# An Approximate Spatio-Temporal Bayesian Model for Alberta Wheat Yield

by

Evan Popoff

B.Sc. Hons., The University of British Columbia, 2012

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE COLLEGE OF GRADUATE STUDIES

(Interdisciplinary Studies)

THE UNIVERSITY OF BRITISH COLUMBIA

(Okanagan)

June 2014

# Abstract

Crop forecasting models are very valuable to a number of agricultural and government agencies. We investigated the effect of spatial and temporal environmental climate covariates on the growth of crop yield (wheat) at the regional scale across the province of Alberta. Model fitting was accomplished using data collected during the growing season from climate stations across Alberta provided by Agriculture and Agri-Food Canada (AAFC). A best fitting model was selected which takes into account simplicity (number of covariates used) and accuracy (predictive capability based on two selection criteria).

There have been a number of Bayesian methods for predicting wheat yield. However, many of these methods typically involve extensive algorithms such as a Metropolis-Hastings Markov Chain Monte Carlo (MCMC) that adds substantial computational complexity and run-time. We investigated the application of a spatio-temporal Bayesian model entitled the Integrated Nested Laplace Approximation (INLA). This method offers a computationally cheaper alternative to the MCMC approach and is capable of handling large data requiring interpolation (data sparsity) with relative ease. By structuring the model to have a sparse precision matrix, INLA is able to simplify posterior marginal estimation of the parameters by incorporating the Laplace approximation.

The INLA model demonstrated strong predictive capabilities when predicting for one year in advance or hind-casting for a *single* previous year. However, when multiple years of data were removed or predictions were made for multiple years in advance, INLA struggled to make predictions which deviated considerably from the mean of the remaining data. Predictive performance in the best fitting model saw a 40% increase in root mean

squared error (RMSE) when moving from one year to two and another 6% increase when moving from two to three years. We conclude that the INLA model structure offers valuable information when examining one year in advance but caution should be taken when attempting to forecast for multiple years in advance.

# Table of Contents

# List of Tables

# List of Figures

# Acknowledgements

First of all, I would like to thank my supervisor, Dr. Jason Loeppky, who supported and guided me throughout the entirety of my Masters Degree. You showed me the way to statistics and opened the door to a great opportunity to work at Agriculture and Agri-Food Canada. Thank you for your patience and for always having my best interests in mind. The lessons and knowledge I have learnt from you will stick with me for a lifetime.

Thank you to Dr. Nathaniel Newlands for allowing me the amazing work opportunity at Agriculture and Agri-Food Canada and for all of your helpful and encouraging insight throughout my Masters. This opportunity and your insight made me excited about statistics and this thesis would have not been possible without either of them.

Thank you to Crystal Parras for being an outstanding research and work partner throughout the summer of 2013. Your enthusiasm and dedication is what pushed the project through and I couldn't have wished for a more organized, understanding, and supportive partner.

# Chapter 1

# Introduction

## 1.1 Motivation

A reliable model for forecasting crop yield at the regional-scale is crucial for guiding decision making by farmers, governments, and agricultural industries [1]. For farmers, it provides a reference in order to estimate their expected profits for the year. It also helps them judge how to care for their crops. For example, a model can provide information as to what the optimal amount, and combination, of chemicals in the fertilizer will help their crops grow best, how much water to provide their crops throughout the season, and information on the best time to reseed and replace old crops. Forecasts also provide farmers with important marketing and production information that help with decisions influencing their financial well-being further on in the year [1]. Governments use agricultural forecasts as a way to adequately adjust policies and pricing and also to predict profits, which influence things such as Gross Domestic Product (GDP) for the country. In some countries, governments use forecasts to intervene in the interest of protecting their domestic agriculture [1]. Agricultural journalists make up another portion of the population who has a need for the forecasts. Agricultural journalists provide another way for farmers and the general public to keep up-to-date on crop forecasts. Food producers and others in the food industry require agricultural forecasts in order to adequately make purchasing decisions and to aid in storage decisions [1]. For example, if a producer requires a crop and purchases a large quantity of that crop in advance, it can be expensive for that producer to store all the extra product. However, if a forecast determines that a good harvest is approaching, the producer can then wait and buy at a later time to avoid the extra storage costs [1].

There are different models available for predicting wheat yield. They range from basic linear and quadratic regression models, to more sophisticated methods [4, 9]. Newlands and Zamar [9] partition the province of Alberta into Agricultural Statistical Census Regions (CARs) and take advantage of the data from neighbouring CARs in order to obtain a predicted mean wheat yield for each CAR. They use techniques such as a robust least angle regression (LARS), a Markov Chain Monte Carlo (MCMC) while incorporating a Metropolis-Hastings step, and a Random Forest Algorithm. A more detailed description of their methods can be found in [9]. Newlands et al. then go on to improve upon this method in a more recent publication [10]. Carew et al. employ the use of a three-stage generalized least squares procedure. This procedure is a form of weighted regression, which was used to account for the heteroscedasticity present between certain variables [4].

There has also been some work in comparing various wheat forecasting methods. These include the methods of Michel and Makowski which compare time series models at regional and national scales [8], as well as the methods of Prost et al. which involve comparing stepwise model selection with Bayesian model averaging [12].

In what follows, we consider a spatio-temporal model within a Bayesian framework. Bayesian methods are beneficial as they allow prior information to be included into the analysis which accounts for results learned from previous studies and/or previous knowledge. Bayesian methods account for all sources of uncertainty. Furthermore, a Bayesian approach allows for easily obtainable information about populations parameters. For example, under a Bayesian approach it is easy to obtain the probability that the population mean is greater than a certain value. Having such capabilities allows for much more intuitive results than what a frequentist analysis would present. Parameter estimation is accomplished using the Integrated Nested Laplace Approximation (INLA) which is computationally feasible for large data. The INLA method has been used before to construct models for spatial and spatio-temporal data, as seen in [3]. To our knowledge, this is the first time the INLA method has been employed for the modelling of crop yield. INLA incorporates the spatial correlation between regions

as a neighbourhood structure while operating under the Markovian property. This property says that two regions share a non-zero correlation if and only if the two regions are neighbours of each other. This implies that non-neighbouring regions are assumed to be completely independent of one another. Spatial structures of this form are well documented and referred to as Gaussian Markov Random Fields (GMRF)[13]. Unlike previous MCMC methods which randomly sample to eventually converge on the marginal posterior distribution of the parameters in the model; INLA exploits its model assumptions to obtain a numerical approximation to the marginal posteriors in question by using a Laplace approximation [3]. A full layout of how the INLA model operates will be introduced in chapter 3. The use of the `R-INLA` package in `R` allows for ease of implementation of the INLA method. The `R-INLA` package can be obtained for free download from `www.r-inla.org`.

The INLA method allows for a more efficient, computationally cheaper alternative to the Markov chain Monte Carlo (MCMC) technique mentioned in the previous paragraph that is often implemented when dealing with Bayesian modelling [3]. MCMC requires intensive computational power along with lengthy simulations in order to converge upon the posterior distribution of the parameters [3]. In some cases the MCMC method may not converge to the target posterior. INLA uses a combination of an analytical approximation and numerical integration to obtain the target distribution while removing the aforementioned convergence problems that can arise as well as the lengthy computation time needed in most cases [16].

The aim of this analysis involves obtaining a model using INLA which accurately describes the properties of wheat yield throughout the province of Alberta. Only regions of Alberta for which data in the covariates are readily available were considered. Environmental covariates that are deemed best at predicting the wheat behaviour were evaluated. When selecting a model, both accuracy and efficiency were considered. A final working INLA model was chosen based on the model selection criterion outlined in chapter 5. The final model will then be used to obtain a forecast of predicted wheat yield for up to three years into the future as well as a year ahead of the range of available data. Also, the probability the wheat yield in each ecodistrict

involved in the model exceeds the mean wheat yield over all ecodistricts will be calculated and included in a map plot. 25% and 75% confidence intervals were produced for wheat yield. These intervals hind cast for all the years of available data, taking into account the spatial and temporal aspects of the data.

The remainder of this thesis is organized as follows. Chapter 2 highlights how the data were obtained, structured and organized, as well as the different methods for dealing with missing data. Chapter 3 then goes on to describe the INLA method in detail, along with all mathematical algorithms behind the method. Chapter 4 discusses the structure of the different models used to fit both spatial and spatio-temporal data as well as describes the different configurations of covariates used in each model structure. Chapter 5 lays out the different methods used for model selection and describes the process of narrowing down the models to one final model. Chapter 6 displays various graphs and results obtained from the final selected model in the previous chapter. Chapter 7 concludes and summarizes upon conclusions drawn from the analysis.

# Chapter 2

# Data Assimilation, Manipulation, and Interpolation

## 2.1 Data Assimilation and Manipulation

All data used throughout the analysis were provided by Agriculture and Agri-Food Canada (hereafter, AAFC). The province of Alberta is divided into 150 ecodistricts as defined by Agriculture and Agri-Food Canada. Data from the years 1999-2004 for mean wheat yield were provided over the growing season for the 150 ecodistricts (with missing data present). The growing season is defined to take place between the months of May to September. Ecodistricts where data for wheat yield were missing throughout the entire time period were removed. This left a total of 92 ecodistricts remaining that contained wheat yield data in at least one year. Wheat yield is measured in bushels/acre and contains the following varieties of wheat: Canadian Western Extra Strong (CWES), Prairie Red, Soft White, Durum, Hard Red, and Winter wheat.

Climate stations are distributed throughout Alberta. Daily year round data for maximum temperature (°C), minimum temperature (°C), and precipitation (mm) are recorded at these stations. A map displaying the ecodistricts and climate stations is shown below in Figure 2.1. The grey ecodistricts represent ecodistricts where we have data on yield for the period of 1999-2004 and the white ecodistricts represent locations where yield data in unavailable. Due to the yearly format of the wheat yield data having only

one value per ecodistrict, the data for the other variables was arranged to also have one yearly value per ecodistrict. Over the days of the growing season the total precipitation is summed and an average of both the maximum and minimum temperature is taken for each climate station. Manipulating the data into a different format may have resulted in different model fits. For example, keeping the temperature and precipitation data as a daily measure over the growing season or averaging these data on a monthly scale over the growing season. In addition, examining precipitation data prior to the growing season may have different results than summing total precipitation over the growing season. Following, the average of the data from all climate stations within each ecodistrict is taken to obtain one value for minimum temperature, maximum temperature and total precipitation for each ecodistrict in each year. Data for water deficit (mm/growing season) and growing degree days (GDD) (°C-day) were also provided for each ecodistrict in each year in the analysis. Water deficit is defined as precipitation minus potential evapotranspiration, where potential evapotranspiration is the representation of the environmental demand for evapotranspiration provided there is adequate water status in the soil profile. Evapotranspiration is the sum of evaporation and plant transpiration from the Earth's land to the atmosphere. Growing degree days are defined as the average of the daily maximum and minimum temperature compared to a base temperature (usually 10 °C). Defined in an equation:

$$GDD = \frac{T_{max} + T_{min}}{2} - T_{base}, \tag{2.1}$$

where $T_{base}$ is the base temperature.

## 2.2 Dealing With Missing Data

The INLA method allows for missing values in the response variable (wheat yield). Therefore, there was no need to account for missing values for mean wheat yield before fitting the models. However, INLA is not able to fit a model with missing values in the covariates and because of this,

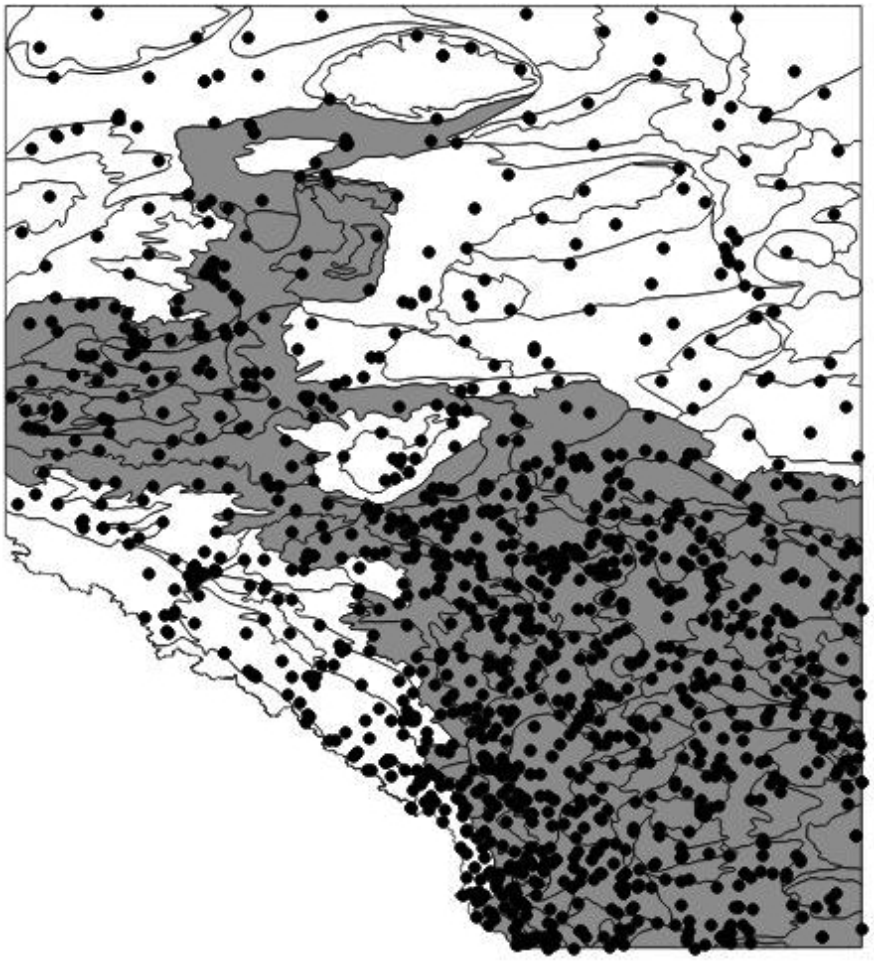

Figure 2.1: Map of 150 ecodistricts in Alberta and 1272 climate stations. Wheat yield data is unavailable for all white ecodistricts and is available for all 92 ecodistricts shaded in grey over the period of 1999-2004.

missing values for temperature, water deficit, GDD, and precipitation had to be imputed. Three different methods were considered when addressing the issue of missing covariate data. The benefits of each of these approaches are discussed in chapter 5.

1. *Nearest neighbours*

   For each ecodistrict in Alberta with missing data, the average of the data from the four nearest neighbouring ecodistricts were used to impute the missing value.

   The nearest neighbours were determined by calculating the distances between the centre point of the current ecodistrict and the centre points of the other ecodistricts. This resulted in selecting ecodistricts with the four shortest distances from the current ecodistrict's centre, otherwise known as the nearest neighbours.

   Problems with this method stem from the fact that the missing data in Alberta is not randomly scattered throughout the province. The missing data are most often observed over a large block of ecodistricts. Hence, there were many instances (approximately 1/3 of the time) where the four nearest neighbours of an ecodistrict with missing data also contained no data and therefore, the nearest neighbours method was not used.

2. *Averaging data*

   For each year, the average of the covariate data for each ecodistrict was taken (minimum temperature, maximum temperature, water deficit, GDD, and total precipitation). This value was used to fill in all the missing data for the corresponding covariate.

3. *Prediction using the INLA method*

   For each covariate in the model, maximum temperature, minimum temperature, water deficit, GDD, and total precipitation, the INLA method was used to predict missing values. Each covariate was modelled separately. For the spatial only model the covariate was the

response variable and the spatial structure ID was the sole predictor variable and was modelled using the Besag, York, and Molie (BYM) model [2]. Details on this model will be included in chapter 4 to follow. For the spatio-temporal model the covariate was again used as the response variable. This time the spatial structure ID and the year variables were both used as predictor variables for the response. The spatial ID was again modelled using a BYM model and the year variable was modelled using a first order random walk [11], which will also be described in more detail in the section to follow.

For both the spatial and spatio-temporal models, the fitted values from the resulting models were extracted and used to fill in the missing data for the respective ecodistricts in their given years.

After examining the data via histogram for total precipitation, it was deemed necessary to attempt to correct for normality as this precipitation data is to be used as a response variable in the imputation process. The two transformations considered prior to fitting were a square root transformation and a log(data+1) transformation. Both of these transformations allow for a value of zero to be mapped back to zero without any difficulties. After viewing histograms of both transformations, the square root transformation was chosen as it better normalized the data as well as preserving the spread of the data. These properties proved favourable when random effects were introduced via the first order random walk. The square root and log transformations can be seen in figure 2.2 to follow. When the fitted values were extracted, each of the values was squared to reverse the transformation.

Because the nearest neighbour method was deemed infeasible, a detailed description comparing the two remaining data imputation methods will be outlined in Chapter 5. To summarize, in terms of the Deviance Information Criterion (DIC) it can be seen that the model fits using INLA imputed values performed better than the model fits using the averaged data in all cases.

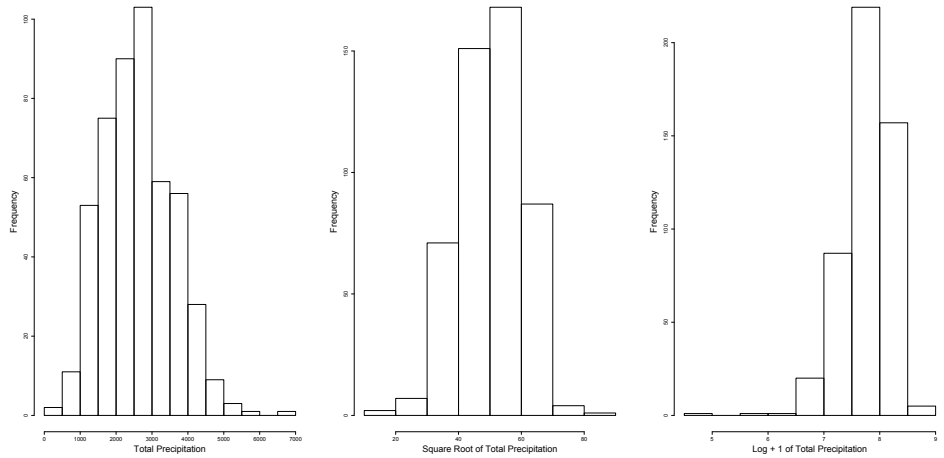


Figure 2.2: Total precipitation normalization. The square root transformation better normalized the precipitation data opposed to the Log + 1 transformation.

This also held true for essentially every case when examining the various cross-validated root mean squared errors (CV RMSEs). For these reasons, imputing values using INLA was selected as favourable over the data averaging method.

# Chapter 3

# Model Methodology

## 3.1 Bayesian Spatio-Temporal Model

Here we examine methods allowing for the incorporation of both spatial and spatio-temporal data. Spatial data in two-dimensional space are indexed by Y(s)$\equiv \{y(s), s \in \mathcal{D} \in \mathbb{R}^2\}$, and the spatio-temporal data we observe are indexed by Y(s,t)$\equiv \{y(s,t), (s,t) \in \mathcal{D} \subset \mathbb{R}^2 \times \mathbb{R}\}$. Here, $\mathcal{D}$ can represent either a set of spatial units in the case of point level data, or a continuous surface in the case of area level data [3]. In what follows, we will consider only area level spatio-temporal data. In this context, the set of spatial locations $(s_{11}, ..., s_{nT})$ were used to examine the spatio-temporal pattern of mean wheat yield over Alberta in $n = 92$ spatial ecodistricts and $T = 6$ time points representative of the years 1999 through 2004. The observed mean wheat yield data was represented as $\mathbf{y(s)} = \{y(s_{11}), ..., y(s_{nT})\}$.

We define a spatio-temporal model under a Bayesian framework. Let $\boldsymbol{\theta}$ be the set of parameters representing the expected mean wheat yield response at each spatio-temporal location. These parameters are indexed in such a way that the spatial correlation between ecodistricts is accounted for. The subscript notation $i, j$ will be used to indicate a spatio-temporal point opposed to using the $s_{ij}$ index.

As the methods involved in this analysis only consider area level data (as opposed to data situated at fixed points), it is possible to apply a neighbourhood structure to the data. This is accomplished by operating under the Markovian property, which states that any two elements of the parameter vector $\boldsymbol{\theta}$ share a non-zero correlation if and only if the two ecodistricts indexed in the parameter vector are neighbours of each other. Defining the precision matrix for the elements of $\boldsymbol{\theta}$ to be $\mathbf{Q} = \mathbf{\Sigma^{-1}}$, we have $\mathbf{Q_{ij}}$ repre-

senting the entry for any pair of elements $(i, j)$. Then, under the Markovian property, $\mathbf{Q_{ij}} \neq \mathbf{0}$ only if $j \in \{i, \mathcal{N}(i)\}$, where $\mathcal{N}(i)$ represents the set of neighbours of $\theta_i$ [3]. Spatial structures of this form are referred to as Gaussian Markov Random Fields (GMRF) [13].

The structure of the Bayesian model is then as follows. The mean of the $i$-th unit, $\mu_i$, is linked to an additive predictor $\eta_i$ by means of an appropriate link function $g(\cdot)$, such that $g(\mu_i) = \eta_i$ (for example, a logistic link function for binomial data). $\eta_i$ is then modelled additively using the following structure:

$$\eta_i = \alpha + \sum_{m=1}^{M} \beta_m x_{mi} + \sum_{l=1}^{L} f_l(z_{li}). \tag{3.1}$$

In our context, the link function used will simply be the identity function, so that the mean for observation $i$ ($\mu_i = \eta_i$). Here, $\alpha$ represents the intercept, $\boldsymbol{\beta} = (\beta_1, ..., \beta_M)$ represents the vector of linear coefficients corresponding to the covariates $\mathbf{x} = (x_1, ..., x_M)$, and $\mathbf{f} = (f_1(\cdot), ..., f_M(\cdot))$ represents a set of functions defined on another set of covariates $\mathbf{z} = (z_1, ..., z_M)$ which may or may not be the same as those in $\mathbf{x}$.

Looking back to equation (3.1), we can now define $\boldsymbol{\theta}$ more concretely. The parameters of interest are given by $\boldsymbol{\theta} = \{\alpha, \boldsymbol{\beta}, \mathbf{f}\}$. To return to the previous discussion in this chapter, a GMRF prior is placed on the parameter vector $\boldsymbol{\theta}$. This prior has mean $\mathbf{0}$ and precision matrix $\mathbf{Q}$. Furthermore, a vector of K hyper-parameters $\boldsymbol{\psi} = (\psi_1, ..., \psi_K)$ contained in the precision matrix $\mathbf{Q}(\boldsymbol{\psi})$ is specified. This corresponds to $\boldsymbol{\theta}$ having a multivariate normal distribution with mean vector $\boldsymbol{\mu}^*$, whose $i$th element is equal to $\alpha + \sum_{m=1}^{M} \beta_m x_{mi}$ and covariance matrix $\mathbf{Q}^{-1}(\boldsymbol{\psi})$.

The objective of the analysis would ideally be to obtain the joint posterior distribution of the parameters given the data, $\pi(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y})$. However, this not only does not have a closed form known distribution, it is also computationally infeasible to obtain a direct estimate of this distribution. Therefore, we must settle for a computationally feasible alternative method that involves estimating the marginal posterior distributions for each of the

elements in both the parameter vector $\boldsymbol{\theta}$ and the hyper-parameter vector $\boldsymbol{\psi}$.

These marginals are represented respectively below [3]:

$$\pi(\theta_i|\mathbf{y}) = \int \pi(\boldsymbol{\psi}|\mathbf{y})\pi(\theta_i|\boldsymbol{\psi},\mathbf{y})\, d\boldsymbol{\psi} \tag{3.2}$$

$$\pi(\psi_k|\mathbf{y}) = \int \pi(\boldsymbol{\psi}|\mathbf{y})\, d\boldsymbol{\psi}_{-k}. \tag{3.3}$$

To accomplish this, we will examine two different techniques. The traditionally used approach of Gibbs sampling, and the Integrated Nested Laplace Approximation (INLA). Gibbs sampling, while practical and proven to be effective, requires a lot of implementation and execution time especially when dealing with large data. INLA offers a computationally cheaper technique that is structured in a matter that allows large data to be handled with relative ease.

The reason for wishing to to stray away from traditional methodology and to apply the methodology of the INLA approach is to test the technique on the limited region of the province of Alberta in order to investigate whether these methods can offer any computational and accuracy improvements when compared to the traditional Gibbs sampling approach on this region. If this is the case, this would offer valuable insight for longer-term operational goals regarding agricultural risk management and decision support. The INLA methodology would hope to be extended to a larger project by Agriculture and Agri-Foods Canada that would expand the model region to the entirety of the agricultural landscape of Canada.

## 3.2 Gibbs Sampling

Gibbs Sampling has been the go to method when dealing with spatial model parameter estimation. WinBUGS is the traditional statistical software developed for use in Bayesian methods and for incorporating MCMC methods such as Gibbs Sampling. The Gibbs Sampling technique proceeds as follows.

To begin, we assume that all of the marginals represented by equations

3.2 and 3.3 are of known form (for example Gaussian, Gamma, etc.). In this senerio, Gibbs sampling proceeds to sample from the aforementioned marginals via the following steps:

1. Set arbitrary initial values for $\boldsymbol{\theta}^{(0)}, \boldsymbol{\psi}^{(0)}$

2. For t=1, ..., T proceed with the following:

   **Step 1**: Randomly sample $\theta_1^{(t)}$ from
   $\pi(\theta_1|\theta_2^{(t-1)}, ..., \theta_I^{(t-1)}, \psi_1^{(t-1)}, ..., \psi_K^{(t-1)}, \mathbf{y})$

   .
   .
   .

   **Step I**: Randomly sample $\theta_I^{(t)}$ from
   $\pi(\theta_I|\theta_1^{(t)}, ..., \theta_{I-1}^{(t)}, \psi_1^{(t-1)}, ..., \psi_K^{(t-1)}, \mathbf{y})$
   **Step I + 1**: Randomly sample $\psi_1^{(t)}$ from
   $\pi(\psi_1|\theta_1^{(t)}, ..., \theta_I^{(t)}, \psi_2^{(t-1)}, ..., \psi_K^{(t-1)}, \mathbf{y})$

   .
   .
   .

   **Step I + K**: Randomly sample $\psi_K^{(t)}$ from
   $\pi(\psi_K|\theta_1^{(t)}, ..., \theta_I^{(t)}, \psi_1^{(t)}, ..., \psi_{K-1}^{(t)}, \mathbf{y})$

Gibbs Sampling operates under the fact that as $t$ gets large, samples from each of the marginal distributions of the elements of the parameter vector will converge to a sample from the joint posterior distribution of the elements of the parameter vector.

However, it is often the case that we do not have a recognizable distribution for some or all of the marginals in the steps highlighted above. Therefore, at any step in which the marginal is of unknown form, we must employ the use of the Metropolis Hastings Algorithm to sample for those marginals.

The Metropolis Hastings Algorithm proceeds as follows. Note that this is an example for $\theta_i$, the same would hold true if needed for any of the $\psi_k$'s

1. Choose an arbitrary starting value for $\theta_i$ (called $\theta_i^{(0)}$)
2. Specify a proposal density function for $\theta_i$ (presented as $q(\theta_i^*|\theta_i^{(t)})$)
3. Draw $\theta_i^*$ from $q(\theta_i^*|\theta_i^{(t-1)})$
4. Compute the ratio $r=\frac{\pi(\theta_i^*)q(\theta_i^{(t-1)}|\theta_i^*)}{\pi(\theta_i^{(t-1)})q(\theta_i^*|\theta_i^{(t-1)})}$
   In the case that the proposal density is symmetric, we would have that:
   $r=\frac{\pi(\theta_i^*)}{\pi(\theta_i^{(t-1)})}$.
5. If $r\geq 1$, this means that the proposal draw has a higher probability of occurring than the current value of $\theta_i$ (that is $\theta_i^{(t-1)}$) and we will thus accept the proposal draw and set $\theta_i^{(t)}=\theta_i^*$. If $r\leq 1$, what happens essentially is that a random coin is flipped. With this, we will set
   $$\theta_i^{(t)}=\begin{cases}\theta_i^* & \text{with probability r}\\ \theta_i^{(t-1)} & \text{with probability 1-r}\end{cases}$$

Finally, once all the iterations of this chain are complete using direct sample or a Metropolis step, in order to obtain the final estimates, typically what is done is that the mode of each sample is determined and this value is taken as our final estimate for each parameter in each model.

## 3.3 The Integrated Nested Laplace Approximation (INLA)

The Integrated Nested Laplace Approximation [14] is a method developed for use on a group of structured additive regression models, entitled *latent Gaussian models*. These models span a large variety of structures including generalized linear models and spatial and spatio-temporal models. Due to the flexible nature of these models, they have been used in a wide range of applications and fields [3].

The first step in the INLA approach is to find a nested approximation for $\pi(\boldsymbol{\psi}|\mathbf{y})$. From this, it will then be possible to obtain the marginals $\pi(\psi_k|\mathbf{y})$. Due to the conditional independence of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ given $\mathbf{y}$ induced by the GMRF [13], it can be seen that:

$$\pi(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y}) = \pi(\boldsymbol{\theta}|\mathbf{y})\pi(\boldsymbol{\psi}|\mathbf{y}).$$

Again, due to the conditional independence relationship, $\pi(\boldsymbol{\theta}|\mathbf{y}) = \pi(\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{y})$. Therefore, we obtain:

$$\pi(\boldsymbol{\psi}|\mathbf{y}) = \frac{\pi(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y})}{\pi(\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{y})}. \tag{3.4}$$

The above marginal is then approximated by:

$$\tilde{\pi}(\boldsymbol{\psi}|\mathbf{y}) = \left.\frac{\pi(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y})}{\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{y})}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*(\boldsymbol{\psi})}, \tag{3.5}$$

where $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{y})$ is the Gaussian approximation of $\pi(\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{y})$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^*(\boldsymbol{\psi})$ is the mode of $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{y})$.

The Gaussian approximation is constructed based on densities of the form [14]:

$$\pi(\boldsymbol{\theta}) \propto \left\{ -\frac{1}{2}\boldsymbol{\theta}^T \mathbf{Q}\boldsymbol{\theta} + \sum_{i \in \mathcal{I}} \log\left\{\pi(y_i|\theta_i, \boldsymbol{\psi})\right\} \right\}. \tag{3.6}$$

For ease of readability, let $\log\left\{\pi(y_i|\theta_i, \boldsymbol{\psi})\right\} = g_i(\theta_i)$. The Gaussian approximation is then constructed by matching the mode of the distribution and the shape of the distribution to that of a Gaussian distribution. The mode is calculated by using either a Newton-Raphson scoring algorithm or the Fisher scoring algorithm [6]. These algorithms are conducted via the following steps:

1. Produce an initial estimate, $\boldsymbol{\mu}^{(0)}$, for the mode of $\pi(\boldsymbol{\theta})$.

2. Conduct a Taylor series expansion of $g_i(\theta_i)$ around $\mu_i^{(0)}$ truncating any

terms higher than second order. This is highlighted below:

$$g_i(\theta_i) \approx g_i(\mu_i^{(0)}) + b_i\theta_i - \frac{1}{2}c_i\theta_i^2 \tag{3.7}$$

where $b_i = f_1(\boldsymbol{\mu}^{(0)})$ and $c_i = f_2(\boldsymbol{\mu}^{(0)})$.

3. Obtain a Gaussian approximation of $\pi(\boldsymbol{\theta})$ with precision matrix $\mathbf{Q}$ + diag($\mathbf{c}$) and mode $\boldsymbol{\mu}^{(1)}$ found as the solution of $\{\mathbf{Q} + \text{diag}(\mathbf{c})\}\boldsymbol{\mu}^{(1)}$=$\mathbf{b}$.

4. Continue until the approximation converges to a Gaussian distribution. Define the mean of this converged distribution to be $\boldsymbol{\theta}^*$ and precision matrix $\mathbf{Q}^* = \mathbf{Q} + \text{diag}(\mathbf{c}^*)$ [14].

The next step is to locate the mode of $\tilde{\pi}(\boldsymbol{\psi}|\mathbf{y})$, by optimizing $\log\{\tilde{\pi}(\boldsymbol{\psi}|\mathbf{y})\}$ with respect to $\boldsymbol{\psi}$ using a quasi-Newton method [7]. This updates the second derivatives of $\log\{\tilde{\pi}(\boldsymbol{\psi}|\mathbf{y})\}$ using successive gradient vectors to converge to the target mode. A popular quasi-Newton method to achieve this is the Broyden-Flether-Goldbarb-Shanno (BFGS) method. This method starts with an initial estimate of the mode $x_0$ for a function $f(x)$, and an approximate Hessian matrix $B_0$ is obtained using finite differences to approximate all the second derivatives of $f(x)$. Define the gradient of $f$ to be $\nabla f$, which is also approximated using finite differences. These steps are completed until convergence is achieved:

1. Obtain a direction $\mathbf{p_k}$ by solving the equation $B_k\mathbf{p_k} = -\nabla f(\mathbf{x_k})$ for $\mathbf{p_k}$.

2. Minimize the function $f(\mathbf{x_k} + \alpha\mathbf{p_k})$ for $\alpha \in \mathbb{R}_+$ to obtain the minimum $\alpha_k$. Then update $\mathbf{x_{k+1}} = \mathbf{x_k} + \alpha_k\mathbf{p_k}$.

3. Set $\mathbf{s_k} = \alpha_k\mathbf{p_k}$.

4. Obtain the difference of successive gradient vectors, $\mathbf{y_k} = \nabla f(\mathbf{x_{k+1}}) - \nabla f(\mathbf{x_k})$

5. Update the approximate Hessian matrix, $B_{k+1} = B_k + \frac{\mathbf{y_k}\mathbf{y_k^T}}{\mathbf{y_k^T}\mathbf{s_k}} - \frac{B_k\mathbf{s_k}\mathbf{s_k^T}B_k}{\mathbf{s_k^T}B_k\mathbf{s_k}}$.

Once convergence is achieved for $\tilde{\pi}(\boldsymbol{\psi}|\mathbf{y})$, define the values of the function at the mode to be $\boldsymbol{\psi}^*$.

Following, the negative Hessian matrix is computed at $\boldsymbol{\psi}^*$. This is again accomplished using finite differences. Define this matrix as $\mathbf{H^{-1}}$.

## 3.4 Obtaining the marginals $\pi(\psi_k|\mathbf{y})$

As $\boldsymbol{\psi}$ is Gaussian distributed, we set $\boldsymbol{\Sigma} = \mathbf{H^{-1}}$ to be its covariance matrix. In order to more efficiently explore the distribution of $\log\{\tilde{\pi}(\boldsymbol{\psi}|\mathbf{y})\}$, standardized variables $\mathbf{z}$ will be used instead of $\boldsymbol{\psi}$. This is accomplished by setting $\boldsymbol{\Sigma} = \mathbf{V\Lambda V^T}$ to be the eigendecomposition of $\boldsymbol{\Sigma}$, and defining $\boldsymbol{\psi}$ as follows (with, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$):

$$\boldsymbol{\psi}(\mathbf{z}) = \boldsymbol{\psi}^* + \mathbf{V}\boldsymbol{\Lambda}^{1/2}\mathbf{z}. \tag{3.8}$$

$\log\{\tilde{\pi}(\boldsymbol{\psi}|\mathbf{y})\}$ is now explored using the $\mathbf{z}$-configuration as shown in figure 3.1 for a two dimensional vector $\boldsymbol{\psi}$ [14]. The intersection of the two z1 and z2 axis represents the mode ($\mathbf{z=0}$). In order to explore $\log\{\tilde{\pi}(\boldsymbol{\psi}|\mathbf{y})\}$ and find the maximum probability density, the following grid search is completed. Beginning at the mode, the density is explored in the positive z1 direction using a pre-specified step width $\delta_z$ as long as the condition

$$\log\left[\tilde{\pi}\left\{\boldsymbol{\psi}(\mathbf{0})|\mathbf{y}\right\}\right] - \log\left[\tilde{\pi}\left\{\boldsymbol{\psi}(\mathbf{z})|\mathbf{y}\right\}\right] < \delta_\pi \tag{3.9}$$

holds. The smaller the value of $\delta_z > 0$, the more accurate the exploration is. Where $\delta_\pi$=2.5 is a pre-defined value depending on the threshold accuracy.

Following, the negative z1 direction is explored the same way, and finally the z2 axis is explored in the same matter as the z1 axis. The result of this exploration produces the estimates represented by the black dots shown in figure 3.1. Afterwards, all combinations in the grid of black dots are evaluated and are included as grey dots provided that they satisfy the condition in equation (3.9). As all of the points are laid out on a rectangular grid, all of the area weights $\Delta_k$ are taken to be equal. These weights will be used during a numerical integration as seen in equation (3.10) to follow.

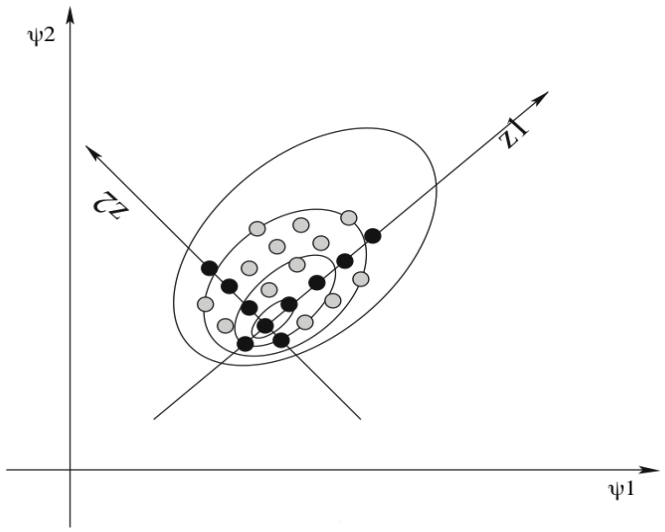


Figure 3.1: Exploration of the posterior marginal for $\boldsymbol{\psi}$. Assuming two dimensions, a grid search is performed along the standardized axes [14].

These evaluated points $\{\psi_k\}$ will now be used in a numerical integration to evaluate the posterior marginals $\pi(\theta_i|\mathbf{y})$. This is accomplished by using the points to construct an interpolant to $\log\{\tilde{\pi}(\boldsymbol{\psi}|\mathbf{y})\}$ and evaluating the marginals $\tilde{\pi}(\psi_k|\mathbf{y})$ via numerical integration from this interpolant.

## 3.5 Obtaining the marginals $\pi(\theta_i|\mathbf{y})$

The next task for the INLA approach is to obtain an estimate for the marginals $\pi(\theta_i|\mathbf{y})$. This was accomplished using the following numerical integration:

$$\tilde{\pi}(\theta_i|\mathbf{y}) \approx \sum_{k=1}^{K} \tilde{\pi}(\theta_i|\psi_k, \mathbf{y})\tilde{\pi}(\psi_k|\mathbf{y})\Delta_k. \tag{3.10}$$

Having just calculated the $\pi(\psi_k|\mathbf{y})$ and the weights $\Delta_k$, the only remaining task is to estimate the marginals $\pi(\theta_i|\boldsymbol{\psi}, \mathbf{y})$.
There are multiple methods for which to estimate these marginals: using

Gaussian approximations, Laplace approximations, or a simplified Laplace Approximation.

### 3.5.1 Using Gaussian Approximations

Using Gaussian approximations does not require much more calculation time from what has already been computed. However, because of this, the method lacks accuracy when compared to the others. When constructing the estimate for $\tilde{\pi}(\boldsymbol{\psi}|\mathbf{y})$, the Gaussian approximation for $\pi(\boldsymbol{\theta}|\boldsymbol{\psi},\mathbf{y})$ was already calculated. Therefore, all that would be left to calculate are the marginal variances for each $\theta_i$.

### 3.5.2 Using the Laplace Approximation

A more accurate and computationally intensive method then the Gaussian approximation is to compute the Laplace approximation. Note: the Laplace approximation will not be used directly for any computations with INLA but some preliminary explanations must be made in order to properly lay out the structure of the simplified Laplace approximation which will be used for calculations. We partitioned the parameter vector as $\boldsymbol{\theta} = (\theta_i, \boldsymbol{\theta}_{-i})$ and the following was calculated using the Laplace approximation:

$$\pi(\theta_i|\boldsymbol{\psi},\mathbf{y}) = \frac{\pi((\theta_i,\boldsymbol{\theta}_{-i})|\boldsymbol{\psi},\mathbf{y})}{\pi(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\psi},\mathbf{y})}. \tag{3.11}$$

This is approximated at the modal configuration as per equation (3.5). We denote the Laplace approximation at the modal configuration using the subscript $LA$:

$$\pi(\theta_i|\boldsymbol{\psi},\mathbf{y}) \approx \tilde{\pi}_{LA}(\theta_i|\boldsymbol{\psi},\mathbf{y}) = \left.\frac{\pi(\boldsymbol{\theta},\boldsymbol{\psi},\mathbf{y})}{\tilde{\tilde{\pi}}(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\psi},\mathbf{y})}\right|_{\boldsymbol{\theta}_{-i}=\boldsymbol{\theta}^*_{-i}(\theta_i,\boldsymbol{\psi})}, \tag{3.12}$$

where $\tilde{\tilde{\pi}}(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\psi},\mathbf{y})$ is the Gaussian approximation of $\pi(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\psi},\mathbf{y})$ (this is different than the Gaussian approximation $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{\psi},\mathbf{y})$ computed previously), and $\boldsymbol{\theta}^*_{-i}(\theta_i,\boldsymbol{\psi})$ is the modal configuration.

The computational complexity of equation (3.12) arises from the fact that $\tilde{\tilde{\pi}}(\boldsymbol{\theta}_{-i}|\theta_i, \boldsymbol{\psi}, \mathbf{y})$ must be re-computed for every value of $\theta_i$. We simplified this procedure by proposing either to remove the need for $\tilde{\tilde{\pi}}(\boldsymbol{\theta}_{-i}|\theta_i, \boldsymbol{\psi}, \mathbf{y})$ to be optimized for every value of $\theta_i$ altogether or to downgrade the optimization to just a subset of the $\theta_i$'s.

The first method removes the optimization step by estimating the modal configuration from the previously computed Gaussian approximation $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{y})$:

$$\boldsymbol{\theta}^*_{-i}(\theta_i, \boldsymbol{\psi}) \approx E_{\tilde{\pi}}[\boldsymbol{\theta}_{-i}|\theta_i], \tag{3.13}$$

where the expected value on the right hand side is computed from $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{y})$.

The second method involves choosing only the $\theta_j$'s that are close to $\theta_i$ as they have the most effect on the marginal of $\theta_i$. The correlation between regions $\theta_j$ and $\theta_i$ decreases as the distances between the nodes increases. We defined a space as the set $R_i(\boldsymbol{\psi})$ where we only include the regions $\theta_j$s that fall within a specified distance and use only these regions to compute the marginal of $\theta_i$. The points found with the region $R_i(\boldsymbol{\psi})$ are then standardized using the marginal mean and standard deviation from the Gaussian approximation found in equation (3.5) and are represented as $\theta_i^{(s)}$. These points are then used in the construction of the approximation of the marginal for $\theta_i$.

In the computation of the Laplace approximation in equation (3.12) there are more details which can be found in [14], but for the purposes of explaining the simplified Laplace approximation for use in INLA, we have all the details we need at this point.

### 3.5.3 Using the Simplified Laplace Approximation

The method used by INLA which best balances the need for efficiency and accuracy is the Simplified Laplace Approximation. This method proceeds by computing a Taylor's series expansion of the Laplace Approximation and only retaining the first few terms. This Taylor's series is then manipulated via a skewness function which better fits the approximation to the target distribution.

A general guideline of how the Simplified Laplace Approximation is constructed will be laid out to follow. However, for more details please refer to [14].

As specified, the natural logarithm of the numerator of equation (3.12) is approximated via a Taylor's series expansion truncated to any terms less than or equal to order three.

$$\log\{\pi(\boldsymbol{\theta},\boldsymbol{\psi},\mathbf{y})\}|_{\boldsymbol{\theta}_{-i}=E_{\tilde{\pi}}[\boldsymbol{\theta}_{-i}|\theta_i]} = -\frac{1}{2}\left(\theta_i^{(s)}\right)^2 + \frac{1}{6}\left(\theta_i^{(s)}\right)^3 \sum_{j\in\mathcal{I}\setminus i} \mathrm{d}_j^{(3)}\{\mu_i(\boldsymbol{\psi}),\boldsymbol{\psi}\}\{\sigma_j(\boldsymbol{\psi})a_{ij}(\boldsymbol{\psi})\}^3 + ... \tag{3.14}$$

As we have Gaussian distributed data, the denominator of (3.12) simply reduces to a constant [14].

$$\log\left\{\tilde{\tilde{\pi}}(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\psi},\mathbf{y})\right\}\big|_{\boldsymbol{\theta}_{-i}=E_{\tilde{\pi}}[\boldsymbol{\theta}_{-i}|\theta_i]} = \text{constant} \tag{3.15}$$

Setting

$$\gamma_i^{(3)}(\boldsymbol{\psi}) = \sum_{j\in\mathcal{I}\setminus i} \mathrm{d}_j^{(3)}\{\mu_i(\boldsymbol{\psi}),\boldsymbol{\psi}\}\{\sigma_j(\boldsymbol{\psi})a_{ij}(\boldsymbol{\psi})\}^3,$$

and combining equations (3.14) and (3.15), we obtain a simplified approximation of equation (3.12) as:

$$\log\left\{\tilde{\pi}_{SLA}(\theta_i^{(s)}|\boldsymbol{\psi},\mathbf{y})\right\} = \text{constant} - \frac{1}{2}\left(\theta_i^{(s)}\right)^2 + \frac{1}{6}\left(\theta_i^{(s)}\right)^3 \gamma_i^{(3)}(\boldsymbol{\psi}) + ... \tag{3.16}$$

Finally, to complete the approximation, a skew-normal distribution is introduced upon (3.16) in order to better emulate the desired distribution. This distribution is given by:

$$\pi_{SN}(z) = \frac{2}{\omega}\,\phi\left(\frac{z-\xi}{\omega}\right)\Phi\left(a\frac{z-\xi}{\omega}\right), \tag{3.17}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal cumulative distribution and

probability density functions respectively. Furthermore, $\xi$, $\omega > 0$, and $a$ are the location, scale, and skewness parameters.

# Chapter 4

# Model Application Methodology

## 4.1 Latent Model Descriptions Used in R-INLA

The structuring of INLA allows for each covariate to be fit using a different model or added linearly, with the final model being a combination of all models used for the covariates and the linear terms (as displayed is equation 3.1). The models used in the fitting of the spatial and the spatio-temporal INLA models were: the Besag, York, and Molie (BYM) model [2], an independent random noise (iid) model, and a first order random walk (rw1) [11]. A list of all the latent models available for R-INLA can be found at www.r-inla.org/models/latent-models.

The BYM model is simply a summation of two models, namely, the Besag model [2] and the iid model. Each of these two models are used to account for the fixed and random effects of spatial interaction, respectively.

The distribution for the Besag model is given by:

$$z_i | z_j, i \neq j, \tau \sim \mathcal{N}\left(\frac{1}{n_i}\sum_{i\sim j} z_j, \frac{1}{n_i \tau}\right). \tag{4.1}$$

Here, $z_i$, represents the index of ecodistrict $i$, and follows a normal distribution with mean $\frac{1}{n_i}\sum_{i\sim j} z_j$ and variance $\frac{1}{n_i\tau}$. The number of neighbours of ecodistrict $i$ is represented by $n_i$ and $i \sim j$ indicates that ecodistricts $i$ and $j$ are neighbours [2]. The precision parameter, $\tau$, is given a log gamma prior distribution, as follows:

$$\pi(\tau) = \frac{\ln(\tau)^{\alpha-1}}{\tau\beta^{\alpha}\Gamma(\alpha)} \exp(-\ln(\tau)/\beta), \tag{4.2}$$

where $\alpha$ and $\beta$ are shape parameters taking on values greater than zero. The density function for the iid model is given by:

$$\pi(\mathbf{z}|\tau) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \sqrt{s_i\tau} \exp\left(\frac{1}{2} s_i \tau z_i^2\right). \tag{4.3}$$

Here, $\mathbf{z}$ represents the vector of indices for all ecodistricts, $s_i$ is a scaling constant (the default which is used being equal to one), and $z_i$ and $\tau$ are defined in equation (4.1).

For a Gaussian vector, $\mathbf{x} = (x_1, ..., x_n)$, the first-order random walk is specified by [11]:

$$\Delta x_i = x_i - x_{i+1} \sim \mathcal{N}(0, \tau^{-1}). \tag{4.4}$$

.

## 4.2 Imputing Missing Covariate Values

Missing data for ecodistricts $i$ and years $j$ must be imputed. For ease of readability, we defined the covariates whose values need to be imputed as:

- $Z1$ = Growing Season Maximum Temperature
- $Z2$ = Growing Season Minimum Temperature
- $Z3$ = Growing Season Water Deficit
- $Z4$ = Growing Season Growing Degree Days
- $Z5$ = Growing Season Total Precipitation

Spatial imputing and spatio-temporal imputing were modelled in different forms, with a temporal component being added to the base spatial imputing

model when the aspect of time was introduced. For the spatial only model the covariates were imputed using the following models:

$$ZI_i = \alpha + \upsilon_i + \nu_i, \qquad I = 1, 2, \tag{4.5a}$$

$$\eta 5_i = \sqrt{Z5_i} = \alpha + \upsilon_i + \nu_i, \tag{4.5b}$$

where, $\eta 5_i$ is the link function as described at the beginning of chapter 3, $\alpha$ represents the intercept, $\upsilon$ represents the spatially structured component, and $\nu$ represents the random spatial component (each modelled as described in 2.2).

Continuing, the covariates for the spatio-temporal models were imputed using the following models:

$$ZI_{ij} = \alpha + \upsilon_i + \nu_i + \gamma_j + \phi_j, \qquad I = 1, 2, 3, 4, \tag{4.6a}$$

$$\eta 5_{ij} = \sqrt{Z5_{ij}} = \alpha + \upsilon_i + \nu_i + \gamma_j + \phi_j, \tag{4.6b}$$

where, $\gamma_j$ represents the temporally structured component, $\phi_j$ represents the random temporal component and all other components are as defined as above (each modelled as described in 2.2).

## 4.3 Spatial Model

**SPATIAL**

The structure of the spatial only model is as follows:

$$y_i = \alpha + \upsilon_i + \nu_i + f_1(Z1_i, \text{iid}) + f_2(Z2_i, \text{iid}) + f_2(Z3_i, \text{rw1}) \tag{4.7}$$

where $y_i$ represents the mean wheat yield for ecodistrict $i$, $\alpha$ represents the intercept, $\upsilon$ represents the spatially structured component, $\nu$ represents the random spatial component, and the $f(Z, M)$s represent the model $M$ for which the covariate $Z$ is fit [3]. The models are as described in section

4.1.

## 4.4 Spatio-Temporal Models

When examining a spatio-temporal structure, there were three base models considered. For each of these three bases, five different covariate configurations were tested on each base. Thus, bringing the total number of models considered to 15. The model selection process involved selecting models using a subset of the following covariates: maximum temperature, minimum temperature, total precipitation, water deficit, and growing degree days. The base structure of each model is described below.

**Model 2**

$$\begin{aligned} y_{ij} &= \alpha + \upsilon_i + \nu_i + (\beta + \delta_i) \cdot j \\ &\quad + \text{covariate functions}, \end{aligned} \tag{4.8}$$

where $y_{ij}$ represents the mean wheat yield for ecodistrict $i$ in the year $j$, $\alpha$ represents the intercept, $\upsilon$ represents the spatially structured component, $\nu$ represents the random spatial component, $\beta$ represents the linear global time effect, $\delta$ represents the interaction between time and space, and $j$ is a variable for year [3].

The spatial components were modelled using a Besag, York, and Molie (BYM) model. Which, as explained previously, separates the structured and random effects of space into the Besag and iid models, respectively. The model component $\delta_i$ was obtained by modelling the spatial component as a independent random noise (iid) using the years as weights. Finally, $\beta$ was obtained by simply introducing the temporal component as a linear covariate.

The third model includes a spatially structured and spatially random component as in model 2. However, instead of introducing the temporal component linearly, the temporal component is modelled in a similar matter

to the spatial component. That is, both a temporally structured and a temporally random component are introduced. This configuration assumes an independence between the spatial and temporal components. This model is specified as follows:

**Model 3**

$$\begin{aligned} y_{ij} &= \alpha + \upsilon_i + \nu_i + \gamma_j + \phi_j \\ &\quad + \text{covariate functions}, \end{aligned} \tag{4.9}$$

here $\gamma_j$ represents the temporally structured component, $\phi_j$ represents the random temporal component and all other variables are as defined in equation (4.8) [3].

The spatial components were modelled in the same matter as the model specified by equation (4.8). The temporally structured component $\gamma_j$ was modelled using a first order random walk (rw1) while incorporating a neighbourhood structure. The temporally random component $\phi_j$ was modelled simply as an iid.

The first model has the same structure as mentioned in the third model except it also incorporates a term to take into account the interaction between time and space. This model can be seen below:

**Model 1**

$$\begin{aligned} y_{ij} &= \alpha + \upsilon_i + \nu_i + \gamma_j + \phi_j + \delta_{ij} \\ &\quad + \text{covariate functions}, \end{aligned} \tag{4.10}$$

here $\delta_{ij}$ represents the interaction term between space and time. This interaction term assumes that the two random effects on space and time ($\nu_i$ and $\phi_j$) interact. This assumption also specifies neither a spatial nor temporal structure on the interaction term. i.e. $\delta_{ij}$ is distributed normally with a mean of 0 and variance $\tau_\delta$ ($\delta_{ij} \sim \mathcal{N}(0, \tau_\delta)$). All other variables are defined as in equation (4.9) [3].

The interaction term $\delta_{it}$ was modelled using an iid model. This was done by specifying an index number for each observation in the whole data set (552 total representing data for 6 years from 92 ecodistricts) to be inputted into the iid model. All other variables were modelled as described in the model defined in equation (4.9).

When being modelled, the various covariates were fit using the following models:

- **Maximum Temperature** - Independent Random Noise
- **Minimum Temperature** - Independent Random Noise
- **Total Precipitation** - First Order Random Walk
- **Water Deficit** - First Order Random Walk
- **Growing Degree Days** - Independent Random Noise

In table 4.1 on the following page, the covariates used in each configuration are displayed. Afterwards, three configurations were attempted but either achieved much worst results or failed to work at all. These configurations are referenced by NOT USED under the configuration column.

The covariates measuring aspects of temperature (i.e. maximum temperature, minimum temperature, and growing degree days) were modelled using an independent random noise model because it was hypothesized that temperature follows a fairly random trend throughout both time and space. Therefore, using this specification, it is assumed that the temperature in one ecodistrict is effected very minimally by the temperature of the ecodistricts surrounding it in the current year and the temperature of that same ecodistrict in past years.

Covariates measuring aspects of precipitation (i.e. total precipitation and water deficit) were modelled using a first order random walk as it was deemed fit that precipitation follows a fairly random trend in terms of starting and stopping. However, due to random weather patterns throughout space it was appropriate to capture these patterns within the random walk structure. The structure of the random walk also takes into account random

trends in precipitation throughout time. For example, if total precipitation increased for multiple consecutive years within a specific ecodistrict then the total precipitation for the ecodistrict in the following year would be given a higher probability to be modelled as increasing further and the same goes for a steadily declining total precipitation within a given ecodistrict throughout time.

Table 4.1: Covariate Configurations.

| Configuration | Covariates Used |
|---|---|
| I | Maximum Temperature, Minimum Temperature, Total Precipitation |
| II | Maximum Temperature, Minimum Temperature, Total Precipitation, Water Deficit |
| III | Maximum Temperature, Minimum Temperature, Total Precipitation, Growing Degree Days |
| IV | Maximum Temperature, Minimum Temperature, Total Precipitation, Water Deficit, Growing Degree Days |
| V | Maximum Temperature, Minimum Temperature, Water Deficit |
| | |
| NOT USED | Water Deficit, Growing Degree Days |
| NOT USED | Total Precipitation, Growing Degree Days |
| NOT USED | Maximum Temperature, Minimum Temperature, Water Deficit, Growing Degree Days |

# Chapter 5

# Model Selection

Model selection was performed via the use of two criterion: the deviance information criterion (DIC) [15] and a k-fold cross validated root mean squared error (CVRMSE).

The deviance information criterion is a commonly used criterion in Bayesian model selection. The general rule for determining which model contains a better fit is to choose the model with the smallest DIC [15]. For each of the models fit, the DIC was calculated using all available data.

The DIC is calculated as follows:

$$DIC = p_D + \bar{D}. \tag{5.1}$$

Here, $p_D$ is a measure of the number of effective parameters in the model and penalizes models with more complexity (i.e. a larger number of parameters) and is defined as:

$$p_D = \bar{D} - D(\bar{\theta}). \tag{5.2}$$

$D(\theta)$ is known as the deviance and is defined as:

$$D(\theta) = -2\log(p(y|\theta)) + C, \tag{5.3}$$

where $y$ are the data, $\theta$ are the unknown parameters in the model, $p(y|\theta)$ is the likelihood and $C$ is a constant with respect to $\theta$ which always cancels out when comparing models and therefore does not need to be defined.

Furthermore, $\bar{\theta}$ is the expectation of $\theta$ and $\bar{D}$ is the expectation of $D(\theta)$ with respect to $\theta$. Therefore, (5.2) can be viewed as the mean deviance minus the deviance of the means. $\bar{D}$ will always decrease as the number of

the parameters in a model increases, which is why $p_D$ is used to compensate for this and favours models with fewer parameters [15]. The method in which the DIC is calculated remains unchanged between the spatial and spatio-temporal models and is included as an output within the `R-INLA` package for each model.

Even though DIC is a measure of deviance, it is possible to obtain negative values for this criterion. This is true because the probability density $p(y|\theta)$ can be greater than one if it has a very small standard deviation. In this case, the log of the density would be positive thus making equation 5.3 negative [15]. Due to this, it is the difference in DIC values between the models that we look at and not the absolute value of DIC for each model. Thus, implying the lowest DIC value regardless of sign defines the best fitting model.

Cross validation was calculated differently for the spatial only and the spatio-temporal models.

For the spatial only model, cross validation was performed by randomly sampling 10% of the ecodistricts that contained wheat yield data. This sample was removed from the data frame and the model was fit on the remaining data. The predicted values for the sample of ecodistricts were compared to the observed data by calculating the mean squared error. This process was repeated 5000 times and the k-fold CVRMSE was calculated as the determining factor of model fit.

Mean squared error (MSE) is defined to be:

$$MSE_{spatial} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}, \tag{5.4}$$

where $y_i$ are the observed mean wheat yield data, $\hat{y}_i$ are the predicted mean wheat yield data and $n$ is the size of the 10% sample. The CVRMSE is then calculated as:

$$CVRMSE_{spatial} = \sqrt{\frac{\sum_{j=1}^{5000} MSE_j}{5000}}, \tag{5.5}$$

where $MSE_j$ refers to the MSE of the $j$th sample.

For the spatio-temporal model, instead of sampling and removing part of the data frame, each year was used to partition the data. When calculating MSE, either one or multiple years were removed and predicted for during each of the terms in the calculation. Each possible combination of years were removed during each calculation. Therefore, equations (5.4) and (5.5) update to:

$$MSE_{temporal} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n_2}(y_{ij} - \hat{y}_{ij})^2}{n_2}, \tag{5.6}$$

$$CVRMSE_{temporal}\sqrt{\frac{\sum_{k=1}^{K} MSE_k}{K}}. \tag{5.7}$$

Where $n_2 \leq 92$ are the number of observed mean wheat yield data points in the year $j$, $m$ are the number of years removed during each iteration, and K= $\binom{6}{m}$ (where 6 is the number of years with obtained data).

When performing the cross-validation, there were instances in which certain years were removed. INLA was unable to predict the wheat yield for those missing years. The spacing of the available data in the adjacent years to the year in question caused issues for the INLA method in certain situations. When there are large clumps of ecodistricts with missing data throughout the map, INLA has a difficult time predicting the wheat yield for these ecodistricts. Typically when this occurred, it was in an iteration where it was particularly difficult for the models to predict accurately. These iterations were removed from all the models in order to avoid biasing the

validation results.

When calculating the CVRMSE for both the spatial only and spatio-temporal models, inconceivable outliers for MSE were produced in both certain iterations of the spatial only model and for certain year combinations in the spatio-temporal models. Outliers of this sort were removed using the rule-of-thumb outlier rule when plotting the MSE values in a box plot for each of the respective models. That is, any outliers outside of the range [Q1-(1.5)IQR, Q3+(1.5)IQR] of the MSE data for each of the models were removed. Where Q1 is the first quartile, Q3 is the third quartile, and IQR(interquartile range)=Q3-Q1.

In table 5.1 to follow, Model refers to the base model structure as defined in equations (4.10), (4.8), and (4.9) respectively. The numbers I, II, III, IV, and V refer to the respective covariate configuration as defined in table 4.1. DIC refers to the respective model-configuration pairing's DIC value, and CVRMSE1, CVRMSE2, and CVRMSE3 refer to the cross-validation values calculated by removing all subsets of 1, 2, and 3 years respectively from each model-configuration pairing.

Table 5.1: Deviance information criterion (DIC) and cross validated root mean squared error (CVRMSE) values for the various models fitted. Smaller values of DIC and CVRMSE indicate a better model fit so Model 2 under configuration III had the best DIC value and Model 1 under configuration II had the best overall CVRMSE value.

| Model | I | II | III |
|---|---|---|---|
| Model 1-DIC | -2292.42 | -1698.69 | -1798.81 |
| Model 2-DIC | 855.53 | 64.94 | -2708.32 |
| Model 3-DIC | -2040.65 | -1688.83 | -2054.89 |
| Model 1-CVRMSE1 | 40.55 | 37.09 | 43.24 |
| Model 2-CVRMSE1 | 72.96 | 55.50 | 69.01 |
| Model 3-CVRMSE1 | 51.88 | 51.36 | 38.25 |
| Model 1-CVRMSE2 | 56.90 | 63.25 | 56.37 |
| Model 2-CVRMSE2 | 78.69 | 115.29 | 79.20 |
| Model 3-CVRMSE2 | 58.76 | 60.48 | 58.12 |
| Model 1-CVRMSE3 | 60.51 | 65.79 | 60.88 |
| Model 2-CVRMSE3 | 143.72 | 99.87 | 162.94 |
| Model 3-CVRMSE3 | 62.10 | 59.15 | 60.53 |

| Model | IV | V |
|---|---|---|
| Model 1-DIC | 3786.07 | -2506.65 |
| Model 2-DIC | -2243.12 | 4381.46 |
| Model 3-DIC | 3884.25 | 4360.94 |
| Model 1-CVRMSE1 | 44.39 | 61.26 |
| Model 2-CVRMSE1 | 142.51 | 61.65 |
| Model 3-CVRMSE1 | 50.08 | 60.80 |
| Model 1-CVRMSE2 | 55.35 | 55.91 |
| Model 2-CVRMSE2 | 107.16 | 75.77 |
| Model 3-CVRMSE2 | 65.95 | 58.13 |
| Model 1-CVRMSE3 | 62.10 | 55.21 |
| Model 2-CVRMSE3 | 112.03 | 188.72 |
| Model 3-CVRMSE3 | 63.94 | 56.52 |

To determine if there were redundant covariates (i.e. including all redundant covariates would achieve no further gains in accuracy), correlation values were calculated between covariates measuring similar properties. Total precipitation and water deficit essentially are measuring the same quantity. The correlation coefficient between total precipitation and water deficit was 0.8366. This suggests a high collinearity between the two covariates that cannot be ignored. Collinearity between explanatory variables offers several undesirable consequences under a regression context [5]. Furthermore, since maximum temperature, minimum temperature, and growing degree days are all measuring certain aspects of temperature, their correlation values amongst the covariates were also examined. The correlations between maximum temperature, minimum temperature, and growing degree days are summarized in table 5.2 below.

Table 5.2: Correlation matrix for maximum temperature, minimum temperature, and growing degree days. Small values indicate no potential problems with collinearity.

| | Max Temp | Min Temp | GDD |
|---|---|---|---|
| Max Temp | 1 | 0.6434 | 0.5748 |
| Min Temp | 0.6434 | 1 | 0.4054 |
| GDD | 0.5748 | 0.4054 | 1 |

After observing the values in table 5.2, there does not appear to be any troublesome combination of variables that need to be examined further. However, the high correlation value between the total precipitation and water deficit covariates must be addressed. Since total precipitation and water deficit are highly positively correlated and both these covariates contribute positively towards wheat growth, there is a high potential that their collinearity will cause a problem in the INLA regression model. It becomes difficult to distinguish the effect each covariate has individually on the response variable (mean wheat growth), meaning that the respective contributions of the covariates could overlap. Hence, leading to skewed results by under or overestimating mean wheat growth in certain ecodistricts.

It is unwise to include both total precipitation and water deficit within the same model. Therefore, configurations II and IV in table 5.1 will no longer be considered. Thus, the following configurations in table 5.3 below remain.

Table 5.3: Deviance information criterion (DIC) and cross validated root mean squared error (CVRMSE) values for a subset of models fitted. Model 2 under configuration III had the best overall DIC value and model 1 under configuration I had the best overall CVRMSE value.

| Model | I | III | V |
|---|---|---|---|
| Model 1-DIC | -2292.42 | -1798.81 | -2506.65 |
| Model 2-DIC | 855.53 | -2708.32 | 4381.46 |
| Model 3-DIC | -2040.65 | -2054.89 | 4360.94 |
| Model 1-CVRMSE1 | 40.55 | 43.24 | 61.26 |
| Model 2-CVRMSE1 | 72.96 | 69.01 | 61.65 |
| Model 3-CVRMSE1 | 51.88 | 38.25 | 60.80 |
| Model 1-CVRMSE2 | 56.90 | 56.37 | 55.91 |
| Model 2-CVRMSE2 | 78.69 | 79.20 | 75.77 |
| Model 3-CVRMSE2 | 58.76 | 58.12 | 58.13 |
| Model 1-CVRMSE3 | 60.51 | 60.88 | 55.21 |
| Model 2-CVRMSE3 | 143.72 | 162.94 | 188.72 |
| Model 3-CVRMSE3 | 62.10 | 60.53 | 56.52 |

Comparing the three remaining configurations in 5.3, it can be seen that configuration V performs the worst. The only clear stand out position where configuration V was better was in the DIC of model 1. However, due to the significantly better (lower) CVRMSE1 values in configurations I and III, this configuration is outdone. Also note that the CVRMSE1 value holds much more precedence over both the CVRMSE2 and CVRMSE3 values as it becomes much harder for each of the models to predict as more years are removed. Other than the DIC of model 1, there are no other significant places where configuration V is strictly better. Due to this, configuration V will no longer be considered.

Comparing the remaining two configurations I and III, model 1 is better under configuration I, model 2 is better under configuration III, and model 3 is strictly better under configuration III. To expand, better is defined as having the majority of lower DIC, CVRMSE1, CVRMSE2, and CVRMSE3 values under one configuration over the other and strictly better is defined as having all lower values in the respectively categories under a given configura-

tion. While model 2 does have the lowest DIC value under configuration III when compared to the best fits of model 1 under configuration I and model 3 under configuration III, model 2 has a significantly worse CVRMSE under all categories than the other best model fits. Therefore, the final two models to be considered are model 1 under configuration I and model 3 under configuration III.

Comparing these two fits it can been seen that model 1 under configuration I has a slightly better DIC, CVRMSE2, and CVRMSE3 values, while model 3 under configuration III has a slightly better CVRMSE1 value. To further decide which model is to be chosen, it is seen that model 1 under configuration I has three covariates included in the model (Maximum Temperature, Minimum Temperature, and Total Precipitation) where model 3 under configuration III has four covariates included in the model (Maximum Temperature, Minimum Temperature, Total Precipitation, and Growing Degree Days). Therefore, if the models were extremely similar in terms of fit, judging solely on complexity, model 1 under configuration I is favoured as it includes less covariates and is thus a less complex model. Furthermore, the DIC (as described in the beginning of this chapter) is a value created specifically for Bayesian models and is included as part of the `R-INLA` package for comparing different model fits and will hence be favoured in terms of model selection judgement over CVRMSE when the differences between two models are not drastically different in either category. Therefore, model 1 under configuration I is to be the chosen model to fit.

To further investigate the chosen model fit, a different method of imputing the missing data was attempted. As described in section 2.2, this method was averaging all the collected data from each appropriate covariate rather than using INLA to impute the values. These results are displayed in table 5.4 on the following page.

The column titled "Model" refers to the model's reference name. Here **SPATIAL** and **Model 1** are self explanatory by their reference equation numbers in the following column and **AVG** is the same model structure as **Model 1** except using the averaged data to impute missing data rather than using the INLA model to impute missing data. Eqn is referring to the

equation in which that model structure is defined in chapter 4. CVRMSE, CVRMSE2, and CVRMSE3 refer to the cross-validation values calculated by removing all subsets of 1, 2, and 3 years respectively. The SDE1, SDE2, and SDE3 values were calculated as a standardized measure in order to give the CVRMSE values for 1, 2, and 3 years more context. First, a dummy value was calculated by simply predicting the wheat yield for each ecodistrict using the mean of the wheat yields from the entire data set. Meaning that when a year was removed, every value was filled in using the mean. From this, a CV RMSE was calculated using the same method as described in equation (5.7) for both the **Model 1** and **AVG** models. Finally, the SDE values were calculated as the ratio of the CVRMSE from each model over the respective dummy value. Given this specification, an SDE value of less than one indicates that the given model is better at predicting than the dummy model which just predicts using the overall mean. Dummy values were calculated as 59.28, 59.26, and 59.25 when one, two, and three years were removed when calculating CVRMSE respectively.

Table 5.4: Model validation for spatial only, INLA imputed missing values, and average imputed missing values. The model using the INLA imputed missing values displayed the best results in terms of both DIC and CVRMSE.

| **Model:** | **SPATIAL** | **Model 1** | **AVG** |
|---|---|---|---|
| Eqn | (4.7) | (4.10) | (4.10) |
| DIC | -372.17 | -2292.42 | 122.81 |
| CVRMSE1 | 71.16* | 40.55 | 44.02 |
| SDE1 | NA | 0.68 | 0.74 |
| CVRMSE2 | NA | 56.90 | 56.56 |
| SDE2 | NA | 0.96 | 0.95 |
| CVRMSE3 | NA | 60.51 | 62.91 |
| SDE3 | NA | 1.02 | 1.06 |

**Model 1** had a significantly better DIC compared to both the spatial only model and the **AVG** model. The CVRMSE for the spatial only model was significantly worse than both the other models and is even worse than any of the temporal models dummy value CVRMSE values. The spatial only

model performed worse than the temporal models in terms of model prediction capabilities. When examining the CVRMSE values between **Model 1** and **AVG**, they compare fairly similarly in magnitude with **Model 1** having the slight edge. It should also be stated that when removing one year at a time, both models predict significantly better than the dummy value using only the mean. This is seen by SDE1 having a value lower than 1. However, when two and three years are removed, both models tend to predict very close to the mean value (indicated by SDE2 and SDE3 values close to 1). Hence, caution should be taken when trying to predict for data with two or more missing years.

With **Model 1** emerging as the best model in terms of both model selection criterion, it can be seen that imputing missing values using INLA is a better method than computing the missing values using the average of the entire data set. **Model 1** will be the final model selected from which to display results from in chapter 6 to follow.

# Chapter 6

# Results: Regional Distribution of Wheat Yield

The plots displayed in figures 6.1 and 6.2 respectively represent the percentiles for mean wheat yield when examining the spatial and temporal components of the model separately.

Figure 6.1 shows the spatial aspect of the model. The two subsequent plots represent the 25th and 75th spatial percentiles, respectively. These values were calculated by taking the 25th and 75th percentile of the six available fitted values (due to the six available years of data) for each ecodistrict. This plot reveals the highest density of wheat yield observed throughout the central belt of Alberta, with relatively high yields in the southeast and lower yields throughout the northwest portion of the province. The plot also displays the significant variation in wheat yield over neighbouring ecosdistricts in addition to considerable variation in relation to the expected yield.

Figure 6.2 shows the temporal aspect of the model. Highlighted is the temporal variability of wheat yield over the years of model fitted values. The blue line represents the mean predicted wheat yield over all ecodistricts in the respective year. The lower and upper portions of the grey band represent the 25th and 75th percentiles of the wheat yield over all ecodistricts in the given year, respectively. This plot reveals the greatest mean wheat yield occurring in the boundary years of 1999 and 2004, with decreasing mean wheat yield when proceeding towards the middle years finally resulting in the minimum trough year of 2002.

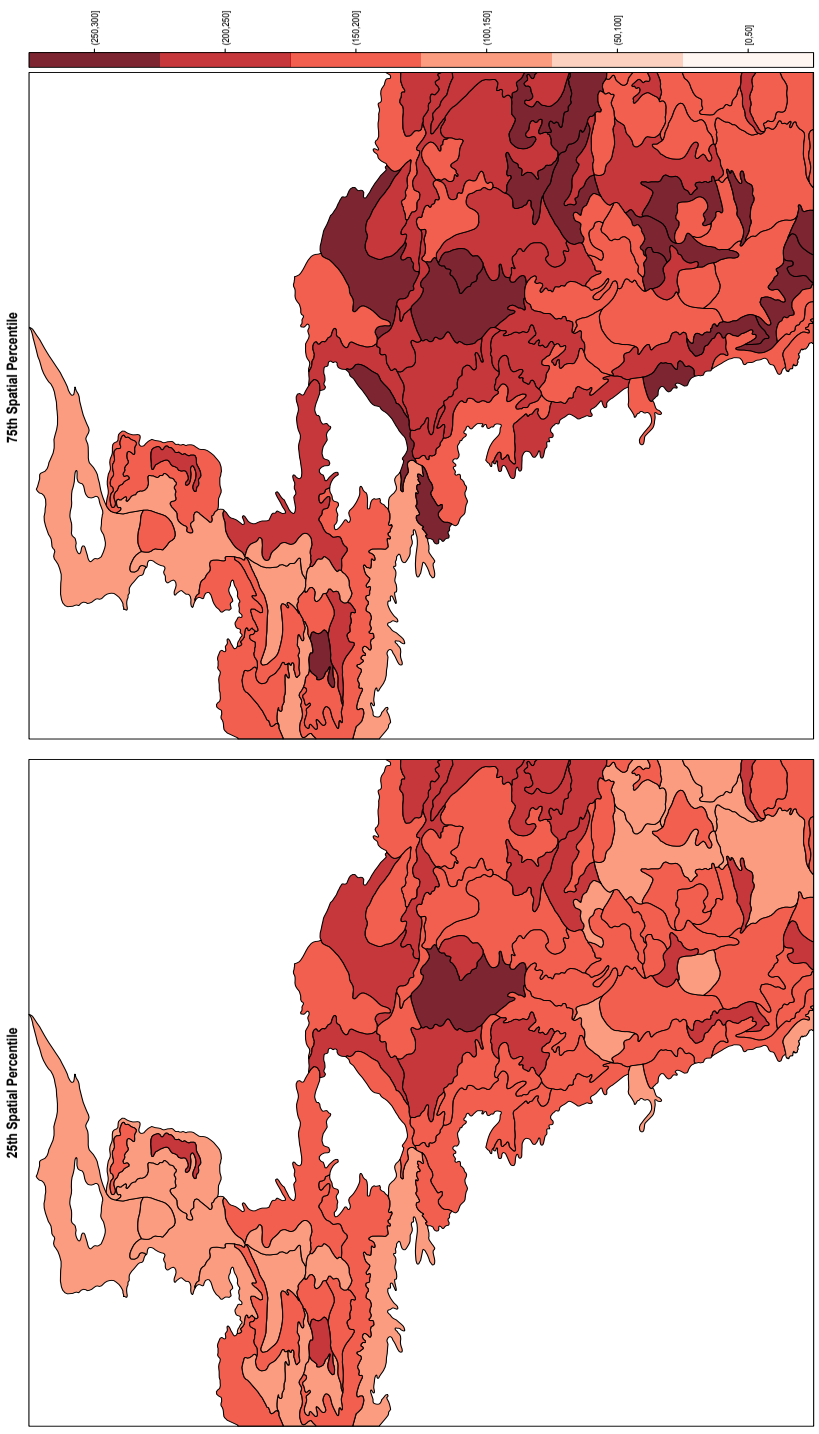


Figure 6.1: 25th and 75th Spatial Percentiles for Fitted Mean Wheat Yield over Alberta. Highest wheat yield density is observed throughout the central belt of Alberta. Significant variation of wheat yield is observed over neighbouring ecosdistricts.

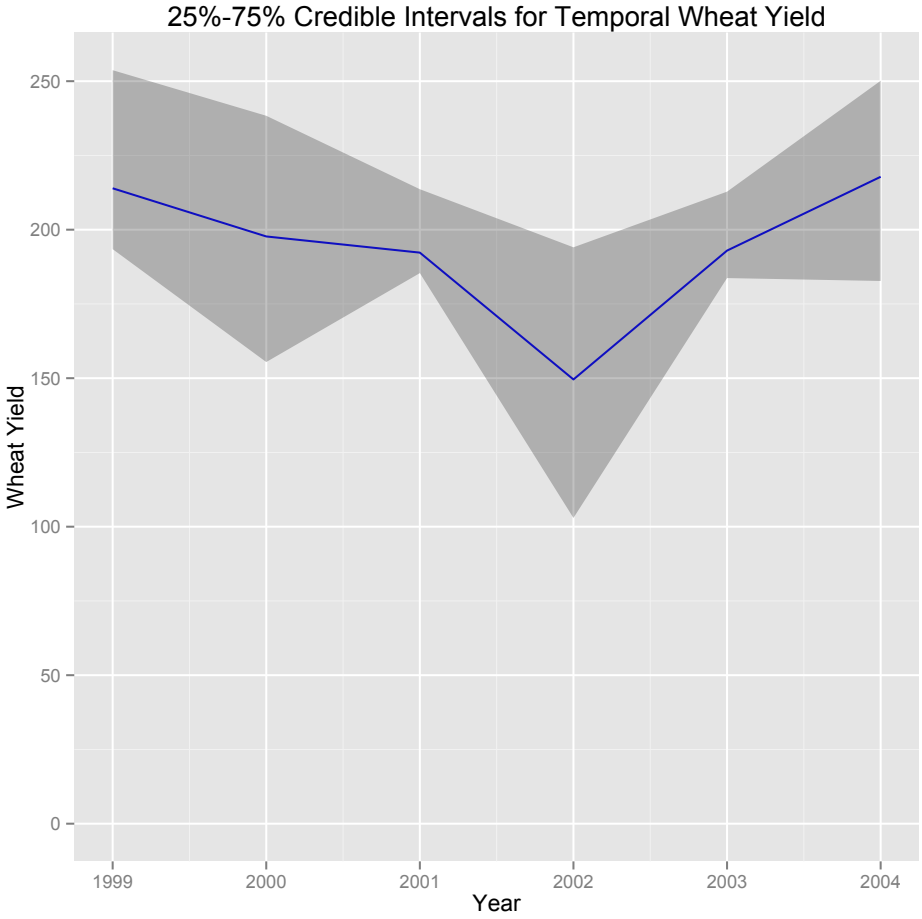


Figure 6.2: 25th and 75th temporal percentiles for fitted mean wheat yield over Alberta covering the years 1999-2004. Highlighted is the temporal variability. The blue line represents the mean predicted wheat yield with the lower and upper portions of the grey band representing the 25th and 75th percentiles, respectively. The greatest mean wheat yield is observed in 1999 and 2004, with a minimum trough year of 2002.

Figure 6.3 considers only the spatial aspect of the model. The numbers shown represent the probability that the wheat yield in the given ecodistrict will be above the mean wheat yield taken over all ecodistricts for which observed data were available. This plot is consistent with the discussion highlighted for the plot portrayed in 6.1 in that the highest probabilities are observed throughout the central belt of Alberta, with lower probabilities observed when moving to the northern and southern portions of the province. In addition, we again see large variation between neighbouring ecodistricts.

Figures 6.4, 6.5, and 6.6 all represent the mean wheat yield predicted for the year 2004. For each plot, the predicted values are displayed in the segment to the left, where as the observed values are displayed in the segment to the right. Figure 6.4 represents the prediction when the wheat yield data for 2004 had been removed (i.e. one year in advance), figure 6.5 for when the wheat yield data for 2003 and 2004 had been removed (i.e. two years in advance), and finally figure 6.6 for when the wheat yield data for 2002, 2003, and 2004 had been removed (i.e. three years in advance). Note that an "NA" value in the right column represents that wheat yield data for that ecodistrict were missing in the year 2004. Consistent with the numbers for SDE in table 5.4, we can see from the subsequent plots that as we predict for more years in advance the predictions tend to converge more and more to the mean of the remaining overall data.

Finally, 6.7 represents the mean wheat yield forecasted by the model using all six years of available data for one year ahead of known data (2005). This plot is meant as an example to demonstrate the prediction capabilities of the model. Due to no data being available in 2005, there is nothing to compare this result to.

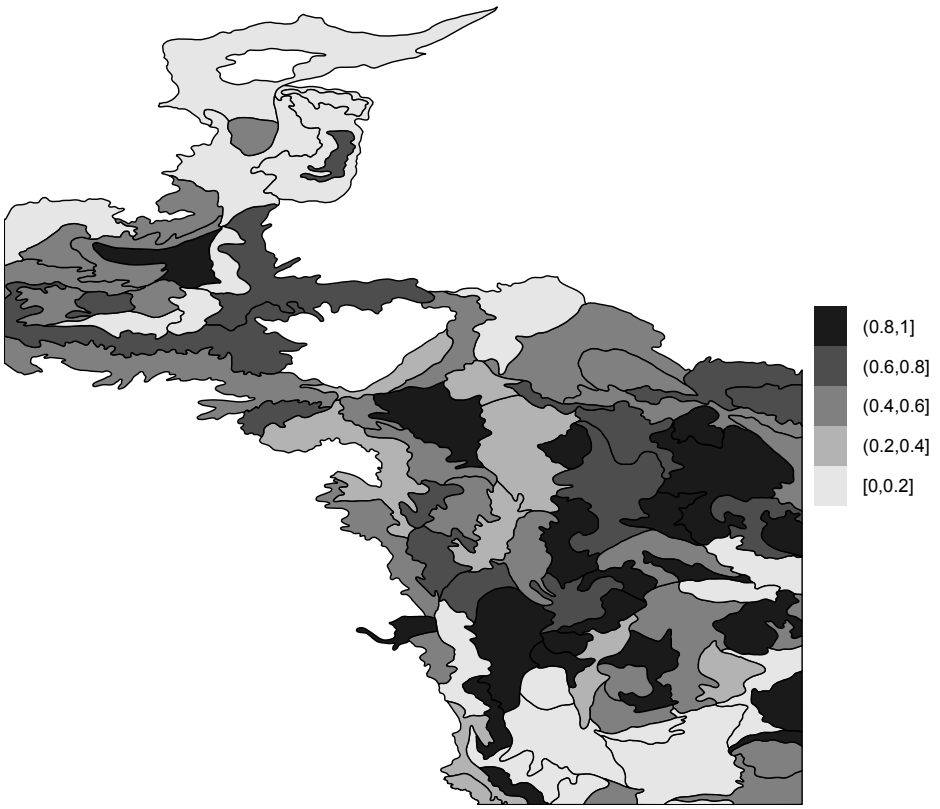


Figure 6.3: Probability of ecodistrict wheat yield in Alberta exceeding overall mean threshold. Consistencies with the spatial variability in 6.1 are displayed. With highest probabilities observed throughout the central belt of Alberta and large variation between neighbouring ecodistricts.

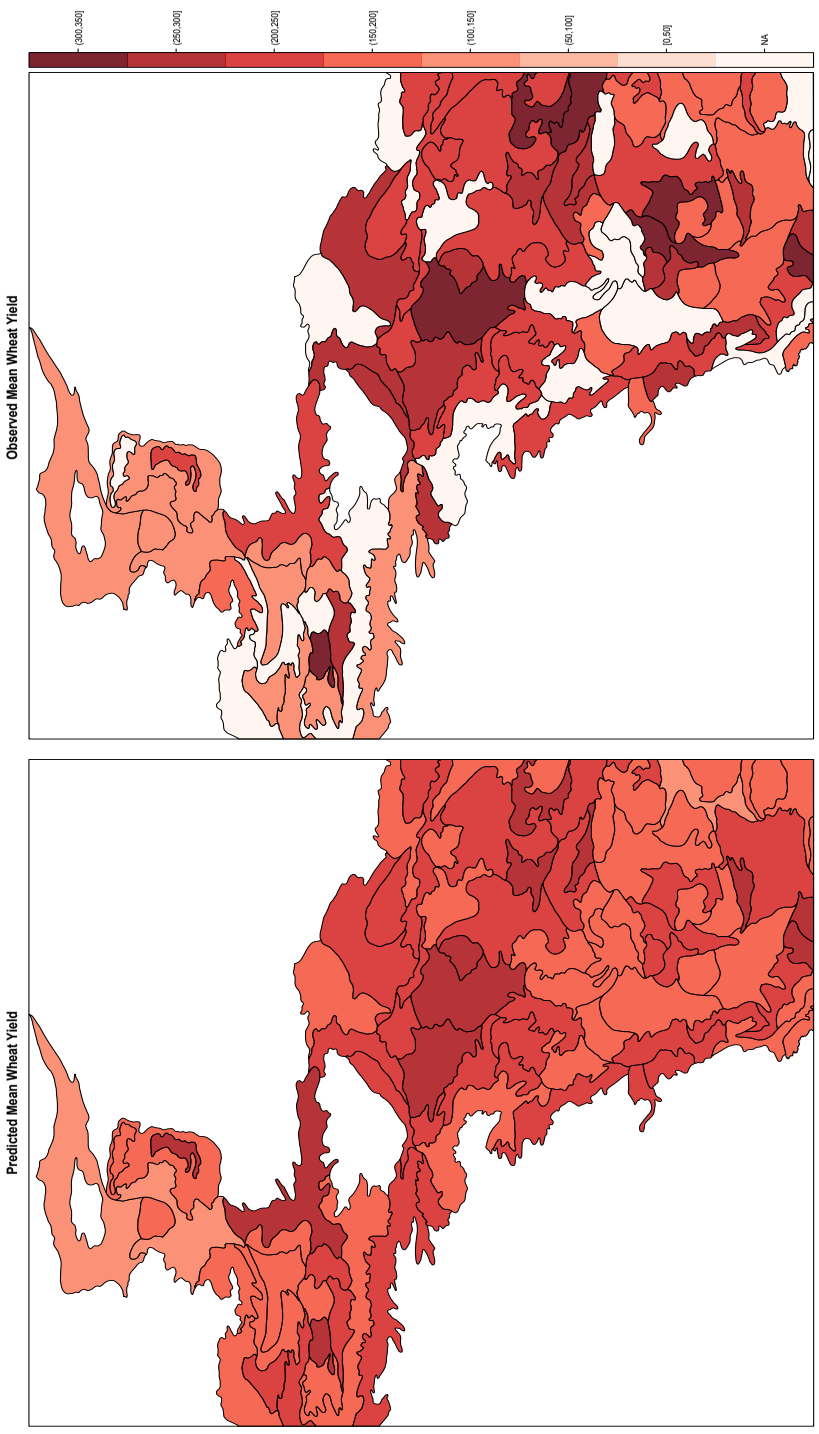


Figure 6.4: Mean wheat yield predictions over Alberta for the year 2004 - one year in advance. Predicted values displayed on the left are fitted using the data from 1999-2003. White ecodistricts within the observed data on the right indicate missing data for the year 2004.

Figure 6.5: Mean Wheat Yield Predictions over Alberta for the year 2004-two years in advance. Predicted values displayed on the left are fitted using the data from 1999-2002. White ecodistricts within the observed data on the right indicate missing data for the year 2004.

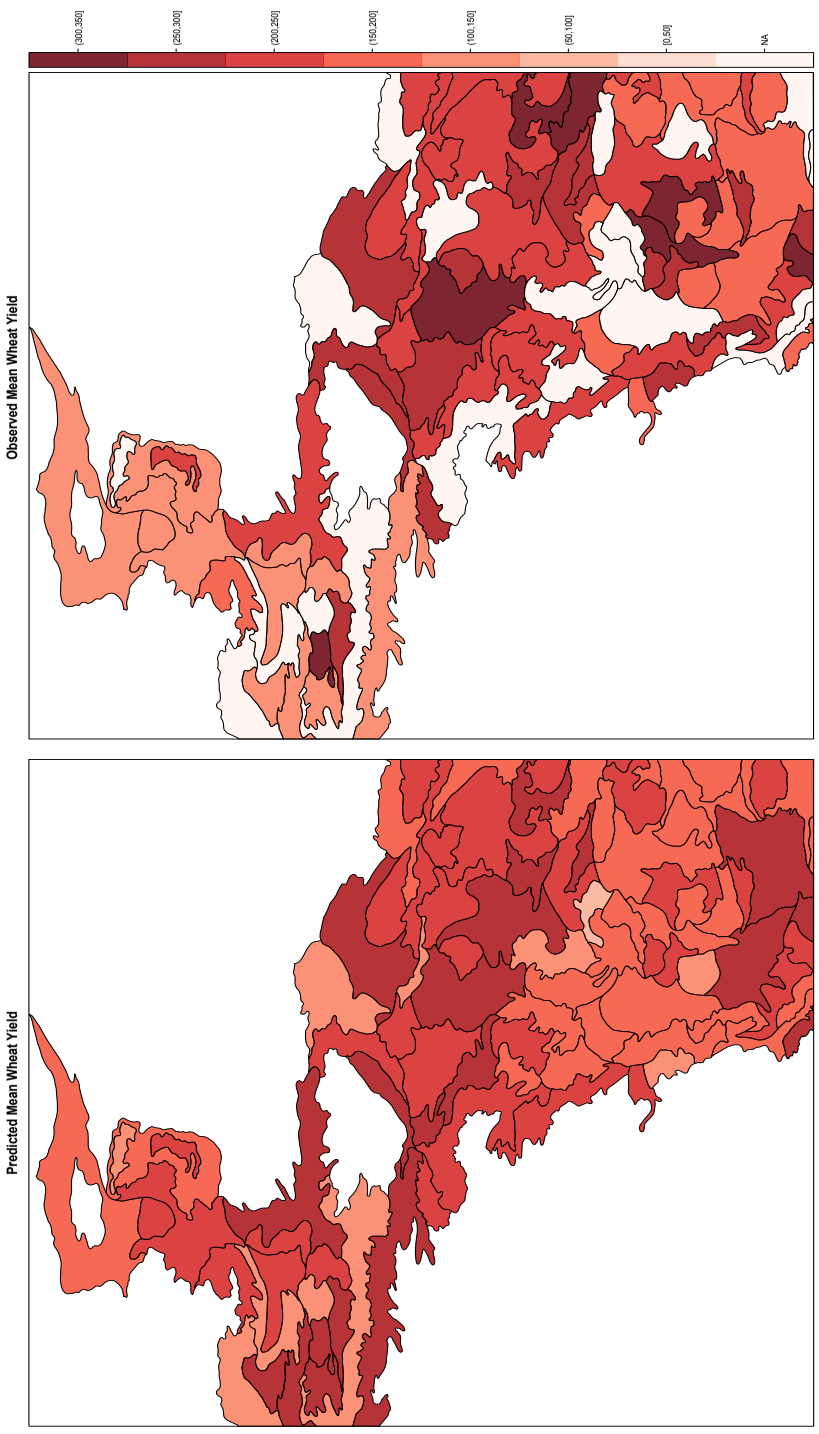


Figure 6.6: Mean Wheat Yield Predictions over Alberta for the year 2004-three years in advance. Predicted values displayed on the left are fitted using the data from 1999-2001. White ecodistricts within the observed data on the right indicate missing data for the year 2004.

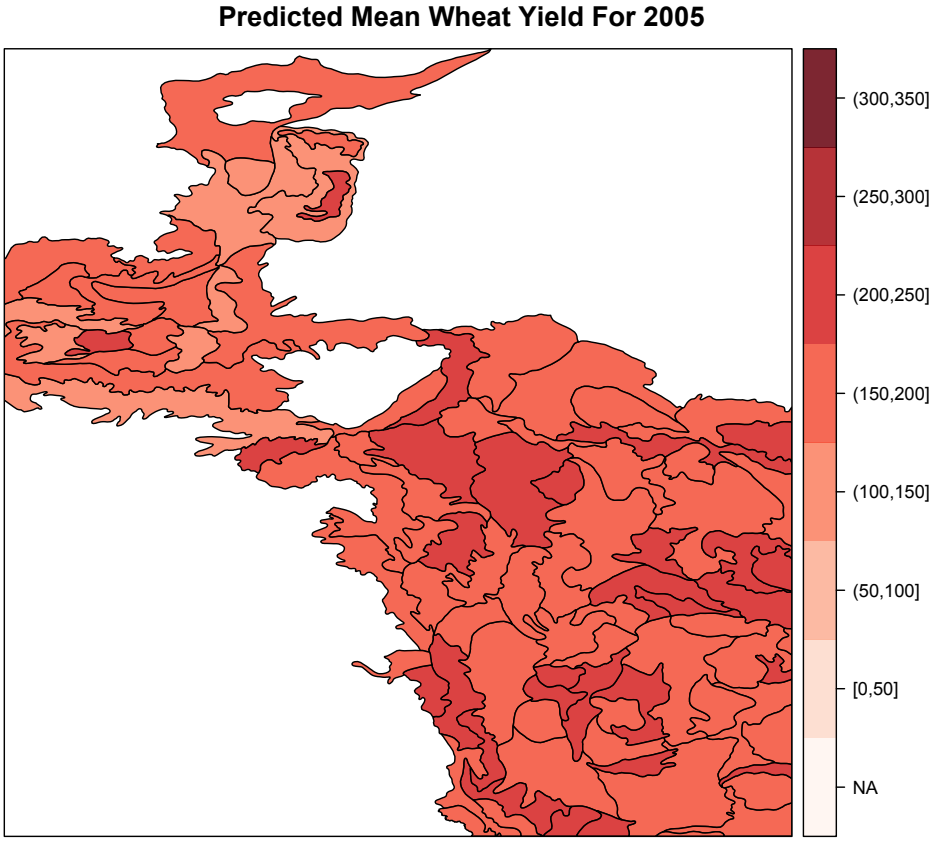


Figure 6.7: Mean Wheat Yield Predictions over Alberta for the year 2005- one year ahead of known data. Predicted values are fitted using all six years of available data from 1999-2004.

# Chapter 7

# Conclusion

The main objective of the research and analysis was to develop a computationally cheaper and accurate alternative method for modelling crop yield different from the traditional MCMC algorithms for estimating model parameters. Data for wheat yield in Alberta was used to test and validate this objective. This was accomplished using a spatio-temporal Bayesian model incorporating a neighbourhood structure via the use of the INLA method. Using the Laplace approximation and exploiting the Gaussian Markov Random Field properties of the ecodistrict neighbourhood structure, the INLA method allowed for a computationally reasonable algorithm for obtaining the posterior marginals of the model parameters.

Crop yield and climate data provided by AAFC was assimilated, manipulated, and interpolated. This was because yield data were provided in the form of one data point per year over the growing season at the ecodistrict level, this meant that data collected from the multiple climate stations in each ecodistrict had to be aggregated to the ecodistrict level. Data were provided from each climate station for each day throughout the year meaning that a single value for each covariate had to be obtained over the growing season for each year. Finally, some ecodistricts contained no data for wheat yield altogether (either because no wheat grows within that ecodistrict or the data were just missing) and hence data taken from the climate stations within these ecodistricts had to be removed and the shape file of Alberta that was created for use in `R-INLA` had to be adapted to exclude these ecodistricts.

Missing data in the covariates (minimum temperature, maximum temperature, water deficit, GDD, and total precipitation) was also resolved. This was a concern due to the fact that the INLA method can deal with

missing data in the response variable, however it is unable to handle missing data in the covariates. Three methods were considered to fill in the missing data. These included taking the average of the four nearest neighbours to each ecodistrict with missing data, taking the average of the entire data set for each covariate, and using INLA models treating the covariates as response variables and then using the fitted values from these models to replace the missing data. The nearest neighbour method was dismissed as it was found that the missing data tended to appear in blocks and in some occasions no average of neighbouring data could be computed. From the remaining two methods, the prediction of the missing data using the INLA method performed better than the averaging method in terms of model selection criterion and was hence selected as best.

Using the Deviance Information Criterion and a K-fold cross validated root mean squared error as the two model selection criterion it was deduced that the set of covariates most suited for explaining the properties of wheat growth in Alberta while also paying attention to model complexity were maximum temperature, minimum temperature and total precipitation. Furthermore, the model structure which was chosen involved including four separate components for structured and random spatial and temporal effects as well as an interaction effect between space and time. This structure was chosen opposed to two other model structures where the first excluded the interaction effect, and the second included a linear time effect and a space time interaction effect instead of including separate temporal effects.

The INLA method performed quite well in hind casting for a single previously removed year or forecasting for one year in advance. However, as more years are removed, the INLA model tends to predict closer and closer to the mean of the remaining data. For example, when three years are removed in figure 6.6, the model only has three years of data left to build from and the predictions will all be very close to the mean of the wheat yield from the years 1999 through 2001 (which is 203.56). Confidence can be placed in the one year predictions and caution should be taken when viewing predictions for multiple years.

In future work, a model which allowed for in season forecasting could

prove extremely useful. If wheat yield and covariate data were provided on a monthly basis, the INLA model could take on some added complexity by further separating the time component into monthly fractions. For example, in July this would allow one to predict the midseason forecast for wheat yield in August while using the already collected data from May and June instead of only allowing for a one year in advance prediction. Additional economic indicators relating to crop prices and distribution can also be included as covariates.

# Bibliography


[1] Allen, R. G. (1994), "Economic forecasting in agriculture," *International Journal of Forecasting*, 10, 81–135. → pages 1

[2] Besag, J., York, J., and Mollié, A. (1991), "Bayesian image restoration, with two applications in spatial statistics," *Annals of the Institute of Statistical Mathematics*, 43, 1–20. → pages 9, 24

[3] Blangiardo, M., Cameletti, M., Baio, G., and Rue, H. (2012), "Spatial and Spatio-Temporal models with R-INLA," *Spatial and Spatio-Temporal Epidemiology*, 4, 33–49. → pages 2, 3, 11, 12, 13, 15, 26, 27, 28

[4] Carew, R., Smith, E. G., and Grant, C. (2009), "Factors Influencing Wheat Yield and Variability," *Journal of Agricultural and Applied Economics*, 41, 625–639. → pages 2

[5] Dobson, A. J. . and Barnett, A. G. . (2008), *An Introduction to Generalized Linear Models*, Boca Raton, Florida: Chapman & Hall / CRC. → pages 36

[6] Fahrmeir, L. and Tutz, G. (2001), *Multivariate Statistical Modelling Based on Generalized Linear Models*, Berlin: Springer. → pages 16

[7] Lewis, A. S. and Overton, M. L. (2012), "Nonsmooth optimization via quasi-Newton methods," *Springer and Mathematical Optimization Society*, 141, 135–163. → pages 17

[8] Michel, L. and Makowski, D. . (2013), "Comparison of Statistical Models for Analyzing Wheat Yield Time Series," *PLoS ONE*. → pages 2

[9] Newlands, N. K. and Zamar, D. S. (2012), "In-season probabilistic crop yield forecasting, integrating agro-climate, remote-sensing and crop phenology data," *2012 Joint Statistical Meetings (2012 JSM)*. → pages 2

[10] Newlands, N. K., Zamar, D. S., Kouadio, L. ., Zhang, Y. ., Chipanshi, A. C. ., Potgieter, A. ., Toure, S. ., and Hill, H. S. . (2014), "An integrated, probabilistic model for improved seasonal forecasting of agricultural crop yield under environmental uncertainty," *Frontiers in Environmental Science*. → pages 2

[11] Pearson, K. (1905), "The Problem of the Random Walk," *Nature (London)*, 72, 318. → pages 9, 24, 25

[12] Prost, L., Makowski, D. ., and Jeuffroy, M. H. . (2008), "Comparison of stepwise selection and Bayesian model averaging for yield gap analysis," *Ecological Modelling*, 219, 66–76. → pages 2

[13] Rue, H. and Held, L. (2005), *Gaussian Markov Random Fields*, Boca Raton, Florida: Chapman & Hall / CRC. → pages 3, 12, 16

[14] Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations," *Journal of the Royal Statistical Society*, 71, 319–392. → pages vii, 15, 16, 17, 18, 19, 21, 22

[15] Spiegelhalter, D. J., Best, N. G., Carlin, P., and van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64, 583–639. → pages 31, 32

[16] Taylor, B. M. and Diggle, P. D. (2012), "INLA or MCMC? A Tutorial and Comparative Evaluation for Spatial Prediction in log-Gaussian Cox Processes," *eprint arXiv:1202.1738*. → pages 3