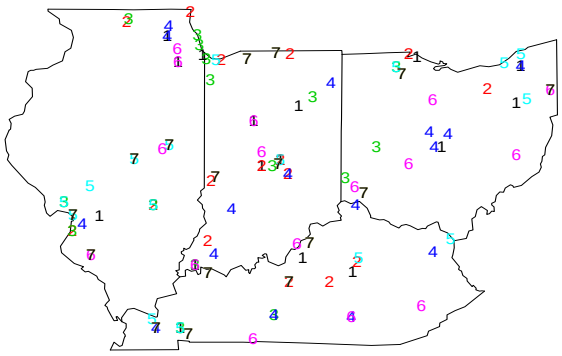


Figure 6.3: Different sets of hold-out validation sites are numbered in the map of the four states.

From Table 6.6 we observe that for 107 knot sizes, the RMSE of the different hold-out data sets varies from 6.07 to 6.20. We also observe the prediction interval varies from 80.60% to 93.98%.

6.7 Comparison with the Hierarchical AR Models

This section is devoted to comparing the proposed GPP approximation models with the hierarchical AR models (see Section 3.4.5). Note that the proposed model does not directly approximate the AR model and hence the latter is not likely to be uniformly better than the former, and therefore this comparison is

	Validation Criteria				
Data set	RMSE	MAE	rBIAS	rMSEP	Coverage
1	6.09	4.42	-0.007	0.15	93.98%
2	6.20	4.55	-0.005	0.16	80.60%
3	6.08	4.42	-0.004	0.15	90.26%
4	6.07	4.37	-0.009	0.14	92.88%
5	6.11	4.52	-0.008	0.16	84.67%
6	6.10	4.50	-0.009	0.16	84.87%
7	6.07	4.36	-0.009	0.14	91.58%

Table 6.6: Validation criteria for different sets of hold-out sites using 107 knots for the four states example.

meaningful. Both the models use the same three covariates, namely, maximum temperature (MaxTemp), relative humidity (RH) and wind speed (WDSP). In both the cases we also adopt the same prior distributions and use the Metropolis-Hastings sampling algorithm for sampling the spatial decay parameters. In the GPP based proposed model we use 107 knot points as decided in Section 6.6.1.

The estimates of the parameters of the two models are provided in Table 6.7, except for the parameters μ_l and σ_l^2 under the hierarchical AR model and σ_l^2 under the GPP based model, $l = 1, ..., r$ since those estimates are not interesting for model comparison purposes. Both the models show significant effect of the three covariates, although the effects get attenuated under the hierarchical AR model due to the presence of the temporal auto-regression.

Further discussion about these effects is provided in Section 6.8. However, there are large differences between the two models as regards to the estimates of spatial and temporal correlations. The temporal correlation under the AR model (0.523) is much larger than the same for the GPP based model (0.102). This is due to the fact that the auto-regressive model for the Sahu *et al.* version is assumed for the true ozone levels which are highly temporally correlated, whereas the GPP based model assumes the auto-regression for the latent random effects which are also significantly temporally correlated but at a magnitude lower than that for the true ozone levels in AR model. However, to compensate for this low value of temporal correlation, the GPP based model has estimated a much higher level of spatial correlation since the spatial decay of 0.0036 is much smaller for this model compared to the same, 0.012, for the full hierarchical AR model. The estimates of the variance components, under both the models, show that more

variation is explained by the spatial effects than the pure error.

The two models are compared using the PMCC and the two model validation criteria: RMSE and MAE. We also report the nominal coverage of the 95% prediction intervals for the out of sample validation data. These three validation statistics are based on 21,008 (=24480–3472) daily observations (see details in Chapter 4).

Parameter	Mean	sd	2.5%	97.5%
		AR Model		
Intercept	4.447	0.061	4.346	4.541
Max.Temp.	0.016	0.001	0.013	0.018
RH	–0.314	0.010	–0.325	–0.302
WDSP	–0.061	0.002	–0.065	–0.057
ρ	0.523	0.002	0.519	0.526
σ^2_ϵ	0.056	0.001	0.055	0.058
σ^2_η	0.537	0.038	0.527	0.540
ϕ	0.0120	0.0006	0.0119	0.0121
		GPP based Model		
Intercept	6.353	0.056	6.224	6.445
MaxTemp	0.060	0.001	0.057	0.063
RH	–0.179	0.009	–0.198	–0.160
WDSP	–0.033	0.001	–0.036	–0.031
ρ	0.102	0.003	0.095	0.109
σ^2_ϵ	0.169	0.001	0.167	0.171
σ^2_w	0.457	0.004	0.449	0.466
ϕ	0.0036	0.0001	0.0030	0.0041

Table 6.7: Parameter estimates of the two AR models.

	P	G	P+G	RMSE	MAE	Coverage (%)
Full	90,807.32	41,077.80	131,885.10	6.82	5.04	93.50
GPP	40,893.51	28,143.46	69,036.97	6.07	4.37	92.88

Table 6.8: Model comparison results for the hierarchical AR and GPP based models.

Model comparison results presented in Table 6.8 almost uniformly give evidence in favour of the the proposed GPP based models. Components of the PMCC show that the GPP based model provides a much better fit than the AR model. Both the RMSE and MAE are also better for the proposed model. However, the nominal coverage is slightly smaller for the proposed method, but this is not much of a cause for concern since both are close to 95%. The GPP

based model fitting requires about 2.24 hours of computing time while the full AR model takes about 7.86 hours. Thus the GPP based model implementation requires less than a third of the computing time needed for fitting the AR model. In conclusion, the GPP based model not only provides a faster and better fit but also validates better than the hierarchical AR model. In the next section, for the full eastern US data, we shall only consider the GPP based model.

6.8 Analysis for the Eastern US Ozone Concentration Levels

In this section we analyse the full eastern US data set introduced in Chapter 4. We use data from 622 monitoring sites to model and the data for the remaining 69 sites are set aside for validation, see Figure 4.1.

We continue to use the three meteorological variables as covariates in the model. We choose the same prior and the Metropolis-Hastings sampling method for the spatial decay parameter ϕ. To select the number of knots we start with regular grid sizes of 12×12, 15×15, 20×20 and 25×25 and then only retain the points inside the land boundary of the eastern US that gives us 68, 105, 156 and 269 points respectively. As in previous section we fit and predict using the model with these knot sizes and obtain the two validation statistics: RMSE and MAE in Table 6.9.

	Knot Sizes			
	269	156	105	68
RMSE	6.41	6.42	6.78	7.09
MAE	4.73	4.75	5.02	5.26

Table 6.9: Two model validation criteria for different knot sizes

As has already been seen in Section 6.6, the performance gets better with increasing grid sizes, but the improvement in performance is only marginal when the grid size goes up to 269 from 156. The much smaller computational burden with 156 knot points outweighs this marginal improvement in the validation statistics. Henceforth we proceed with grid size 156 in our analysis. Figure 6.4 provides a map of the eastern US with these grid points superimposed.

Parameter estimates of the fitted model with 156 knot points are provided in

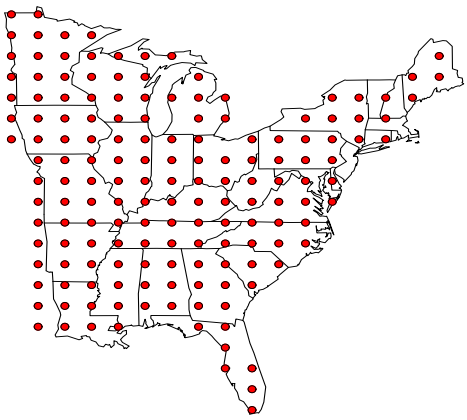


Figure 6.4: A map of the eastern US with 156 grid knot points superimposed.

Table 6.10. All three covariates, Max.Temp., WDSP, and RH, remain significant in the spatio-temporal model with a positive effect of MaxTemp and negative effects of the other two. This is in accordance with the results in the literature in ozone modelling, see e.g. in Section 1.5. The auto-regressive parameter is also significant and the pure error variance σ^2_ϵ is estimated smaller than the spatial variance σ^2_w. The spatial decay parameter is estimated to be 0.0018 which corresponds to an effective spatial range (Sahu, 2011) of 1666.7 kilometres that is about half of the maximum distance between any two locations inside the study region.

	Mean	sd	2.5%	97.5%
Intercept	6.817	0.101	6.604	6.991
MaxTemp	0.027	0.001	0.025	0.029
RH	–0.243	0.004	–0.251	–0.234
WDSP	–0.009	0.002	–0.013	–0.006
ρ	0.132	0.002	0.128	0.136
σ^2_ϵ	0.266	0.001	0.265	0.267
σ^2_w	0.729	0.014	0.708	0.770
ϕ	0.0018	0.0001	0.0017	0.0019

Table 6.10: Parameter estimates of the fitted GPP based AR model for the eastern US data.

We now turn to the validation of the ozone summaries: the annual 4th highest

maximum and the 3-year rolling average (see Section 1.2) of these. Table 6.11 provides the validation statistics. We also report the validation statistics for these summaries obtained using simple kriging using the `fields` package (Fields Development Team 2006). In this method the daily ozone levels are first kriged and then those are aggregated up to the annual levels. It is remarkable that the proposed method is able to perform better in out of sample predictions than standard kriging which is well known to be difficult to beat using model based approaches (Liu *et al.*, 2011). This shows that the model is very accurate in predicting the ozone standard based on the annual summaries.

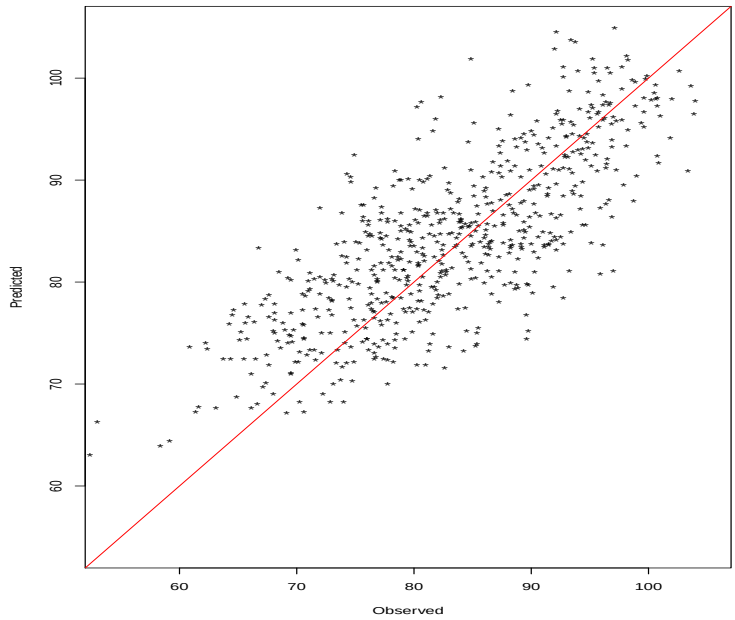
	Annual 4th highest		3-year average	
	Kriging	Model	Kriging	Model
RMSE	5.41	5.24	4.27	4.21
MAE	4.38	4.17	3.51	3.36

Table 6.11: Two validation criteria for the annual ozone summaries

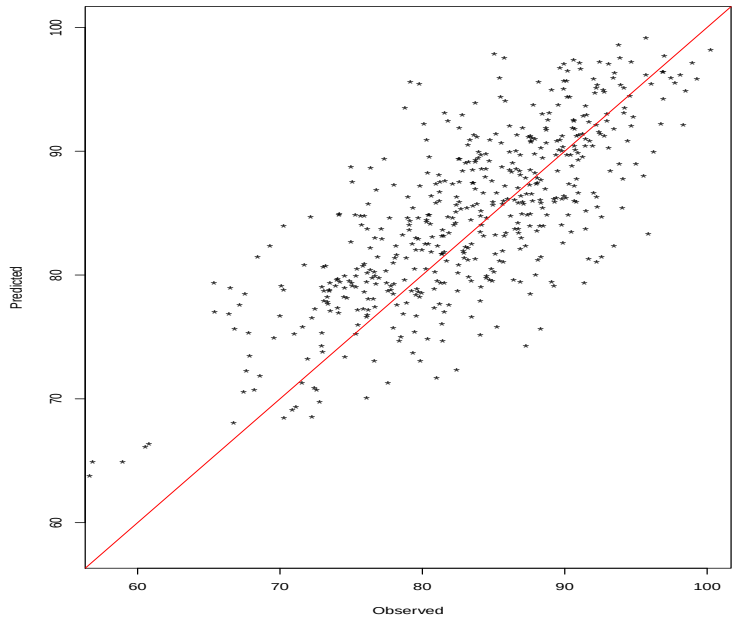
Figure 6.5 examines this in more detail where the predicted values of these summaries are plotted against the observed values. The plot provides evidence of accurate prediction with a slight tendency to over predict. The actual over prediction percentage for the annual 4th highest maximum is 52% while the same for the 3-year rolling averages is slightly higher at 56% which are reasonable. Hence we proceed to make predictive inference for the ozone standard based on these model based annual summaries.

We perform predictions at 936 locations inside the land-boundary of the eastern US obtained from a regular grid. At each of these sites we spatially interpolate the daily maximum 8-hour average ozone level on each of 153 days in every year using the details in Section 6.5. These daily levels are then aggregated up to the annual levels. Figures 6.6 to 6.10 provide the model based interpolated maps of annual 4th highest maximum ozone levels for the years 1997-2006. Observed values of these annual maxima from a selected number of sites (data from all the 691 sites are not plotted to avoid clutter) are also superimposed and those show reasonably good agreement with the predicted values.

Similarly, Figures 6.11 to 6.14 plot the model based interpolated maps of the 3-year rolling averages of the annual 4th highest maximum ozone concentration levels for the years 1999-2006. The plotted observed values of these rolling

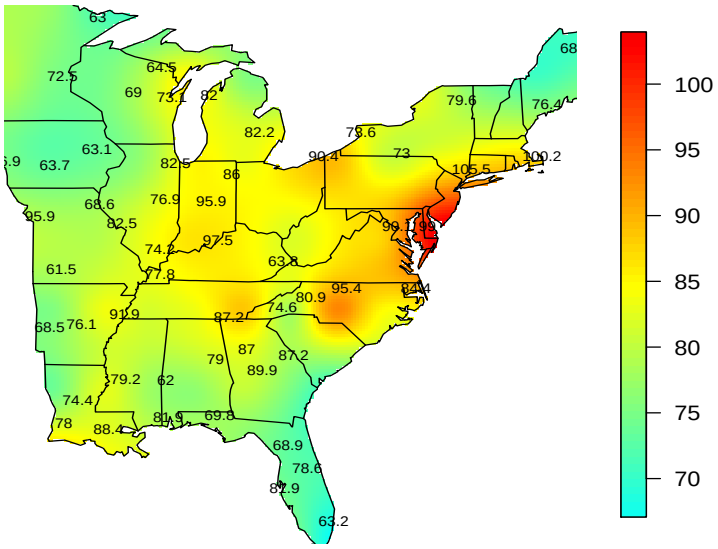


(a)

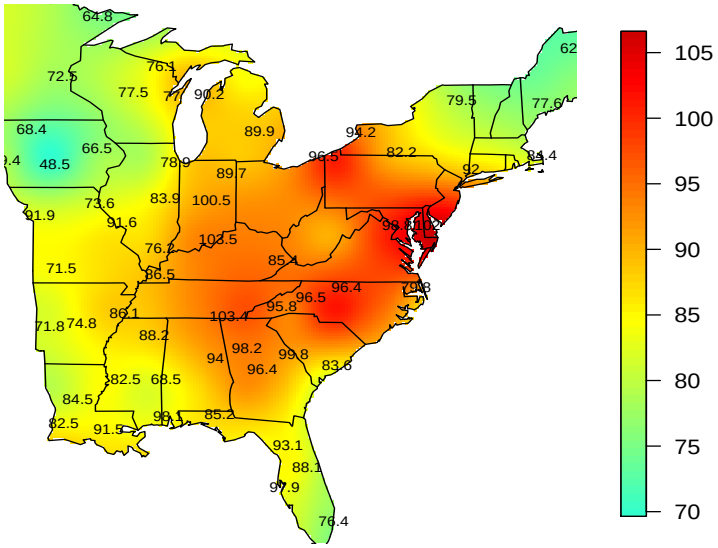


(b)

Figure 6.5: Scatter plots of the prediction against the observed values, (a): annual 4th highest maximum, (b) 3-year rolling average of the annual 4th highest maximum. The $y = x$ line is superimposed.

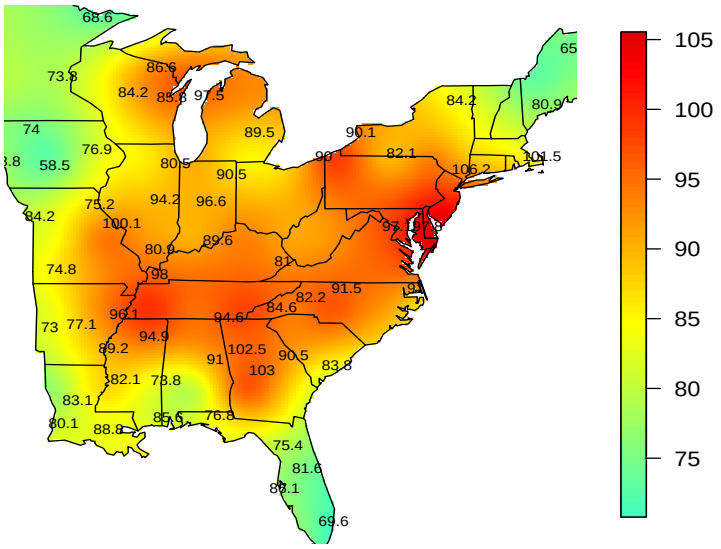


(a)

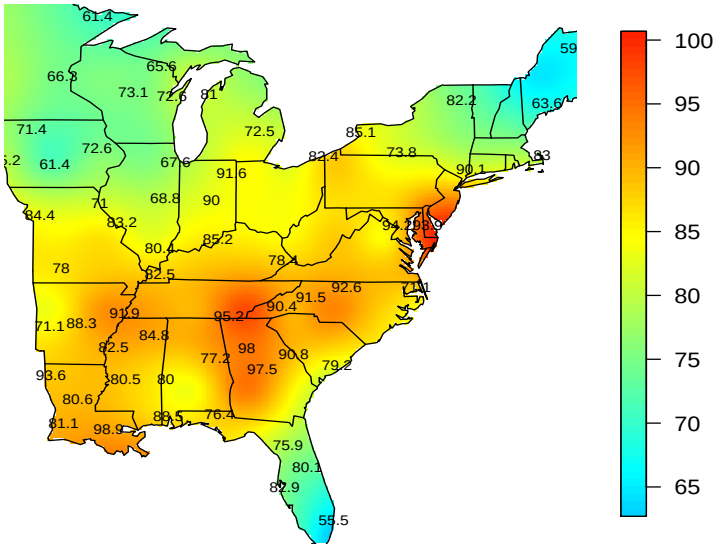


(b)

Figure 6.6: Model based interpolation of the annual 4th highest maximum ozone levels, panel (a) for 1997 and (b) for 1998. Observed data from a few selected sites, to enhance readability, are superimposed.

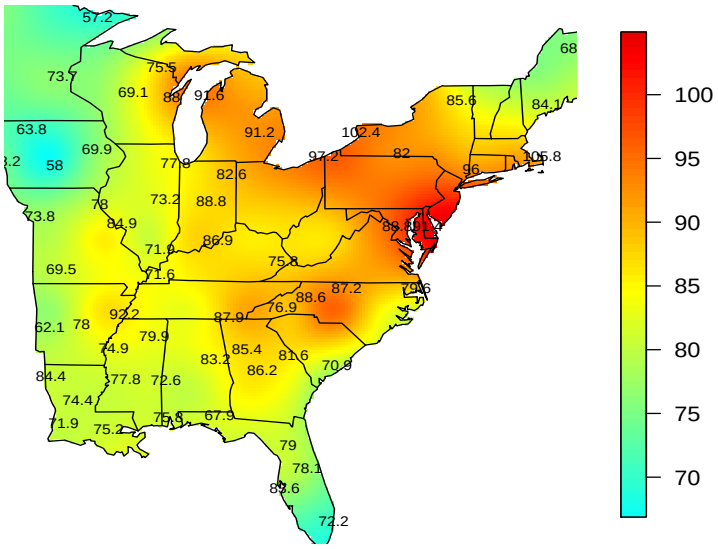


(a)

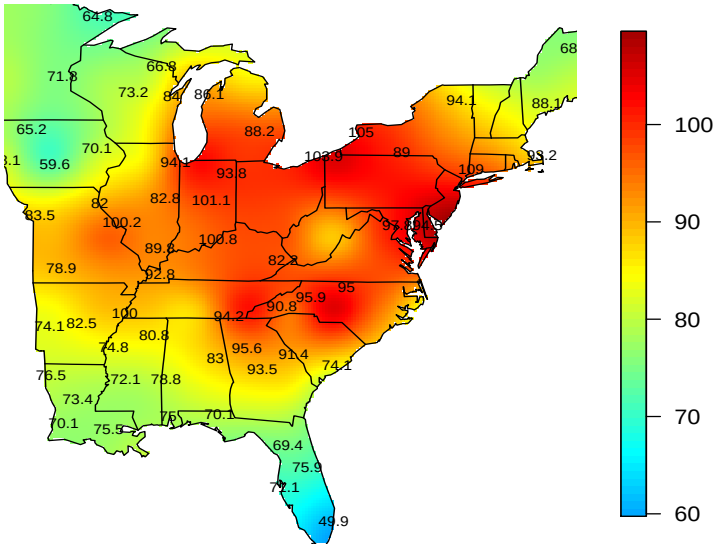


(b)

Figure 6.7: Model based interpolation of the annual 4th highest maximum ozone levels, panel (a) for 1999 and (b) for 2000. Observed data from a few selected sites, to enhance readability, are superimposed.

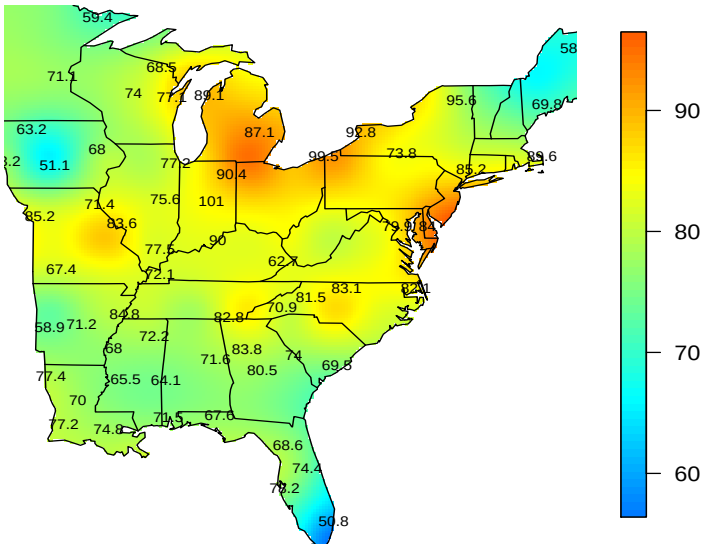


(a)

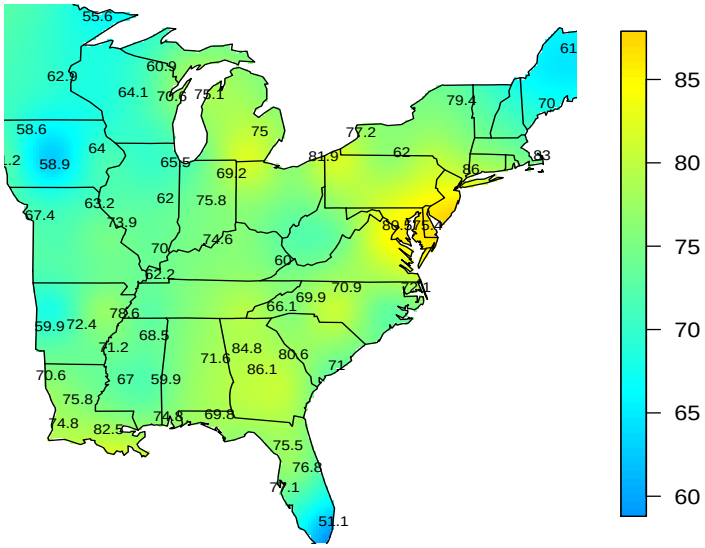


(b)

Figure 6.8: Model based interpolation of the annual 4th highest maximum ozone levels, panel (a) for 2001 and (b) for 2002. Observed data from a few selected sites, to enhance readability, are superimposed.

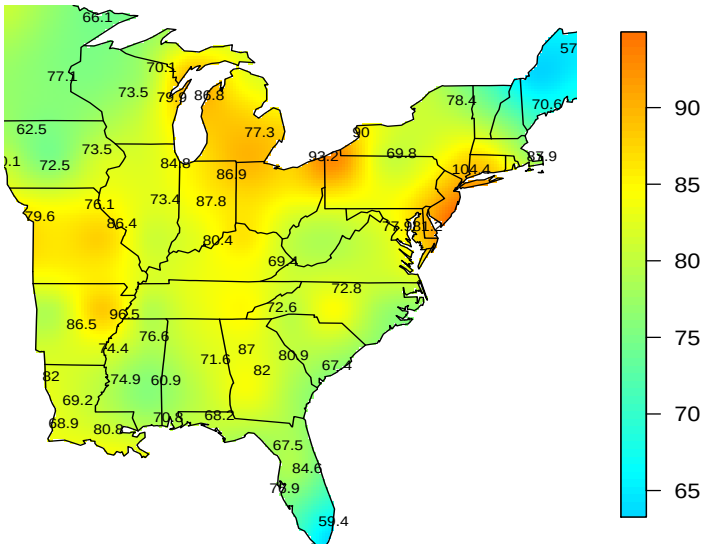


(a)

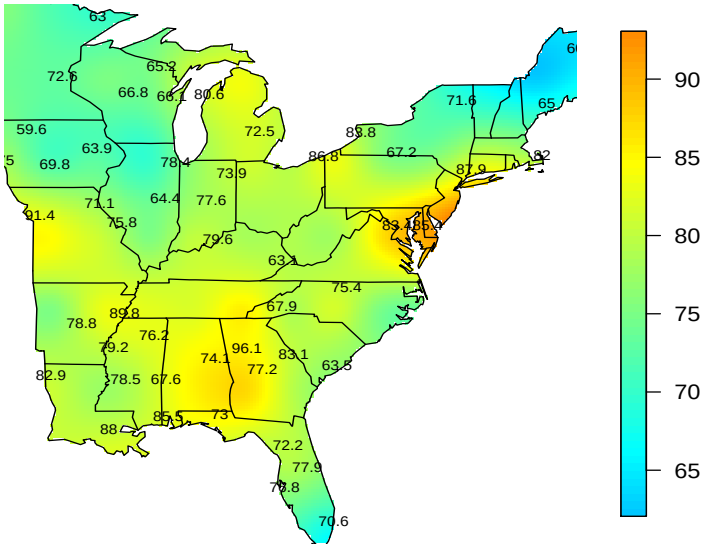


(b)

Figure 6.9: Model based interpolation of the annual 4th highest maximum ozone levels, panel (a) for 2003 and (b) for 2004. Observed data from a few selected sites, to enhance readability, are superimposed.



(a)



(b)

Figure 6.10: Model based interpolation of the annual 4th highest maximum ozone levels, panel (a) for 2005 and (b) for 2006. Observed data from a few selected sites, to enhance readability, are superimposed.

averages are also in good agreement with the predicted values. The uncertainty maps corresponding to the prediction maps in Figures 6.6 and 6.14 showed larger uncertainty for the locations which are farther away from the monitoring sites.

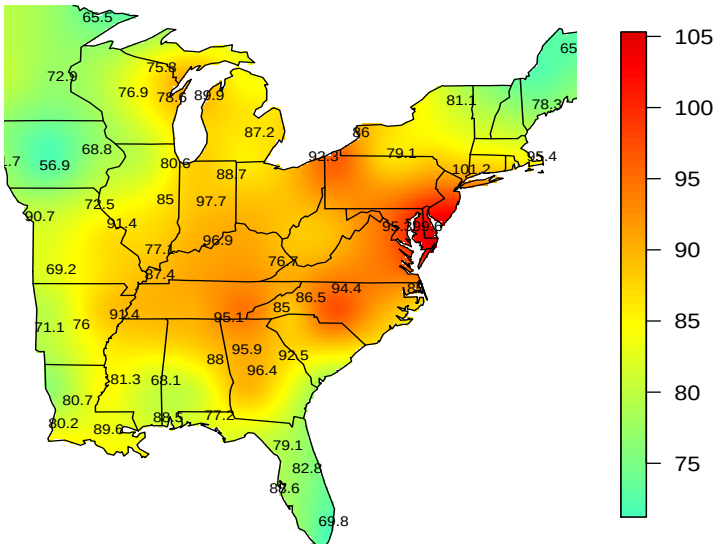
We study the relative percentage trends both for the meteorologically adjusted and the unadjusted levels in Figure 6.15 for 1997-2006. We observe that for most locations the trends are negatively significant.

The model based predictive maps of the probability that the 3-year rolling average of the annual 4th highest maximum ozone level is greater than 85 ppb, i.e. non-compliance with respect to the primary ozone standard, are provided in Figures 6.16 to 6.19. The plots show that many areas were out of compliance in the earlier years 1999-2003. However, starting in 2004 most areas started to comply with the primary ozone standard and except for some areas covering the New York City area and the north-east corner of the state of Ohio.

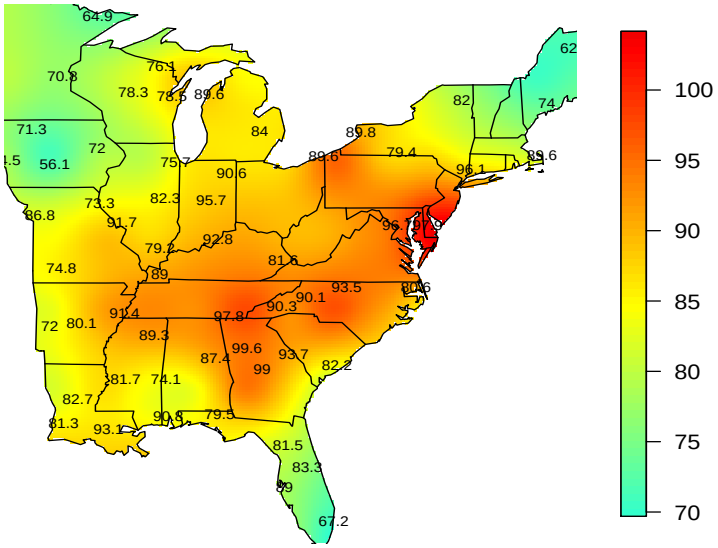
6.9 Summary

A fast hierarchical Bayesian auto-regressive model for both spatially and temporally rich data sets has been developed in this chapter. The methods have been shown to be accurate and feasible for simultaneous modelling and analysis of a large data set with more than a million observations using computationally intensive MCMC sampling algorithms. The hierarchical auto-regressive models have been shown to validate well for completely out of sample predictions.

Specifically, the methods have been illustrated for evaluating meteorologically adjusted trends in the primary ozone standard in the eastern US over a 10 year period from 1997-2006. To our knowledge no such Bayesian model based analysis exists for the same data and the same modelling purposes. An important utility of the high resolution space-time model lies in the ability to predict the primary ozone standard at any given location for the modelled period. This helps in understanding spatial patterns in ozone levels and trends both at the meteorologically adjusted and unadjusted levels which in turn will help evaluating the industrial emission reduction policies.

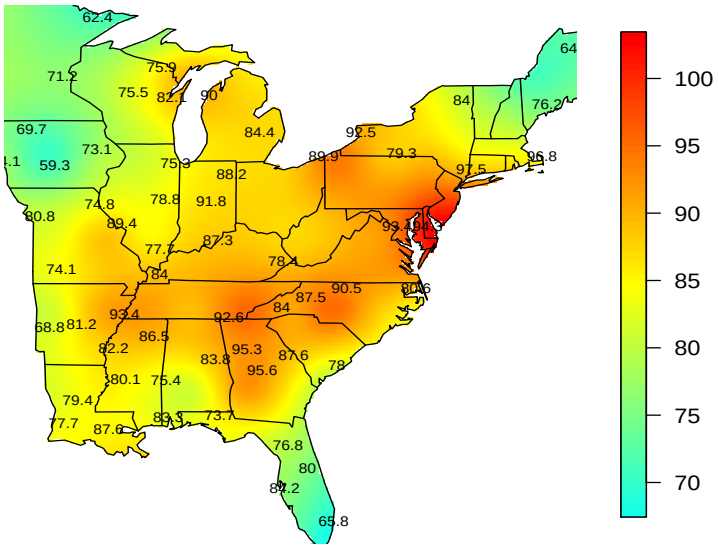


(a)

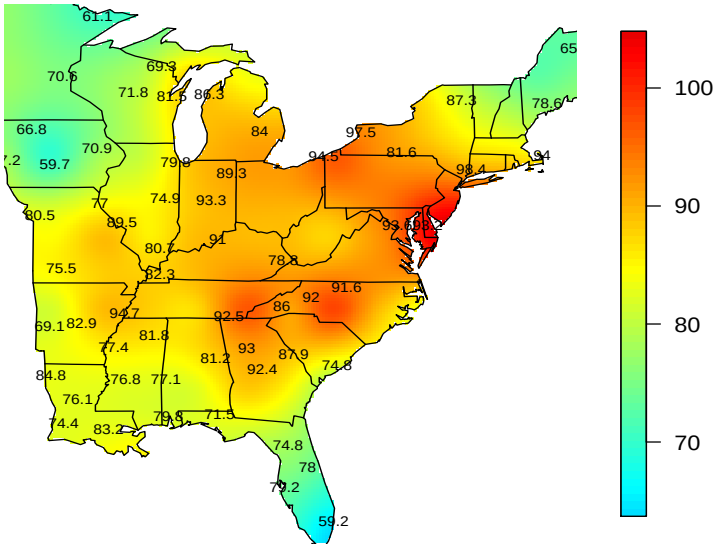


(b)

Figure 6.11: Model based interpolation of the 3-year rolling average of the annual 4th highest maximum ozone levels for 8 years panel (a) for 1999 and panel (b) for 2000. Observed data from a few selected sites, to enhance readability, are superimposed.

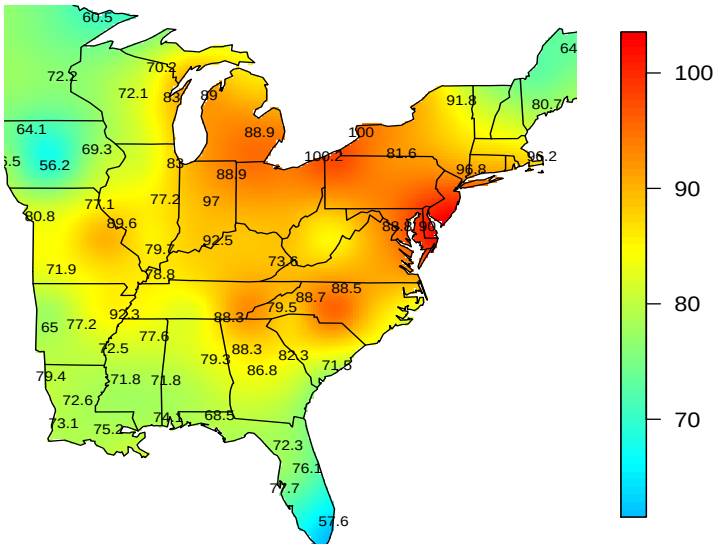


(a)

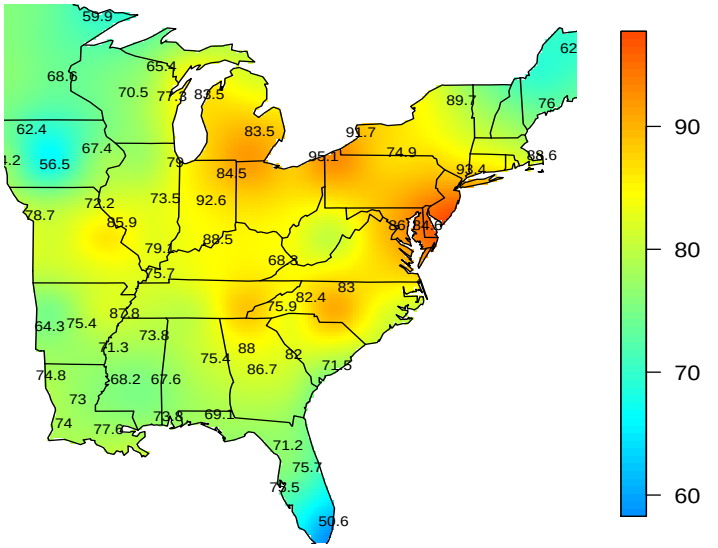


(b)

Figure 6.12: Model based interpolation of the 3-year rolling average of the annual 4th highest maximum ozone levels for 8 years panel (a) for 2001 and panel (b) for 2002. Observed data from a few selected sites, to enhance readability, are superimposed.

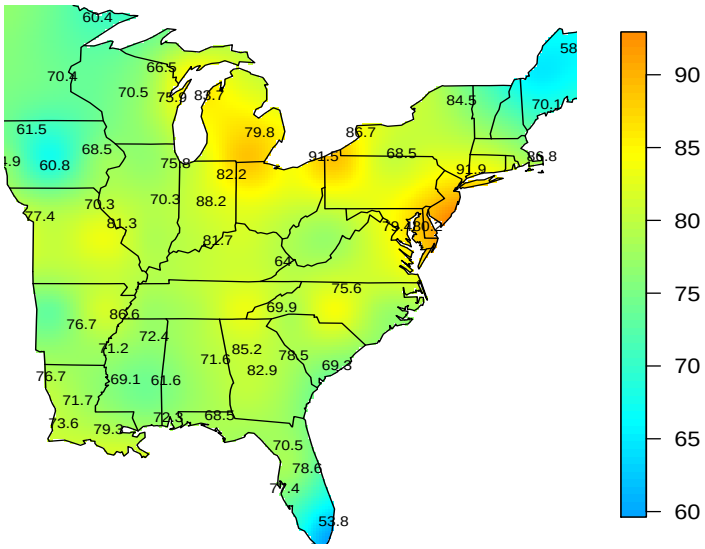


(a)

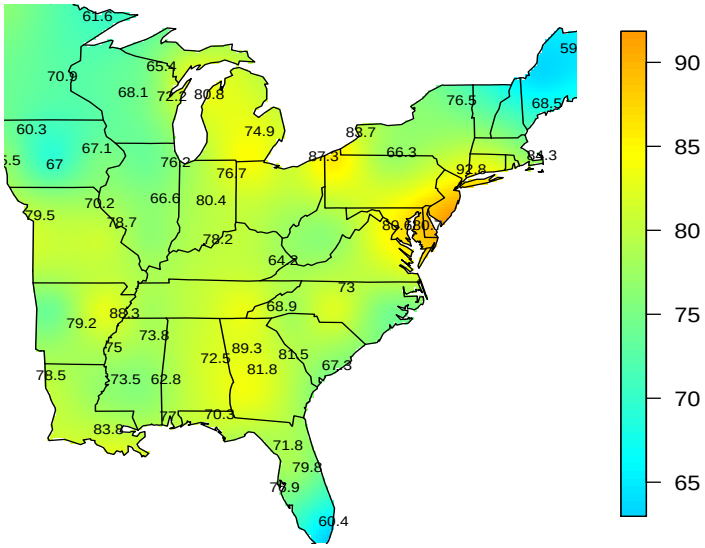


(b)

Figure 6.13: Model based interpolation of the 3-year rolling average of the annual 4th highest maximum ozone levels for 8 years panel (a) for 2003 and panel (b) for 2004. Observed data from a few selected sites, to enhance readability, are superimposed.

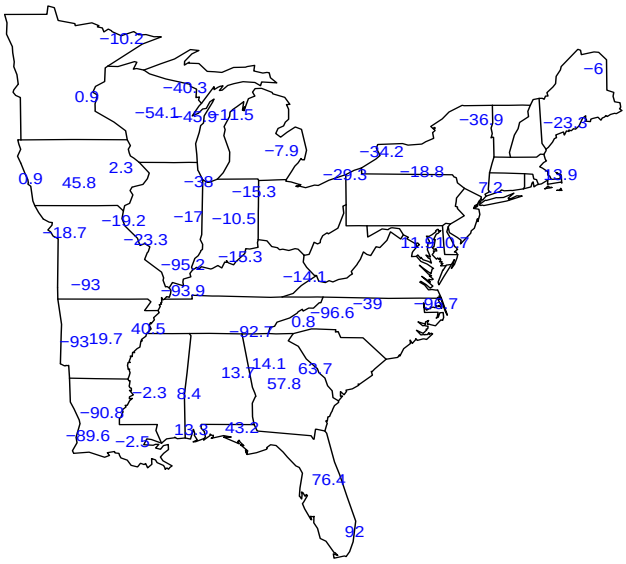


(a)

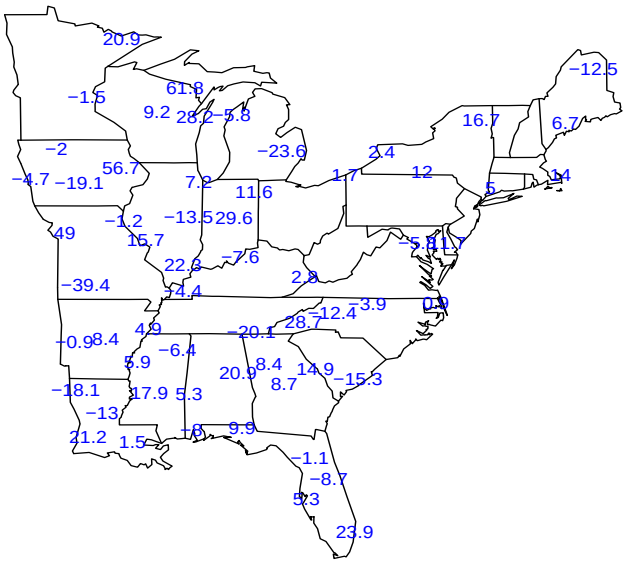


(b)

Figure 6.14: Model based interpolation of the 3-year rolling average of the annual 4th highest maximum ozone levels for 8 years panel (a) for 2005 and panel (b) for 2006. Observed data from a few selected sites, to enhance readability, are superimposed.

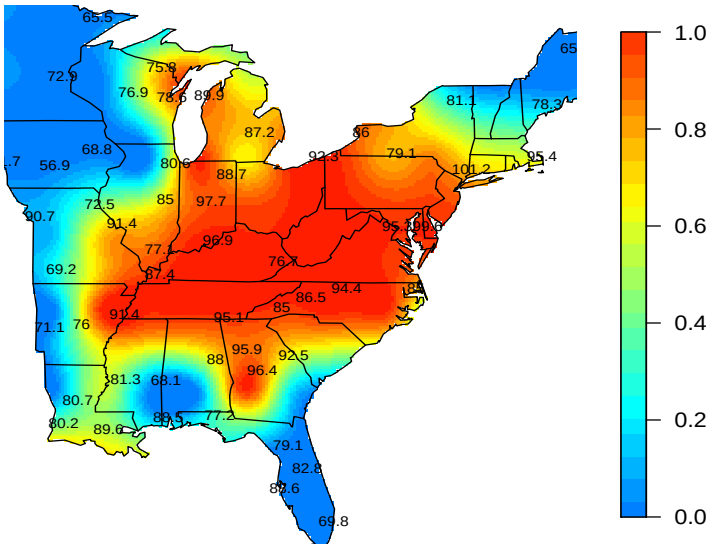


(a)

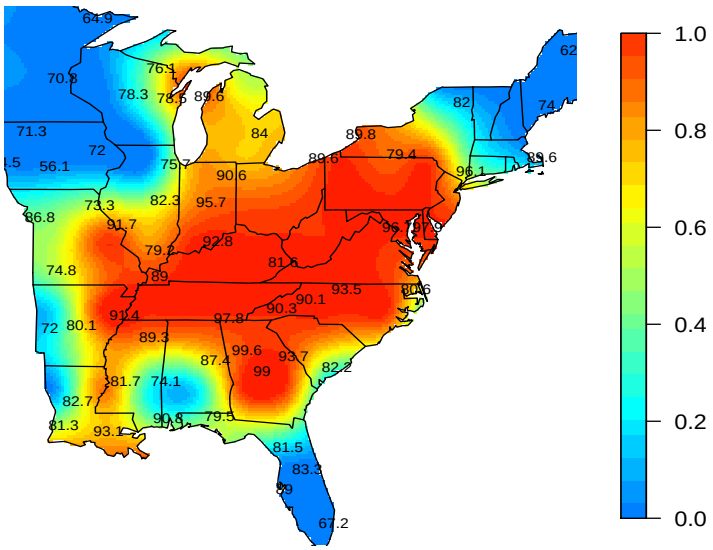


(b)

Figure 6.15: Plots of the relative percentage change between years 1997 and 2006: (a) Meteorologically adjusted and (b) unadjusted.

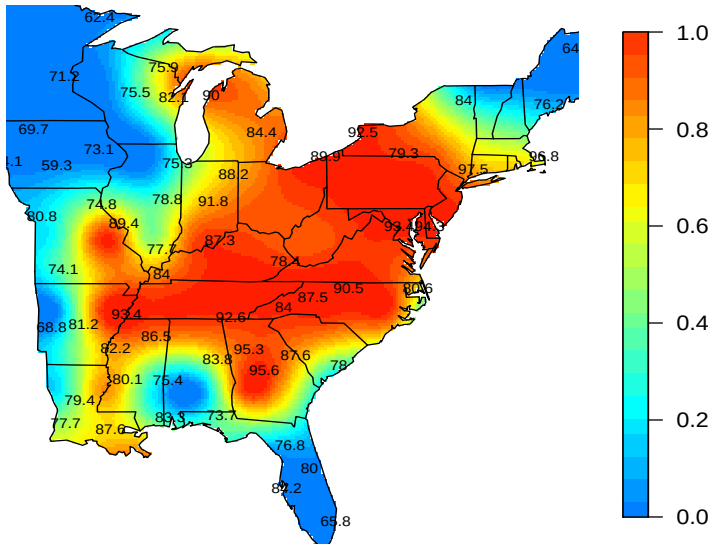


(a)

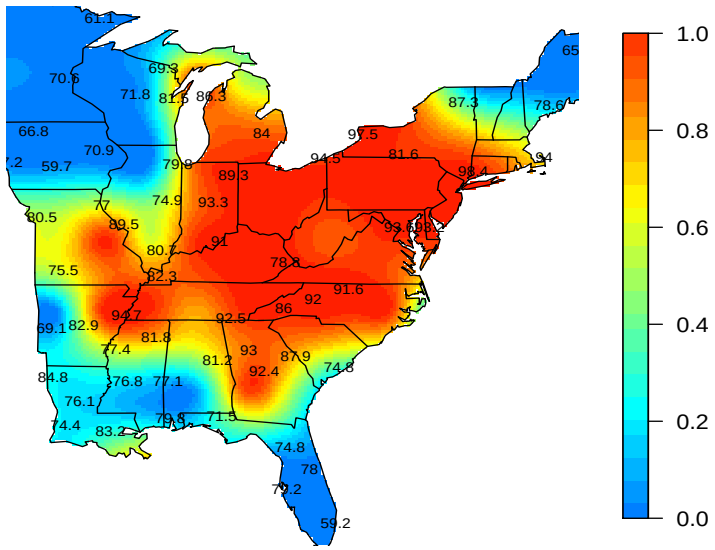


(b)

Figure 6.16: Model based interpolated maps of the probability that the 3-year rolling average of the annual 4th highest maximum ozone level is greater than 85 ppb for the years 1999 panel (a) and 2000 panel (b). Observed 3-year averages from a few selected sites, to enhance readability, are superimposed.

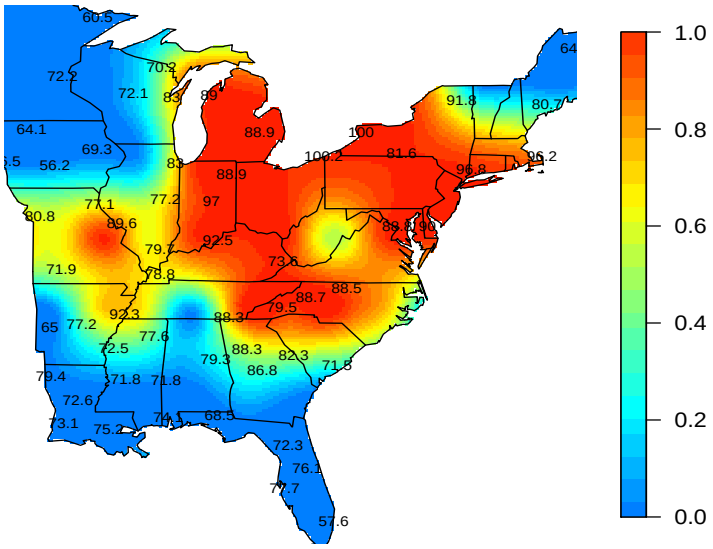


(a)

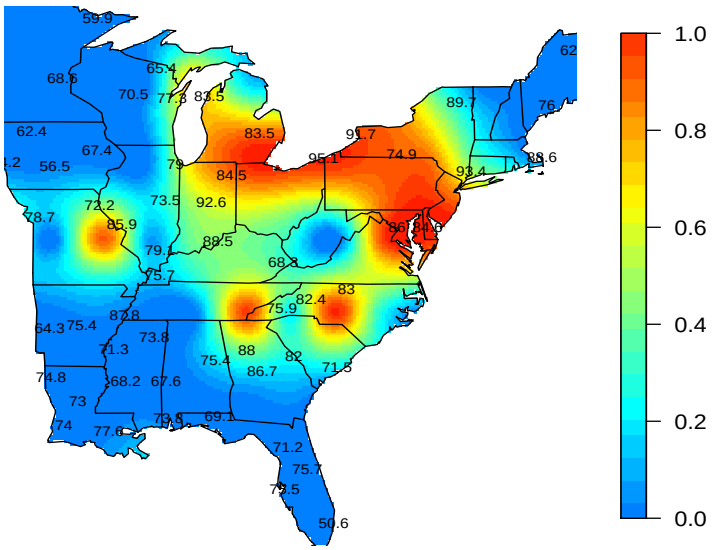


(b)

Figure 6.17: Model based interpolated maps of the probability that the 3-year rolling average of the annual 4th highest maximum ozone level is greater than 85 ppb for the years 2001 panel (a) to 2002 panel (b). Observed 3-year averages from a few selected sites, to enhance readability, are superimposed.

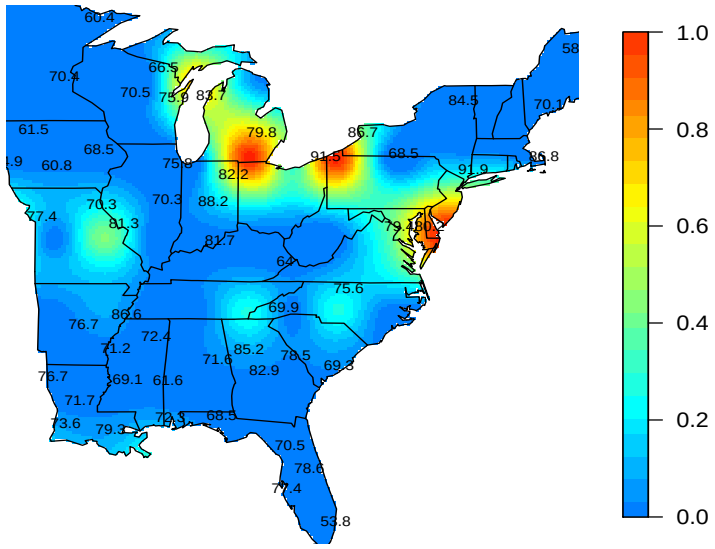


(a)

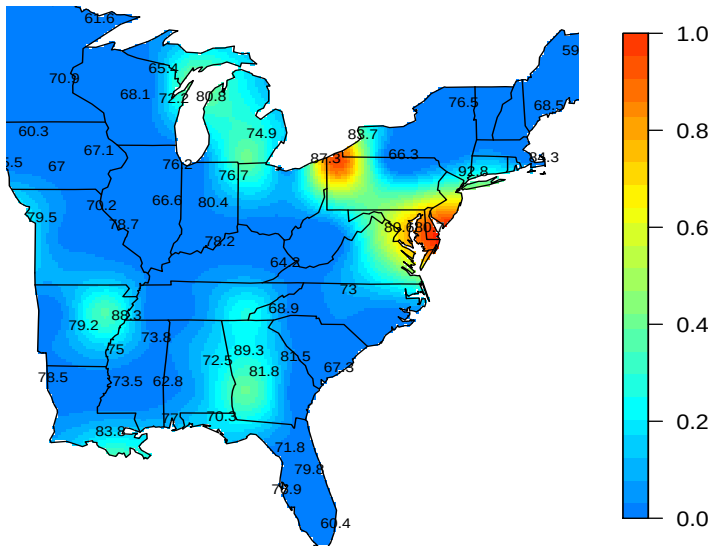


(b)

Figure 6.18: Model based interpolated maps of the probability that the 3-year rolling average of the annual 4th highest maximum ozone level is greater than 85 ppb for the years 2003 panel (a) and 2004 panel (b). Observed 3-year averages from a few selected sites, to enhance readability, are superimposed.



(a)



(b)

Figure 6.19: Model based interpolated maps of the probability that the 3-year rolling average of the annual 4th highest maximum ozone level is greater than 85 ppb for the years 2005 panel (a) and 2006 panel (b). Observed 3-year averages from a few selected sites, to enhance readability, are superimposed.

Chapter 7

Forecasting of the Daily Eight-Hour Maximum Ozone Levels

7.1 Introduction

In this chapter we use forecasting methods to estimate values of ozone levels at future time points. Forecasting can be done at any spatial location where there is no monitoring station, and also at any ozone monitoring site. We develop the forecasting methods using a number of spatio-temporal models introduced in the previous chapters. The spatio-temporal Gaussian process (GP) linear regression models (Section 3.4.3), the dynamic linear models (DLM) in Section 3.4.4, and the auto-regressive (AR) models discussed in Section 3.4.5, are well suited for forecasting ozone concentrations (Huerta *et al.*, 2004; Sahu *et al.*, 2009). However, these are very expensive computationally and are sometimes infeasible when the number of monitoring sites are large. This is also known as the *big-n problem* as discussed in Section 1.6. To solve the *big-n problem* we have developed a spatio-temporal AR modelling strategy based on Gaussian predictive processes (GPP) approximation that has been discussed earlier in Chapter 6. In this chapter, we illustrate the GPP based model for forecasting large dimensional ozone data obtained from the eastern US and compare forecast performance with the other models discussed in this thesis.

The rest of this chapter is organised as follows: Section 7.2 describes the

forecasting methods for the spatio-temporal models. A comparison of the forecasting performances is provided in Section 7.3 for a smaller data set consisting of monitoring data from four states in the eastern US. In Section 7.4 we illustrate the GPP based models for forecasting with sensitivity analyses. Section 7.5 analyses the full eastern US data and obtains the next days forecasts using the GPP approximation model. Finally in Section 7.6 we provide a number of conclusions.

7.2 Forecasting Methods

7.2.1 Forecasting using GP Models

Recall that the spatio-temporal Gaussian process (GP) models as discussed in Section 3.4.3 are given by:

$$Z_l(\mathbf{s}_i, t) = O_l(\mathbf{s}_i, t) + \epsilon_l(\mathbf{s}_i, t), \tag{7.1}$$

$$O_l(\mathbf{s}_i, t) = \mathbf{x}_l(\mathbf{s}_i, t)\boldsymbol{\beta} + \eta_l(\mathbf{s}_i, t) \tag{7.2}$$

where, $Z_l(\mathbf{s}_i, t)$ and $O_l(\mathbf{s}_i, t)$ are the observed and true values at site $\mathbf{s}_i$, day t and year l respectively, $i = 1, ..., n$, $t = 1, ..., T$ and $l = 1, ..., r$. The term $\mathbf{x}_l(\mathbf{s}_i, t)$ is the $1 \times p$ vector of co-variates, and $\boldsymbol{\beta}$ contains the unknown regression coefficients. $\epsilon_l(\mathbf{s}_i, t)$ is the pure error processes and $\eta_l(\mathbf{s}_i, t)$ is the spatial random effect. Details are given in Section 3.4.3.

We obtain one step ahead forecast distribution $Z_l(\mathbf{s}', T+1)$ at any unobserved location $\mathbf{s}'$ at time $T + 1$ as:

$$\begin{aligned} Z_l(\mathbf{s}', T+1) &= O_l(\mathbf{s}', T+1) + \epsilon_l(\mathbf{s}', T+1), \\ O_l(\mathbf{s}', T+1) &= \mathbf{x}_l(\mathbf{s}', T+1)\boldsymbol{\beta} + \eta_l(\mathbf{s}', T+1) \end{aligned}$$

For the MCMC algorithm samples are drawn from the forecast distribution by composition. We first obtain the joint distribution of $\mathbf{O}_{lT+1}$, which is given by the distribution $N(\mathbf{x}_{lT+1}\boldsymbol{\beta}, \Sigma_\eta)$. Then we obtain the conditional distribution of $\pi(O_l(\mathbf{s}', T+1)|\mathbf{O}_{lT+1})$, which is normally distributed with mean

$$\mathbf{x}_l(\mathbf{s}', T+1)\boldsymbol{\beta} + \Sigma_{\eta 12}\Sigma_\eta^{-1}(\mathbf{O}_{lT+1} - \mathbf{x}_{lT+1}\boldsymbol{\beta})$$

and variance

$$\sigma_\eta^2(1 - \Sigma_{\eta,12}\Sigma_\eta^{-1}\Sigma_{\eta,21})$$

Thus, at each MCMC iteration we draw $\mathbf{O}_{lT+1}$. Then we draw $O_l^{(j)}(\mathbf{s}', T+1)$ conditionally given $\mathbf{O}_{lT+1}$, and finally draw $Z_l^{(j)}(\mathbf{s}', T+1)$ from $N(O_l^{(j)}(\mathbf{s}', T+1), \sigma_\epsilon^{2(j)})$, where $j \geq 1$ is the iteration index.

Now for forecasting at any observed site $\mathbf{s}_i$ at time $T+1$ we obtain:

$$\begin{aligned} Z_l(\mathbf{s}_i, T+1) &= O_l(\mathbf{s}_i, T+1) + \epsilon_l(\mathbf{s}_i, T+1), \\ O_l(\mathbf{s}_i, T+1) &= \mathbf{x}_l(\mathbf{s}_i, T+1)\boldsymbol{\beta} + \eta_l(\mathbf{s}_i, T+1) \end{aligned}$$

Thus at each iteration we draw a forecast iterate $Z_l^{(j)}(\mathbf{s}_i, T+1)$ from the normal distribution with mean $O_l^{(j)}(\mathbf{s}_i, T+1) = \mathbf{x}_l(\mathbf{s}_i, T+1)\boldsymbol{\beta}^{(j)} + \eta_l^{(j)}(\mathbf{s}_i, T+1)$ and variance $\sigma_\epsilon^{2(j)}$, where $j \geq 1$.

7.2.2 Forecasting using DLM

The spatio-temporal DLM has been described in Section 3.4.4. We recall the model as:

$$\begin{aligned} Z_l(\mathbf{s}_i, t) &= \mathbf{x}_l(\mathbf{s}_i, t)\boldsymbol{\theta}_t + \nu_l(\mathbf{s}_i, t), \\ \boldsymbol{\theta}_t &= \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad t \geq 1, \end{aligned}$$

The model terms have been defined earlier. We also obtain the initial information $\boldsymbol{\theta}_0$ from $N(\boldsymbol{\mu}, \sigma_\theta^2\mathbf{I})$. The posterior distribution of the state parameter at time point T is obtained as $\boldsymbol{\theta}_T|. \sim N(\boldsymbol{\mu}_{T-1}, \sigma_\omega^2\mathbf{I})$. Hence, one step ahead forecast distribution at new site $\mathbf{s}'$ at time $T+1$ is written as:

$$\begin{aligned} Z_l(\mathbf{s}', T+1) &= \mathbf{x}_l(\mathbf{s}', T+1)\boldsymbol{\theta}_{T+1} + \nu_l(\mathbf{s}', T+1) \\ \boldsymbol{\theta}_{T+1} &= \boldsymbol{\theta}_T + \boldsymbol{\omega}_{T+1}, \quad t \geq 1, \end{aligned}$$

Thus we get,

$$\pi(Z_l(\mathbf{s}', T+1)|\boldsymbol{\theta}_T, \sigma_\nu^2, \sigma_\omega^2, \phi) \sim N(\mathbf{x}_l(\mathbf{s}', T+1)\boldsymbol{\mu}_T, \sigma_\omega^2(\mathbf{x}_{lT+1}\mathbf{x}'_{lT+1}) + \Sigma_\nu)$$

where, $\boldsymbol{\mu}_T$ and σ^2_ω are the mean and variances of $\boldsymbol{\theta}$ at time $T+1$. That is, at each iteration j we draw $Z_l^{(j)}(\mathbf{s}', T+1)$ from the above distribution. Similarly, forecasting at the observation sites we draw samples for $Z_l^{(j)}(\mathbf{s}_i, T+1)$ at each iteration $j \geq 1$.

7.2.3 Forecasting using AR Models

Recall from Section 3.4.5, we write the AR models as:

$$\begin{aligned} Z_l(\mathbf{s}_i, t) &= O_l(\mathbf{s}_i, t) + \epsilon_l(\mathbf{s}_i, t) \\ O_l(\mathbf{s}_i, t) &= \rho O_l(\mathbf{s}_i, t-1) + \mathbf{x}_l(\mathbf{s}_i, t)\boldsymbol{\beta} + \eta_l(\mathbf{s}_i, t) \end{aligned}$$

The terms of the equations are defined in Section 3.4.5. In the AR models the predictive distribution of $Z(\mathbf{s}', T+1)$ is determined by the true forecast value $O_l(\mathbf{s}', T+1)$. Thus according to (3.10) we simulate $O_l(\mathbf{s}', T+1)$ from marginal distribution with mean given by $\rho O_l(\mathbf{s}', T) + \mathbf{x}_l(\mathbf{s}', T+1)\boldsymbol{\beta}$ with site invariant variance σ^2_η. We use marginal distribution instead of conditional distribution because we already obtain the conditional distribution given observed information upto time T and r at the monitoring sites $\mathbf{s}_1, ..., \mathbf{s}_n$, and at the future time $T+1$ there is no new available information to condition on except for the new regressor values $\mathbf{x}_l(\mathbf{s}', T+1)$ in the model.

Thus, in each iteration j, we obtain the forecast of $Z_l^{(j)}(\mathbf{s}', T+1)$ with mean $O_l^{(j)}(\mathbf{s}', T+1) = \rho^{(j)} O_l^{(j)}(\mathbf{s}', T) + \mathbf{x}_l(\mathbf{s}', T+1)\boldsymbol{\beta}^{(j)}$ and variance $\sigma^{2(j)}_\epsilon$, $j \geq 1$. Henceforth, for forecasting at the observed locations $\mathbf{s}_i$ we need to draw samples for $Z_l^{(j)}(\mathbf{s}_i, T+1)$ following similar steps discussed earlier.

7.2.4 Forecasting using Models Based on GPP Approximations

Recall from Section 6.2, we write the AR models based on GPP approximation as:

$$Z_l(\mathbf{s}_i, t) = \mathbf{x}_l(\mathbf{s}_i, t)\boldsymbol{\beta} + \tilde{\eta}_l(\mathbf{s}_i, t) + \epsilon_l(\mathbf{s}_i, t), \tag{7.3}$$

$$\tilde{\eta}_l(\mathbf{s}_i, t) = A w_l(\mathbf{s}_i, t) \tag{7.4}$$

$$w_l(\mathbf{s}_i, t) = \rho w_l(\mathbf{s}_i, t-1) + \xi_l(\mathbf{s}_i, t) \tag{7.5}$$

The notations are defined in Section 6.2. Under a Bayesian hierarchical setup, we use the MCMC algorithm to obtain estimates of the model parameters (see Chapter 3).

At an unobserved location $\mathbf{s}'$, the one step ahead Bayesian forecast is given by the predictive distribution of $Z_l(\mathbf{s}', T+1)$, that we determine form the equation (6.8) replacing t with $T+1$. Thus the one step ahead forecast distribution has the mean

$$\mathbf{x}_l(\mathbf{s}', T+1)'\boldsymbol{\beta} + \tilde{\eta}_l(\mathbf{s}', T+1)$$

and variance σ^2_ϵ when all the parameters are known. We also obtain $\tilde{\eta}_l(\mathbf{s}', T+1)$ from (6.9) as:

$$\tilde{\eta}_l(\mathbf{s}', T+1) = \mathbf{c}'(\mathbf{s}')S_w^{-1}\mathbf{w}_{lT+1}$$

where we get $\mathbf{w}_{lT+1}$ from (7.5). Thus, at each MCMC iteration we draw a forecast value $Z_l^{(j)}(\mathbf{s}', T+1)$ from the normal distribution with mean $\mathbf{X}_l(\mathbf{s}', T+1)\boldsymbol{\beta}^{(j)} + \tilde{\eta}_l^{(j)}(\mathbf{s}', T+1)$ and variance $\sigma_\epsilon^{2(j)}$, $j \geq 1$. We obtain the forecasts $Z_l^{(j)}(\mathbf{s}_i, T+1)$ at the observation location $\mathbf{s}_i$, $i = 1, ..., n$, similarly.

In the following sections, models based on GPP approximation are used to obtain forecasts and these forecasts are compared with the ones obtained from the GP, DLM and the AR models. We also obtain forecast results for both fitting and validation sites using the models.

In all our illustrations below we have diagnosed MCMC convergence by visual examination of the time-series plots of the MCMC iterates. We also used multiple parallel runs and calculated the Gelman and Rubin statistics (Gelman and Rubin 1992), which we found to be satisfactory. We have used 15,000 iterations to make inference after discarding first 5000 iterates to mitigate the effect of initial values.

7.3 Comparison of the Forecast Models

7.3.1 Example: Four States Data Set

In this section the GPP approximation model is compared with the GP, DLM and AR models for forecasting (see Section 7.2). Similar to Section 6.6, a smaller subset of the whole eastern US data consisting of four states, Illinois, Indiana, Ohio and Kentucky (see Figure 7.1) is used to facilitate the comparison. We have 147 ozone monitoring sites inside these four states during 14th June to 8th

July, 2010, to illustrate the forecasts. We set aside data from 20 randomly chosen locations for the forecast validation. This choice is also repeated 7 times in an experiment to observe the effect of the choice of these validation sites. For the GPP based models knot size is taken as 107, that has been chosen from the sensitivity analysis we have performed in Section 7.4.1. We use 7 and 14 consecutive days observation from 1 July to 7 July and 24 June to 7 July and obtain forecast on 8th July 2010. Different model validation criteria (see Section 3.3) are used to compare the models and also for performing sensitivity analysis.

The CMAQ grid output (see Section 1.4) in the eastern US for this period is also used in the model for ozone data. The CMAQ output values are used as a covariate in the ozone model in the same fashion as we have discussed previously in Chapter 5. This modelling technique is also known as the downscaling method, see Berrocal *et al.* (2010a, 2010b); and also as data assimilation method as discussed in Section 1.4.3.

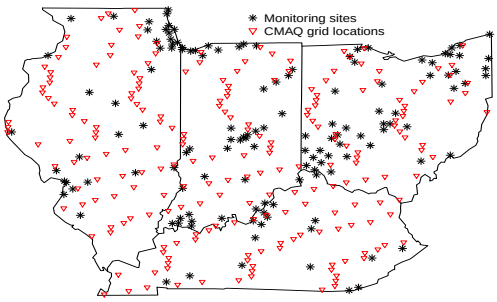


Figure 7.1: A map of the four states, Ohio, Indiana, Illinois and Kentucky. 147 ozone monitoring locations are superimposed.

7.3.2 Comparison Results

In this section we compare the models using both the predictive model choice criteria (PMCC) and the forecast validation criteria. It is observed from Table 7.1 that forecasting using GPP approximation model is the best. The PMCC values provided in Table 7.2 also confirm that the GPP based model is the best among all the spatio-temporal models that we consider here.

Table 7.3 provides the nominal coverage for the 95% intervals for the forecasts of ozone levels at the hold-out sites. We observe that GP and GPP based approximation models show coverage probability near 95%, whereas the DLM has the smallest nominal coverage and AR has a little bit higher nominal coverage than 95%.

	Forecast Validation					
	7 Days Data			14 Days Data		
Models	RMSE	MAE	rMSEP	RMSE	MAE	rMSEP
GP	12.85	10.75	0.82	12.78	10.70	0.82
DLM	12.10	10.06	0.78	12.05	10.01	0.77
AR	10.58	8.65	0.70	10.55	8.64	0.70
GPP	9.10	7.04	0.65	9.06	7.00	0.65

Table 7.1: Values of the forecast validation criteria for the GP, the DLM, the AR, and the models based on GPP approximations.

	7 Days Data			14 Days Data		
Models	GoF	P	PMCC	GoF	P	PMCC
GP	1388.43	1704.65	3093.08	2375.78	2986.44	5362.22
DLM	1150.98	1644.56	2795.54	2083.95	2865.88	4949.83
AR	796.45	1857.64	2654.09	1409.68	3271.02	4680.70
GPP	865.55	1074.86	1940.41	1749.09	2113.3	3862.39

Table 7.2: Values of the PMCC for the GP, DLM, AR, and the models based on GPP approximations. Here, GoF is the goodness of fit and P is the penalty.

	Nominal coverage (95% intervals) for the Hold-out Sites							
	Using 7 Days Data				Using 14 Days Data			
Day	GP	DLM	AR	GPP	GP	DLM	AR	GPP
08/07	94.44	84.10	97.42	93.95	94.65	85.76	97.81	94.55

Table 7.3: Nominal coverage of the 95% intervals for the one-step ahead forecasts at the 20 randomly chosen validation sites.

Finally, we conclude that the GPP based model is the best among all the models we have considered here. Henceforth, in the following sections, we only use the GPP based approximation model and study its sensitivity for forecasting.

7.4 Sensitivity Analysis of the Forecasts Based on GPP Models

7.4.1 Sensitivity of Knot Sizes

Similar to Section 6.6, we define five different sets of regular grid locations starting from 6×6, 8×8, 10×10, 12×12 and 14×14 over the four states. These grid sizes lead to knot-sizes of 26, 40, 60, 107 and 138 respectively that are inside the boundary of the four states.

For different set of knot sizes and for 7 & 14 consecutive days data the values of the forecast validation criteria for the forecasts made on 8 July 2010, are given in Table 7.4. As expected, the RMSE decreases as the knot size increases and we observe that the differences between the validation criteria for knot sizes 107 and 138 are very small. Thus, similar to the spatial interpolation results in Section 6.6.1 we observe that after a particular choice of knot size the forecasting performance of the models are approximately same.

	Forecast Validation					
	7 Days Data			14 Days Data		
Knots	RMSE	MAE	rMSEP	RMSE	MAE	rMSEP
26	10.11	8.35	0.67	10.08	8.30	0.66
40	9.75	7.82	0.66	9.70	7.79	0.66
60	9.48	7.36	0.66	9.47	7.36	0.65
107	9.10	7.04	0.65	9.06	7.01	0.65
138	9.08	7.02	0.65	9.06	7.00	0.65

Table 7.4: Values of the forecast validation criteria for different knot sizes for the 7 and 14 days data in the four states example on 8 July, 2010.

7.4.2 Sensitivity of Prior Selection

In this section, different hyper-parameter values of the prior distributions are used for the GPP based models with knot size 107. Table 7.5 shows different forecast validation criteria for four different sets of values of a and b, the hyper-parameters of the gamma prior distribution for the variance components. The validation criteria values are not very sensitive to the choice of the a and b and the combination $a = 2$ and $b = 1$ provides the best results. Henceforth, this choice will be adopted in our analysis.

	Forecast Validation					
	7 Days Data			14 Days Data		
(a,b)	RMSE	MAE	rMSEP	RMSE	MAE	rMSEP
(2,1)	9.10	7.04	0.65	9.06	7.00	0.65
(1,1)	9.11	7.10	0.66	9.10	7.08	0.65
(2,2)	9.15	7.18	0.66	9.13	7.16	0.66
(10,10)	9.35	7.34	0.68	9.30	7.29	0.67

Table 7.5: Values of the forecast validation criteria for different hyper-parameter values for the GPP based models fitted to 7 and 14 days data from the four states. The forecasts are made for 8th July 2010 in both the model fitting cases.

7.4.3 Choice of the Sampling Method for the Spatial Decay Parameter

Here we compare the Metropolis-Hastings sampling method for the spatial decay parameter, ϕ with the discrete sampling method corresponding to the assumption of a discrete prior for ϕ, see Section 3.2. We tune the variance of the proposal distribution for the Metropolis-Hastings to achieve an acceptance rate of 29.85%. In the case of the discrete sampling for ϕ we assume the discrete uniform prior distribution on 50 equally spaced values in the interval 0.001 to 0.1. Table 7.6 represents the results from these two sampling schemes, where better performance is observed for the random walk approach.

	Forecast Validation					
	7 Days Data			14 Days Data		
	RMSE	MAE	rMSEP	RMSE	MAE	rMSEP
Discrete	9.14	7.17	0.68	9.13	7.15	0.68
Continuous	9.10	7.04	0.65	9.06	7.00	0.65

Table 7.6: Values of the forecast validation criteria for different sampling approaches of ϕ for the 7 and 14 days data in the four states example on 8 July 2010.

7.4.4 Results for Different Sets of Hold-Out Sites

We randomly choose 7 hold-out data sets, each of which consists of ozone concentrations from 20 monitoring sites. From Table 7.7 we observe that when the knot size is 107 the RMSE varies between 9.10 and 11.50 for forecasting made using the 7-days data. For forecasting using 14 days data the RMSE varies be-

tween 9.06 and 11.33. We also observe that the nominal coverage varies between 93.15% to 96.20%, showing very good accuracy of the forecasts.

	Forecast Validation					
	7 Days Data			14 Days Data		
Data set	RMSE	MAE	rMSEP	RMSE	MAE	rMSEP
1	9.10	7.04	0.65	9.06	7.00	0.65
2	10.98	8.88	0.72	10.85	8.82	0.72
3	11.50	9.94	0.74	11.33	9.68	0.74
4	9.76	7.85	0.68	9.72	7.80	0.68
5	9.88	7.96	0.68	9.85	7.88	0.67
6	9.15	7.19	0.66	9.14	7.19	0.66
7	9.34	7.30	0.66	9.32	7.28	0.65

Table 7.7: Values of the forecast validation criteria for different sets of hold-out sites using 107 knots.

7.5 Analysis of the Full Eastern US Data

In this section we evaluate the forecasting performance using daily ozone concentration data for the three week study period from 23 June to 14 July, 2010. Data are available from 639 ozone monitoring sites in the eastern US and we use data from 577 sites to fit our models and the data from 62 sites are set aside for validation purposes. Details of the ozone concentration data with summary statistics are given in Section 4.6.

As in Section 7.3.1, we use a running window of data for 7 and 14 days observations to fit the models and provide the one step ahead forecasts. For example, 7 days data, say from 1 July to 7 July are used to forecast for 8th July. Similarly when the models are fitted using 14 days data, say from 24 June to 7 July the forecasts are made for the ozone concentration levels on the 8th of July. This is done to see the performance of forecasts models for the same day based on 7 and 14 days data.

Similar to the illustration given in Section 7.3.1, we include CMAQ grid output as a covariate in the models. As expected, we see from Figure 7.2 that there are similarities in the patterns between the observed and CMAQ gird output, however in some cases, the CMAQ output over estimates the actual observations in the eastern US study region.

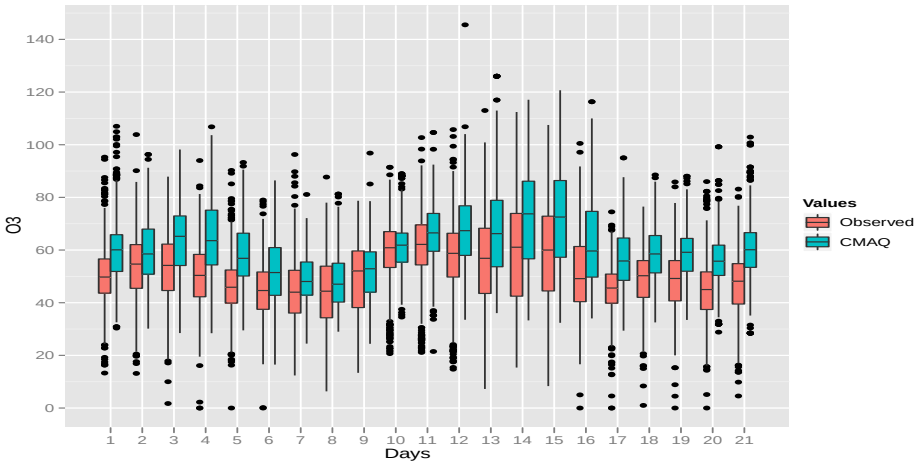


Figure 7.2: Box-plot for the observed and CMAQ grid output for 21 days from all 639 sites in the eastern US.

In this section we only use the GPP based model, because the other models are not suitable for analysing large dimensional data set (see Chapter 6). The forecasts made by the GPP based model are then compared with the CMAQ forecasts as discussed in Section 1.4.

7.5.1 Knot Size Selection

As seen previously in Section 7.4 increase in the number of knot locations yields better forecasting for the GPP approximation models. This is also confirmed here by the results presented in Table 7.8.

The number of knots used in Section 6.8 for the predictions are also considered here starting from 68, 105, 156, and 269, that are inside the boundary of the eastern US study region. We observe the model performance is only marginally improved when knot size goes up to 269 from 156. Henceforth, we proceed with the knot size 156, similar to Section 6.8 which has a much smaller computational burden.

7.5.2 Parameter Estimates

In this section we discuss the parameter estimates for the GPP based model for the eastern US study region. Table 7.9 provides the MCMC summary statistics for the model parameters using 7 days data from 1 July–7 July. We observe that the CMAQ variable is a significant predictor since β_1 is significant because

Forecast Validation								
	7 Days Data				14 Days Data			
Knots	RMSE	MAE	rBIAS	rMSEP	RMSE	MAE	rBIAS	rMSEP
68	12.98	11.98	-0.11	0.82	12.85	11.78	-0.11	0.80
105	12.01	10.05	-0.10	0.78	12.01	10.04	-0.10	0.75
156	11.17	9.22	-0.09	0.65	11.11	9.18	-0.09	0.65
269	11.15	9.21	-0.09	0.65	11.10	9.18	-0.09	0.65

Table 7.8: Values of the forecast validation criteria for different knot sizes for the GPP based models fitted to 7 and 14 days data from the four states. The forecasts are made for 8th July 2010 in both the model fitting cases.

the 95% interval does not contain zero. The temporal correlation of the latent random effects is also statistically significant. The estimate of the spatial decay parameter is 0.0024, that corresponds to an effective range of 1250 kilometers. As expected, the estimates of the variance components show that the nugget effect has smaller variability than the spatial error variance σ_w^2. Similar parameter estimates are obtained when the model is fitted to other data sets, e.g., data from 2–8 July and so on, see Table 7.10. We omit other summary statistics of the model parameters for brevity.

Table 7.11 shows the estimates of the model parameters using 14 consecutive days observations. We observe that the estimates are approximately same as the estimates reported in Table 7.9 and 7.10 obtained using 7 days data sets.

	Mean	Median	sd	95% interval
β_0	4.3974	4.3984	0.1768	(4.0525, 4.7452)
β_1	0.3264	0.3263	0.0213	(0.2854, 0.3684)
ρ	0.2109	0.2108	0.0451	(0.1232, 0.2998)
σ_ϵ^2	0.2477	0.2476	0.0061	(0.2358, 0.2597)
σ_w^2	0.5291	0.5261	0.0574	(0.4271, 0.6503)
ϕ	0.0024	0.0024	0.0003	(0.0019, 0.0030)

Table 7.9: Parameter estimates for the proposed AR models based on GPP approximation, fitted with 7 days observations from 1 July–7 July, 2010.

7.5.3 Comparison with the CMAQ Output

This section is devoted to comparing the proposed models with the CMAQ forecasts. Table 7.12 represents the RMSE values (see details in Section 3.3) for the models based on GPP approximations. The RMSE obtained from CMAQ output

Days			Parameters					
Fitted	Forecast		β_0	β_1	ρ	σ^2_ϵ	σ^2_w	ϕ
2/7-8/7	9/7	Mean	4.34	0.33	0.20	0.26	0.49	0.0024
		sd	0.16	0.02	0.05	0.007	0.06	0.0004
3/7-9/7	10/7	Mean	4.18	0.35	0.20	0.26	0.49	0.0024
		sd	0.18	0.03	0.05	0.006	0.05	0.0003
4/7-10/7	11/7	Mean	4.15	0.35	0.19	0.30	0.46	0.0024
		sd	0.17	0.03	0.04	0.007	0.06	0.0003
5/7-11/7	12/7	Mean	4.14	0.34	0.19	0.30	0.45	0.0024
		sd	0.17	0.02	0.04	0.006	0.06	0.0004
6/7-12/7	13/7	Mean	4.06	0.35	0.17	0.30	0.45	0.0024
		sd	0.18	0.04	0.06	0.007	0.07	0.0004
7/7-13/7	14/7	Mean	4.12	0.33	0.16	0.31	0.46	0.0024
		sd	0.17	0.03	0.05	0.005	0.06	0.0003

Table 7.10: Parameter estimates (mean and sd) for the models based on GPP approximation fitted using 7 consecutive days observations.

Days			Parameters					
Fitted	Forecast		β_0	β_1	ρ	σ^2_ϵ	σ^2_w	ϕ
24/6-7/7	8/7	Mean	3.90	0.38	0.21	0.24	0.42	0.0025
		sd	0.19	0.02	0.08	0.008	0.06	0.0002
25/6-8/7	9/7	Mean	3.92	0.37	0.21	0.25	0.43	0.0025
		sd	0.15	0.03	0.07	0.006	0.05	0.0003
26/6-9/7	10/7	Mean	3.95	0.37	0.21	0.26	0.43	0.0025
		sd	0.21	0.02	0.08	0.006	0.06	0.0002
27/6-10/7	11/7	Mean	4.01	0.36	0.21	0.26	0.43	0.0025
		sd	0.20	0.03	0.07	0.009	0.07	0.0002
28/6-11/7	12/7	Mean	4.05	0.35	0.20	0.26	0.42	0.0025
		sd	0.19	0.03	0.09	0.006	0.05	0.0002
29/6-12/7	13/7	Mean	4.11	0.34	0.20	0.26	0.43	0.0025
		sd	0.18	0.03	0.08	0.007	0.05	0.0003
30/6-13/7	14/7	Mean	4.24	0.33	0.19	0.27	0.43	0.0025
		sd	0.22	0.04	0.09	0.009	0.09	0.0003

Table 7.11: Parameter estimates (mean and sd) for the models based on GPP approximation fitted with 14 days observations starting from 24 June to 13 July, 2010.

values are also given. Forecast validation results are also obtained for 7 and 14 consecutive days data sets starting from 23 June to 13 July 2010. These forecast results are obtained from both hold-out and fitted sites.

As expected, we observe much better performance of the GPP based models compared to the CMAQ model. The RMSE is smaller for the GPP approximation model in all forecast days (i.e., 8–14 July) than the CMAQ models. We also observe that the RMSEs are smaller for the GPP based model for both hold-out and fitted data. Table 7.12 also indicates that the RMSEs are smaller for the data sets fitted with 14 days data compared to 7 days data using the GPP based models. This is because 14 days data provides more information to model fitting than 7 days data.

		7 Days Data Set			
Fitted	Forecast	Hold-out Sites		Fitted Sites	
		GPP	CMAQ	GPP	CMAQ
1/7-7/7	8/7	11.17	20.52	11.09	17.42
2/7-8/7	9/7	10.79	19.68	10.63	16.02
3/7-9/7	10/7	8.59	16.36	8.96	15.27
4/7-10/7	11/7	8.18	15.51	9.00	13.46
5/7-11/7	12/7	8.67	13.12	9.16	13.71
6/7-12/7	13/7	11.24	20.36	11.18	16.29
7/7-13/7	14/7	9.21	18.10	11.01	17.56
		14 Days Data Set			
24/6-7/7	8/7	10.07	20.52	11.05	17.42
25/6-8/7	9/7	10.57	19.68	10.29	16.02
26/6-9/7	10/07	7.85	16.36	8.94	15.27
27/6-10/7	11/07	8.11	15.51	8.49	13.46
28/6-11/7	12/07	8.31	13.12	9.00	13.71
29/6-12/7	13/07	10.70	20.36	11.65	16.29
30/6-13/7	14/07	8.61	18.10	10.07	17.56

Table 7.12: Values of the forecast RMSE for the models based on GPP approximation and the CMAQ output in the hold-out and fitted sites.

The nominal coverages for the hold-out data are given in Table 7.13 for the GPP approximation models. We observe that these are close to 95% for all the days, which indicates that the uncertainties in the forecasts are about right.

Figure 7.3 shows the scatter plot of the forecasts against observed values for the hold-out locations. The proposed model shows better forecasting performance compared to the CMAQ models. We also observe from Table 7.14 that the CMAQ

	Nominal coverage (95% interval)						
	8/7	9/7	10/7	11/7	12/7	13/7	14/7
7 Days	93.55	93.75	94.96	95.16	94.96	93.75	95.56
14 Days	94.62	94.30	94.84	95.05	94.62	94.84	94.84

Table 7.13: Nominal coverage of the 95% intervals for the hold-out data for the models based on GPP approximations.

model over estimates the actual observations.

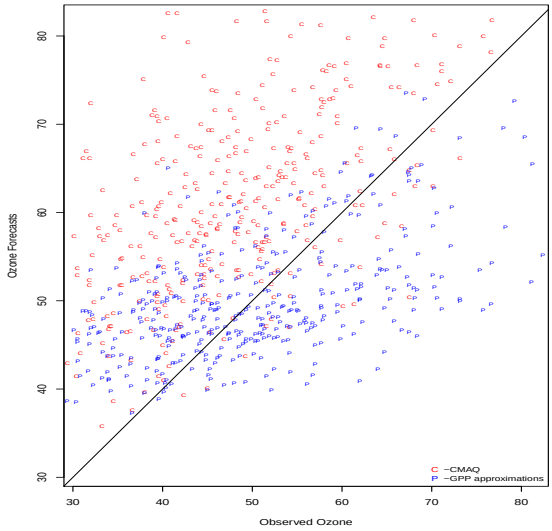
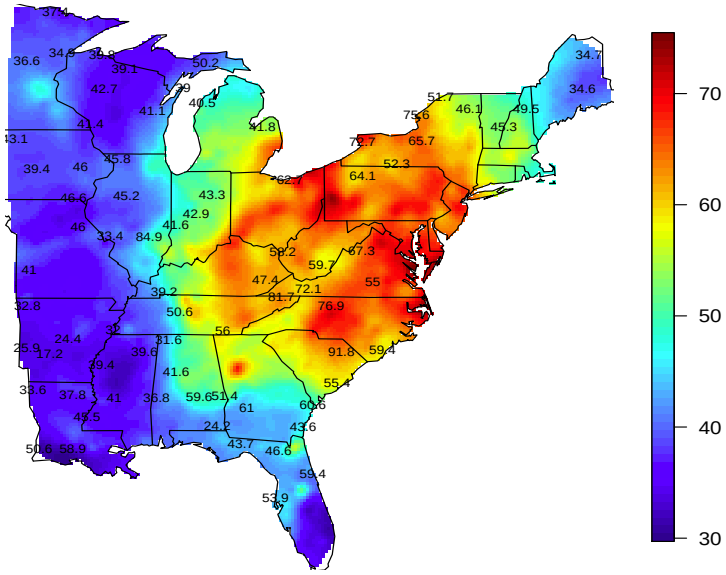


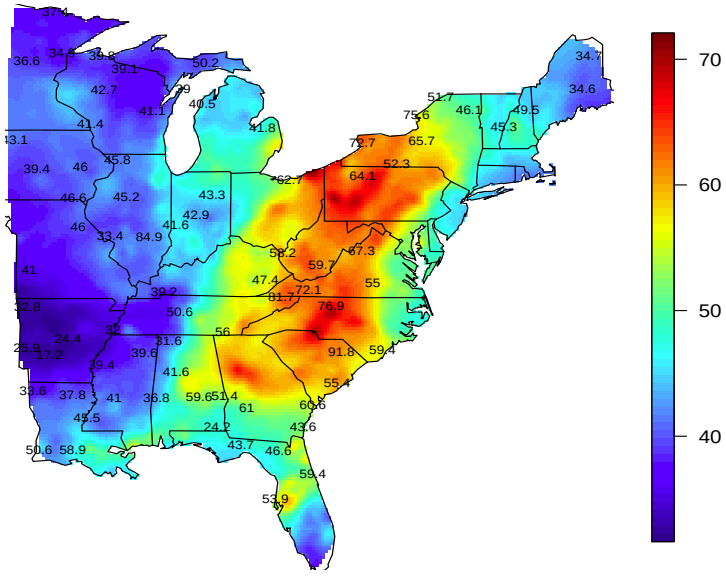
Figure 7.3: A scatter plot of forecasts against observations in the 62 hold-out sites. The symbols 'C' and 'P' represents the CMAQ output and the GPP based models respectively.

7.5.4 Forecast Maps

Figures 7.4 to 7.7 show the mean surface plot of the forecast for 8–14 July using the GPP based approximate models. Observed values of the ozone levels from a selected number of sites are also superimposed. Data from all the monitoring sites are not plotted to avoid clutter. We observe that the forecast values show a very good agreement with the actual values. The uncertainties in forecast are presented in Figures 7.8 to 7.11 using the standard deviation for the forecasts. Forecast maps of sd in 8–14 July shows that the overall sd varies between 6.0 to 9.0. The CMAQ output is also presented in Figures 7.12 to 7.15, where we observe that it over estimates the actual values in most of the areas.

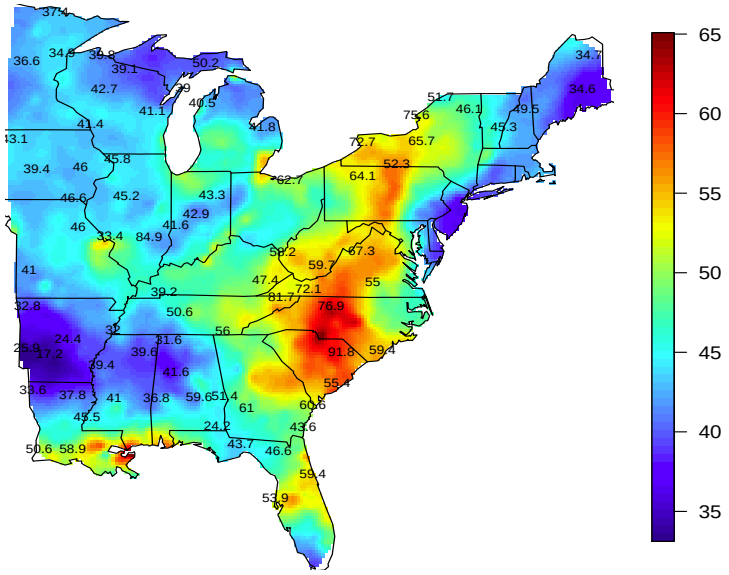


(a)

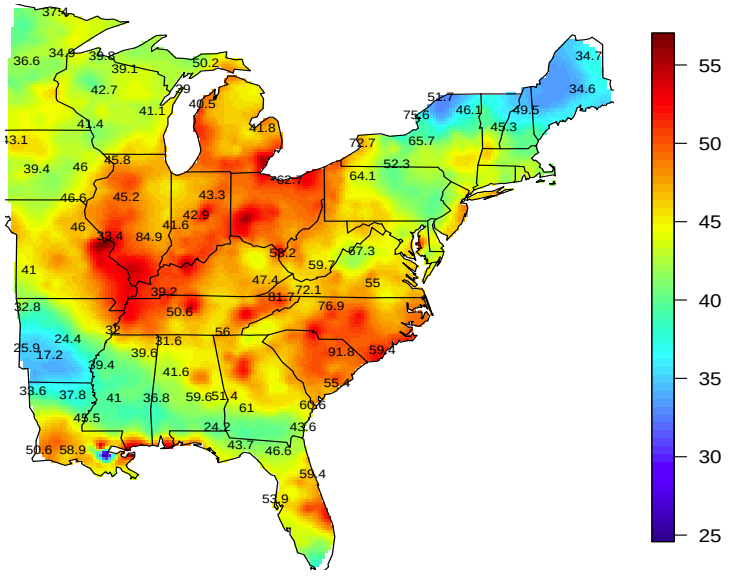


(b)

Figure 7.4: Forecast maps of the average daily ozone levels using the GPP based model for 7 days, panel (a) for 8 July and (b) for 9 July. Actual observations are also superimposed. The colour scheme is different for different maps.

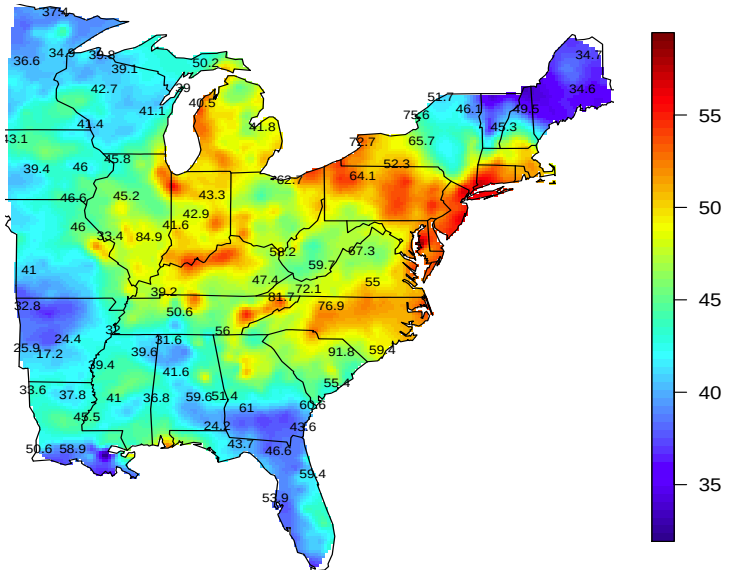


(a)

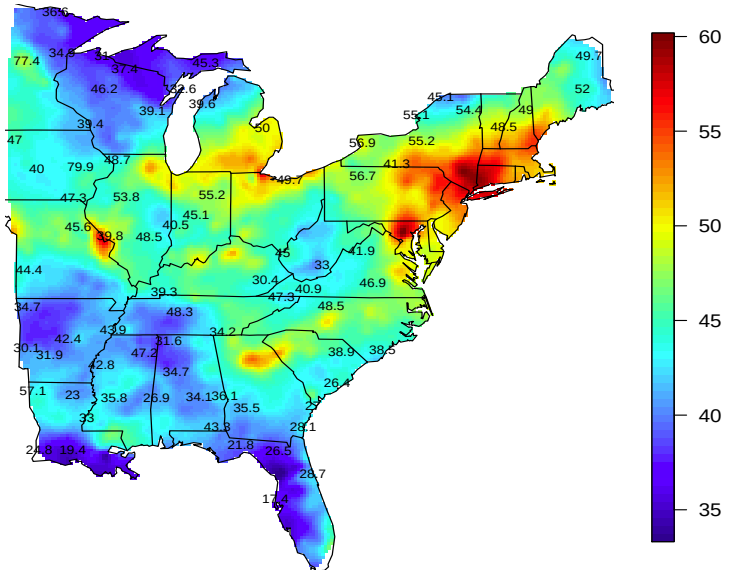


(b)

Figure 7.5: Forecast maps of the average daily ozone levels using the GPP based model for 7 days, panel (a) for 10 July and (b) for 11 July. Actual observations are also superimposed. The colour scheme is different for different maps.



(a)



(b)

Figure 7.6: Forecast maps of the average daily ozone levels using the GPP based model for 7 days, panel (a) for 12 July and (b) for 13 July. Actual observations are also superimposed. The colour scheme is different for different maps.

	GPP	CMAQ
Over estimation	51.65%	91.60%
Under estimation	48.35%	8.40%

Table 7.14: Percentage of over and under estimation of forecasts in the hold-out locations, for the models based on GPP approximation and the CMAQ output.

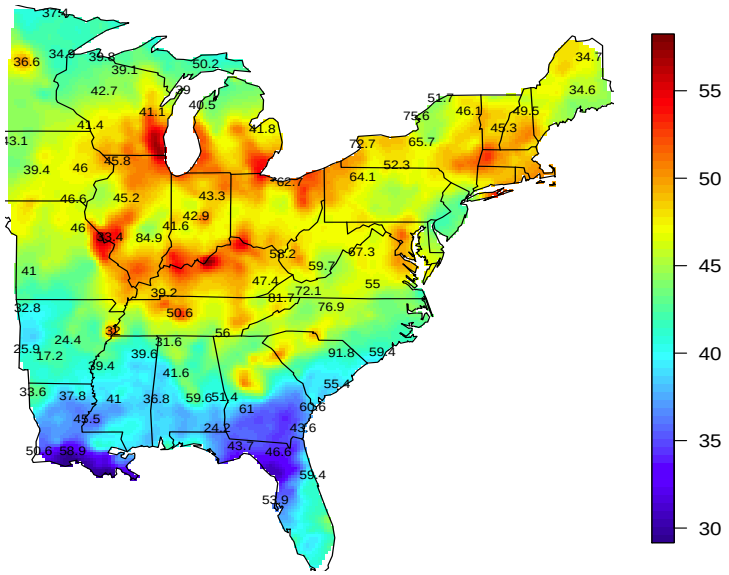
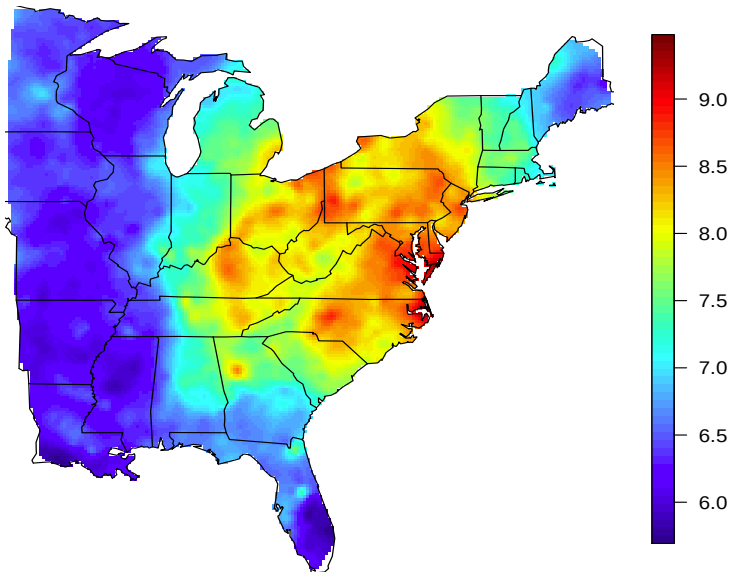
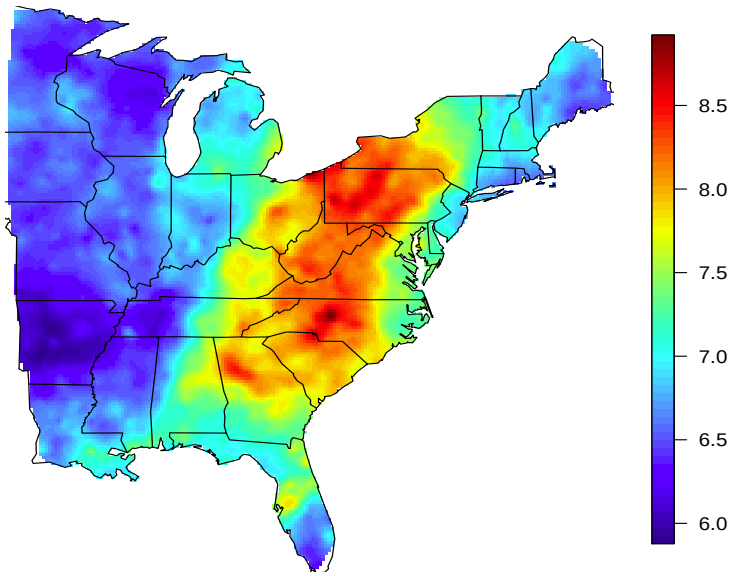


Figure 7.7: Forecast maps of the average daily ozone levels using the GPP based model for 7 days for 14 July. Actual observations are also superimposed.

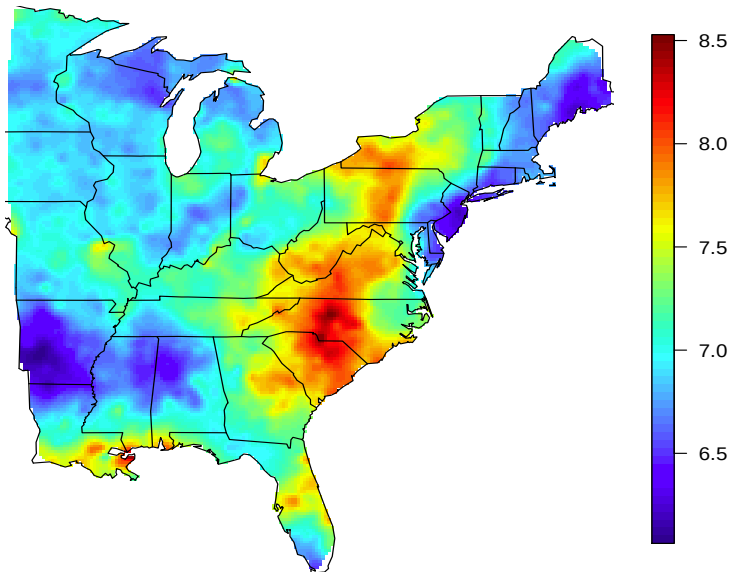


(a)

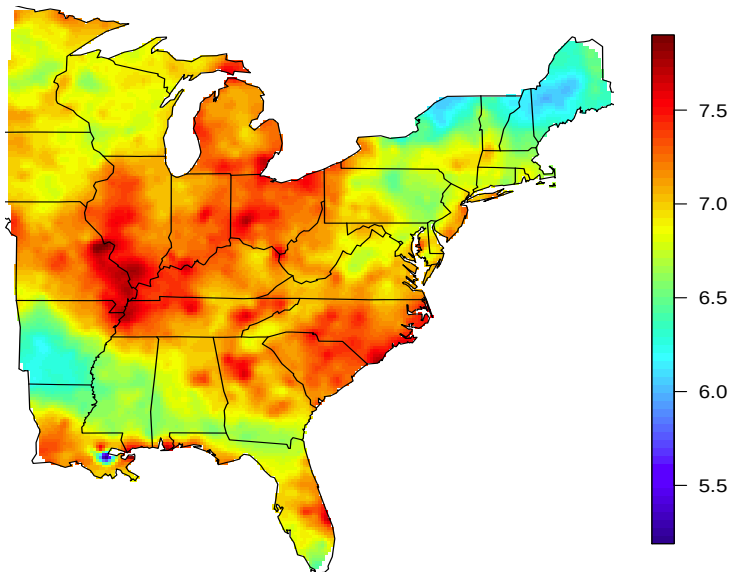


(b)

Figure 7.8: Forecast uncertainty (standard deviations) maps for the eastern US, using the GPP based model for 7 days, panel (a) for 8 July and (b) for 9 July. The colour scheme is different for different maps.

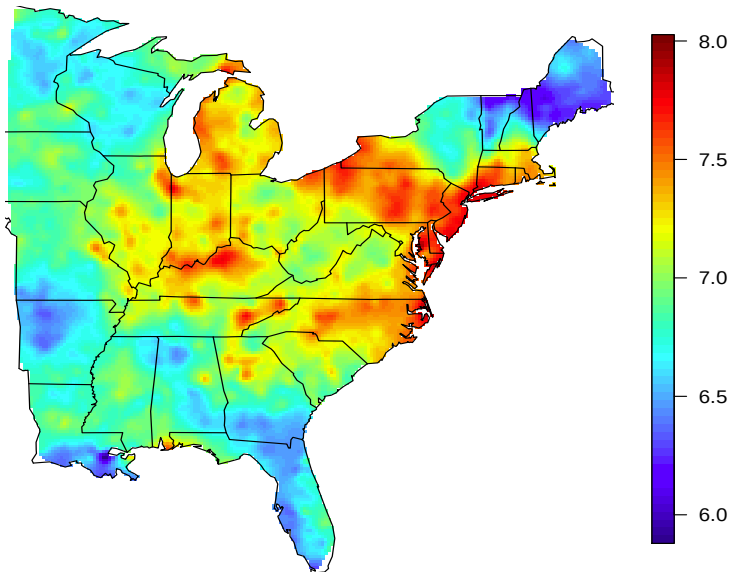


(a)

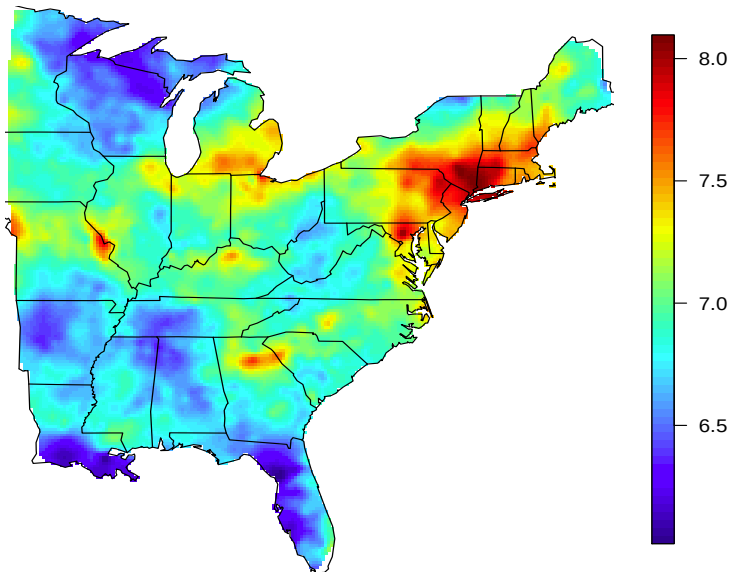


(b)

Figure 7.9: Forecast uncertainty (standard deviations) maps for the eastern US, using the GPP based model for 7 days, panel (a) for 10 July and (b) for 11 July. The colour scheme is different for different maps.

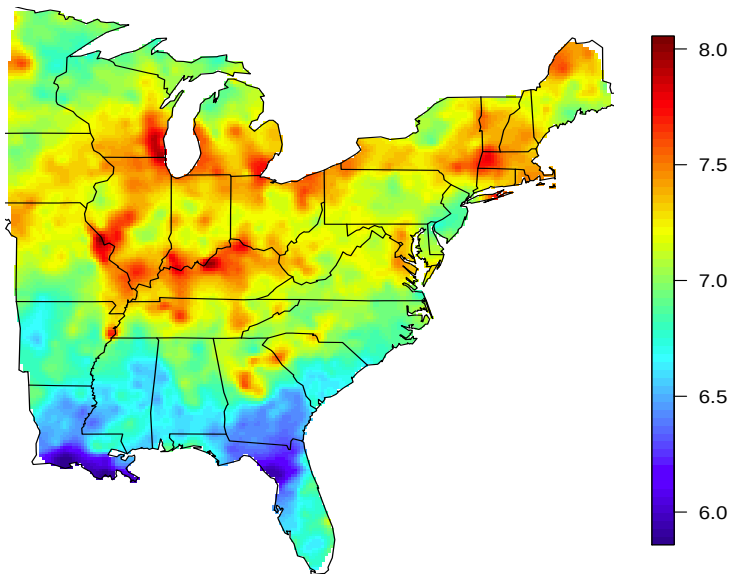


(a)



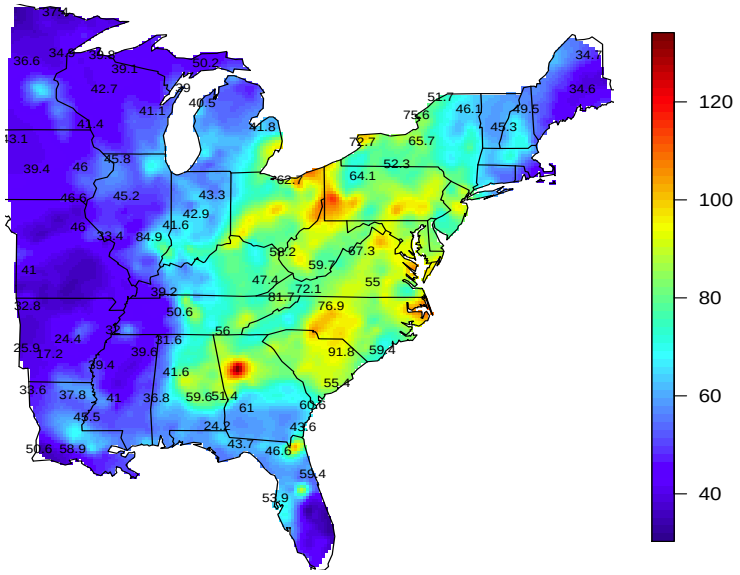
(b)

Figure 7.10: Forecast uncertainty (standard deviations) maps for the eastern US, using the GPP based model for 7 days, panel (a) for 12 July and (b) for 13 July. The colour scheme is different for different maps.

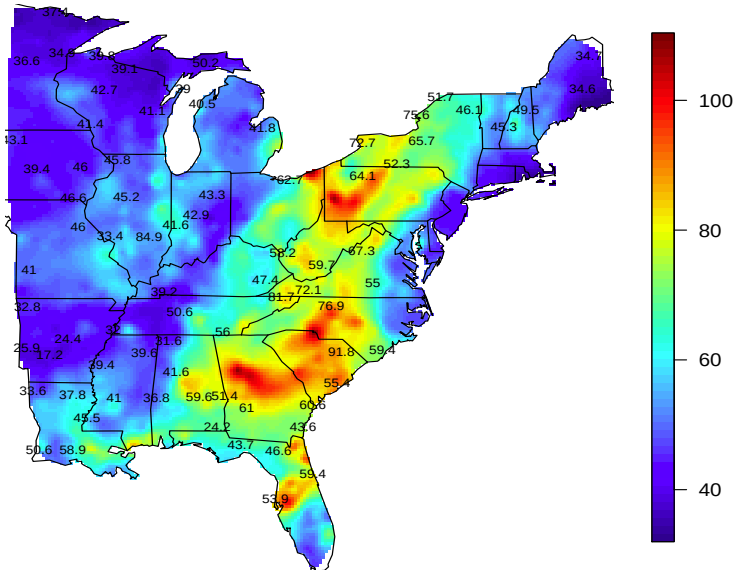


(a)

Figure 7.11: Forecast uncertainty (standard deviations) maps for the eastern US, using the GPP based model for 7 days for 14 July.

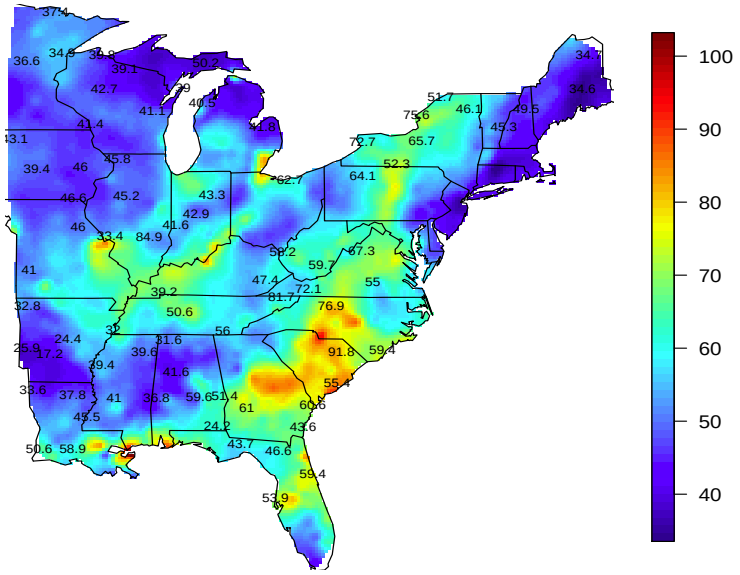


(a)

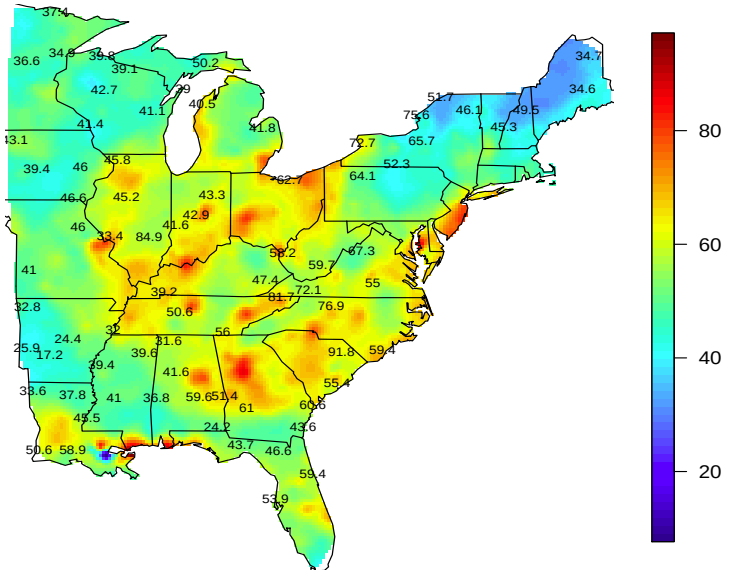


(b)

Figure 7.12: Forecast maps of the average daily ozone levels using the CMAQ model for 7 days, panel (a) for 8 July and (b) for 9 July. Actual observations are also superimposed. The colour scheme is different for different maps.

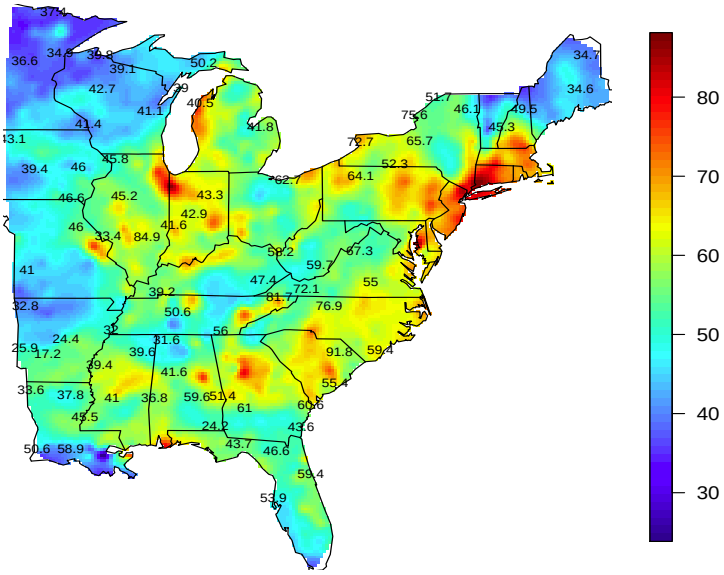


(a)

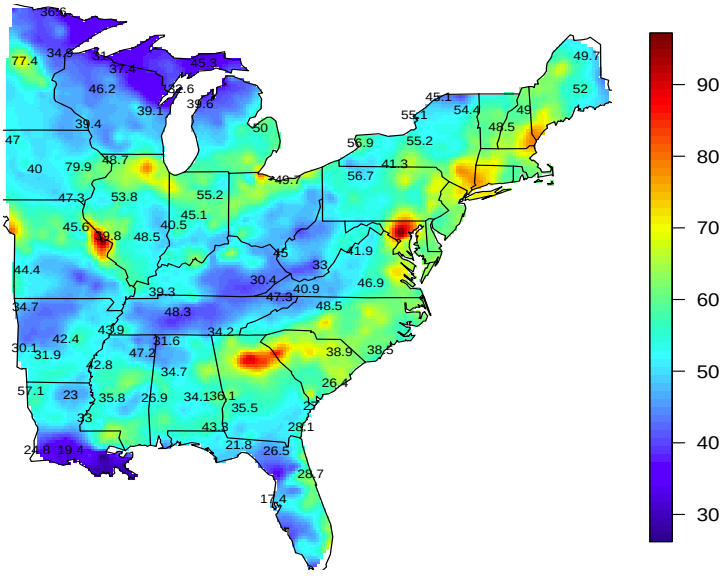


(b)

Figure 7.13: Forecast maps of the average daily ozone levels using the CMAQ model for 7 days, panel (a) for 10 July and (b) for 11 July. Actual observations are also superimposed. The colour scheme is different for different maps.

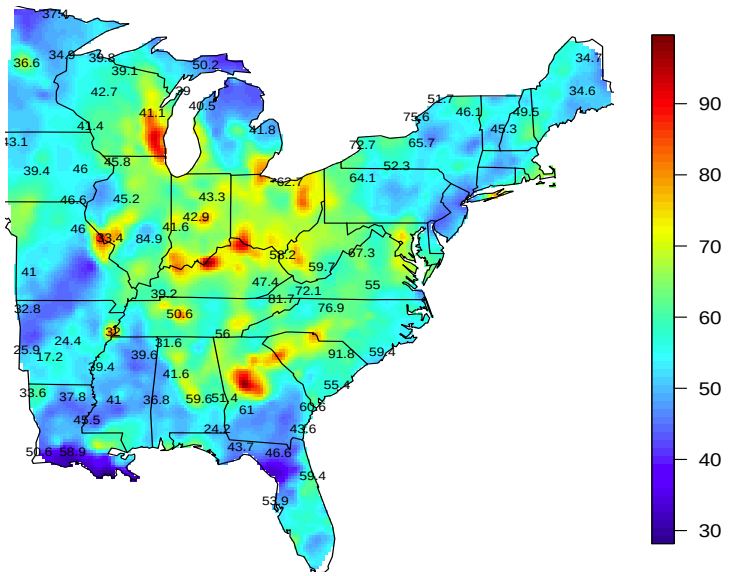


(a)



(b)

Figure 7.14: Forecast maps of the average daily ozone levels using the CMAQ model for 7 days, panel (a) for 12 July and (g) for 13 July. Actual observations are also superimposed. The colour scheme is different for different maps.



(a)

Figure 7.15: Forecast maps of the average daily ozone levels using the CMAQ model for 7 days for 14 July. Actual observations are also superimposed.

7.6 Summary

In this chapter we perform forecasting for next day's daily 8-hour maximum ozone concentration levels. We use different forecast methods in this context, namely the CMAQ, GP, DLM, AR, and the GPP based approximation models. Specifically, these methods have been illustrated and compared for a relatively small set of fitting data from four states in the eastern US. The results indicate a better forecasting performance for the GPP approximation model than the others. Unlike the GP, DLM and the AR models, the GPP based model is suitable for analysing large volumes of ozone concentration data. The GPP approximate model is also compared with the CMAQ output values in the eastern US, where we see that the proposed GPP based model performs much better than the CMAQ forecasts.

Chapter 8

spTimer: Spatio-Temporal Bayesian Modelling Using R

8.1 Introduction

This chapter illustrates the implementation of the previously introduced Bayesian spatio-temporal models in R using the package spTimer developed as a part of this thesis. The models included in this package are the Gaussian process (GP) models (Cressie, 1993; Stein, 1999; Banerjee *et al.*, 2004), the autoregressive (AR) models as introduced in Sahu *et al.* (2007), and the Gaussian predictive processes (GPP) based model for analysing large dimensional data as introduced in Chapter 6.

There are several R packages are available for model based analysis of spatial data under the Bayesian setup, for example, spBayes (Finley *et al.*, 2007), geoR (Ribeiro Jr and Diggle, 2001), geoRglm (Christensen and Ribeiro Jr, 2002). However these packages are not able to analyse spatio-temporal data, although the spBayes package can model some space-time data by using some multivariate spatial models, yet this is not feasible even for moderately sized data sets. Moreover, time series models such as the AR models cannot be implemented in spBayes.

The package spTimer is developed using the C-language, that enables much faster computation than the high level R language.

The main objective of this chapter is to verify the code of spTimer using simulation. We simulate data sets from each of the three implemented models

and then estimate the parameters using `spTimer` that enables us to verify the model fitting routines. We also verify the spatial interpolation and temporal forecasting routines using cross-validation. The code for this chapter is provided in the accompanying compact disk (CD).

The Gibbs sampler is implemented for all the models. Convergence of the Gibbs sampler has been assessed by using the Gelman and Rubin statistics (Gelman and Rubin, 1992) calculated from several parallel runs. We also have examined the time-series and auto-correlation plots of the MCMC samples and the chains converged rapidly for all models.

The rest of this chapter is organised as follows: Section 8.2 discusses the main functions and routines developed in the package `spTimer`. The details for simulating data from the GP, AR and GPP based models are provided in Section 8.3. Sections 8.4, 8.5, and 8.6 fit and analyse the simulated data sets from the GP, AR, and GPP based models respectively. These sections also provide prediction and forecast results for the models using `spTimer`. Finally, we conclude with a brief discussion in Section 8.7.

8.2 The Main Functions in `spTimer`

There are three main functions in `spTimer` package, namely, `spT.Gibbs` for model fitting, `spT.prediction` to obtain predictions based on the fitted models and `spT.forecast` to obtain forecast in future time points.

8.2.1 `spT.Gibbs`

The function `spT.Gibbs` is used to fit all three models using Gibbs sampling approach. Here is the list of arguments that can be sent to `spT.Gibbs`.

- The required argument `formula` is used to specify the linear part of the model. Its documentation is same as that for the formula argument of the R function `lm` used to fit linear regression models.

- The argument `data` provides the data set used for model fitting. The data set must be ordered by the location index and under each location data must be ordered by time. Time-series data with more than one segments, for example, T daily observations in each of r years, must be ordered first

by year and then by days in each year. The length of the segments must be same in each year and also for each location. Currently the package cannot handle any irregular time-series. Missing data should have the standard `NA` identifier. For each missing data point, standard Bayesian technique of treating it as an unknown and sampling this unknown parameter at each Gibbs iteration is employed. For the covariates no missing values are allowed.

- The argument `time.data` defines the time-series, and we use another function `spT.time` to specify this. See documentation in Section 8.2.4 for details.

- The required argument `model` specifies the intended model to be fitted and this can be one of the three, `GP`, `AR`, and `GPP`. The default is `GP`.

- The argument `coords` is used for providing the spatial locations, e.g., longitude and latitude, or easting and northing. This must be supplied as an $n \times 2$ matrix, where n is the number of locations in the data.

- The optional input `knots.coords` is only used for the models based on GPP approximations, i.e., when `model="GPP"`. This input must be an $m \times 2$ matrix as `coords`, where m is the number of knot locations and $m < n$.

- The prior distributions in `spT.Gibbs` are provided using argument `priors`. If we choose `priors=NULL` then the routine `spT.Gibbs` automatically takes proper prior distributions for the model parameters. Prior configuration can also be defined using the output of the function `spT.priors`. See Section 8.2.4 for details.

- Initial values for the model parameters are defined in `spT.Gibbs` using the function `spT.initials`. Details of this argument is given in Section 8.2.4. In addition, writing `initials=NULL` yields default choices of input values for the model parameters. The default values for spatial variance is 0.1 and for nugget effect, it is 0.01. For spatial decay, default is calculated using $(-\log(0.05)/d_{max})$, where d_{max} is the maximum distance calculated

from the coordinates. The initial parameters for the covariates including the auto-regressive term are calculated using a simple linear model.

- The argument `its` specifies the number of iterations in the Gibbs sampler.
- The argument `burnin` is the number of initial iterations to be discarded before making inference.
- The argument `report` is the number of reports printed on screen to monitor the progress of the Gibbs sampler. The default is report equals to one for printing information only once after finishing all iterations.
- The argument `distance.method` specifies the method to calculate the distance between any two locations. This argument can take any if the values `"geodetic:km"` for distance in kilometres, `"geodetic:mile"` for distance in miles, `"euclidean"` for Euclidean. See Section 2.6.1 for details regarding geodetic distances.
- To ensure the non-singularity of the covariance matrices, we can also define the minimum allowed distance between two locations out of those specified by the coordinates. For example, `tol.dist=2` implies the allowed distance as 2 units of measurement. The default unit is 0.005. The programme will exit if the minimum distance is less than the non-zero specified values.
- The choice of the spatial covariance function is provided by the required argument `cov.fnc`. This argument can take one of the values: `"exponential"`, `"gaussian"`, `"spherical"`, and `"matern"`. See Section 2.4.4 for more details regarding the covariance functions.
- There are three options for handling the spatial decay parameter ϕ using the argument `spatial.decay` in `spT.Gibbs`. The function `spT.decay` sets up the options for this. See details in Section 8.2.4.
- A particular scale transform of the response can be provided using the optional argument `scale.transform`. Currently, it can take one of the values `"NONE"`, `"LOG"` or `"SQRT"`. The default is `"NONE"`. Note that all the predictions and forecast will be made on the original scale. Further, this transformation does not apply to any of the covariates.

The function `spT.Gibbs` will both fit and predict if two further optional arguments, `pred.coords` and `pred.data` are provided. These are described as follows:

- `pred.coords` is a $q \times 2$ matrix of prediction locations similar to the `coords` argument, where q is the number of prediction locations.
- `pred.data` should be a data frame with the same space-time structure as the fitted data frame.

In this combined approach there is an option to obtain summary statistics by aggregating different time segments. For example, if data set has 30 days observations for 5 years, then use of `annual.aggregation="ave"` yields annual average, and `"an4th"` yields annual 4th highest value. Currently we have the options `"ave"`, `"an4th"` and `"NONE"`, where `"NONE"` represents no annual summary statistics. Obviously this input is only meaningful if `spT.time` has input more than one segment and when fit and predict are done together.

Output of the Function `spT.Gibbs`

The output of `spT.Gibbs` is a list containing various information. Some of the members of this list themselves are list or matrices. These are described as follows:

- All MCMC samples of the model parameters.
- The `spT.Gibbs` also provides the `PMCC` that is the predictive model choice criteria discussed in Section 3.3. Both penalty and goodness of fit values are obtained from the output.
- In the output-list, object `X` and `Y` represents the design matrix and the independent variables that have been used in the model fitting.
- `call` provides the formula that has been used for model fitting.
- The `spT.Gibbs` also provides output that can identify the distance method (`distance.method`), the name of the covariance function (`cov.fnc`), the type of scale transformation (`scale`), and the approach used for sampling the spatial decay parameter (`sampling.sp.decay`).

- There are also output lists of the prior distributions (`priors`) and initial values (`initials`) used in the fitted model.
- The MCMC control parameters, i.e., number of iterations (`its`), burn-in from the output (`burnin`) can be obtained from the output.
- In addition, we can also recall the computational time elapsed (`computation.time`) in the model fitting using Gibbs algorithm.

Text Output of the Function `spT.Gibbs`

We have already mentioned that model fitting and prediction can be done together using function `spT.Gibbs`. In this case, the `spT.Gibbs` also writes out some output values in text files in the current working directory.

- MCMC values of the model parameters are given in a text file whose name is specified according to the model.
- We also get MCMC samples for prediction by the file name "OutMODEL-Values-Prediction.txt" and the file name changes in different models as described in the previous paragraph.
- Mean and standard deviations of the predicted values are also written in the text file as "OutMODEL-Stats-PredValue.txt". This text file is useful when the data set is very large.
- Similarly we get the text output of the fitted summary statistics by "OutMODEL-Stats-FittedValue.txt" and so on.
- Particularly, for the AR models we obtain one more text file that is for the summary statistics of true ($O_l(\mathbf{s}_i, t)$, see equation 3.10) underlying values as "OutAR-Stats-TrueValue.txt"
- If `annual.aggregation` is equal to `"ave"`, then we get another text file that is the MCMC values for the annual averages and the name of the file is "OutMODEL-Annual-Average-Prediction.txt". `annual.aggregation="an4th"` yields the text file of the MCMC values for the annual 4th highest values and is written on the file "OutMODEL-Annual-4th-Highest-Prediction.txt". Similarly, the first part of the file name changes for different models.

8.2.2 `spT.prediction`

The function `spT.prediction` is used to obtain predictions at unmonitored locations based on the results obtained from the routine `spT.Gibbs`. This function will not work if fit and predict has already been done in the `spT.Gibbs`.

Arguments for the Function `spT.prediction`

- `pred.coords` and `pred.data` are same as defined in Section 8.2.1.
- The required argument `posterior` in `spT.prediction` must be the output of the model fitting routine `spT.Gibbs`.
- The minimum separation distance between the fitted and prediction sites is defined by `tol.dist` and discussed in Section 8.2.1. The default is 0.005. The programme will exit if the minimum distance is less than the specified values.
- A logical expression is used to get summary statistics by writing `Summary=TRUE`. Default is `TRUE`.
- There is also an option to include further burn-in if necessary using `burnin` argument. For example, if `burnin` is 5000 and burn-in in `spT.Gibbs` is 1000 then it will remove 6000 iterations altogether, and if the total number of iterations are less than 6000 then it will stop the programme and provide related warning messages.

Output of the Function `spT.prediction`

- The MCMC prediction samples are also available through `predicted.samples`.
- If `Summary=TRUE` in the routine `spT.prediction`, then the output includes `Mean`, `Median`, `SD` (i.e., standard deviations), and 95% lower and upper prediction intervals of the prediction samples.
- Some other output, e.g., `distance.method`, `cov.fnc`, and `computation.time` are also obtained from the function `spT.prediction`.

8.2.3 `spT.forecast`

To obtain forecast using package `spTimer`, we use the function `spT.forecast`. This function can calculate K-step ahead forecasts.

Input for the Function `spT.forecast`

- In `spT.forecast` there is option to get K-step ahead forecast using `K`.
- `burnin` option is also available in this forecast function.
- There is option to include the forecast covariate values by `fore.data`, and the forecast coordinates using `fore.coords` as discussed in the previous sections.
- Similar to prediction, the output of the `spT.Gibbs` are used as input in `posteriors`.
- The forecast summary can be obtained writing `Summary=TRUE`.

Output of the Function `spT.forecast`

- `forecast.samples` are for the forecast MCMC output.
- Similar to prediction, if `Summary=TRUE`, we get the summary statistics for the forecasts that includes `Mean`, `Median`, `SD` (i.e., standard deviations), and 95% lower and upper forecast intervals.
- `distance.method`, `cov.fnc`, and `computation.time` are also obtained similar to prediction output.

8.2.4 Some Other Functions

In this package some other utility functions are also provided that are often needed to get summary. Some of these are discussed below:

- The number of years and the length of the segments in the time-series are provided by argument `spT.time`. For example, if we have $30(= T)$ days of observations in $5(= r)$ years, then we define the `spT.time` function as:

```
> t.data <- spT.time(t.series = 30, segments = 5)
```

The function `spT.time` can also be used to define hourly time-series data. For example, we can define 24 hours as `t.series=24` and 5 days as `segments=5`. There is no default given for `t.series` and for `segments` the default is 1.

- `spT.priors` routine has inputs to define the hyper-parameter values of the prior distribution. For example, for model variances and spatial decay parameters we consider Gamma prior distribution with the hyper-parameters $a = 2$ and $b = 1$; for regression coefficient β and auto-regressive parameter ρ we consider Normal prior distribution with mean zero and variance 10^4. We write the input for `spT.priors` for model `AR` as:

```
> prior <- spT.priors(model="AR", var.prior=Gam(a=2,b=1),
           beta.prior=Nor(0,10^4), rho.prior=Nor(0,10^4),
           phi.prior=Gam(a=2,b=1))
```

  For other models we need to change the name in the argument `model` as described in Section 8.2.1. If any argument in `spT.priors` is not given then for that option by default a proper prior specification will be made.

- We can provide the initial values of the model parameters through the `spT.initials` argument. For example, we input the initial values of $\sigma^2_\epsilon = 0.01$, $\sigma^2_\eta = 0.5$, $\rho = 0.2$, $\boldsymbol{\beta} = (1.8, 0.3)'$ and $\phi = 0.01$ for the model AR as:

```
> initials <- spT.initials(model="AR", sig2ep=0.01,
              sig2eta=0.5,rho=0.2,beta=c(1.8,0.3),
              phi=0.01)
```

  Similar to the `spT.priors` we can choose the models and any input defined as `NULL` will take the initial values described in Section 8.2.1.

- `spT.decay` is used to handling the sampling method of the ϕ parameter. The function select one of the three options described below:

  1. **Fixed**: The first choice is to fix ϕ at a particular value. This is achieved by writing `type="FIXED"` in the argument. For example, for fixing ϕ at 0.01 we write:

     ```
     > spatial.decay <- spT.decay(type="FIXED", value=0.01)
     ```

2. **Discrete**: This option corresponds to assuming a discrete uniform prior for ϕ in a specified interval. Then the full conditional distribution of ϕ will be discrete and Gibbs sampler will sample from this distribution. A typical specification is provided below:

   ```
   > spatial.decay <- spT.decay(type="DISCRETE",
                       limit=c(.01,.02), segments=10)
   ```

   where, the `segments` argument specifies the number of support points in the prior distribution under this option. The prior for ϕ in `spT.priors` will be ignored.

3. **Metropolis-Hastings**: this is the most general method for sampling ϕ. A random-walk-Metropolis sampling algorithm (see Section 3.2) is used to sample ϕ. The tuning parameter of normal distribution (the standard deviation of the proposal) is also needed for this approach. This algorithm must be specified as follows:

   ```
   > spatial.decay <- spT.decay(type="MH", tuning=0.08)
   ```

   where, the desired tuning parameter (see Section 3.2.5) is supplied by the `tuning` argument.

Currently, no default choice is available for the routine `spT.decay`.

- `spT.MCMC.stat` is used to obtain the MCMC summary statistics of the model parameters. For example, we write:

  ```
  > spT.MCMC.stat(posteriors, burnin=1000)
  ```

  where, `posterior` argument is the output of the routine `spT.Gibbs`. This will produce the summary statistics for the model parameter with 1000 burn-in.

- Similarly, we obtain the MCMC trace plots with density and auto-correlation and partial auto-correlation plots of the model parameters using function `spT.MCMC.plot`. Typically we write:

  ```
  > spT.MCMC.plot(posteriors, burnin=1000, ACF=TRUE,
                  PARTIAL.acf=TRUE)
  ```

- The function `spT.geodist` is used to calculate the geodetic distance between two locations using the formula stated in Section 2.6.1.
- The validation criteria defined in Section 3.3.4 are calculated using the routine `spT.validation`. The output of this function includes VMSE, RMSE, MAE, rBIAS, and rMSEP.
- The nominal coverages are calculated using the function `spT.pCOVER`.

For details see the `spTimer` documentation provided in the attached CD.

8.3 Simulation Study

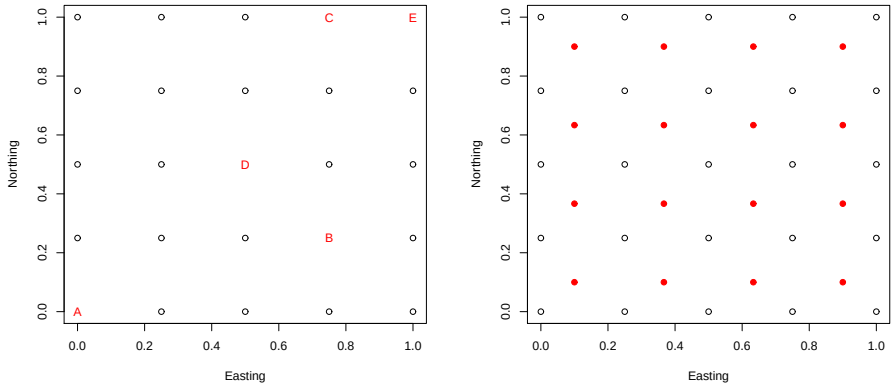
We perform a simulation study to validate the `spTimer` code. Data sets are simulated from each of the three models, and then the package `spTimer` is used to estimate the model parameters. Prediction at the unmonitored locations and forecasts at future time points are also performed in this simulation study for GP, AR, and GPP based approximation models.

Sensitivity analysis for the prior distributions of the model parameters are also considered in the simulation study. We use only Metropolis-Hastings method for sampling the spatial decay parameter ϕ.

8.3.1 Simulation Design

We simulate data sets from each of the three spatio-temporal models. A regular spatial grid size of 5×5 in the unit square $(0, 1) \times (0, 1)$ is used for simulating data. We simulate data for 31 days in each for 2 years, thus for each data set we have total $25 \times 31 \times 2 = 1550$ observations. Figure 8.1(a) shows the grid location points that are used in this section. We use only exponential covariance function to simulate data sets and also for model fitting.

For prediction validation, we set aside data from 5 randomly chosen locations see for example, Figure 8.1(a). Similarly, for forecast validation we consider 30 days observation in each year for model fitting and obtain forecast for day 31 of each year for the hold-out locations.



(a) (b)

Figure 8.1: A representation of the 25 regular grid locations for the replicated data. (a) Five locations A-E are chosen randomly and set aside for validation. (b) Locations in solid circle are 16 knot points used for GPP based approximation models.

8.3.2 True Parameter Values for the GP Models

Data set for the GP models are simulated with a true value of the intercept $\beta = 5.0$. The variance parameters, i.e., the nugget effect and the spatial error variance of the models are set at: $\sigma^2_\epsilon = 0.001$ and $\sigma^2_\eta = 0.1$ respectively. Spatial-decay parameter (ϕ) is taken as 0.01, and we use the Euclidean distance to obtain the spatial correlation.

8.3.3 True Parameter Values for the AR Models

For the AR model we use the same intercept and variance components as for the GP models. In addition, the auto-regressive parameter for the AR model is set at $\rho = 0.2$. The initial mean and variance for $O_l(\mathbf{s}_i, 0)$ are taken to be $\mu_l = 5.0$ and $\sigma^2_l = 0.5$, that are same for each year l, where $l = 1, 2$.

8.3.4 True Parameter Values for the GPP based Models

In GPP based approximation models we simulate data set using same location points that have been used for the GP and AR models. In addition, we define $m = 16$ knot points that is smaller to the actual locations $n = 25$, see Figure 8.1(b). The temporal auto-correlation for the spatial random effect term is considered

Parameters		β_0	σ^2_ϵ	σ^2_η	ϕ
True values		5.0000	0.0010	0.1000	0.0100
Hyper-prior		Estimates			
(a=2,b=1)	Low	4.9649	0.0008	0.0440	0.0011
	Mean	5.0951	0.0012	0.2405	0.0077
	Up	5.2288	0.0017	0.9103	0.0224
(a=1,b=1)	Low	4.9586	0.0009	0.0468	0.0011
	Mean	5.0957	0.0014	0.2587	0.0073
	Up	5.2292	0.0019	0.9204	0.0202
(a=10,b=10)	Low	4.9504	0.0010	0.0990	0.0012
	Mean	5.0961	0.0198	0.3232	0.0044
	Up	5.2451	0.0292	0.7644	0.0102

Table 8.1: Posterior mean and 95% credible interval of the GP model parameters for different hyper-prior values for the simulated data set obtained from the GP model.

as $\rho = 0.2$. We also assume that the initial mean $\mu_l = 0$ and $\sigma^2_l = 0.5$ for the spatial random effect. Other parameters of the GPP based model are assumed to be as above for the GP and AR models.

8.4 Simulation Example: GP Models

8.4.1 Sensitivity of Prior Distribution

Table 8.1 provides estimated values of the GP model parameters with 95% credible interval. We perform the sensitivity study for the Gamma prior distributions changing its hyper-parameter values. It is observed that all the parameters are close to the true simulation values and all the 95% credible intervals contain these true values.

8.4.2 Predictions and Forecasts

In this section we discuss predictions and forecasts using GP models. As mentioned in Section 8.3, we randomly select 5 locations out of 25 locations from one of the simulated data set and set aside for validation purpose, see Figure 8.1(a). We obtain forecast at the prediction locations at day 31 analysing the 30 days observations for both years, (i.e., $l = 1, 2$). Table 8.2 provides the root mean squared error (RMSE) and mean absolute error (MAE) (see details in Section 3.3) for prediction and forecast validations for the 5 hold-out sites. In Figure 8.2 we

represent the 30 days prediction and next day forecast estimates with 95% intervals for one hold-out site. As expected the 95% prediction interval is smaller compared to the 95% forecast interval.

	Prediction		Forecast	
Location	RMSE	MAE	RMSE	MAE
A	0.0300	0.0229	0.4022	0.3306
B	0.0223	0.0166	0.3786	0.2925
C	0.0300	0.0239	0.3813	0.3074
D	0.0346	0.0292	0.3994	0.3369
E	0.0374	0.0299	0.3876	0.3154
All	0.0300	0.0245	0.3900	0.3166

Table 8.2: Prediction validations for the GP model for simulated data set obtained from the GP model.

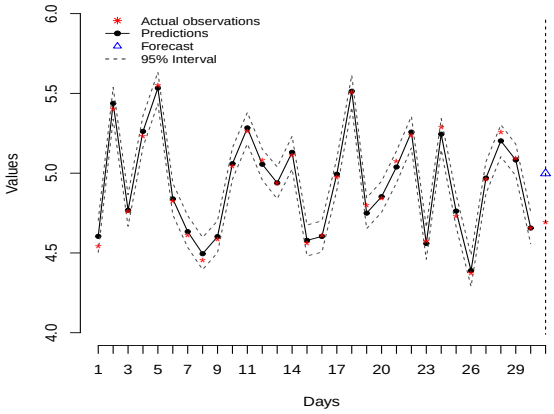


Figure 8.2: Prediction and forecast results for first 31 days in a hold-out site for the GP models. 95% prediction and forecast intervals are also superimposed.

8.5 Simulation Example: AR Models

To reconstruct the true parameters of the auto-regressive models we model the simulated data set obtained from the AR models (see Section 8.3.1).

8.5.1 Sensitivity of Prior Distribution

Table 8.3 provides estimated values of the AR model parameters with 95% credible interval for different hyper-parameter values of the Gamma prior distribution. These estimates here are more sensitive to the larger hyper parameter values than the same under GP models. The initial parameters (μ_l, σ_l^2) for the true values $(\mathbf{O}_{l0})$ are also estimated and presented in Table 8.4 for different hyper-parameters.

Parameters		β_0	ρ	σ_ϵ^2	σ_η^2	ϕ
True values		5.0000	0.2000	0.0010	0.1000	0.0100
Hyper-prior		Estimates				
(a=2,b=1)	Low	4.8900	0.1013	0.0009	0.0802	0.0060
	Mean	5.2599	0.1601	0.0012	0.1195	0.0092
	Up	5.6344	0.2173	0.0016	0.1747	0.0133
(a=1,b=1)	Low	4.9270	0.0961	0.0010	0.0871	0.0053
	Mean	5.2969	0.1543	0.0015	0.1274	0.0086
	Up	5.6720	0.2121	0.0019	0.1968	0.0122
(a=10,b=10)	Low	4.8773	0.0965	0.0152	0.2087	0.0028
	Mean	5.2753	0.1578	0.0164	0.2748	0.0040
	Up	5.6839	0.2169	0.0178	0.3733	0.0052

Table 8.3: Posterior mean and 95% credible interval of the AR model parameters for different hyper-parameter values for the simulated data set obtained from the AR model.

Parameters		μ_1	μ_2	σ_1^2	σ_2^2
True values		5.0000	5.0000	0.5000	0.5000
Hyper-prior		Estimates			
(a=2,b=1)	Low	1.2089	0.9713	0.1693	0.2136
	Mean	6.0585	4.2187	0.7238	1.3025
	Up	10.8059	9.2562	2.3946	4.7667
(a=1,b=1)	Low	0.4329	0.1726	0.2409	0.3420
	Mean	5.9849	4.0058	1.3284	2.8524
	Up	11.1200	9.8612	4.6837	11.1200
(a=10,b=10)	Low	-1.1072	-3.8097	0.5734	0.6094
	Mean	6.2430	3.9120	1.0874	1.1874
	Up	13.9117	11.2282	2.0201	2.2974

Table 8.4: Posterior mean and 95% credible interval of the AR model parameters μ_l and σ_l^2 for different hyper-parameter values for the simulated data set obtained from the AR model.

8.5.2 Predictions and Forecasts

The RMSE and MAE validation results are given in Table 8.5. Figure 8.3 represents the prediction and forecast estimates with 95% intervals for a hold-out site for the AR models. Similar to GP models and for simplicity we represent 30 days prediction and next day forecast. As expected the 95% prediction interval is smaller compared to the 95% forecast interval.

	Prediction		Forecast	
Location	RMSE	MAE	RMSE	MAE
A	1.1612	1.0271	1.0957	1.0448
B	1.2641	1.1637	1.1172	1.0577
C	1.2661	1.1662	1.0981	1.0459
D	1.2704	1.1670	1.0632	1.0004
E	1.2653	1.1635	1.0497	0.9964
All	1.2669	1.1660	1.0851	1.0290

Table 8.5: Prediction validations for the GP model for simulated data set obtained from the AR model.

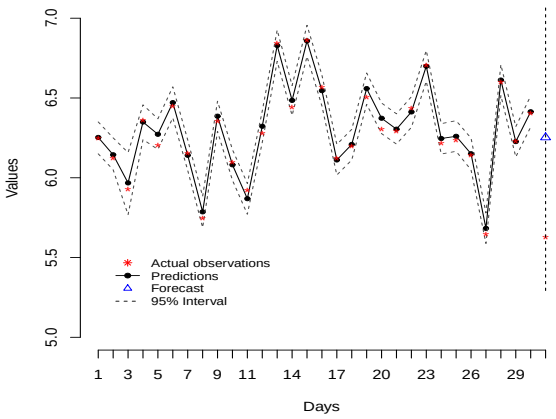


Figure 8.3: Prediction and forecast results for first 31 days in a hold-out site for the AR models. 95% prediction and forecast intervals are also superimposed.

8.6 Simulation Example: GPP based Models

In this section, similar to GP and AR models we first conduct sensitivity analysis and then obtain results on prediction and forecasts using package `spTimer`.

8.6.1 Sensitivity of Prior Distribution

Similar to GP and AR models we conduct sensitivity study for the prior distributions of the GPP based model. Table 8.6 represents the parameter estimated. We observe parameter estimates are sensitive for large hyper-parameter values.

Parameters		β_0	ρ	σ^2_ϵ	σ^2_η	ϕ	σ^2_1	σ^2_2
True values		5.0000	0.2000	0.0010	0.1000	0.0100	0.5000	0.5000
Hyper-prior		Estimates						
(a=2,b=1)	Low	4.7800	0.0652	0.0025	0.0998	0.0046	0.0595	0.0602
	Mean	4.8566	0.1564	0.0027	0.1529	0.0074	0.1061	0.1059
	Up	5.0215	0.2445	0.0030	0.2155	0.0117	0.1840	0.1814
(a=1,b=1)	Low	4.8075	0.0659	0.0025	0.1102	0.0043	0.0623	0.0618
	Mean	4.8604	0.1555	0.0027	0.1590	0.0071	0.1127	0.1124
	Up	4.9232	0.2442	0.0030	0.2230	0.0110	0.1957	0.1936
(a=10,b=10)	Low	4.7955	-0.0624	0.0176	0.2218	0.0019	0.2912	0.2915
	Mean	4.8923	0.0295	0.0191	0.2948	0.0045	0.4329	0.4324
	Up	5.0122	0.1239	0.0207	0.3959	0.0079	0.6475	0.6323

Table 8.6: Posterior mean and 95% credible interval of the GPP based model parameters for different hyper-parameters.

8.6.2 Predictions and Forecasts

Similar to the GP and AR models, prediction and forecast results are also obtained. Table 8.7 represents the validation results for the simulated data set for both prediction and forecasts. In Figure 8.4 we see that the prediction and prediction intervals are well suited to the actual values, in addition the 95% forecast interval is much larger compared to the 95% prediction intervals.

	Prediction		Forecast	
Location	RMSE	MAE	RMSE	MAE
A	0.4930	0.4310	0.1845	0.1341
B	0.0307	0.0257	0.1951	0.1385
C	0.0265	0.0206	0.1696	0.1455
D	0.0643	0.0526	0.2043	0.1699
E	0.0590	0.0475	0.1497	0.1352
All	0.0438	0.0323	0.1817	0.1446

Table 8.7: Prediction validations for the GP model for simulated data set obtained from the GPP based model.

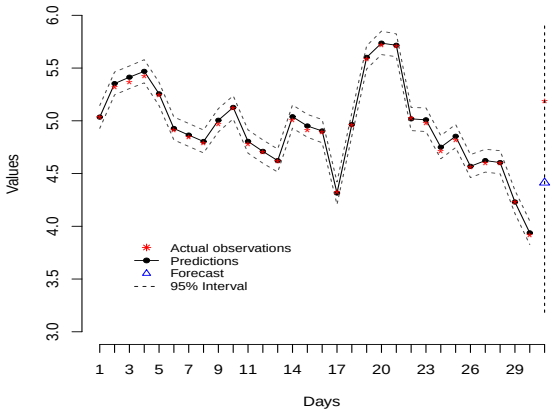


Figure 8.4: Prediction and forecast results for first 31 days in a hold-out site for the GPP based models. 95% prediction and forecast intervals are also superimposed.

8.7 Summary

In this chapter we discuss the `spTimer` package in `R` that is developed using the `C` language. Currently this package is suitable for analysing data using three different types of spatio-temporal models, i.e., the GP, AR and GPP based approximation models. We provide several simulation examples to validate the code developed for the package `spTimer` for all three models. It has been observed that the estimated model parameters are close to the true simulation values. In addition, the 95% credible intervals of the estimated parameters always include the true values of the simulated data sets. We also provide different sensitivity analysis that includes the sensitivity of the prior distributions. Prediction at unmonitored locations and forecasts in future time points are also discussed using the package `spTimer` for the three spatio-temporal models. The simulation examples, presented in this chapter, validate the `spTimer` code by correctly re-estimating the simulation parameters. These example also validate the code for spatial interpolation and temporal forecasting.

Chapter 9

Conclusion and Future Work

This last chapter contains the summary of the thesis and outlines some future work that can be extended based on the approaches adopted here. The results are summarised in Section 9.1 In addition, we discuss some limitations of the analysis in Section 9.1.1.

9.1 Thesis Summary

This thesis is motivated by the need to analyse and obtain long term trends in ozone concentration levels using Bayesian hierarchical spatio-temporal models. In this context, we use data obtained from a vast region of the eastern US for 10 years. As a part of this analysis we have done editing and cleaning of the raw ozone concentration data that have been obtained from the USEPA (see Chapter 4). This thesis addresses several challenges related to ozone modelling. We model ozone levels using hierarchical structure and also provide predictions at unmonitored locations. In addition, we forecast at future time points at those unmonitored locations. There are also issues of choosing appropriate modelling strategies for analysing ozone levels (see Chapter 5) and challenges to handle large dimensional spatio-temporal data (see Chapter 6) that have also been discussed in this thesis. Some of these major issues are as follows:

- **Comparison of Rich Hierarchical Spatio-Temporal Models:**
  In this thesis we compare two well-known Bayesian hierarchical spatio-temporal modelling strategies, the DLM and the AR models (see Chapter 5). Theoretical model comparison of these approaches are adopted

based on their correlation and covariance structure. We observe that the AR model theoretically gives better result compared to the DLM. Model comparison has also been completed using simulation studies and a real life example of the ozone concentration levels obtained from the state of NY.

- **Spatio-Temporal Models for the big-n Problem:**
  A major problem in analysing large dimensional space-time data comes from the need to invert high dimensional variance-covariance matrices, that is also known as the *big-n* problem. In Bayesian hierarchical context, repeated inversion of this matrix is almost infeasible. In this thesis, we adopt the concept of predictive processes approximation and propose a rich hierarchical spatio-temporal model (see Chapter 6) to analyse ozone levels in the vast eastern US study region. We provide spatial interpolation as well as temporal forecasting (see Chapter 7) using the proposed model based on the GPP approximation. Long term meteorology adjusted and unadjusted trends in ozone levels are also obtained from 1997 to 2006 in the eastern US region that has never been done before.

- **Adoption of Data Assimilation Techniques:**
  The deterministic computer simulation model output are also used in this thesis to model the observed ozone levels. We use grid output of the CMAQ model in the NY data example (see Chapter 5) and also in the forecast models (see Chapter 7) to analyse the eastern US data set. This type of data assimilation leads us to adopt the downscaler models, when grid-level deterministic model output is used as a covariate in the statistical models (see Section 1.4.3). This type of covariate information enriches the models we developed in this thesis.

- **Software for the Models:**
  As a part of this thesis we have developed a software package `spTimer` in `R`. Currently there is no package available to analyse data using Bayesian hierarchical spatio-temporal models. This package is written in low-level `C` language that facilitates fast model fitting. Currently, three Bayesian hierarchical models can be fit using `spTimer`. These are the GP spatio-temporal linear regression models, the AR models and the GPP based approximation models. Details of code validation have been presented in

Chapter 8.

9.1.1 Limitations

Some limitations related to this thesis are discussed below:

- The models used in this thesis particularly dealt with the Gaussian process approaches for analysing ground level ozone concentrations. In addition these models are not suitable for irregularly observed time-series.
- The `spTimer` package is currently able to fit only three Gaussian process models. In addition, input data for the package should have a particular structure, where time points are ordered and regular for each spatial locations.

9.2 Future Work

- **Challenges with Irregular Time-Series Data:**
  In this thesis we consider only the regular time-series data. For example, in each year we have observations for 153 days in each of the spatial locations in the eastern US study region. There is scope for extending the methods and the software for irregular observed data. For example at each time point a different location may be sampled, see e.g., Sahu and Challenor (2008). In addition, the length of the segments may be different at each location and irregularly sampled in time.
- **Increase in the Number of Lags in the AR Models:**
  The AR models we used in this thesis have auto-regressive patterns with just lag one. It is possible to increase the number of lags used in the model.
- **Other Approaches to Solve the big-n Problem:**
  To tackle the *big-n problem*, we use predictive process approximation. In addition, we can also use the fixed rank kriging method proposed by Cressie and Johannesson (2008) with usual basis functions, details are discussed in Section 1.6.
- **Multivariate Extension of the Spatio-Temporal Models:**
  The models adopted in this thesis are for univariate spatio-temporal data.

These models can be extended to the multivariate settings, where at each spatial location we have temporal observations of two or more response variables. The multivariate space-time random effect can be specified using a linear model of coregionalisation, see e.g., Gelfand *et al.* (2004); Reich and Fuentes (2007). An alternative of this method is to specify the multivariate response conditionally, see e.g., Daniels *et al.* (2006) where ozone concentration levels and particulate matter data have been modelled jointly.

Currently, `spTimer` cannot fit multivariate space-time models. So, it is possible to extend our spatio-temporal package `spTimer` to model multivariate space-time data.

- **Extend Models for Spatial Misalignment:**
  In this thesis, spatial misalignments between predictor and predictand are handled using independent kriging (see Section 6.6.4). However, new models can be proposed that can take care of the misalignment through sampling from the joint posterior distribution of the parameters of the joint spatio-temporal model (Sahu *et al.*, 2007; Sahu and Nicolis, 2009; Lopiano *et al.*, 2011).

- **Non-Gaussian Models:**
  In this thesis, we only discuss the Gaussian process modelling methodology for analysing ozone concentration data. However, it is possible to extend the models for non-Gaussian distributions (e.g., generalised linear models) at the first stage of modelling hierarchy. Following Salway *et al.* (2010) we can also model the latent process using AR and moving average (MA) techniques.

  Simple regression type non-Gaussian models are available in the package `spBayes` for modelling spatial data sets. Henceforth, our package `spTimer` can be improved by including the non-Gaussian models in the first-stage of modelling for the space-time data.

- **Modelling the Extreme Observations:**
  The ozone data set used in this thesis is positively skewed for the high variability in the data. We use square root transformation of the original data

to make the Gaussian assumption appropriate (see Section 1.5.3). However, it is possible to use models based on extreme value theory to analyse the extremes in ozone levels. For example, Ghosh and Mallick (2011) used hierarchical spatio-temporal model to incorporate spatial correlation in the likelihood and used temporal component at the second level of hierarchy to analyse monthly rainfall data.

- **Extend Space-Time Models for Stream Networks:**
  In modelling observations obtained from a river network, the Euclidean and geodetic distances may not be valid because of the pattern of the water flow. Hoef and Peterson (2010) developed spatial moving average approach to model stream networks using spatial covariance function that is based on stream distances. Henceforth, we can extend our models in this context and also improve our package `spTimer`.

- **Other Application Areas:**
  The spatio-temporal models developed in this thesis have been applied to analyse ozone concentration levels. These models and their modifications, however, can be applied to model data for other air pollutants such as particulate matter. Other types of spatio-temporal data such as many meteorological and climate observations, such as rainfall, can also be modelled and analysed using these models.

Bibliography

Ashmore, M.R. (2005). Assessing the future global impacts of ozone on vegetation. *Plant Cell Environment*, **28**, 949-964.

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

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, Series: B*, **70**, 825-848.

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

Berrocal, V.J., Gelfand, A.E., and Holland, D.M. (2010a). A spatio-temporal downcsaler for outputs from numerical models. *Journal of Agricultural, Biological and Environmental Statistics*, **15**, 176-197.

Berrocal, V.J., Gelfand, A.E., and Holland, D.M. (2010b). A bivariate space-time downcsaler under space and time misalignment. *Annuals of Applied Statistics*, **4**, 1942-1975.

Black, T.L. (1994). The new NMC mesoscale Eta Model: Description and forecast examples. *Weather and Forecasting*, **9**, 265-278.

Bloomfield, P., Royle, A., and Yang, Q. (1996) Accounting for Meteorological Effects in Measuring Urban Ozone Levels and Trends. *Atmospheric Environment*, **30**, 3067-3077.

Bochner, S. (1955). *Harmonic Analysis and the Theory of Probability*. University of California, Berkley and Los Angeles.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees*. Chapman and Hall, London.

Bruno, F., Guttorp, P., Sampson, P.D. and Cocchi, D. (2009). A Simple Non-separable, Non-stationary Spatiotemporal Model for Ozone. *Environmental and Ecological Statistics*, **16**, 515-529.

Camalier, L. Cox, W. and Dolwick, P. (2007). The effects of meteorology on ozone in urban areas and their use in assessing ozone trends. *Atmospheric Environment*, **41**, 7127-7137.

Cassmassi, J.C., and Bassett, M. (1991). Air quality trends in the South Coast Air Basin, in Southern California Air Quality Study Data Analysis: Proceedings of an International Specialty Conference, *Air and Waste Management Association*, Pittsburgh.

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.

Chan, C.K. and Yao, X. (2008). Air pollution in mega cities in China. *Atmospheric Environment*, **42(1)**, 1–42.

Chen, M., Saho, Q, and Ibrahim, J.G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag, New York.

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

Christensen O.F., and Ribeiro, Jr P.J. (2002). `georglm`: A Package for Generalised Linear Spatial Models. *R News*, **2(2)**, 26–28.

Ching, J., and Byun D. (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.*

Papamichael, C. (2011). Bayesian Spatial-Temporal Modelling of Air Pollution. *PhD Thesis*. University of Bath, UK.

Cocchi, D., Fabrizi, E., and Trivisano, C. (2005). A Stratified Model for the Assessment of Meteorologically Adjusted Trends of Surface Ozone. *Environmental and Ecological Statistics*, **12**, 195-208.

Cogliani, E. (2001). Air pollution forecast in cities by an air pollution index highly correlated with meteorological variables. *Atmospheric Environment*, **35(16)**, 2871–2877.

Condit, R. (1998). *Tropical Forest Census Plots*. Springer-Verlag, Berlin and R.G. Landes Company, Georgetown, Texas.

Cox, W.M., and Chu, S.H. (1993). Meteorological Adjusted Trends in Urban Areas, a Probabilistic Approach. *Atmospheric Environment*, **27B**, 425-434.

Cressie, N.A.C. (1993). *Statistics for Spatial Data*. Revised Edition. John Wiley and Sons, New York.

Cressie, N.A.C. and Huang, H.C. (1999). Classes of Nonseparable, Spatio-temporal Stationary Covariance Functions. *Journal of the American Statistical Association*, **94**, 1330-1340.

Cressie, N.A.C. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society, Series: B*, **70**, 209-226.

Daniels, M.J., Zhou, Z. G., and Zou, H. (2006) Conditionally specified spacetime models for multivariate processes. *Journal of Computational and Graphical Statistics*, **15**, 157-177.

Davis, J., Eder, B., Nychka, D. and Yang, Q. (1998). Modeling the effects of meteorology on ozone in Houston using cluster analysis and generalized additive models. *Atmospheric Science*,**32**, 2505-2520.

Davis, J.M. and Speckman, P. (1999). A model for predicting maximum and 8 h average ozone in Houston. *Atmospheric Environment*, **33**, 2487-2500.

Dempster, A.P. (1974). The direct use of likelihood for significance testing. In *Proceedings of Conference on Foundational Questions in Statistical Inference*, Eds. O. Barndorff-Nielsen, P. Baesild, & G. Schou, 335-352. Department of Theoretical Statistics, University of Aarhus.

De Oliveria, V., Kedem, B. and Short, D.A. (1997). Bayesian Prediction of Transformed Gaussian Random Fields. *Journal of the American Statistical Association*, **92**, 1422-1433.

Diggle, P.J., and Lophaven, S. (2006). Bayesian geostatistical design. *Scandinavian Journal of Statistics*. **33**, 55-64.

Diggle, P.J. and Ribeiro, P.J. (2007). *Model-based Geostatistics*. Springer, New York.

Dingenena, R.V., Dentenera, F.J., Raesa, F., Krolb, M.C., Embersonc, L. and Cofalad, J. (2009). The global impact of ozone on agricultural crop yields under current and future air quality legislation. *Atmospheric Environment*, **43(3)**, 604-618.

Dou, Y., Le, N.D. and Zidek J.V. (2010). Modeling Hourly Ozone Concentration Fields. *Annals of Applied Statistics*, **4**, 1183-1213.

Dou, Y., Le, N.D. and Zidek J.V. (2011). Temporal prediction with a Bayesian spatial predictor: An application to ozone fields. To appear.

Ecker, M.D. and Gelfand, A.E. (1997). Bayesian Variogram Modelling for an Isotropic Spatial Process. *Journal of Agricultural, Biological and Environmental Statistics*, **2**, 347-369.

Feister, U., and Balzer, K. (1991). Surface ozone and meteorological predictors on a subregional scale. *Atmospheric Environment*, **25**, 1781-1790.

Fields Development Team (2006). `fields`: Tools for Spatial Data. National Center for Atmospheric Research, Boulder, CO. URL http://www.cgd.ucar.edu/Software/Fields.

Finkenstadt, B., Held, L., and Valerie, I. (2007). *Statistical Methods for Spatio-temporal Systems*. Monograph on Statistics and Applied Probability 107, Chapman and Hall/CRC.

Finley, A.O., Banerjee, S., and Carlin, B.P. (2007). `spBayes`: An `R` package for univariate and multivariate hierarchical point-referenced spatial models. `Journal of Statistical Software`, **19(4)**, 1-24.

Finley, A.O., Sang, H., Banerjee, S., and Gelfand, A.E. (2009). Improving the performance of predictive process modeling for large datasets. *Computational Statistics and Data Anlaysis*, **53**, 2873-2884.

Fioletov, V.E., Bodeker, G.E., Miller, A.J., McPeters, R.D., and Stolarski, R. (2002). Global and zonal total ozone variations estimated from ground-based and satellite measurements: 1964-2000, *Journal of Geophysical Research (Atmospheres)*, **107(D22)**, ACH 21-1, CiteID 4647.

Fiore, A.M., Jacob, D.J., Logan, J.A., Yin, J.H. (1998). Long-term trends in ground level ozone over the contiguous United States, 1980-1995. *Journal of Geophysical Research*, **103**, 1471-1480.

Folinsbee L.J., McDonnell, W.F., and Horstman, D.H. (1988). Pulmonary function and symptom responses after 6.6-hour exposure to 0.12 ppm ozone with moderate exercise. *Journal of Air Pollution Control Association*, **38(1)**, 28-35.

Fuentes, M. (2002). Spectral Methods for Nonstationary Spatial Process. *Biometrika*, **89**, 197-210.

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.

Galbally, I.E., Miller, A.J., Hoy, R.D., Ahmet, S., Joynt, R.C., and Attwood, D. (1986). Surface ozone at rural sites in the Latrobe Valley and Cape Grim, Australia. *Atmospheric Environment*, **20**, 2403-2422.

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

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, **85**, 398-409.

Gelfand, A. E. Schmidt, A. M. Banerjee, S. and Sirmans, C. F. (2004). Nonstationary Multivariate Process Modelling through Spatially Varying Coregionalization (with discussion). *Test*, **2**, 1–50.

Gelfand, A.E., Diggle, P.J., Guttorp, P., and Fuentes, M. (2010). *Handbook of Spatial Statistics*. Chapman and Hall/CRC Handbooks of Modern Statistical Methods, Boca Raton.

Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457-472.

Gelman, A., Gilks, W.R., and Roberts, G.O. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, **7**, 110-120.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004). *Bayesian Data Analysis*, Second edition, Chapman and Hall/CRC.

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.

Ghosh, S., and Mallick, B.K. (2011). A hierarchical Bayesian spatio-temporal model for extreme precipitation events. *Environmetrics*, **22(2)**, 192-204.

Gneiting, T. (2002). Nonseparable, Stationary Covariance Functions for Space-time Data. *Journal of the American Statistical Association*, **97**, 590-600.

Guhaniyogi, R., Finley, A.O., Banerjee, S. and Gelfand, A.E. (2011). Adaptive Gaussian predictive process models for large spatial datasets. it Environmetrics, **22(8)**, 997–1007.

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

Handcock, M.S. and Stein, M.L. (1993). A Bayesian Analysis of Kriging. *Technometrics*, **35**, 403-410.

Hartman, L., and Hossjer, O. (2008). Fast kriging of large data sets with Gaussian Markov random fields. *Computational Statistics and Data Analysis*, **52**, 2331-2349.

Hass, T.C. (1995). Local prediction of a spatio-temporal process with an application to wet sulfet deposition. *Journal of the Americal Statistical Association*, **90**, 1189-1199.

Hastie, T. (1996). Pseudosplines. *Journal of the Royal Statistical Society, Series: B*, **58**, 379-396.

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

Hoef, J.M.V., and Peterson, E. (2010). A Moving Average Approach for Spatial Statistical Models of Stream Networks. *Journal of the American Statistical Association*, **105(489)**, 6-18.

Horstman, D.H., Folinsbee, L.J., Ives, P.J., Abdul-Salaam S., and McDonnell, W.F. (1990). Ozone concentration and pulmonary response relationships for 6.6-hour exposures with five hours of moderate exercise to 0.08, 0.10, and 0.12 ppm. *American Review of Respiratory Disease*, **142(5)**, 1158-1163.

Houyoux M.R., Vukovich, J., and Brandmeyer, J. (2000). Sparse Matrix Operator Kernel Emissions Modeling System (SMOKE) user manual. *MCNC-North Carolina Supercomputing Center*, Environmental Programs, Research Triangle Park, NC.

Huang, L., and Smith, R.L. (1999). Meteorologically-dependent trends in urban ozone. *Environmetrics*. **10**, 103-118.

Huang, H.C., Cressie, N., and Gabrosek, J. (2002). Fast, resolution-consistent spatial prediction of global processes from satellite data. *Journal of Computational and Graphical Statistics*, **11**, 63-88.

Huerta, G., Sanso, B., and Stroud, J.R. (2004). A Spatiotemporal Model for Maxico City Ozone Levels. *Journal of the Royal Statistical Society, Series: C*, **53(2)**, 231-248.

IPCC (Intergovernmental Panel on Climate Change). (2007a). *Climate Change 2007 – The Physical Science Basis*. Fourth Assessment Report, Working Group-I, Geneva.

IPCC (Intergovernmental Panel on Climate Change). (2007b). *Climate Change 2007 – Synthesis Report*. Fourth Assessment Report, Geneva.

Johnson, M.E., Moore, L.M., and Ylvisaker, D. (1990). Minimum and maximum distance designs. *Journal of Statistical Planning and Inference*, **26**, 131-148.

Johannesson, G. and Cressie, N. (2004). Finding large-scale spatial trends in massive, global, environmental datasets. *Environmetrics*, **15**, 1-44.

Johannesson, G., Cressie, N. and Huang, H.C. (2007). Dynamic multi-resolution spatial models. *Environmental and Ecological Statistics*, **14**, 5-25.

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

Kalman, R.E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, Transaction of the American Society Of Mechanical Engineers(ASME), 35-45.

Kammann, E.E., and Wand, M.P. (2003). Geoadditive models. *Journal of the Royal Statistical Society, Series: C*, **52(1)**, 1-18.

Khanna, N. (2000). Measuring environmental quality: an index of pollution. *Ecological Economics*, **35(2)**, 191-202.

Kitanidis, P.K. (1986). Parameter Uncertainty in Estimation of Spatial Functions: Bayesian Analysis. *Water Resources Res*, **22**, 449-507.

Korsog, P.E., and Wolff, G.T. (1991). An examination of urban ozone trends in the northeastern US (1973-1983) using a robust statistical method. *Atmospheric Environment*, **25**, 47-57.

Krige, D.G. (1951). A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, **52**, 119-139.

Le, N.D., and Zidek, J.Z. (1992). Interpolation with Uncertain Spatial Covariance - A Bayesian Alternative to Kriging. *Journal of Multivariate Analysis*. **43(2)**, 351-374.

Le, N.D. and Zidek, J.Z. (2006). *Statistical Analysis of Environmental Space-time process*. Springer, New York.

Lee, D., Ferguson, C. and Scott, E.M. (2011). Constructing representative air quality indicators with measures of uncertainty. *Journal of the Royal Statistical Society, Series: A*, **174(1)**, 109-126.

Lindgren, F., and Rue, H. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society, Series: B*, **73(4)**, 423-498.

Liu, Z., Le, N. D., and Zidek, J. V. (2011). An empirical assessment of Bayesian melding for mapping ozone pollution. *Environmetrics*, **22**, 340-353.

Lopiano, K.K., Young, L.J., and Gotway, C.A. (2011). A comparison of errors in variables methods for use in regression models with spatially misaligned data. *Statistical Methods in Medical Research*, **20**, 29–47.

Lorence, A.C. (1986). Analysis methods for numerical weather prediction. *The Quarterly Journal of the Royal Meteorological Society*, **112**, 1177-1194.

Matérn, B. (1986). *Spatial variation* (2nd ed.), Lecture notes in statistics, Springer, Berlin.

Matheron, G. (1963). Principles of Geostatistics. *Economic Geology*, **58(8)**, 1246-1266.

McDonnell, W.F., Kehrl, H.R., Abdul-Salaam, S., Ives, P.J., Folinsbee, L.J., Devlin, R.B., O'Neil, J.J., and Horstman D.H. (1991). Respiratory response of humans exposed to low levels of ozone for 6.6 hours. *Archives of Environmental Health*, **46(3)**, 145–150.

McGarth, W.D., and Norrish, R.G.W. (1957). The flash photolysis of ozone. *Proceedings of the Royal Society of London. Series A*, Mathematical and Physical Sciences, **242(1230)**, 265-276.

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.

Mohanakumar, K. (2008). *Stratosphere Troposphere Interactions: An Introduction*. Springer Science, Business Media B.V.

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, Series. B*, **56**, 3-48.

NRDC (National Resources Defense Council). (2004). *Heat advisory: how global warming causes more bad air days*. Washington, DC.

Nychka, D., Bailey, B., Ellner, S., Haaland, P., and O'Connell, M. (1996). *FUNFITS: Data analysis and statistical tools for estimating functions*. Raleigh: North Carolina State University.

Ott, W.R. (1978). *Environmental indices: theory and practice*. Ann Arbor Science Publishers, Inc.,Ann Arbor, MI.

Otte, T.L., Pouliot, G., Pleim, J.E., Young, J.O., Schere, K.L., Wong, D.C., Lee, P.C.S, Tsidulko, M., McQueen, J.T, Davidson, P., Mathur, R., Chuang, H., DiMego, G., and Seaman, N.L. (2005). Linking the Eta Model with the Community Multiscale Air Quality (CMAQ) Modeling System to Build a National Air Quality Forecasting System. *Weather and Forecasting*, **20(3)**, 367-384.

Paciorek, C.J. (2007). Computational techniques for spatila logistic regression with large datasets. *Computational Statistics and Data Analysis*, **51**, 3631-3653.

Pardo-Iguzquiza, E., and Dowd, P.A. (1997). AMLE3D: A Computer Program for the Interface of Spatial Covariance Parameters by Approximate Maximum Likelihood Estimation. *Computers and Geosciences*, **23**, 793-805.

Reich, B.J., and Fuentes, M. (2007). A Multivariate Semiparametric Bayesian Spatial Modeling Framework for Hurricane Surface Wind Fields. *The Annals of Applied Statistics*, **1(1)**, 249-264.

Reich, B.J., Fuentes, M., and Dunson, D.B. (2011). Bayesian Spatial Quantile Regression. *Journal of the American Statistical Association*, **106**, 6-20.

Ribeiro, Jr. P.J., and Diggle, P.J. (2001). `geoR`: A Package for Geostatistical Analysis. *R News*, **1(2)**, 14–18.

Rogers, E., Deaven, D.G., and DiMego, G.J. (1996). The regional analysis system for the operational eta model: Original 80 km configuration and recent changes. *Weather Forecasting*, **10**, 810–825.

Rouhani, S., and Myers D.E. (1990). Problems in Space-time Kriging of Geohydrological Data. *Mathematical Geology*. **22**, 611-623.

Rue, H., and Held, L. (2006). *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton: Chapman and Hall/CRC.

Salway, R., Lee, D., Shaddick, G., and Walker, S. (2010). Bayesian latent variable modelling in studies of air pollution and health. *Statistics in Medicine*, **29**, 2732-2742.

Sahu, S.K., Gelfand, A.E., and Holland, D.M. (2007). High-Resolution Space-Time Ozone Modeling for Assessing Trends. *Journal of the American Statistical Association*, **102**, 1221-1234.

Sahu, S.K., and Challenor, P. (2008). A space-time model for joint modeling of ocean temperature and salinity levels as measured by Argo floats. *Environmetrics*, **19**, 509-528.

Sahu, S.K., and Nicolis, O. (2009). An evaluation of European air pollution regulations for particulate matter monitored from a heterogeneous network. *Environmetrics*,**20(8)**, 943–961,

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. (2011). Hierarchical Bayesian models for space-time air pollution data In Handbook of Statistics-Vol 30. *Time Series Analysis, Methods and Applications*. Editors: T Subba Rao and C R Rao. Elsevier Publishers, Holland. To appear.

Shekhar, S., and Xiong, H. (2008). *Encyclopedia of GIS*. Springer, New York.

Sitch, S., Cox, P.M., Collins, W.J., and Huntingford, C. (2007). Indirect Radiative Forcing of Climate Change through Ozone Effects on the Land-Carbon Sink. *Nature*, **448**, 791-794.

Sousa, S.I.V., Pires, J.C.M., Martins, F.G., Pereira, M.C., and Alvim-Ferraz, M.C.M. (2009). Potentialities of quantile regression to predict ozone concentrations. *Environmetrics*, **20**, 147-158.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and Van der Linde (2002). Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society: Series B*, **64**, 538-616.

Stein, M.L. (1999). *Statistical Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer.

Stein, M.L. (2005). Space-time Covariance Functions. *Journal of the American Statistical Association*, **100**, 310-321.

Stein, M.L. (2007). Spatial variation of total column ozone on a global scale. *Annals of Applied Statistics*, **1**, 191-210.

Stein, M.L. (2008). A modelling approach for large spatial datasets. *Journal of the Korian Statistical Society*, **37**, 3-10.

Stephenson, J. (2006). Non-stationary Spatial Statistics in the Geosciences. *PhD Thesis*. Department of Earth Science and Engineering, Imperial College London.

Stroud, J.R., Muller, P., and Sanso, B. (2001). Dynamic Models for Spatio-Temporal Data. *Journal of the Royal Statistical Society: Series B*, **63**, 673-689.

USEPA (United States Environmental Protection Agency), (1996). *Air quality criteria for O_3 and related photochemical oxidants*. EPA/600/P-93/004a-cF, Washington, DC.

USEPA (United States Environmental Protection Agency), (1998). *Guideline on data handling conventions for the 8-hour ozone*. EPA-454/R-98-017, Washington, DC.

USEPA (United States Environmental Protection Agency), (1999a). *Smog Who Does It Hurt? What You Need to Know About Ozone and Your Health*. EPA-452/K-99-001, Washington, DC.

USEPA (United States Environmental Protection Agency), (1999b). *Air quality index reporting; final rule*. Federal Register, Part III, 40 CFR Part 58, Washington, DC.

USEPA (United States Environmental Protection Agency), (2004). *The Ozone Report: Measuring Progress through 2003*. EPA-454/K-04-001, Washington, DC.

West, M., and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models, 2nd Edition*. New York: Springer.

WHO (World Health Organization), (1979). Photochemical Oxidants. *Environmental Health Criteria 7*. Geneva.

WHO (World Health Organization), (1987). *Air Quality Guidelines for Europe*. Copenhagen: WHO Regional Office for Europe.

Yip, C.Y. (2009). Bayesian Spatio-temporal Modelling for Forecasting Ground Level Ozone Concentrations. *PhD Thesis*. School of Mathematics, University of Southampton.

Xia, G., and Gelfand, A.E. (2006). Stationary Process Approximation for the Analysis of Large Spatial Datasets. *Technical Report*. Institute of Statistical and Decision Sciences, Duke University, Durham, USA.

Zheng, J., Swall, J. Cox, W.M., and Devis, J.M. (2007). Interannual Variation in Meteorological Adjusted Ozone Levels in the Eastern United State: A Comparison of two Approaches. *Atmospheric Environment*, **41**, 705-716.

Zidek, J.V., Le, N.D., and Liu, Z. (2011). Combining data and simulated data for space-time fields: application to ozone. *Environmental and Ecological Statistics*. In Press.

Appendix A

Proofs for Chapter 5

A.1 Results Related to the Correlation Function

For the autoregressive (AR) models, the correlation between the data points $Z(s_i,t)$ and $Z(s_j,t)$ can be written as:

$$\begin{aligned}\mathrm{Cor}(Z(s_i,t),Z(s_j,t)) &= \frac{\mathrm{Cov}(Z(s_i,t),Z(s_j,t))}{\sqrt{\mathrm{Var}(Z(s_i,t))\times\mathrm{Var}(Z(s_j,t))}} \\ &= \frac{\rho^{2t}\sigma_o^2\exp[-\phi_0 d_{ij}]+\left[\frac{1-\rho^{2t}}{1-\rho^2}\right]\sigma_\eta^2\exp[-\phi_\eta d_{ij}]}{\rho^{2t}\sigma_0^2+\left[\frac{1-\rho^{2t}}{1-\rho^2}\right]\sigma_\eta^2+\sigma_\epsilon^2},\end{aligned}$$

where, $i \neq j$, and t=1,2,...T.

- For increase in time t, the correlation between the observations $Z(s_i,t)$ and $Z(s_j,t)$ is written by simple calculation as;

$$\lim_{t\to\infty}\mathrm{Cor}(Z(s_i,t),Z(s_j,t)) = \frac{\sigma_\eta^2\exp(-\phi_\eta d_{ij})}{\sigma_\epsilon^2(1-\rho^2)+\sigma_\eta^2}.$$

- For the AR models, the correlation between the observations $Z(s_i,t)$ and $Z(s_j,t)$ tends to one, for the decrease in the distance d_{ij} to zero, where d_{ij} is the distance between the observations of sites s_i and s_j, $i \neq j$ i.e.,

$$\lim_{d_{ij}\to 0}\mathrm{Cor}(Z(s_i,t),Z(s_j,t)) = 1.$$

- When d_{ij} increases to infinity, the correlation between the observations

$Z(s_i,t)$ and $Z(s_j,t)$ tends to zero, i.e.,

$$\lim_{d_{ij}\to\infty} \mathrm{Cor}(Z(s_i,t), Z(s_j,t)) = 0, \quad i \neq j.$$

A.2 Expression for the Conditional Variances

We can write the predictive conditional variance of $Z(s_0,1)$ given observations $Z(s_1,1)$, $Z(s_1,2)$ as:

$$\mathrm{Var}(Z(s_0,1)|z(s_1,1), z(s_1,2)) = \frac{M_1}{\Delta_2}, \tag{A.1}$$

and the predictive conditional variance of $Z(s_0,2)$ given observations $Z(s_1,1)$, $Z(s_1,2)$ can be written as:

$$\mathrm{Var}(Z(s_0,2)|z(s_1,1), z(s_1,2)) = \frac{M_2}{\Delta_2}, \tag{A.2}$$

where,

$$\begin{aligned}
M_1 &= (1-\zeta^2)\sigma_\eta^6 + (3+\rho^2-\zeta^2(1+\rho^2))\sigma_\eta^4\sigma_\epsilon^2 + (3+\rho^2)\sigma_\eta^2\sigma_\epsilon^4 + \sigma_\epsilon^6 + \\
&\quad (1-\zeta^2)\sigma_0^4\rho^4(\sigma_\eta^2 + \sigma_\epsilon^2 + \rho^2\sigma_\epsilon^2) + \\
&\quad \rho^2\sigma_0^2(2(1-\zeta^2)\sigma_\eta^4 - 2(\zeta^2(1+\rho^2) - 2 - \rho^2)\sigma_\eta^2\sigma_\epsilon^2) + (2+\rho^2)\sigma_\epsilon^4. \\
M_2 &= (1-\zeta^2)(1+\rho^2)\sigma_\eta^6 + (3+3\rho^2+\rho^4-\zeta^2(1+3\rho^2+\rho^4))\sigma_\eta^4\sigma_\epsilon^2 + (3+2\rho^2)\sigma_\eta^2\sigma_\epsilon^4 + \\
&\quad \sigma_\epsilon^6 + (1-\zeta^2)\sigma_0^4\rho^6(\sigma_\eta^2 + \sigma_\epsilon^2 + \rho^2\sigma_\epsilon^2) + \rho^2\sigma_0^2(2(1-\zeta^2)(1+2\rho^2)\sigma_\eta^4 + \\
&\quad 2(1+2(1-\zeta^2)\rho^2 + (1-\zeta^2)\rho^4)\sigma_\eta^2\sigma_\epsilon^2 + (1+2\rho^2)\sigma_\epsilon^4). \\
\Delta_2 &= \sigma_\eta^4 + (2+\rho^2)\sigma_\eta^2\sigma_\epsilon^2 + \sigma_\epsilon^4 + \rho^2\sigma_0^2(\sigma_\eta^2 + \sigma_\epsilon^2 + \rho^2\sigma_\epsilon^2).
\end{aligned}$$

where, $\zeta = \exp(-\phi d_{01})$ and $\phi = \phi_0 = \phi_\eta$.

A.3 Proof of Inequalities

A.3.1 Inequalities Related to Predictions

(i) *For the AR models, conditioned on the same amount of data, the predictive variance of $Z(s_0, 1)$ would be no greater than that of $Z(s_0, 2)$, that is,*

$$Var(Z(s_0,1)|Z(s_1,1), Z(s_1,2)) \leq Var(Z(s_0,2)|Z(s_1,1), Z(s_1,2)).$$

if the following condition holds:

$$\frac{\sigma_\eta^2}{\sigma_0^2} \geq 1 - \rho^2.$$

Proof. The difference between the terms of the equations (A.1) and (A.2) are:

$$\text{Var}(Z(s_0,2)|z(s_1,1), z(s_1,2)) - \text{Var}(Z(s_0,1)|z(s_1,1), z(s_1,2)) = \frac{(\sigma_0^2(-1+\rho^2) + \sigma_\eta^2) \times \text{A}}{\Delta_2}$$

where, Δ_2 is defined above and

$$\text{A} = \rho^2((1-\zeta^2)\sigma_\eta^4 + (1-\zeta^2)(2+\rho^2)\sigma_\eta^2\sigma_\epsilon^2 - \sigma_\epsilon^4 + (1-\zeta^2)\rho^2\sigma_0^2(\sigma_\eta^2 + \sigma_\epsilon^2 + \rho^2\sigma_\epsilon^2))$$

The terms A and Δ_2 are always positive for all values of $\sigma_0^2 \geq 0$, $\sigma_\eta^2 \geq 0$, $\sigma_\epsilon^2 \geq 0$, $0 < \rho < 1$ and $0 < \zeta < 1$. So, the predictive variance differences

$$\text{Var}(Z(s_0,2)|z(s_1,1), z(s_1,2)) - \text{Var}(Z(s_0,1)|z(s_1,1), z(s_1,2)) \geq 0$$

iff the following condition holds:

$$\frac{\sigma_\eta^2}{\sigma_0^2} \geq 1 - \rho^2.$$

□

(ii) *For the AR models, following the equation 5.9, we can write,*

$$\text{Dif} = \text{Var}(Z(s_0,1)|Z(s_1,1)) - \text{Var}(Z(s_0,2)|Z(s_1,1), Z(s_1,2)) > 0.$$

(a) Dif < 0, as $\sigma_0^2 \to 0$.

(b) Dif > 0, as $\zeta \to 1$, iff $\frac{\sigma_\epsilon^2}{\sigma_0^2} < \frac{\kappa+\rho^2}{\kappa-(1-\rho^2)}$.

(c) Dif < 0, as $\zeta \to 0$, and $\sigma_0^2 = (1-\rho^2)\sigma_\eta^2$.

(d) Dif > 0, as $\zeta \to 1$, and $\sigma_0^2 = (1-\rho^2)\sigma_\eta^2$ iff $\frac{\sigma_\epsilon^2}{\sigma_\eta^2} < \frac{1-2\rho^4+\rho^6}{\rho^2(2-\rho^2)}$.

Proof. Proofs of (a) to (d) are given below:

(a) For $\sigma_0^2 \to 0$, we obtain the term Dif as:

$$\lim_{\sigma_0^2 \to 0} \text{Dif} = \frac{\rho^2\sigma_\eta^2((\zeta^2-1)\sigma_\eta^2(1+(3+\rho^2)\sigma_\epsilon^2(\sigma_\eta^2+\sigma_\epsilon^2))-\sigma_\epsilon^6)}{(\sigma_\eta^2+\sigma_\epsilon^2)(\sigma_\eta^4+(2+\rho^2)\sigma_\eta^2\sigma_\epsilon^2+\sigma_\epsilon^4)}$$

Clearly, the numerator of the above equation is negative as $0 < \zeta < 1$, and the denominator is positive for all values of $\sigma_\eta^2 > 0$, $\sigma_\epsilon^2 > 0$, and $0 < \rho < 1$. So, we can write, Dif < 0 as $\sigma_0^2 \to 0$.

(b) For $\zeta \to 1$, the Dif is:

$$\lim_{\zeta \to 1} \text{Dif} = \frac{\rho^2\sigma_\epsilon^2(\rho^2\sigma_0^4-\sigma_\eta^2\sigma_\epsilon^2+\sigma_0^2(\sigma_\eta^2+\sigma_\epsilon^2-\rho^2\sigma_\epsilon^2))}{(\rho^2\sigma_0^2+\sigma_\eta^2+\sigma_\epsilon^2)(\sigma_\eta^4+(2+\rho^2)\sigma_\eta^2\sigma_\epsilon^2+\sigma_\epsilon^4+\rho^2\sigma_0^2(\sigma_\eta^2+\sigma_\epsilon^2+\rho^2\sigma_\epsilon^2))}$$

The denominator of the above equation is positive and the numerator will be positive iff:

$$\rho^2\sigma_\epsilon^2(\rho^2\sigma_0^4-\sigma_\eta^2\sigma_\epsilon^2+\sigma_0^2(\sigma_\eta^2+\sigma_\epsilon^2-\rho^2\sigma_\epsilon^2)) > 0$$

A simple calculation yields the above term is positive iff

$$\frac{\sigma_\epsilon^2}{\sigma_0^2} < \frac{\kappa+\rho^2}{\kappa-(1-\rho^2)}$$

where, $\kappa = \frac{\sigma_\eta^2}{\sigma_0^2}$.

(c) Considering σ_0^2 to be the equilibrium variance $(1-\rho^2)\sigma_\eta^2$, and spatial correlation $\zeta \to 0$ as zero, we get:

$$\lim_{\zeta \to 0} \text{Dif} = -\rho^4(2-\rho^2)\sigma_\eta^2, \quad \text{when} \quad \sigma_0^2 = (1-\rho^2)\sigma_\eta^2$$

Which is always negative.

(d) Again, for $\sigma_0^2 = (1-\rho^2)\sigma_\eta^2$, and large spatial correlation $\zeta \to 1$, we get:

$$\lim_{\zeta \to 0} \text{Dif} = \frac{\rho^2\sigma_\eta^2\sigma_\epsilon^4((\rho^6-2\rho^4+1)\sigma_\eta^2+\rho^2\sigma_\epsilon^2(\rho^2-2))}{((\rho^4-\rho^2-1)\sigma_\eta^2-\sigma_\epsilon^2)((\rho^4-\rho^2-1)\sigma_\eta^4+(\rho^6-2\rho^2-2)\sigma_\eta^2\sigma_\epsilon^2-\sigma_\epsilon^4)}$$

It can be easily calculate that the denominator of the above equation is positive for values of $0 < \rho < 1$, σ_η^2 and σ_ϵ^2. Hence the inequality Dif > 0, iff the numerator of the above equation is positive. Simple calculation leads, it is positive iff:

$$\frac{\sigma_\epsilon^2}{\sigma_\eta^2} < \frac{1 - 2\rho^4 + \rho^6}{\rho^2(2 - \rho^2)}.$$

Hence, all propositions are proved. □

A.3.2 Inequalities Related to Forecasts

(i) *We obtain the following inequalities for both the DLM and the AR models.*

$$\begin{aligned} Var(Z(s_0,3)|Z(s_1,1),Z(s_1,2)) &\leq Var(Z(s_0,3)|Z(s_1,2)) \\ Var(Z(s_0,3)|Z(s_1,2),Z(s_2,2)) &\leq Var(Z(s_0,3)|Z(s_1,2)). \end{aligned}$$

Proof. **DLM**
For the DLM we can write:

$$\begin{aligned} \mathrm{Var}(Z(s_0,3)|Z(s_1,1),Z(s_1,2)) &= \frac{\sigma_\nu^2 + 6\sigma_\nu^4\sigma_\omega^2 + 5\sigma_\nu^2\sigma_\omega^4 + \sigma_\omega^6 + \sigma_\theta^2(3\sigma_\nu^4 + 4\sigma_\nu^2\sigma_\omega^2 + \sigma_\omega^4)}{\sigma_\nu^2 + 3\sigma_\nu^2\sigma_\omega^2 + \sigma_\omega^4 + \sigma_\theta^2(2\sigma_\nu^2 + \sigma_\omega^2)} \\ \mathrm{Var}(Z(s_0,3)|Z(s_1,2),Z(s_2,2)) &= \frac{(1+\zeta_0)\sigma_\nu^4 + (7+3\zeta_0)\sigma_\nu^2\sigma_\omega^2 + 4\sigma_\omega^4 + \sigma_\theta^2((3+\zeta_0)\sigma_\nu^2 + 2\sigma_\omega^2)}{2\sigma_\theta^2 + \sigma_\nu^2 + \zeta_0\sigma_\nu^2 + 4\sigma_\omega^2} \end{aligned}$$

So, the difference between $\mathrm{Var}(Z(s_0,3)|Z(s_1,2))$ and $\mathrm{Var}(Z(s_0,3)|Z(s_1,1),Z(s_1,2))$ and $\mathrm{Var}(Z(s_0,3)|Z(s_1,2))$ and $\mathrm{Var}(Z(s_0,3)|Z(s_1,2),Z(s_2,2))$ can be written as:

$$\begin{aligned} \mathrm{Var}(Z(s_0,3)|Z(s_1,2)) &- \mathrm{Var}(Z(s_0,3)|Z(s_1,1),Z(s_1,2)) = \\ &\frac{\sigma_\nu^4(\sigma_\omega^4 + \sigma_\theta^2)^2}{(\sigma_\theta^2 + \sigma_\nu^2 + 2\sigma_\omega^2)(\sigma_\nu^4 + 3\sigma_\nu^2\sigma_\omega^2 + \sigma_\omega^4 + \sigma_\theta^2(2\sigma_\nu^2 + \sigma_\omega^2))}. \\ \mathrm{Var}(Z(s_0,3)|Z(s_1,2)) &- \mathrm{Var}(Z(s_0,3)|Z(s_2,1),Z(s_2,2)) = \\ &\frac{\sigma_\nu^2(1-\zeta_0)(\sigma_\theta^2 + 2\sigma_\omega^2)^2}{(\sigma_\theta^2 + \sigma_\nu^2 + 2\sigma_\omega^2)(2\sigma_\theta^2 + \sigma_\nu^2 + \zeta_0\sigma_\nu^2 + 4\sigma_\theta^2)} \end{aligned}$$

where, $\zeta_0 = \exp(-\phi d_{12})$ is the spatial correlation between the observations at sites s_1 and s_2. These two equations are always positive for $0 < \zeta_0 < 1$, $\sigma_\theta^2 > 0$, $\sigma_\nu^2 > 0$ and $\sigma_\omega^2 > 0$. Hence this proofs the inequality. □

Proof. **AR models**

Similarly for the AR models we can write the variance differences as:

$$\mathrm{Var}(Z(s_0,3)|Z(s_1,2)) - \mathrm{Var}(Z(s_0,3)|Z(s_1,1),Z(s_1,2)) = \frac{\zeta^2\rho^4\sigma_\epsilon^4(\rho^2\sigma_0^2+\sigma_\eta^2)^2}{(\rho^4\sigma_0^2+\sigma_\eta^2+\rho^2\sigma_\eta^2+\sigma_\epsilon^2)(\sigma_\eta^4+(2+\rho^2)\sigma_\eta^2\sigma_\epsilon^2+\sigma_\epsilon^4+\rho^2\sigma_0^2(\sigma_\eta^2+\sigma_\epsilon^2+\rho^2\sigma_\epsilon^2))}.$$

$$\mathrm{Var}(Z(s_0,3)|Z(s_1,2)) - \mathrm{Var}(Z(s_0,3)|Z(s_1,2),Z(s_2,2)) = \frac{\rho^2(\rho^4\sigma_0^2+\sigma_\eta^2+\rho^2\sigma_\eta^2)^2(\zeta\zeta_0(\rho^4\sigma_0^2+\sigma_\eta^2+\rho^2\sigma_\eta^2)-\zeta_1(\rho^4\sigma_0^2+\sigma_\eta^2+\rho^2\sigma_\eta^2+\sigma_\epsilon^2))^2}{D}$$

where,

$$\begin{aligned} D &= (\rho^4\sigma_0^2+\sigma_\eta^2+\rho^2\sigma_\eta^2+\sigma_\epsilon^2)((1-\zeta_0^2)\rho^8\sigma_0^4+(1-\zeta_0^2)(1+\rho^2)^2\sigma_\eta^2+ \\ &\quad 2\rho^4\sigma_0^2((1-\zeta_0^2)(1+\rho^2)\sigma_\eta^2-\sigma_\epsilon^2)+2(1+\rho^2)\sigma_\eta^2\sigma_\epsilon^2+\sigma_\epsilon^4), \end{aligned}$$

and $\zeta_1 = \exp(-\phi d_{02})$. So, this term is also positive for values $0<\rho<1$, $\sigma_0^2>0$, $\sigma_\eta^2>0$, $\sigma_\epsilon^2>0$, $0<\zeta<1$, $0<\zeta_0<1$, and $0<\zeta_1<1$. Hence, this proofs the inequality. □

(ii) *For forecasts of the AR models, following the equation 5.11, we can write,*

$$\mathrm{Dif} = \mathrm{Var}(Z(s_0,2)|Z(s_1,1)) - \mathrm{Var}(Z(s_0,3)|Z(s_1,1),Z(s_1,2)) > 0$$

(a) Dif >0, as $\sigma_0^2\to\infty$.

(b) Dif >0, as $\zeta\to 1$, iff $\frac{\sigma_\epsilon^2}{\sigma_0^2} < \frac{\kappa+\rho^2}{\kappa-(1-\rho^2)}$.

(c) Dif <0, as $\zeta\to 0$, and $\sigma_0^2=(1-\rho^2)\sigma_\eta^2$.

Proof. Proofs of (a) to (c) are given below:

(a) For $\sigma_0^2\to\infty$, the straight forward calculation of difference of the conditional variances for forecasts leads it to $\lim_{\sigma_0^2\to\infty}\mathrm{Dif}=\infty$.

(b) Similar to the proof of the conditional variance for prediction in Appendix A.3.1, we obtain

$$\lim_{\zeta\to 1}\mathrm{Dif} = \frac{\rho^2\sigma_\epsilon^4(\rho^2\sigma_0^4-\sigma_\eta^2\sigma_\epsilon^2+\sigma_0^2(\sigma_\eta^2+\sigma_\epsilon^2-\rho^2\sigma_\epsilon^2))}{(\rho^2\sigma_0^2+\sigma_\eta^2+\sigma_\epsilon^2)(\sigma_\eta^4+(2+\rho^2)\sigma_\eta^2\sigma_\epsilon^2+\sigma_\epsilon^4+\rho^2\sigma_0^2(\sigma_\eta^2+\sigma_\epsilon^2+\rho^2\sigma_\epsilon^2))}$$

A simple calculation yields the above term is positive iff

$$\frac{\sigma_\epsilon^2}{\sigma_0^2} < \frac{\kappa + \rho^2}{\kappa - (1-\rho^2)}$$

where, $\kappa = \frac{\sigma_\eta^2}{\sigma_0^2}$.

(c) For $\sigma_0^2 = (1-\rho^2)\sigma_\eta^2$, and spatial correlation $\zeta \to 0$, we get:

$$\lim_{\zeta \to 0} \text{Dif} = -\rho^6(2-\rho^2)\sigma_\eta^2, \quad \text{when} \quad \sigma_0^2 = (1-\rho^2)\sigma_\eta^2$$

Which is always negative.

□

A.4 Monotone Functions of the Conditional Variances

A.4.1 For Predictions

The first partial derivative of the predictive conditional variance of $Z(s_0, 1)$ given observation $Z(s_1, 1)$, with respect to ζ can be written as:

$$\frac{\delta}{\delta\zeta}[\text{Var}(Z(s_0,1)|Z(s_1,1))] = -\frac{2\zeta(\rho^2\sigma_0^2 + \sigma_\eta^2)^2}{\rho^2\sigma_0^2 + \sigma_\eta^2 + \sigma_\epsilon^2}$$

The first partial derivative of the predictive conditional variance of $Z(s_0, 1)$ given observations $Z(s_1, 1)$ and $Z(s_1, 2)$, with respect to ζ can be written as:

$$\frac{\delta}{\delta\zeta}[\text{Var}(Z(s_0,1)|Z(s_1,1), Z(s_1,2))] = -\frac{2\zeta(\rho^2\sigma_0^2 + \sigma_\eta^2)^2(\sigma_\eta^2 + \sigma_\epsilon^2 + \rho^2\sigma_\epsilon^2)}{\sigma_\eta^4 + (2+\rho^2)\sigma_\eta^2\sigma_\epsilon^2 + \sigma_\epsilon^4 + \rho^2\sigma_0^2(\sigma_\eta^2 + \sigma_\epsilon^2 + \rho^2\sigma_\epsilon^2)}$$

First partial derivatives of both conditional variances with respect to ζ are negative. This implies that the variance functions are monotonically decreasing function of the spatial correlation ζ, or in terms we can say the variances are monotonically decreasing function of the distance between sites s_0 and s_1.

A.4.2 For Forecasts

For the forecast, the partial derivative of the conditional variance of $Z(s_0, 3)$ given data point $Z(s_1, 2)$ with respect to ζ is:

$$\frac{\delta}{\delta\zeta}[\mathrm{Var}(Z(s_0,3)|Z(s_1,2))] = -\frac{2\zeta\rho^2(\rho^3\sigma_0^2 + \sigma_\eta^2 + \rho^2\sigma_\eta^2)^2}{\rho^4\sigma_0^2 + \sigma_\eta^2 + \rho^2\sigma_\eta^2 + \sigma_\epsilon^2}$$

The for the conditional variance of forecast $\mathrm{Var}(Z(s_0, 3)|Z(s_1, 1), Z(s_1, 2))$, we get the partial derivative with respect to ζ as:

$$\frac{\delta}{\delta\zeta}[\mathrm{Var}(Z(s_0,3)|Z(s_1,1),Z(s_1,2))] = \frac{-U}{\sigma_\eta^4 + (2+\rho^2)\sigma_\eta^2\sigma_\epsilon^2 + \sigma_\epsilon^4 + \rho^2\sigma_0^2(\sigma_\eta^2 + \sigma_\epsilon^2 + \rho^2\sigma_\epsilon^2)},$$

where,

$$\begin{aligned} U &= 2\zeta\rho^2(\rho^6\sigma_0^4(\sigma_\eta^2 + \sigma_\epsilon^2 + \rho^2\sigma_\epsilon^2) + \sigma_\eta^4(\sigma_\eta^2 + \rho^2\sigma_\eta^2 + \sigma_\epsilon^2 + 3\rho^2\sigma_\epsilon^2 + \rho^4\sigma_\epsilon^2) \\ &\quad +\rho^2\sigma_\eta^2\sigma_0^2(\sigma_\eta^2 + 2\rho^2\sigma_\eta^2 + 2\rho^2(2+\rho^2)\sigma_\epsilon^2)) \end{aligned}$$

The both partial derivatives for the conditional variances of forecasts are negative for the values $\sigma_0^2 > 0$, $\sigma_\eta^2 > 0$, σ_ϵ^2 and $0 < \rho < 1$, so they are monotonic decreasing function of the spatial correlation ζ.