The Design of Weather Index Insurance for Forage: The Case of Basis Risk for the Canadian Province of Ontario

by

Shuo Wang

A thesis submitted to
The Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements
of the degree of

Master of Science

I.H. Asper School of Business
Warren Centre for Actuarial Studies and Research
The University of Manitoba
Winnipeg, Manitoba, Canada
October 2015

The Design of Weather Index Insurance for Forage: The Case of Basis Risk for the Canadian Province of Ontario

Abstract

This thesis examines weather index area-yield basis risk for forage insurance. The first focus of the research is to determine which weather variables (e.g. rainfall, temperature, sunshine, etc.) should be included in the multivariable weather index, given the limited yield data and multicollinearity among weather variables. The second focus is to analyze the effect of the geographical scale (number of counties used in the index) on basis risk. Daily weather data and actual forage yield are from Ontario?s rainfall index-based forage insurance plan. Both principal component regression (PCR) and partial least squares regression (PLSR) are used to select the weather index variables. Results show that the two regression models generate similar weather variable selection for the weather index. Both models can be considered suitable depending on the choice of criteria. Further, the results show that as the number of counties of the index decreases, area-yield basis risk is reduced substantially.

Keywords: weather index insurance, multivariable weather index, rainfall, temperature, sunshine, basis risk, regression, forage, geographical scale, county, Canada

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my gratitude to my advisors, Professor Lysa Porth and Professor Milton Boyd, for their useful comments, remarks and engagement through the learning process of this master thesis. Furthermore, I would like to thank my committee, Professors Barry Coyle and Jeffrey Pai, for their insights, expertise as well for their support on the way. I also want to express my gratitude to Professor Kenseng Tan from the University of Waterloo for his valuable suggestions. I would like to thank my parents and all the people who have supported me along the way.

Chapter 1

Introduction

Background of Agricultural Sector

Expansion in agriculture is important for the increasing population worldwide. According to Food and Agricultural Organization of the United Nations (FAO), global food production needs to increase by 70% in order to feed the population by 2050. When farmers are exposed to risks without sufficient external risk protection or transferring mechanisms, they may use less inputs, such as fertilizer, resulting in lower production (Barret et al., 2007). Thus, the agricultural sector may have difficulty attaining its maximum capacity without a well functioning agricultural insurance system in place.

Introduction to Self-insurance

Some production risks are self-insurable through saving, improving irrigation systems, innovating farming technologies, sharing tenancy, etc. (Barnett and

Mahul, 2007). However, these approaches are not always cost-effective. For example, Rosenzweig and Binswanger (1993) show that he risk premiums of some farmers' self-risk-coping approaches are as high as 35% in rural India. Further, due to severely covariate and infrequent adverse weather events, some risks cannot be managed through risk retention. When households in an area all experience losses at the same time, traditional risk-coping approaches, like risk sharing among households, lose the ability to hedge farmers' losses. In many countries, including Canada and the US, transferring risk to the insurance market has been successful.

Farmers face significant risks caused by adverse weather conditions. Some research shows that weather can explain as much as 90% of the variability in crop yields (Climate Change Cell, 2009). Extreme weather conditions, such as droughts and floods, tend to occur over large areas and are hard to predict. Thus, weather risks may be difficult to manage without agricultural insurance (Giné et al., 2010).

1.1 Weather Index Forage Insurance

Introduction to Crop Insurance

Traditional indemnity-based insurance contracts are exposed to moral hazard. For instance, farmers may intentionally reduce the level of management without the concern of reduction in yield as a claim is paid once production level decreases. The problem of adverse selection is also of concern in traditional crop

insurance since mispricing may occur when lower risk farmers cross-subsidize higher risk farmers. Further, administration and claim verification costs add to the expense of delivering indemnity-based crop insurance, and this may impact the sustainability of the insurance program.

Introduction to Index-Based Crop Insurance

Index-based insurance (IBI) can be a feasible and cost-efficient tool for hedging weather risks and also serves as a promising alternative to traditional crop insurance. IBI avoids the critical issues of moral hazard and adverse selection since the indemnity depends on the performance of an underlying index (e.g., cumulative rainfall, maximum temperature, etc.), which is independent of the policyholders' behaviors and cannot be manipulated by the insured. Another advantage of IBI is that it does not require on-field inspection, which can be extremely costly. This is an particularly important feature for developing countries where there are often a large number of very small farms, and some crops may have yields difficult to track.

Problems of Index-Based Forage Insurance

A major difficulty with IBI, however, has been low participation rates. The participation rates of IBI contracts tends to be very low, particularly when compared to indemnity-based crop insurance. In Canada, for example forage plantation accounts for 44% of the the farming land, and in Ontario alone forage accounts for about 1,990,000 acres of the total seeded land (Ontario Ministry of Agriculture

Food and Rural Affairs, 2014). Despite forage representing the second largest crop in Canada, the uptake of forage insurance has generally been quite low. In particular, Ontario offers an index-based forage insurance program, only about 10% of the total forage acres are insured (Agricorp, 2014). Further research is needed to help address this low demand and support the success of IBI generally, and forage insurance specifically.

Potential Causes for the Low Demand

Mispricing is one of the possible reasons for the low level of demand for IBI. Some research has been carried out to examine at which premium rate farms may decide to participate in agricultural insurance (Cole et al., 2009; Giné and Yang, 2009). Giné et al. (2010) show that price and liquidity constraints all impede the demand of rainfall index insurance.

Another explanation for low demand of IBI is basis risk, which refers to when the index does not fully reflect the actual individual farmer loss in yield. Elabed and Carter (2015) show that IBI appears to be a compound lottery to purchasers, where the first lottery corresponds to whether an adverse event occurs, and the second lottery is whether the policy is triggered when a loss occurs. With basis risk, therefore, a farmer may suffer a loss on the farm, but not receive an indemnity payment. In this scenario, the farmer is worse off due to the premium paid for the insurance. Basis risk can also occur when a farmer receives an indemnity payment despite no actual loss on the farm, and this scenario is problematic for the insurer.

Thus, it is believed reducing basis risk is important for successfully implementing IBI. (Carter et al., 2014; Clarke, 2011; Elabed et al., 2013; Jensen et al., 2014; Skees, 2008). Research by Wakker et al. (1997) also shows that purchasers of insurance require a reduction in the premium of about 30% for each percent increase in basis risk.

Basis risk, in general, can be classified into three main categories including spatial, temporal, and variable basis risk. Spatial basis risk is caused by the mismatch between the insured farm and the location of the index(i.e. weather station). This may be particularly prevalent in developing countries, where there tends to be limited weather stations, and farm sizes tend to be smaller. Temporal basis risk occurs when the effects of weather variables over the growth cycle are not taken into account by the index. Variable basis risk occurs when wrong or not enough weather variables are selected for the index construction, and thus, the index and yield are not sufficiently correlated.

Basis Risk of the Current Rainfall Index in Ontario

Ontario Association of Cattleman(2014) criticizes the current forage index plan in Ontario for only relying on rainfall. However, insufficient yield data has been a tremendous issue for forage which constrains the ability to analyze the relationship between weather variables and yields, especially, with a large set of complex weather variables. Moreover, multicollinearity is also a big concern for such type of analysis. Thus, the construction of multivariate weather indexes has been

limited by these data constraints. Another controversial issue is that the all the counties in the province of Ontario are covered by the same rainfall index.

A second difficulty regarding the index-based forage insurance plan in Ontario is regarding the geographical scale of the index. Currently in Ontario, one index covers the entire province, yet, the province tends to experience much different climatic effects depending on the region. For example, some regions in Ontario may experience too much rain, while other regions experience too little. Elabed et al. (2013) show that the amount of the basis risk of an index depends heavily on the scale of the geographical area (e.g., number of counties) it covers. Therefore, it is important to examine the impact of geographical scale (number of counties covered by an index) on the basis risk of an index. The most accurate approach to build an index is to develop a unique index for each county. However, this approach can lead to considerable administration and underwriting costs. Therefore, it is interesting to consider the optimal geographical scale for an index is, which is one of the aims of this study.

Objectives

The first objective of the thesis is to determine the significant weather variables for forage in order to construct a multivariable indexes in the presence of limited yield data The multicollinearity of the explanatory variables are considered using Principal Component Regression(PCR) and Partial Least Squares Regression(PLSR). The second objective of the thesis is to examine the impact of ge-

ographical scale on the average area-yield basis risk of the multivariate weather index structure. **Methodology**

Part 1

Part one of the paper uses and compares PCR and PLSR models to overcome the problem of high dimension and multicollinearity of weather variables for variable selection. Their efficiencies are compared with respect to two criteria-dimensions needed to achieve the lowest prediction error and the magnitude of the smallest prediction error.

Part 2

Part two examines the effect of geographical scale on variable basis risk with the variable selection results from PCR and PLSR are both used. Indexes are constructed for different number of counties according to a multivariate weather index structure defined in this paper. A procedure called pairwise correlation maximization is applied to optimize the weights of weather variables in the multivariable index so that the correlations between the index and yields are maximized but not too biased towards a specific county.

The remainder of this thesis is organized as follows. First, the paper provides a detailed introduction and explanation of two regression models considered for weather variable selection. Then, important procedures for the construction of a multivariate weather index structure are proposed. Following this, based on the data collected from Agricorp, the application results of these models are analyzed.

Data processing is discussed first, which involves data choices and data stationarity concerns. Next, results of variable selection based on both dimension reduction models are compared regarding their discrepancies and efficiencies. After significant variables are determined for each county in sample, with a simple variable filtering principle, the multivariate weather index is constructed based on the correlation optimization at various geographical coverage levels. Then, changes in variable basis risk are measured as the geographical scale of the index varies. Finally, the multivariate weather index developed in this paper is compared to the index-based forage insurance plan offered in Ontario, Canada. Additionally, with some numerical results, the importance of county clustering is discussed. Based on the results of the multivariate weather index, some insights are provided for both commercial as well as governmental institutions for the design of future IBI schemes.

1.2 Background

Background and Issues with the Index-Based Forage Insurance Plan in Ontario

Ontario initiated the first forage insurance plan, called simulated forage yield (SIMFOY) in 1981. The plan modelled forage yield with a combination of various weather variables, including temperature, rainfall, soil type, etc. However, producers found the contract structure was too complex, and the claim computation was difficult to understand. This caused a tremendous barrier for the up-scaling

of SIMFOY, and the plan was terminated in 2004.

The current rainfall index-based forage insurance plan (RIFI) was launched as a pilot insurance program in 2000 to 2002, and in 2005, the plan was officially introduced as a province-wide plan. Under this plan, farmers choose to insure forage crops against the events of insufficient rainfall, excess rainfall, or both. In general, producers found this plan easier to understand compared to SIMFOY. However, one drawback is that claims only depend on cumulative rainfall, rather than the multiple weather variables adopted by SIMFOY. It is unknown what impact this may have on basis risk and demand. For example, for a location with sufficient rainfall for most of the year, temperature or sunshine may also be important in explaining yield. Intuitively, it seems that a single variable may have a limited ability to explain loss in yield. Therefore, this paper addresses variable basis risk by considering a multivariate weather index design to incorporate the effects of different weather variables into the structure of the index.

Further, dominant weather variables may vary from an area to another. Therefore, this is an important consideration in constructing the index across different geographical combinations (i.e. areas formed by different mixes of counties). It is important to understand the definition of "area" in this study. In this paper, a geographical area is not limited to the continuity on a map, rather an area could be a specific county, two counties or more that do not share borders, yet, share similar weather characteristics. This is because counties do not have to be adjacent to each

other to experience similar meteorological and spatial conditions. Therefore, areas refer to combinations of counties that do not necessarily have shared borders. In fact, in this paper, indexes are constructed for all the possible combinations of the sample counties. The correlation results can be an indicator of the similarity among the sampled counties.

Limited Yield Data

A majority of forage production is used for feeding cattle with a minority of various end-uses where forage sometimes is traded in the market. Moreover, Ontario has been offering IBI since 2000. Thus, tracking forage yield is difficult because yield inspection is avoided by implementing IBI, which is a main advantage over products that require loss adjusting. In most provinces in Canada, the scarcity of forage yield data is also a big challenge. This creates difficulty in the analysis of the relationship between weather and yield, particularly, for regression type methods. As a result, it becomes important to efficiently utilize existing yield data, and supplement it with rich daily weather data to carry out the analysis of the relationship between the weather and yield (Zhu, 2015). In this paper, this issue is addressed with two special regression approaches, as both models are good at extracting the most important information with a limited sample size through their abilities of dimension reduction. The weather selection results are compared to suggest the most suitable model for the task of variable selection under the scarcity of yield data.

Relationship Between Weather and Yield

The rainfall index-based forage insurance plan assumes a linear relationship between cumulative rainfall and yield. This assumption has significant implications. On one hand, a non-linear relationship between the index and yield can cause more confusion for farmers regarding contract structure. However, on the other hand, a non-linear relationship may be more representative, thus, further reducing the basis risk. In this paper, the proposed index structure optimizes the weights of selected weather variables based on the maximization of pearson's correlation. However, future analysis could consider a more generalized approach based on a non-linear relationship between the index and yield, and correlation structures such as Kendall's taw and Spearman's rho can be easily generalized to the procedures introduced in this paper.

Index-Based Forage Insurance Plan in Ontario

The rainfall index-based forage insurance plan consists of four basic plans including Base, Three-month, Bi-monthly, and Monthly. Under most situations, forage is cut twice or three times a year. In this paper, the analysis is restricted to first cut yield data only, given the data limitations.

With more detailed data that is broken down by each cut over the entire growing period, the proposed index structure in this paper could more accurately be compared to Ontario's existing rainfall index-based forage insurance plans. Therefore, this paper focuses on the Base plan for comparison, given that it is the

Plan	Index
Actual Base Plan	Cumulative rainfall from May to August of the policy year divided by the county's long term cumulative rainfall from May to August
Base Plan due to data limitations	Cumulative rainfall from May to June of the policy year divided by the county's long term cumulative rainfall from May to June

Table 1.1: Proxy Plan Information

most popular and accounts for 49% of all the policies sold since 2006. The Base plan for the purpose of comparison in this paper is defined in the Table 1.1.

From Table 1.1, it can be seen that the rainfall index in the proxy plan is defined as the cumulative rainfalls in the May and June. The main reason to define the proxy index in this way for the Base plan is that the Base plan assumes forage growth starts in May and harvests of the first cut of forage are finished by July 1^{st} according to Agricorp.

1.3 Literature

1.3.1 Spatial Basis Risk

Spatial basis risk is one of the major challenges in successfully implementing IBI. Chen and Liu (2012) utilize inverse distance weighting (IDW) to interpolate weather data at 12 weather stations in Taiwan. They also determine the optimal

research radius for IDW around an ungauged area in their research. The paper shows that the correlation coefficients between interpolated weather data are 0.95 at all 12 weather stations. This demonstrates the potential of IDW for estimating weather data at places without an allocated weather station to reduce spatial basis risk. Although the correlation is high in this study, it is generally agreed that IDW cannot handle the scenario when there are sparse weather stations and the geographical arrangement is relatively complicated (e.g. mountainous areas) (Dirks et al., 1998; Mair and Fares, 2010). For example, in mountainous areas where microclimates exist it is realistic to consider the distance between two locations only. Therefore, alternative approaches are necessary considering spatial correlation between different locations.

Cao et al. (2014) use data from a large area in China to test two forecasting methods with different spatial interpolation approaches. They compute interpolated weather data using a kriging model and forecasted weather data with time series analysis. A dynamic semiparametric factor model (DSFM), which considers both major spatial and temporal factors of spatial-temporal weather data, is applied to weather data for both interpolation and forecasting. The article shows that the kriging model produces smaller absolute error than DSFM.

Ritter et al. (2012) suggest a multi-site rainfall model approach (MRM) where a payoff is generated according to the weighted payoff of a portfolio consisting of weather derivatives from different locations around a production field. The

multi-site rainfall model is calibrated according to the historical daily rainfall data measured in 49 weather stations in Germany. The model is then used to simulate future rainfall predictions that are utilized for computing weights assigned to weather derivatives. They compare multi-site rainfall model with simpler approaches including IDW and historical simulation in which future weights of payoffs is assumed to be same as historical weights. The result shows that the MRM in general performs better than other two approaches, reducing basis risk by 20% to 40%.

1.3.2 Variable Basis Risk

Contract design is a crucial component of the success of weather IBI. Elabed et al. (2013) proposed a multi-scale yield index-based contract to help reduce basis risk and effectively control moral hazard, through employing double indexes at different geographical scales. They show that with the addition of the second index, IBI gains a 40% increase in demand using simulation of data of Malian cotton farmers. Weber et al. (2014) design IBI contracts using data from central Asia based on indexes aggregated at different levels (i.e. county level, farm level). They conclude that as risk aggregation level increases, the hedging effectiveness of weather index insurance increases.

While improvements in contract design can help to reduce basis risk, developing alternative indexes based on various weather variables can also play a ma-

jor role in basis risk reduction. (Deng et al., 2008) examine two possible indexes to predict yields, including a cooling degree-day production model (CDD) using historical data, and a CEREZ-Maize model under the Decision Support System for Agrotechonology Transfer (DSSAT). Bobojonov and Sommer (2011) test satellite-based farm-level and area-level Normalized Difference Vegetation Index (FNDVI/ANDVI) as potential candidates for yield estimation. They show that FNDVI has a high correlation with yield data, and ANDVI only shows moderate correlation with yield data., which implies ANDVI has more basis risk than FNDVI. Zhu (2015) combined principal component analysis and screening regression (PCASR) to explore the relationship between yield data and more than 160 weather predictors according to past municipal yield and weather data in Manitoba. The research also compares alternative forms of PCASR and pure screening regression. it is agreed that PCASR provides superior out-of-sample prediction and the most dominant weather variables are also identified for different municipalities.

1.3.3 Other Solutions for Low Participation

While basis risk may be a major challenge in IBI, some research has focused on other sources of low demand. For example, Public-Private Partnership (PPP) approaches may be an important framework for agricultural index insurance (Carter et al., 2014). Under a PPP frameworks, risks are classified into three categories, in-

cluding a retention risk layer, a commercial risk layer, and a catastrophic risk layer. By efficiently subsidizing or insuring catastrophic risk layer, where most ambiguity and uncertainty of loss exists, amount of risk loading can be reduced while guaranteeing a minimum market volume for insurers. In addition, government could enhance demand for IBI products in several ways. Boyd et al. (2011) test the significance of eight variables in terms of determining the demand for IBI using probit regression model. They show that knowledge of crop insurance, trust of the crop insurance company, importance of low crop insurance premium, and government as the main information source for crop insurance are relatively important factors in this study. Therefore, government intervention and regulation are extremely important for elevating the uptake of crop IBI. Governments play a crucial role in distributing essential knowledge related to IBI and sharing risk and costs of crop insurance to ensure an affordable premium. Moreover, governments should regulate crop insurance market to enhance the trust that farmers grant to insurers and, thus, crop insurance products. Similarly, Lin et al. (2015) use probit model to examine the significance of 15 weather variables affecting the willingness to buy weather index insurance in Hainan province of China. They find that farmers prefer a private company over the government for the delivery of weather index insurance.

Chapter 2

Data and Methodology

The study of forage yield is faced with a major challenge of limited yield observations over time. Comparatively, weather data is typically plentiful, and often available in daily increments. while farm-level forage yield data was available, the resulting time series at the farm-level was too short to conduct a meaningful analysis. Therefore, the farm-level forage yield data was aggregated to the municipality level to produce a longer time series. Thus, in this study, 24 years of forage yield data is utilized at the county-level.

With traditional regression methods, this sample size is insufficient for studying the impact of many complex weather effects on forage yield. Specifically, if the degree of freedom is not sufficiently high in regression-type models, the results of variable selection varies significantly from sample to sample. Moreover, if the number of covariates is higher than the sample size, the system of functions

of the regression model cannot be solved. Another issue with regression analysis is that researchers often face the tradeoff between avoiding multicollinearity and losing useful information.

Another important consideration is deciding the geographical scale of the index. For example, as the number of counties covered by an index increases, it can have significant impact on basis risk. In Ontario, the current rainfall forage index-based insurance plan uses each index to cover the entire province. However, research is needed to examine the optimal geographical scale (number of counties covered by the index).

Part one of the methodology section focuses on the first objective of this thesis, which is variable selection. To select the significant weather variables for constructing weather indexes, two regression-type models are applied to help overcome some of the challenges associated with the limited yield data.

Part two of the methodology section focuses on the second objective, which examins the impact of geographical scale on the county-level average basis risk. The influence of geographical scale is examined by contrasting the correlation results of multivariable weather indexes optimized for different groupings of counties according to a pairwise correlation maximization procedure and the variable selection results achieved in the first objective.

In this chapter, datasets and corresponding weather variables are introduced in Section 2.1. In Sections 2.2.1 and 2.2.2, principal component regression (PCR)

and partial least squares regression (PLSR) models are discussed in detail. Following these two sections, important concepts and methods are introduced, including leave-one-out cross-validation and bootstrapping for variable selection in sections 2.2.3 and 2.2.4. In the last two sections, a multivariable weather index is defined and the pairwise correlation optimization procedure is introduced to maximize the correlation between the multivariable weather index and the yields of the corresponding counties.

2.1 Data

Forage yield data from the simulated forage yield (SIMFOY) plan over the period of 1981 to 2004 for seven counties, including Algoma, Cochrane, Leeds Grenville, Niagara (NG), Nipissing (NP), Rainy River (RR), and Thunder Bay (TB) are collected. Yield data comprises the first cut of forage only, however, the first cut yield of forage accounts for approximately 55% to 65% of the total annual yield. The methodology introduced in this paper can easily be generalized to the second cut and third cut, as well as the total annual yield.The seven locations considered cover the entire province of Ontario. However, since sufficient farm-level forage yield data was not available, data was aggregated at the county level. As a results, the basis risk analyzed in this paper correspond to the average county area-yield rather than individual farmer basis risk.

Climate data from 1981 to 2004 is obtained from SIMFOY dataset, which in-

cludes daily precipitation, daily maximum and minimum temperatures, and daily hours of sunshine. Further, the accuracy of the data is verified by professional third parties and compared to the records of Environment Canada. The use of this rich dataset of weather data helps to overcome the problem of missing weather data which is normally a major limitation in crop insurance studies.

2.1.1 Design Matrix of Weather Variables

Research regarding the relationship between crop yield and weather typically uses common indexes such as cooling degree day (CDD) and heating degree day (CDD), and growing degree day (GDD). These three indexes only consider the heat accumulated during the growing period of forage. The index currently utilized by Ontario for the index-based forage insurance plan only considers rainfall. However, Ontario Cattlemen's Association (2013) has suggested that the current index implemented in Ontario does not contain the complexity of some important weather variables, such as soil moisture. Therefore, a multivariable weather index structure that can incorporate the effects of multiple weather variables is proposed in this paper. In order to construct the multivariable weather index, a large set of weather variables must be analyzed simultaneously. Thus, to construct the design matrix, three types of weather variables are considered, including rainfall, temperature, and hours of sunshine, as well as a large number of synthetic variables that are based on these three basic types of variables to try to replicate

the soil moisture.

In April, May, and June, during a normal year with relatively normal amount of rainfall and moderate temperature, soil has the best capability to preserve moisture. It is one of the main reasons why first cut yield accounts for more than 55% of the total yield. Thus, because soil moisture data is not available, we decided to incorporate the ratio between rainfall and temperature, as well as the ratio between rainfall and hours of sunshine, in the model as indicators of soil moisture. For most variables, not only are the monthly quantities considered, but also the total quantities over the period from April 15th to June 30th. For example, discussions with forage specialists from Ontario indicated that to achieve a satisfying level of first cut yield, both monthly cumulative rainfall and seasonal cumulative rainfall must reach satisfactory levels. Thus, while analyzing weather effects, weather variables are analyzed at both individual monthly and aggregated levels. Moreover, the period from April 15th to June 30th is chosen as the sample period for analysis because it covers the growing period related to the first cut of forage for most farmers in Ontario. A summary of the weather variables defined in this study are outlined in Tables 2.1 and 2.2.

Given that more than 50 variables are defined, the screening procedure in Zhu (2015) is applied to filter out variables that have low correlations with the yield data. These variables are considered unlikely to provide valuable information.

Table 2.1: Single Type Weather Variables

Rainfall Variables	**Temperature Variable**	**Sunshine**
Cumulative rainfall in April	Minimum of minimum temperature in April	Average hours of sunshine in April
Cumulative rainfall in May	Minimum of minimum temperature in May	Average hours of sunshine in May
Cumulative rainfall in June	Minimum of minimum temperature in June	Average hours of sunshine in June
Seasonal cumulative rainfall	Average of maximum temperature in April	
Number of days having rainfall in April	Average of maximum temperature in May	
Number of days having rainfall in May	Average of maximum temperature in June	
Number of days having rainfall in June	Average of minimum temperature in April	
Number of days having rainfall through the season	Average of minimum temperature in May	
	Average of minimum temperature in June	

2.2 Part 1: Weather Variable Selection

The methodologies for the objective of variable selection under limited yield data and multicollinearity are introduced and explained in this section. Princi-

Table 2.2: Synthetic Soil Moisture Variables

Rainfall to Temperature Ratio	**Rainfall to Sunshine Ratio**
Maximum rainfall to temperature ratio in April	Cumulative rainfall to sunshine ratio in April
Maximum rainfall to temperature ratio in May	Cumulative rainfall to sunshine ratio in May
Maximum rainfall to temperature ratio in June	Cumulative rainfall to sunshine ratio in June
Cumulative rainfall to temperature ratio in April	Seasonal cumulative rainfall to sunshine ratio
Cumulative rainfall to temperature ratio in May	
Cumulative rainfall to temperature ratio in June	
Seasonal cumulative rainfall to temperature ratio	

pal component regression (PCR) and partial least squares regression (PLSR) are introduced for the purpose of variable selection. Non-linear iterative partial least algorithms for solving PCR and PLSR models are explained using power methods.

Leave-one-out cross-validation is applied to determine the dimension (number of projected weather variables) needed to generate the lowest prediction error to reduce the dimension of the original design matrix. To determine the significant weather variables, the bootstrapping is used for both PCR and PLSR models to construct confidence intervals for the weather variables.

2.2.1 Principal Component Regression

PCR is a popular regression model that has been used in various areas of studies. To deal with the problem of multicollinearity, the approach project the original variables onto a new coordinate space where the resulted new variables, which are called principal components (PC), not only contain the variation in the original variables, but are orthogonal to each other as well. However, PCs have different variances that can be interpreted as the information from the original variables they explain. Intuitively, the PCs with large variation are kept as they represent a large amount of information from the original system. The ones with small variation can only explain noise. Therefore, some components can be eliminated to realize dimension reduction. To do so, it requires the directions with the largest variances and covariances.

In order to perform PCR, the matrix of weather variables, which is call design matrix, is standardized. Thus, weather variables are subtracted by their means and divided by their standard deviation. Standardization is a common step in statistical regression models, as it removes the effect of varying scales and units among variables that can impact the allocation of variables to each PC. Thus, the standardized design matrix are always used in PCR. To understand how to find the directions of the projection, the next section discusses the concepts of eigenspace.

Eigenspace and Singular Value Decompostion

X is a n by p standardized design matrix with a p by p covariance matrix Σ, where n is the number of objects and p is the number of variables. The objective is to find a p by 1 dimensional vector $\vec{a}$ such that the projection, $X\vec{a}$, has the biggest covariance.

$$\max_{\vec{a}\in\mathbb{R}^p} \quad cov(X\vec{a}) = \max_{\vec{a}\in\mathbb{R}^p} \quad \vec{a}^T\Sigma\vec{a} \tag{2.1}$$

In fact, it is easy to prove that if Σ is positive definite with eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p > 0$, and corresponding eigenvectors $\vec{e}_1, \vec{e}_2, ..., \vec{e}_p$, where $\|\vec{e}_i\| = 1$ for $i = 1, 2, ..., p$, following results are true,

$$\max_{\vec{a}\in\mathbb{R}^p} \quad \vec{a}^T\Sigma\vec{a} = \lambda_1 \quad when \quad \vec{a} = \vec{e}_1 \tag{2.2}$$

$$\min_{\vec{a}\in\mathbb{R}^p} \quad \vec{a}^T\Sigma\vec{a} = \lambda_p \quad when \quad \vec{a} = \vec{e}_p \tag{2.3}$$

$$\max_{\vec{a}\perp\vec{e}_1,\vec{e}_2,...\vec{e}_k} \quad \vec{a}^T\Sigma\vec{a} = \lambda_{k+1} \quad when \quad \vec{a} = \vec{e}_{k+1} \tag{2.4}$$

Thus, the first result shows the direction explaining the most of the variation in the original variables is the eigenvector with the biggest eigenvalue. The amount of variation projected onto the first eigenvector is equal to the magnitude of the first eigenvalue. The third result demonstrates the $(k+1)$th largest variance ex-

plained is in the direction of the $(k+1)$th eigenvector with the magnitude of the $(k+1)$th eigenvalue. One important property of eigenvectors is that a set of orthogonal eigenvectors can always be found for a symmetric matrix. Therefore, it can be checked that projections of X onto two different eigenvectors are orthogonal to each other. Therefore, using eigenvectors, a new coordinate system can be formed such that the original design matrix can be transformed into an orthogonal design matrix. This can be represented in the following form,

$$X = TP^T + E \tag{2.5}$$

where P is a p by k matrix consisting of eigenvectors as columns; k is the number of the dominant eigenvectors; T is a n by k matrix, called score matrix; and E is the residual.

The eigenvectors with small variances can be ignored as they only explain noise. More importantly, if there is relatively strong multicollinearity in the design matrix, the information contained in the variables can be comprised by the directions of the first few eigenvectors. In general, the eigenvectors and eigenvalues are computed by using singular value decomposition (SVD).

$$X = USV^T \tag{2.6}$$

where U is the matrix of the left eigenvectors of X from XX^T; V is the the matrix

of the right eigenvectors of X from X^TX; S is the matrix with diagonals being the non-zero singular values of X.

It is known that $cov(X) = \Sigma = \frac{X^TX}{n}$ since the design matrix is standardized. Thus, the eigenvectors of Σ is the same as eigenvectors of X^TX. From Equation 2.6, by comparison, it is clear that the eigenvectors of Σ are equal to the columns of V. It is also easy to see that the columns of the score matrix T are equal to the eigenvectors in U multiplied by the corresponding eigenvalues. The eigenvalues and eigenvectors can be solved using system of equations, however, the process is computationally expensive when large matrices are involved in the calculation. For PCR, different algorithms exits for calculating T and P in Equation 2.5 by avoiding large matrix operations. One of the most popular algorithms is called non-linear iterative least squares (NIPALS). In order to understand the algorithm, power method for solving eigenproblems is introduced in the next section.

Power Method

The power method is one of the common algorithms for solving eigen problems. It is used to attain the dominant eigenvector and eigenvalue of a matrix. Assume A is a n by n square matrix with eigenvalues, $|\lambda_1| \geq |\lambda_2| \geq ... \geq |\lambda_n|$ and corresponding eigenvectors, $\vec{v}_1, \vec{v}_2, ..., \vec{v}_n$, respectively. The algorithm is initiated by starting at $\vec{b}_0 \in \mathbb{R}^n$, and it is known that

$$\vec{b}_0 = \beta_1\vec{v}_1 + \beta_2\vec{v}_2 + \beta_3\vec{v}_3 + ... + \beta_n\vec{v}_n \tag{2.7}$$

where $\forall\beta_i \in \mathbb{R}$.

If $\vec{b}_0$ is multiplied by matrix A, repetitively,

$$\begin{aligned} A\vec{b}_0 &= \beta_1\lambda_1\vec{v}_1 + \beta_2\lambda_2\vec{v}_2 + \beta_3\lambda_3\vec{v}_3 + ... + \beta_n\lambda_n\vec{v}_n \\ &= \lambda_1(\beta_1\vec{v}_1 + \beta_2\frac{\lambda_2}{\lambda_1}\vec{v}_2 + \beta_3\frac{\lambda_3}{\lambda_1}\vec{v}_3 + .. + \beta_n\frac{\lambda_n}{\lambda_1}\vec{v}_n) \end{aligned} \tag{2.8}$$

$$\begin{aligned} \vec{b}_2 = A^{(2)}\vec{b}_0 &= \beta_1\lambda_1^2\vec{v}_1 + \beta_2\lambda_2^2\vec{v}_2 + \beta_3\lambda_3^2\vec{v}_3 + ... + \beta_n\lambda_n^2\vec{v}_n \\ &= \lambda_1^2(\beta_1\vec{v}_1 + \beta_2(\frac{\lambda_2}{\lambda_1})^2\vec{v}_2 + \beta_3(\frac{\lambda_3}{\lambda_1})^2\vec{v}_3 + .. + \beta_n(\frac{\lambda_n}{\lambda_1})^2\vec{v}_n) \end{aligned} \tag{2.9}$$

In general, it is true that

$$\begin{aligned} \vec{b}_k = A^{(k)}\vec{b}_0 &= \beta_1\lambda_1^k\vec{v}_1 + \beta_2\lambda_2^k\vec{v}_2 + \beta_3\lambda_3^k\vec{v}_3 + ... + \beta_n\lambda_n^k\vec{v}_n \\ &= \lambda_1^k(\beta_1\vec{v}_1 + \beta_2(\frac{\lambda_2}{\lambda_1})^k\vec{v}_2 + \beta_3(\frac{\lambda_3}{\lambda_1})^k\vec{v}_3 + .. + \beta_n(\frac{\lambda_n}{\lambda_1})^k\vec{v}_n) \end{aligned} \tag{2.10}$$

Obviously, $\lim_{k\to\infty}\frac{\vec{b}_k}{\lambda_1^k} = \beta_1\vec{v}_1$ for $\beta_1 \neq 0$ since $\frac{\lambda_i}{\lambda_1} < 1$ for $i = 2, 3, ...n$. Thus, the process converges to the dominant eigenvector of matrix A because a multiple of an eigenvector is still an eigenvector of A. However, problems of overflow and

underflow are of concern since the limit goes to ∞ or $-\infty$ when $|\lambda_1| > 1$ and 0 when $0 < |\lambda_1| < 1$. To overcome the issue, after each multiplication by A, the $A^{(k)}\vec{b}_0$ is scaled by dividing it by its own length. To find the corresponding eigenvalue, the Rayleigh quotient is used. The NIPALS algorithm generalize the power method to non-symmetric matrices and SVD. In the next section, NIPALS algorithm is explained in detail.

NIPALS Algorithm for PCR

NIPALS algorithm is one of the most popular methods for computing eigenvectors, designed specifically for principal component analysis at first. However, it was further developed by Wold for partial least squares regression. First, some notations are defined and reviewed. X is the standardized n by p design matrix; t is defined as the score vector of a PC in an iteration; p is the p by 1 loading of a PC in an iteration; E_i is the residual of X after $(i-1)$ iterations; and at the beginning of each iteration, let t be a column of the residual as a starting point for a iteration. At the first iteration, let $E_0 = X$, and $t =$a column of E_0

- Step 1: Compute $p = \frac{E_{i-1}^T t}{t^T t}$
- Step 2: Normalize $p = \frac{p}{\sqrt{p^T p}}$
- Step 3: $t = E_{i-1} p$

- Step 4: Check the convergence of t and t_{old},$\|t - t_{old}\| < \epsilon$, where ϵ is the criterion of convergence. If convergence is achieved, then t is put into T as a column, and p into P as a column, then go to Step 5. Otherwise, set $t = t_{old}$ and return to Step 1.

- Step 5: Deflate E_{i-1}, $E_i = E_{i-1} - tp^T$

The process stops after attaining the number of PCs needed.

Next, the generalized power method (GPM) is introduced first. Then, the interpretation of NIPALS algorithm is outlined in five steps.

For a n by m matrix M, the relationship between the ith left eigenvector, u_i, right eigenvector, v_i, and ordered singular value, σ_i, which are ordered according to the magnitude of their absolute values, $i = 1, 2..., q$ with q being the numbers of non-zero singular values, is given by,

$$M^T u_i = \sigma_i v_i \tag{2.11}$$

$$Mv_i = \sigma_i u_i \tag{2.12}$$

Let $_0 be a linear combination of all the left eigenvectors with non-zero singular values.$

$$\vec{z}_0 = \alpha_1 \vec{u}_1 + \alpha_2 \vec{u}_2 + \alpha_3 \vec{u}_3 + ... + \alpha_m \vec{u}_q \tag{2.13}$$

where $\vec{z}_0 \in \mathbb{R}^n$

Then,

$$\begin{aligned} M^T\vec{z}_0 &= \alpha_1\sigma_1\vec{v}_1 + \alpha_2\sigma_2\vec{v}_2 + \alpha_3\sigma_3\vec{v}_3 + ... + \alpha_n\sigma_q\vec{v}_q \\ &= \sigma_1(\alpha_1\vec{v}_1 + \alpha_2\frac{\sigma_2}{\alpha_1}\vec{v}_2 + \alpha_3\frac{\sigma_3}{\sigma_1}\vec{v}_3 + .. + \alpha_m\frac{\sigma_q}{\sigma_1}\vec{v}_q) \end{aligned} \tag{2.14}$$

$$\begin{aligned} MM^T\vec{z}_0 &= \alpha_1\sigma_1^2\vec{u}_1 + \alpha_2\sigma_2^2\vec{u}_2 + \alpha_3\sigma_3^2\vec{u}_3 + ... + \alpha_q\sigma_q^2\vec{u}_q \\ &= \sigma_1^2(\sigma_1\vec{u}_1 + \alpha_2(\frac{\sigma_2}{\sigma_1})^2\vec{u}_2 + \alpha_3(\frac{\sigma_3}{\sigma_1})^2\vec{u}_3 + .. + \alpha_q(\frac{\sigma_q}{\sigma_1})^2\vec{u}_q) \end{aligned} \tag{2.15}$$

$$\begin{aligned} M^{(k)}\vec{z}_0 = M^TM...MM^T\vec{z}_0 &= \alpha_1\sigma_1^k\vec{v}_1 + \alpha_2\sigma_2^k\vec{v}_2 + \alpha_3\sigma_3^k\vec{v}_3 + ... + \alpha_q\sigma_q^k\vec{v}_q \\ &= \sigma_1^k(\alpha_1\vec{v}_1 + \alpha_2(\frac{\sigma_2}{\sigma_1})^k\vec{v}_2 + \alpha_3(\frac{\sigma_3}{\sigma_1})^k\vec{v}_3 + .. + \alpha_q(\frac{\sigma_q}{\sigma_1})^k\vec{v}_q) \end{aligned} \tag{2.16}$$

$$\begin{aligned} M^{(k+1)}\vec{z}_0 &= MM^T...MM^T\vec{z}_0 \\ &= \alpha_1\sigma_1^{k+1}\vec{u}_1 + \alpha_2\sigma_2^{k+1}\vec{u}_2 + \alpha_3\sigma_3^{k+1}\vec{u}_3 + ... + \alpha_m\sigma_m^{k+1}\vec{u}_m \\ &= \sigma_1^{k+1}(\alpha_1\vec{u}_1 + \alpha_2(\frac{\sigma_2}{\sigma_1})^{(k+1)}\vec{u}_2 + \alpha_3(\frac{\sigma_3}{\sigma_1})^{(k+1)}\vec{u}_3 + .. + \alpha_q(\frac{\sigma_q}{\sigma_1})^{(k+1)}\vec{u}_q) \end{aligned} \tag{2.17}$$

where k is an odd number.

As k converges to infinity, Equation 2.16 converges to a multiple of the dominant right eigenvector of M. Equation 2.17 converges to a multiple of the dominant left eigenvector of M.

- Step 1: Starting at a column of E_{i-1} as it is a linear combination of left eigenvectors of E_{i-1}. Then, generalized power method is used to calculate the left and right eigenvectors of E_{i-1}. The first step is same as what is done in Equation 2.16.

- Step 2: This step is, first, applied to avoid overflow and underflow problems and, second, to standardize the eigenvector from Step 1 for recording.

- Step 3: In this step, Equation 2.17 is applied. The resulted vector converges to the multiple of dominant left eigenvector with the corresponding eigenvalue.

- Step 4: This step is self-explanatory.

- Step 5: The deflation is according the spectral decomposition of non-symmetric matrix,

$$M = \sum_{\forall j} \sigma_j u_j v_j^T \tag{2.18}$$

  Thus, with the deflation procedure in Step 5, the matrix is reduced such that the dominant eigenvector is removed, and then, the next dominant eigenvector can be computed following the same procedures.

To link the response variable with the original variables for the purpose of variable selection, it can be shown the regression coefficient matrix of original variables is

$$Y = TB + \epsilon_{pcr} = XPB + \epsilon_{pcr} \tag{2.19}$$

where B is the regression coefficients of T; and ϵ_{per} is the error term of the regression.

$$B_{pcr} = PB \tag{2.20}$$

Thus,

$$B_{pcr} = P(P^T X^T XP)^{-1} P^T X^T Y \tag{2.21}$$

PCR solely considers the covariances of the explanatory variables during the transformation and does not rely on the covariance between the design matrix and the response variable. The PLSR model introduced in the next section not only considers covariances in the design matrix, but also the covariance between explanatory variables and the response variables.

2.2.2 Partial Least Squares Regression

PLSR was introduced as an econometric analysis tool by Wold in the 1960s. Similar to PCR, PLSR relies on a set of new variables that are called latent vari-

ables (LV) selected to be orthogonal to each other. However, LVs are different from PCs in the way that the computation of LVs takes into the account the covariance between the design matrix and the response variables. Due to this feature of PLSR, it is referred to as a supervised model. Therefore, the application of PLSR seems to be more reasonable since the main focus of the paper is to analyze the relationship between the weather variables and the response variable. Normally, one would expect PLSR to produce a similar level of in-sample and out-of-sample error as PCR, however with less LVs. Therefore, it is possible to save some degree of freedom to stabilize the result of the regression. Next, the important notations for PLSR model are defined. X is standardized n by p design matrix; Y is a n by q matrix of response variables; E and F are corresponding residual matrices of X and Y; columns of W are weight vectors that are computed to maximize the correlation between score vectors with the response variable; B is a diagonal matrix of the regression coefficients of scores on columns of U, respectively; and L is a residual matrix; t and u are the scores of LVs in each iteration for X and Y, respectively.

$$X = TP^T + E \tag{2.22}$$

$$Y = UQ^T + F \tag{2.23}$$

$$T = XW \tag{2.24}$$

$$U = TB + L \tag{2.25}$$

$$T = X - scores \quad P = X - loadings \tag{2.26}$$

$$U = Y - scores \quad Q = Y - loadings \tag{2.27}$$

Variabilities of X and Y are described by their corresponding LVs. PLSR seeks to find T that maximizes the correlation between T and U. Each column, t and u, of scores, T and U, are linear combinations of the residual information of X and Y respectively. Thus, for example, the first LVs are $t = Xw$ and $u = Yc$, where w and c are called loading weights. Therefore, t and u are computed so that

$$|cov(t,u)| = |cov(Xw,Yc)| = \max_{|a|=|b|=1} |cov(Xa,Yb)| \tag{2.28}$$

It is important to realize that the purpose is to extract T and utilize it to predict Y.

NIPALS Algorithm for PLSR

There are in general two types of PLSR models, PLS1 and PLS2. PLS1 means that there is only one response variable as the target of the analysis. PLS2 represents a multiple multivariate case of the regression model. In this paper, the only response variable is the first cut yield of forage, but PLS2 model is considered since PLS1 is just a special case of PLS2. Before getting into the algorithm, notations are clarified and also assume X is a centred and scaled design matrix.There are different algorithms to compute LVs. However, all of them are iterative, and the difference lies in whether variables are normalized. Nonlinear iterative paritial least squares(NIPALS) approach is one of the most popular algorithms.

The following gives a concise overview of NIPALS algorithm:

Set u =a column of F_i, $E_0 = X$, and $F_0 = Y$.

- Step 1: $w = E_{i-1}^T u$
- Step 2: $w = \frac{w}{||w||}$
- Step 3: $t = E_{i-1} w$
- Step 4: $c = \frac{F_{i-1}^T t}{t^T t}$
- Step 5: $c = \frac{c}{\|c\|}$
- Step 6: $u = F_{i-1} c$

- Step 7: Check convergence of t. If convergence is achieved, continue with Step 8. Record t into T, and c into C. Otherwise, go back to Step 1.
- Step 8: $p = \frac{E_{i-1}^T t}{t^T t}$ Record p into P
- Step 9: $b = (t^T t)^{-1} t^T u$.
- Step 10: Deflate X and Y into $E_i = E_{i-1} - tp^T$ and $F_i = F_{i-1} - tc^T$, then use E_i and F_i to go through all the steps until X the number of LVs required is achieved.

To understand NIPALS algorithm, it is first explained from the perspective of regression. Then, it is shown how the algorithm is related to eigenproblems.

- Step 1: The first step chooses the weight according to the covariances of variables with the response variable.
- Step 2: This step normalizes the weight vector.
- Step 3: With the weights, a score vector is calculated.
- Step 4: With the score vector, response variable is regressed on the score to calculate coefficient c. This can be derived from Equation 2.23 and 2.25 by

$$\begin{aligned} Y &= TBQ^T + LQ^T + F \\ &= TC^T + F^* \end{aligned} \tag{2.29}$$

Thus, C is just the coefficients of the regression of Y on T.

- Step 5: Normalize vector c.
- Step 6: Compute the Y score u.
- Step 7: The step is self-explanatory.
- Step 8: Columns of the E_{i-1} is regressed on the score to find corresponding loading vector.
- Step 9: Regress u on t to get the regress coefficient b to record in the diagonal matrix B.
- Step 10: This step is self-explanatory.

Now, it can be shown how scores and weights vectors in PLSR relate to eigen problems.

Without loss of generality, all the scalar terms are ignored in each step of the algorithm. Taking the Steps in the algorithm as conditions for Equation 2.28, it can be shown that

$$\begin{aligned} \underset{w}{argmax} \quad cov(t,u) &= cov(Xw, Yc) \quad \text{Conditions by Step 3 and Step 6} \\ &\propto cov(Xw, YY^TXw) \quad \text{Condition by Step 4} \\ &\propto w^T cov(X, YY^TX)w \\ &\propto w^T X^T YY^T Xw \end{aligned} \tag{2.30}$$

According to the maximization of quadratic form, the covariance of t and u is maximized when w is the eigenvector of X^TYY^TX. This shows w is a dominant eigenvector of X^TYY^TX (or the dominant left eigenvector of X^TY). To check whether the algorithm is computing the eigenvector of X^TYY^TX, the algorithm is followed step by step. $w \propto X^Tu$(Step 1)$\propto X^TYc$(Step 6 from last iteration)$\propto X^TYY^Tt$(Step 4 from last iteration)$\propto X^TYY^TXw$(Step 3 from last iteration). Thus, This is exactly an application of generalized power method for solving the dominant left eigenvector of X^TY.

Similarly,

$$\begin{aligned} \underset{c}{argmax} \quad cov(t,u) &= cov(Xw, Yc) \quad \text{Conditions by Step 3 and Step 6} \\ &\propto cov(XX^TYc, Yc) \quad \text{Condition by Step 1} \\ &\propto c^T cov(XX^TY, Y)c \\ &\propto c^TY^TXX^TYc \end{aligned} \tag{2.31}$$

Again, according to maximization of quadratic form, the covariance is maximized when c is the dominant right eigenvector of X^TY. Following the steps of NIPALS algorithm, $c \propto Y^Tt$(Step 4)$\propto Y^TXw$(Step 3)$\propto Y^TXX^Tu$(Step 1)$\propto Y^TXX^TYc$(Step 6 from last iteration).

Following the algorithm, similar results are shown for score vectors, t and u.

$$\begin{aligned} t &= Xw \quad \text{Step 3} \\ &\propto XX^T u \quad \text{Step 1} \\ &\propto XX^T Yc \quad \text{Step 6} \\ &\propto XX^T YY^T t \quad \text{Step 4} \end{aligned} \tag{2.32}$$

Therefore, t is in fact the dominant eigenvector of $XX^T YY^T$.

Similarly,

$$\begin{aligned} u &= Yc \quad \text{Step 6} \\ &\propto YY^T t \quad \text{Step 4} \\ &\propto YY^T Xw \quad \text{Step 3} \\ &\propto YY^T XX^T u \quad \text{Step 1} \end{aligned} \tag{2.33}$$

Thus, u is the dominant eigenvector of $YY^T XX^T$.

Overall, NIPALS algorithm of PLSR is in fact a series of SVD operations. w and c can be computed using one generalized power method. It should be noticed that the first step generates a vector which is a linear combination of the left eigenvectors of $X^T Y$. Then, Step 3 and Step 4 leads to the right eigenvector of $X^T Y$, thus, the computation of w and c is carried out using generalized power method. t and

u are calculated using the simple power method. Thus, NIPALS integrated three power methods into one algorithm.

After each iteration, the information in the direction of the score vector, t, is removed from the residual matrices of X and Y. Intuitively, the deflation step eliminates all the information of X and Y in the direction in the previous dominant eigenvector of XX^TYY^T. Thus, in the next iteration, power method leads to the next dominant eigenvector of the symmetric matrix $XX^TY_{i-i}Y_{i-i}^T$. The properties of a positive definite symmetric matrix ensures the score vectors are orthogonal to each other and allow us to create an orthogonal design matrix.

The coefficients of original variables are found to facilitate variable selection, and the coefficient matrix is given in Manne (1987) as follows:

$$B_{pls} = W(P^TW)^{-1}C^T \tag{2.34}$$

2.2.3 Cross-validation for Dimension Reduction

In general, there are different ways of choosing the number of PCs and LVs to be included in the model. However, all these methods are arbitrary. For example, with a PCR model, the amount of variance explained by PCs is considered to determine the appropriate number of components to be included. The cut-off point of the explained variance is subjective and often depends on personal preference. If one wants 100% of the information, all PCs are used. For the PLSR model, root

mean squared error of prediction (RMSEP) is often used as an intermediary for deciding the number of LVs to be included. However, this process is less arbitrary than looking at the explained variance since the number of LVs is chosen according to the lowest RMSEP. The arbitrary part of checking RMSEP is when a larger number of LVs generates a smaller out-of-sample error, it is subjective to decide whether it is worth sacrificing the degree of freedom for the lower error. In this study, to compare these two models, choices of number of PCs and LVs are both based on the out-of-sample error.

In statistics, out-of-sample error is normally calculated using cross-validation. In this paper, leave-one-out cross-validation (LOOCV) is applied to select the optimal numbers of PCs and LVs to be included in the PCR and PLSR, respectively. Thus, dimension reduction is achieved by discarding the PCs and LVs that only explain a very small proportion of the variations in the explanatory and the response variables. Cross-validation also improves the issue of overfitting for the purpose of variable selection.

To use LOOCV, the training dataset contains n-1 observations. One data point is left out in the testing set, and the error of prediction for the testing set is calculated. This process continues until all the observations are left out once. Then, sum of the the square roots of the prediction errors is calculated. Thus, PCR and PLSR are each calibrated n times, where n is the sample size. Root mean squared

error of prediction (RMSEP) is used as the criteria of model selection.

$$RMSEP = \sum_{i=1}^{n} \sqrt{\frac{(\hat{y}_i - y_i)^2}{n}} \tag{2.35}$$

As described above, the principle adopted for determining the optimal numbers of PCs and LVs are similar to the parsimony principle of choosing regression models (e.g., forward or backward selection).

2.2.4 Bootstrapping for Variable Selection Using Confidence Interval

Overall, the goal is not to find the PCs and LVs to be incorporated in the model, but, rather to determine the weather variables that have significant effect on forage yield using the variables selected under PCR and PLSR models. Thus, it is important to find an approach that allows us to select variables. In PCR and PLSR traditional measures such as p-values of weather variables cannot be calculated. Therefore, a non-parametric method is used to determine which variables are significant. The confidence interval of the coefficient of each weather variable is constructed, the significant weather variables can be identified by checking whether zero lies in its confidence interval.

The sample can be treated as the best representation of the whole population when the information about the population is unknown. Bootstrapping is a re-

sampling technique that takes advantage of this concept. The technique resamples the dataset with replacement for a number of times to approximate the distribution of an interested property of the population. By calibrating the models selected using LOOCV for a large number of times, confidence intervals are employed to check the significance of the original weather variables according to the estimated distribution of their coefficients.

In this study, the first order normal approximation of confidence intervals is considered as the standard of variable selection. To compute the confidence intervals, the R package 'boot' for bootstrapping is applied.

2.3 Part 2: Analysis of Average County Area-Yield Basis Risk

This section focuses on the methodologies for the second objective of basis risk analysis. First, a multivariable weather index is defined such that the index does not depend on a single weather variable and the results of the variable selection results of PCR and PLSR models can be utilized. Then, pairwise correlation maximization is used to optimize the multivariable weather indexes covering different number of counties. To analyze the changes in basis risk, the maximized correlations are compared as the number of counties covered by the index changes.

2.3.1 Multivariable Weather Index

In order to utilize the results from PCR and PLSR variable selection results, a multivariable weather index is defined in this paper. The effect of geographical scale (number of counties covered by an index) will be tested based on this multivariable weather index structure.

An important part of an index is the underlying index structure. The index structure can either be linear or non-linear. In the paper, a linear structure of the index is applied with the selected variables. Assume that $V_1, V_2, ..., V_l$ are the variables that are significant for a covered area, the new index, I, is defined as,

$$I = \sum_{i=1}^{l} w_i V_i \tag{2.36}$$

where w_i is the weight assigned to the ith variable.

2.3.2 Pairwise Correlation Maximization

The weights in the index need to be optimized according to a standard. In this paper, the multivariable weather index is optimized by determining the weights that maximize the pearson's correlations between the index and yields of the counties it covers. Therefore, the correlations between the index and yields are optimized using the optimization functions below. For an index covering a single

county,

$$\underset{\forall w_i}{argmax} \quad cor(I,Y) = \underset{\forall w_i}{argmax} \quad \rho = \underset{\forall w_i}{argmax} \frac{\sum\limits_{\#of years} (I_j - \bar{I})(Y_j - \bar{Y})}{\sqrt{\sum\limits_{\#of years} (I_j - \bar{I})^2} \sqrt{\sum\limits_{\#of years} (Y_j - \bar{Y})^2}} \tag{2.37}$$

With the procedures introduced above, the remainder of the paper empirically examine the methodology and their implications using first cut forage yield data in Ontario, Canada.

Chapter 3

Results

In this chapter, results of data processing are presented, including data detrending and stationarity testing to ensure the correctness of subsequent results. Then, the dimension reduction results of PCR and PLSR models are illustrated in Section 3.2.1. Following this, the variable selection results under PCR and PLSR are compared to check whether there are discrepancies between the two models. Based on the results of variable selection of each model, the new indexes are constructed for various geographical scales (numbers of counites) to test to what extent average county area-yield basis risk changes. In the last section of this chapter, the new indexes are compared with the rainfall index-based forage insurance plan offered in Ontario to provide possible suggestions on improving the current index.

3.1 Results of Data Stationarity for Yield

Stationarity of data is important for yield and weather analysis. Depending on the nature of the data, various techniques can be applied for the purpose of detrending, such as simple linear regression, piecewise spline regression, ARIMA, etc. In this study, all 7 counties in the dataset are found to follow simple linear trends, and no significant heteroskadasticity can be observed in the data. Thus, first, simple linear regression of yield on time are used to remove the trends for the counties with significant linear trends. Then, Kwiatkowski-Phillips-Schmidt-Skin(KPSS) tests are employed to check the stationarity of the data after detrending. The results for all 7 counties are shown in Table 3.1.

For counties that are not detrended including Algoma, Cochrane, Nipissing, Rainy River, and Thunder Bay, the small p-values as well as the small coefficients of linear trends both indicate that there are insignificant trends. This implies that forage yield has remained fairly consistent on average over the1980's to the middle of 2000's at most locations in Ontario. Limited literature discusses the historical trend of forage yield. However, Ottman et al. (2013) find that the trend of alfalfa, which is normally the biggest component of forage plantation, has not changed since the 1980's in the western 11 states in the US. Their finding agrees with the findings in this paper based on the Ontario forage data, which covers the same period of time as their research.

Table 3.1: Data Stationarity Summary

County	**Trend**	***P*-value**	**KPSS**
Algoma	-0.003	0.167	> 0.1
Cochrane	-0.001	0.921	> 0.1
Leeds Grenville	-0.032	0.003	> 0.1
Niagara	0.026	0.022	> 0.1
Nipising	-0.003	0.786	> 0.1
Rainy River	-0.005	0.733	> 0.1
Thunder Bay	-0.001	0.921	> 0.1

Note: Table above shows the coefficients of linear trend of yields, p-values of the coefficients, and p-values of Kwiatkowski-Phillips-Schmidt-Skin (KPSS) tests at all seven counties (after the removal of trend if trend is significant). Yield is only detrended for the counties with significant linear trends including Leeds Grenville and Niagara. For counties with larger p-values, the coefficients of linear trends also reassure the insignificance of trends by their small magnitudes.

3.2 Part 1: Weather Variable Selection

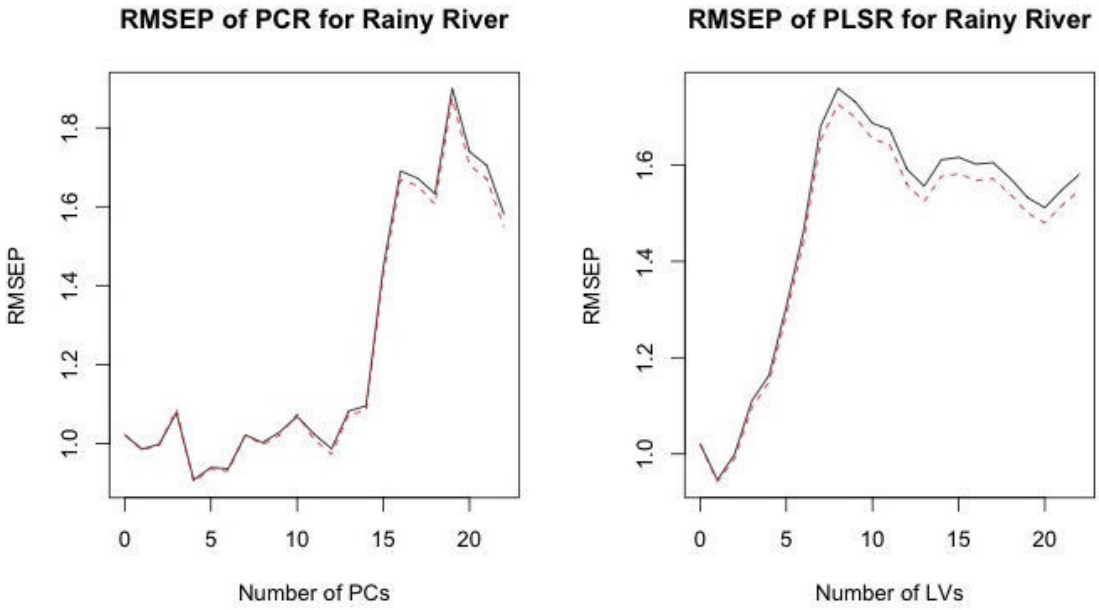
Part one of the results chapter summarizes the weather variable selection results of principal component regression (PCR) and partial least squares regression (PLSR) models. Weather variable selection results of the two models are presented

and compared to determine the most suitable model for weather variable selection with insufficient yield data and strong multicollinearity of weather variables. The results also provide insight regarding the weather variables that are most dominant in explaining forage yield.

3.2.1 Results of Dimension Reduction of PCR and PLSR for Weather Variables

Both PCR and PLSR are often applied for the task of dimension reduction and resolving multicollinearity. However, the two models are based on different underlying ideas. Regression results at four counties including Rainy River, Niagara, Nipissing, and Thunder Bay are compared to analyze the efficiencies of both models, as well as the differences between the models, including their variable selection results. For the purpose of dimension reduction, the numbers of projected variables (PCs in PCR model and LVs in PLSR model) that are incorporated in the two models are determined by out-of-sample prediction error attained from leave-one-out cross-validation (LOOCV). The performance of each model is based on the number of projected weather variables required to achieve the lowest prediction error as well as the magnitude of prediction error. The results of root mean squared error of prediction (RMSEP) against the number of projected weather variables are shown in Figure 3.1, Figure 3.2, Figure 3.3, and Figure 3.4, for each of the four counties.

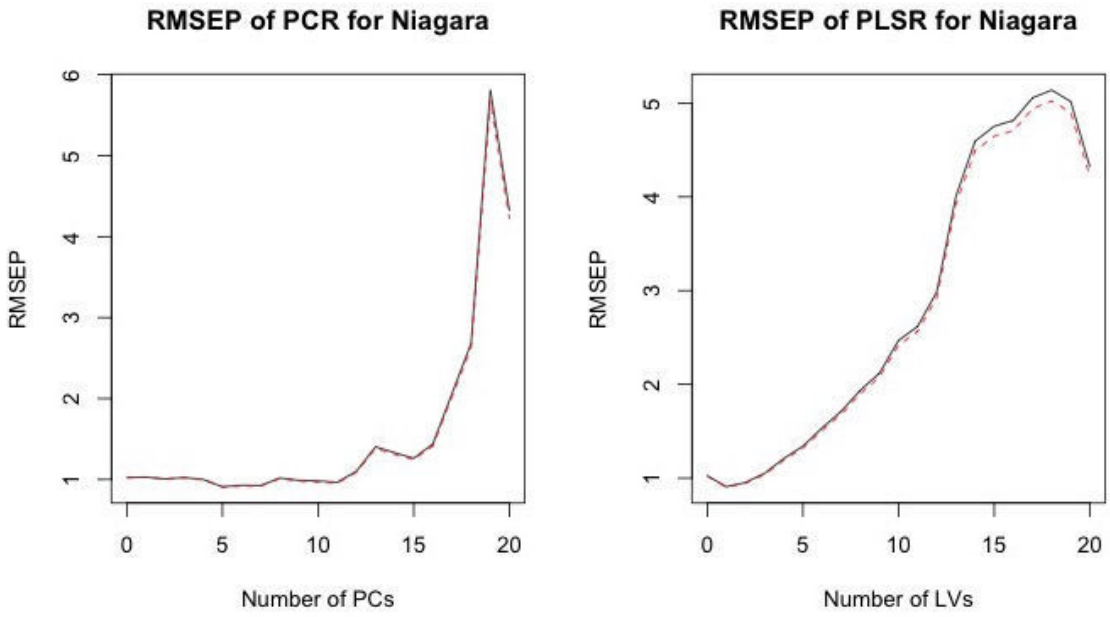
Figure 3.1: RMSEP at Rainy River



Note: Figure on the left shows the values of root mean squared error of prediction (RMSEP) at each number of principal components (PCs) in the principal component regression (PCR) model for Rainy River. The lowest RMSEP occurs with the first four PCs at 0.9077. Figure on the left shows the values of RMSEP at each number of latent variables (LVs) in the partial least squares regression (PLSR) model for Rainy River. The lowest RMSEP occurs with the first LV at 0.9461.

Table 3.2 summarizes the dimension reduction results of PCR and PLSR, and it can be shown that the dimension of the design matrix is reduced substantially for each county. For example, for Rainy River, the lowest RMSEP of 0.9077 is reached with only four PCs. Therefore, only four PCs are needed for the regression model, compared to the original dimension of 31 weather variables, which significantly improves the degree of freedom of the regression the model. In the case of Thun-

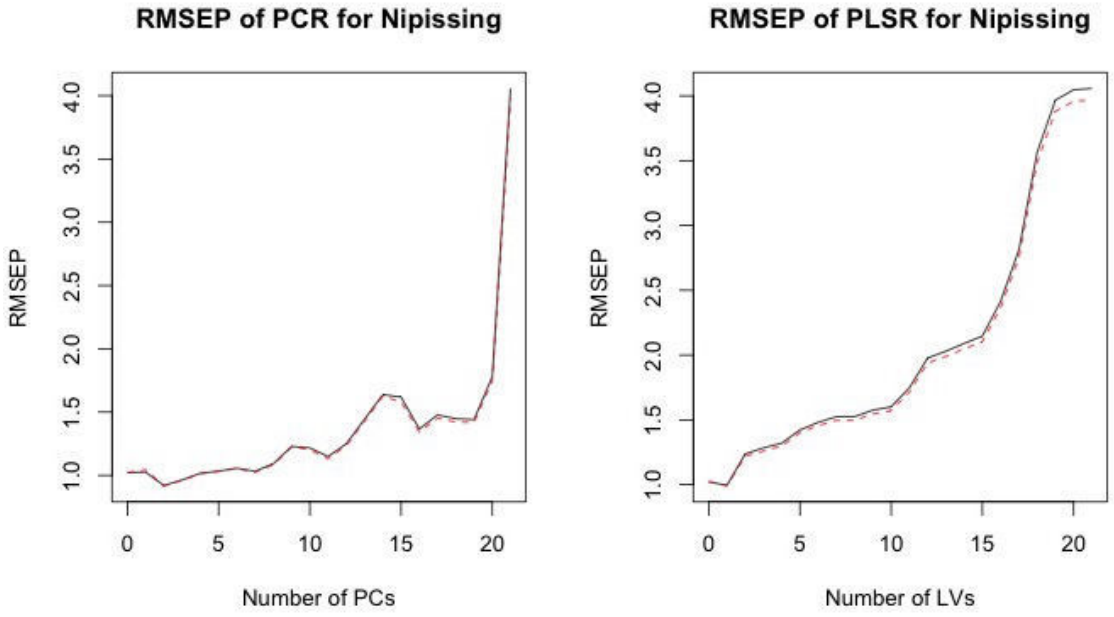
Figure 3.2: RMSEP at Niagara



Note: Figure on the left shows the values of root mean squared error of prediction (RMSEP) at each number of principal components (PCs) in the principal component regression (PCR) model for Niagara. The lowest RMSEP occurs with the first five PCs at 0.9100. Figure on the left shows the values of RMSEP at each number of latent variables (LVs) in the partial least squares regression (PLSR) model for Niagara. The lowest RMSEP occurs with the first LV at 0.9083.

der Bay, the lowest RMSEP of 0.7622 is attained by incorporating eight PCs in the model. However, the RMSEP is 0.7913 with six PCs, which is only 3.8% higher compared to eight PCs with two degrees of freedom saved. Thus, six PCs are perserved for Thunder Bay. Under PLSR, the lowest RMSEPs are normally obtained with no more than one LV compared to 31 variables in the original design matrix. One exception is Thunder Bay with RMSEP of 0.8172 with one LV, com-

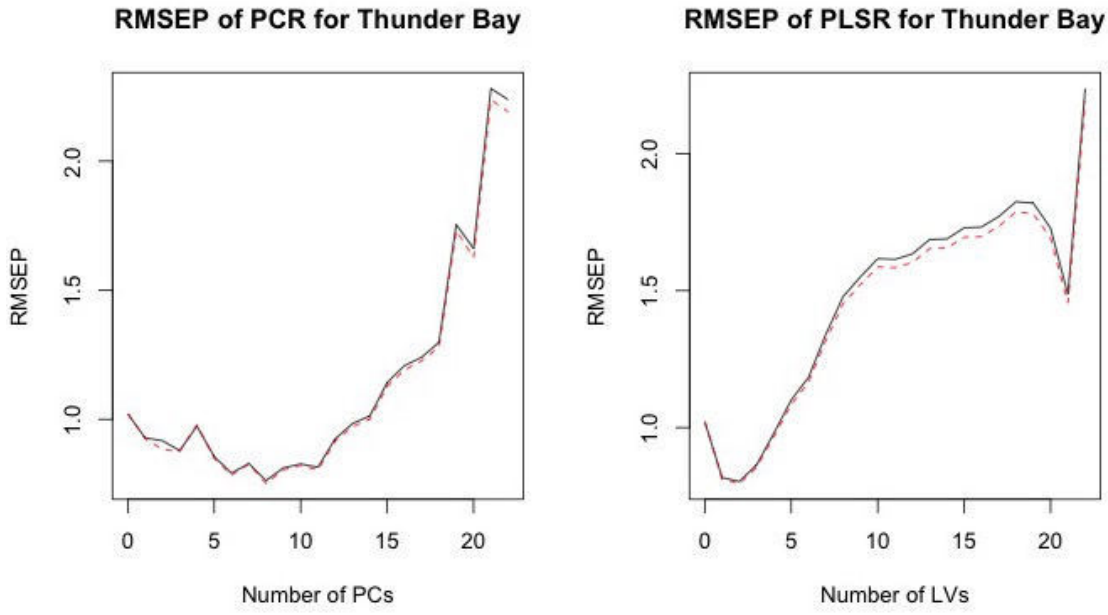
Figure 3.3: RMSEP at Nipissing



Note: Figure on the left shows the values of root mean squared error of prediction (RMSEP) at each number of principal components (PCs) in the principal component regression (PCR) model for Nipissing. The lowest RMSEP occurs with the first two PCs at 0.9196. Figure on the left shows the values of RMSEP at each number of latent variables (LVs) in the partial least squares regression (PLSR) model for Nipissing. The lowest RMSEP occurs with the first LV at 0.9948.

pared to the minimum RMSEP of 0.8030 when two LVs are incorporated. However, adding the additional LV only explains 1.7% more information than the first LV. Therefore, only the first LV is kept to save an extra degree of freedom. On average, PCR models are able to reduce the dimension of the weather variables from 31 to an average of 4.25, whereas the PLSR models can reduce the dimension to around one on average. Thus, based on the measure of numbers of PCs and

Figure 3.4: RMSEP at Thunder Bay



Note: Figure on the left shows the values of root mean squared error of prediction (RMSEP) at each number of principal components (PCs) in the principal component regression (PCR) model for Thunder Bay. The lowest RMSEP occurs with the first eight PCs at 0.7622. Figure on the left shows the values of RMSEP at each number of latent variables (LVs) in the partial least squares regression (PLSR) model for Thunder Bay. The lowest RMSEP occurs with the first two LVs at 0.8030.

LVs needed to achieve the lowest prediction error, the PLSR models are better at dimension reduction compared to PCR model.

However, for three of the four counties, PCR generates the lowest RMSEPs even though for these counties, the RMSEPs of PLSR models are less than 5% higher than PCR models. For Niagara, the PLSR model leads to a RMSEP that is even lower than the PCR model and saves four degrees of freedom. According to

Table 3.2: Dimension Reduction Summary

	PCR		**PLSR**	
	PCs	**RMSEP**	**LVs**	**RMSEP**
Rainy River	4	0.9077	1	0.9461
Niagara	5	0.9100	1	0.9083
Nipissing	2	0.9196	1	0.9948
Thunder Bay	6	0.7913	1	0.8172

Note: The table shows the number of principal components (PCs) and latent variables (LVs) kept for principal component regression (PCR) and partial least squares regression (PLSR) models for all four counties, respectively. PCR reduces dimension to an average of around four PCs. PLSR model reduces the average dimension to one. For Thunder Bay, although first eight PCs generate the lowest root mean squared error of prediction (RMSEP) for PCR model, with six PCs, the RMSEP is only 3.8% lower, thus, only six PCs are kept to save two extra degrees of freedom. Likewise, Lowest RMSEP is generated with first two LVs, however, only the first one is kept to save a degree of freedom, which only leads to an increase in RMSEP of 1.7%..

the performance measure of the lowest prediction error, the PCR model is almost always able to attain the lowest prediction error with samples in this study. Overall, which model to use depends on an individual's preference for a higher degree of freedom (PLSR) or a lower prediction error (PCR). Therefore, the subsequent numerical results are based on both PCR and PLSR models.

3.2.2 Results of Variable Selection of PCR and PLSR

So far, three types of variables have been introduced, the original weather variables, PCs of the weather variables, and the LVs of the weather variables. To avoid confusions among these different types of variables, the original weather variables are rotated so that they are independent as the PCs in the PCR model and LVs in the PLSR model. The regression is based on the PCs and LVs. However, for the purpose of variable selection, the coefficients of the PCs and LVs are transformed back to the coefficients of the original weather variables according to the relationships in Equation 2.21 and 2.32. Thereafter, bootstrapping is used to construct confidence intervals for the original variables instead of the PCs and LVs.

Variable selection is implemented using both PCR and PLSR with bootstrapping. Despite the discrepancies in terms of dimension reduction and magnitudes of RMSEPs, the results of variable selection under both models are very close and consistent for each county. For Rainy River, ten weather variables are chosen by the PLSR model with eight of them also chosen by the PCR model as well. The PCR model determines that 19 weather variables are significant at Thunder Bay, and 18 of these variables are selected by the PLSR model as well. similar results are found for the other two counties as well. The pattern is expected as PCR model is based on the variances and covariances of weather variables, while PLSR is based on the covariance between the weather variables and the yield. Weather effects with relatively large variations are also most likely the ones imposing dom-

inant effects on the forage yield, and thus, with the largest covariances with the yield.

In order to design an index for different geographical scales (i.e. different numbers of counties), a standard is needed for 'filtering' out variables for the construction of a multivariable weather index covering these counties simultaneously. As the index is structured to cover all the counties forming the area, it is reasonable to use the weather variables that are commonly significant for the counties within the area. For example, for an area formed by Rainy River and Niagara, Rainy River has 16 significant weather variables and Niagara has 15 significant variables, but only nine of these variables are significant for both counties. Thus, the index for this area is constructed upon these 9 variables only. If the index covers an area defined by only one county, all the variables selected by PCR or PLSR for this county are used for the construction of the index. With this simple variable filtering principle, in the next section, the results regarding the effect of geographical scale of an index on its average county area-yield basis risk is shown.

3.3 Part 2: Analysis of Average County Area-Yield Basis Risk

In this section, the results regarding the second objective of the paper are presented. The implication of geographical scale (i.e. number of counties) on the av-

erage county area-yield basis risk of an index is analyzed. Multivariable weather indexes are constructed with the results of variable selection and a simple variable filtering principle for different numbers of counties. The results of pairwise correlation maximization, which optimizes the correlations between the multivariable weather index and the yields of counties covered by the index, are shown in Tables 3.3 and 3.4. Then, the correlations of indexes covering different geographical scales are compared to measure the changes in average county area-yield basis risk. Moreover, yield samples from Algoma, Cochrane, and Leeds Grenville are used to check the overfitting issue as well as the ability of the indexes to capture the weather effects correctly.

3.3.1 Results of Multivariable Weather Indexes for a Single County

When an index is defined for only one county, the correlation between the index and the forage yield is expected to be high since the variables selected and the weights assigned to the selected variables tend to be locationally specific. In the analysis as shown in Tables 3.3 and 3.4 for areas formed by a single county, this finding is observed. For example, the average correlation between the index and the county yield is above 0.88 under PCR and PLSR models. In Table 3.3, Thunder Bay has a correlation of as high as 0.956 and Nipissing has a correlation of as high as 0.959.

In regression models, when the R-squared value is extremely high, or the in-

sample error is extremely low, the problem of overfitting is, in general, a concern. Similarly, the overfitting issue should be checked when the optimized correlations are considerably high. For Thunder Bay, the sample size of the forage yield is 23 years, but 22 variables are selected for the construction of index. In this case, the problem of overfitting exists as the overall rank of the selected variables is high. Thus, the index is overfitted for the sample and may not be representative for the out-of-sample data and therefore, not useful for projecting losses in the future. The overfitting problem can be overcome by constructing indexes covering more than one county.

3.3.2 Results of Multivariable Weather Indexes for Multiple Counties

As shown in Table 3.3 and 3.4, when the number of counties covered by the index changes from one county to two counties, the correlations between the index and the forage yield decrease substantially under both PCR and PLSR approaches. With variable selection results from the PCR model, by changing from a one-county index to a two-county index, the correlations decrease by 23.5%, 16.8%, 29.8%, and 26.0% for RR, NG, NP, and TB, respectively. As the geographical scale increases from one county to two, the correlations decrease by 29.7% on average for all locations under PLSR. Since the overfitting issue prevails among the one-county index, these significant decreases in the correlations tend to over-

estimate the increase in the basis risk. Therefore, it is more reasonable to compare the indexes covering multiple counties.

When the indexes are built for combinations of three counties, the correlations are 16.5%, 11.9%, 16.9%, and 14.4% lower than the two-county indexes for Rainy River, Niagara, Nipissing, and Thunder Bay, respectively, under PCR approach. Similarly, under PLSR approach, the two-county indexes lead to improvements of 25.2%, 17.6%, 5.9%, and 12.1% over the three-county index for Rainy River, Niagara, Nipissing, and Thunder Bay, respectively. When a four-county index is used instead of the three-county index, an average fall of 16.6% in the correlations can be seen under PCR and PLSR approaches. Thus, the average county area-yield basis risk of the indexes increase due to the decreasing correlations, likely because the index becomes less representative of the yields for all counties. Overall, the findings align with the statement by Elabed et al. (2013) that basis risk increases with a broader geographical scale. However, caution must be taken since the basis risk measured in this study is the average county- area-yield basis risk, rather than the individual farmer basis risk, which is commonly used to address basis risk concerns in other studies. Therefore, the basis risk analyzed in this study may underestimate the basis risk relative to the farmer level basis risk.

Another interesting finding is that as the geographical scale of the indexes rise (i.e adding more counties to the index), although the overall correlations drop for all the counties, the indexes demonstrate an improved ability to more consistently

capture weather effects for forage. For example, the correlations of the out-of-area counties carry the same sign with the optimized correlations of the in-area counties. This means the index structure captures an aggregated weather effect that has either significant positive or negative impact on forage yield regardless of the counties. For an overfitted index such as the indexes for Thunder Bay and Nipissing in the single-county index case under PCR and PLSR approachs, the weights in the index structure optimized do not lead to the same signs at all the out-of-area counties. For instance, the out-of-area correlations of the index defined for Thunder Bay are random and volatile. Moreover, the correlations for some out-of-area counties are very close to zero when the index is overfitted. For example, in Table 3.3, with the single-county index for Nipissing, the correlations for Cochrane and Leeds Grenvilles are both very insignficiant from zero. Therefore, the index does not manifest a consistent weather effect at all the counties and the overfitting issue is of serious concern.

3.3.3 Comparison between the Multivariable Weather Index and the Rainfall Index-based Forage Insurance Plan in Ontario

The index in Ontario is defined as the cumulative rainfall during May and June, with adjustments in recognition of rainfall levels with limited effect on forage production. For example, the daily rainfall of 1mm is counted as 0mm, and the daily cap is 50mm, and the monthly cumulative rainfall is capped at 125% of the

Table 3.3: PCR Index Correlations

	In-area				Out-of-area		
Area	**County 1**	**County 2**	**County 3**	**County 4**	**Algoma**	**Cochrane**	**Leeds Grenville**
RR	0.888				0.525	0.147	-0.005
NG	0.861				0.083	-0.241	0.391
NP	0.959				0.242	-0.099	-0.09
TB	0.956				0.082	-0.164	-0.070
RR&NG	-0.639	-0.639			-0.33	-0.34	-0.23
RR&NP	-0.671	-0.581			-0.361	-0.359	-0.198
RR&TB	-0.728	-0.672			-0.364	-0.207	-0.082
NG&NP	0.817	0.764			0.406	-0.053	0.312
NG&TB	0.691	0.741			0.091	-0.124	0.301
RR&NG&NP	0.500	0.697	0.534		0.296	0.342	0.240
RR&NG&TB	0.521	0.527	0.491		0.293	0.104	0.312
RR&NP&TB	-0.680	-0.463	-0.669		-0.326	-0.256	-0.129
NG&NP&TB	0.668	0.683	0.675		0.206	-0.215	0.300
RR&NG&NP&TB	0.462	0.574	0.482	0.468	0.282	0.186	0.318

Note: The table summarizes the optimized correlations with the variable selection results from PCR model for Rainy River, Niagara, Nipissing, and Thunder Bay under multivariable weather indexes covering different numbers of counties. When the index covers a single county, the correlations are extremely high, however, the problem of overfitting should be aware of. The problem of overfilling is reflected by small magnitude of out-of-area correlations as well as the inconsistent signs among the three out-of-area counties' correlations. As the index covers more than one county, the problem of overfitting becomes much less problematic. Moreover, as the geographical coverage of the index increases one county, the average optimized correlations of all the counties decrease significantly.

Table 3.4: PLSR Index Correlations

	In-area				Out-of-area		
Area	**County 1**	**County 2**	**County 3**	**County 4**	**Algoma**	**Cochrane**	**Leeds Grenville**
RR	0.831				0.343	0.145	0.372
NG	0.921				0.140	-0.328	0.127
NP	0.698				0.226	0.109	-0.334
TB	0.964				0.083	-0.177	0.025
RR&NG	-0.675	-0.469			-0.409	-0.304	-0.233
RR&NP	-0.492	-0.508			-0.339	-0.478	-0.209
RR&TB	-0.678	-0.724			-0.335	-0.194	-0.383
NG&NP	0.484	0.536			0.382	0.396	0.249
NG&TB	0.664	0.701			0.457	0.170	0.303
RR&NG&NP	-0.455	-0.418	-0.460		-0.382	-0.566	-0.183
RR&NG&TB	0.533	0.419	0.453		0.428	0.506	0.340
RR&NP&TB	0.394	0.541	0.698		0.294	-0.049	0.267
NG&NP&TB	-0.495	-0.479	-0.726		-0.374	-0.444	-0.272
RR&NG&NP&TB	0.394	0.326	0.521	0.430	0.369	0.484	0.263

Note: The table summarizes the optimized correlations with the variable selection results from PLSR model for Rainy River, Niagara, Nipissing, and Thunder Bay under multivariable weather indexes covering different numbers of counties. Similar to the results under PCR model, when the index covers a single county, there is a strong tendency for overfitting. The problem of overfilling is reflected by small magnitude of out-of-area correlations and the inconsistent signs among the three out-of-area counties' correlations. As the index covers more than one county, the problem of overfitting becomes much less problematic. Again, as the geographical coverage of the index increases one county, the average optimized correlations of all the counties drops dramatically. This indicates a higher basis risk since the index becomes less reflective of the forage yield at all the covered counties.

Table 3.5: Ontario's Index Correlations

Index	RR	NG	NP	TB	AG	CH	LG
Ontario	0.459	0.228	0.641	0.597	0.350	0.385	0.425
Three-county	0.567 23.5%	0.631 176.8%	0.560 (12.6%)	0.612 2.5%			
Two-county	0.679 47.9%	0.716 214.0%	0.673 5%	0.707 18.4%			

Note: The table shows the correlations of Ontario's index with yields compared to the average optimized correlations with PCR results at two and three-county geographical levels. The percentages show by how much the two and three-county indexes outperform Ontario's index. Ontario's index structure is optimized for all the counties within the province. The results show again how much correlations improve when geographical scale decreases. Thus, a reduction in geographical scale may improve the performance of the current Ontario's rainfall index.

long-term monthly rainfall, where the long-term rainfalls are determined according to the last 45 years weather data. The summary of the correlations between yields and Ontario's index is summarized in the Table 3.5.

An important advantage of Ontario's index is that it is simple to understand. At the same time, as shown in Table 3.5, the correlations present the same sign across all the counties. This means Ontario's index is able to pick up a significant rainfall impacts that affects the forage yield at all seven counties in the same direction. In fact, the construction of Ontario's index is similar to the construction of the multi-county index but it covers all the counties in Ontario, thus, the geographical scale of Ontario's index is very large. The index was officially launched in 2005, therefore, the data in this paper was considered for the development of

this poxy index structure, and the comparison between Ontario's index and multivariable weather multiple-county index should be relatively fair. Although the new index structure is a little more complex, the linear relationship between the index and the yield is able to reduce the ambiguity in terms of claim computation. The new index also has the advantage of merging various weather effects into a single indicator. Indexes are compared without considering the magnitude of the correlations at Algoma, Cochrane, and Leeds Grenville as the purpose of the new index is to cover only the in-area counties.

Compared to Ontario's index, the three-county indexes under PCR improve the correlations between the index and the yield for Rainy River and Niagara, significantly, and result in slightly higher correlations for Thunder Bay. The correlations are 23.5%,176.8%, and 2.5% higher than Ontario's index for Rainy River, Niagara, Nipissing, and Thunder Bay, respectively. Although the correlation for Nipissing is 12.6% lower under the three-county indexes than Ontario's index, when the number of counties continue to decrease, all the indexes constructed have significantly higher correlations under both variable selection approaches compared to Ontario's index. For example, RR has correlations of 0.679 and 0.617 on average under PCR and PLSR, respectively, compared to 0.459 for Ontario's index, the improvement in the correlation is more than 30%. For Niagara, Nipissing, and Thunder Bay, the improvements in correlations under the two-county index are at least 5% higher than Ontario's index. Under the PCR approach, the

average correlation for NG under the two-county index is 214.0% higher than the correlation achieved by Ontario's index.

3.4 Additional Results Regarding the Importance of Reasonable Clustering of Counties

In addition to the effect of geographical scale, the grouping of counties for an area is also important. For example, under the PLSR approach, for the area, RR&NP, Rainy River has a correlation of -0.492, and for the area, NG&TB, Thunder Bay has an correlation of 0.701. However, by clustering Rainy River and Thunder Bay in the same area, the correlations between the index and the yields increase by 37.8% and 3.3%, respectively.

For Ontario's index, Nipissing has a correlation which is dramatically higher than the correlations at other counties. This means that the the rainfall index fits Nipissing better than other locations. It could also reflect that Nipissing has some meteorological conditions different from other locations. The finding is supported by the multivariable weather index results as well. For example, considering two-counties multivariable weather indexes in Table 3.4, the correlations are generally lower for areas formed by pairing NP with other counties. Moreover, for the four-county index in Table 3.4, the correlation for Nipissing is the highest, and Niagara has the lowest correlation. This pattern is the same as Ontario's index, suggesting

Nipissing is dominating the optimization and stealing the correlations at other counties. Therefore, Nipissing may need to be considered separately for the index construction or clustered with other counties with similar meteorological conditions. Overall, it is necessary to construct the index on an appropriate cluster of counties.

Chapter 4

Summary

4.1 Summary of Results

IBI is a valuable substitute to traditional insurance for forage as the yield of forage is difficult to track. The low demand of forage IBI has been a barrier for the development of this type of insurance, suggesting more research is needed. Basis risk has been recognized as one of the main causes for the low participation rate, and this paper focuses on the analysis of basis risk at the average county level.

With respect to the forage industry in Ontario, there are several major issues associated with the analysis of basis risk. The first problem is that the current forage index only depends on rainfall. This has already served as the basis for some criticism from cattlemen and farmers in Ontario, but the construction of an index with multiple variables requires the analysis of the joint effect of a large set

of weather variables on forage yield. However, this kind of analysis is impeded by the insufficient forage yield data. Therefore, the problem of variable selection with limited yield data is addressed by the first objective of this paper. The second problem is that the currently offered rainfall index in Ontario covers all the counties in the entire province. This may have tremendous implications on basis risk, and this issue is examined by the second objective of this study.

The first objective of this paper is to select weather variables, that have significant effects on forage yield, as candidates for the construction of weather indexes by using principal component regression (PCR) and partial least squares regression (PLSR) models to address challenges of high dimension and multicollinearity of weather variables.

The second objective is to analyze the trend of average basis risk as the number of counties covered by the index increases under a multivariable weather index structure, using the variable selection results from PCR and PLSR models.

Data

Farm-level forage yields are aggregated to the county level for first cut forage in Ontario. Data over the period 1981-2004 is used, resulting in approximately 25 observations in the time series for each of seven counties that make up the province of Ontario in this study. Daily weather data, including maximum and minimum temperatures, rainfall, and hours of sunshine, is attained from Agricorp, Ontario. The problem of missing weather data is limited with weather ob-

servations verified by professional third parties and rechecked with Environment Canada's records. With the weather data, 31 weather variables are included in the analysis, such as monthly maximum temperature, monthly cumulative rainfall, seasonal cumulative rainfall, etc.

Methodology

PCR and PLSR are both applied for the task of variable selection. These two models are compared according to the criteria of dimensions needed to attain the lowest prediction error and magnitudes of the lowest prediction error. Variable selection results are also compared to suggest the best model for the task of variable selection under the constraint of limited yield data.

For the second objective of analyzing average basis risk, a multivariable weather index structure is defined in this paper to take advantage of the variable selection results of the PCR and PLSR models, The multivariable weather indexes are constructed according to a pairwise correlation optimization procedure for different numbers of counties to examine the impact of geographical scale (i.e. number of counties) on average county area-yield basis risk. This basis risk analysis may understate basis risk that is more commonly analyzed at the farm level. The correlations between the multivariable weather indexes and the yields are compared as the geographical scale of the index changes. These results are compared to the current rainfall index-based forage insurance plan in Ontario, which is based on an extremely large geographical scale that covers all counties in Ontario. This is

compared to the multivariable index developed in this paper that covers two, or three counties. Therefore, the issue of basis risk with respect to the geographical scale of the index is examined.

Results

4.1.1 Part 1: Summary of Variable Selection

The objective of variable selection under yield data scarcity and multicollinearity is addressed in part one of this paper, using principal component regression(PCR) and partial least squares regression(PLSR) models. The paper first provides a detailed introduction as well as an interpretation of PCR and PLSR models. It is shown how PCR and PLSR models can be transferred into eigenproblems, while power methods are used in the non-linear iterative partial least squares (NIPALS) algorithm for solving eigenproblems in these two models.

Two main results were found regarding the first objective by comparing performances of these two regression models. The first result shows that PLSR almost surely generates the lowest prediction error with no more than one latent variable (LV or projected weather variable) for the counties in this study. Thus, it substantially lowers the dimension of weather variables from 31 to around one. However, in contrast to the PLSR model, the lowest root mean squared error of predictions (RMSEPs) are lower under the PCR model, but, with more principal components (PC or projected weather variables). Therefore, the PLSR model is better at main-

taining a relatively low dimension of weather variables, whereas the PCR model does a better job with achieving lower prediction errors with samples in this study. For the purpose of variable selection, bootstrapping is applied to construct confidence intervals for each weather variables in the weather variable matrix.

The second result shows that although there are some discrepancies between PCR and PLSR models, they are able to provide very similar and consistent variable selection results for all the counties considered in this study. For example, for Thunder Bay in the sample of this paper, the PCR model finds 19 significant weather variables, among which, 18 of these weather variables are also determined by the PLSR model. Thus, deciding on which model to be used depends on personal preferences for a lower prediction error or a lower degree of freedom.

4.1.2 Part 2: Summary for Analysis of Average County Area-Yield Basis Risk

The second part of this paper tests the effect of geographical scale(number of counties) on an index's average county area-yield basis risk. A multivariable weather index structure is defined to utilize the variables selected by either the PCR or the PLSR model. The index is optimized for all the counties covered by it using a pairwise correlation function which is inspired by the pairwise likelihood function. The multivariable weather indexes are constructed for different numbers of counties to measure the impact of geographical scale covered by the index

on average county area-yield basis risk.

Results show that as the geographical scale of the index changes from four counties to three counties, the correlations of all the counties increase by 16.4% on average. With a further reduction from three counties to two counties, the average correlations between the index and yields of all the counties increase by 14.7% and 15.2% with the PCR and the PLSR variable selection results, respectively. When the index covers only one county, the average correlations again increase by more than 20% for most counties under both PCR and PLSR results.

The rainfall index currently offered in Ontario is compared with the multivariable weather indexes developed in this paper. The geographical coverage of the Ontario's index is extremely large. Thus, as expected, the correlations in Table 3.5 show that the Ontario's index reflects a consistently positive rainfall effect on forage yield. However, compared to the Ontario's index, it can be seen that with a multivariable weather index covering only three counties, there are significant increases in correlations for some counties, such as Rainy River and Niagara, where the correlations are 23.5% and 176.8% higher than Ontario's index, respectively. With a further reduction in the number of counties, the average correlations of all the counties are much higher than the Ontario's index which covers the entire province.

4.1.3 Summary of Additional Results

In addition, the grouping of areas covered by the index is important. In the results, it is shown that different groupings of counties can significantly affect the optimal correlations that can be achieved. Thus, to construct an index more reflective of the forage yield of an area, counties with similar meteorological and spatial conditions should be grouped so that more common significant weather variables can be used to construct the index.

4.1.4 Conclusion

Overall, to select weather variables with limited yield data and strong multicollinearity, results show that the PLSR model is able to reduce more dimensions of the weather variables compared to the PCR model, while the PCR model is capable of attaining lower prediction errors than the PLSR model. However, with these discrepancies between the PCR and PLSR models, the weather variable selection results for each county in the sample of this study are very close under the PCR and the PLSR models. The choice of model for weather variable selection depends on preferences on lower dimension of weather variables or lower prediction errors of the model.

The second part of this paper suggests that indexes covering a larger number of areas tend to suffer more average basis risk because of the dropping correlations between the index and forage yields. This result coincides with the state-

ment made by Elabed et al. (2013). However, caution should be taken in interpreting these results since this study examines average basis risk at the county level, compared to other studies that tend to focus on farmer level basis risk. Therefore, the basis risk measured in this study may be understated compared to farm-level basis risk.

Furthermore, groupings of counties according to their weather and geographical conditions can ensure that the index is more meteorologically and spatially specific, which can lead to a further reduction in basis risk.

The results will help commercial and governmental insurance institutions with analyzing the meteorological effects of multiple weather variables on crops with limited yield data. Moreover, it provides an insight to the optimal geographical index coverage of future index design for the minimal basis risk.

4.2 Limitations and Future Research

The approaches proposed in this paper focus on pearson's correlation to maintain a linear relationship between weather variables and yields. On one hand, a non-linear relationship between the index and yield can cause more confusion for farmers regarding contract structure. However, on the other hand, a non-linear relationship may be more representative, thus, further reducing the basis risk. However, the procedures can be easily generalized by incorporating non-linear correlations, such as rank correlations, into the optimization procedure. Future

study will focus on the non-linear relationship. At the same time, a more interpretable index structure will be developed to minimize the ambiguity associated with the non-linear relationship.

The correlations optimized in this paper may not be sufficiently high to implement in practice. This can be attributed to several issues. First, more detailed agronomic information regarding weather variables that are critical for the growth of forage can be chosen for the initial construction of the design matrix. Second, the linear relationship between a weather variable and the forage yield usually changes at a certain level. For example, when excess rainfall leads to a flood, the linear relationship between cumulative rainfall and forage yield can alter from a positive gradient to a negative gradient. Therefore, this problem can further examined by consulting agronomical researchers or using the correlation optimization procedure for each weather variable before using them for the construction of the index. For instance, for the maximum temperature, a threshold parameter can be set for the maximum temperature, and the correlation between the maximum temperature and the yield can be optimized first using the pearson's correlations optimization procedure in this paper. In this case, the index will instead be the maximum temperature as a function of the threshold parameter. In addition, statistical clustering approaches can also be applied for the grouping of the counties with similar weather conditions and geographical conditions to improve the optimization results that further reduce the problem of basis risk.

List of References

B. Barnett and O. Mahul. Weather index insurance for agriculture and rural areas in lower– income countries. *American Journal of Agricultural Economics*, 89(5): 1241–1247, 2007.

C. B. Barret, J. B. Barnett, M. R. Carter, S. Chantarat, J. W. Hansen, A. G. Mude, D. E. Osgood, J. R. Skees, C. G. Turvey, and M. N. Ward. Poverty traps and climate risk: limitations and opportunities of index-based risk financing. *International Research Institute for Climate and Society Technical Report 07-02*, 2007.

I. Bobojonov and R. Sommer. Alternative insurance indexes for drought risk in developing countries. *Presentation at the EAAE Congress*, 2011.

M. Boyd, J. Pai, Q. Zhang, H. H. Wang, and K. Wang. Factors affecting crop insurance purchases in china: the inner mongolia region. *China Agricultural Economic Review*, 3(4):441–450, 2011.

X. Cao, O. Okhrin, M. Odening, and M. Ritter. Modelling spatio-temporal variability of temperature. *Computational Statistics*, pages 1–22, 2014.

M. Carter, A. de Janvry, E. Sadoulet, and A. Sarris. Index-based weather insurance for developing countries: A review of evidence and a set of propositions for up-scaling". *FERDI working Paper 112*, 2014.

F. W. Chen and C. W. Liu. Estimation of the spatial rainfall distribution using inverse distance weighting (idw) in the middle of taiwan. *Paddy and Water Environment*, 10(3):209–222, Semptember 2012.

D. Clarke. A theory of rational demand for index insurance. *Department of Economics Discussion Paper Series, Oxford University, No. 572*, 2011.

S. Cole, T. Sampson, and B. Zia. Money or knowledge? what drives demand for financial services in emerging markets? *Harvard Business School Working Paper*, 9:117, 2009.

X. Deng, B. J. Barnett, G. Hoogenboom, Y. Yu, and A. G. Garcia. Alternative crop insurance indexes. *Journal of Agricutural and Applied Economics*, 40(1):223–237, April 2008.

K. N. Dirks, J. E. Hay, C. D. Stow, and D. Harris. High-resolution studies of rainfall on norfolk island: Part ii: Interpolation of rainfall data. *Journal of Hydrology*, 208 (3):187–193, 1998.

G. Elabed and M. R. Carter. Compound-risk aversion, ambiguity and the willingness to pay for microinsurance. *Journal of Economic Behavior and Organization*, 2015.

G. Elabed, M. F. Bellemare, M. R. Carter, and C. Guirkinger. Managing basis risk with multi-scale index insurance. *Agricultural Economics*, 44:419–431, 2013.

X. Giné and D. Yang. Insurance credit, and technology adoption: Field experimental evidence from malawi. j. of dev. *Econ.*, 89:1–11, 2009.

X. Giné, M. Lev, T. Robert, and V. James. Microinsurance: a case study of the indian rainfall index insurance market. *World Bank Policy Research Working Paper Series, No. 5459. Washington, D.C.: World Bank*, 2010.

N. Jensen, A. Mude, and C. B. Barnett. How basis risk and spatiotemporal adverse selection influence demand for index insurance: Evidence from northern kenya. *Mineo. Ithaca: Cornell University*, 2014.

J. Lin, M. Boyd, J. Pai, L. Porth, Q. Zhang, and K. Wang. Factors affecting farmers willingness to purchase weather index insruance in the hainan province of china. *Agricultural Finance Review*, 75(1):103–113, 2015.

A. Mair and A. Fares. Assessing rainfall data homogeneity and estimating missing records in makaha valley, o'ahu, hawaii. *Journal of Hydrologic Engineering*, 15(1): 61–66, 2010.

R. Manne. Analysis of two partial-least-squares algorithms for multivariate calibration. *Chemometrics and Intelligent Laboratory Systems*, 2:187–197, 1987.

M. Ottman, D. Putnam, V. Barlow, J. Brummer, M. Bohle, E. Creech, S. Foster,

R. Long, M. Marsalis, S. Norberg, S. Orloff, and G. Shewmaker. Long term trends and the future of the alfalfa and forage industry. *Western Alfalfa and Forage Symposium, December 11-13*, 2013.

M. Ritter, O. Mußhoff, and M. Odening. Minimizing geographical basis risk of weather derivatives using a multi-site rainfall model. In *Proceedings of the 123rd EAAE Seminar*, Dublin, Ireland, Feburary 23-24 2012.

M. Rosenzweig and H. Binswanger. Wealth, weather risk and the composition and profitability of agricultural investments. *Economic Journal*, 103:56–78, 1993.

J. Skees. Innovations in index insurance for the poor in lower income countries. *Agricultural and Resource Economics Review*, 37:1–15, 2008.

P. Wakker, R. Thaler, and A. Tversky. Probabilistic insurance. *Journal of Risk Uncertainty*, 15(1):7–28, 1997.

W. Zhu. *Actuarial Ratemaking in Agricultural Insurance*. PhD thesis, University of Waterloo, 2015.