THE UNIVERSITY OF BRITISH COLUMBIA

DEPARTMENT OF STATISTICS

TECHNICAL REPORT #258

# Efficient Stabilization of Crop Yield Prediction in the Canadian Prairies

By

Luke Bornn

James V Zidek

April 2010

# Efficient Stabilization of Crop Yield Prediction in the Canadian Prairies

*Luke Bornn and James V Zidek*

**Abstract**

This report describes how spatial dependence can be incorporated into statistical models for crop yield along with the dangers of ignoring it. In particular, approaches that ignore this dependence suffer in their ability to capture (and predict) the underlying phenomena. By judiciously selecting biophysically based explanatory variables and using spatially-determined prior probability distributions, a Bayesian model for crop yield is created that not only allows for increased modelling flexibility but also for improved prediction over existing least-squares methods. The model is focussed on providing efficient predictions which stabilize the effects of noisy data. Prior distributions are developed to accommodate the spatial non-stationarity arising from distinct between-region differences in agricultural policy and practice. In addition, a range of possible dimension-reduction schemes are examined in the pursuit of improved prediction.

## 1 Introduction

This report presents a method for forecasting wheat crop yields in the Canadian Prairie Provinces – a challenging task due to dramatic variability in yield over space and time. Its importance, however, should not be understated: wheat is one of Canada's primary exports, accounting for 12 percent of wheat and barley traded in the world market. Thus variation in yield has considerable impact both within and beyond Canadian borders ([18]). Enabling effective crop management, handling, and marketing, thus requires accurate predictions of crop yield that account for and explain these variations. For example, these forecasts are helpful in setting insurance premiums and futures prices as well as in managing grain transport. Since spatial and temporal climate variability affect crop yields ([19]; [12]), a crop yield forecasting method must include climate as an essential component if it is to be successful.

Several process–based models have been successfully used for crop yield prediction including the Agricultural Production Systems Simulator (APSIM) in Australia ([9]) as well as a web-based tool developed by the United States' Southeast Climate Consortium ([8]). These process–based models typically employ tunable and user adjustable deterministic and stochastic models to simulate biological and physical processes related to crop yield. While these models use knowledge pertaining to the individual processes, they often require significant

input from the user, including a wide range of meteorological and environmental variables which may be difficult or expensive to obtain.

In contrast to the above, traditional statistical techniques are purely empirical. While these methods may result in accurate predictions, they typically lack the interpretability of process–based models ([2]. As a result of this criticism, recent years have seen the development of statistical models that also provide interpretation of the underlying biophysical process. One such process knowledge based approach involves water stress indices (SIs; [13], [12]; [14], [15]). While these developments have resulted in improved crop yield models, the majority are deficient in: a) not providing an efficient dimension reduction of explanatory variables; b) not accounting for uncertainty in the estimated technology trend; c) ignoring spatial correlation between regions.

This report describes the results of a project coordinated by Agriculture and Agri-foods Canada to develop a model that explains and predicts wheat yield and its relation to climatic variables. With plans for an online implementation in the future, efficiency was required as a feature of the model, as was the ability to stabilize the effects of noisy measurements. Building on earlier work, we employ a crop water SI to provide explanatory power for a new crop yield predictor. To improve prediction over existing approaches, we extract a sensitive yet low-dimensional summary of this stress index. We then demonstrate its improved prediction performance compared to currently used windowed average approaches. In contrast to previous work which models each agricultural region separately, we create a unified model that allows strength to be borrowed from adjacent and nearby regions, thus stabilizing both inference and prediction. By employing a spatially-motivated context-specific prior distribution on the parameters of interest, we account for and use spatial correlation between sites while smoothing and consequently improving predictions.

Following this introduction, Section 2 describes the crop yield forecasting problem and available data. This section works through a series of successively improved models, eventually leading to a Bayesian model in Section 3 which jointly models all regions simultaneously. Model testing and diagnostics are explored in Section 4. Lastly, Section 5 concludes the work.

## 2 Modelling Crop Yield with Biophysically Based Explanatory Variables

This report models crop yield in the Canadian Prairies as a function of climate-related explanatory variables. The data include annual wheat yields (in bushels per acre) along with associated measurements of a crop water stress index and growing degree day (both described later) for 40 agricultural regions across the Canadian Prairies from 1976-2006. The agricultural regions are those used in the 2006 Canadian Census of Agriculture and are determined from climate and soil information. For each of the 31 years and 40 regions, yield is an aggregated average across the the region. Likewise, stress index and growing degree day are calculated regionally, but on a daily basis throughout the growing season

(April 1 to September 30).

## 2.1 Incorporating Soil Water

The well recognized influence of soil water on crop yields dictates its inclusion in any yield prediction model ([4]). However, due to the time consuming and costly process of measuring soil water content, in practice its effects must be inferred from more widely available environmental variables such as precipitation, temperature, and easily measured crop and soil-related factors. A suite of models have been developed which attempt to understand soil water availability in the context of these environmental variables. Beginning with simple water balance approaches that balance precipitation and soil water storage with evapotranspiration and water runoff, these models have increased in their complexity over the years. We focus on budget models, which build on the premise that above a certain threshold (called the 'field capacity'), soil cannot absorb any more water and therefore any additional water is drained off through runoff or drainage. Also, if the soil water fails to be replenished through precipitation, irrigation, or other sources, the soil reaches a point where plant roots are no longer capable of uptaking water. This stage is known as the 'wilting point'.

Evapotranspiration, which describes the sum of evaporation and plant transpiration, measures the water lost from plants, soil, and other land surfaces into the atmosphere. There are two key components in the budget model, potential evapotranspiration (PET) and actual evapotranspiration (AET). PET represents the atmospheric demand for evapotranspiration; specifically, it accounts for the energy available to evaporate water and transport it into the lower atmosphere. AET is the actual water content available for evaporation and transpiration, and relies on plant physiology and soil characteristics for its calculation. When the soil has ample water, the actual evapotranspiration (AET) can equal the PET. However when the soil is not at its field capacity, AET will be less than PET.

Budget models are straightforward to implement since they require a minimum of meteorological data as well as soil field capacities and wilting points. While more advanced models have been built which include soil hydraulic characteristics and more complex relationships between soil, plant, and meteorological systems, these models requires considerably more information from the user, including detailed soil and plant characteristics. Because of the additional variables required by these models, we employ a budget model in the remainder of this work. Our model uses crop water stress index (SI), defined as 1-AET/PET. This quantity will be near 0 when water is plentiful in the soil and near 1 when the plant is stressed by a lack of available moisture. Intuition might suggest directly including precipitation, temperature, soil and plant information into the model. However, doing so would add a large number of variables, especially considered that many of these variables are observed for every day of the growing season. Using the SI instead provides an economical reduction in the dimensionality of the description space in a way that respects the biophysical processes involved in soil water movement and availability.

### 2.1.1 Predicting Yield with SI

We begin by detailing the process of fitting a regression model to crop yield using least squares (LS). First let $y_j$, $j = 1, \ldots, 40$ be the yield from region $j$ for years $t = 1976, \ldots, 2006$. Since SI is a daily value, we create an annual average for each year and region; let the vector $\overline{si}_j$ denote these means for each region. We begin by fitting a common regression model to all regions, specifically

$$y_j = \beta_0 + \beta_1 t + \beta_2 \overline{si}_j + \epsilon_j. \tag{1}$$

While previously developed statistical models for crop yield account for technology trend by first fitting a regression on time and then modelling the residuals, such approaches yield little understanding about the uncertainty associated with forecasting. In particular, while forecasts that use detrended data may be similar, their associated variances will be biased as uncertainty in the technology trend is ignored. As a result, to properly account for all sources of variability technology trend should be an integral part of any forecasting model.

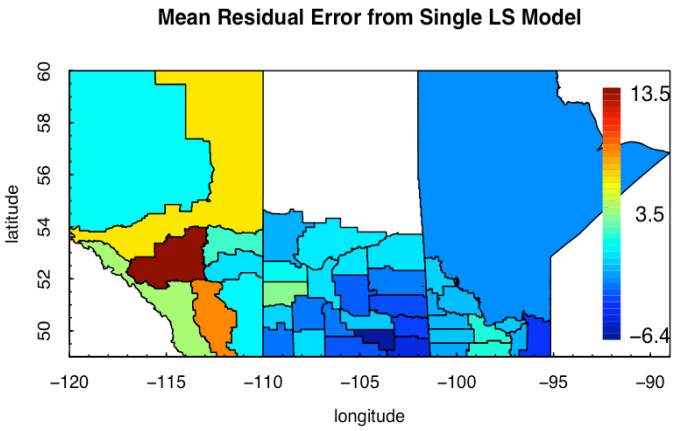
To begin, note that the simple model in Equation 1 relies on only 3 parameters – all regions are described by the same formula. The assumptions of such a model include, for instance, that the errors $\epsilon_j$ are stochastically independent for all $j$. To test this assumption, we plot the mean residual (averaged over the 31 years) for each of the 40 stations in Figure 1. This figure makes it clear that the residuals are spatially correlated. For instance, the residuals in Alberta (the western-most Prairie Province) are much larger than the other two provinces, highlighting the fact that the model is biased, particularly in central Alberta. Considering the mean and standard deviation of crop yield across the prairies is 30.9 and 8.2 respectively, the average residual value of 13.5 in this region indicates that the model is consistently underestimating the crop yield there.

To gain descriptive power, researchers have expanded the above model by fitting a different regression model to each region, specifically

$$y_j = \beta_{0,j} + \beta_{1,j} t + \beta_{2,j} \overline{si}_j + \epsilon_j. \tag{2}$$

The expanded model now accounts for 61% of crop yield variation, compared to 33% for (1), albeit at the expense of additional parameters. In fact, by assigning a unique parameter to each region, this expanded model has $3 \times 40 = 120$ parameters. By using such models, albeit with potentially modified/additional explanatory variables, several authors have been able to create fairly accurate predictions of crop yield ([13]; [15]). It is important to note that the large number of predictor variables (120) makes this model prone to overfitting; while some authors have used cross-validation to prevent this (i.e. [15]), others have exacerbated the problem by conducting extensive calibration to tune the explanatory variables (i.e. [13]). It is well understood that smoothed, or penalized, models have better prediction properties than larger, more variable models ([7]). This leads us to prefer the most parsimonious model yielding accurate forecasts and to select explanatory variables which provide optimal prediction power for crop yield.

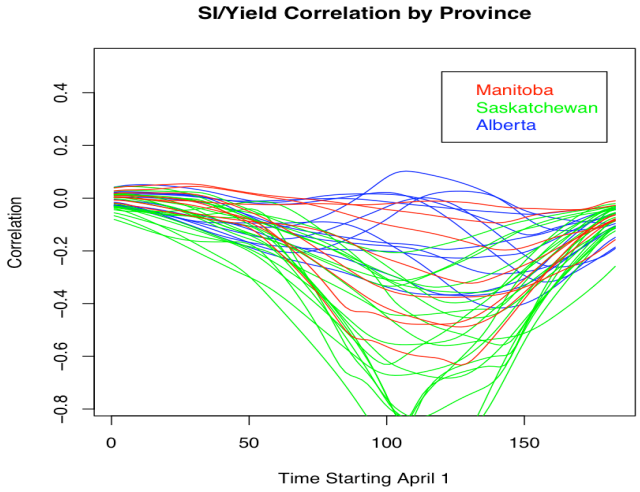
Fig. 1: Mean residuals from model (1).




The availability of SIs for every day of the growing season (in our case April 1 to September 30) means its vector of measured values is of very large dimension. Good modelling practice requires that this dimension be reduced before introducing the vector into the regression model. At one extreme, we could do what we did previously, and use just the mean of these daily SI values over the growing season, a one-dimensional feature, as our explanatory variable. However, that would oversimplify the SI's role, since plant growth is influenced more at certain times than others during the growing season. As an extreme example, if the crop is harvested in early September, the SI values in late September would aid little in predicting crop yield. To find a low-dimensional feature that provides good predictive power for crop yield, we could average over a reduced window, that is, exclude SI values early and late in the season ([15]). This reflects the point just made that SIs early and late in the season may not be correlated with crop yield. Figure 2 shows this correlation between SI and crop yield for each day in the summer, organized by province. This figure suggests we average over days 80 through 160, rather than the entire growing season. However, this produces only a modest improvement, 60.72% of crop yield's variability now being explained instead of 60.56% using the average over the entire season as before. This plot also reveals spatial variability, particularly between provinces. We explore this issue in more detail later.

There exists considerable scope for tuning this window; for instance [13] select unique window start and end points for each region to achieve an excellent fit – over 75% of variation explained. However such tuning entails much attention to detail. On top of the upper and lower limits for the averaging to take place, [13] calibrate potential available soil water capacities, the maximum

Fig. 2: Correlation of SI and yield over time. Correlations smoothed with Lowess smoothing. From this we see that SI is most correlated with yield in an intermediate part of the season, namely days 80 through 160.



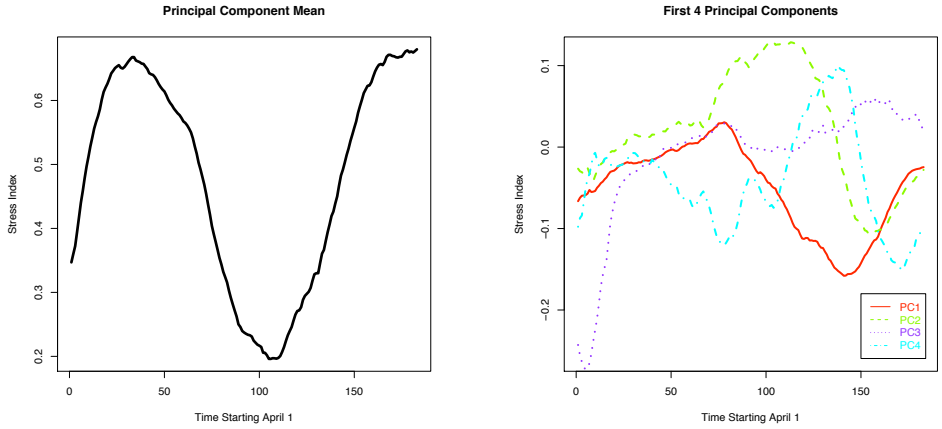

number of sowings and the rainfall amount triggering planting in each region. In other words, in addition to the corresponding regression coefficients, this tuning in effect adds 5 additional parameters per region, which in our case would increase the number of parameters being fitted in (2) from 120 to 480, leading most likely to serious over-fitting. To quote John von Neumann:

> "With four parameters I can fit an elephant and with five I can make him wiggle his trunk."

As such, a preferred alternative would be a lower dimensional feature which captures the key components of the stress index.

To capture more information from the SI values than would be available from simple averaging, we extract the principal components and hence main sources of variation from the stress index. To be more precise, after subtracting the average SI from each day, the first principal component is the linearly transformed vector of growing season SI values that accounts for the most variability in the SI values. The second, which is orthogonal to the first, explains the next largest amount of variation, and so on. Each observation, in this case each region – year combination, also has a set of loadings that, when multiplied by the corresponding principal components, return the original observation. Figure 3 shows the subtracted mean process as well as the first four principal components that together show the SIs history over the growing season of our study. Figure 3(a) reveals firstly the primary shape of the stress index, showing that initially – from April 1 – the stress is moderate, increasing until May, followed by a

Fig. 3: Principal components and mean for SI. This figure depicts the major patterns in the variation of the stress index (unitless) over the growing season. Observe how the first four principal components pick up deviations from the overall pattern in (a), and reveal the peaks and valleys of the stress cycle over the course of the summer induced by things like patterns in precipitation and temperature. Together these four components capture most of the variation in stress in a very economical way and eliminate the need for the high dimensional vector of daily SI values.

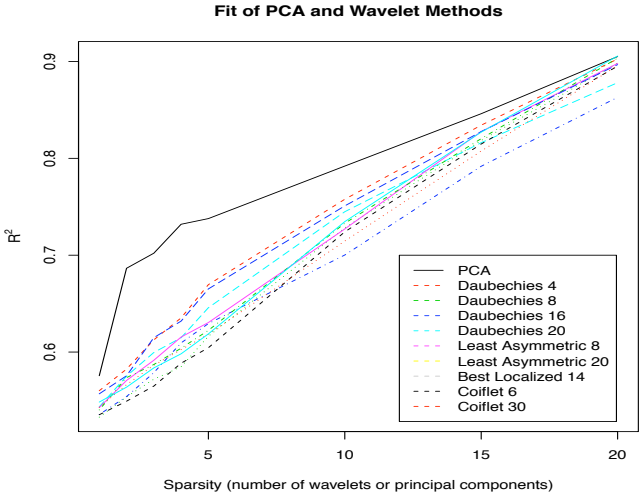



(a) (b)

gradual gradual decline until it bottoms out in July. It then returns to its highest values by the end of September. The first component (Figure 3(b)), which describes 46.9% of the variation in SI, captures a valley in the SI cycle around late August. The second component, which accounts for 14.6% of the variation, shows SI's decline into its July valley followed by its rise to its early September peak. The orthogonality of the first two components is apparent from (b). The third and last major component of SI's variation captures its low April start. Altogether, the first 4 principal components account for 78.5% of the variation in SI over the growing season. Thus by including the loadings for these 4 principal components as explanatory variables, we have created a 4-dimensional feature which accounts for a large proportion of variation in the stress index.

Note that the first SI principal components aren't necessarily the best predictors of yield. However, LASSO – a penalized least squares variable selection method – in fact selects these same four principal components as the best four ([7]). This choice of feature also has a natural biophysical interpretation. For instance, a large and positive regression coefficient for the loadings corresponding to principal component 3 would imply that a reduction in stress in early April is highly connected with increased crop yield. By using this approach,

Fig. 4: $R^2$ of the crop yield model for a range of bases and sparsity levels. From this we notice that principal components (PCA) provide better model fit for all sparsity levels.




the explained variance of the regression model increases from 60.56% from averaging SI over the growing season to 70.06%. Using these principal component loadings, our new model is

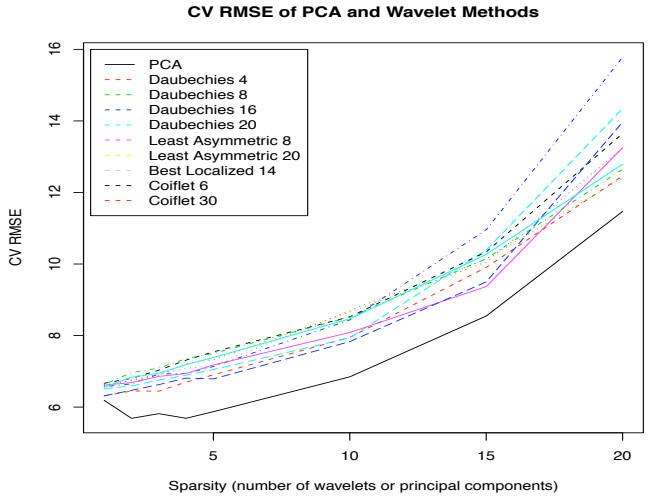
$$y_j = \beta_{0,j} + \beta_{1,j}t + \beta_{2,j}PC1_j + \cdots + \beta_{5,j}PC4_j + \epsilon_j. \tag{3}$$

where $PC1_j$ indicates the loading for principal component 1 in region $j$.

### 2.1.2 Alternative Bases and Levels of Sparsity

Because of their widely documented ability to model complex nonlinear signals while maintaining sparsity, we briefly explore wavelet bases as an alternative to principal components ([11]). Specifically, we examine a variety of different wavelet bases and levels of sparsity both in terms of cross-validated prediction error as well as $R^2$. Figure 4 plots $R^2$ of the yield model for various bases and levels of sparsity. From this plot we see that principal components dominate in terms of model fit. Figure 5 plots cross-validation root mean squared error (in bushels per acre) for each basis and sparsity level. Once again we observe that principal components outperform wavelets. From these figures we conclude that principal components lead to a model with better fit and prediction performance. This example highlights the need to be selective in the choice of basis to represent stress index and other variables in such a model. While wavelets excel at representing piece-wise smooth models in a very sparse way (requiring the storage of only 1 vector – the mother wavelet – as well as a series of indices),

Fig. 5: Cross-validation RMSE (bushels per acre) of the crop yield model for a range of bases and sparsity levels. From this we notice that principal components (PCA) provide better prediction than the wavelet bases for all sparsity levels.




this is also their downfall in some circumstance such as this one which require a more rich representation.

## 2.2 Incorporating Temperature

Temperature affects a plant's development and growth in a variety of ways, in particular its photosynthesis and respiration. In general, temperature affects plant functioning through its action on enzymatic reactions. At low temperatures, enzyme proteins are not sufficiently flexible to complete the conformation necessary for enzymatic reaction. Conversely, high temperatures can coagulate the enzyme leading to similar barriers to the reaction. Alongside a minimum and maximum temperature to allow growth, most plants have an optimum temperature to encourage growth. For instance, [16] conclude that the minimum and optimum temperatures for wheat are respectively 0 and 20-25 degrees celsius. As a result of temperature's influence on plant development, we suspect that its inclusion into the model will result in prediction performance gains.

### 2.2.1 Growing Degree Day

While temperature could go directly into the model, its measurement in hourly or smaller increments creates a considerable amount of data. As a result, some dimension reduction is needed to limit the number of explanatory variables. One could do this using just the maximum and minimum daily temperatures

or better still, a one dimensional summary that combines the two. Thus 'growing degree day' (GDD) measures the heat accumulation in a region based on local weather by taking an average of the daily minimum and maximum and subtracting a base temperature as follows:

$$GDD = \max\left(0, \frac{T_{max} + T_{min}}{2} - T_{base}\right).$$

Thus the GDD measures the daily average temperature but in a way that reflects the extremes more sensitively. The base temperature represents the physiological temperature below which development would be zero.

A day with a high and low of 30 and 15 degrees celsius and a base temperature of 10 degrees would have a GDD value of 12.5 degrees celsius. Thus GDD is a simple, single-dimensional summary for describing the plant's exposure to heat. While GDD is a simple heuristic, it is commonly used by horticulturists to estimate the stages of a plant's growth. As an example, the maturation of wheat corresponds to about 1600 GDDs ([5]). Thus GDD provides us with a simple low-dimension summary of temperature which allows for comparison of the thermal time available in different climatic zones.

### 2.2.2 Predicting Yield with GDD

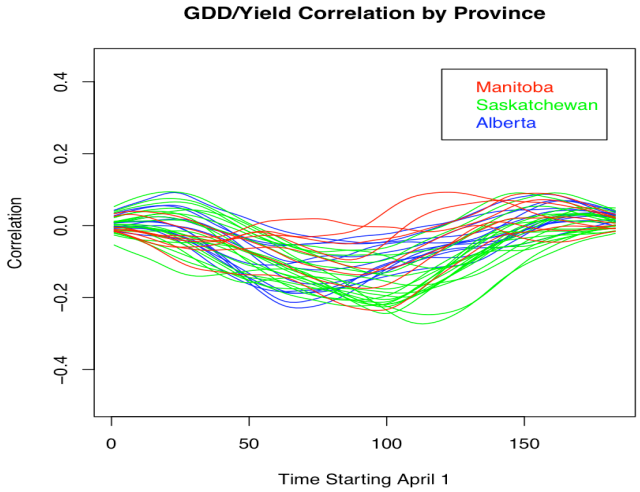
While SI gives scientific insight into the moisture available for plant growth, it says little directly about the heat available to the crop. Thus to improve our model we can also include GDD, which up until now has been used primarily in this context for tuning the explanatory variables ([12]). Like SI, GDD is a daily value, and hence can be treated similarly. Thus through the correlations plotted in Figure 6 we look at the time of season where GDD is most correlated with yield. This figure tells us that an appropriate window would be the one bounded by days 50 through 160. Using a cumulative average over the whole season, the explained variation in yield increases from 70.06% to 73.20%, with the shortened window performing similarly. In addition to averaging over the whole season or a shorter window, we can also use principal components as we did for SI above. While using the first 4 principal components only increases this to 76.48%, the additional 120 variables result in reduced cross-validation prediction performance, hence we prefer using just the windowed average. The expanded LS model 3 then becomes

$$y_j = \beta_{0,j} + \beta_{1,j}t + \beta_{2,j}PC1_j + \cdots + \beta_{5,j}PC4_j + \beta_{6,j}\overline{GDD_j} + \epsilon_j. \qquad (4)$$

where $\overline{GDD_j}$ is the windowed average of GDD in region $j$. It is worth noting that temperature is a component of SI; however, the addition of GDD into the model improves model fit and prediction.

We compare the previous models as well as those developed later in the report in Table 1, showing the features and performance of each successive model. The traditional regression models represented in Table 1, fitted for each region separately, ignore a considerable amount of information. Specifically,

Fig. 6: Correlation of GDD and Yield over Time. Smoothed with lowess smoothing. From this we see that a reduced window average may be appropriate.




Tab. 1: Features of various models. We see that while model 4 has the best fit to the data ($R^2 = .73$), the Bayesian model gives the best prediction performance in terms of cross-validated root mean squared error (in bushels per acre). Effective parameters is defined as $tr(S)$, where $\hat{y} = Sy$, and may be concerned a measure of model complexity ([7]).

| Model | Parameters | Effective Parameters | $R^2$ | CV RMSE |
|---|---|---|---|---|
| 1: Single LS | 3 | 3 | .33 | 6.83 |
| 2: LS with SI | 120 | 120 | .61 | 5.79 |
| 3: LS with PCA | 240 | 240 | .70 | 5.72 |
| 4: LS with PCA + GDD | 280 | 280 | .73 | 5.69 |
| 5: Bayes | 280 | 139 | .70 | 5.35 |

because of the close spatial proximity of the regions, considerable strength may be gained by exploiting the correlation among regions. For instance, use of neighbouring SI values can help stabilize predictions based on SI values, since the latter come from a small set of regional monitoring stations and hence can be fairly noisy. The amount of borrowed strength can be considerable when the correlation between stations is high. In addition, modelling all stations jointly while incorporating spatial information allows us to continue to make predictions even in the presence of missing or noisy data. If a measuring station goes out of operation temporarily, its missing values may be inferred from data collected at nearby regions to yield accurate forecasts. This idea leads into our next section, which focuses on spatial models that look at all regions together in a unified manner.

## 3 A Context-Specific Spatial Bayesian Approach

Classical regression methods rely on the assumption that their model residuals are uncorrelated. Indeed violation of that assumption can have very serious deleterious effects on parameter estimates compared, for example, to violations of the assumption that those residuals have a Gaussian distribution. In our case the residuals are most certainly spatially dependent and thus the actual amount of information in the data can be much less than the assumptions underlying (1) would suggest. The unwary analyst would then be led to make overconfident forecasts with biased parameter estimates having unduly small standard errors.

One work around would model the regions separately. However, this wastes the benefits spatial dependence provide for borrowing strength by telegraphing information across the regions through the wires of correlation for the mutual improvement of all their forecasts. This progression naturally leads us to a Bayesian framework for handling this problem, one which jointly models all regions simultaneously while accounting for their spatial dependence. Thus we move from the frequency paradigm of classical statistics to the Bayesian paradigm of modern statistics.

These two paradigms, which tend to give the same inferences at least for fairly large datasets, are very different in concept. Frequentists see data as being generated by a system governed by some true but unknown parameters. They commonly seek to estimate these true parameters well in some sense, for a variety of inferential purposes such as forecasting. The central tenet of their theory is repeated sampling – in the long run the parameters can be estimated to arbitrarily high levels of precision if the system producing the data were unperturbed. However, Bayesian statisticians reject the notion of repeated sampling as a fundamental construct in their theory, recognizing realistically that most systems can not remain unperturbed and pump out replicate data over an extended sequence of trials. Although their models involve uncertain parameters, these parameters like all uncertain objects such as future data values, are characterized by a probability distribution. Initially that distribution, called a prior, simply reflects the Bayesian's own knowledge. An abundance of such knowledge

would mean a prior concentrated around a single point and a state of near certainty. The information in the data adds to the state of knowledge through the celebrated Bayes theorem. The latter relies on the likelihood function of the uncertain parameters which captures all the information in the data. A likelihood tightly concentrated around a single value would mean the data has eliminated much of the uncertainty about the parameters. However generally, Bayes rule needs to be applied to get the combined effect of data and prior knowledge; this yields the Bayesian's updated prior, or the so-called posterior distribution. Due to its adaptability and ease of use, Bayesian inference has become a prominent fixture in modern spatial statistics, and in particular the modelling of random spatio-temporal fields ([1]; [10]).

## 3.1 Available Prior Information

Consider, for example, the spatial structure discussed above. Even before estimating the parameters in equation 3, we expect parameters in adjacent regions to be similar. Thus we would be surprised if the parameters relating GDD to yield had completely opposite signs in two neighbouring regions. This reflects our prior beliefs about those parameters, namely that knowledge of one would tell us something about the other. More simply, we would see them as stochastically dependent in the language of the probability distribution that characterizes our beliefs about them. We might even have some idea of their approximate magnitudes. For instance, a magnitude of 100 (bushels per acre/degree celsius) for the coefficients $\beta_{6,j}$ for GDD would be completely untenable, since it would mean that changing one cold day to a warm one (adding, say, 10 GDD over the entire cumulative season), would increase the yield by roughly 10 bushels per acre. Thus even without formalizing our beliefs in a prior distribution, loose bounds on parameters are almost always apparent.

Application of the Bayesian approach starts by characterizing our beliefs about the parameters in the form of a prior distribution. In the regression models introduced above, this would amount to a joint prior distribution on each $\beta$ to account for our belief in their dependence (similarity) for adjoining regions. For simplicity, stack all of the coefficients into a vector $\boldsymbol{\beta}$, the first 7 coefficients being for all variables in region 1, the next 7 for region 2, and so on. Assuming a Gaussian distribution as a convenient prior form, we can explicitly write the prior as follows:

$$\boldsymbol{\beta} \sim N\left(0, \Sigma_0 \otimes g\Omega\right). \tag{5}$$

By using such a Kronecker structure, $\Sigma_0$ models the correlation within a given coefficient across space, while $g\Omega$ corresponds to Zellner's $g$-prior ([21]) with $\Omega$ the $7 \times 7$ empirical covariance between explanatory variables. We now specify $\Sigma_0$, the correlation between regions, as

$$\Sigma_0 = \exp(-D/\phi), \tag{6}$$

with a slight abuse of notation where $D$ is the matrix with element $(i, j)$ the Euclidean distance between regions $i$ and $j$ (as measured from the centre of the

region). Here $\phi$ is a parameter controlling the range of the variogram. In this way, $\phi$ controls how spatially smooth the coefficients are, while $g$ controls how tight around zero the coefficients are.

While we suspect neighbouring regions to be similar, Figure 2 highlights the differences between provinces. In fact, the varying irrigation and technology policies in each province result in a sharp boundary between provinces. As such, it is not entirely logical to use a stationary prior ([3]) which assigns correlation between regions solely based on distance without any respect for political boundaries. As a result, we adjust our prior distribution to have reduced correlation between regions in different provinces. While the the obvious approach is to scale down the prior correlation between regions in different provinces with a constant value, this may lead to non-positive definiteness of $\Sigma_0$; alternative methods which do not suffer from this problem are therefore needed. We accomplish this task by deforming the physical space, in effect pushing neighbouring provinces apart. Motivated from [17], this artificial distortion of the space results in a stationary prior in the deformed space, yet a nonstationary one in the original space. The distance $d$ (measured in degrees latitude/longitude) by which the provinces are pushed apart in the artificial space is tuned through cross-validation. Searching over the integers from 1 to 10, we find $d = 4$ to give the best prediction performance (CV RMSE of 5.35 vs 5.39 for $d = 0$), intuitively meaning that Alberta and Manitoba are pushed respectively west and east from Saskatchewan by 4 degrees longitude in the artificial space. The end result is a reduction in the off-diagonal elements of $\Sigma_0$ corresponding to between-province regions while maintaining positive definiteness.

## 3.2 Likelihood and Posterior Distributions

We begin by employing the likelihood corresponding to (3), namely

$$y_j \sim N\left(\beta_{0,j} + \beta_{1,j}t + \beta_{2,j}PC1_j + \cdots + \beta_{5,j}PC4_j + \beta_{6,j}GDD_j, \sigma^2\right). \quad (7)$$

We also assign an Inverse-Gamma prior distribution on $\sigma^2$ with parameters $a$ and $b$ set to be highly noninformative. Before proceeding, we introduce the notation $y$, the column vector of stacked $y_j$, and $X$, the $(31 \times 40) \times 240$ block-diagonal matrix of explanatory variables. Using Bayes theorem to combine our initial knowledge (in the form of prior distributions) and the information provided by the data (in the form of the likelihood), we can obtain the posterior distribution of the parameters. Specifically, for the regression coefficients $\boldsymbol{\beta}$, the marginal posterior is obtained using Bayes Theorem as follows:

$$\pi(\boldsymbol{\beta}|X, y) \propto \int \pi(y|X, \boldsymbol{\beta}, \Sigma)\pi(\boldsymbol{\beta}|\Sigma)\pi(\Sigma)d\Sigma. \quad (8)$$

Due to the conjugate nature of the prior and likelihood, we are able to analytically complete this integral. The resulting distribution is a multivariate

Student-T,

$$\begin{aligned} \boldsymbol{\beta} &\sim T\left(\boldsymbol{\beta}_f, \Psi, n+2a\right) \\ &\text{where} \\ \boldsymbol{\beta}_f &= (X^T X + (\Sigma_0 \otimes g\Omega)^{-1})^{-1}(X^T y) \\ \Psi &= (\, X^T X + (\Sigma_0 \otimes g\Omega)^{-1})(SS + 2b)/(n+2a) \\ SS &= y^T y - \boldsymbol{\beta}_f^T (X^T X + (\Sigma_0 \otimes g\Omega)^{-1})\boldsymbol{\beta}_f^T \end{aligned} \tag{9}$$

From this last expression, we get the posterior mean $\boldsymbol{\beta}_f$, which may be used as a simple estimator for $\boldsymbol{\beta}$. In fact, comparing $\boldsymbol{\beta}_f = (X^T X + (\Sigma_0 \otimes g\Omega)^{-1})^{-1}(X^T y)$ to the LS estimate $(X^T X)^{-1}(X^T y)$, we readily see how the prior covariance affects the parameter estimates. In particular, a diffuse prior distribution adjusts the estimate little, whereas an informative prior distribution – one that is fairly tightly concentrated around zero – shrinks the posterior estimate considerably.

Setting $g = 10$ and $\phi = 10^6$, we obtain coefficient estimates as shown in Figure 7, which also shows the corresponding least squares estimate using (3). We see that the spatial information used in the Bayesian model causes the coefficients to be more correlated across space. In addition, the zero-mean prior distribution leads to some shrinkage in the coefficient estimates. Interestingly, we notice little shrinkage in the estimated coefficient for technology trend, suggesting that the data contains considerable information on this quantity.

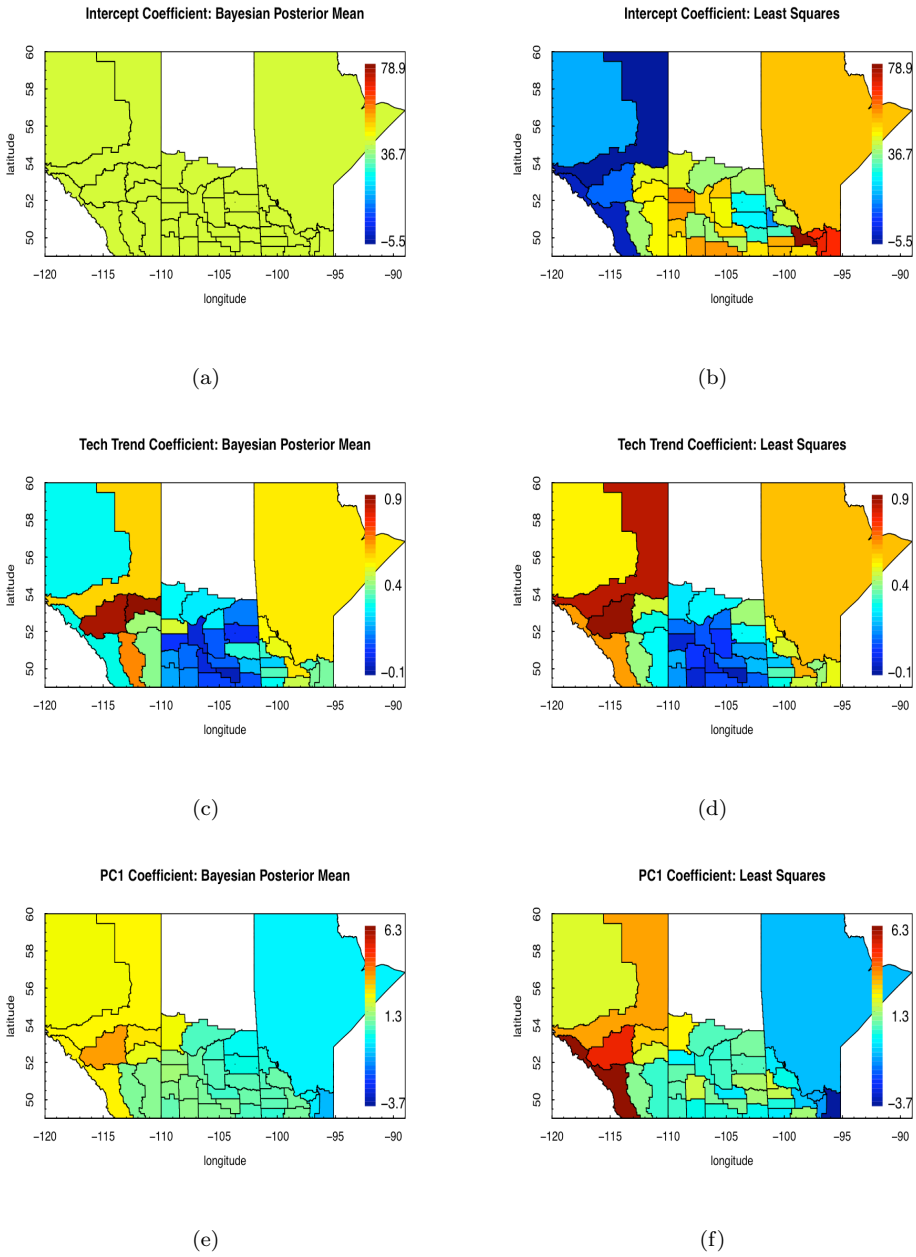
## 4 Model Testing and Diagnostics

We proceed by comparing the prediction performance of the least squares and Bayesian methods. To accomplish this we use leave-one-out cross-validation, removing years one at a time in succession to compare each model's predictive ability. More specifically, we successively remove each year in turn, using the remaining years to find the posterior mean, notated $\hat{\boldsymbol{\beta}}^i$ if year $i$ is removed. This posterior mean is then used to perform prediction on the removed year. From this the root mean squared error (RMSE) is calculated as the square root of the sum of squared prediction errors for each year and region.

$$RMSE = \sqrt{\sum_{i=1}^{31}\sum_{j=1}^{40}(y_{i,j} - X_{i,j}\hat{\boldsymbol{\beta}}^i)^2/(31 \times 40)}. \tag{10}$$

Figure 8 shows the cross-validation root mean squared error (RMSE) of the posterior mean estimate for various settings of $g$ and $\phi$. As $g \to \infty$ and $\phi \to 0$, the Bayesian model converges to the least squares solution, as evidenced by converging cross-validation errors. However, if $g$ is too small, the prior on the regression coefficients is too informative towards zero, and hence the resulting posterior means are overly shrunken, resulting in poor prediction (RMSE $>$ 6). While one could assign prior distributions to these parameters, we prefer finding them through cross-validation for computational efficiency. Specifically,

Fig. 7: Coefficient surfaces for intercept, technology trend, and PC1. Other coefficients are similarly smoothed.

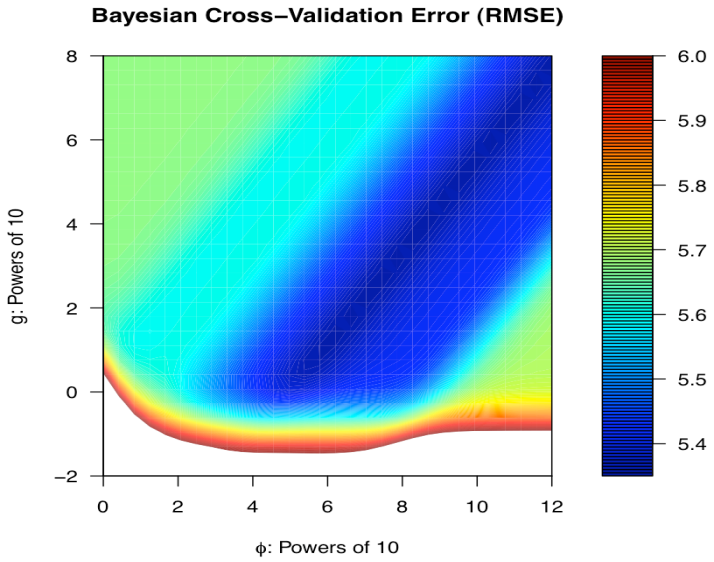



(a) (b)

(c) (d)

(e) (f)

Fig. 8: Cross-validation errors using Bayesian posterior mean. For comparison, the least squares error is 5.69. From this, we observe a ridge of excellent prediction. Hence there is some tradeoff between the two parameters to be tuned.



given the optimal parameters, the model is conjugate, and hence sequential updating and prediction is analytic and therefore nearly instant. It is very interesting to note that the optimal prediction error for the Bayesian model is less than for the least squares model, indicating that prediction is improved with regularization (provided by the zero-mean prior and/or correlation). The area of lowest prediction error occurs along a diagonal of $g$ and $\phi$ and has value approximately 5.35. This is likely due to the fact that an increase in $g$ results in a more diffuse posterior which regularizes less, while increases in $\phi$ result in increased correlation between regions and hence more regularization. Hence the optimal prediction seems to occur for moderate amounts of regularization.

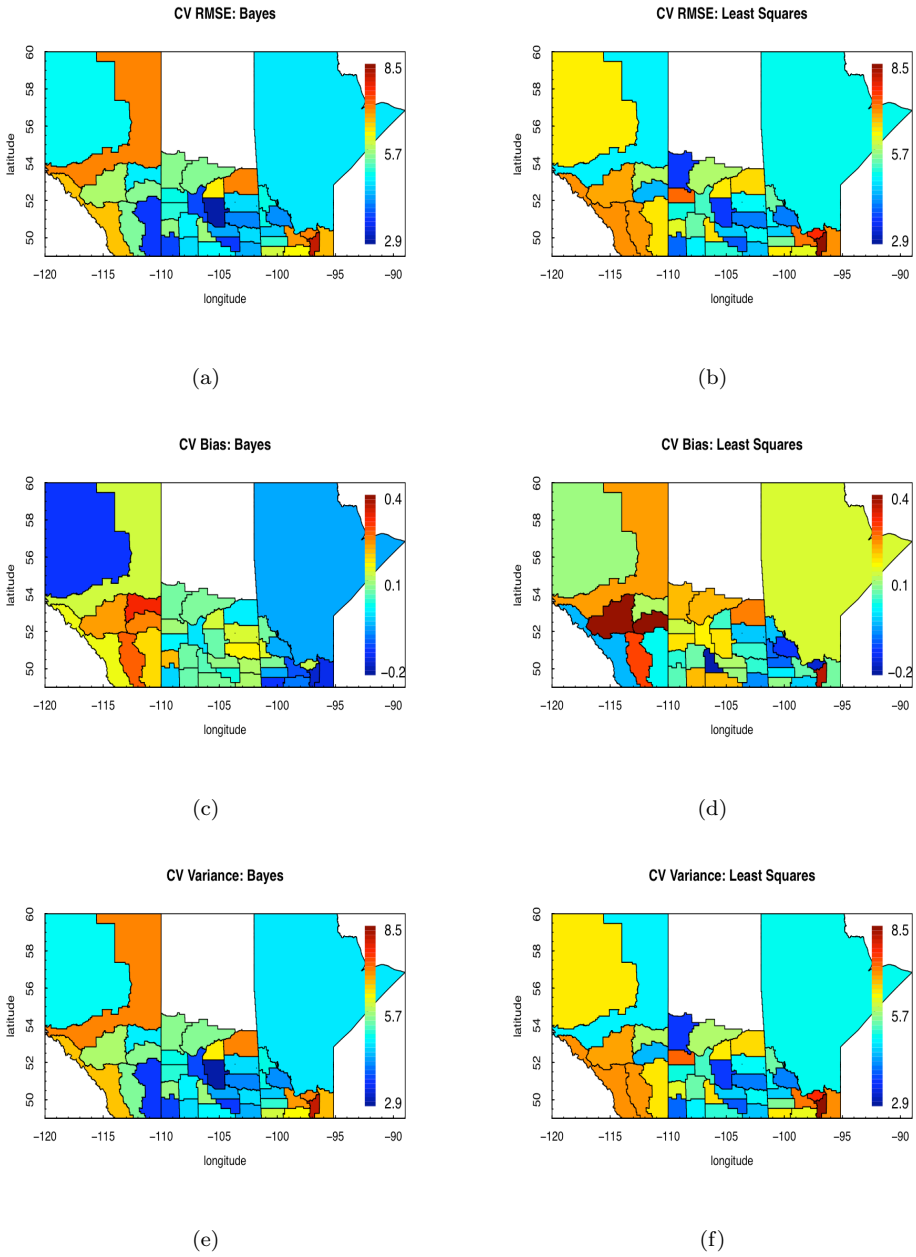
The cross-validation RMSE can also be calculated for each region by summing only over years. In this way we can gain an improved perspective on the model's prediction performance. However, while cross-validation RMSE gives an idea of the prediction performance of a model, it does little to tell of a model's bias. To do this we decompose the RMSE into the model's prediction bias and variance. Doing this for each region, we obtain Figure 9 detailing the prediction RMSE, bias, and variance of the Bayesian and LS models in each region. From this figure we observe that, with the exception of one or two individual regions, the Bayesian model improves RMSE in all areas except for southern Manitoba. Digging deeper, we see a negative bias in this area. Thus the regularization of the model is perhaps not useful in this region due to some systematic differences in this area. Specifically, this section of southern Manitoba is known to use significant irrigation ([6]). As a result, further model development might be explored in this area to account for irrigation.

## 4.1 Conclusion

In this report we have examined the role of SI in predicting crop yields, emphasizing the need to create a judicious low-dimensional summary in order to improve prediction. Simply averaging SI over the entire season is inefficient, as yield may be insensitive to stress in certain parts of the summer. The traditional solution to this problem is to average over a reduced window of data, hence cutting out those areas lacking in sensitivity from the analysis. However, this one dimensional feature is not particularly sensitive to changes in stress indices within that window. For example, a region which has low SI in June but high SI in July might ultimately have the same averaged value as another region which had just the opposite trend. To address this issue, we have implemented principal components analysis to create a set of flexible summary statistics which better describe the variations in CWSI, and as a result improve prediction considerably. We also demonstrated principal components' improved performance over wavelet bases.

We have also shown the importance of incorporating spatial correlation into crop yield models; ignoring this information can lead to bias both in model identification and prediction. Specifically, we observed that a common least squares fit of crop yield on some explanatory variables over the entire region resulted in biased residual errors, and hence violated the assumptions of the model. One

Fig. 9: Cross-Validation results by region. The Bayesian model improves prediction by all standards in the majority of regions.




(a) (b)

(c) (d)

(e) (f)

method to avoid this is to fit each agricultural region with its own model. The problem, however, is that this ignores information between crop regions, and as such we observed reduced prediction power and model identifiability. We addressed this issue through the use of a Bayesian model which modeled all regions together, yet accounted for spatial correlation. This model smooths and stabilizes prediction and also allows for analytic and therefore efficient updating and prediction. In addition, we created a non-stationary prior distribution to address the issue of province to province variability resulting from provincial differences in policy and management. Through cross-validation, we demonstrated this model to achieve improved prediction performance over the least squares model which ignores spatial dependence.

## Acknowledgements

The authors wish to thank Charles Serele, Harvey Hill, and Nathaniel Newlands, of Agriculture and Agri-foods Canada for providing the data used in the report as well as useful suggestions throughout the model's development. The research reported in this report builds on the work of Wendy Wu ([20]) whose work was co-supervised by the second author and the National Agroclimate Information Service (Agriculture and Agri-Foods Canada), as part of a research program supported by Canada's National Institute for Complex Data Structures in partnership with Agriculture and Agro-Foods Canada.

## References

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

[2] V. Barnett. *Environmental statistics: methods and applications*. John Wiley and Sons, 2004.

[3] N.A.C. Cressie. *Statistics for spatial data*. John Wiley & Sons, New York, 1993.

[4] R. De Jong and A. Bootsma. Review of recent developments in soil water simulation models. *Canadian Journal of Soil Science*, 76(3):263–273, 1996.

[5] KA Dolan, RB Lammers, and CJ Vöröosmarty. Pan-Arctic temperatization: A preliminary study of future climate impacts on agriculture opportunities in the Pan-Arctic drainage system. *Eos, Trans. Amer. Geophys. Union*, 87, 2006.

[6] Gaia Consulting Limited. 2006 manitoba irrigation survey. 2007.

[7] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009.

[8] SS Jagtap, JW Jones, P. Hildebrand, D. Letson, JJ O'Brien, G. Podestá, D. Zierden, and F. Zazueta. Responding to stakeholder's demands for climate information: from research to applications in Florida. *Agricultural systems*, 74(3):415–430, 2002.

[9] BA Keating, PS Carberry, GL Hammer, ME Probert, MJ Robertson, D. Holzworth, NI Huth, JNG Hargreaves, H. Meinke, Z. Hochman, et al. An overview of APSIM, a model designed for farming systems simulation. *European Journal of Agronomy*, 18(3-4):267–288, 2003.

[10] N.D. Le and J.V. Zidek. *Statistical analysis of environmental space-time processes*. Springer Verlag, 2006.

[11] S.G. Mallat. *A wavelet tour of signal processing: the sparse way*. Academic Pr, 2009.

[12] AB Potgieter, GL Hammer, and A. Doherty. Oz-wheat: a regional-scale crop yield simulation model for Australian wheat. *Queensland Department of Primary Industries & Fisheries, Information Series No. QI06033, Brisbane, Qld (ISSN 0727-6273)*, 2006.

[13] AB Potgieter, GL Hammer, A. Doherty, and P. De Voil. A simple regional-scale model for forecasting sorghum yield across North-Eastern Australia. *Agricultural and Forest Meteorology*, 132(1-2):143–153, 2005.

[14] B. Qian, R. De Jong, and S. Gameda. Multivariate analysis of water-related agroclimatic factors limiting spring wheat yields on the Canadian prairies. *European Journal of Agronomy*, 30(2):140–150, 2009.

[15] B. Qian, R. De Jong, R. Warren, A. Chipanshi, and H. Hill. Statistical spring wheat yield forecasting for the Canadian prairie provinces. *Agricultural and Forest Meteorology*, 149(6-7):1022–1031, 2009.

[16] JT Ritchie and DS NeSmith. Temperature and crop development. *Modeling plant and soil systems*, pages 5–29, 1991.

[17] P.D. Sampson and P. Guttorp. Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87(417):108–119, 1992.

[18] A. Schmitz and W.H. Furtan. *The Canadian Wheat Board: marketing in the new millennium*. Canadian Plains Research Center, 2000.

[19] R.C. Stone and H. Meinke. Operational seasonal forecasting of crop performance. *Philosophical Transactions B*, 360(1463):2109, 2005.

[20] W. Wu, C. Serele, J.V. Zidek, and N. Newlands. Agroclimatic wheat yield model for canadian prairies. 2009.

[21] A. Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P.K. Goel and A. Zellner, editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno do Dinetti*, pages 233–243. North-Holland, 1986.