Multi-risk modeling for improved agriculture decision-support: predicting crop yield variability and gaps due to climate variability, extreme events, and disease

by

Weixun Lu
B.Sc., University of Manitoba, 2010
M.Sc., University of Victoria, 2013

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Geography

Multi-risk modeling for improved agriculture decision-support: predicting crop yield variability and gaps due to climate variability, extreme events, and disease

by

Weixun Lu
B.Sc., University of Manitoba, 2010
M.Sc., University of Victoria, 2013

Supervisory Committee

Dr. David Atkinson., Supervisor
(Department of Geography)

Dr. Nathaniel Newlands., Supervisor
(Agriculture and Agri-Food Canada)

Dr. Alex Cannon., Outside Member
(Environment and Climate Change Canada)

Supervisory Committee

Dr. David Atkinson., Supervisor
(Department of Geography)

Dr. Nathaniel Newlands., Supervisor
(Agriculture and Agri-Food Canada)

Dr. Alex Cannon., Outside Member
(Environment and Climate Change Canada)

ABSTRACT

The agriculture sectors in Canada are highly vulnerable to a wide range of inter-related weather risks linked to seasonal climate variability (e.g., El Niño Southern Oscillation (ENSO)), short-term extreme weather events (e.g., heatwaves), and emergent disease (e.g., grape powdery mildew). All of these weather-related risks can cause severe crop losses to agricultural crop yield and crop quality as Canada grows a wide range of farm products, and the changing weather conditions mainly drive farming practices. This dissertation presents three machine learning-based statistical models to assess the weather risks on the Canadian agriculture regions and to provide reliable risk forecasting to improve the decision-making of Canadian agricultural producers in farming practices.

The first study presents a multi-scale, cluster-based Principal Component Analysis (PCA) approach to assess the potential seasonal impacts of ENSO to spring wheat and barley on agricultural census regions across the Canada prairies areas. Model prediction skills for annual wheat and barley yield have examined in multi-scale from spatial cluster approaches. The 'best' spatial models were used to define spatial patterns of ENSO forcing on wheat and barley yields. The model comparison of our spatial model to non-spatial models shows spatial clustering and ENSO forcing

have increase model performance of prediction skills in forecasting future cereal crop production.

The second study presents a copula-Bayesian network approach to assess the impact of extreme high-temperature events (heatwave events) on the developments of regional crops across the Canada agricultural regions at the ecodistrict-scale. Relevant weather variables and heatwave variables during heatwave periods have identified and used as input variables for model learning. Both a copula-Bayesian network and Gaussian-based network modeling approach is evaluated and inter-compared. The copula approach based on 'vine copulas' generated the most accurate predictions of heatwave occurrence as a driver of crop heat stress.

The last study presents a stochastic, hybrid-Bayesian machine-learning approach to explore the complex causal relationships between weather, pathogen, and host for grape powdery mildew in an experimental farm in Quebec, Canada. This study explores a high-performance network model for daily disease risk forecast by using estimated development factors of pathogen and host from recorded daily weather variables. A fungicide strategy for disease control has presented by using the model outputs and forecasted future weather variability.

The dissertation findings are beneficial to Canada's agricultural sector. The inter-related weather risks explored by the three separate studies in multi-scales provide a better understanding about the interactions between changing weather conditions, extreme weather, and crop production. The research show cases new insights, methods, and tools for minimizing risk in agricultural decision-making.

Contents

Supervisory Committee **ii**

Abstract **iii**

Table of Contents **v**

List of Tables **viii**

List of Figures **xiv**

Acknowledgements **xxi**

Dedication **xxii**

1 Introduction **1**

2 ENSO risk: Predicting crop yield variability and coherence using cluster-based PCA **6**

2.1 Introduction 7

2.1.1 El Niño-Southern Oscillation (ENSO) 7

2.1.2 Impact of ENSO on crop yields 7

2.1.3 ENSO forcing across the Canadian Prairies 9

2.1.4 ENSO direct and indirect impacts on crop yield distributions . 12

2.1.5 Modeling Objective 14

2.2 Materials and Methods 15

2.2.1 Standard PCA method 15

2.2.2 PCA extensions 16

2.2.3 Data sources 17

2.2.4 Cluster-PCA modeling framework 18

2.2.5 Model and predictor selection 19

2.2.6 Sensitivity and validation 21
2.2.7 Benchmarking 23
2.3 Results 24
2.4 Discussion 30
2.5 Conclusions 34

3 Spatial and temporal analysis of heatwave occurrence across the Canadian agricultural regions using a copula-Bayesian network model 35
3.1 introduction 36
3.1.1 Canada's agriculture regions and heat stress 36
3.1.2 Recent heat stress studies 37
3.1.3 Objectives 37
3.2 Materials and methods 38
3.2.1 Study site 38
3.2.2 Weather data 38
3.2.3 Normalized Difference Vegetation Index (NDVI) 40
3.2.4 Measuring crop heatwave stress 41
3.2.5 Copula-Bayesian networks 43
3.3 Copula-Bayesian Network Learning 46
3.3.1 Data processing 46
3.3.2 Network learning 48
3.3.3 Selection of copula distribution 50
3.3.4 Spatial and temporal analysis 52
3.4 Results 53
3.5 Discussion 64
3.6 Conclusions 65

4 Disease Risk Forecasting with Bayesian Learning Networks: Application to Grape Powdery Mildew (*Erysiphe necator*) in Vineyards 66
4.1 Introduction 67
4.1.1 Economic importance of grapes in North America 67
4.1.2 Grape Powdery Mildew (PM) disease 68
4.1.3 Impacts of weather on grape powdery mildew 69
4.1.4 Management of grape powdery mildew 69

4.1.5 Problem statement 71
4.1.6 Research objective 71
4.2 Materials and methods 72
4.2.1 Study site 72
4.2.2 Global Forecast System (GFS) Ensemble Reforecasts (GEFSR) 75
4.2.3 Grapevine development 77
4.2.4 PM disease development 77
4.3 Bayesian network learning model 82
4.3.1 Supervised and algorithm learning 83
4.3.2 Forecast skill under different learning modes 85
4.4 Model forecast evaluation 88
4.5 Model-based fungicide spray program 89
4.6 Results 90
4.7 Discussion 102
4.8 Conclusions 105

5 Discussions **107**

6 Achieved **112**
6.1 Publications 112
6.2 Leadership 112
6.3 Professional Development and Training 113

Bibliography **114**

List of Tables

Table 2.1 Prediction and cross-validated deviation (MAE) and discrepancy (RMSE) error (MAE - mean absolute error, RMSE - root-mean-squared-error) for cluster-PCA model optimized to each crop type (wheat and barley). These results show the relative gains in prediction skill obtained by considering spatial dependence, clustering at the CAR-scale, and the inclusion of ENSO as a predictor of crop yield. Values are reported to 3 significant figures. * indicates high error from leave-one-out cross-validation (LOOCV) indicating over-fitting. 25

Table 3.1 A set of candidate marginals probability distributions 49
Table 3.2 A range of candidate distributions for bivariate copula. 51

Table 4.1 Calibrated, site-specific parameter values for PM model. 82
Table 4.2 Variables for structural learning by the Grape PM model. 86
Table 4.3 Inter-comparison of model performance in parameter learning from both supervised and algorithmic learning in a total of 32 subsets of network random variables in four different model starting dates. The table shows the best model results of the network random variables and model start date. Besides, a binary factor of drought years (1 as in 2000, 2002, and 2008; 0 as in the rest of study years) was added as a parent to disease incidence to examine the model performance of disease incidence predictions in the hot years. Model performance has compared in mean-absolute-error (MAE) and root-mean-square-error (RMSE) from k-fold cross-validation, and model prediction skills in predicting disease incidence using observed weather variables in 2011. Smaller and non-negative values in both MAE and RMSE indicate higher model performance. 93

Table 4.4 Comparison of model performance of parameter learning starting from the first disease date for 8 subsets of network random variables. Forecast skill (MAE and RMSE) from k-fold cross-validation is listed. Smaller and non-negative values in both MAE and RMSE indicate higher model performance. The dash $-$ indicates there is no available model for forecasting DI using the selected network random variables. 93

Table 4.5 DI values for the forecast model-based fungicide spray program on August 27 in 2011. 102

List of Abbreviations

AAFC Agriculture and Agri-Food Canada

ADR Germination Rate of Ascospore

AI Artificial Intelligence

AIC Akaike Information Criterion

AMIS Agricultural Market Information System

AMR Rate of Ascospore Maturation

AOG Germinated Ascorespore

AUG Ungerminated Ascorespore

AvgPrcntAW Average Percentage of Variable Soil Water Holding Capacity

AvgSI Average Crop Water Stress

AVHRR Advanced Very-High-Resolution Radiometer

BN Bayesian Network

CAR Census of Agriculture Region

CBN Copula Bayesian Network

CCCAP Canadian Crop Condition Assessment Program

CGDD Cumulative Growing Degree Days

CLT Central Limit Theorem

CMAP CPC Merged Analysis of Precipitation

CP Central Pacific

CPC Climate Prediction Center

CSI Crop Stress Index

CTP Cumulative Total Precipitation

DAG Direct Acyclic Graphical

DI Disease Incidence

DR Dispersal Rate

EA Eastern Atlantic

EAWR Eastern Atlantic-Western Russia

EHF Excess Heat Factor

EHI Excess Heat Index

ENSO El Niño Southern Oscillation

EOF Empircial Orthogonal Function

EP Eastern Pacific

EPI El Niño Prediction Index

FPCA Functional Princial Component Analysis

GCM General Circulation Model

GDD Growing-Degree Days

GEFSR Global Forecast System Ensemble Reforecasts

GEOGLAM Group on Earth Observations' Global Agricultural Monitoring

GEV Generalized Extreme Value

GFS Global Forecast System

GLM General Linear Model

GS Grow-Shrink

HC Hill-Climbing

HWMId Heatwave Magnitude Index daily

ICCYF Integrated Canadian Crop Yield Forecaster

JECAM Joint Experiment of Crop Assessment and Monitoring

JRA-55 Japanese 55-year Reanalysis

LOOCV Leave-One-Out Cross Validation

LP Latent Period

MAE Mean Absolute Error

MLR Multiple Linear Regression

MME Multi-Model Ensemble

NAEFS North American Ensemble Forecast System

NAO North Atlantic Oscillation

NCEP National Centers for Environmental Prediction

NCR North Central Region

NDVI Normalized-Difference Vegetation Index

NOAA National Oceanic and Atmosphheric Administration

NRT Near-Real-Time

NWP Numerical Weather Prediction

OIV International Organization of Vine

ONI Oceanic Niño Index

P Precipitation

PAR Proportion of Ascospores Ready for Release

PC Principal Component

PCA Principal Component Analysis

PCR Principle Component Regression

PDO Pacific Decadal Oscillation

PIR Primary Infection Rate

PM Powdery Mildew

PNA Pacific North American

PS Plant Stages

R Relative Humidity

RCP Representative Concentration Pathway

RH Relative Humidity

RMSE Root Mean Squared Error

RSGA Remote Sensing and Geospatial Analysis Section

S14FD S14 Forcing Dataset

SCAND Scandinavian

SIR Secondary Infection Rate

SLP Surface Air Pressure

SST Sea Surface Temperature

T Temperature

TP Total Precipitation

TSS Total Sum of Squares

WD Wetness Duration

WS Wind Speed

List of Figures

Figure 1.1 Representation of main aspects of statistical modeling predictions. 3
Figure 1.2 Relationship of agricultural risk components according to the space and coverage. 5

Figure 2.1 (a) The Canadian Prairies in Western Canada (Provinces of Alberta, Saskatchewan and Manitoba) containing 40 Census of Agriculture regions (CARs) of varying area and boundary delineation. Crop wheat (b) and barley (c) yield (bu/ac) for CARs is from Statistics Canada's Field Crop Survey Reporting Program for 1987-2012. Lower bound of the grey shadow is the minimum yield of the year and the upper bound is the maximum yield. Solid line indicates the average annual yield of the year and the dotted line indicates the year of 2007. 10
Figure 2.2 (a) Variability in the ONI Index (ENSO 3.4 region, 3 month running sea-surface temperature (SST) mean anomolies) showing weak, moderate, strong and very strong ENSO historical events, 1987-2012. Variability in the field correlation (linear) between the ONI Index, and seasonally-averaged (1987-2007) air temperature (b) and precipitation (c) across the Canadian Prairies, impacting crop development (phenology and growth). Air temperature values are from NCEP Reanalysis data [70]. Precipitation values are from The U.S. Climate Prediction Center (CPC) Merged Analysis of Precipitation ("CMAP") dataset merges gauge and the five kinds of satellite estimates (not NCEP Reanalysis values), with monthly mean climatology based on 1979-1995 [143]. Source: NOAA/ESRL Physical Sciences Division, Boulder Colorado, https://www.esrl.noaa.gov/psd/data/correlation/. 11

Figure 2.3 Cluster-PCA statistical modeling framework that includes data transformation (standardization and detrending), K-means clustering, PCA applied at the Provincial and CAR aggregation scales, model selection, sensitivity analysis based on cluster number, cross-validation (hindcasting) and benchmarking of prediction skill. 18

Figure 2.4 Sensitivity of model yield (wheat and barley) (bc/ac) predictions to the number of clusters: (a) yield bias (MAE), (b) yield error variance (RMSE). The prediction error statistics are based on the full cluster-PCA model. 26

Figure 2.5 Predicted CAR-clusters from K-means clustering, for (a) wheat and (b) barley. Results for each subregion are shown in fraction format, where the numerator is the number of yield PC scores used to predict CAR yields and the denominator indicates the total number of PC scores. 28

Figure 2.6 Predicted yield *bias* and *variance* of CAR yield (bu/ac): (a) MAE of wheat, (b) RMSE of wheat, (c) MAE of barley, (d) RMSE of barley. Prediction errors of were computed using prediction yield from $M2$ with $k = 6$ for wheat and $M3$ with $k = 3$ for barley. Small values of MAE and RMSE indicate higher predictive skill. 29

Figure 2.7 Relative gains in yield prediction skill (bu/ac) by including ENSO forcing: (a) Predicted yield *bias* change (ΔMAE) (Wheat), (b) Predicted yield *variance* change (ΔRMSE) (Wheat), (c) Predicted yield *bias* change (ΔMAE) (Barley), and (d) Predicted yield *variance* change (ΔRMSE) (Barley). Positive values indicate a reduction in bias and/or error variance (improved yield prediction), whereas negative values indicate a gain in bias and/or error variance. 31

Figure 3.1 The eco-zones across Canada. The eco-zones in grey are the major agricultural farming areas of Canada. The black circle dots within the agricultural eco-zones are a set of 96 high activities agricultural regions. 39

Figure 3.2 A learning process of the paired-copula Bayesian network. Network variables in the hybrid Bayesian network structure are in discrete data types and have classified into levels for model learning except heat stress variable and NDVI. The causal relationship between heat stress variable and NDVI has learned with the paired-copula approach from the selected candidate distributions and the copula family distributions. 55

Figure 3.3 An example of simulation for NDVI probability predictions. Model simulation has performed using two classified heatwave types, heatwave magnitude with a 3-days in minimum days from HWMId (HWMI.M3) and precipitation with a 4-days in minimum days from CSI. Model simulations for NDVI probability predictions from CBN has compared to the empirical observations and a Gaussian-based Bayesian network. 56

Figure 3.4 Spatial analysis of the estimated starting date of heat stress risk across the eco-zones 7 and 8 (right side figures), and 9, 10, 13, and 14 (left side figures). The 'best' starting date has selected at every first day 01 (empty circles) and the fifth day 15 (solid circle) for the months of May (top), June (middle), and July (bottom). 58

Figure 3.5 A spatial representation of minimum days of heatwaves occurrences on regional crops across Canada. The minimum days has defined from a range of days from 2 to 7 in two-weeks observation intervals. Minimum days of 2-Days (empty circles), 3-Days (solid circles), 4-Days (empty triangles), and 6-Days(solid triangles) have found. A small value in minimum days indicates a regional crop has a low resistance to heatwave events and is more likely to has a higher than normal in heatwave frequency days 59

Figure 3.6 Extreme weather conditions on regional crops during heatwave actives. Extreme weather conditions were characterized by temperature (T), total precipitation (P), and relative humidity humidity (R). Significant weather conditions include temperature (empty circles), temperature and precipitation (solid circles), temperature and humidity (empty triangles), and temperature, precipitation, and humidity (solid triangles). 60

Figure 3.7 Spatial mapping of regional heatwave intensity (top) and frequency (bottom) for the agricultural eco-zones (eco-zones 7 and 8 (right sides), and 9, 10, 13, and 14 (left sides)). The values of heatwave intensity have estimated by averaged maximum daily temperature during heatwave actives. Both values of heatwave intensity and frequency have scaled over the study regions and plotted as circle points. A small size indicates a low heatwave intensity or frequency, respectively, have found over the study period from 1987 to 2012. The solid circles in the heatwave intensity plots indicate the region has a higher than 28 °C on the averaged maximum daily temperature. And the solid circles in the heatwave frequency plots indicate the region has a more than 100 heatwave days in annually. 62

Figure 3.8 Inter-comparisons heatwave intensity (top) and frequency (bottom) change between two five-year periods (1987- 1991 and 2008-2012). The values of change in heatwave intensity and frequency have computed by using the values obtained from the period from 2008 to 2012 minus the values obtained from another period from 1987 to 1991. The circle points in the plot indicate a decreased value has found in heatwave intensity and frequency, respectively. The triangle points indicate an increased value has found in heatwave intensity and frequency, respectively. The size of the points shows the different amounts. 63

Figure 4.1 The experimental farm (top) and the weather station (bottom) used to monitor weather variables at the grape canopy level from 2000 to 2011. 73

Figure 4.2 Seasonal weather variability and disease incidence (DI) (% infected/non-infected leaves) of grape PM through the growing season for northern hybrid grape cultivars with differing susceptibility (Chancellor, Geisenheim-318, Frontenac). The mean (solid line) is shown varying between minimum (lower bound) and maximum (upper bound) observed values (2000-2011). 74

Figure 4.3 Annual first disease date (top) and the maximum DI (bottom) of grape PM for the three susceptible cultivars: Chancellor (circle); Geisenheim-318 (square); and Frontenac (triangle). 76
Figure 4.4 Temperature effect rate plot of the developing rate of PM in response to daily temperature from 6 °C to 32 °C. 79
Figure 4.5 Number of days needed to complete a latent period responses to daily temperatures (°C). The solid line represents the estimated time to complete a latent period from Equation 4.9 in relation to measured days to complete a latent period. 81
Figure 4.6 DAG representation of Bayesian network model structure identified by *supervised* learning of grape PM. The causal relationships between the variable were linked by existing empirical and published scientific peer-reviewed knowledge on the interactions between the weather, host, and pathogen. Variables in the DAG representation includes: Plant stage (PS), Total precipitation (TP), Daily mean temperature (Tmean), Wind speed (WS), Degree-days based risk assessment model $P_{maxacc3}$ (3°C), Latent period (LP), Primary infection rate (PIR), Secondary infection rate (SIR), Dispersal rate (DR), Recent disease incidence (DI_P), Cultivar (Type), and Disease incidence (DI). 94
Figure 4.7 DAG representation of Bayesian network model structure learned by *algorithmic* learning. The structure was learned from bootstrap sampling technique under 5000 iterations with arc strength above 0.8 and arc direction above 0.5. Variables learned in this Bayesian network are a combination of observed weather variables and the estimated development factors of grapevine and the pathogen of grape PM. Variables in the DAG representation includes: Plant stage (PS), Total precipitation (TP), Daily mean temperature (Tmean), Wind speed (WS), Degree-days based risk assessment model $P_{maxacc3}$ (3°C), Latent period (LP), Primary infection rate (PIR), Secondary infection rate (SIR), Dispersal rate (DR), Recent disease incidence (DI_P), Cultivar (Type), and Disease incidence (DI). 95

Figure 4.8 Model forecast skill from k-fold cross-validation (2000-2010) and 2011 year validation in both supervised (circle) and hill-climbing based algorithmic learning (triangle). MAE and RMSE were computed for the three susceptible grape cultivars in the testing year. Smaller values in both MAE and RMSE indicates higher forecast skill. 97

Figure 4.9 Model predictions of DI of PM for the three grape cultivars: High susceptible Chancellor (Top), Medium susceptible Geisenheim-318 (Medium), and the low susceptible Frontenac (bottom) in 2011. Model predictions from both supervised (dotted line) and algorithm (dashed line) learned Bayesian network are shown alongside the observed daily DI. The three vertical lines from left to right are the estimated plant stage of grapevine: flowering, setting, and veraison stages. DI of grape PM in 2011 started from June 29th until the end of the growing season. 98

Figure 4.10 Sensitivity analysis of model predicted DI for Chancellor (top), Geisenheim-318 (middle), and Frontenac (bottom) cultivars, by changing temperature by $2°C$ (warm and cold year scenarios). Daily temperature in 2011 is shown as a reference baseline (solid line). 99

Figure 4.11 Sensitivity of DI prediction error to forecast window size (1-16 days), measured using average MAE and RMSE, under supervised learning and GEFS weather input. 100

Figure 4.12 Fungicide spray programs of the UC-Davis and the degree-days risk assessment model to grape PM in 2011. The dotted line indicates the threshold-based UC score from the UC Davis model. The long dashed line indicates the schedule-based model scores from the degree-day model. The solid line indicates the DI of Chancellor (top); Geisenheim-318 (median); and Frontenac (bottom). 101

Figure 4.133D plot of the fungicide spray strategies of Chancellor (upper-left), Geisenheim-318 (upper-right), and Frontenac (lower center) from the forecast model on August 27th in 2011 for six fungicide spray dates and corresponding 10-day DI based on the 16-day GEFS reforecast weather input. (Lower right) 2D plot of the 10-day daily DI of the optimal spray date for Chancellor (dotted line); Geisenheim-318 (dot-dash line); and Frontenac (long-dash line) based on DI on August 27th 2011 and 16-day GEFS input. The solid line indicated the daily cumulative total precipitation from GEFS on August 28th (2011). Note: Darker color indicates higher disease severity. 103

Figure 5.1 A general overview of analytical approaches and how they were employed in the various models. 110

ACKNOWLEDGEMENTS

Firstly , I would like to express the deepest appreciation to my advisor Dr. Nathaniel Newlands for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. I could not have imagined having a better advisor and mentor for my Ph.D study. He is one of the few who change my life and bring me to the applied science research field. Thanks for his guidance for training me as an independent research scientist to handle researches, to think deeply, critically, and closely about what agriculture producers want and care. I cannot make such a success without him.

Besides my main supervisor , I would like to thank Professor Dr. David Atkinson and Dr. Alex Cannon became my university supervisor and thesis committee at the University of Victoria and for their valuable comments, insights, and guidance, but also for the hard question which guides me to widen my research from various perspectives. Thanks Dr. David Atkinson for supports of the weather and atmosphere and weather station installation training for my future research needs. Thanks Dr. Alex Cannon for the R coding help and weather data support.

I would also like to owe my deepest gratitude to all the friends across the cities of Lethbridge (Alberta), Victoria (BC), Kelowna (BC), and Summerland (BC) as well as the two institutions of Agriculture and Agri-Food Canada (AAFC) and University of Victoria, the climate lab members (Adam, Vida, Mohammed, Chris, Laura, Eric, and Ben), and my family for their help and support along the way of my PhD. Without their persistent help this dissertation would bot have been possible.

This dissertation was funded in part by the by the Growing Forward One Federal research program (Agriculture and Agri-Food Canada, AAFC) (Project No. J-000179.001.02) and assistance of Canada's Federal Research Affiliate Program (RAP).

Weixun Lu, Kelowna, BC, Canada

DEDICATION

My beloved parents,
Vanessa Zhang,
And all my friends,
For your love, endless support, and encouragement.

Chapter 1

Introduction

Canadian agriculture production is highly vulnerable to impacts associated with extreme weather events of high precipitation, hail storms, heatwaves, and meteorological droughts. A wide range of annual and perennial crops are grown across Canada, with the main crops being wheat, corn, soybean, and canola. The relative proportion of cropland by region is 4% in British Columbia, more than 80 % on the Canadian Prairie provinces (31% in Alberta, 38% in Saskatchewan, and 11% in Manitoba), 5% in Quebec, 8% in Ontario, and 3% elsewhere. Weather conditions are warm and wet around the two coastal areas of Canada. Summer seasons on the Canadian Prairies are usually dry and hot with high rainfall during short but intense summer storms. Both the developments of crop plants and farming practices are vulnerable to the impacts of long-term climate change and short-term weather variability. Long-term climate changes are raised by long-term atmospheric behaves (typically 30 years or more). Long-term climate risks such as rising global land temperature, rising sea level, and long-term changes of precipitation (drought and flood years) impact agriculture by lengthening growing seasons and changing land suitability for agriculture. Despite longer-term anticipated changes due to climate change changes to agricultural practices, agricultural producers tend to be more concerned by potential impacts of weather variability across shorter time-scales (i.e., daily, weekly, monthly, seasonal). Crops are sensitive to different types of weather-related stress at different phenological stages. Risks are also inter-related. For example, seasonal climate variability (e.g. ENSO) affects the probability of heatwaves, which can determine probabilities of short-term extreme weather events (heatwaves), which can influence disease occurrence, and in turn, depends on crop phenological stage. Farmers and other agricultural stakeholders require models that can reliably forecast potential agricultural risks and their associ-

ated impacts to provide sufficient lead time to effectively respond to changing weather conditions, extreme weather events, and crop disease.

Machine learning statistical models have become efficient approaches to provide supporting risk forecast information for improving decision-making and farming strategies. Machine learning is an application of artificial intelligence (AI) that provides computer systems the ability to automatically learn and improve from experience without being explicitly programmed. In this thesis, machine-learning was applied to improve crop protection in relation to the causal relationship and uncertainty of environmental, agronomic, crop and disease variables. Fig. 1.1 shows the main aspects of hindcasting and forecasting. The performance of a statistical model can be varied by the bias of historical observation and model training windows (study period). This allows independent assessments to be made of the influences of weather events on the development of regional crops as well as the approaches used in model learning. Model learning approaches can be classified based on their roles in data processing, which consist of: data generation (timing processing, such as daily, weekly, and monthly), geospatial classification (spatial processing), modeling learning, inference learning, model validation, sensitivity testing, and model predictions. A statistical model with a high prediction performance (or 'skill') is then able to provide reliable predictions for hindcasting and forecasting using historical observations and scenario forecasts about future observations, respectively. Model prediction errors are higher in forecasting than in hindcasting because weather events can occur that are outside of the model training exposure and so difficult for the model to fully capture.

In this dissertation, I examined several machine learning statistical approaches to assess three weather-based agricultural risk components, including: seasonal climate variability (El Niño-Southern Oscillation (ENSO) in Chapter 2), short-term extreme weather impacts (heat stress in Chapter 3), and emergent disease (grape powdery mildew in Chapter 4) for summer crops grown in Canada. In each component different type of risks are considered for different crop types; these components are interconnected by explicit consideration of the uncertainty of weather risks through the growing season. The relationship between the three agricultural risks considered in this dissertation according to their space and time coverage is shown in Fig. 1.2.

ENSO is a large, naturally-occurring climate variability mode that affects the seasonal weather conditions (temperature and precipitation) at locations around the globe, including the Canadian Prairies. Spatially, ENSO influences often show up spatially as irregularly-shaped patterns. An ENSO cycle occurs over 2-5 years and

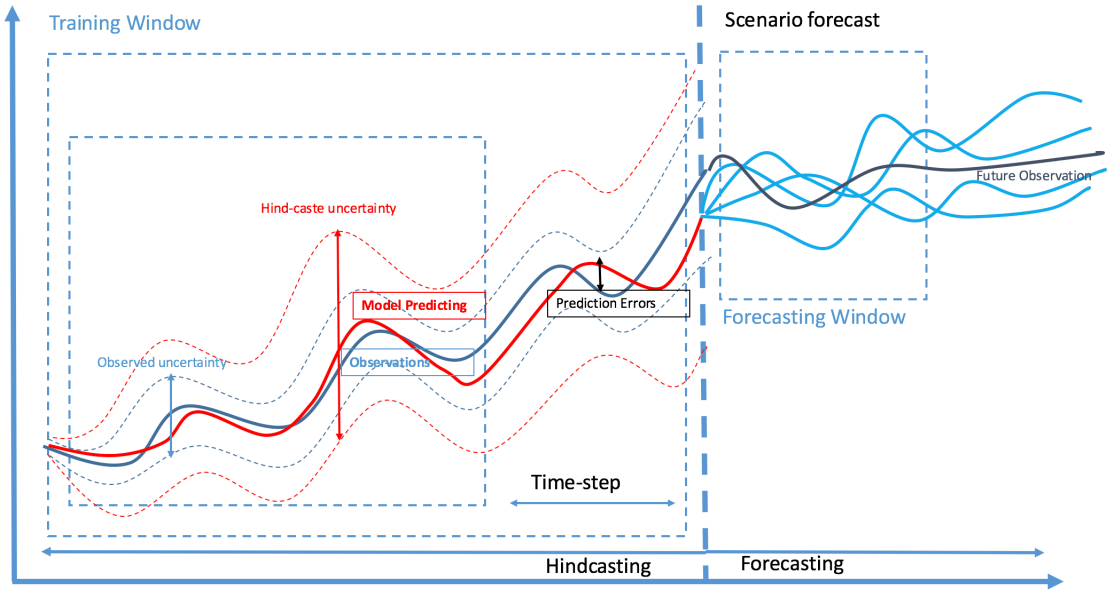


Figure 1.1 Representation of main aspects of statistical modeling predictions.

proceeds through three phases: El Niño (*warmer than normal*), La Niña (*cooler than normal*), and a neutral phase during which it is moving between warm and cold phases. Data sources for this study include monthly mean sea surface temperature (SST) and monthly NDVI and agro-weather indices for each growing season (May to August) to predict seasonal crop yield (wheat and barley) over the prairies. The spatial domain for this study are the 40 census agriculture regions reported from the 2011 agricultural census. For these regions monthly NDVI and agro-weather indices were obtained. ENSO data are provided as an index of ENSO strength, which is based on sea-surface temperature data from a specific region within the primary area of ENSO activity (ENSO 3.4 regions, central south Pacific).

A heatwave event is an extreme weather event defined as a period of elevated temperature lasting more than a few days occurring on a spatial scale of 10s to 100s of kilometers. Its impact on plants, termed 'heat stress', can represent an adverse affect resulting in a potential significant reduction to the final yield. Crop yield loss can be complete when high-intensity heatwaves occur a few days before and after the flowering stage of crop growth. Detecting and responding to heatwaves is complicated by the fact that there are a variety of algorithms for detect an agriculturally-relevant heatwave event, leading to a misunderstanding of the impacts of heat stress on crop growth. Heatwaves in this study are identified using five heatwave indices based on

daily weather variables of temperature and precipitation over the growing season. Data sources include weekly NDVI and daily weather variables, focusing on a spatial domain that includes 96 census agricultural regions across Canada's major agricultural lands.

Powdery mildew (PM) is a fungal disease of grapevines that can grow rapidly in summer, given the right weather conditions, and which can also survive over-winter. This can cause grape yield reductions for years. The development of this fungal disease is driven by hourly changes in temperature, precipitation, and humidity. Anticipating the endemic and epidemic spread and effective growth of disease spores is the major challenge for developing efficiency strategies in disease control. Disease inflection data in this study are collected weekly over the growing season from an experimental farm in Quebec, Canada. Historical weather observations are obtained in hourly resolution and are reduced to daily for risk modeling.

Ultimately, reliable assessment of risks assessments from statistical models can deliver benefits to Canada's agricultural sectors to help crop managers improve crop protection by better anticipating and reducing uncertainty of weather impacts on crops. All statistical modeling work for the three research components is coded and implemented using the statistical programming language R (version 3.1.3 in the ENSO risk; version 3.5.3 in the heatwave risk; and version 3.4.2 in the grape powdery mildew risk research).

This research provides innovative mathematical and statistical methodologies, along with theoretical and applied insights, for quantifying agricultural risks due to weather variability, extreme events, and disease. Computer codes enable the implementation, further validation and extension use of these novel methodologies within the broader scientific community and agricultural sector to improve and enhance agricultural decision support. The three components of this research on agricultural risk explored the following questions:

- Short-term climate variability risk - ENSO variability (seasonal)
    1. Does ENSO have a significant influence on crop growth and yield in Canada?
    2. How does the impact of ENSO vary spatially?
    3. How "best" to predict ENSO's impact on crops in time and space?
- Extreme Event risk - Heatwave events (weekly)

1. What is the "best" way to characterize heatwaves in time and space?
2. How can heatwave impact crop growth and crop-weather response?
3. How "best" to forecast crop heat stress on crops?

- Disease risk - Grape powdery mildew (daily)

1. How "best" to forecast the risk of powdery mildew in time and space?
2. When is the "best" time to start model forecasting of powdery mildew?
3. When is the "best" to apply fungicide spray in disease control for different grapes over the growing season.

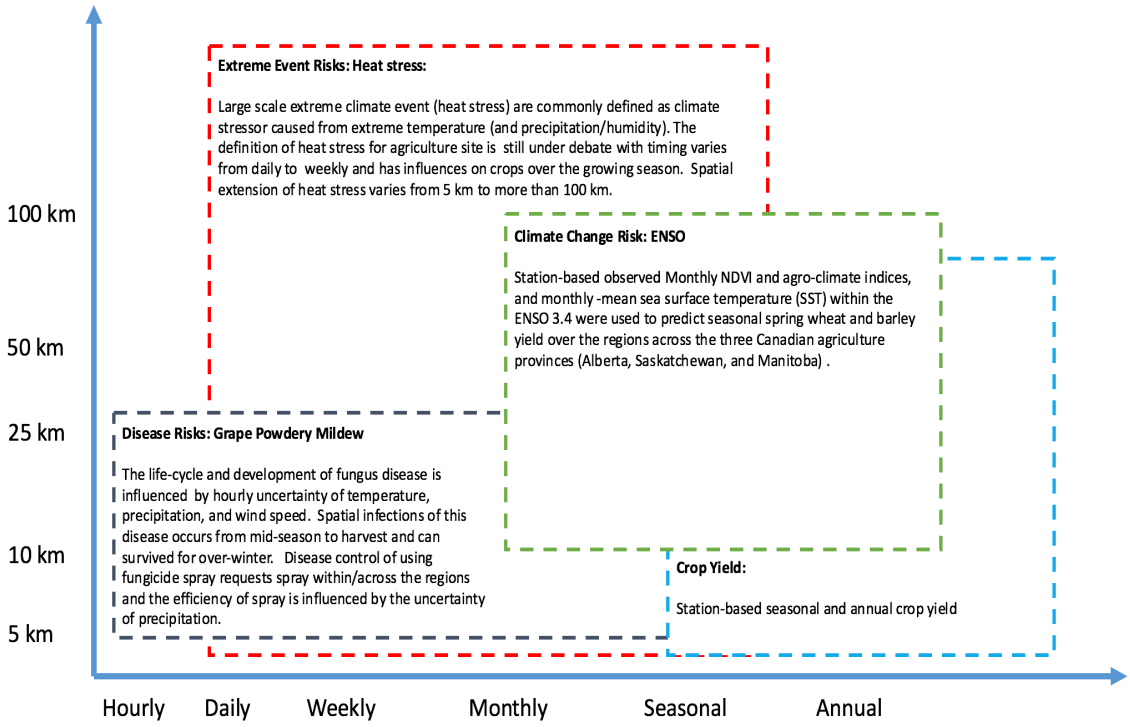


Figure 1.2 Relationship of agricultural risk components according to the space and coverage.

This original research has been published:
[79], [80]

Chapter 2

ENSO risk: Predicting crop yield variability and coherence using cluster-based PCA

The El Niño-Southern Oscillation (ENSO) has, in recent years, contributed to increases in the yields of major agricultural (annual) crops like wheat and barley in Canada. How such forcing alters the pattern of yield variation across different geographic scales and across large agricultural landscapes like the Canadian Prairies is less understood. Yet, such questions are of major importance in forecasting future cereal crop production. In this chapter, we explore the potential impact of ENSO on wheat and barley across the Canadian Prairies/Western Canada using a multi-scale, cluster-based Principal Component Analysis (PCA) model that integrates machine-learning (K-means clustering) to predict areas of high seasonal climate risk, These risk areas are separable clusters of subregions that show similar ENSO-yield correlation response (spatial coherency). Benchmarking this spatial model to non-spatial models indicates that spatial coherency leads to gains in prediction skill. Incorporating spatial coherency increased the skill in predicting crop yield; reducing RMSE error by up to 26-34% (spring wheat) and 2-4% (barley). We infer that accounting for spatial coherency improves the accuracy and reliability of crop yield forecasts. The main result of this chapter has been published in *Modeling Earth Systems and Environment*[1], see [79].

[1]https://link.springer.com/article/10.1007/s40808-017-0382-0?shared-article-renderer

2.1 Introduction

2.1.1 El Niño-Southern Oscillation (ENSO)

ENSO is a naturally occurring, large-scale internal model of climate variability, whereby changes in sea surface temperature (SST) through air-sea coupling, result in changes to atmospheric pressure patterns over the tropical Pacific [110]. ENSO is the leading mode of climate variability at interannual timescales. It starts in the tropical Pacific, but interacts with pressure patterns away from its local source of action (i.e., teleconnections), exerting a broad impact on the hydrological cycle, seasonal weather (e.g., temperature and precipitation) and weather extremes around the globe. ENSO fluctuates between warm, cool and neutral phases: *warmer than normal* (El Niño), *cooler than normal* (La Niña) central and eastern equatorial Pacific SSTs, or average conditions that are more controlled by other climate teleconnections such as the North Atlantic Oscillation (NAO) and Pacific North American (PNA). In the Pacific region, two types of ENSO have been identified (i.e., eastern Pacific (EP) and central Pacific (CP) types). While EP has anomalous centres of action (termed field significance) representing significant correlation and stationarity attributable to ENSO forcing, unlike CP [146]. In the Atlantic region, significant non-stationarity in ENSO forcing and land/crop impacts is also evident, particularly in the North Atlantic eastern region, where its impacts are still poorly understood and controversial [110]. The geographic variation of ENSO impact depends on various factors affecting the timing, intensity and duration of the events. Significant non-stationarity in ENSO anomoly patterns (i.e., drier and wetter areas), is attributed to the modulating influence of Pacific Decadal Oscillation (PDO). When ENSO and PDO are in phase (ENSO and PDO in warm phase), anomalies generally intensify and expand poleward inducing broader and more severe droughts, while when they are out-of-phase (PDO in cold phase), anomolies dampen and dissapate with reduced drought risk [137]. Precipitation patterns across North America show marked changes between La Niña neutral and El Niño winters [19].

2.1.2 Impact of ENSO on crop yields

The ENSO of 2015-16 was one the strongest in the last 30 years. Agricultural impacts included a 5% decrease global wine production (perennial grapevines)[2], but an increase

[2]International Organization of Vine & Wine (OIV)

in yields for major annuals like wheat, maize/corn, rice and soybeans[3]. Even though climate variation associated with teleconnection activity can account for as much as a third of crop yield variability worldwide, the particular response varies greatly by crop type as well as with ENSO phase [106, 62]. European regional climate trends have resulted in stagnated yield increases of ∼10% in Europe over the last two decades: average yields of both wheat and barley were ∼0.12 tons/ha/year during the 1980s and would be 3.5% and 3.8% higher today, if such increases had continued [89]. Crop yields in many regions of the world have become much more variable, and future changes in weather conditions, more frequent and intense extreme weather events, and crop disease risk could drive continued yield variability and loss [63]. A recent agricultural economic review and analysis indicates that cereal production in both tropics and temperate countries is strongly coupled to ENSO behaviour [59]. This analysis measured the dominant state of ENSO in each calendar year by averaging the monthly NINO3.4 Index May-December to construct an annual ENSO index [59], finding that an +1°C increase in the ENSO Index lowers cereal yields by 2 %, total cereal production by 3.5 %, and agricultural income by 1.8 % on average across the tropics, while they rise in temperate countries when the tropical Pacific warms, albeit with a smaller magnitude that is less significant (i.e., increases of 1.7% yield, 2.4% production and 1.6% agricultural income).

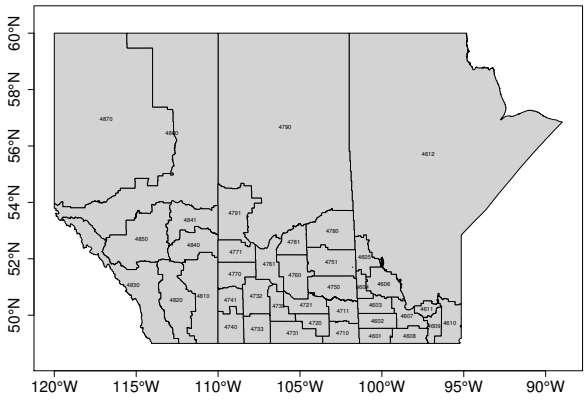
Large-scale atmospheric circulation (i.e., North Atlantic Oscillation (NAO), Eastern Atlantic (EA), Scandinavian (SCAND) and Eastern Atlantic-Western Russia (EAWR) patterns) explain on average 43% of inter-annual winter wheat yield variability, ranging between 20% and 70% across Europe [25]. An analysis of crop yield response to extreme heat stress under plausible weather change futures reveal significant regional disparities in crop yield impacts are likely to occur, with yield increases at high latitudes, but decreases in mid to low latitudes [33]. This suggests that improvements in regional crop mixing practices and the breeding of "weather-resistent" cultivars should be undertaken with some urgency [147]. In tandem, by increasing the prediction skill in how crops respond to weather variability, the robustness and reliability of operational crop seasonal forecasts and outlooks is increased, making them more useful and informative for agricultural decision-making under adaptation.

[3] www.foodsecurityportal.org

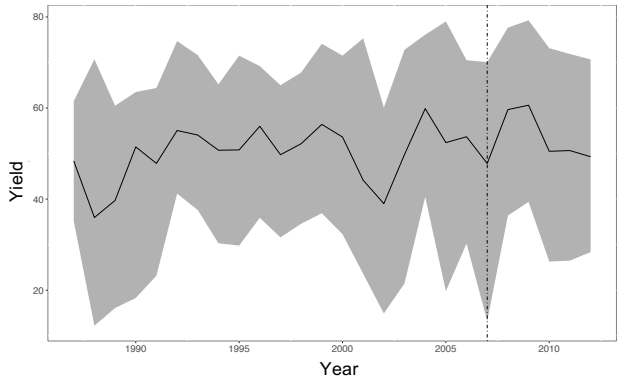
2.1.3 ENSO forcing across the Canadian Prairies

The Canadian Prairies, which include the Canadian provinces of Alberta, Saskatchewan and Manitoba is the center of Canada's cereal production (Fig. 2.1). This region has exhibited considerable variability in yields of wheat (e.g., spring wheat) and barley. Previous work has indicated that several internal modes of climate variability affect yield on the Prairies, including ENSO, the Pacific Decadal Oscillation (PDO); the Pacific North America pattern (PNA); and the North Atlantic Oscillation (NAO), however, ENSO has an especially strong influence on droughts in the Prairies [132, 9, 10, 52, 121, 104, 120]. As highlighted by [125], forecasting involves assumptions on the future behaviour of an agroecosystem (and embedded hydrological system) from dynamics that has been observed to exhibit in the past, generally assuming that the weather and hydrological system does not change over forecast time-window and that historical observation data provides a good baseline for future variability. Spectral (i.e., frequency-based) analysis of reconstructed streamflows (650 years) from moisture-sensitive tree ring chronologies in the Prairies show a highly significant multi-decadal ($\sim$ 65 years) component of variability, together with significant variability at inter-annual time-scales (2-6 years) in the ENSO band, and while the recent past (i.e., 20th century) is representative of drought frequency over the long term, there are droughts of *greater* severity and especially duration in the preinstrumental proxy record [117, 125, 6].

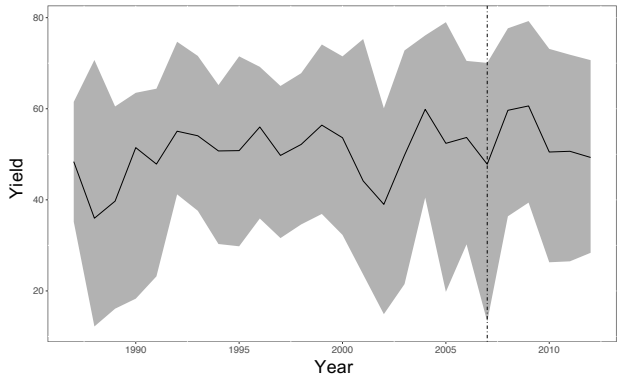
The potential for ENSO forcing on North American climate can be mapped based on the correlation of air temperature and precipitation with the Oceanic Niño Index (ONI) - a three-month running-mean of SST anomalies, based on centered 30-year base periods updated every 5 years. The ONI index is the area-averaged anomalous SST within the Niño 3.4 Region defined by 5°S-5°N, 170°-120°W (Central Pacific), and is considered a principal measure for monitoring, assessing and predicting ENSO used by operational centres worldwide [124]. Figure 2.2 shows the variability of the ONI index over time, and the resulting spatial correlation pattern across the Canadian Prairies in relation to seasonally-averaged temperature and precipitation (1987-2007). SOI has been identified as strong indicator of *between*-season component of yield variability arising from atmospheric variability, whereby strong El Niño have been found to be preceded by a standing wave in planetary Rossby waves within the Southern Hemisphere which are coupled to SOI variability. SOI is defined as the difference in surface air pressure (SLP) between Tahiti and Darwin, Australia, and is an atmospheric based measure of the intensity or strength of the Walker Circulation.



(a) The Canadian Prairies

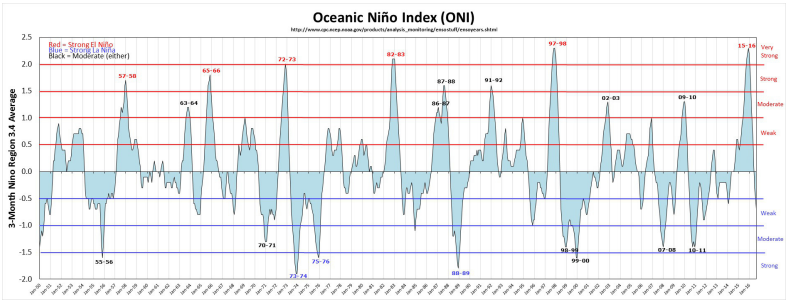


(b) Historical variability in spring wheat yield.

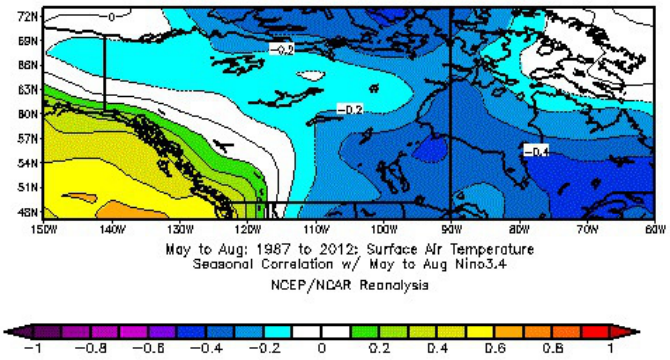


(c) Historical variability in barley yield.

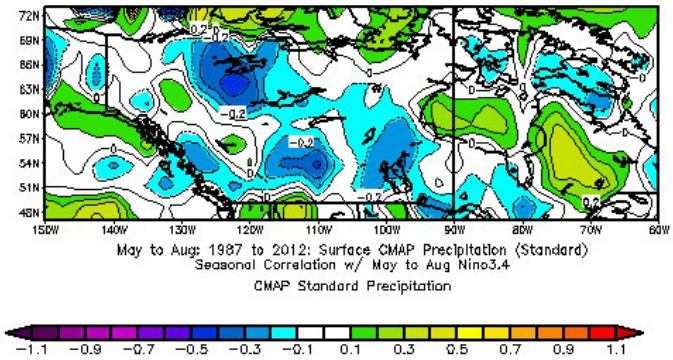
Figure 2.1 (a) The Canadian Prairies in Western Canada (Provinces of Alberta, Saskatchewan and Manitoba) containing 40 Census of Agriculture regions (CARs) of varying area and boundary delineation. Crop wheat (b) and barley (c) yield (bu/ac) for CARs is from Statistics Canada's Field Crop Survey Reporting Program for 1987-2012. Lower bound of the grey shadow is the minimum yield of the year and the upper bound is the maximum yield. Solid line indicates the average annual yield of the year and the dotted line indicates the year of 2007.



(a) The ONI Index



(b) Correlation between ONI and seasonal air temperature



(c) Correlation between ONI and seasonal precipitation

Figure 2.2 (a) Variability in the ONI Index (ENSO 3.4 region, 3 month running sea-surface temperature (SST) mean anomolies) showing weak, moderate, strong and very strong ENSO historical events, 1987-2012. Variability in the field correlation (linear) between the ONI Index, and seasonally-averaged (1987-2007) air temperature (b) and precipitation (c) across the Canadian Prairies, impacting crop development (phenology and growth). Air temperature values are from NCEP Reanalysis data [70]. Precipitation values are from The U.S. Climate Prediction Center (CPC) Merged Analysis of Precipitation ("CMAP") dataset merges gauge and the five kinds of satellite estimates (not NCEP Reanalysis values), with monthly mean climatology based on 1979-1995 [143]. Source: NOAA/ESRL Physical Sciences Division, Boulder Colorado, https://www.esrl.noaa.gov/psd/data/correlation/.

A significant relationship between ENSO and wheat yield across Australia has been identified whereby SOI (April/May phase) is a reliable indicator of *subsequent*-year yield variability, discriminating wheat spatial clusters well (i.e., homologous modes or zones of wheat yield variation each containing specific analog years or year types) [103, 126]. SST variability has also been shown to have a dominant forcing on *within*-season yield variability [103, 60, 126]. SST across a three-month seasonal interval is able to also account for shorter time-lags in SOI variability [62]. The ONI has a similar shape to the SOI, but is of opposite sign and SST and SLP are strongly coupled (i.e., not just collinear, but causally linked). Moreover, a statistical analysis of the causal linkage detectable between ENSO SST and SLP (i.e., SOI) variability provides evidence of this significant causality between these predictors exists, and that multi-collinearity can extend to lead times of 7-22 months [96]. This finding is also congruent with the ability of operational El Niño Prediction Index (EPI) that provides lead times of 12-15 months before El Niño peaks, which occurs within this maximum lead time range of 7-22 months within which significant Granger causality between SST and SLP persists. El Niño conditions are characterized by a positive ONI greater than or equal to +0.5°C, while La Niña, is characterized by a negative ONI less than or equal to -0.5°C, such that conditions must exceed these thresholds for a period of at least five consecutive overlapping three-month seasons to be identified as ENSO events. Seasonal correlation maps of ONI with summer regional climate across Western Canada (1987-2007) show anomalous centres of action (termed field significance) that represent significant correlation and stationarity attributable to ENSO forcing (Fig. 2.2). A similar anomalous correlation pattern is shown in detailed correlation maps of ONI with atmospheric circulation (represented by 500 gPa geopotential Φ_{500} anomolies), surface temperature, and precipitation anomalies [146].

2.1.4 ENSO direct and indirect impacts on crop yield distributions

Crop models that are process-based tend to be complex, involving a large number of parameters and variables. This complexity makes this set of model more difficult to calibrate, and often constrains their use and applicability [16]. Crop models that are statistically-based offer an alternative approach for predicting crop yield by harnessing the power of variable-reduction and optimization methods to quatify and reduce multi-collinearity, autocorrelation and spatial correlation biases and calibration uncertainties,

and incorporate different types of data in different ways to increase prediction skill. Statistical models are, though designed to fit available data (or simulated scenario sets) being highly constrained by such inputs, whereas process-models are more constrained by their process descriptions, interactions and underlying assumptions [138]. A detailed analysis of the impact that data resolution and aggregation can have on regional-scale crop yield model predictions finds that prediction error using aggregated soil data can often be larger than the inter-annual yield variability or differences between models [57]. This study highlights the importance of modeling nonlinear effects that arise from distinct yield patterns that depend on climate conditions and plant available water capacity. Statistical modeling using geographically weighted PCA has previously shown that the influence of Pacific sea-surface temperature (SST) across the Canadian Prairies can inform yield prediction, because low yields seem more predictable than high yields, and that the weather-crop relationship is nonlinear due to asymmetry in the response of ENSO (i.e., cooler and wetter La Niña has a more significant influence on yield than warmer, drier El Niño conditions [60]). [50] also previously identified distinctly different profiles of accumulated monthly ENSO and PNA index for growing season extremes (i.e., hottest and coldest or driest and wettest summers) over the Canadian Prairies and 2-7 month lead times for operational wheat yield forecasting in the Canadian Prairies using multiple (linear) regression analysis (MLR). The seasonal mean Niño 3.4 index is well predicted in a multi-model ensemble (MME), up to a 4-month lead time, but coupled models have particularly low skill in predicting the global SST pattern during weak ENSO events [124]. While there is strong prediction skill generally < 2 years in advance of an ENSO event, stochastic variation dominates at longer time-horizons.

Higher moments of the crop yield distribution characterise the so-called *elasticity* of crop yield. When temperature has a stronger influence on crop yield than precipitation, or when upper extremes are greater than lower extremes, the distribution becomes more 'asymmetric', changing its elasticity [86]. In representing the crop yield distribution, yield elasticity is typically superimposed on an underlying trend in average yield arising from gradual improvements in crop genetics and harvesting technology, including changes in harvested area, crop prices and subsidies. Highs and lows in wheat and barley yield have been demonstrated to coincide with the timing (i.e., occurrence) of ENSO events, but the yield (i.e., crop response under ENSO events of different intensity and duration) shows marked differences in relative average yield change and interannual trends (autocorrelation extent) even in cereal crops (i.e., wheat versus

barley). Barley, for example, is considered better adapted than wheat to these areas because of its *lower* heat unit requirement to reach maturity [102]. Barley is also more sensitive than wheat to daylength changes because it makes more efficient use of photothermal resources for leaf production [35]. Barley also has a smaller phyllochron interval than wheat and attains anthesis/maturity earlier, with an ability also to rapidly fill and ripen under cool conditions. Cereals and oilseed cultivars respond differently depending on conditions, species and variety, agronomic characteristics (e.g., heat units for maturity, shatter resistance, and oil content in the case of oilseeds) are well-suited to the soil type and lower soil moisture/drier conditions [86]. Different crop types, species and cultivar varieties also differ in their susceptibility and resistance to disease. The timing, spread and spatial patterning of disease outbreaks can further confound the attribution of particular crop impacts to climate variability. For example, the inter-annual variations in the emergence and severity of wheat stripe rust *Puccinia striiformis* f. sp. *tritici* disease within the Pacific Northwest of the United States and Western Canada are reportedly significantly lower during El Niño years attributed to decreased annual precipitation and increased winter and spring frost days. While wheat rust disease and SOI show coherent (consistent and significant) patterns with SOI variability at temporal scales matching ENSO (2-10 yrs), such coherence is out-of-phase, suggesting that a latent, causal factor exists [118].

2.1.5 Modeling Objective

In this paper, we apply a statistical modeling technique called cluster-based Principal Component Analysis (PCA) to investigate the influence of ENSO on wheat and barley crop yield across the Canadian Prairies. This involves coupling K-means clustering with the PCA dimension/variable-reduction technique to include spatial dependence and benchmark its prediction skill against existing Multivariate Linear Regression (MLR) model and Principle Component Regression (PCR) models. We hypothesized that crop response to climate variability in this region exhibits distinct spatial patterns of yield trends (i.e. distinct clusters of coherent regions) where crop yield can be predicted well, and other areas where prediction is less skillful. Previous evidence of such a spatially heterogeneous correlation pattern or coherence between ENSO and crop yield exists [104]. Coherency can be used to help infer physical mechanisms that may link teleconnection forcing with crop yield, taking place at finer spatial scales. After identifying the best performing model for each crop type, we profile

the coherence pattern based on yield prediction output associated with and without ENSO forcing.

2.2 Materials and Methods

2.2.1 Standard PCA method

Principal component analysis (PCA) is a multivariate analysis technique for dimension reduction. Its goal is to extract the most important information from presumably non-independent variables, in order to compress the size of the data set, simplify its statistical description, and analyze the structure of the observations and the variables [1, 68, 8]. PCA applies an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearized, uncorrelated variables called principal components (PCs). An individual PC is therefore a linear combination of optimally-weighted observed variables, and the total number of principal components is less than or equal to the number of original variables. Generally, the first component captures the primary modes of variation common to the greatest number of variables. This pattern is removed and the second component captures the next most widespread mode of variation. This continues until a number of components equal to the original number of variables has been identified [34, 131]. PCA makes no assumption about an underlying causal model. When a data distribution is skewed or there are outliers, data transformation is needed, with the choice of the transformation being informed by previous studies, prior knowledge, or trial-and-error. Often, the Box-Cox power transformation that includes the logarithm, square root, and multiplicative inverse as special cases is applied [81].

To summarize more formally; let the original, observed data and a transformed data matrix be denoted $\mathbf{X}$, and $\mathbf{Z}$, respectively. To standardize the data across the different variables a data transformation was applied (Gaussian distribution-based) to standardize it across the different variables,

$$z_{ij} = \left(\frac{x_{ij} - \mu_j}{\sigma_j} \right), \tag{2.1}$$

where μ_j is the mean and σ_j is the standard deviation, and z_{ij} and x_{ij} are the scaled and original data for row i and column j. Both data matrices have the same matrix order with rows representing years of observation and columns representing crop yield

or climate variable. The main PCA equation associated with this transformed data matrix $\mathbf{Z}$ ($m \times p$) is then:

$$\mathbf{F} = \mathbf{ZQ}, \tag{2.2}$$

where $\mathbf{F}$ is an $m \times p$ matrix of principal component *PC scores*, $\mathbf{Q}$ is the $p \times p$ matrix of *PC loadings*, where the bilinear decomposition of Z is given by,

$$\mathbf{Z} = \mathbf{FQ}^{-1} \tag{2.3}$$

where the exponent '-1' denotes the inverse of the matrix $\mathbf{Q}$.

2.2.2 PCA extensions

There are several key methodological extensions to PCA: Primary examples include the following. *geographically-weighted PCA* generates space-time empirical orthogonal functions (EOF) that decompose a signal or data set in terms of orthogonal basis functions [73, 114, 55], principal component regression (PCR) is a regression approach using PCs instead of original predictor variables [16, 68], *functional PCA* (FPCA) is an approach that extends standard PCA to a set of functions using regularization to account for underlying smoothness [105], and *cluster-based PCA* that extends PCA by including *unsupervised* (e.g., K-means) or *supervised* (agglomerative, hierarchical) clustering of the components [71, 136, 107, 56].

Unsupervised clustering based of K-means is a variance-based, unsupervised method for classifying data. It partitions a set of points or regions into K sets (clusters) such that the points in each cluster tend to be near each other. There are a variety of K-mean clustering algorithms available. This clustering method is not to be confused with the k-NN (Nearest Neighbor) classification/regression method and its algorithm that is supervised and classifies a point or region based on a known classification of other points/regions, thereby determine classification by combining the classification of the K nearest points. In contrast, supervised hierarchical clustering generates a tree-like structure (dendrogram) where the leaves are the individual objects (samples or variables) and the algorithm successively pairs together objects showing the highest degree of similarity.

Applications of PCA and these extensions are broad and diverse, for example: PCR has been applied to study crop-weather modeling and has been able to capture significant wheat crop-weather correlation patterns [16, 139], FPCA has also

provided a better understanding of spatial variations in the weather impacts of four teleconnections across British Columbia [9], and cluster-based PCA has recently been applied to understand complex molecular characteristics of cell states to obtain a better understanding of the cell reprogramming process [61] and to identify beneficial agronomic traits when selecting genotypes in a breeding program for white bean, pulse agricultural crops [74]. The later applications highlight how cluster-PCA is receiving increased attention its ability to explore and reduce the complexity of large, complex genetic and environmental datasets.

2.2.3 Data sources

For this study historical crop yield data for Spring Wheat (*Triticum aestivum* L.) and Barley (*Hordeum vulgare* L.) from 1987 to 2012 reported across 40 Census of Agricultural Regions (CARs) were obtained from the Field Crop Reporting Series of the Canadian Crop Condition Assessment Program (CCAP) within Statistics Canada's Agricultural Division. A linear detrending technique was first applied to the crop yield data. A dataset of yield predictors or explanatory variables was used that consisted of three-week moving means of remotely-sensed, normalized-difference vegetation index (NDVI) and agro-weather indices (i.e., seeding days, growing-degree days (GDD) above an ambient airtemperature of 5°C, average crop water stress index (AvgSI), average percentage of variable soil water holding capacity (Avg PrcntAW) and total precipitation (P)). We refer readers to details contained in a previous regional-scale modeling study that used these same input datasets [95].

In this study, for each cropping year (1987-2012), a growing season (i.e., May-August) cumulative temperature (°C) value was computed as a sum of monthly-mean sea surface temperature (SST) within the ENSO 3.4 region (monthly OISST v.2 (1981-2010 base period) available from the US Climate Prediction Center (CPC)), National Centers for Environmental Prediction (NCEP) of NOAA's National Weather Service. These monthly averages of ENSO 3.4 values account for ENSO events that span more than one year, and constructing such an annualized index based on monthly aggregate values captures the dominant state of ENSO within a given year [59]. ENSO correlations were considered to be contemporaneous (rather than at lag) with changes in other seasonal variables affecting yield and at lag.

2.2.4 Cluster-PCA modeling framework

A flowchart of our modeling framework and its statistical methodology is shown in Fig.2.3 that includes data transformation (standardization and detrending), K-means clustering, PCA applied at the Provincial and CAR aggregation scales, model selection, sensitivity analysis based on cluster number, cross-validation (hindcasting) and benchmarking of prediction skill. The modeling framework was coded and implemented using the R Statistical Language (Version 3.1.3, R Development Core Team) (R Foundation, 2016) making use of validated, open source algorithm libraries: pracma (detrending), stats (K-means, PCA, PCR), clustering (classInt), bestglm (GLM), Metrics (calculating MSE, RMSE), including other geospatial, plotting packages (i.e., raster, rgdal, rgeos, ggplot2, gridExtra, RColorBrewer).

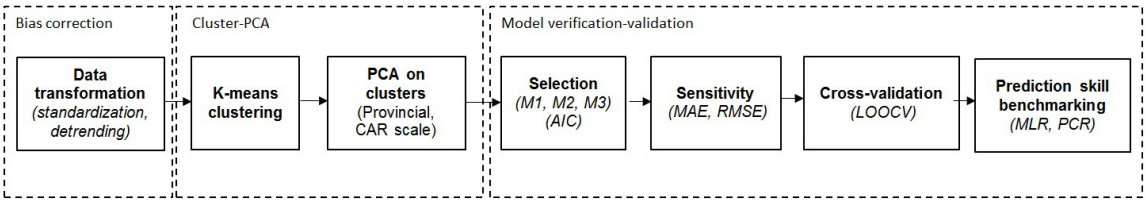


Figure 2.3 Cluster-PCA statistical modeling framework that includes data transformation (standardization and detrending), K-means clustering, PCA applied at the Provincial and CAR aggregation scales, model selection, sensitivity analysis based on cluster number, cross-validation (hindcasting) and benchmarking of prediction skill.

Clustering was performed at two aggregation levels, namely a non-CAR scale (i.e., provincial) and CAR-scale. At the Provincial scale, clustering was supervised, and the number of clusters was fixed to match the number of provinces (i.e., k of 3), with CARs assigned to each cluster before cluster-PCA was performed. This was generated to provide a non-CAR scale baseline of model performance, above which the relative gain of the unsupervised, CAR-scale clustering could be measured. This non-CAR test was also associated with a *null hypothesis* that finer CAR-specific information than the coarser provincial scale (with/without ENSO forcing considered) *is not* needed to explain observed yield-weather patterning, versus the *alternate hypothesis* that CAR-specific information *is* needed to explain historically observed yield-weather spatial patterning.

At the CAR-scale, clustering was unsupervised. An initial (distinct) number of clusters, k is specified within the range of 2 to 39. Average crop yields from 1987 to 2007 were computed for clustering and CAR subregions are then randomly selected

(i.e., 'random starts' in k-means R library code that implements the Hartigan and Wong clustering algorithm) and assigned to a given cluster based on an iterative, two-step process: 1) each CAR is assigned to a cluster based on the minimum distance between the CAR centroid and cluster centre locations. The total sum-of-squares (TSS) is minimized. TSS is defined as the sum of the squared differences of each observation (i.e., separation distances between CAR centroids and cluster centres), from the overall mean. In the next step, 2) cluster centre locations are updated as CARs are added to a given cluster. CARs from any part of the Prairies study regions were able to be assigned to a given cluster. Convergence of this clustering was reached when there were no further CAR assignments to a cluster or exchanges of CARs between clusters. When k is sufficiently small (i.e., 2 or 3), clusters can contain more CARs than years, thus violating the requirement of PCA that the number of CAR subregions must be less than or equal to the number of years (i.e., 20), these k values were not considered in this study. Findings of a recent study identified optimal values of cluster number k between 2-6 for the Canadian Prairies that resulted in high prediction skill [67]. PCA was performed with crop yield as the response variable, and NDVI, GDD, AvgSI, Avg PrcntAW, P as predictor variables. Separate PC scores were computed for crop yield response to each of the predictors. The total variance (i.e., cumulative variation) explained by a cluster was computed based on the cluster-average of RMSE values of all the CARs contained in it. Also, some CARs can contribute more variance than others.

2.2.5 Model and predictor selection

The accuracy of a statistical model can be highly dependent on the scale of its input data [83]. To account for the possibility of confounding scale-dependent effects, three different PCA *model types* were compared that involved different assumptions about the relationship between the response (crop yield) and predictor (weather) variables. Comparing models under provincial-scale (i.e., non-CAR scale) versus CAR-scale was used to assess the relative benefit that clustering and PCA separately had on overall cluster-PCA model performance.

- **Model** $M1$ (Cluster scale) assumes spatial variation in crop yield response (PC score) and weather predictors at the *cluster*-scale and assumes no finer, CAR-scale sub-regional variability.

- **Model** $M2$ (Multi-scale: Cluster and CAR), like **M1**, assumes spatial variation at the cluster-scale for *weather* predictors, but not for crop yield response. Instead, crop yield variability response is predicted at the CAR-scale. For this reason, PC scores for crop yield response are not used, as for Model $M1$, but instead standardized crop yield (Z value).

- **Model** $M3$ (CAR scale) assumes spatial variation in crop yield response and weather predictors at the finer, CAR-scale sub-regional variability. Standardized crop yield (Z value) was used.

Model type selection was coupled with predictor *variable* selection in the following way: statistical likelihoods (i.e., conditional probability of a dataset given a model of a given complexity and inter-linked assumptions) for each of the competing models ($M1$,$M2$, and $M3$) were computed and the best- fitting (i.e., optimized) model was selected based on the Akaike information criterion (AIC) ($AIC = 2k - 2lnL$) where L is the maximum value of the likelihood function of a model fit and k is the number of model parameters estimated. AIC provides a measure of the relative quality of statistical models for a given set of data, such that given a set of candidate models for the data the preferred model is the one with the minimum AIC value. Model response of $M1$ (F^*) can be transformed to the same model response (Z^*) as $M2$ and $M3$ using Eq. (2.4) with selection of number of PC yield scores.

$$Z^* = F^* Q^{*-1} \tag{2.4}$$

Where F^* is a $m \times h$ matrix consisting of the first h PC scores, selected from j possible choices, $0 < h \leq p$, from F from Eq. (2.3) and Q^{*-1} is a $h \times p$ matrix consisting of the first h rows, $0 < h \leq p$, from Q^{-1}. The resulted Z^* represents an amount of percentage of the variations about the original standard yield matrix Z depending on the number of h PC scores that was picked from F. Retaining more PC scores increases the amount of variation about the standard yield matrix Z that is contained in Z^*. Z^* will contain all the information about Z when $h = p$.

Both partial and complete representative sets of all the predictor variables were considered when selecting the best performing model (i.e., $M1$,$M2$,and $M3$ types). This involved *stepwise* regression that is an automated variable selection method that combines the efficiency of forward and backward inclusion of predictor variables and

has been shown to significantly improve crop yield prediction skill [95, 104]. This stepwise procedure involved selecting a first, leading predictor, evaluating model performance based on AIC criterion, and iteratively adding additional predictors and testing for any further reduction of AIC or improvement in model performance. If a predictor does not reduce AIC, then it is removed from consideration. Such partial models are comparable to full/complete models that comprise a fixed set of all the possible predictors of the single response variable (crop yield).

2.2.6 Sensitivity and validation

The study CARs were first clustered into regions based on province or clusters identified by K-means with k ranging from 2 to 39. The crop yield data was detrended to remove any bias from a technology historical trend, assumed to be linear. Weather variables are available as a three dimension matrix with axes are years, CAR clusters/province, and weather variables by CAR cluster or province. Weather variables by CAR are aggregated onto CAR clusters or provinces by taking the average of the weather variables for the constituent CARs for each aggregate zone. The PCA is applied to a two-dimensional matrix: years by CAR aggregated-weather variable. While weather variables in CAR scales, each CAR has its own weather matrix with observed years and the local weather variables as on the matrix rows and columns, respectively.. All data, including the raw crop yield, detrended crop yield, and both regional and CAR scales weather variables are standardized using Eq. (2.1). The standardized matrices were then trained by PCA and the year periods from 1987 to 2007 are used as models inputs for the three models in the system.

Model sensitivity analysis involved multiple runs of the cluster-based PCA for a range of k and then computing and generating corresponding profiles of the model performance measures of absolute error (MAE) and root-mean-squared-error (RMSE) across the simulated range. MAE is a quantity that measures deviation, while RMSE measures discrepancy in how close model predictions or forecasts are to 'true' outcomes represented as historically observed values (hindcasting) or future scenario or projected outcomes (forecasting),

$$MAE = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} |e_{ij}|, \tag{2.5}$$

$$RMSE = \sqrt{\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}(e_{ij})^2}, \tag{2.6}$$

$$\Delta MAE = (MAE_{ENSO^-} - MAE_{ENSO^+}), \tag{2.7}$$

$$\Delta RMSE = (RMSE_{ENSO^-} - RMSE_{ENSO^+}) \tag{2.8}$$

where e_{ij} is the estimate error from yield prediction and leave-one-out cross validation (LOOCV), respectively, over n years of crop yield response and m CARs. ΔMAE (bias) and ΔRMSE (variance) are differences in the statistics without (ENSO$^-$) and with (ENSO$^+$) ENSO, whereby a positive value of this difference indicates ENSO improves prediction, relative to model error without ENSO. Similarly, negative values indicate model predictions without ENSO are lower than with ENSO.

MAE and RMSE performance criteria were used to provide reliable measures of overall model stability and robustness for the full coupled (i.e., k-means clustering and PCA) model. At the CAR-scale, the RMSE and MAE of the first h PC scores for a given k from the yield model prediction and its cross-validation (LOOCV) were inter-compared, and the given k associated with the minimal RMSE was selected as optimal. In this way, the choice of optimal k guarantees minimal RMSE over all the CAR subregions.To evaluate prediction skill and model robustness, LOOCV to assesses how the omission of single years of data affected the cluster-based model predictions. In addition, the cluster-PCA model prediction skills was evaluated against several non-clustering regression-based benchmarks, namely multivariate (linear) regression (MLR) and principal component regression (PCR).

While cluster models can be evaluated based on how robust and/or stable they are, currently there is no consensus on minimal evaluation requirements because it is difficult to identify criteria that work for all datasets and statistical features. Moreover, devising clustering criteria that capture what users need that are compatible with both the content and context of a given dataset is difficult. Different types of criteria and procedures for assessing clustering optimality and stability could involve cluster crispness, shape and size attributes (particularly for asymmetric clusters), however searching for a single, optimal clustering is inappropriate, especially when correct clustering criteria cannot be specified in advance [24]. In light of this an integrated statistical methodology was selected that combined an assessment of model optimality, sensitivity, prediction skill/cross-validation and benchmarking to provide a more comprehensive evaluation of our cluster-based PCA model's performance and

prediction skill. This enabled a sufficiently broad assessment of clustering quality in terms of how well the dataset was partitioned and how well the selected clustering performs when used to predict.

Over-fitting is a potential problem, especially in multivariate (linear) regression (MLR). Over-fitting can occur when a model that is fitted or 'trained' using a given dataset and perform well when fitting to data, performs poorly with low prediction skill when 'validated' to a new, partial set of data becomes available, or when it is fully re-applied to an completely independent dataset. To avoid model over-fitting, Leave-One-Out-Cross-Validation (LOOCV) was applied [88, 11], whereby a complete year of yield and weather data was randomly selected and omitted, and the models were fitted to the remaining data. MAE and RMSE validation statistics are then computed based on the omitted year of data containing the 'true' response of crop yield and the model prediction containing data for all the other remaining years. The process of randomly selecting a given year to omit is repeated until all observations have been omitted at least once. MAE and RMSE are then obtained over all these model runs with smaller values indicating higher model prediction skill. In this study, LOOCV has applied to both benchmark models and our clustered-PCA modes. The resulted MAE and RMSE from LOOCV were used for model comparisons of model performance in crop yield prediction skills. In the non-PCA benchmark MLR model, LOOCV has applied to the observed climate variables to examine model performance of yield prediction skills directly from non-cluster and non-PCA climate variables. In this study, we assumed most of the crop yield variations in future short-terms can be captured by the PC scores from the Principal component analysis on historical long-term observation (1987-2007). For the PCA-based models in PCR and the three PCA model types (M1, M2, and M3), LOOCV has applied to the PC scores from PCA analysis on both crop yield and climate variables, respectively., for same years period. MAE and RMSE were calculated between the observed crop yields and the crop yield predictions from the PCR and the clustered-PCA models, respectively, for model comparisons in crop yield prediction skills.

2.2.7 Benchmarking

In addition to evaluate the benefits conferred by the CAR scale of spatial aggregation by comparing the cluster-PCA at the CAR and non-CAR (i.e., Provincial) scales, two additional models that did not involve any clustering at all, were run to compare with

the cluster-PCA model, to evaluate the benefit of clustering and benchmark model performance. The multivariate linear regression(MLR) in Eq. (2.9) is a generalization of linear regression by considering more than one independent variable (i.e. weather predictors), and a specific case of a general linear model (GLM) formed by restricting the number of dependent variables to one (i.e., crop yield). A MLR model was fit to the individual time-series for each CAR subregion.

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{2.9}$$

where $\mathbf{Y}$ is crop yield response matrix, $\mathbf{X}$ is weather predictor matrix, and β is the coefficient matrix, and ϵ is the random residual matrix, assumed to be normally-distributed with mean 0 and standard deviation of σ. In this model, predictors are assumed to be independent and no multi-collinearity exists among predictors. The second benchmark model was PCR. Use of a PCR model as a benchmark provides an indirect test for multicollinearity (highly correlated predictors) that is not provided by MLR [16], u. For this reason, PCR was included as an additional prediction skill benchmark. In PCR, the high correlated weather variables are transformed into PC scores by using PCA according to,

$$\mathbf{Y} = \mathbf{F}\gamma + \epsilon \tag{2.10}$$

where $\mathbf{Y}$ is crop yield response matrix, $\mathbf{F}$ is the transformed PC scores matrix for each CAR and γ is a coefficients matrix.

2.3 Results

An inter-comparison of wheat and barley yield prediction skill achieved by the competing model types ($M1$, $M2$, and $M3$), in comparison to the benchmark models (MLR, PCR) (under no ENSO forcing), is summarized in Table 2.1, in terms of the MAE and RMSE validation metrics. These models were run across CAR subregions within the Canadian Prairies (1987-2007), with the best performing model for wheat being $M2$ with 6 clusters, and $M3$ with 3 clusters for barley (i.e., both MAE and RMSE were at their minimum).

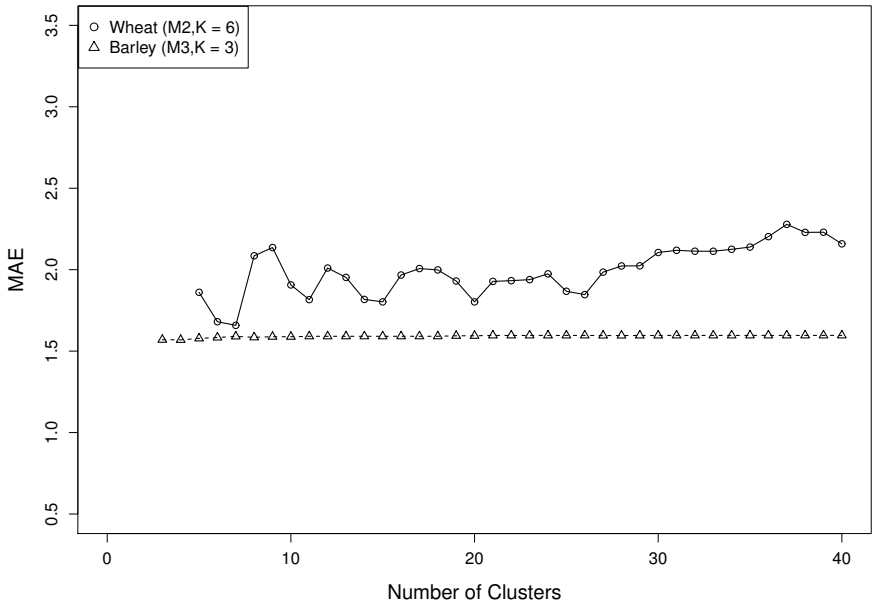
Prediction variance (RMSE) had reduced up to by 26-34 % for wheat and 2-4 % for barley from MLR and PCR. Validation RMSE decreased from 163.44 in MLR to 10.62 in $M2$ for wheat. Validation RMSE for barley also decreased from 144.87 in

Table 2.1 Prediction and cross-validated deviation (MAE) and discrepancy (RMSE) error (MAE - mean absolute error, RMSE - root-mean-squared-error) for cluster-PCA model optimized to each crop type (wheat and barley). These results show the relative gains in prediction skill obtained by considering spatial dependence, clustering at the CAR-scale, and the inclusion of ENSO as a predictor of crop yield. Values are reported to 3 significant figures. * indicates high error from leave-one-out cross-validation (LOOCV) indicating over-fitting.

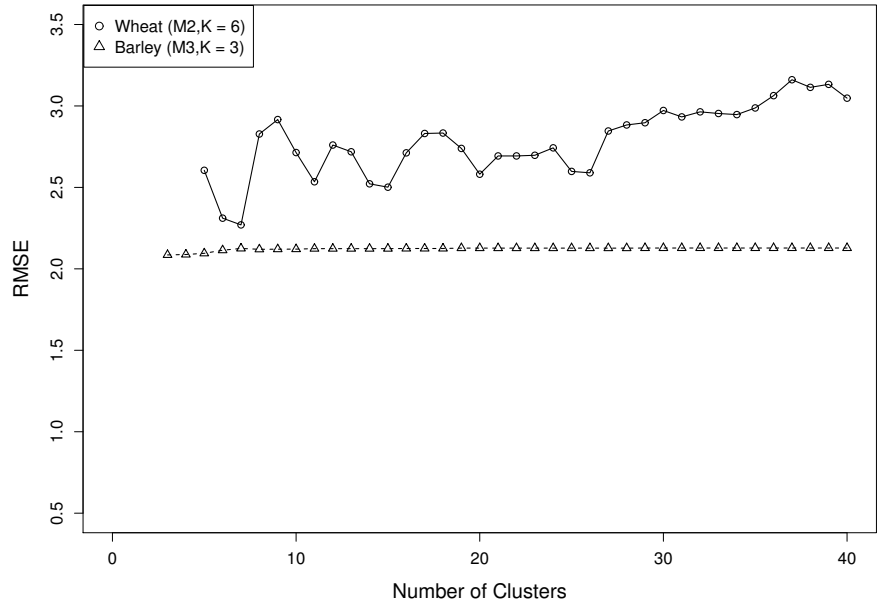
Spring Wheat					Barley				
	Prediction		LOOCV			Prediction		LOOCV	
Model	MAE	RMSE	MAE	RMSE	Model	MAE	RMSE	MAE	RMSE
MLR	2.06	2.87	61.61*	163.44*	MLR	1.62	2.17	54.06*	109.63*
PCR	2.16	3.05	40.13*	104.52*	PCR	1.6	2.13	60.06*	144.87*
$M2$ (P)	2.53	3.6	6.12	10.62	$M3$ (P)	1.57	2.09	5.83	11.82
$M2$ $(K=6)$	1.66	2.27	6.45	14.23	$M3$ $(K=3)$	1.57	2.09	6.53	16.53
$M2$ $(K=6, ENSO)$	1.51	2.14	6.11	12.61	$M3$ $(K=3, ENSO)$	1.25	1.76	8.77	32.39

PCR to 11.82 for the $M3$ model. Note that low values in the model prediction error, but with large values in leave-one-out cross-validation (LOOCV) error, indicates that over-fitting occurred; while these model types predict well based on the historical training data, they have low predictive skill under LOOCV prediction. Over-fitting did not occur with the other modeling approaches based on the error statistics ($M2$ for wheat and $M3$ for barley), whereby both non-CAR scale and CAR-scale clustering. For barley, prediction errors were same for both clustering types (Provincial and K-means $k = 3$ clustering) with non-CAR provincial clustering ($M3(P)$) shows a lower validation RMSE than with K-means clustering ($M3(K = 3)$). ENSO forcing was applied to the $M3$ model for both clustering types. $M3$ with provincial clustering has prediction MAE (1.25), RMSE (1.75) and LOOCV MAE (8.98) and RMSE (32.88), which were very closely to the result from ENSO forcing in $M3(k = 3)$.

Sensitivity testing of prediction bias (MAE) and variance (RMSE) over a range of possible k values (from 2 to 39 including $k = 40$, which treated each CAR as an individual cluster) is shown in Fig. 2.4. The MAE (left two) and RMSE (right two) values resulted from prediction (top insets) and LOOCV (bottom insets) for spring wheat and barley. Sensitivity of the best-performing clustered- PCA models under both progress versus step-wise variable selection procedures are shown. The process procedure uses an increasing number of predictors with the number of predictors from 1 to all when modeling with no model reduction process, rather than a standard step-wise approach. The minimum values in prediction MAE and RMSE with low values in averaged LOOCV MAE and RMSE indicates that the optimal and robust or 'best' model for spring wheat was $M2$, which used deterended step-wise modeling



(a) Sensitivity in yield prediction bias (MAE).



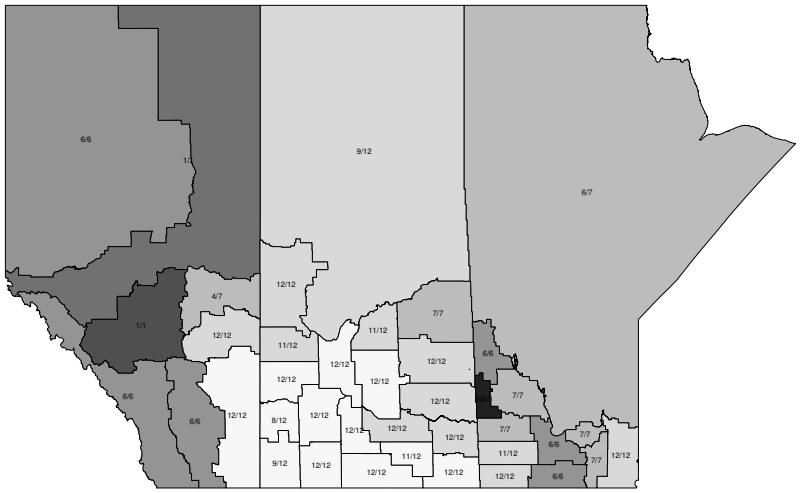
(b) Sensitivity in yield error variance (RMSE).

Figure 2.4 Sensitivity of model yield (wheat and barley) (bc/ac) predictions to the number of clusters: (a) yield bias (MAE), (b) yield error variance (RMSE). The prediction error statistics are based on the full cluster-PCA model.

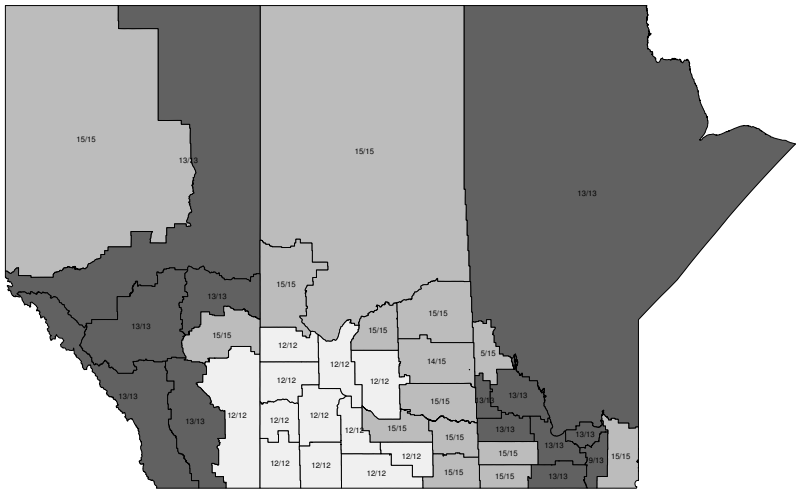
on CARs aggregated into $k = 6$ clusters and $M3$, which used de-trended step-wise modeling on CARs aggregated into $k = 3$ for barley.

Both prediction MAE and RMSE plots show similar plot shapes. MAE and RMSE are both slightly smaller than those from the step-wise modeling, averaged LOOCV MAE and RMSE for both spring wheat and barley are large compared to the results using step-wise modeling. The averaged LOCOV MAE are from 10.78 to 26.69 for spring wheat and 6.55 for all k values for barley. The averaged LOOCV RMSE are from from 25.06 to 91.39 for spring wheat and from 14.17 to 24.02 for barley. such that, we decided $M2$ with $k = 6$ and $M3$ with $k = 3$ both using step-wise modeling are the best model for spring wheat and barley, respectively.

Fig. 2.5 shows the predicted clusters based on yield variability - for wheat and barley (units of bu/ac i.e., bushels/acre) - from left to right, across the Canadian Provinces of Alberta, Saskatchewan and Manitoba. The number of PC scores used to predict yield within each CAR, alongside the total number of PC scores is indicated. CAR 4604, highlighted in Fig.2.1a and Fig.2.1b (see legend), did not have sufficient record length of historical wheat yield for modeling purposes. The crop yield variation result shows (not shown) these total number of PC scores had explain more than 93 % of the regional crop yield variations in CAR-scale for both wheat and barley, indicates the high performance of PCA in data transformation and improves model performance in yield prediction skills. In the case of barley, a more defined spatial pattern of yield variability is evident, showing some similarity to the pattern of different soil types in the Canadian Prairies. Both crops shows a similar core cluster in terms of both spatial extent and the specific CARs that belong to the cluster. This core cluster is centered in the southern portion of Saskatchewan. Wheat shows more predicted yield variability, especially at more northern latitudes or CAR regions. The clusters identified by the cluster-PCA method comprise CARs with a similar pattern of yield variability (i.e., yield response to weather including ENSO forcing), as measured in terms of the number of PC scores required to explain observed yield variance. In this way, this fraction involving PC scores is an indicator of spatial coherence in yield response to weather. The number of PC scores associated with highest prediction skill provides an indicator of spatial coherence in yield; the variability in yield *within* clusters is less than *between* clusters. Predicted cluster-PCA model bias and error uncertainty under ENSO forcing is shown in Fig. 2.6. Estimated bias and error variance (bu/ac) provide a reasonable level of accuracy. CARs with high bias do not necessarily have high error, and likewise for low bias and error - as evident from these

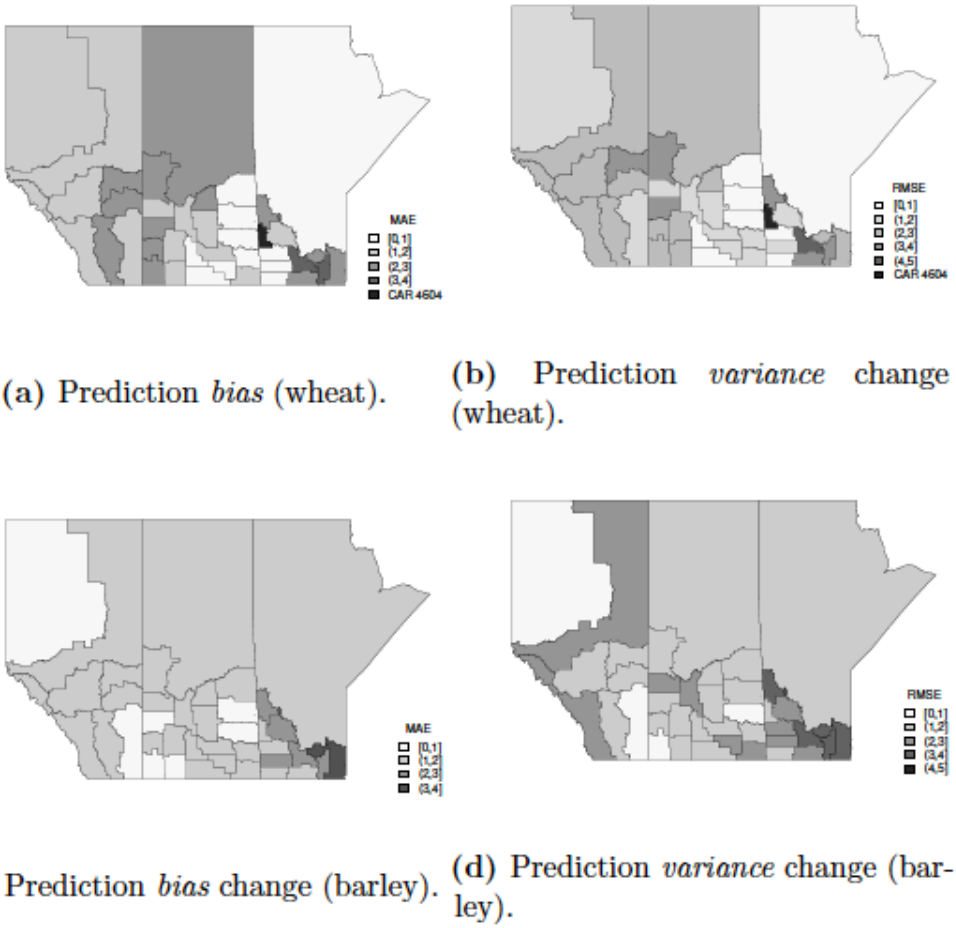


(a) PC scores - wheat (M2 model, k=6 clusters).



(b) PC scores - barley (M3 model, k=3 clusters).

Figure 2.5 Predicted CAR-clusters from K-means clustering, for (a) wheat and (b) barley. Results for each subregion are shown in fraction format, where the numerator is the number of yield PC scores used to predict CAR yields and the denominator indicates the total number of PC scores.



(a) Prediction *bias* (wheat).

(b) Prediction *variance* change (wheat).

(c) Prediction *bias* change (barley).

(d) Prediction *variance* change (barley).

Figure 2.6 Predicted yield *bias* and *variance* of CAR yield (bu/ac): (a) MAE of wheat, (b) RMSE of wheat, (c) MAE of barley, (d) RMSE of barley. Prediction errors of were computed using prediction yield from $M2$ with $k = 6$ for wheat and $M3$ with $k = 3$ for barley. Small values of MAE and RMSE indicate higher predictive skill.

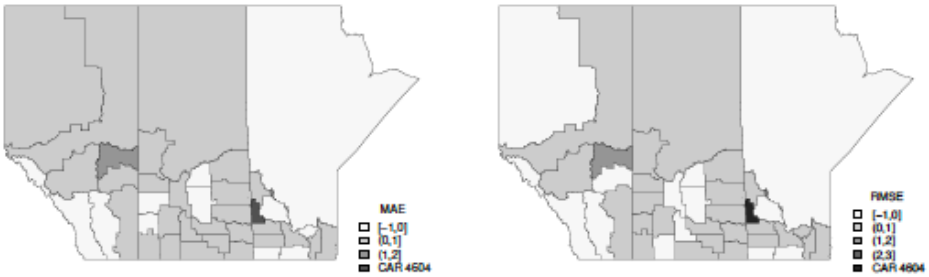
model output maps. Bias was higher for CARs in Manitoba for both wheat and barley. While accounting for ENSO statistically improved yield prediction over all CARs, no clear spatial pattern in prediction uncertainty is evident across CARs. Separate from ENSO climate forcing, other spatial variables (confounding or latent factors) such as soil type and agronomic management play a role in driving such heterogeneity and are considered implicitly (as opposed to explicitly) in the cluster-based PCA statistical model approach. While separate clusters of CARs were identified and such clustering improved the accuracy yield prediction overall, the clustering pattern was not evident in terms of the bias and variance contributions to accuracy (Fig. 2.7). In other words, changes to bias in some CARs improved yield prediction, while for

other CARs, reductions in variance were achieved by the cluster-PCA model. Put simply, both bias and variance reductions explain why the cluster-PCA model is able to achieve higher accuracy over other approaches.

2.4 Discussion

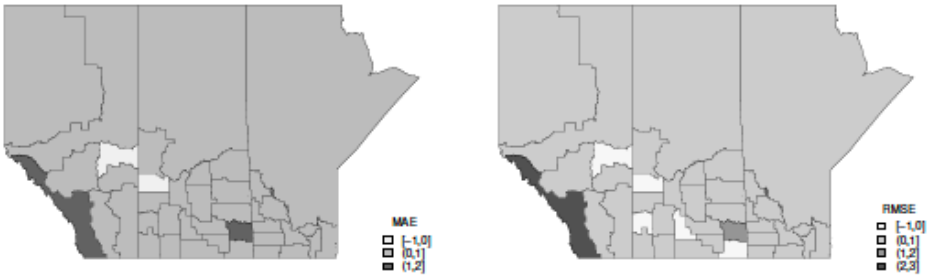
The model performance results indicate that a multi-scale ($M2$ model) is the best-performing model for wheat, and the CAR-scale model $M3$ for barley. As shown in Fig. 2.4, sensitivity of model yield prediction vary by crop types, wheat yield prediction is more sensitive to spatial-scale (and latitude variation) than barley, suggesting a considerable gain in predictive skill using a clustering approach for wheat over the large agricultural landscape of the Canadian Prairies. The minimum value of prediction MAE and RMSE were found at $k = 6$ for wheat and $k = 3$ for barley, which this result matched the findings from the previous modeling study [16]. Averaged crop yield for 1987-2007 were used to evaluate model predictive skill, nonetheless, historically this period includes some years with major drought, such as 1988, 2002, and 2007, where significant yield gaps/declines occurred, as shown in Figs 2.1(b) and (c) and the spatial varied drops of crop yield affect the result of CAR-scale clustering. As a result, some CAR regions have different selection of cluster PC yield score for yield prediction than others (refer to Fig. 2.5). A core or dominant cluster of CARs with a similar pattern of yield variability is evident in our findings, which roughly overlaps with a region of low correlation between the ONI index and seasonal precipitation (1987-2012) (southern 'blue' region shown in Fig. 2.2c) - this indicates that this cluster of CARs is a region with a significantly higher ENSO climate (e.g., drought) risk. In study, the best cluster-PCA models for both wheat and barley yield predictions are selected by the minimum values of AIC in model selections. [12] suggested AIC or AIC in small-sample equivalent may not be the optimal model performance measurements in model selection in all cases, especially with a less relevant in the case when model assessments is not related to predictions.

Overall, cluster-PCA model prediction achieved higher (and more robust) prediction of the various benchmarks/non-clustering approaches. In addition, the inclusion of ENSO forcing led to further gains of reductions of prediction and validation errors for wheat, thereby improving crop yield prediction. Compared to the MLR and PCR models, the model types from the cluster-PCA ($M1$, $M2$, and $M3$) provide more accurate crop yield prediction at the CAR scale. Model performance of prediction and



(a) Predicted yield *bias* change (wheat).

(b) Predicted yield *variance* change (wheat).



(c) Predicted yield *bias* change (barley).

(d) Predicted yield *variance* change (barley).

Figure 2.7 Relative gains in yield prediction skill (bu/ac) by including ENSO forcing: (a) Predicted yield *bias* change (ΔMAE) (Wheat), (b) Predicted yield *variance* change (ΔRMSE) (Wheat), (c) Predicted yield *bias* change (ΔMAE) (Barley), and (d) Predicted yield *variance* change (ΔRMSE) (Barley). Positive values indicate a reduction in bias and/or error variance (improved yield prediction), whereas negative values indicate a gain in bias and/or error variance.

validation results were shown in Table 2.1. Both prediction and validation errors have been reduced from the benchmark models (MLR and PCR), specially the reduction of validation error, over-fitting problems were avoided. This indicates the importance of scale in predicting crop yield. Clustering of CAR regions led to gains in yield prediction of wheat, while also being sensitive to crop type. The clustering result of barley in $M3$, with 3 clusters was shown in Fig. 2.5. The closely prediction and LOOCV validation errors from two different clustering results (Provincial and CAR-scale clusterings) indicates barley has more resistance and well-adapted to geospatial environment conditions.

In this study, we assumed the influence of ENSO forcing on the regional crop yield across the Canadian prairies is stationary in time. SST variability (ONI Index) was found to explain most of the observed crop yield variability of wheat and barley across the Canadian Prairies. Our cluster-based PCA analysis reveals that such ENSO variability has a nonlinear (i.e., clustered not directional linear gradient-based) forcing on crop yield variability at the regional scale. This nonlinear pattern for wheat yield across the study region (Canadian Prairies) was found to comprise six clusters. ENSO forcing has one step ahead improve model performance of modeling wheat yields, provide evidences of the importance of the relationship of ENSO and wheat yields. The changes of prediction errors shown in Fig. 2.7 also provides evidence of a spatial pattern of ENSO influence on wheat, across CARs situated in the north-west to south-east of the Canadian Prairies. CARs located from the Western of Alberta to the Northern of Manitoba are impacted less from ENSO, as indicated by negative values in the change of prediction error. The reduction of prediction error, but increasing LOOCV error for the selected cluster-PCA model, indicates that ENSO has less significant influence on barley yield across the Canadian Prairies. This finding suggests that future agriculture practices for wheat in CARs located within the identified ENSO forcing regions should consider more the climate risk and potential drought and flood impacts. Also, a significant non-stationary influence of ENSO on annual maximum precipitation was found on the southern Canadian Prairies. Annual maximum precipitation was higher in El Niño years than in La Niña years [120, 129].

New crop monitoring and reporting systems like the Agricultural Market Information System (AMIS) and Early Warning Crop Monitors, developed by The Group on Earth Observations' Global Agricultural Monitoring (GEOGLAM) initiative, seek to establish consensus through international assessment of crop growing conditions,

status, and agro-climatic conditions that are likely to impact major crop production[4]. An overarching goal of the Joint Experiment of Crop Assessment and Monitoring (JECAM) global initiative of GEOGLAM is to develop monitoring and reporting protcols and best practices for agricultural systems, while facilitating a convergence of modeling methodologies. Linked with the opportunities for breeding crops for the future, the value-added benefit of such improved crop yield prediction is both to gauge major climate impacts (like ENSO forcing) on crop yield and improved information to crop breeding programs that look far into the future across major production areas, and require more spatially-explicit and accurate information to delineate crop-weather adaptation zones and establish yield targets [147].

'Next generation' agroecosystem models that are now being developed will also require automated statistical algorithms to increase the reliability and consistency of multi-scale (field, region, globe) operational forecasts for agriculture, incorporating open collaboration and innovation requirements and capabilities (e.g., Agricultural Production Systems sIMulator, APSIM) [63, 76, 58]. Causality modeling that considers conditional probabilities and higher-order variable interaction or multi-scale effects can increase confidence in agricultural crop forecasts and weather adaptation decisions [96, 95, 93]. Multi-dimensional impact (i.e., integrated social, environmental and economic) assessments increasingly employ statistical, spatially-explicit (i.e., dependent) approaches when modeling spatially heterogeneous patterns of agricultural farming activities, changes in rural landscapes, urban-rural interactions and changing environmental conditions. Cluster-based PCA, as a trend-based, dynamic clustering approach, provides a method for the automated, optimal identification of risk areas (i.e., clusters, CARs) having similar levels of vulnerability/risk. This is useful for agricultural risk practitioners, assessors and managers in assessing crop insurance risks, by better capturing spatio-temporal dynamics of crop yield at more than one scale (i.e., multi-scale, multi-resolution). It also offers a consistent methodology for generating geospatial predictive maps of crop yield and its uncertainty, accounting for seasonal climate variability (e.g., ENSO), differing crop responses, and changes in network station density. Here we demonstrate the ability of this approach to boost the performance of crop yield prediction.

[4]G20 Action Plan on global food price volatility, www.foodsecurityportal.org/g20-action-plan-highlights-agriculture-and-food-price-volatility

2.5 Conclusions

ENSO affects crop yield in a nonlinear way, with factors like other then those represented by the teleconnection and SST variables (such as disease and agronomic management) also contribute to variation in crop yields.Significant evidences of influence of ENSO to barley yield prediction has not obviously observed. The inclusion of variables measuring soil fertility, soil moisture, plant diseases, insects, and weeds would undoubtedly help to reduce the unexplained variance. The usefulness of any indicator of crop yields is greater when it is known with sufficient time so as to help farmers in their agronomic decisions, for instance prior to the establishment of a given crop. Supporting earlier finding by [104] our cluster modeling identified coherent regions within the study area where the crop-yield model was quite effective at explaining crop yields (northern and western sections) and areas where it was less effective (central and eastern sections). In fact, no two crop districts used the exact same combination of variables to predict crop yield. These variations in crop prediction models, model validations, and patterns of correlation indicate that the physical mechanisms that link some teleconnections and, to a lesser extent, weather variables with crop yield operate on a relatively small spatial scale. This work also informs the model-based selection of predictors, inclusion of crop phenology and ENSO forcing, enhancing the capabilities and reliability of crop yield seasonal forecasting across Canada linked with the Integrated Canadian Crop Yield Forecaster (ICCYF) decision-support tool [27, 95, 127, 11, 88].

Chapter 3

Spatial and temporal analysis of heatwave occurrence across the Canadian agricultural regions using a copula-Bayesian network model

Heat stress, which is caused by prolonged periods of temperature exceeding a defined problematic threshold (heatwaves), has become one of the major disaster risks for agriculture production in Canada. In recent decades, an increasing frequency of heatwaves and their associated impacts have caused severely impacted crop yield losses, especially when associated with other weather variables (precipitation and humidity) during heatwave actives. Many approaches exist to identify heat wave occurrence and to categorize crop response to heat waves; this contributes to misunderstanding in how heat stress affects crops. Moreover, the low frequency of heatwave occurrence limits the application of some Gaussian-based statistical models that require relatively large frequencies of heat wave event occurrence for (linear) model training; this creates challenges to the development of efficient model-based strategies for heat stress control. In this chapter, we present a statistical and machine-learning approach, called the vine-copula Bayesian model, to explore the complex influences of heat stress on regional crops across Canada. This approach, based on conditional probability, combines daily environmental factors from a post-processed reanalysis climate data (Japan 55-year Reanalysis, JRA-55), modeled crop growth assessments (Advanced Very-High-Resolution Radiometer normalized difference vegetation index, AVHRR NDVI) at 1

km resolution and five candidate heatwave indices. This approach is validated in 96 Canadian agricultural regions by simulating crop growth using highlighted heat stress variables from probability model outputs. Results showed an improvement over the more typically used Gaussian-based Bayesian model. The outputs of this approach provide a better understanding of crop growth under extremely hot conditions for guiding decisions when faced with heat stress events.

3.1 introduction

3.1.1 Canada's agriculture regions and heat stress

Heatwaves, defined as daily temperature raised rapidly for a minimum of some defined multi-day period, and its negative impacts (heat stress), is one of the major extreme weather stressors for Canada's agricultural sector. Canada is a major global agricultural producer, growing a wide range of agricultural products includes jam berries (blue berry, cranberry), grains and oil-seeds (wheat, canola, and barley), fruit plants (cherry and apple), and vine grapes. Most of these agricultural products are summer outdoor plants. The development of these plants is a function of several variables, especially temperature, precipitation and humidity. Summer weather conditions vary across Canada, with milder climates on the coasts, hot and dry on the prairies, and wet in central Canada. Summer temperatures can often reach higher than 35 °C and have crept up in the last decades. As reported, the rate of temperature increase in in Canada associated with climate warming has been greater than the globally averaged temperature increased of about 0.85 °C) over the period from 1880 to 2012 (IPCC, 2013. `www.ec.gc.ca`). Extremely high temperatures during the growing season can cause significant crop yield reduction, the worst even can cause a 100 % in yield losses. Agricultural meteorology studies have raised concerns about the global yield losses from summer heatwaves. [97, 130, 77]. Studies in Canada have found crop response to heatwaves varies by crop types, regions, and heatwave types. Greater yield losses are noted when high temperature is associated with a lack of precipitation [90, 5, 13]. Conditions of high temperature and low precipitation also raise the hazard of secondary risks, such as environmental events such as wildfires and droughts.

3.1.2 Recent heat stress studies

There are many indices available for identifying heatwave events. Adverse impacts of high temperatures are usually felt on crops in as little as 2 to 7 days. Most heat stress impact studies have developed their index to assess specifically the influence of summer heat stress on crops at large scale [97, 130], which often leave uncertainty and misunderstanding of heat stress on less-common regional crops. Globally, [130] assessed yield losses from heat stress to the four major global crops: wheat, maize, rice, and soybean by using the magnitude and frequency information from heatwaves computed by a temperature threshold-based index with a minimum day threshold. This study identified spatial patterns of heat stress intensity for the four major crops around the world and concluded that Canadian agricultural areas are at a high risk of being affected by heat stress. The extreme weather disasters study by [77], which evaluates the influences of extreme weather disasters (droughts, floods, and extreme temperature) using an average national per-disaster cereal production loss (1964 to 2007), found extreme heat has caused a 7.0 to 8.1 % reduction of cereal yield reductions. These studies provided great insight into evaluating the yield loss caused by the intensity and frequency of heatwaves over the growing seasons. Modern agriculture practices have an increasing interest in developing efficient strategies to minimize yield losses from heat stress, which requires a better understanding of heat stress affects among the development of crop growth and the weather conditions associated with heatwave activities. A model which can exhibit skill in assessing the impact of heat stress on summer crop growth and that can identify the significant weather variables for efficient strategies can make a great benefit to agricultural practices across Canada.

3.1.3 Objectives

The objective of this work is to better understand the effects of heat stress on Canada's regional crops. To do this a statistical model is developed, which represents the risks of heat stress on regional crop growth. In particular, a copula-based Bayesian model is implemented, which will allow the following specific tasks to be completed: 1) Detect heatwaves for regional crops across Canada's agriculture regions and ascertain heatwave risk (intensity and frequency) across the country, 2) Highlight weather conditions that affect crop growth during heatwave actives, 3) Evaluate the starting date of heatwave effects, and 4) Improve model performance in probability predictions for regional crop growth under low occurrence of heatwaves. This will provide helpful

information for agriculture advisors/producers to improve their decision-making in their efforts to manage risk exposure to high summer temperatures.

3.2 Materials and methods

3.2.1 Study site

The study sites used in this study were a set of 96 agriculture regions selected from Canada's eco-district farming areas based on the findings from a recent study by [94], which had downscaled climate scenarios to a high resolution of 10 km for a set of selected Canada agricultural regions. Specific regions were retained for study based on the following criteria: 1) Regions have high levels of agricultural activity, 2) Regions are fully within the agricultural lands in the distinct eco-zones of Canada, zones which are in turn defined by climate conditions, and 3) Regions contain a maximum number of high-quality, long-term climate monitoring stations, and a maximum number of climate model scenario evaluations for long-term studies. Regions were confirmed to be agriculture regions within the eco-zones of Canada through the Google mapping website (`https://www.google.ca/maps`). Fig. 3.1 show the geo-spatial mapping of the study regions (black circle dots) located within the eco-zones (grey areas) of 7, 8, 9, 10, 13, and 14 across the country. These eco-zones are in the south of Canada, producing a wide range of agricultural productions driven by the uncertain and changing climate conditions over summer seasons.

3.2.2 Weather data

Daily weather variables (temperature, precipitation, and relative humidity) over the summer seasons (April to September) during a 26-years study period from 1987 to 2012 were the bias-corrected data from the Japanese 55-year Reanalysis (JRA-55) in the S14 retrospective meteorological forcing dataset (S14FD, `http://h08.nies.go.jp/s14/`). Daily weather data from the JRA-55 are high-quality reanalysis climate dataset generated by using a highly sophisticated operational Data Assimilation (DA) system and newly prepared dataset of historical observations. weather datasets have homogenized in time (up to 3-hourly) and space (around 50km) since 1958 to current dates. JRA-55 provides a wide range of weather variables for multiple weather purpose use includes meteorological agriculture factors (temperature, precipitation,

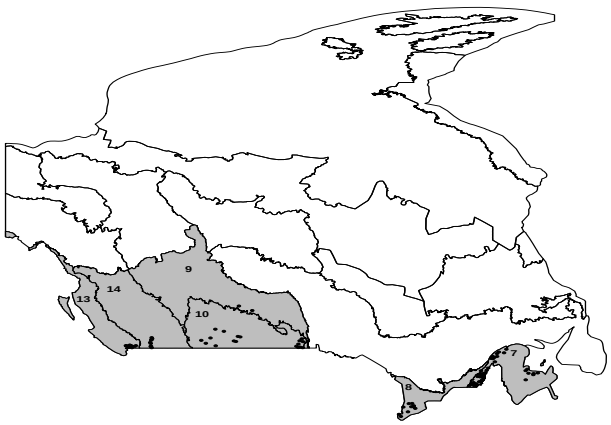


Figure 3.1 The eco-zones across Canada. The eco-zones in grey are the major agricultural farming areas of Canada. The black circle dots within the agricultural eco-zones are a set of 96 high activities agricultural regions.

and humidity) at near surface level, long term seasonal forecasts, extreme weather analysis (heat stress), and climate monitoring. The S14FD was developed by the organization of Institute for Agro-Environmental Sciences, National Agriculture and Food Research Institute in Japan (`http://metadata.diasjp.net`). It offers 11 bias-corrected daily climate variables over the globe from 1958 to 2013 by using a set of bias-correction methods and forcing data sets in different General Circulation Model (GCMs) and Representative Concentration Pathways (RCPs). Daily weather variables include daily temperatures(minimum, maximum, and mean) at 2m, relative/specific humidity (2m), total precipitation, mean downward shortwave/long-wave radiation flux, mean vapor/surface pressure, and wind speed (10m) are all at surface levels. The bias-corrected weather variables from S14FD has shown to have advantages in improving the accuracy of observed temperature and precipitation extremes in recent decades (1961-2000 and 1979-2008) than other forcing data sets [64]. Daily weather variables from JRA-55 data in biased-corrected S14 can provide high-quality historical weather data for the studies of agricultural climatology in weather extremes. Daily weather variables of minimum/maximum temperature, relative humidity, and total precipitation from the bias-corrected JRA-55 in S14FD were used in this study to detect heat stress events and model the impacts of heat stress on crop developments.

3.2.3 Normalized Difference Vegetation Index (NDVI)

Normalized Difference Vegetation Index (NDVI) was obtained from historical satellite images of Advanced Very High-Resolution Radiometer (AVHRR) in the Crop Condition Assessment Program (CCAP)(`https://open.canada.ca/data/`). AVHRR continuous providing four- to six-band multi-spectral data from the Administration (NOAA) polar-orbiting satellite series for global vegetation monitoring in a time intervals of every 7- and 14-days, and in a spatial resolution of 1.1 kilometers since 1979 to present. CCAP is developed and maintained by the Remote Sensing and Geospatial Analysis Section (RSGA) within the Agriculture Division, which keeps providing reliable, objective, and detailed information on crop and pasture conditions using mapping images for the whole Canadian agricultural area and the northern portion of the United States. CCAP supports AVHRR data since 1987 for most of the growing weeks in Canada in summer seasons (between Julian weeks 15 and 41). The NDVI values from AVHRR 1 km data have been widely used for environmental monitoring and extreme weather monitoring [98, 36, 128]. The conversion equation

from AVHRR to NDVI has shown as below:

$$NDVI = (AVHRR - 10000)/10000 \tag{3.1}$$

where $AVHRR$ is the historical satellite image values. A time-series NDVI data in a 14-days interval over the summer growing seasons from 1987 to 2012 have used to assess the development of crop plants over the study regions in this study.

3.2.4 Measuring crop heatwave stress

Evaluations of heat stress on crop developments in this study were obtained from a set of heatwaves detected from the five selected heatwave index. These heatwave index were all driven by daily weather variables of temperature and precipitation, which can be easily obtained and apply for most of the regional crops across Canada. The first widely used heat stress index is the 95th percentile index, which a Heatwaves has defined when the daily maximum temperature is over the 95th percentile value in long-term monitoring. This pure and straightforward temperature-based index is one of the most commonly used index in heatwave impact studies. It has advantages in assessing final crop yield losses due to the disaster impacts of extremely high temperature but may remain uncertainty in heatwave frequency and duration as a heatwave event maybe shifted from one short-term period to another by the drought years or warmer years in the study period. The formula of the 95th percentile index has shown below:

$$F_X(x) = Pr(X \leq x) = 0.95 \tag{3.2}$$

where F_X is the cumulative density function for daily maximum temperature and x is the temperature threshold that higher than 95th percentile of the daily maximum temperature over the summer seasons from 1987 to 2012.

Another widely used index is the constant heat stress index, in which a heatwave event has detected when the daily maximum temperature is over a constant threshold temperature. Crop plants can develop their resistance against cold and hot environmental conditions. Most major crops have their plant resistance for high temperature extreme at a range between 28 and 35 °C [130, 97]. High temperature over 28 °C in the flowering stage of some major crops, such as wheat, can result a significant reduction in final yield. A constant threshold of daily maximum temperature at 28 °C

has used in this study. This constant-based heatwave index can be easily applied in heatwave detection for an individual year without a long-term observation. Summer temperature reaches, or over than 28 °C are often can be found at late summers for regions located at a higher latitude and can be earlier for a lower latitude regions.

The Crop Stress Index (CSI) introduced by [48] is a crop heat stress index computed by using accumulate daily temperature and precipitation in a month-long period. It was developed for major cereal crops and has been correlated with wheat yields and grasshopper populations in Saskatchewan in Canada, a cotton pest in Australia, and regional drought in the U.S. North Central Region (NCR) [49, 54, 47]. The application of CSI in the NCR regions has shown a negative relationship between the values of CSI and yield losses with each unit increase in the CSI has resulted in a 0.14 and 0.004 Mg/ha yield reduction for corn and soybean, respectively [48]. The equation of the CSI is:

$$CSI = \frac{CGDD_{10}}{(CTP + 1)} \tag{3.3}$$

where $CGDD$ and CTP are the cumulative *growing degree days* (GDD) at a 10 °C base temperature and total precipitation (mm), respectively, in a month period. GDD can be calculated as:

$$GDD = \frac{T_{max} + T_{min}}{2} - 10$$

where T_{max} and T_{min} are the daily maximum and minimum temperature, respectively.

Change-point detection approach has widely used in heatwave detection by comparing the changes of averaged daily temperature in a short-term to a longer-term observation. Two Change-point detection based heat index, the Excess Heat Factor (EHF) [91], and the heatwave Magnitude Index daily (HWMId)[113], were used. The calculation of EHF is a combination of another two Change-point-based index, the significance index (EHI_{sig}) and the acclimatization index (EHI_{accl}). The EHI_{sig} has computes the changes of an average daily mean temperature at three-day period (the current day and the next two days) to the 95th percentile threshold a full set of observation while the EHI_{accl} has computes the changes of an averaged temperature at three-days period to longer period from 30 days ago. The equations of EHI_{sig} and EHI_{accl} can be expressed as:

$$EHI_{sig} = \frac{(T_i + T_{i+1} + T_{i+2})}{3} - T95$$

$$EHI_{acc} = \frac{(T_i + T_{i+1} + T_{i+2})}{3} - \frac{(T_{i-1} + \cdots + T_{i-30})}{30}$$

where T_i is the mean daily temperature for day i and $T95$ is the 95th percentile daily mean temperature calculated by using Eq. 3.2. The equation of EHF can be obtained by:

$$EHF = max(0, EHI_{sig}) * max(1, EHI_{acc}) \tag{3.4}$$

where $max(a, b)$ is the maximum value between a and b. A heatwave event is defined when $EHF > 0$. A three days period from day i to day $i + 2$ are classified as event days for every positive value of EHF at day i.

The Heatwave Magnitude Index daily (HWMId) introduced by [112] is an improvement of the Heatwave Magnitude Index (HWMI). The HWMId is a percentile- based index with its temperature threshold is obtained at a 90th percentile value from the daily temperature in a 31 days window over the years from 1987 to 2012. Let A_d denotes as a subset of daily maximum temperature for a summer day i, then A_d can be expressed as:

$$A_d = \cap_{1987}^{2012} \cap_{d-15}^{d+15} T_{y,i} \tag{3.5}$$

where $T_{y,i}$ is the maximum daily temperature at day i in year y. A_d containing all maximum daily temperature from day $i - 15$ to day $i + 15$ over the years for $T_{y,i}$. A heatwave day is defined if a maximum daily temperature at day i is higher than the 90th percentile of A_d.

3.2.5 Copula-Bayesian networks

Bayesian networks (BNs) are direct acyclic graphical (DAG) models that used to represent the complex joint probability distribution between network variables. The complex causal relationships between random variables X_i are split into multiple local distributions and represented as a DAG. Independent random variables within a DAG are represented as nodes and linked by edges from the direct causes as determined by the conditional dependencies probability P. Each variable is conditionally dependent on its effects and independent of its non-effects. Let X denoted as a set of random

variables in a DAG, then the full joint probability distribution has expressed as:

$$P(X) = P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|\pi_i) \tag{3.6}$$

with the conditional local distribution can be express as:

$$P(x_i|\pi_i) = \frac{P(x_i, \pi_i)}{P(\pi_i)} \tag{3.7}$$

where x_i and π_i, $i = 1, 2, \ldots, n$, are child and parent nodes, respectively. The distribution of child nodes depends only on their parent nodes, the number of which is limited. The global joint distribution is split into local distributions as network structure and parameters are learned from such a structure using the network observations.

There are several advantages of using a Bayesian network over traditional regression methods in environmental apply science studies. Firstly, the flexibility of BNs allows divers of data types, including discrete, continuous, and mix (discrete and continuous) in model learning. Secondly, BN models can capture the inter-dependencies between non-linear environmental variables. Lastly, the graphical nature of a BN makes the dependence configuration more explicit between environmental variables, which enable one to make probability inferences of local distributions from a complex DAG [100, 119].

In structure learning, independent random variables can be linked by dependent evidence from expert's experience on the field, study findings, or dependent algorithms. Model structure of a Bayesian has represented by child and parent nodes. Data type in parent nodes can either be discrete or continuous, but only continuous data are allowed in child nodes. The performance of a BN in parameter learning is dependent on the sample size of the observation. Independent random variables have generally assumed to be Gaussian distributed in model learning when the sample size of observation is large as followed by the Central Limit Theorem (CLT) that the distribution of sample means will approximate to a Gaussian distribution when the sample size is large enough. This approximation provides tremendous convenience in parameter learning from a complex model structure, but may be be the case in heat stress studies as the low frequency of heatwave occurrence. Copula approach was applied to improve the model performance in parameter learning by modeling the dependent structures between child and parent nodes within a BN network.

Copula approaches are multivariate cumulative distributions often used to describe

the dependence structures between random variables. It is on an n-dimensional unit cube with the marginal probability distribution of its n random variables are uniformly distributed within an interval of $[0, 1]$ [66, 108, 116, 7]. Copula approach is an excellent tool for modeling and simulating correlated random variables, which it can present a multivariate dependence of the marginals as a function of univariate marginals and a copula link function C [92, 78, 72].

The development of copulas has rooted in Sklar's theorem [123] that any multivariate distribution can be represented as a copula function of its marginals. To explain the concepts of the copula in more detail, let define $F(x_1, \ldots, x_n)$ be any multivariate distribution over its random variables from x_1 to x_n, then there exists a copula function such that:

$$F(x_1, \ldots, x_n) = c(F_{X_1}(x_1), \ldots, F_{X_n}(x_n)). \tag{3.8}$$

where $F_{X_i}(x_i)$ is the univariate cumulative distribution function of the random variables X_i and $C(\cdot)$ is the copula that capturing the dependence structure between random variables from X_1 to X_n. If each $F(x_i)$ is continuous then C is unique. When the random variables X are continuous, and by using the chain rule, the joint probability density function $f(x)$ can be expressed as:

$$f(x) = \frac{\partial^n C(F(x_1), \ldots, F(x_n))}{\partial F(x_1) \ldots \partial F(x_n)} \prod_i f(x_i) = c(F(x_1), \ldots, F(x_n)) \prod_i f(x_i) \tag{3.9}$$

where $C(F(x_1), \ldots, F(x_n))$ is called the *copula density function.*

Compares to most linearly dependent based approaches, such as multivariate Gaussian distribution, copula provides more useful results in representing non-linear dependent structures in multivariate elliptical distributions. When copula approaches apply to the local distributions within the DAG of a Bayesian network (CBNs), it often results in a significant improvement in representing the dependent relationships between child and parents sets, especially when the sample size is small [37, 148]. By applying Eq. 3.9 in Eq. 3.7, [37] have shown the local conditional probability density $f(x_i|\pi_i)$ for a continuous random variable x_i can be expressed as:

$$f(x_i|\pi_i) = R_c(F(x_i), F(y_{i1}), \ldots, F(y_{ik}))f(x_i) \tag{3.10}$$

where R_c is a ration function of:

$$R_c(F(x_i), F(y_{i1}), \ldots, F(y_{ik})) = \frac{c(F(x_i), F(y_{i1}), \ldots, F(y_{ik}))}{\frac{\partial^k C(1, F(y_{i1}), \ldots, F(y_{ik}))}{\partial F(y_{i1}) \ldots \partial F(y_{ik})}} \tag{3.11}$$

where R_c is equal to 1 when $\pi_i = 0$. If x_i has only one parent in π_i, such that $k = 1$. Let $\pi_i = y_{i1}$ as the only parent for child x_i, then the conditional density function of $f(x_i|\pi_i)$ can be simplified as:

$$f(x_i|y_{i1}) = R_c(F(x_i), F(y_{i1}))f(x_i) = c_i(F(x_i), F(y_{i1}))f(x_i) \tag{3.12}$$

where c_i is a local copula modeling the dependence structures between the child node x_i and its parent y_i1, such a copula also called vine-copula. The heatwaves used in this study from different heatwave index were in different periods. Using vine-copula rather than multivariate copula can provides a better understanding of the influence of heat stress on crop growth from each heatwave index. Multivariate copula often requires a large sample size from long-term monitoring in model learning, which vine-copula has a better fit to the low occurrence heatwaves as it requires much less observations than multivariate copula. The copula Bayesian model was coded and implemented using the statistical programming language R (version 3.5.3) making use of libraries: bnlearn (Bayesian network learning), copula (multivariate distributions constructed from copulas), VineCopula (maximum likelihood estimation of vine copula families), and other plotting packages (i.e., ggplot2, raster, ggspatial).

3.3 Copula-Bayesian Network Learning

3.3.1 Data processing

At each of the growing season, AVHRR satellite mapping for the regional crops across Canada starts from different starting dates based on long-term monitoring. NDVI data were obtained from the weekly observed AVHRR using Eq. 3.1 and has used as an assessment for regional crop growth. Heatwave detection started on the same date as AVHRR monitoring started. Crop plants have their minimum days (2 7 days) in resistance to be affected by heatwaves. Regional heatwaves have detected from the five heatwave index as discussed in section 3.2.4 in different minimum days using weather observations from the post-processed JRA-55 in S14FD from 1987 to 2012. To obtain

more heat stress data for modeling learning, observations have generated in two-weeks intervals. That is, the observation of NDVI at the current week has compared to the observation from two weeks ago, and the differences of NDVI was used as heat stress assessment. In such a way, a heatwave can have a maximum duration of 14 days, and multiple heatwaves can be detected within each two-weeks intervals. A heatwave with a duration across observation will be cut-shorts by the current week but can be captured again by the next observation on next week. The influence of heat stress on regional crops has assessed by the heat stress variables (heatwave intensity and daily weather variables (temperature, precipitation, and humidity)) over the heatwave actives. Heatwave intensity were assessed in two ways, heatwave magnitude and heatwave degrees. The heatwave magnitude a temperature-based index, it was first introduced by [113] used to calculate the daily heat intensity in a European study. The heatwave magnitude is counting the accumulated degrees of daily maximum temperature from the 25th and 75th percentile value from the historical observation. The equation for the heatwave magnitude has expressed as:

$$M_d(T_d) = \begin{cases} \frac{T_d - T_{26y25p}}{T_{26y75p} - T26y25p} & if\ T_d > T_{26y25p} \\ 0 & if\ T_d \leq T_{26y25p} \end{cases} \tag{3.13}$$

where T_d is the maximum daily temperature on day d. T_{26y25p} and T_{26y75p} are the 25th and 75th percentile values, respectively, from the observed maximum daily temperature from 1987 to 2012. M_d is a non-negative value and only taking the degree values into account when the maximum daily temperature is higher than the 25th percentile from the observation. The heatwave degrees a heatwave index-based assessment, it has counting the degrees of heat stress index changes over each two-weeks intervals. The calculation of heatwave degrees is:

$$V_d = X_d - Base_d \tag{3.14}$$

where X_d and $Base_d$ are the daily value and the base value, which varied in heatwave index. For temperature threshold based heat index, such as *Percentile*, *Constant*, and HWMId, X_d is the maximum daily temperature and $Base_d$ is the temperature threshold from the heatwave index. While for CSI and EHF, X_d is the index value of CSI and EHF, and $Base_d$ is equal to 0 as the values of heatwave index lower than 0 indicates there is no heatwave occurs. Others, the averaged daily maximum temperature, precipitation, and relative humidity over heatwaves actives has computed

and used for heat stress risks assessment.

3.3.2 Network learning

Three critical factors have considered when copula applied in Bayesian networks. Firstly, dependence structure modeling using copula approaches has restricted by the random variables showing in an independent relationship or shallow relationship, independent test between the two variables may need before the dependence structure modeling. Secondly, copula requires knowing the marginal distributions of its random variables for copula selection. The commonly used of normality distributed assumption for continuous random variables within a Bayesian network may not suit for the low occurrence of heatwaves. Marginal distributions need to be identified by the empirical observations. In lastly, copula are well representing the dependence structures for continuous random variables in the Bayesian network [72]. When only continuous variables are in a Bayesian network structure, the use of copula can representing the joint probability and dependence structure between the nodes in the DAG in an efficient way, which allows separating the selection of a multivariate distribution from multiple one-dimensional marginal distributions and enables a high flexibility cumulative distributions [30]. Network modifications may be needed in network structure learning when network variables contains discrete data types in a DAG.

The relationship between the heat stress variables and the NDVI has computed using Spearman's Rank-order correlation. Compare to Pearson correlation in evaluating the linear relationships between the variables, Spearman correlation tends to evaluate the monotonic relationship between two continuous or ordinal variables, which allows the correlated variables to changes together at a non-linear rate. In Spearman's correlation, variables are ranked in orders and used to compute the coefficient of the correlation. The equation used to calculate the coefficient of Spearman's Rank-Order correlation is:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{3.15}$$

where d is the rank difference of the two variables and n is the number of observations. ρ is the correlation coefficient having values between -1 to 1 with values more close to -1 and 1 indicates the two variables are in a more definite negative and positive relationships, respectively. While the values of ρ closes to 0 indicates a fragile relationship. To obtain better observations for model learning, we defined a heatwave

Table 3.1 A set of candidate marginals probability distributions

Distribution	Probability density function	Parameters	Description
Normal	$f(x) = \frac{1}{\sqrt{(2\pi)}\sigma} exp\{-(x-\mu)^2/(2\sigma^2)\}$	μ, σ^2	sample mean, sample variance
Log-Normal	$f(x) = \frac{1}{\sqrt{(2\pi)}\sigma x} exp\{-(lnx-\mu)^2/(2\sigma^2)\}$	μ, σ^2	sample mean, sample variance
Gamma	$f(x;\alpha,\beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$	α, β	shape parameter, inverse scale parameter
GEV	$f(s;\xi) = \begin{cases} (1+\xi s)^{(-1/\xi)-1} exp(-(1+\xi s)^{-1/\xi}) & if\ \xi \neq 0 \\ exp(-s) exp(-exp(-s)) & if\ \xi = 0 \end{cases}$	s, ξ	$s = \frac{x-\mu}{\sigma}$, shape parameter

variable is significant if it has an absolute value of $|\rho|$ higher than 0.2 with NDVI and has a higher than 50 occurrence from 1987 to 2012, which has an averaged frequency of higher than two times in annually.

Evaluations of marginal distributions for both heat stress variables and NDVI has done using the Shapiro-Wilk Goodness-of-Fit tests on a set of candidate probability distributions. Heatwave variables are non-negative values. Probability distributions have chosen from Gaussian, gamma, log-normal, and generalized extreme value (GEV) distributions. While for NDVI, best fit distribution have chosen from Gaussian and generalized extreme value distributions. The probability density functions and the parameter description of the candidate probability distributions have shown in Table 3.1.

The Shapiro-Wilk test has introduced by [122] for testing the goodness of fit between the expected distribution on the observation with an null hypothesis of a sample variable is came from a normally distributed population. The W-statistic of the Shapiro-Wilk is:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.16}$$

where $x_(i)$ is the ith order element in the sample variable, a_i is the coefficient for $x_{(i)}$, and $\bar{x}$ is the sample mean. The null hypothesis is rejected if the returned *p-value* from the W-statistic is less than 0.05, and one can conclude that there is no evidence to support the sample variable is normally distributed. Shapiro-Wilk test has applied to select the best empirical distribution for both heatwave variable and NDVI, respectively, from the candidate distributions listed in Table. 3.1. The best fit distribution is chosen by the largest value of *p-value*, which the selected candidate distribution has the best fit to the observation. If all returned *p-value* are less than 0.05, none of the candidate distribution have a good fit to the observation.

Independent test between the heat stress variable and NDVI has performed in

a Kendall's τ-based bivariate asymptotic independence test. The equation of the bivariate asymptotic independent test is:

$$T = \left(\frac{9N(N-1)}{2(2N+5)}\right)^{0.5} * |\tau| \tag{3.17}$$

where N is the number of observations and $\hat{\tau}$ is the empirical Kendall's tau of the data vectors $u1$ and $u2$. $u1$ and $u2$ are the data vectors of heat stress variable and NDVI, respectively, with values are within an interval of [0,1]. The p-value of the null hypothesis of bivariate independence can be calculated as:

$$p.value = 2 * (1 - \phi(\tau)) \tag{3.18}$$

where τ is the coefficient of Kendall's correlation and can be calculated as:

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} sgn(x_i - x_j) sgn(y_i - y_j) \tag{3.19}$$

The hypothesis of which the two variables are independent will be rejected if P.value is less than 0.05. The bivariate independence test has applied to test the relationship between each heat stress variable and NDVI. A heat stress variable will be removed if it shows an independent relationship with NDVI.

3.3.3 Selection of copula distribution

Suppose F_1 and F_2 are representing the marginal distribution for random heat stress variable and NDVI, respectively. Then we can have:

$$F_i(F_i^{-1}(y)) = y \tag{3.20}$$

where $F^{-1}(\cdot)$ denotes as the inverse of $F(\cdot)$. By evaluating Eq. 3.8 at $x_i = F_i^{-1}(u_i)$ and use Eq. 3.20, we can obtain a copula distribution from Eq. 3.8 as:

$$F(x_1, x_2) = C(u_1, u_2) = F(F_1^{-1}(u_1), F_2^{-1}(u_2)) \tag{3.21}$$

By using Eq. 3.21, the best fit copula distribution has selected from a set of five candidate marginal distributions as shown in Table. 3.2, including Gaussian, Student-t, Clayton, Gumbel, and Frank distributions. In copula selection, the parameters of the candidate distribution have fitted using maximum likelihood estimation and use

Table 3.2 A range of candidate distributions for bivariate copula.

Distribution	Copula	parameters	Descriptions
Gaussian	$\Phi_P(\Phi^{-1}(u_1), \Phi^{-1}(u_2))$	P	correlation matrix
Student-t	$t_{\nu,P}(t_\nu^{-1}(u_1), t_\nu^{-1}(u_2))$	ν	degrees of freedom
Gumbel	$exp\left(-((-\ln u_1)^\theta + (-\ln u_2)^\theta)^{\frac{1}{\theta}}\right)$	θ	parameter, $\theta \in [1, \infty)$
Clayton	$(u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}$	θ	parameter, $\theta \in [-1, \infty) \backslash \{0\}$
Frank	$\frac{\theta e^{-\theta(u+v)}(e^{-\theta}-1)}{(e^{-\theta(u+v)}-e^{-\theta u}-e^{-\theta v}e^{-\theta})}$	θ	parameter, $\theta \in (-\infty, \infty) \backslash \{0\}$

to compute the model fit measurement of Akaike Information Criteria (AIC). The 'best' copula families is selected by the one which has the minimum value in AIC. The Akaike Information Criteria (AIC) has introducted by [2]. It is a measurement used for estimating the relative amount of information lost by a given model with the trade-off between the goodness-of-fit of the model and the simplicity of the model. A smaller value of AIC indicates the candidate model has a better fit to the observations. For observation u_{ij}, $i = 1, \ldots, N$, $j = 1, 2$, the AIC of a bivariate copula family c with parameters θ is defined as

$$AIC := -2 \sum_{i=1}^{N} ln[c(u_{i,1}, u_{i,2}|\theta)] + 2k \tag{3.22}$$

where $k = 1$ for one parameter copula and $k = 2$ for the two parameter (t-copula). Given the distributions of the marginals and copula, a set of heat stress variables has identified for dependence structure modeling for regional crop growth during heatwave actives using Eq. 3.12.

In this study, the causal relationships between heat stress on regional crop growth from daily weather changes has modeled using a hybrid Bayesian network, which a Bayesian network contains both discrete and continuous random variables.

Network structure has learned based on the causal relationships and outputs from statistical tests, as discussed above. To apply vine-copula in the hybrid data type Bayesian network model, values of discrete variables have classified into levels except for heat stress variable and NDVI. Such that, the dependent modeling between a heat stress variable and NDVI can be obtained when given the levels of the discrete parent. Model performance in probability prediction of crop growth using heat stress variables has examined using simulation. A Bayesian network model with high performance in probability prediction should have the same high skills in representing the dependence

structures of local distributions within the network structure [7, 37, 148, 75]. Simulation results from CBN have compared to the results from a Gaussian-based Bayesian network with the same network structure, which local distributions are assumed to follow a Gaussian distribution. The 'best' network model is then used to explore heat stress risks for regional crops.

3.3.4 Spatial and temporal analysis

Although heatwaves typically occur in late summer months when the temperature reaches the highest period of the year (i.e., July, August, and September), recent studies have found heatwave occurrence tends to be more frequent and earlier over recent decades due to the increasing global temperature [28]. Earlier heatwave occurrence can result in severe damage to young crops that have a low resistance to high temperatures. Knowing the affecting date of heatwaves on regional crops can help agriculture producers do better anticipate and respond to heat stress risk to lessen its impact/s. Six candidate starting dates from May to July (May 1, May 15, June 1, June 15, July 1, and July 15) have tested using Spearman's correlations from Eq. 3.15. Spearman's correlations have performed to each regional heat stress variable and NDVI on the starting dates. A correlation coefficient is replaced by 0 if it has the absolute Spearman's coefficient is less than 0.2. Averaged absolute value of the coefficients has computed for comparison, and the best starting date has selected by the one has the highest value.

Spatial analysis of minimum day affecting by heat stress was performed from a set of the significant regional heatwaves from the best model outputs. A heatwave event is significant if it has at least one heat stress variable that has a dependence relationship with regional crops. A region with a smaller value in minimum days is likely to have higher heatwave risks in frequency than other regions. The result of this spatial analysis provides a better understanding of regional crop resistance to heatwaves and environmental heat stress risks from heatwave frequency.

Weather conditions during heatwave actives are playing a key role in the development of regional crops. Identify regional extreme weather conditions provides a better understanding of regional crop growth under heat stress. Identification of extreme weather variables for regional crops has done from the causal relationships in the network structure from the best Bayesian network model. Extreme weather variables include daily maximum temperature, precipitation, and relative humidity. As the daily

maximum temperature is the primary driver for detecting heatwaves, it is the base extreme weather variable for all study regions. The computation of CSI contains both temperature and precipitation in its equation. Precipitation is identified if a region has any CSI-based heatwave. Other than that, precipitation and relative humidity can be identified if it shows a dependence relationship with DNVI during heatwave actives.

Long-term sensitivity of heat stress risks has performed using heatwave intensity and frequency. Heatwave intensity has measured using the averaged daily maximum temperature during heatwaves over the study period from 1987 to 2012. A region has high values in both heatwave frequency and intensity is likely to have more severe damage from heat stress risks. A region has a high frequency of heatwave and small heatwave intensity suggests the regional crop is facing a higher than normal heatwave occurrence in the region. While a low heatwave frequency and high heatwave intensity suggests agricultural producers should pay attentions on the detected heatwaves on their regions due to the disaster damage of heatwaves to crops. This long-term test can provide useful information for long-term agriculture planning from extreme hot temperature risks.

Inter-comparison of short-term sensitivity of heat stress risk has done using heatwave frequency and intensity on the last five study years (2008 to 2012) minus the heatwaves occur on the first five study years (1987 to 1991). Increased heatwave frequency and intensity suggests the region is going to have an increasing disaster risk from heatwaves. While decreased values in heatwave frequency and intensity suggests a decreasing heatwave impact in the region. This inter-comparison can provides a future heatwave scenario for agriculture producers based on the evidence from the historical observation.

3.4 Results

The structure of the copula Bayesian network (CBN) is shown in Fig. 3.2. Network random variables include marginal distribution, minimum days, variable type, heatwave index, coefficient of Spearman correlation, independent test, and frequency are in discrete data type and heat stress variable, and NDVI are in continuous data type. Levels of the discrete network variables have shown in the dashed window below each discrete variable. The values of the coefficient of Spearman's correlation has split into two levels at an absolute threshold of 0.2 as we found copula works better in

dependence modelings when the variables have an absolute coefficient higher than 0.2. The independent test between the heat stress variable and NDVI was based on the p-value obtained from Eq. 3.18. If the p-value was less than 0.05, the two variables have a dependent relationship. Observation of heat stress variable in the network is determined by the levels of heatwave index, variable type and minimum days. Example of a heat stress variable has a CSI in heatwave index, heat degrees in variable type, and four days in minimum days indicates the heat stress variable has the observation of heat degrees obtained from the CSI-based heatwaves with a minimum heat duration of four days. NDVI was linked by the parent nodes of heat stress variable, marginal distribution and region. Dependence structure modeling between heat stress variable and NDVI using vine-copula has done using the marginal distributions and a copula distribution from the minimum AIC from Eq. 3.22.

Model outputs of copula Bayesian network (CBN) highlight a set of heatwaves that are affecting regional crops. We classified all the detected heatwaves from model output by heatwave index and minimum days. If heatwaves were detected from the same heatwave index, but in different minimum days, the smaller value would be used. In such a way, agriculture producers can easily classify heatwaves by index and knowing when a detected heatwave will start affecting their regional crops. Model simulations of probability predictions of crop growth under heat stress risks has performed using the heat stress variables from the classified heatwaves. The comparison result of crop growth simulation in the CBN and Gaussian-based Bayesian network model have performed. The comparison result shows an improvement in the model performance of NDVI probability prediction has found in CBN. A sample comparison result for a prairies region located in eco-zone 10 has shown in Fig. 3.3. The development of regional crop over the detected heatwaves has predicted using two heatwave variables, heatwave magnitude with a 3-days in minimum days from HWMId (HWMI.M3) and precipitation with a 4-days in minimum days from CSI. Heatwaves detected from CSI have a higher frequency of occurrence than HWMId. The simulation result shows that the dependence structure modelings between NDVI and heat stress variables were better in CBN than in the Gaussian-based Bayesian network. The simulation results in CBN are very similar to the empirical observations, which can provide high accuracy in probability prediction. CBN has chosen as the best model for heat stress risks modeling.

Spatial analysis of the 'best' starting date of heat stress risk is shown in Fig. 3.4. The 'best' starting date was selected by the highest value of averaged absolute

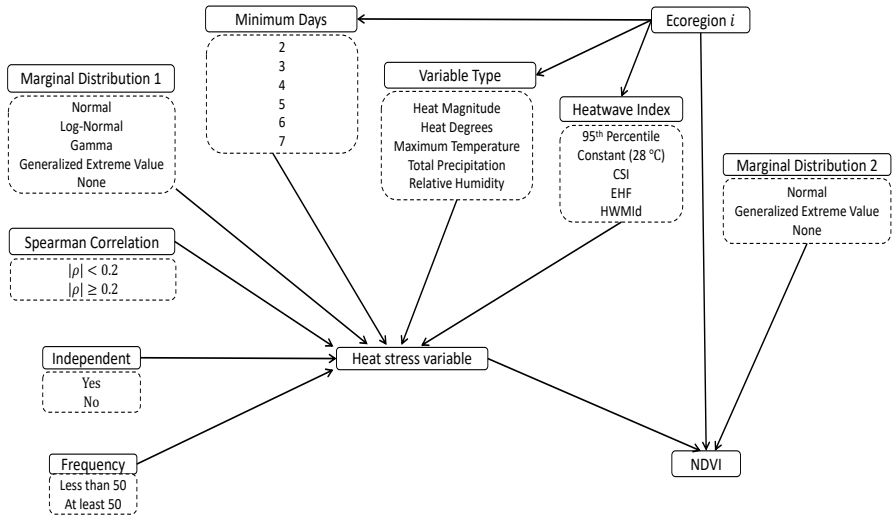


Figure 3.2 A learning process of the paired-copula Bayesian network. Network variables in the hybrid Bayesian network structure are in discrete data types and have classified into levels for model learning except heat stress variable and NDVI. The causal relationship between heat stress variable and NDVI has learned with the paired-copula approach from the selected candidate distributions and the copula family distributions.

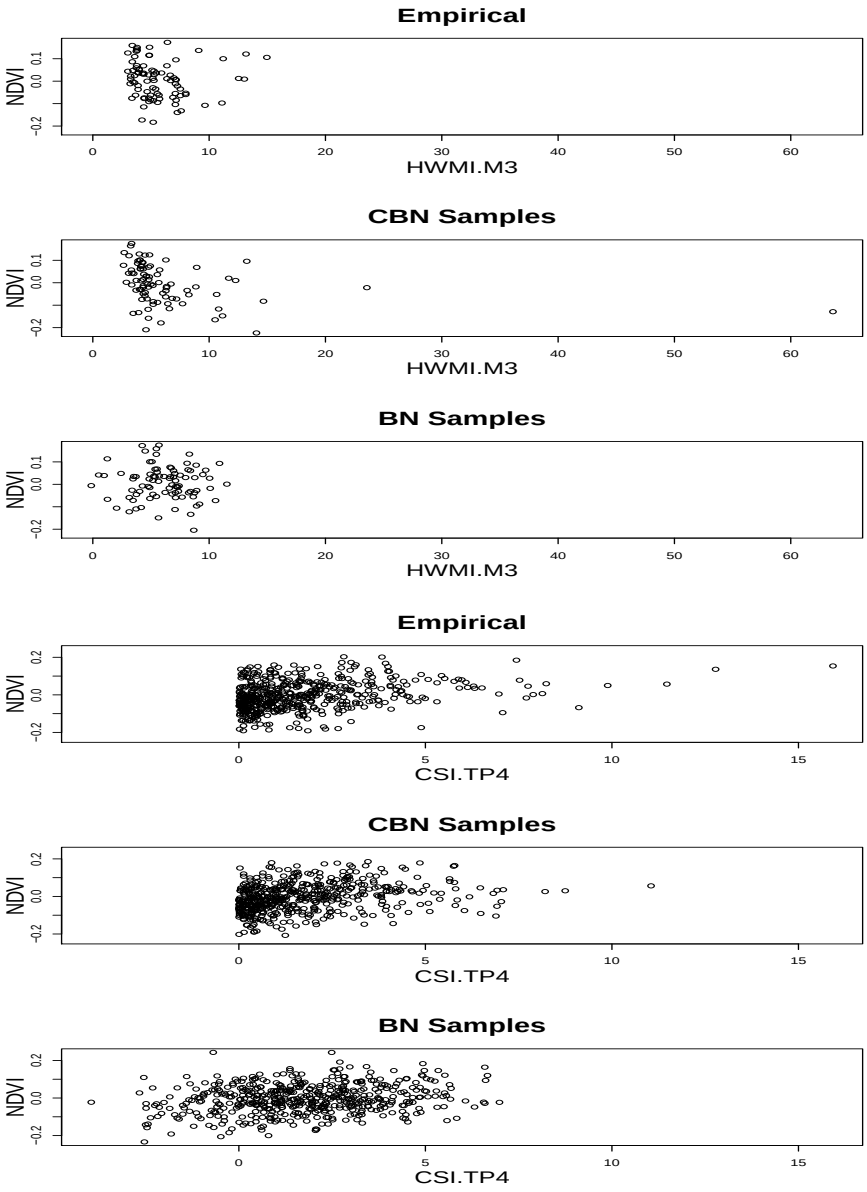


Figure 3.3 An example of simulation for NDVI probability predictions. Model simulation has performed using two classified heatwave types, heatwave magnitude with a 3-days in minimum days from HWMId (HWMI.M3) and precipitation with a 4-days in minimum days from CSI. Model simulations for NDVI probability predictions from CBN has compared to the empirical observations and a Gaussian-based Bayesian network.

coefficient from Spearman's correlation in a range of dates at the beginning (01) and the middle (15th) day of the month from May to July. The. Study regions located in eco-zones 9, 10, 13, and 14 were shown in the left and eco-zone 7 and 8 were shown on the right. Subfigures on Fig. 3.4 from top to bottom are May, June, and July. The result shows Canada's eco-zones has more than half of its agriculture regions are affected by early heatwaves of the growing season. The earliest affecting date at May 01 has found in most of the eco-zones except eco-zone 8. Spatial crops paired by eco-zones 7 and 13, 8 and 14, 9 and 10, have a similar response to the affecting date from heat stress. A spatial pattern on the affecting date has found. Heat stress is starting to affect the crop growth on the regions located on the coastal eco-zones 7 and 13 and the center eco-zones 9 and 10 in May, then extend to the nearby neighbor eco-zones 8 and 14 in June and forward to eco-zones 7, 8, 13, and 14 in July. The crop grows on the Canadian prairies (eco-zones 9 and 10) are very sensitive to heat stress, most of the crops within these two eco-zones has heat stress from May. While crops grow on the two coastal eco-zones tend to have two kinds of crops, the more sensitive crops have heat stress affects from May and the less sensitive crops are from July.

Fig. 3.5 shows the spatial analysis result of the minimum days of regional crop affected by a heatwave. The minimum days has selected from a range value from 2- to 7- days from the heatwaves detected in two-weeks observation intervals. The result shows that local crops grow across Canada's agriculture eco-zones has minimum days of 2-, 3-, 4- and 6-days affected by a heatwave. Most of the regions have a minimum day between 2 and 3 days, except two regions in eco-zone 13 are in 4-days and one region in eco-zone 7 is in 6-days.

Fig. 3.6 shows the significant weather conditions affecting regional crop growth during heatwave actives. Weather conditions include temperature only (T, empty circles), temperature and precipitation (TP, solid circles), temperature and relative humidity (TR, empty triangles), and all three weather variables (TPR, solid triangles). Heat stress affecting the regional crop growths in eco-zone 13 tends to relate temperature with few regions that have affected by precipitation and humidity. None of the crops in eco-zone 14 has affected by temperature and precipitation conditions during heatwaves actives. The growth of crops in eco-zone 8, 9, and 10 are more likely affected by complex weather conditions of temperature and precipitation, or temperature, precipitation, and humidity. The development of crop grows on eco-zone 7 are likely affected by humidity conditions (TR, and TPR) during heatwave actives.

Long-term spatial comparison of heatwave intensity (averaged maximum tempera-

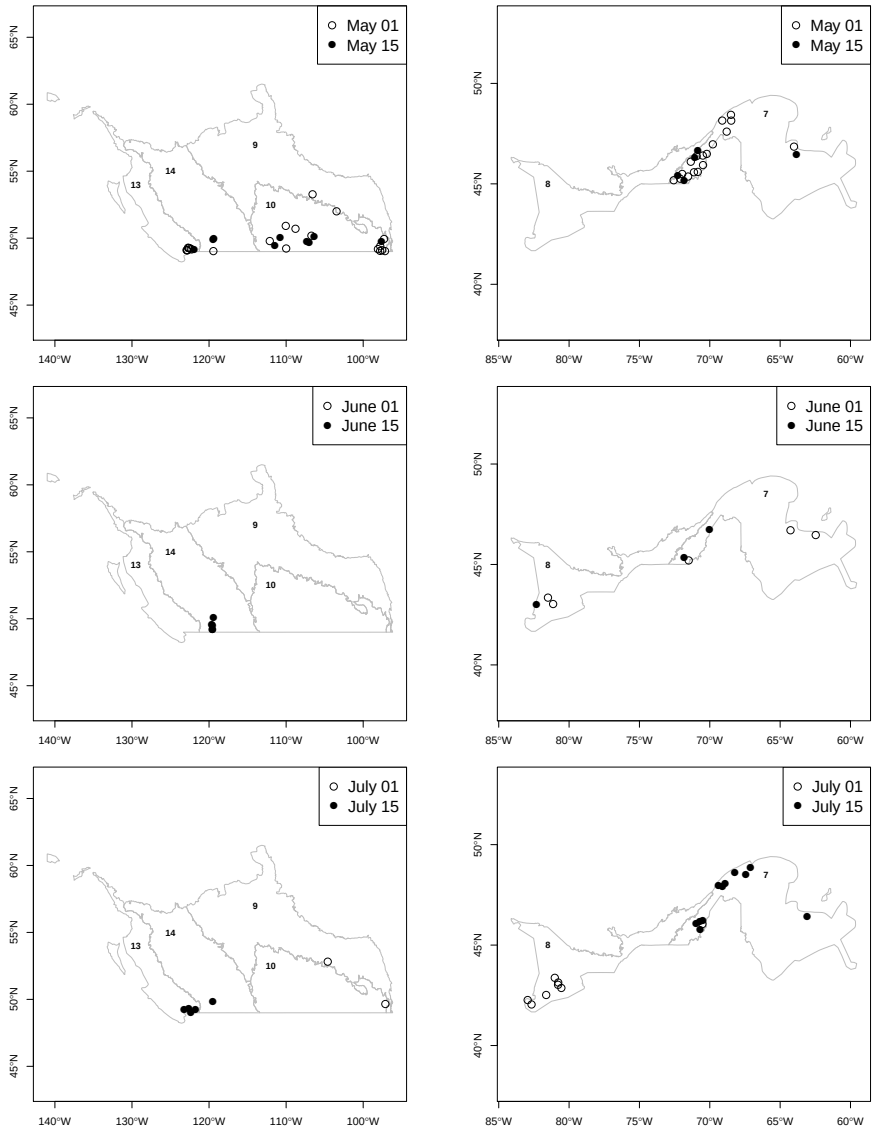


Figure 3.4 Spatial analysis of the estimated starting date of heat stress risk across the eco-zones 7 and 8 (right side figures), and 9, 10, 13, and 14 (left side figures). The 'best' starting date has selected at every first day 01 (empty circles) and the fifth day 15 (solid circle) for the months of May (top), June (middle), and July (bottom).

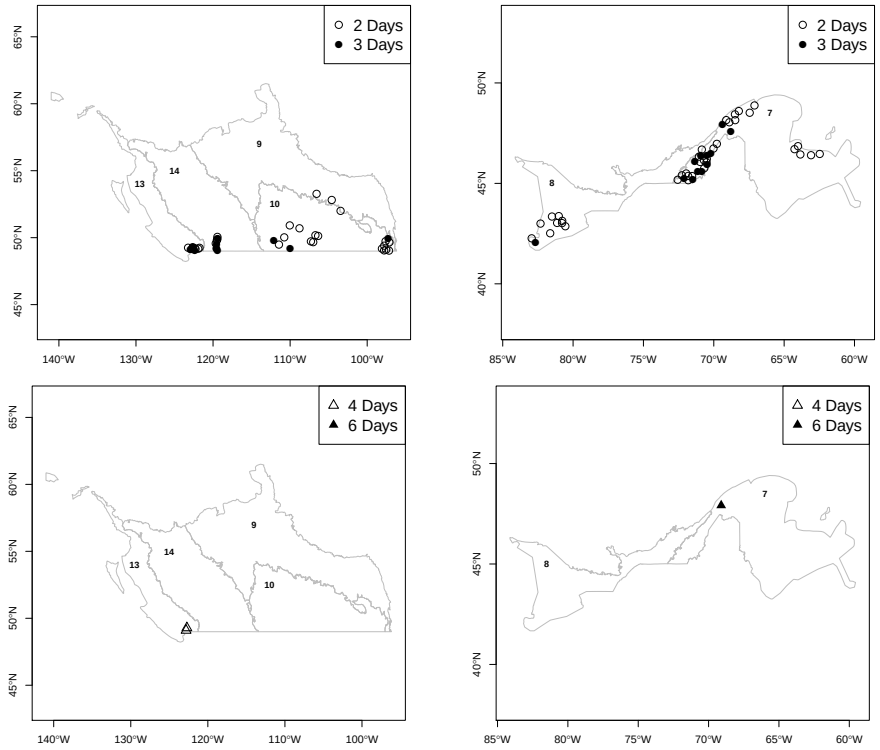


Figure 3.5 A spatial representation of minimum days of heatwaves occurrences on regional crops across Canada. The minimum days has defined from a range of days from 2 to 7 in two-weeks observation intervals. Minimum days of 2-Days (empty circles), 3-Days (solid circles), 4-Days (empty triangles), and 6-Days(solid triangles) have found. A small value in minimum days indicates a regional crop has a low resistance to heatwave events and is more likely to has a higher than normal in heatwave frequency days

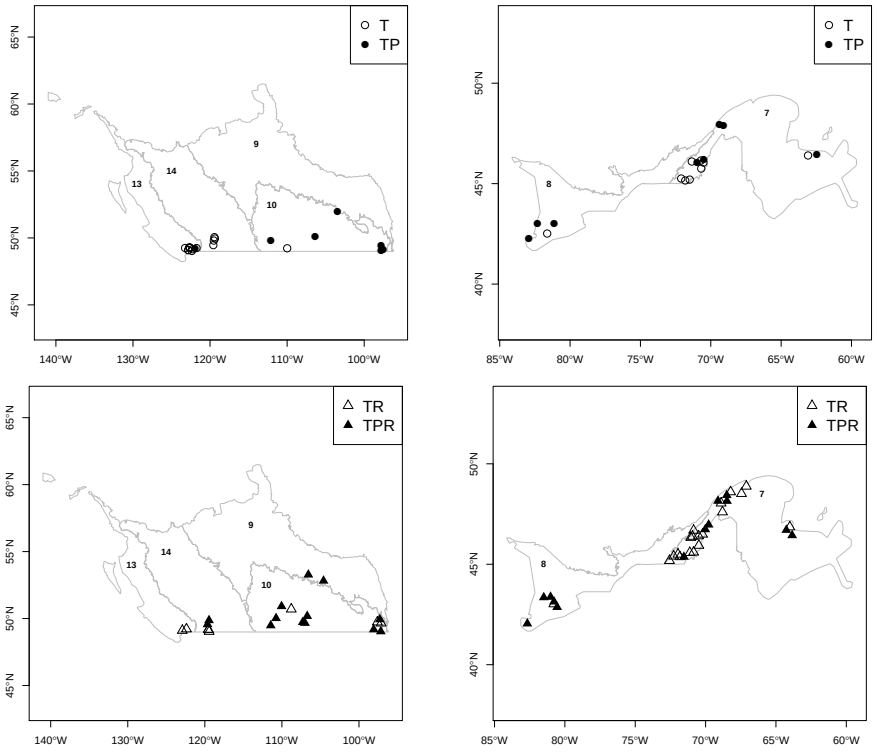


Figure 3.6 Extreme weather conditions on regional crops during heatwave actives. Extreme weather conditions were characterized by temperature (T), total precipitation (P), and relative humidity humidity (R). Significant weather conditions include temperature (empty circles), temperature and precipitation (solid circles), temperature and humidity (empty triangles), and temperature, precipitation, and humidity (solid triangles).

ture) (top) and frequency (bottom) have shown in Fig. 3.7. Heatwave intensity has evaluated by the averaged maximum daily temperature over the heatwave days. The heatwave frequency has evaluated as the annual heatwave rates in days. The heatwave intensity shows the impact score of a heatwave day on the growth of regional crops. A high value of heatwave intensity indicates that the regional crop has a high heat stress risk on heatwave days. The heatwave frequency shows the average number of heatwave days in the agricultural region. A region with a high value in heatwave frequency indicates that it has a high frequency of attacking by heatwaves during hot summers. The values of averaged maximum temperature and frequency in the plots have scaled in circle points and has highlighted in solid circles. A large size circle point in heatwave intensity or frequency indicates a region has a high risk of high daily maximum temperature or annual heatwave days, respectively, in hot summers. The solid circle points in the heatwave intensity and frequency plots indicate the region has a daily maximum daily temperature higher than 28 °C on average and at a rate of more than 100 heatwave days annually, respectively. For heatwave intensity, regions with high daily temperature on heatwave days have found on the south-east of eco-zone 7 and 8, the center and the east of eco-zone 10, and southern of eco-zone 13 and 14. Regions in eco-zone 9 at low risk of heatwave intensity. For heatwave frequency, the most heatwave actives regions have found from eco-zones 9 and 10, which most of the regions in these eco-zones have more than 100 heatwave days in annually. Regions on eco-zones 8, 13, and 14 tend to in a low risk of heatwave frequency on average.

Inter-comparison of short-term changes in heatwave intensity and intensity have performed between two periods (1981- 1991 and 2008- 2012). The result has shown in Fig. 3.8 for the changes in intensity (top) and frequency (bottom). The circle points indicate an increasing heatwave intensity and frequency, respectively, has found. The triangle points indicate a decrease in changes has found. The size of the points indicates the amount of difference. For most of the regions in eco-zones 7, 8, 9, and 14, there is a slight increase in changes in heatwave intensity and frequency have found. Partially, some significant changes in heatwave intensity and frequency have found in eco-zone 10 and 13. Most of the regions on the center of eco-zone 10 have a slight decrease in both heatwave intensity and frequency, but two regions showing a decrease in heatwave intensity and an increase in heatwave frequency. Significant reductions on both heatwave intensity and frequency have found on the east of eco-zone 10. Most of the regions in eco-zone 13 are showing a significant increase in both heatwave intensity and frequency.

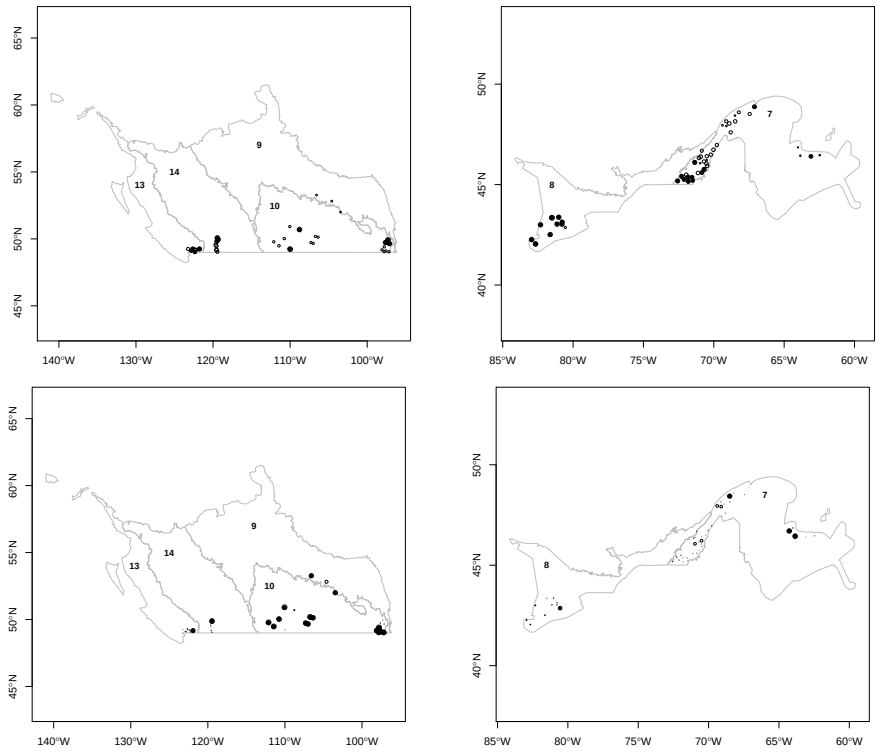


Figure 3.7 Spatial mapping of regional heatwave intensity (top) and frequency (bottom) for the agricultural eco-zones (eco-zones 7 and 8 (right sides), and 9, 10, 13, and 14 (left sides)). The values of heatwave intensity have estimated by averaged maximum daily temperature during heatwave actives. Both values of heatwave intensity and frequency have scaled over the study regions and plotted as circle points. A small size indicates a low heatwave intensity or frequency, respectively, have found over the study period from 1987 to 2012. The solid circles in the heatwave intensity plots indicate the region has a higher than 28 °C on the averaged maximum daily temperature. And the solid circles in the heatwave frequency plots indicate the region has a more than 100 heatwave days in annually.

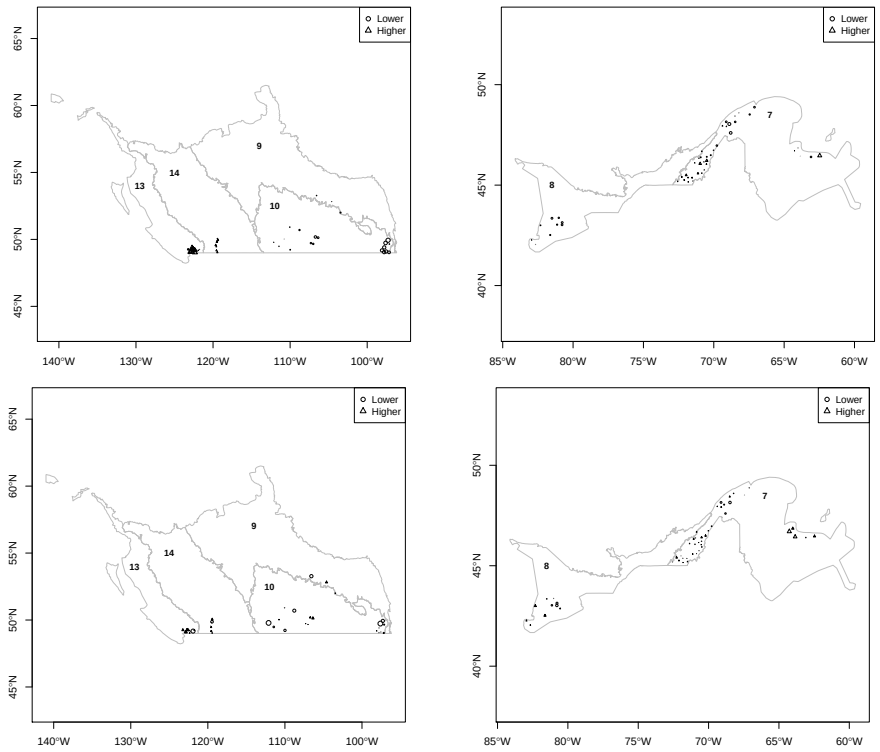


Figure 3.8 Inter-comparisons heatwave intensity (top) and frequency (bottom) change between two five-year periods (1987- 1991 and 2008- 2012). The values of change in heatwave intensity and frequency have computed by using the values obtained from the period from 2008 to 2012 minus the values obtained from another period from 1987 to 1991. The circle points in the plot indicate a decreased value has found in heatwave intensity and frequency, respectively. The triangle points indicate an increased value has found in heatwave intensity and frequency, respectively. The size of the points shows the different amounts.

3.5 Discussion

The simulation result for probability predictions of crop growth on a sample region shown in Fig. 3.3 shows regional heatwaves can be detected by a combination of multiple heatwave index, and model performance in probability predictions has improved in a copula-Bayesian network (CBN) from the Gaussian-based Bayesian network. The simulations of NDVI changes in CBN have high accuracy to the empirical observations. The simulations from the Gaussian-based Bayesian network shows model accuracy can be biased by the observations when the sample size is small. The Gaussian assumption in the Bayesian network does not work well in tail dependence modeling, especially for small events.

Fig. 3.4 shows the estimated starting date of heatwave affections on regional crops. The spatial plot shows that crops grow across the Canada agricultural regions and have different affecting date from heatwave because environmental heat stress and crop resistance for heat stress varied in regions. Early affecting date has found on eco-zones 7, 9, 10, and 13 in May, indicating regions in these eco-zones are having a long period of heat stress. The temporal pattern of starting date had found from early in the eco-zones on the center (9 and 10) and the two coastal eco-zones (7 and 13) of Canada in May, to the neighbors of eco-zones 8 and 14 in June, and last to eco-zones 7, 8, 13, and 14 in July. This time pattern provides some pieces of evidence for the heat stress movement. The minimum day result shown in Fig. 3.5 indicates that most of the crops grow in Canada's agriculture land are at a high risk of summer high temperature as they have small value (2 and 3 days) in minimum heatwave days. The results from Fig. 3.6 shows spatial heat stress risks across Canada varied in eco-zones. Regions in eco-zone depended by temperature; regions in eco-zones 8, 9, 10, and 14 depended by a combination of temperature, precipitation, and humidity; while eco-zone 7 tends to depend by temperature and humidity.

Spatial analysis of long-term heat stress in heatwave intensity and frequency shown in Figs.3.7 highlighted the regions at high heat stress risks (intensity and frequency). Most of the regions in Canada are in a high risk of heatwave intensity. The regions located on Canada's most agricultural active area (eco-zones 9 and 10) have high risks in heatwave frequency. The inter-comparison of heat stress risks between the two periods in Fig. 3.8 shows a slight increasing heat stress risks (heatwave intensity and frequency) have found for most of the regions in Canada. Eco-zone 13 has both a significant increase in heatwave intensity and frequency, indicating the regions in

this zone are facing increase risks from heat stress. Although eco-zones 9 and 10 are showing in high heat stress risks in the long-term, as shown in Fig. 3.7, the result in Fig. 3.8 shows a decrease in both heatwave intensity and frequency, which indicating that heat stress risks in these regions have decreased by time.

3.6 Conclusions

Heat stress is one of the major disaster risks responsible for significant yield losses for regional crops across Canada. Crop resistance to heat stress varies in regions and crop types. With uncertainty in heatwave detection, the influence of heat stress on the development of regional crop are often in misunderstanding, making it challenging to develop efficient strategies for heat stress management. In this study, we developed a copula Bayesian network to identify heat stress variables for Canada's regional crops using heatwaves detected from the five selected heatwave index. The identified heat stress variables were used to simulate and provide probability predictions for the possible changes in crop development under Heatwaves. The simulation result shows an improvement in model predictions of crop development changes from a Gaussian-based Bayesian network. Using outputs from the CBN model, this study highlights essential heat stress information for regional crops includes heatwave affecting dates, extreme weather conditions, heat stress sensitive, intensity, and frequency. By providing regional heat stress risks in both long-term and short-term, agriculture producers can have a better understanding of heat stress risks for their regional crops. With uncertainty in crop types and strategies information, future work stemming from the current study will require additional data such as: regional crop types, yield loss, plant growth periods (flowering and berries), strategies efficiency, and water-use. The developed CBN model aims to provides skillful strategies to minimize the yield losses from regional heat stress by optimizing the timing of water-use. We actively hoping the result of this study address a better understanding of heat stress risks on Canada's agricultural regions and helps agriculture producers improve their decision-making, in protecting their crops from summer heat stress.

Chapter 4

Disease Risk Forecasting with Bayesian Learning Networks: Application to Grape Powdery Mildew (*Erysiphe necator*) in Vineyards

Powdery mildew (*Erysiphe necator*) is a fungal disease causing significant loss of grape yield in commercial vineyards. The rate of development of this disease varies annually and is driven by complex interactions between the pathogen, its host, and environmental conditions. The long term impacts of weather and climate variability on disease development is not well understood, making the development of efficient and durable strategies for disease management challenging, especially under northern conditions. In this chapter, we present a probabilistic, Bayesian learning network model to explore the complex causal interactions between environment, pathogen, and host for three different susceptible northern grape cultivars in Quebec, Canada. This approach combines environmental (weather, climate), pathogen (development stages), and host (crop cultivar-specific susceptibility) factors. The model is evaluated in an operational forecast mode with supervised and algorithm model learning and integrating Global Forecast System (GFS) Ensemble Reforecasts (GEFSR). A model-guided fungicide spray strategy is validated for guiding spray decisions up to 6 days with a 10-day forecast of potential spray efficacy under rain washed off conditions.

The model-guided strategy improves fungicide spray decisions; decreasing the number of sprays, and identifying the optimal time to spray to increase spray effectiveness. The main result of this chapter has been accepted for publication in *Agronomy*, see [80].

4.1 Introduction

4.1.1 Economic importance of grapes in North America

Grapevine growers across North America grow a wide range of grapes, mostly producing table grapes, raisins, and wines. Grapes and their derivative products represent a significant contribution to the economy of North America. As reported for 2015 in California, the primary U.S. wine-grape growing region, grapes and their derivative products contributed $57.6 billion to the state's economy and $114.1 billion to the U.S. economy. In Canada, grapes are produced primarily in Ontario (66% of acreage), British Columbia (33% of acreage), Quebec (5% of acreage) and Nova Scotia (3 % of acreage). In 2015, the Canadian wine and grape industry contributed $9 billion to the Canadian national economy, with the province of Quebec contributing $1 billion [39].

The grapevine production areas in Quebec are characterized by very cold winter temperatures as low as $-30\,^{\circ}\mathrm{C}$ to $-35\,^{\circ}\mathrm{C}$, a cold and wet spring, hot summers, followed by potentially a cold fall. This weather influences the selection of grape varieties and the type of wine to be produced. In this weather berries can reach maturity within the period without frost (i.e., bud emergence after latest spring frosts and berry ripening prior to the first killing frost in the fall) and survive cold winter conditions [21]. Most grapevines are still protected during the winter months with soil or geotextiles [21]. In the spring, such winter protection is removed after the last risk of spring frost to prevent frost damages to primary and secondary flower buds [32]. These conditions influence both grapevine growth and disease development. The grapevine development cycle is characterized by rapid vegetative growth from late May to end of July, follow by slower growth in August and September. The northern weather also influences grape diseases including powdery mildew caused by *Eriysphe necator* (Schw.) Burr., (synonym *Uncinula necator*). Powdery mildew (hereafter, PM) progresses slowly from late May to end of July and rapidly in August and September. Hence PM incidence, when expressed as the proportion of disease leaves is very low, until generally late July to early August, constraining earlier scouting and disease forecasting.

4.1.2 Grape Powdery Mildew (PM) disease

Grape PM caused by *Eriysphe necator* (Schw.) Burr., (synonym *Uncinula necator*) is an obligate parasite affecting only plants in the genus *Vitis*. For more than 150 years, PM has been a significant challenge for grape production [99]. Since the 1850s, research on this disease has been undertaken in response to several major epidemics in Europe. The disease causes both direct (crop losses) and indirect (reduced vine vigor) damages. Moderate to severe disease epidemics cause reduced yield (lower berry weight), delayed berry maturity, and altered wine composition and sensory characters [17, 46, 101]. Current guidance for disease mitigation calls for the application of fungicide sprays at a repeating 7 to 14 day interval. This approach tends to be of limited effectiveness as it does not consider the complex interaction between this pathogen and its host in relation to local weather and broader, regional climate variability. Over-spray is costly and harmful as the chemical can remain on berries, and the local environment. Effective management of grape PM requires the development of efficient and durable strategies to fine-tune fungicide application timings and amounts, taking into account the effect of environmental variability on disease development and pathogen dispersal.

The complex interactions between the pathogen and the host, influenced by climate and weather, drives the rate of PM development and its impact severity. Growth stage (ontogenic resistance) [38, 45] and grape cultivar genetics both influence grapevine susceptibility to this disease. Cultural practices that favour vegetative vigour may predispose the host to an increased development of PM. High grapevine vigour can also modify ontogenic resistance of leaves, delaying grapevine phenological stages such as veraison or harvest, or stretching the duration of the flowering, fruit set or bunch closure periods [134]. Grape PM can affect all above-ground parts of grapevine. A typical disease progression begins on the leaves, where lesions found on the undersides of leaves are the first visual indication of an infection. As the disease progresses, lesions become apparent on the upper sides of the leaves as well. In the absence of control, these lesions will increase both in size and number, causing premature leaf drop. On shoots, symptoms are brown to black irregular lesions that vary in size. On inflorescences and rachis, PM has the appearance of a grey to whitish powder. Severe infections of the rachis can result in premature drop of clusters. The disease can attack berries immediately after bloom through four weeks post-bloom. Infected berries are ash grey and quickly become covered with spores, giving them a floury appearance. Berries infected later during the period of susceptibility are prone to

splitting, making them susceptible to infection by other grapevine pathogens [44].

4.1.3 Impacts of weather on grape powdery mildew

PM is a polycyclic disease that evolves in two distinct phases: the primary infection, caused by ascospores (sexual spores), and secondary infections, caused by conidia (asexual spores). Its epidemiology is well-studied [40, 41, 42, 43, 29, 46, 65, 141, 140, 142]. The pathogen overwinters as cleistothecia which contain immature ascospores. In Eastern Canada, the cleistothecia are the likely source of primary inoculum. Dehiscence of the cleistothecia commonly starts at bud break and continues until the beginning of flowering [42]. Ascospores are released from cleistothecia in response to rain (greater than 2.5 mm) when the temperatures are between 6 and 24 °C; infection will not occur outside of this temperature range. Once released, ascospores that fall on young leaves cause the primary infections. Ascospore germination requires free water or high relative humidity. Infection takes place within an optimal temperature range of 20 to 25 °C, if there is sufficiently high leaf wetness over a duration of 4 hours. Once an infection is established, lesions will develop on infected leaves within 6 to 30 days, depending on temperature variability. Potentially large amounts of conidia are produced on these lesions. They are primarily dispersed by the wind [20, 22, 141, 140]. Unlike ascospores, conidia do not need free water for germination. Their germination is controlled by temperature, relative humidity, and light intensity. The optimal temperature for germination of conidia is 25 °C [31]. Most conidia germinate at a relative humidity of 40 to 100%. Relative humidity is often not a limiting factor for germination [23]. At temperatures of 23 to 30 °C, secondary infection cycles can be completed within 5 to 7 days [26]. At the end of the growing season, cleistothecia appear on infected leaves and berries. Vineyards are more likely to experience a severe damage (and suffer more extensive damage) if the initial infection occurs early in the season with temperatures suitable for the development of the initial lesion and ensuing conidia outbreak.

4.1.4 Management of grape powdery mildew

Current PM disease control strategies consist almost exclusively of schedule-based fungicide spray applications, involving grape growers regularly applying fungicide spray under constant or regular time intervals (e.g., ranging from 7 to 21 days) from the beginning to the end of the growing season. This is generally effective in managing

PM disease, but raises production costs, promotes fungicide resistance, and can be detrimental to both human and environmental health. There have been numerous attempts to link spray application timing and fungicide selection to weather and environmental information. The Gubler-Thomas program, developed at UC-Davis[1], was designed to create a disease risk index using daily average temperature and measured leaf wetness hours for both ascospores and conidial infections to guide fungicide spray strategies for several widely-used fungicide products (i.e., sulfur dust, micronized sulfur, and DMI fungicides (DeMethylation Inhibitors)).

The impact of PM on grape yield loss is closely related to the severity of conidial infection on both leaves and berries. Thus, application of fungicide spray when the first conidia infection occurs, can intercept and stop a disease outbreak before it can establish itself. This spray timing can also reduce the total number of applications required over a growing season and, in turn, spray costs. Caffi et al. (2011) developed a mechanistic model to predict the initiation time of a conidia infection using stochastic and dynamic process models of the lifecycle of overwintered ascospore, as influenced by daily weather conditions [15]. This model was validated over four years (2005 - 2008) in 26 vineyards in Italy and had a 94% accuracy in correctly predicting outbreaks. The host-pathogen model of Calonnec et al. (2008) provides spatial and temporal dispersal information of conidial infections in relation to factors such as the number, age, and pattern of healthy and infected organs, infectious leaf area, and the density of spores released [18].

The Gubler-Thomas program has become a primary disease management tool in California for grape disease control, but is not well-suited for PM disease control in Quebec because of this region's weather variability [20]. Instead, a model based on degree-days, developed by Carisse et al. (2009) and has been shown to reduce the number of spray applications by up to 55%, by determining the initial date to start the application of fungicide spray. This initial date matches the Eichhorn-Lorenz grape phenological stage 7 which is the stage in plant development when there are 2-3 young leaves fully expanded from shoot tips. The model is currently operationally deployed across Quebec and attains a disease control efficiency similar to that of a calendar/schedule-based fungicide program based on regular intervals rather than a model-based schedule [20, 22].

[1] `http://ipm.ucanr.edu/DISEASE/DATABASE/grapepowderymildew.html`

4.1.5 Problem statement

PM management tools have primarily been developed for temperate weather grapevine production areas. More complex, multi-variate models are needed to improve our understanding of PM epidemiology under northern weather and to improve the accuracy of model-based forecasting of the optimal timing of fungicide application. While schedule-based programs can help to significantly reduce the frequency of fungicide sprays over a growing season, contact fungicides are susceptible to being washed off when cumulative rainfall or irrigation after a fungicide spray reaches 25 mm or more. This is an major issue that many existing fungicide spray programs do not consider. Models also need to consider how the efficacy of fungicide spray impacts PM epidemiology over longer time frames. An improved model-guided approach to PM management is needed that can be implemented operationally by grape producers to reduce production costs and environmental (i.e., weather and climate-related) risks. A model-guided approach could increase the long-term economic and environmental sustainability of viticulture under current and future climate.

4.1.6 Research objective

The primary research objective is to develop a probabilistic model that can provide reliable forecasting of PM risk to guide fungicide spray strategies under local weather conditions and changing regional weather variability. Such a model needs to include relevant measurement parameters and variables, so that it can be calibrated and operational deployed in vineyards. It also needs to integrate weather and weather information in near-real-time (NRT) in order to ensure fungicide management is more effective, efficient, and less costly. We present a novel probabilistic model (i.e., Bayesian network model) developed to forecast the risk of PM, validating it against experimental data from Quebec, Canada for three northern grape cultivars - Chancellor, Geisenheim-318-57[2] (hereafter, Geisenheim-318), and Frontenac. These cultivars represent high, medium, and low susceptibilities to PM, respectively. The model combines information from observational data and weather models, integrating weather, host stage, cultivar and autoregressive life-cycle process considerations and factors involved in their complex interaction. The model is tested under two competing learning modes (i.e., supervised and algorithm) that alters the model's structure (i.e.,

[2]Vitis International Variety Catalog: `http://www.vivc.de/index.php?r=passport%2Fview&id=4710`

association of factors and the strength of their influence). This network learning generates insights on essential variables and complex interactions that exist for specific sites (i.e., vineyard ecosystems). By using both structural and parameter optimization, and allowing the model to determine the best model design in a probabilistic and dynamic way, this approach extends existing, mechanistic methods. Such methods typically consider far fewer possible variables/factors, are static, and are calibrated by tuning a fixed number of parameters so are difficult to be applied across different sites and regions, and do not have the ability to predict dynamics, nor generate future forecasts. The forecast-skill of the best-performing model in predicting disease risk or future disease incidence is assessed using Global Forecast System (GFS) Ensemble Reforecasts [53]. GEFSR provides an 11-member ensemble of historical reforecasts and real-time forecasts of weather conditions with lead times of up to 16 days. A model-based fungicide spray program providing up to 10 days of forecast disease incidence with fungicide application guidance for up to 6 days is showcased driven by GEFSR weather forecasts.

4.2 Materials and methods

4.2.1 Study site

The study site was the Agriculture and Agri-Food Canada (AAFC) experimental farm located in Frelighsburg, Quebec, Canada (lat. N 45.05° and long. W 72.86°) (Fig. 4.1). Grapevines are arranged in spatial grid units with 3m $\times$ 0.9m ($row \times column$). Hourly weather data, including temperature, wind velocity, relative humidity, and rain intensity, were monitored at the canopy level (around 1.5m height) within the grapevine canopy and collected through the growing season (May - September) from 2000-2011. The lower part of Chancellor and Geisenheim-318 were covered with 40-60 cm of soil during winter seasons, and the soil was removed at the beginning of the next growing season. Disease sprays (other than PM) for downy mildew and Botrytis bunch rot and insecticides to control flea beetle were applied. Other cultural practices were performed according to standard commercial viticulture practices. PM disease incidence (hereafter, DI) was measured as the ratio of infected leaves over the total number of sampled leaves. PM incidence (Hereafter, DI) was assessed twice weekly from bud break (mid-May) until harvest (mid to late September) by looking at two shoots on eight vines per plot, selected randomly each time. At each sampling

Figure 4.1 The experimental farm (top) and the weather station (bottom) used to monitor weather variables at the grape canopy level from 2000 to 2011.

time, the total number of leaves and total diseased leaves were recorded. A leaf was considered to be diseased if it had one or more lesions. DI was used as the response variable instead of disease severity because visual estimation of severity, expressed a proportion of leaf area diseased, is very difficult and not sufficiently reliable when done under field conditions.

The variability of seasonal weather (May to September) and DI for grape PM for each of the three susceptible cultivars is shown (Fig. 4.2). Typically, summer temperature at the experimental farm varied between 15 °C in May to a maximum of around 20 °C in July and August, reducing slightly to around 15 °C in September. Daily temperatures from July to September varied from a minimum of 10 °C to a maximum of around 27 °C, which is suitable for infections of both ascospore and conidia. Summertime at the experimental farm was humid with many rainfalls. Daily rainfall totals of over 25mm were recorded. Daily relative humidity varied between 60 to 95%, with an average value of around 70% from July to September. Daily wind speed varied from as low as 2 km/h up to around 20 km/h, with average wind speed around 7 km/h. The local weather and regional weather conditions at this site are suitable for the development of grape PM, but vary enough that standard spray

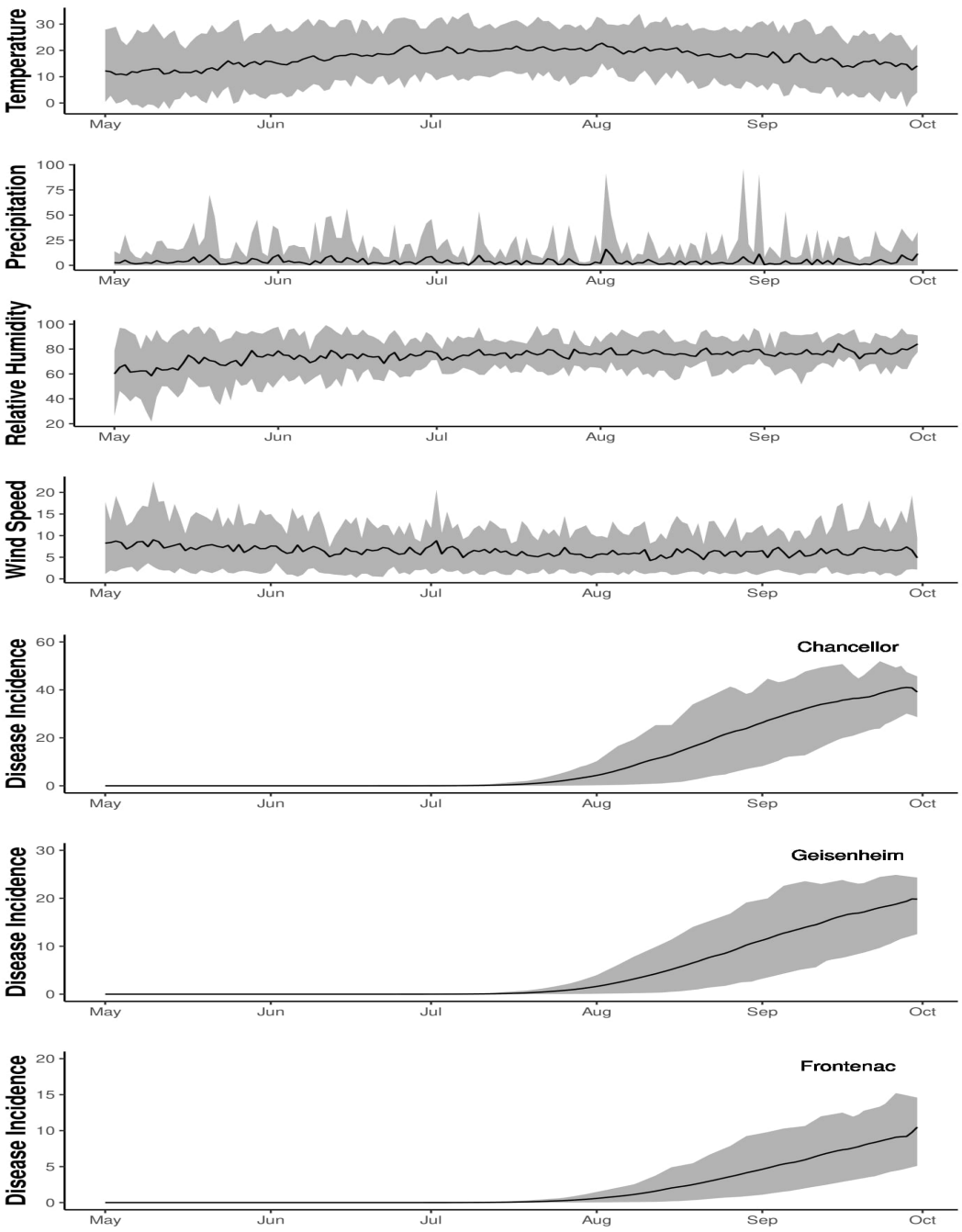


Figure 4.2 Seasonal weather variability and disease incidence (DI) (% infected/non-infected leaves) of grape PM through the growing season for northern hybrid grape cultivars with differing susceptibility (Chancellor, Geisenheim-318, Frontenac). The mean (solid line) is shown varying between minimum (lower bound) and maximum (upper bound) observed values (2000-2011).

regimes (based on fixed interval or basic weather-guided protocols) cannot provide adequate protection. Suitable temperatures and frequent rainfalls early in the growing season causes ascospore to be released from over-wintered cleistothecia during a bud-break and dispersed within the vineyard over distances that depend on the prevailing wind conditions. Applied fungicide is also washed off by frequency and/or intense rainfalls, leading to high DI, even if a regular spray regime is followed. Measured DI for the northern grape cultivars vary with weather and climate, exhibiting different levels of susceptibility (Fig. 4.2) . Although disease monitoring starts in May, the first signs of infection are detected in the beginning of July. DI increases rapidly in August, reaching a peak in mid-September, then decreasing until the end of the growing season. The maximum DI over a growing season is dependent on the timing of disease outbreak (Fig. 4.3). Earlier disease outbreaks result in higher DI in some years (i.e., 2001 and 2010); the reverse pattern was also observed in other years (i.e., 2004 and 2009).

4.2.2 Global Forecast System (GFS) Ensemble Reforecasts (GEFSR)

Reforecasts, also known as hindcasts, are retrospective numerical weather prediction (NWP) forecasts made for extended periods using a fixed version of a weather forecast model and data assimilation system. Ideally, the reforecast NWP system is the same as one that has been used operationally. Reforecasts provide valuable data for developing statistical forecast models for environmental variables. The availability of an extended, relatively homogeneous reforecast dataset – one that matches the statistical characteristics of an operational forecast system – allows robust calibration of statistical forecast model parameters.

The US National Oceanic and Atmospheric Administration (NOAA)'s 2nd. generation GEFSR dataset provides meteorological variables used as predictors in the disease risk forecast model [53]. GEFSR produces retrospective ensemble NWP forecasts every day at 0000 UTC using the National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System (GEFS) (version 9.0.1 ca. 2012). GEFS is one of the models that contributes to the North American Ensemble Forecast System (NAEFS). This is a joint project to provide long-range, probabilistic weather forecasts by the national weather services of Canada, United States, and Mexico. The GEFSR ensemble consists of 1 control forecast and 10 perturbed ensemble members, with

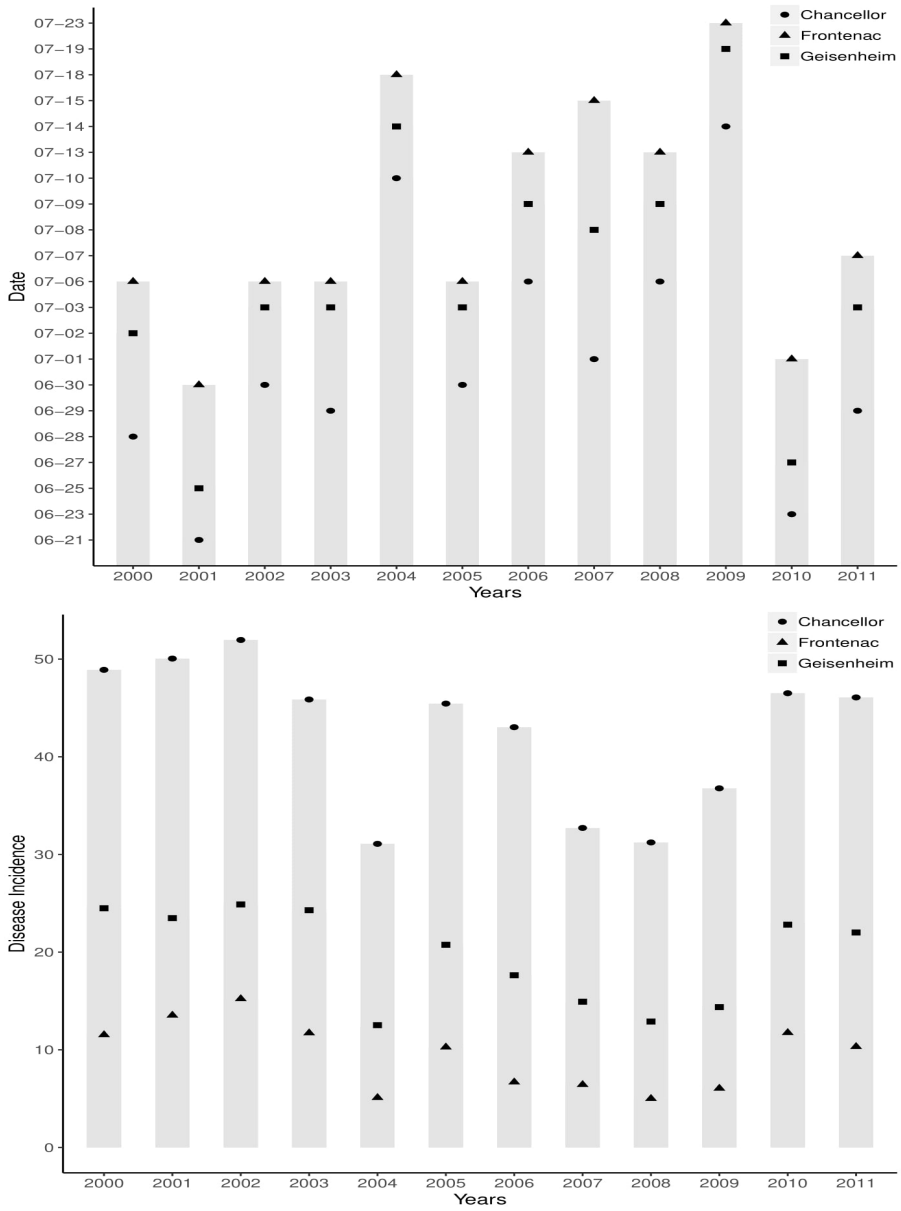


Figure 4.3 Annual first disease date (top) and the maximum DI (bottom) of grape PM for the three susceptible cultivars: Chancellor (circle); Geisenheim-318 (square); and Frontenac (triangle).

archived reforecasts available from December 1984 until the present. Reforecasts are recorded at 3-hourly intervals for lead times from 0 to 72 hours and at 6-hourly intervals after 72 hours. During the first 8 days of the GEFS reforecasts, the model is run at T254L42 resolution (equivalent grid spacing of 40 km at 40 degrees latitude and 42 vertical levels). From 8 to 16 days, the model is run at T190L42 resolution (54 km horizontal grid spacing). Reforecasts from GEFSR are statistically consistent with forecasts from the operational 00 UTC run of GEFS [53]. All model predictors are ensemble averages over the 11 forecast members. For PM disease risk forecasting, a set of available reforecasted weather and climate variables were selected from the GEFSR archive (i.e., minimum temperature, maximum temperature, total precipitation, sea-level pressure, specific humidity, U-component, and V-component wind speed (near-surface)), with relative humidity computed from specific humidity and sea-level pressure.

4.2.3 Grapevine development

The model assumes the susceptibility of grapevines to PM is variable through a growing season. The development of grapevines can be divided into five stages using interval values from the commonly used plant growth index called Cumulative Growing Degree Days (CGDD). CGDD is a cumulative sum of daily GDD (growing degree days) calculated by the daily temperature degree above a specific base temperature (T_{base}), usually taking values between 0 °C and 10 °C [87, 144, 85], a base temperature of 10 °C is mostly used for grapevines [69, 109]. CGDD is given by,

$$CGDD_i = \sum_i (T_i - T_{base}) \tag{4.1}$$

where T_i is the daily mean temperature calculated as the average value between the daily maximum and minimum temperature. Accumulation of daily values starts on April 1st for a given growing season. The specific phases of grapevine are defined as: *Off-Season* ($CGDD < 20$), *Bud break* ($20 \geq CGDD < 254$), *Flowering* ($254 \geq CGDD < 680$), *Setting* ($680 \geq CGDD < 1100$), and *Veraison* ($1100 \geq GDD$).

4.2.4 PM disease development

The development of grape PM is highly dependent on grapevine development, local weather conditions, and regional climate variability. Empirical equations that track

the development of grape PM were calibrated and validated for our study site for later integration into our probabilistic forecast model. Parameter estimates were either tuned or fixed based on published scientific literature and internal Agriculture and Agri-Food Canada (AAFC) reports. It was assumed that the pathogen over its life-cycle has the same temperature response to changing daily temperature, but the time (days) it takes to complete a full development cycle can change, such that some spore development phases may take more than a day to be completed under optimal weather conditions.

A temperature effect rate function was specified to relate the daily rate of PM development (as a percentage) to daily mean temperature. The temperature effect rate function is the inverse of the latent period equation [3, 4].

$$\theta(T) = \begin{cases} \frac{15}{138-7T} & 6\,^{\circ}\mathrm{C} \leq T < 17\,^{\circ}\mathrm{C} \\ \frac{(m+n)^{(m+n)}}{n^n m^m} T_n^n (1-T_n)^m & 17\,^{\circ}\mathrm{C} \leq T \leq 32\,^{\circ}\mathrm{C} \end{cases} \tag{4.2}$$

with,

$$T_n = \frac{T - T_{min}}{T_{max} - T_{min}} \tag{4.3}$$

where T is the daily mean temperature. T_{max} and T_{min} are the temperatures at maximum and minimum thresholds for either mycelial growth or spore infection, respectively. $\theta(T)$ equals zero when $T \leq 6\,^{\circ}\mathrm{C}$ or $T > 32\,^{\circ}\mathrm{C}$. It was validated with AAFC data on the number of days required to complete a latent period under varying daily temperatures. PM development rapidly increases when the temperature is between $6\,^{\circ}\mathrm{C}$ and $17\,^{\circ}\mathrm{C}$ because the curve of the effect rate function is concave up (Fig. 4.4). The rate of temperature effect is high (above 75 %) when the temperature is between $17\,^{\circ}\mathrm{C}$ and $31\,^{\circ}\mathrm{C}$, and reaches a peak when daily mean temperature at $26\,^{\circ}\mathrm{C}$ matches the optimal temperature for PM development. The rate of temperature effect then drops at higher temperatures.

The cumulative proportion of ascospores ready for release (PAR) is a proportion of over-wintered ascospores within a vineyard and is computed as a function related to CGDD from Equation 4.1. We assume that the vineyard has the same amount of over-wintered ascospores population, N, every year, so that N can be omitted from the computation, and the primary infection caused by ascospores would only depend

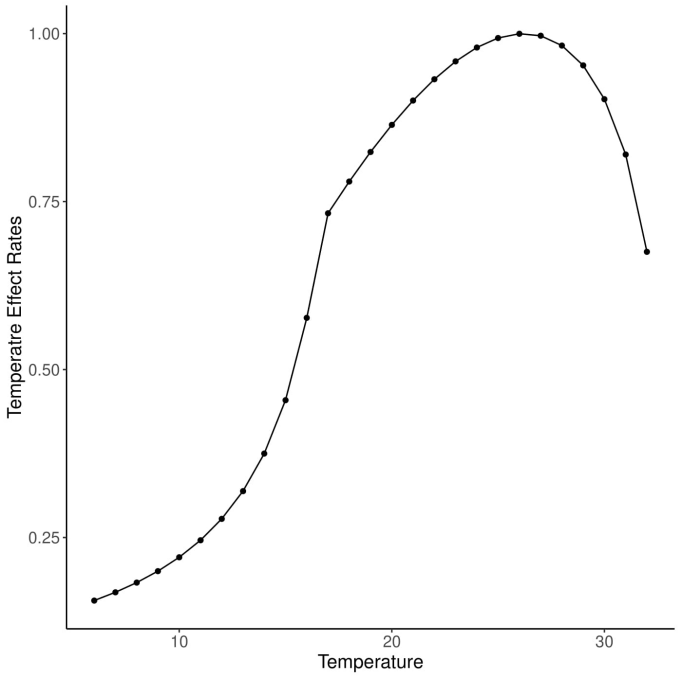


Figure 4.4 Temperature effect rate plot of the developing rate of PM in response to daily temperature from 6 °C to 32 °C.

on the values of PAR. The equation of PAR is:

$$PAR_i = \exp[-\alpha \cdot \exp(-\beta \cdot CGDD_i/100)] \tag{4.4}$$

where $CGDD_i$ is the Cumulative Growing Degree Days for day i. The proportion of PAR is a rate value that falls into an interval of $[0, 1]$. The process of ascospore release is considered to have stopped when PAR reaches 1. Equation 4.4 was first introduced for ascospores infection modeling at the bud break stage [15]. The daily rate of ascospore maturation (AMR) is calculated as the daily change of PAR within the day in $(i-1, i)$ and is expressed as the daily rate of ascospores ready for release.

Ascospores released from leaves during the AMR stage will germinate when there is at least 2 mm of rainfall R_i [42, 111], and the daily temperature is between $6\,^{\circ}\text{C} \geq T \leq 31\,^{\circ}\text{C}$ [20, 22, 18]. The germination rate of ascospores ADR depends on the leaf wetness duration in hours (WD) and canopy temperature (T). Caffi et al. (2011), using data from Gadoury et al. (1990)derived the following ascospore germination rate equation (ADR) [15, 42],

$$ADR_i = 1 - \delta \cdot \exp(-\lambda \cdot T_i^2 \cdot WD_i) \tag{4.5}$$

where $R_i \geq 2$ mm and T is the daily mean temperature with $4\,^{\circ}\text{C} \leq T < 30\,^{\circ}\text{C}$. $ADR_i = 0$ otherwise. Through the process of germination, released ascospores are transferred into ungerminated (AUG) and germinated (AOG) ascospores. AUG spores remain on the leaf leave and wait for the next germination opportunity when environmental conditions are suitable, and AOG is preparing for primary infection. The equations of AUG and AOG depend on the amount of ungerminated ascospore at day $(i-1)$, the amount of newly released ascospores, and germination rate at day i.

$$AUG_i = (AUG_{i-1} + AMR_i) \cdot (1 - ADR_i) \tag{4.6}$$

$$AOG_i = (AUG_{i-1} + AMR_i) \cdot ADR_i \tag{4.7}$$

The primary infection rate (PIR) at any given day i can be calculated as a function of Equations 4.2 and 4.7:

$$PIR_i = AOG_i \cdot \theta(T) \tag{4.8}$$

where $\theta(T)$ is computed from Equation 4.2. Once primary infection has occurred,

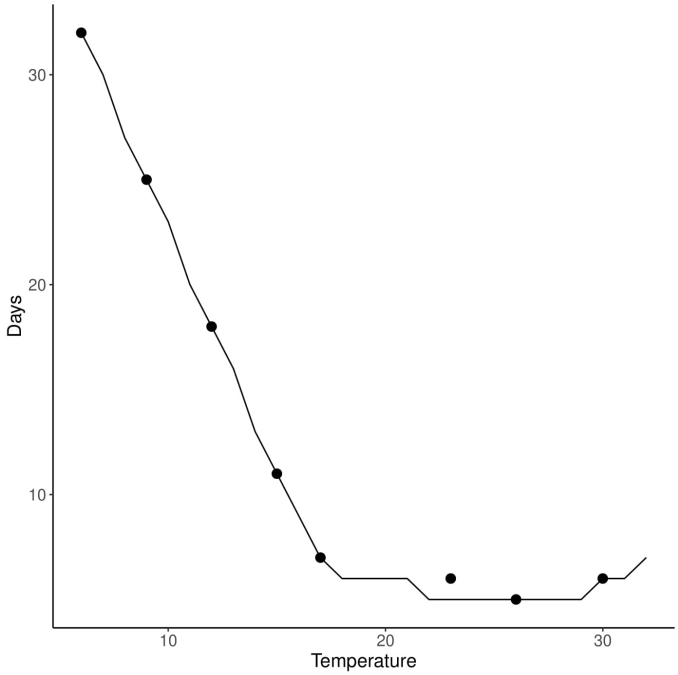


Figure 4.5 Number of days needed to complete a latent period responses to daily temperatures (°C). The solid line represents the estimated time to complete a latent period from Equation 4.9 in relation to measured days to complete a latent period.

lesions are generated on infected leaves to produce conidia following a latent period $\rho(t)$. The latent period $\rho(t)$ was calculated using the minimum time (days) required to complete a latent period and the daily temperature effect rate from Equation 4.2 [4, 51, 18]. The latent period equation is:

$$\rho(T) = \begin{cases} 46 - \frac{7}{3 \cdot T} & 6\,^{\circ}\mathrm{C} \leq T < 17\,^{\circ}\mathrm{C} \\ \frac{\rho_{min}}{\theta(T)} & 17\,^{\circ}\mathrm{C} \leq T \leq 32\,^{\circ}\mathrm{C} \end{cases} \tag{4.9}$$

A latent period is considered complete when $\sum[1/\rho(T)] = 1$; this releases a large amount of conidia spores for dispersal, which cause the secondary infection. Figure 4.5 shows the latent period response to temperature described by Equation 4.9 (solid line), calibrated to parameter estimates from the AAFC vineyard data.

The spatial dispersal of conidia depends on the condition of wind. Willocquet et al. (1998) provide a dispersal rate equation of conidia spore using wind duration, wind speed, and the ages in days of the lesions [140]. The dispersal rate (DR) equation is:

$$DR = \exp(r \times u + b)/(1 + \exp(r \times u + b)) \tag{4.10}$$

Table 4.1 Calibrated, site-specific parameter values for PM model.

Parameter	Description	Eq.	Values	Reference
n_I, m_I	Shape parameters	4.2	1.055-0.338	[3, 4]
T_{max}, T_{min}	Temperature thresholds for pathogen infection	4.2	33-5 ($^\circ$C)	[31], [115]
n_g, m_g	Shape parameters	4.2	1.24-0.27	[3, 4]
T_{max}, T_{min}	Temperature threshold for latent period	4.2	33-5 ($^\circ$C)	[3], [4]
α, β	Location Parameter, Growth Rate Parameter	4.4	1.95 - 1.91	[15]
δ, λ	Location Parameter, Growth Rate Parameter	4.5	0.969- 0.0004	[15]
ρ_{min}	Minimum time to complete a latent period	4.9	5 (days)	AAFC
I_0	Maximum infection rate	4.11	0.53	[18]
r, b	Spore Dispersal Rate by wind, Capacity Rate of a lesion to withhold conidia	4.10	[140] (Table 2)	[140]

where u is the daily wind speed (km/h). r and b are the parameters of the equation. Values of r and b vary depending on the age of a lesion [140]. The infection by conidia is a rain-free process and depends only on daily temperature. The daily infection rate of conidia (Secondary Infection Rate) SIR can be expressed as a function of the temperature effect rate from Equation 4.2. Calonnec et al. (2008) indicate that there is a maximum threshold of around 53% of the total conidia spores that can cause infection on leaves given optimal temperature condition [18]. SIR is given by,

$$SIR = I_0 \cdot \theta(T) \tag{4.11}$$

where $\theta(T)$ is from Equation 4.2, and T_0 is the maximum daily infection threshold of conidia under optimal temperature, $T_0 = 0.53$. Infection of conidia starts from the day of the first completed latent period, which is the day of conidia dispersal starts, until the end of the growing season.

Table 4.1 provides the values and detailed descriptions of the equation parameters. Factors retained for inclusion in the network modeling of disease risk of grape PM include: primary infection PIR, secondary infection SIR, the daily number of completed latent periods LP, and dispersal rate DR.

4.3 Bayesian network learning model

Bayesian networks (BNs), also known as belief networks, are directed acyclic graphical (DAG) probabilistic models that are widely used to represent the complex joint probability distributions between network variables. The complex causal relationships between random variables X_i are split into multiple local distributions and represented as a DAG. Independent random variables within a DAG are represented as nodes. They are linked by edges from the direct causes as determined by the conditional

dependencies probability P. Each variable is conditionally dependent on its effects and independent of its non-effects. Conditional dependencies can be estimated either by using statistical and computational methods or incorporation of expert experience. Let X denoted as a set of random variables in a DAG, then the full joint probability distribution has expressed as:

$$P(X) = P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|\pi_i) \tag{4.12}$$

where x_i and π_i, $i = 1, 2, \ldots, n$, are child and parent nodes, respectively. The distribution of child nodes depends only on their parent nodes, the number of which is limited. The global joint distribution is split into local distributions. The forecast model was coded and implemented using the statistical programming language R (version 3.4.2) making use of validated, open-source algorithm libraries: bnlearn (structure learning), mltool (calculating MSE), and other plotting packages (i.e., Rgraphviz, graph, plot3D, ggplot2, RColorBrewer).

4.3.1 Supervised and algorithm learning

Fitting a Bayesian network to data, usually called *learning* has two forms: structural learning and inference learning. Structural learning also has various forms, and can be based on expert knowledge, termed supervised learning, or based on statistical or computational learning algorithms, termed algorithmic learning. This later approach uses multiple learning algorithms to link random variables within a network into a DAG. Random variables within a network can be discrete, continuous, or hybrid data types, where hybrid is a network containing both discrete and continuous random variables. In a hybrid Bayesian network, network learning to establish causal relationships proceeds under certain restrictions: a child of continuous data type can have either discrete or continuous data type parent nodes, but a child of discrete data type must have discrete parent nodes. This logical connection between child and parent nodes helps to avoid linking problems in structure learning, providing better estimations for parameter learning.

Both supervised and algorithmic learned were employed for the Bayesian network to learn the complex interactions between a grapevine and PM disease under varying environmental conditions. For the supervised learning, the causal relationships within the Bayesian network were linked to the calibrated empirically-derived equations

described previously in Section 4.2.3 and 4.2.4.

The grape PM risk assessment model, P_{maxacc} [20], was used to specify airborne conidium concentration for structural learning of the model. This model is developed from the Richards model using CGDD based on a 6 °C threshold. The Richards model was selected because it had used the observations from the same experimental farm site in this study, which was responsible for reducing fungicide spray for disease control by as much as 40 % from 2004 to 2007. We examined the Pearson correlations between the DI for the three susceptible grape cultivars. The PM risk assessment model had different base temperatures ranging from 1 °C to 13 °C. 3 °C was selected as the base temperature because it had the highest value in the correlations. The PM risk assessment model is:

$$P_{maxacc} = 1.0755(1 + e^{-0.0042 \times CGDD})^{1/(1-1.0169)} \tag{4.13}$$

where $CGDD$ is the cumulative growing degree days from Equation 4.1.

Numerous approaches exist for implementing algorithm learning, with the two main approaches being constraint-based and score-based learning. Constraint-based learning involves developing relationships based on the framework from [135], which uses a conditional independence test inductive causation (IC) to learn the Bayesian network structure. Structural learning involves determining the so-called Markov blanket of each node in the network and is defined as containing the only knowledge needed to predict the behavior of a node and its children within a larger network. Learning algorithms include the Grow-Shrink (GS) [84], the Incremental Association (IAMB) [133], the Fast Incremental Association (Fast-IAMB) [145], and the Interleaved Incremental Association (Inter-IAMB) [133]. In score-based learning, a network is learned from the network's best goodness of fit by assigning a network score from the application of a general heuristic optimization technique to each candidate network. A commonly-used example of a score-based learning algorithm, for example, $Hill - Climbing$ (HC), learns a network structure using a step progress process. In HC learning, a graphical score is computed for a randomly assigned network structure as a network score at the beginning of the process. A new score is computed from the assigned network structure by adding, deleting, or reversing an arc's direction one at a time. The new graphical score will become the network score if it has a higher value than the previous one. The process of the computation repeats again and again until the network score reaches to its maximum. The corresponding network structure, which

has the maximum network score, has the best goodness of fit to the network data. HC is chosen as the algorithm learning approach in this study because of its flexibility and power to handle the hybrid data types existing in our network data. To enhance the robustness of the network structure generated from score-based causal relationships between child and parents sets, a bootstrap sampling technique with an iteration of 5000 is performed, and two criteria – conditional dependence strength of above 80 % and arc directions appearing in more than 50 % of the iterations – are applied to the HC structure learning.

The first disease infection likely occurs between the end of June and July, which is during the flowering stage of the grape (Fig. 4.3). Establishing the initial modeling date is important because this can influence the model accuracy of disease risk. Modeling results were examined for four selected starting dates to determine the best date for initial disease risk modeling. These dates were selected under different assumptions: 1) the flowering date as the time of primary infection onset; 2) the mid-flowering date as the likely onset of secondary infection; 3) July 1st when the infection spreads to other plants, and 4) using the actual observed date of infection start as the estimated model start date may improve disease risk predictions. At the beginning of each growing season (April 1st), the timing of the grapevine growing and mid-flowering stages were also estimated by the CGDD index (Equation 4.1) using daily mean temperature. The estimation of the PM growth factor is based on selected dates using Equations presented in Section 4.2.4 and are specified for both the supervised and algorithm learning.

Table 4.2 summarizes the random variables used for network learning, including the additional considered day-to-day variability of relative humidity. Bayesian network modeling in both supervised and algorithm modes are trained for the three susceptible grape cultivars using observational data from 2000 to 2010 and then tested against actual pathogen occurrence and progression in 2011. Model forecast accuracy was evaluated comparing supervised and algorithm learning and a range of starting dates. The best-performing model was then used to examine the forecast performance and optimal forecast window length using GEFS input data.

4.3.2 Forecast skill under different learning modes

The performance of the forecast model under supervised and algorithm learning was evaluated and inter-compared using: 1) model skill in goodness-of-fit testing, which

Table 4.2 Variables for structural learning by the Grape PM model.

Parameters	Descriptions	Data Type
DI	Disease incidence	Continuous
DI_P	Recent disease incidence before a latent period	Continuous
DR	Dispersal rate	Continuous
LP	Number of latent period are done at current day	Discrete
PIR	Primary infection rate	Continuous
$P_{maxacc3}$	Degree-Days based risk assessment model (3 °C)	Continuous
RH	Relative humidity	Continuous
PS	Plant stage	Discrete
SIR	Secondary infection rate	Continuous
Type	Susceptible cultivar type	Discrete
T_{mean}	Daily mean temperature	Continuous
TP	Daily total precipitation	Continuous
WS	Wind speed	Continuous

examined how well the model worked in accurately reproducing training data; and 2) model performance in prediction, providing benchmark information about model accuracy using forecast weather data. The learned model structure identified a subset of random variables listed in Table 4.2. Random variables in a subset were selected by removing one or more unclear but considerable factors: the degree-days based assessment model (*Pmaxacc*), relative humidity (RH), and plant stages (PS) one at a time from the full data. A total of eight cases were examined: Case 1, full data; Case 2, remove RH; Case 3, remove PS; Case 4, remove *Pmaxacc*; Case 5, remove PS and RH; Case 6, remove RH and *Pmaxacc*; Case 7, remove PS and *Pmaxacc*; and Case 8, remove PS, *Pmaxacc*, and RH. Over-learning or over-fitting is a common problem in parameter learning, in which the network model is forced to accommodate data or parameters that do not contribute information; this results in degraded model predictive skill. k-fold cross-validation, where k is the number of years in the training data, has been applied to prevent over-learning. In this approach mean-absolute-error (MAE) and root-mean-squared-error (RMSE) are used as metrics to identify prediction bias and for variance comparison. Metrics used for model skill comparison for prediction results similarly include the MAE and RMSE generated for the 2011 growing season.

Cross-validation using the k-fold method is a re-sampling technique often used to evaluate the prediction performance of a machine learning model using limited data sample as training data. In k-fold cross-validation, the training data are split into k subsets of equal size which are used to assess how the omission of one subset affects the learning of a Bayesian network from the rest of the subsets, by generating model predictions for the omitted subset. In this study, the training data set were split into k subsets by years. A single year is randomly removed to act as testing data, and the rest of the data are used for parameter learning using both supervised and algorithm learning approaches. The corresponding results of bias (MAE) and variance (RMSE) loss function can be applied to measures deviation and discrepancy, separately, to determine how close model predictions were to the actual outcomes. Both MAE and RMSE can be defines as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y^{(i)} - \hat{y}^{(i)}| \tag{4.14}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2} \tag{4.15}$$

where $(y^{(i)}) - \hat{y}^{(i)}$ is the residual or error between the model predictions and the actual values. A lower value of MAE and RMSE indicates a model exhibiting better performance.

4.4 Model forecast evaluation

Sensitivity analysis (i.e., for a static or time-independent model) involves varying input variables, typically one at a time, to measure how it impacts a model's output. Scenario analysis involves varying input variables to measure its impact on a future value either as future model predictions (if internally forced), or as projections (if externally forced). In the context of dynamic or time-dependent (i.e., forecast) models, shifting an input (e.g., climate or weather input variable) that is time-dependent generates variation over time. Furthermore, when forecast models have cumulative variables or interactions between multiple variables, future model outcomes can also become dependent and coupled across time. In such cases, sensitivity and scenario analysis becomes more integrated and less independently defined.

To understand the effect of a warmer and colder weather on DI model forecasts for the three northern cultivars, we varied the input daily mean temperature by 2 °C in 2011. The runs with warmer and colder temperature were referred to as warm and cold year scenarios, respectively, and compared to the actual or baseline temperature at the study site in 2011. We also varied the model's forecasting window length from 1, out to 16 days, to assess how robust the model's forecasts of DI are over time, how sensitive its error is to the length of the forecasting window, and evaluate the GEFS reforecasted weather data as a model input for generating forecasts out to 16 days in advance. This analysis involved comparing the model's forecast error statistics (MAE and RMSE) over time (days). The Bayesian network model with supervised learning, was used with GEFS reforecasted weather data for climate/weather variable input. Starting April 1, 2011, host plant stages were estimated using historical weather data and used to identify the mid-flowering date (July 2nd) to initiate the forecast model. Canopy-height daily maximum/minimum temperature, total precipitation, air pressure, specific humidity, U/V component of wind were selected. Scalar mean daily wind speed was computed using the U/V-component wind variables. The daily mean temperature was calculated as the mean of the daily maximum and minimum temperature. The relative humidity was calculated from specific humidity using daily mean temperature and air pressure as inputs to the statistical R programming

function $SH2RH$ from the package of "*humidity*". Weather data was averaged from 11 ensembles and used as model inputs for DI forecasting at a 16-day window. For each day after July 2nd, the calibration of the Bayesian network model was performed using historical data from April 1st up to the present day and used for DI forecasting using data generated from GEFS.

4.5 Model-based fungicide spray program

A fungicide spray program was identified for maximizing spray efficiency for disease control based on the Bayesian network model and forecasting windows (refer to Sections 4.3.2 and 4.4). The program guides grape advisor or growers by providing the best times to spray fungicide for PM control in relation to disease risk (future disease incidence), local weather, and regional climate variability.

An optimal model-based fungicide spray program was generated under the following assumptions: 1) The rate of disease incidence given application of fungicide spray follows an exponential decay function of the form $A = A_0 * \exp(kt)$, where A_0, A, k, and t are the disease incidence at the current day, initial disease incidence, decay rate of disease incidence, and time in days, respectively; 2) Fungicide spray is washed off when cumulative precipitation reaches 25 mm or more from the date of spray application. The decay rate of fungicide spray efficiency remaining as a function of precipitation amount was simply estimated using a linear function of $y = -0.04 * P$, where y is the fungicide "effectiveness rate" under precipitation and takes a value between $[0, 1]$. p is the cumulative total precipitation and takes values from $[0, 25]$, with $p = 25$ if cumulative precipitation exceeds 25mm; and 3) The occurrence of precipitation events is independent of fungicide application events. That is, the occurrence of precipitation is not affected by fungicide spraying. By setting the amount by which an application of fungicide can reduce the incidence of powdery mildew from $A_0 = 20$ % to $A = 8$ % for 10 continuous days with no rain $t = 10$. We generate the value for $k = -0.0916$ and the equation used to estimate the disease incidence under fungicide spray control influenced by the uncertainty of precipitation is:

$$A(t) = A_0 * (1 - (1 - \exp(-0.0916t) * (1 - 0.04p)) \tag{4.16}$$

where $A(t)$ is the disease incidence after t days of fungicide spray. A_0 is the DI on the fungicide spray day. A disease incidence profile was then generated for fungicide spray

using forecast disease incidence and Equation 4.16.

4.6 Results

The performance of the forecast model was evaluated by varying the starting date (i.e., 4 selected starting dates) using a subset of variables (Table 4.3). After determining the best model starting date (i.e., first disease date) simulation runs using 8 different variable subsets were performed (Table 4.4). These tables provide a summary of the resulting mean-absolute-error (MAE) and root-mean-squared-error (RMSE) validation metrics for the northern grape cultivars. The forecast model was trained using data from 2000-2010 and validated using data for 2011.

Model forecast skill was evaluated based on k-fold validation statistics (MAE, RMSE) across the years 2000-2011 which removes k years at a time from the data and re-assesses forecast skill. The model was also validated by training the model on all of the historical data (2000-2010) (i.e., removing only data for year 2011) and then comparing its forecasts against 2011 data. Smaller values in both MAE and RMSE indicate higher forecast skill.

For supervised learning, the optimal initial date for start disease risk prediction was starting from first disease date in Case 1, containing all the network variables in the modeling structure. This run had the smallest values of cross-validated RMSE and predictive error in 2011 (MAE and RMSE). The smallest value of MAE (1.83) for training data was in case 6 with model initialed from the flowering date but with higher values in RMSE for training data and MAE and RMSE for prediction data in the first disease date. The highest values of MAE and RMSE in both training and prediction data were from Mid-Flowering data with Case 6, which does not include relative humidity and degree-days based risk assessment model in the model structure.

In algorithm learning, case 2 and 3 performed the best, which do not contain relative humidity or plant stage. The best model performance was from the first disease date in case 2 that had the smallest values of MAE and RMSE in both cross-validation and prediction. The highest values of cross-validated MAE (3.9) and RMSE (6.13) were for flowering date in case 2, which had small MAE (0.57) and RMSE (0.99), as with first disease date. The highest MAE and RMSE values in prediction was for flowering, which had high values of MAE and RMSE similar to mid-flowering. In general, cross-validated MAE and RMSE values were higher than the prediction values of these metrics in historical drought years of 2000, 2002, and 2008. Model accuracy of

disease prediction was also evaluated with and without drought conditions, by adding a binary factor identifying historical years of drought (2000, 2002, 2008). The best performing model for supervised learning with drought years factor being case 3 with model initialed from the first disease date, which has MAE and RMSE of 2.36 and 4.16 from training data and 1.13 and 1.52 from predictions, in separately. The best performing network structure in algorithm learning being Case 2 from the first disease date which has smallest values in MAE and RMSE from both training and prediction data than other selected dates. For both supervised and algorithmic learning, The model without drought had smaller values of MAE and RMSE than those for drought years.

Model performance under structural learning for the 8 network variable sets from the first disease date are shown in Table 4.4. We describe these specific cases here below in greater detail. In supervised learning, the smallest values of MAE and RMSE were from case 1 and 2, which contains plant stage and the degree-days risk assessment submodel. MAE and RMSE were higher when the model structure does not contain plant stage (cases 3 and 5) or degree-days risk assessment model (cases 4 and 6) and was at the highest values when model structures (case 7 and 8) did not contain both plant stage and degree-days risk assessment model. It has noted that network structures in algorithm learning are learned based on network scores from an algorithm; a child was linked only by the parents with significant influences. A child can have the same network structures from two sets of network random variables, with both contain the same parents with significant influences. In algorithm learning, the best performing network model is Case 1 and 2 with MAE (2.11) and RMSE (3.69) from cross-validation and MAE (0.56) and RMSE (0.87) from model prediction. There was no local distribution learned for disease prediction when the plant stage has removed from network variables (cases 3 and 5). While network variable does not contain both plant stage and degree-days assessment model (cases 7 and 8), MAE and RMSE were three times higher for training data (MAE 6.88 and RMSE 9.59) and more than ten times higher for prediction (MAE 5.21 and RMSE 7.34) than cases 1 and 2. When comparing the best model results between supervised and algorithm learning, supervised learning gained smaller values in MAE and RMSE from the training data from 2000 to 2011 with prediction MAE and RMSE in 2011 slightly higher than from algorithm learning.

The representative DAG for the best-performing model under supervised (case 2) and algorithm (case 2) learning is shown in Fig. 4.6 and 4.7. In the case of supervised

learning, the DAG is a plant stage based network. This network structure assumes: 1) the susceptibility of the grapevine to the PM development is influenced by changing weather and climate that varies by plant stage, 2) DI is independently affected by a set of factors including: precipitation (TP), primary infection rate (PIR), secondary infection rate (SIR), wind-speed (WS) influenced dispersal rate (DR), plant stage (PS), the degree-days based disease risk assessment model ($P_{maxxacc3}$), latent period (LP), past DI (DI_P); and genes cultivar types (Type). The network structure learned from supervised learning shows causal relationships based on existing knowledge, with dispersal rate (DR) of grape PM spores being influenced by wind-speed conditions, secondary infection (SIR), and temperature (Fig. 4.6). In algorithmic learning, causal relationships were generated from the independent random variables to maximum the network score of a network structure from bootstrap sampling technique with strength above 0.8 and direction above 0.5 from 5000 interactions. The learned model structure and linkages between variables from algorithmic learning (Fig. 4.7) shows DR not linked to DI as it was not identified as a significantly strong predictor. This suggests that DR needs to be better represented and that the current DR equation does not represent the observed PM epidemic. Significant correlations between grape cultivar types, plant stage (PS), and the risk assessment model ($P_{maxacc3}$) with DI was detected from the algorithmic learning.

The performance of the forecast model under supervised and algorithmic learning for the three northern grape cultivars for training (2000-2010), and prediction in 2011, is shown in Fig. 4.8. Both supervised and algorithm learning have similar model performance (both in MAE and RMSE). Model accuracy was high in the recorded drought years of 2000, 2002, and 2008 with values of MAE and RMSE were varying from 3 to 5 and from 4 to 7, in separately. The model performance worked well in non-drought years with MAE and RMSE were varying around from 0.56 to 2.6 and from 0.86 to 3.7, in separately. Model performance in both supervised and algorithm learning tended to be more and more accurate (smaller values in MAE and RMSE) in disease prediction in drought and non-drought years as the Bayesian network model has learned from more and more reliable input data. Disease performance in algorithm learning tends to be more sensitive in weather-related disease predictions than in supervised learning with MAE and RMSE in algorithm learning were higher than in supervised learning in drought years and smaller in non-drought years. The smallest value of MAE 0.99 and 0.56 in supervised and algorithm learning, in separately, found in 2011. The smallest value of RMSE 1.57 in supervised learning found in 2010 and

Table 4.3 Inter-comparison of model performance in parameter learning from both supervised and algorithmic learning in a total of 32 subsets of network random variables in four different model starting dates. The table shows the best model results of the network random variables and model start date. Besides, a binary factor of drought years (1 as in 2000, 2002, and 2008; 0 as in the rest of study years) was added as a parent to disease incidence to examine the model performance of disease incidence predictions in the hot years. Model performance has compared in mean-absolute-error (MAE) and root-mean-square-error (RMSE) from k-fold cross-validation, and model prediction skills in predicting disease incidence using observed weather variables in 2011. Smaller and non-negative values in both MAE and RMSE indicate higher model performance.

	Supervised					Algorithmic				
		Cross Validation		Prediction			Cross Validation		Prediction	
Without Drought	Data Case	MAE	RMSE	MAE	RMSE	Data Case	MAE	RMSE	MAE	RMSE
Flowering	6	1.83	3.71	1.45	2.65	3	3.31	5.46	2.03	2.7
Mid-Flowering	6	2.44	4.33	1.87	3.03	2	3.9	6.13	0.57	0.99
July 1st	1	2.13	3.82	1.4	2.27	3	3.51	5.36	0.95	1.37
First Disease Date	1	1.86	3.23	0.99	1.57	2	2.11	3.69	0.56	0.87
	Supervised					Algorithmic				
		Cross Validation		Prediction			Cross Validation		Prediction	
With Drought	Data Case	MAE	RMSE	MAE	RMSE	Data Case	MAE	RMSE	MAE	RMSE
Flowering	4	2.24	4.65	1.33	2.5	3	3.9	6.6	2.03	2.7
Mid-Flowering	4	2.93	5.41	1.71	2.83	2	4.32	7.78	0.58	0.99
July 1st	4	2.57	4.94	1.66	2.8	5	3.92	6.2	1.04	1.48
First Disease Date	3	2.36	4.16	1.13	1.52	2	2.64	5.09	0.55	0.86

Table 4.4 Comparison of model performance of parameter learning starting from the first disease date for 8 subsets of network random variables. Forecast skill (MAE and RMSE) from k-fold cross-validation is listed. Smaller and non-negative values in both MAE and RMSE indicate higher model performance. The dash − indicates there is no available model for forecasting DI using the selected network random variables.

	Supervised				Algorithmic			
	Validation (2000-10)		2011 Forecast		Validation (2000-10)		2011 Forecast	
Case	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
1	1.86	3.23	0.99	1.57	2.11	3.69	0.56	0.87
2	1.86	3.24	0.99	1.55	2.11	3.69	0.56	0.87
3	2.22	3.52	1.46	2.02	-	-	-	-
4	2.25	4.12	1.82	2.96	4.2	6.83	2.43	4.09
5	2.23	3.52	1.43	2.01	-	-	-	-
6	2.24	4.1	1.81	2.94	4.2	6.83	2.43	4.09
7	3.75	6.48	3.08	5.98	6.88	9.59	5.21	7.34
8	3.74	6.48	3.1	6.04	6.88	9.59	5.21	7.34

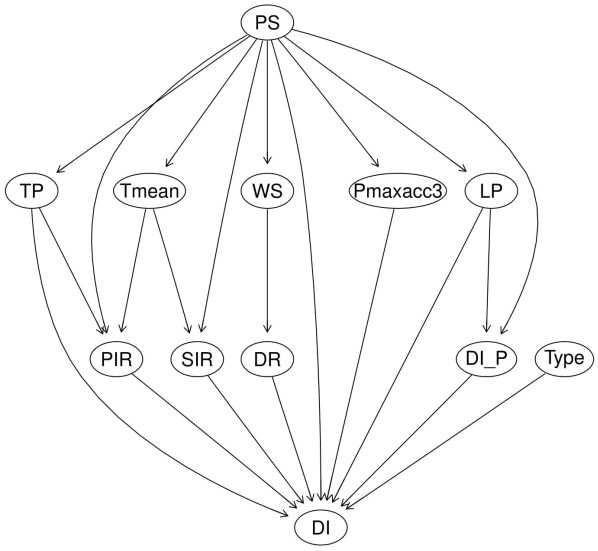


Figure 4.6 DAG representation of Bayesian network model structure identified by *supervised* learning of grape PM. The causal relationships between the variable were linked by existing empirical and published scientific peer-reviewed knowledge on the interactions between the weather, host, and pathogen. Variables in the DAG representation includes: Plant stage (PS), Total precipitation (TP), Daily mean temperature (Tmean), Wind speed (WS), Degree-days based risk assessment model $P_{maxacc3}$ (3°C), Latent period (LP), Primary infection rate (PIR), Secondary infection rate (SIR), Dispersal rate (DR), Recent disease incidence (DI_P), Cultivar (Type), and Disease incidence (DI).

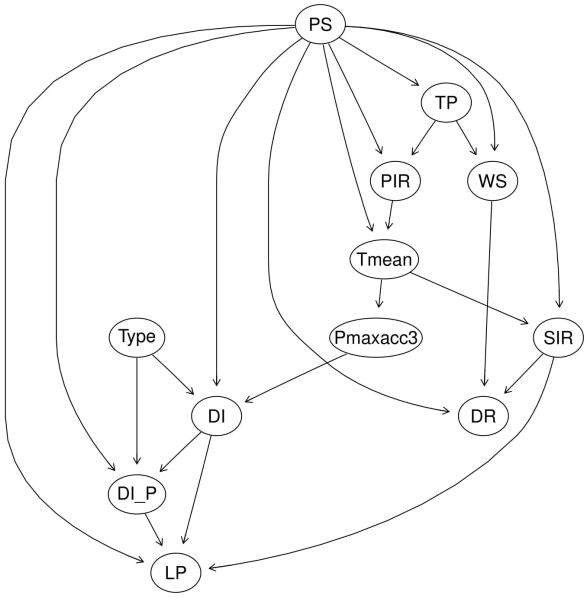


Figure 4.7 DAG representation of Bayesian network model structure learned by *algorithmic* learning. The structure was learned from bootstrap sampling technique under 5000 iterations with arc strength above 0.8 and arc direction above 0.5. Variables learned in this Bayesian network are a combination of observed weather variables and the estimated development factors of grapevine and the pathogen of grape PM. Variables in the DAG representation includes: Plant stage (PS), Total precipitation (TP), Daily mean temperature (Tmean), Wind speed (WS), Degree-days based risk assessment model $P_{maxacc3}$ (3°C), Latent period (LP), Primary infection rate (PIR), Secondary infection rate (SIR), Dispersal rate (DR), Recent disease incidence (DI_P), Cultivar (Type), and Disease incidence (DI).

0.87 in algorithm learning found in 2011.

Model predictions of grape PM for the three cultivars from both supervised learning in case 2 (dotted line) and algorithmic learning in case 2 (dashed line) across the three plant stages of flowering, setting, and veraison are shown in Fig. 4.9. The predictions are shown compared to observed DI in 2011 (solid line). DI of conidial infection generally starts from the end of June until the end of the growing season. For Geisenheim-318 and Frontenac, the Bayesian network in both supervised and algorithm learning did work well in predicting disease incidence in 2011 with supervised learning has slightly better performance the hill-climbing-based algorithm learning in predicting DI over the growing season. Overall, the best model is case 2 with supervised learning for which the model has learned using existing knowledge about disease development of grape PM using the full set of random variables but without relative humidity (as listed in Table 4.2).

Forecast evaluation plots of DI in warm (dotted line) and cold (dashed line) years for the three cultivars: Chancellor (top), Geisenheim-318 (middle), and Frontenac (bottom) in 2011 by increasing and decreasing daily temperature in 2001 by 2 °C are shown in Fig. 4.10. The three grape cultivars have a similar response to the changing temperature but differ in the susceptibility to grape PM. In the warm year, DI tended to be higher than the normal for most of the growing seasons except in mid-August. The maximum DI over the growing season was higher in warm year than in the normal. while in the cold year, model prediction of DI has tended to lower than the normal except the beginning of August. Evaluation of the optimal forecasting window for DI prediction of grape PM using the up to 16 days high-performance forecast data in GEFS is shown in Fig. 4.11. Averaged MAE and RMSE were calculated by taking the average of MAE and RMSE, separately, computed from the model predictions and actual disease incidence in different forecast windows. Lower values of MAE and RMSE indicate lower model uncertainty and higher model forecasting skill. The smallest values of averaged MAE were for 6 days forecasting (0.846), and the most significant values was for 16 days forecasting (0.926). Averaged MAE tended to decreased from 0.863 to the minimum values of 0.846 when forecasting windows changes from 1 to 6 days and increased to the maximum values of 0.926 when forecasting windows changes to 16 days. For averaged RSME, the smallest values were for one-day forecasting (1.029) and tended to the maximum values (1.344) when forecasting windows changes from 1 to 16 days. Overall, both averaged MAE and RMSE in up to 16 days forecasting windows were considerably small, our fungicide spray program has developed using

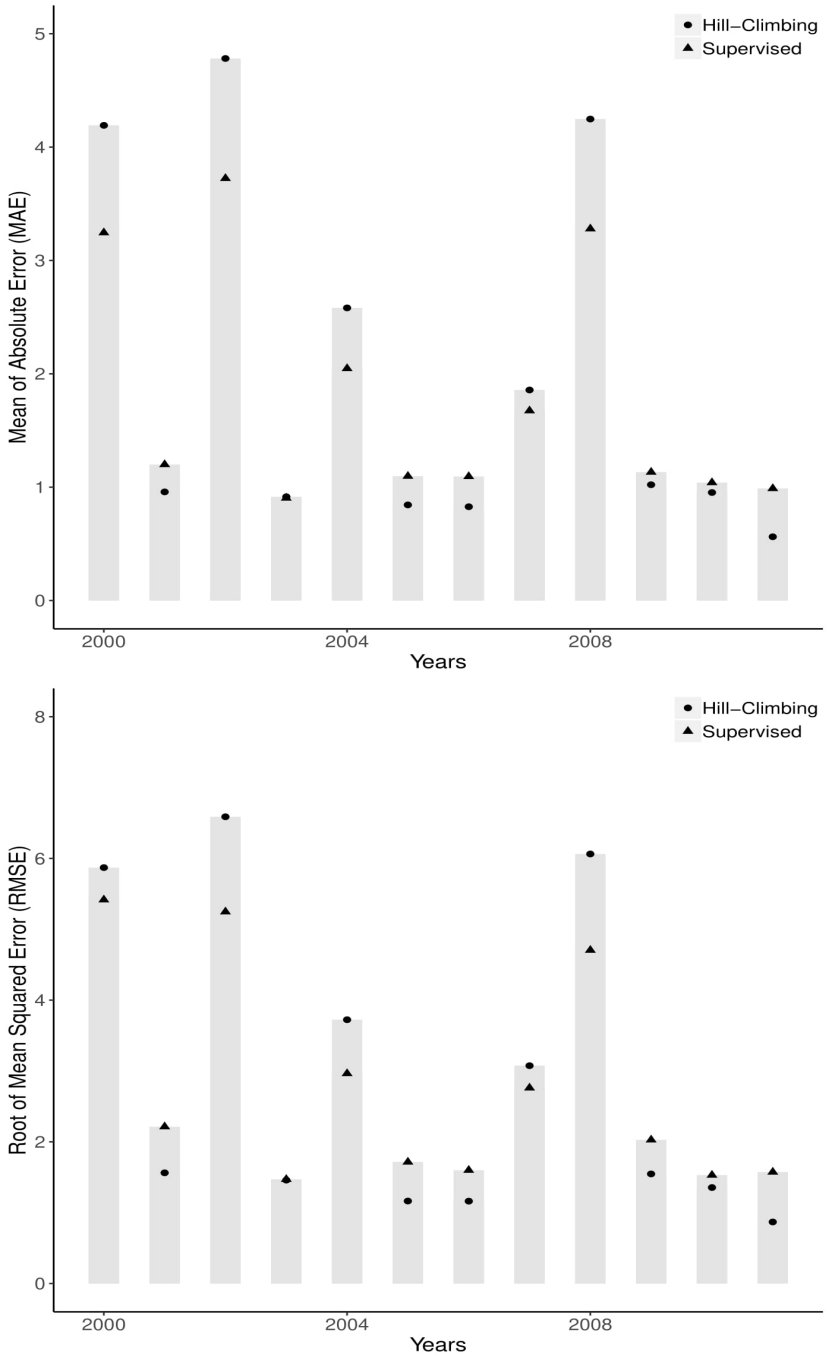


Figure 4.8 Model forecast skill from k-fold cross-validation (2000-2010) and 2011 year validation in both supervised (circle) and hill-climbing based algorithmic learning (triangle). MAE and RMSE were computed for the three susceptible grape cultivars in the testing year. Smaller values in both MAE and RMSE indicates higher forecast skill.

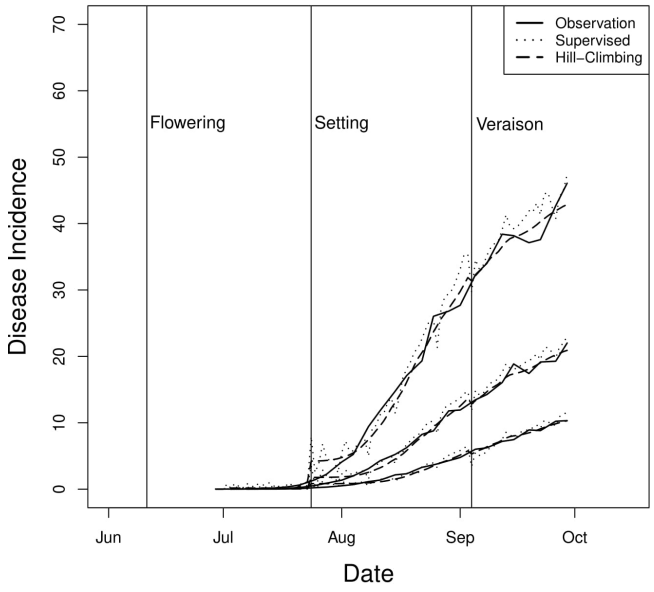


Figure 4.9 Model predictions of DI of PM for the three grape cultivars: High susceptible Chancellor (Top), Medium susceptible Geisenheim-318 (Medium), and the low susceptible Frontenac (bottom) in 2011. Model predictions from both supervised (dotted line) and algorithm (dashed line) learned Bayesian network are shown alongside the observed daily DI. The three vertical lines from left to right are the estimated plant stage of grapevine: flowering, setting, and veraison stages. DI of grape PM in 2011 started from June 29th until the end of the growing season.

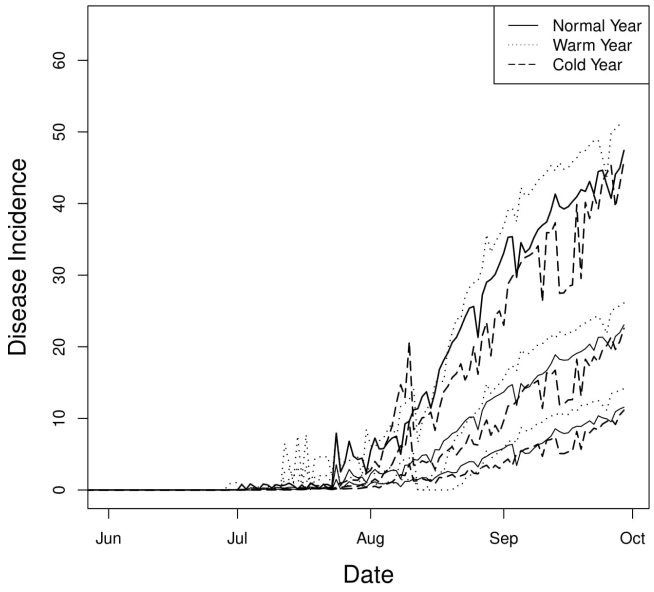


Figure 4.10 Sensitivity analysis of model predicted DI for Chancellor (top), Geisenheim-318 (middle), and Frontenac (bottom) cultivars, by changing temperature by $2°C$ (warm and cold year scenarios). Daily temperature in 2011 is shown as a reference baseline (solid line).

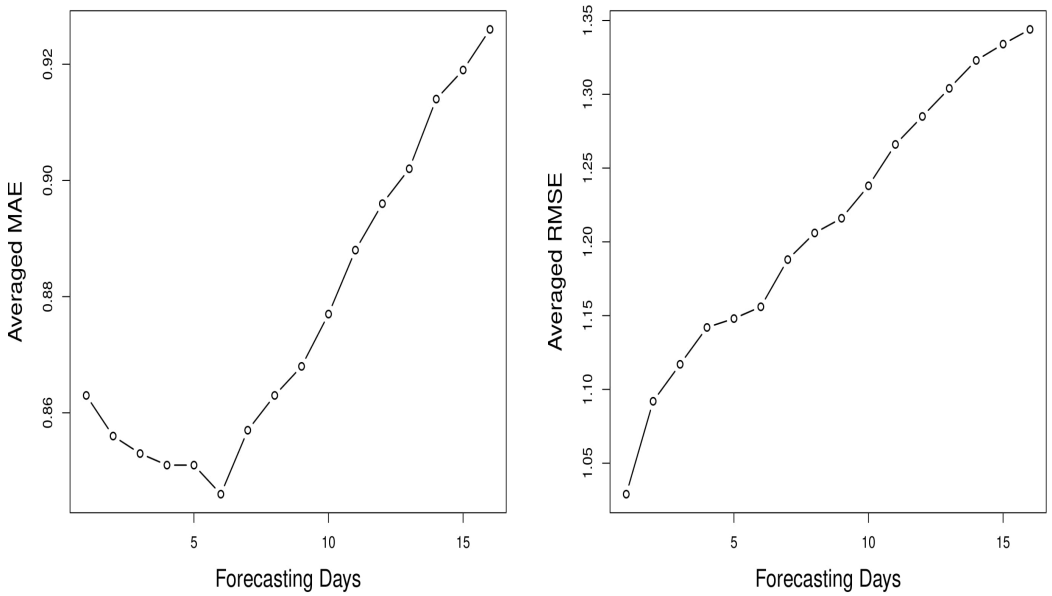


Figure 4.11 Sensitivity of DI prediction error to forecast window size (1-16 days), measured using average MAE and RMSE, under supervised learning and GEFS weather input.

the disease incidence predictions and weather variables from the GEFS in 16 days forecasting window.

We further compared our fungicide spray program to two existing benchmark programs, the UC-Davis, and the degree-days (with a 6 °C base temperature) based risk assessment model. The UC-Davis program is a score based program to provide suggestions for fungicide spray in different spray schedules. The UC score is a daily temperature-based model, which the model initials from the day of the first primary infection. The UC-Davis program applies fungicide spray with an interval of 14 to 21 days, if the UC score is less or equals to 30; a 10 to 17 days interval, if the UC score is from 40 to 50; or a 7-days interval, if the UC score is above 60. The degree-days based risk assessment model specified a spray schedule interval of 7 to 14 days, if $P_{maxacc6}$ is above 1% in Chancellor and Geisenheim-318, and 0.5%, for Frontenac. Figure 4.12 represents the application of the UC-Davis program and the degree-days based risk assessment model for disease control of grape PM at the experimental farm in Quebec in 2011. Fungicide spray applications from the UC-Davis exhibited a case of over-spraying in relation to the actual disease severity for each of the three cultivars.

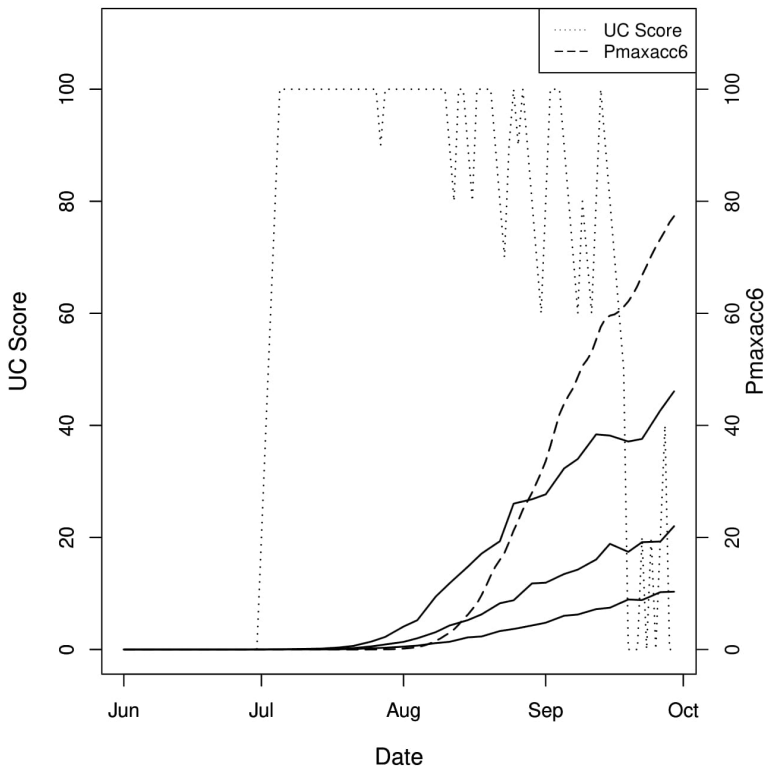


Figure 4.12 Fungicide spray programs of the UC-Davis and the degree-days risk assessment model to grape PM in 2011. The dotted line indicates the threshold-based UC score from the UC Davis model. The long dashed line indicates the schedule-based model scores from the degree-day model. The solid line indicates the DI of Chancellor (top); Geisenheim-318 (median); and Frontenac (bottom).

The UC scores were above 60 from July 09 until September 18 and suggests a high frequency of fungicide sprays. The degree-days based risk assessment model provided better disease control than the UC-Davis model by reducing the number of scheduled fungicide spray. This program suggested initial fungicide spray starts on August 07 for Chancellor, August 10 for Geisenheim-318, and August 12 for Frontenac.

The model-based fungicide spray program provides a spray strategy up to 6 days in advance and forecasts disease risk with a lead time of up to 10 days to optimize the efficiency of fungicide spray under the uncertainty of precipitation. An example of such a model-based fungicide spray program is for the date of August 27, when a high rainfall event was forecast to occur on August 29 and 30 with precipitation of 46.46 mm and 121.129 mm, respectively (Table 4.5). The table shows the average DI from 10 days of disease control under 6 different fungicide spray dates (August

Table 4.5 DI values for the forecast model-based fungicide spray program on August 27 in 2011.

Fungicide Spray Date	Chancellor	Geisenheim-318	Frontenac
August 28	31.59	13.27	5.21
August 29	32.09	13.52	5.26
August 30	32.55	13.74	5.32
August 31	27.91	11.92	4.94
September 1	22.72	9.47	3.82
September 2	31.04	13	5.5

28 to September 2) for the three susceptible cultivars. The best spray day for the three grape cultivars was September 01 having an average DI of 22.72 for Chancellor, 9.47 for Geisenheim-318, and 3.82 for Frontenac. Figure 4.13 shows 3D plots of 10-day daily DI for Chancellor (top-left), Geisenheim-318 (top-right), and Frontenac (bottom left) for the six different spray days, from a low to high level of disease severity. The 2D plot (bottom-right) shows the forecast daily DI for Chancellor (dotted line), Geisenheim-318 (dot-dashed line), and Frontenac (dashed line) on the optimal spray day.

4.7 Discussion

The best performing model for grape PM disease risk forecasting at the experimental farm in Quebec was case 2 with model structure learned using supervised learning. Figure 4.6 shows the learned structure found for this model. DI for the three cultivars was relatively consistent with the degree-days risk assessment model having a 3 °C base temperature and plant stage, but not with relative humidity included. Instead, the best local distribution of the causal relationship of DI involved total precipitation, primary infection rate, secondary infection rate, dispersal rate, plant stage, risk assessment model, latent period, and vary in susceptible cultivar type. Figure 4.9 shows that the supervised Bayesian network has a very high performance in disease risk forecasting for the three susceptible cultivars over the plant stage that are affected by PM disease.

Model validation of DI shown in Fig. 4.9 shows the resultant model network structure from both supervised and algorithm learning. Supervised learning had a slightly better performance in predicting DI of grape PM over 2000-2011. Smaller values of MAE and RMSE in non-drought years than in drought years (i.e., 2000, 2002, 2008) indicates that the model has better forecasting skill in non-drought years,

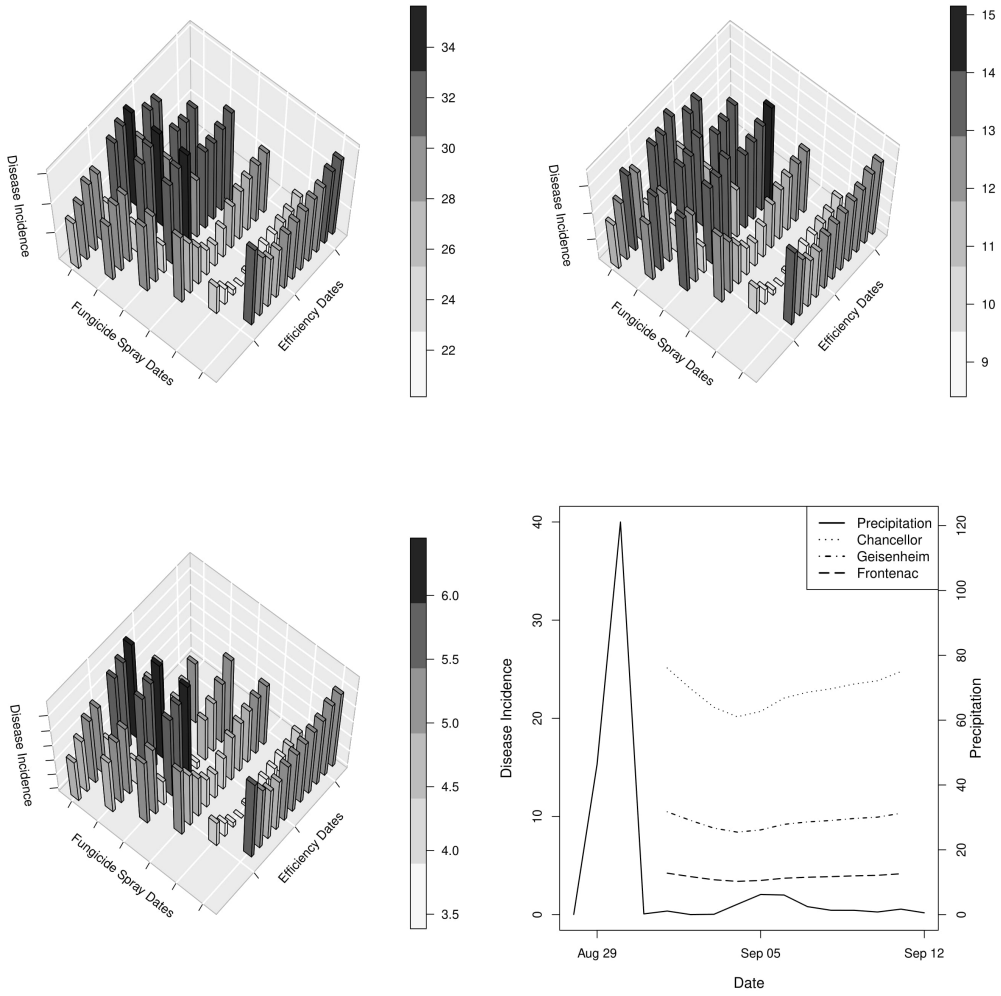


Figure 4.13 3D plot of the fungicide spray strategies of Chancellor (upper-left), Geisenheim-318 (upper-right), and Frontenac (lower center) from the forecast model on August 27th in 2011 for six fungicide spray dates and corresponding 10-day DI based on the 16-day GEFS reforecast weather input. (Lower right) 2D plot of the 10-day daily DI of the optimal spray date for Chancellor (dotted line); Geisenheim-318 (dot-dash line); and Frontenac (long-dash line) based on DI on August 27th 2011 and 16-day GEFS input. The solid line indicated the daily cumulative total precipitation from GEFS on August 28th (2011). Note: Darker color indicates higher disease severity.

and the complex weather conditions significantly influence the development of DI in drought years, summarized in Table 4.3. The development of PM in droughts years was not well explained by the use of a simple binary factor suggesting that a more complex model of DI involving autocorrelation across multiple years (i.e., not just a single year factor) may be needed. The comparison result of the four selected model starting dates indicates the best time to initiate the modeling of disease incidence is when the disease occurs on the farm.

Model forecast evaluation results are shown in Fig. 4.10, whereby the three cultivars have a similar response to the changing temperature. There is a significant difference in DI between the normal year and the warm and cold years, which by adding or dropping the daily temperature by 2 °C, separately. DI tends to be more severe in warm years and less severe in cold years when compared to the DI in the normal year, which explains the complex interactions between the development of the pathogen of powdery mildew and its host under the influences of heat stress conditions. The warm temperature at the beginning of the growing season advances the bud break of the host as well as the development of ascospores from over-wintered cleistothecia. Then dispersal and infection, caused higher than normal DI.

The sensitivity analysis of model prediction of grape PM for the three cultivars in 2011 using the 16 days weather variable from GEFS shown in Fig. 4.11 indicates the GEFS data has very high performance in predicting DI in Quebec. Although the averaged MAE varied from 0.863 to 0.926 and averaged RMSE varied from 1.029 to 1.344 when forecasting window changes from day 1 to day 16, both averaged MAE and RMSE are all small (less than one percentage in averaged MAE and 1.5 percentage in averaged RMSE). The UC-Davis fungicide spray program shown in Fig. 4.12 shows an over-estimate of DI of PM of most of the growing season, results from over-spray suggestions of scheduled fungicide spray application in disease control.

The degree-days based risk assessment model provides better suggestions for disease control than the UC-Davis by delaying the initiation of the fungicide spray program for disease control. However, it does not consider the difference in susceptibility of cultivars. Our fungicide spray program was developed from the best-performing Bayesian network forecast model with supervised learning out to 16 days guided by GEFS reforecast weather. This program provides fungicide spray suggestions up to 6 spray day strategies with 10-days disease control forecasting by using information about forecast DI and the efficiency of fungicide spray under uncertainty weather changes (precipitations). An example of this is presented for August 17, 2011. This

example data was selected because there are two significant forecast rainfall events on August 29 (46.46 mm) and 30 (121.129 mm) and a few small rainfalls from August 31 to September 12. Table 4.5 shows the best spray day for the three susceptible cultivars is on September 1st with results average DI of 22.72, 9.47, 3.82 for Chancellor, Geisenheim-318, and Frontenac, in separately, in 10 days after the spray has applied. The 3D plot in Fig. 4.13 shows model forecasts of DI extending out to 10 days in the future for the different spray days. The 2D plot shows the daily DI based on a current day for fungicide spray under variation to changing precipitation. With these forecast curves or foresight information output from the forecast model, grape producers can make a better decision about the best timing to apply fungicide spray for different grape cultivars in order to maximize the spray efficiency and to protect grapevines over the growing season.

4.8 Conclusions

Grape PM is one of the most common diseases responsible for significant reduction in grape yield in North America. The rate of development and epidemiology of this disease are influenced by regional-scale, longer-term climate and localized, shorter-term weather uncertainty. It also varies with growth stage and the genetics of a grape cultivar, making it difficult to develop efficient strategies for disease management. Most research on epidemiology and modeling of PM has been conducted for temperate weathers. In this study, using 13 years of data, we developed and tested a novel Bayesian network model to forecast disease risk (i.e., the development of DI of conidial infection in grape PM on leaves) for three susceptible cultivars calibrated to site-specific data obtained from an experimental vineyard in Quebec. The forecast model generates high prediction (based on in-sample validation testing) also out to 16 days ahead using GEFS, and enables a reliable fungicide strategy for disease control with a 6-days forecasting window. Fully-independent validation data is needed to evaluate how well our model can forecast grape PM disease in other vineyards with differing grape cultivars, cultural practices, and environmental conditions.

There is observed uncertainty in DI that is not fully explained by our current model. Our modeling focuses on leaf PM epidemic and resistance at a time when the grape cluster is most susceptible to the presence of PM, whereby leaf infection is usually the first warning and signal to the vines treatment and leaf protection is a limiting factor to grape yield. Nonetheless, relying solely on leaf resistance is a current shortcoming

of the forecast model. An operational protection model would need to consider leaves and grape berry clusters. This would require extending the current model to be spatially-explicit, because berry clusters are sufficiently complex in terms of their development and exhibit susceptibility that is spatially heterogeneous within vineyards and strongly dependent on phenological age [18]. Recent spatial model simulations of fungicide use show that applying a fungicide early at flowering may significantly reduce PM diseased area, by up to 81% at the end of the season by delaying the epidemic onset. It also helps to maintain a low level of disease and to minimize potential PM dispersion from leaves to bunches [82]. Burie et al. (2011) demonstrate the strong dependence of PM disease progression on grapevine growth rate (vigour) and regional weather using a multi-scale dynamic grapevine-PM model [14]. Future work stemming from the current study will require spatio-temporal vineyard data and additional data from other measurement variables such as: canopy leaf wetness, fungicide type, and efficiency, spatial locations of susceptible grape cultivars, and berries infection, yield loss.

The developed model-based fungicide spray program provides an improved way to minimize the total number of sprays and their timing for optimizing grape PM spray efficiency and its recommendations could, in the future, be used by vineyard producers through the growing season, as the total proportion of the infections of over-wintered ascospore is a critical factor in determining the outbreak of conidia over an entire season. Nonetheless, skillful protection of grapevines by contact, translaminar or systemic fungicides also helps to address unexplained uncertainty and current shortcomings of the model-based forecasts.

Chapter 5

Discussions

In this dissertation, three statistical models are presented for improving planning and decision-making support for Canadian agricultural producers, with a focus on disease risk in the explicit context of climate change uncertainty as it affected by short- and long-term weather events. Each model employs various machine learning approaches in model learning, which consist of: time series analysis, geospatial analysis, statistical modeling, and model forecasting (Fig. 5.1).

Chapter 2 describes a cluster-based Principal Component Analysis (PCA) model to predict crop yields of wheat and barley grown on the Canadian Prairies. A multi-scale spatial approach was utilized to predict yield of wheat and barley by using k-means clustering, principle component analysis, and generalized linear model approaches. The study results show that the cluster-PCA model achieved higher model performance in yield predictions for both wheat and barley when compared to the results from non-clustering approaches of multivariate (linear) regression (MLR) and principal component regression (PCR). It was also noted that wheat yield prediction is more sensitive to spatial-scale than barley. The cluster-based PCA analysis shows that ENSO variability influences crop yield variability in a non-linear manner across the Canadian prairies, and that ENSO forcing exerts a greater influence on wheat yields. A spatial pattern of ENSO influence has found for wheat yield across the Canadian prairies. EOSO is likely affecting wheat yield in CARs scale located from the North-west extend to the South-east of the Canadian prairies, where wheat yield in the Western of Alberta to the Northern of Manitoba have less affect from ENSO.

Chapter 3 describes a copula-Bayesian network to define heatwave events and extreme weather conditions that have significant impacts on crops in different Canadian agriculture zones. A common approach - identification of extremes using an experience-

based threshold - may not be suitable for real-world studies when a forecasting model is complex. Extracting Information for model forecasting using artificial intelligence (AI) may improve model predicting skills. This research investigated the relationships between Canadian crop growth and extreme weather conditions driven by high temperatures. Model outputs of the copula-Bayesian network model show regional heatwave events can be characterized by multiple heatwave indices. Spatial analysis of heatwave mapping (Fig. 3.4) shows heatwave events tend to occur earlier in the two coastal zones and the Canadian prairies in May; extend to the neighboring zones in June, and back to the coastal zones in July. Most agricultural regions in can be affected by heatwaves from 2 to 3 days from the characterized heatwave events, indicating crops in these regions are highly sensitive to regional heatwave events. The spatial mapping (Fig. 3.6) identifies other extreme weather conditions (temperature, precipitation, and humidity) that occurred during heatwave actives. The spatial mapping of long-term heatwave intensity and frequency (Fig. 3.7) shows some regions on the Canadian Prairies are at a high risk of heat stress with large values of heatwave intensity as well as frequency. An assessment of model performance in probability prediction using model simulations of crop growth based on the following heat stress variables heatwave intensity and daily weather variables (temperature, precipitation, and humidity). Fig. 3.3 shows superior model prediction performance for the copula-Bayesian model, with results approximating empirical observations. This suggests the copula-Bayesian model offers useful skill for prediction of the impacts of heatwave events on regional crop growth.

Chapter 4 proposed a Bayesian network model to forecast the daily risk of powdery mildew on three grape cultivars in the experiment farm in Quebec. Powdery mildew has infected on the three grapes cultivars over the years of the study period and varies in cultivar susceptibility and annually. Model performance of daily disease incidence prediction has examined in different network random variables, learning approaches (supervised and algorithm), model starting dates, drought years, and temperature scenarios (warm and cold years). The 'best' Bayesian network model has selected based on the model outputs of mean-absolute-error (MAE) and root-mean-squared-error (RMSE) from both K-fold cross-validation and prediction. The comparison results show both supervised and algorithmic-based learning have similar high performance in predicting daily powdery mildew of the three grape cultivars. The overall 'best' model has found in case 2 with supervised learning for which the model has learned using existing knowledge about disease development of grape powdery mildew using the full

set of random variables but without relative humidity from the date when the disease has found on the field. Model performance of disease risk forecasting windows in 2011 using high-performance forecast data in Global Ensemble Forecast System (GEFS) shows both averaged MAE and RMSE in up to 16 days forecasting window were in considerably small, suggesting model forecasting of grape powdery mildew can up to 16 days windows when using the climate variables from the GEFS. We developed a model-based fungicide spray program to provide fungicide spray suggestions up to 6 spray day strategies with 10-days disease control forecasting by combining information about daily forecasting of disease incidence and the efficiency of fungicide spray (spray types, spray wash-off, spray decays) under uncertainty climate changes. Advantages of using our fungicide spray program for disease control include: 1) The 6 spray day strategies provide an overview of averaged disease incidence forecasting for the next 10-days window from the spray day, which provides agriculture producers the 'best' time for upcoming spray event, 2) The 10-days daily disease control forecasting form the spray day provides detail information of disease control under the efficiency of fungicide spray from the uncertainty of future precipitation, which provides agriculture producers the 'best' time for next spray date, and 3) The flexibility of this fungicide spray program does not limit a unique spray type, allowing agriculture producers to examine different spray types from their farming practices. An example of our model-based fungicide spray program has shown in Table 4.5.

There are still open issues for future research in the filed of climate-agriculture risk modeling. We will mention two issues directly related to this dissertation.

One issue is about to collect long-term high-quality data support for statistical model training. Model framework and model accuracy in assessing climate-agriculture risks often require long-term observations in a well descriptive of the causal-relationship between climate changes and agriculture risks, which is hard to meet. Many climate models exist to support historical, reforecast, or forecast climate data in different spatial resolutions for multiple purposes used. Long-term observations of agriculture data often collected in the same way as it has designed at the beginning of the study, which this kind of data may not suit other study use. Observation bias is strongly affecting the accuracy of a statistical model in risk predicting. I felt fortunate to have experiences in weather station installation under the help of my supervisor Dr. Nathaniel Newlands. A weather station has installed to automatically collect continuous hourly climate data for disease monitoring in a local vineyard in Kelowna, BC, Canada.

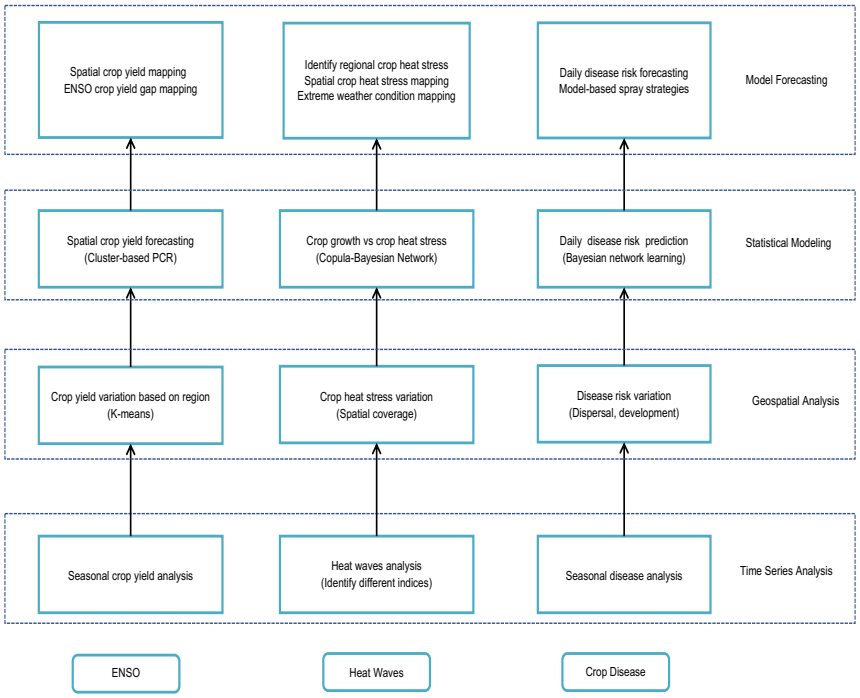


Figure 5.1 A general overview of analytical approaches and how they were employed in the various models.

Another issue is about developing a farming strategy that considers a set of climate risks together to improve Canadian agriculture producer's decision makings. In this dissertation, we had assessed three separate climate risks on Canada's agricultural areas in seasonal (ENSO), short-term (heatwaves), and emergent disease (powdery mildew). Inter-relationship between these climate risks and their impacts on Canada's crops still not clear. In Chapter 2, we had shown a spatial pattern of significant ENSO forcing on crop yield variability on the Canadian prairies. Strong ENSO years often results in a more variability of temperature and precipitation during the growing season, which can cause more uncertainty in heatwave and disease risks on crops. In Chapter 3, we assessed the impacts of heat stress on the development of regional crops across the Canada sectors and the disease risk had assessed in Chapter 4. Model performance of the Bayesian network in disease risk prediction in Chapter 4 shows model performance in higher in non-drought years. Model sensitivity of disease incidence in warm and cold years indicates the development of disease incidence varied by temperature variability. In Chapter 4, we presented a model-based fungicide spray program for disease control. The efficiency of fungicide spray has varied by the uncertainty of precipitation/irrigation, which is also a common strategy for reducing the negative impacts of heat stress on crop growth. How to generalize an optimal strategy that considering a set of multiple weather risks together and support Canada's agricultural producer to make better decisions in their farming practices is worth future attention. Besides, the communications between scientific experts and Canadian agricultural producers and work together to support future researches also worth attention. The fungicide spray program we presented in Chapter 4 provides suggestions for fungicide spray up to 6 strategy days. The open suggestions provided from our model-based fungicide program allows agricultural producers to improve decision-making by combing their agricultural behaviors and agricultural benefits with model-based fungicide spray suggestions rather than focus on suggestions from experts to apply for fungicide spray program in a specific way and time. We strongly hope this can help to improve the communications between scientific experts and Canadian agriculture producers work together to improve agricultural practices. Scientific experts can have a better understanding of the concerns from agricultural producers, helps to guide future researches, and improves the applications of model-based programs. At the same time, agricultural producers know better the potential agricultural risks on their farms and help to support future researches by providing application feedbacks and long-term high-quality data.

Chapter 6

Achieved

6.1 Publications

1. Weixun Lu, Nathaniel K Newlands, Odile Carisse, David E Atkinson, and Alex J Cannon. Disease risk forecasting with bayesian learning networks: Application to grape powdery mildew (erysiphe necator) in vineyards. *Agronomy*, 10(5):622, 2020

2. Weixun Lu, David E Atkinson, and Nathaniel K Newlands. Enso climate risk: predicting crop yield variability and coherence using cluster-based pca. *Modeling Earth Systems and Environment*, 3(4):1343–1359, 2017

3. Nathaniel K Newlands, Weixun Lu, and Tracy A Porcelli. Downscaling of regional climate scenarios within agricultural areas across canada with a multivariate, multisite model. In *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science*, pages 335–340. Springer, 2015

6.2 Leadership

1. Section co-chair and organizer. Organized a presentation session and provide two presentation talks of ENSO and Grape disease risk forecasting in the TIES 2019 conference (2019).

2. Supervising of a Heatwave Activities project, "Spatial and temporal analysis of heatwave activities across the Canadian agricultural regions using a copula-Bayesian network model" (In processing, 2020).

3. Supervising of a Powdery Mildew Risk Forecasting project, "Disease Risk Forecasting with Bayesian Learning Networks: Application to Grape Powdery Mildew (Erysiphe necator) in Vineyards", which includes supervising an internship student for spatial-temporal mapping and climate data mining using R (2019).

4. R code supervising of a coffee leaf rust forecasting model. Open government R statistical code licence and provide R code helps to other agriculture research groups (2019).

5. Member of American Statistical Association. Provided a presentation talk of powdery mildew diseaes risk forecasting in the JSM 2018 conference (2018).

6. Co-Supervising of a Crop Disease Risk Modeling project, "Model-Based Forecasting of Agricultural Crop Disease Risk at the Regional Scale, Integrating Airborne Inoculum, Environmental, and Satellite-Based Monitoring Data" 2018, which includes supervising a co-op student from other research team for disease modeling using R and spatial-temporal mapping (2018).

6.3 Professional Development and Training

1. TIES 2019 (The International Environmetrics Society), Kunming, China, (August 26th - August 27th).

2. Weather station installation at a winery in Okanagan region. Kelowna, BC (May 2019).

3. JSM 2018 (Joint the Statistics Meeting 2018), Vancouver, BC (July 27th - August 4th).

4. Summer School on Mathematical and Statistical Model Uncertainty 2018, Vancouver (July 22nd - July27th).

5. Security Awareness Web-based Training (AGR-620) 2018, Agriculture and Agri-Food Canada (Feb 7th).

6. Security Awareness Training (A-230) 2018, The Canada School of Public Service (CSPS) (Jan 30th).

Bibliography

[1] H Abdi and LJ Williams. Principal component analysis. *Wiley Interdiscip Rev Comput Stat*, 2(4):433–459, 2010.

[2] Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

[3] S Analytis. Über die relation zwischen biologischer entwicklung und temperatur bei phytopathogenen pilzen. *Journal of Phytopathology*, 90(1):64–76, 1977.

[4] S Analytis. Obtaining of sub-models for modeling the entire life cycle of a pathogen/über die erlangung von sub-modellen, die zur beschreibung eines gesamten lebenszyklus eines krankheitserregers dienen. *Zeitschrift für Pflanzenkrankheiten und Pflanzenschutz/Journal of Plant Diseases and Protection*, pages 371–382, 1980.

[5] SV Angadi, HW Cutforth, PR Miller, BG McConkey, MH Entz, SA Brandt, and KM Volkmar. Response of three brassica species to high temperature stress during reproductive growth. *Canadian Journal of Plant Science*, 80(4):693–701, 2000.

[6] JN Axelson, DJ Sauchyn, and J Barichivich. New reconstructions of streamflow variability in the South Saskatchewan River Basin from a network of tree ring chronologies, Alberta, Canada. *Water Resour Res*, 45(9), 2009.

[7] A Bárdossy and GGS Pegram. Copula based multisite model for daily precipitation simulation. *Hydrology & Earth System Sciences*, 13(12), 2009.

[8] P Bickel, P Diggle, S Fienberg, K Krickeberg, I Olkin, N Wermuth, and S Zeger. Principal component analysis. *Springer Verlang*, 2:37–52, 2002.

[9] SJ Bonner, NK Newlands, and NE Heckman. Modeling regional impacts of climate teleconnections using functional data analysis. *Environ Ecol Stat*, 21(1):1–26, 2014.

[10] B Bonsal and A Shabbar. Large-scale climate oscillations influencing Canada, 1900-2008. Technical report, Canadian Biodiversity: Ecosystem Status and Trends 2010, Report 4, 2011.

[11] L. Bornn and J.V. Zidek. Efficient stabilization of crop yield prediction in the Canadian Prairies. *Agric. For. Meteorol.*, 152:223–232, 2012.

[12] Mark J Brewer, Adam Butler, and Susan L Cooksley. The relative performance of aic, aicc and bic in the presence of unobserved heterogeneity. *Methods in Ecology and Evolution*, 7(6):679–692, 2016.

[13] Rosalind A Bueckert and John M Clarke. Annual crop adaptation to abiotic stress on the canadian prairies: Six case studies. *Canadian Journal of Plant Science*, 93(3):375–385, 2013.

[14] Jean-Baptiste Burie, Michel Langlais, and Agnes Calonnec. Switching from a mechanistic model to a continuous model to study at different scales the effect of vine growth on the dynamic of a powdery mildew epidemic. *Annals of Botany*, 107(5):885–895, 2011.

[15] Tito Caffi, Vittorio Rossi, Sara Elisabetta Legler, and Riccardo Bugiani. A mechanistic model simulating ascosporic infections by *Erysiphe necator*, the powdery mildew fungus of grapevine. *Plant Pathology*, 60(3):522–531, 2011.

[16] R. Cai, J.D. Mullen, J.C. Bergstrom, W.D. Shurley, and M.E. Wetzstein. Using a climate index to measure crop yield response. *J Agric. and Appl. Econ.*, 45(4):719–737, 2013.

[17] A Calonnec, P Cartolaro, C Poupot, D Dubourdieu, and P Darriet. Effects of *Uncinula necator* on the yield and quality of grapes (*Vitis vinifera*) and wine. *Plant Pathology*, 53(4):434–445, 2004.

[18] Agnes Calonnec, Philippe Cartolaro, J-M Naulin, D Bailey, and Michel Langlais. A host-pathogen simulation model: powdery mildew of grapevine. *Plant Pathology*, 57(3):493–508, 2008.

[19] Alex J Cannon. Revisiting the nonlinear relationship between ENSO and winter extreme station precipitation in North America. *Int J Climatology*, 35(13):4001–4014, 2015.

[20] O Carisse, R Bacon, A Lefebvre, and K Lessard. A degree-day model to initiate fungicide spray programs for management of grape powdery mildew (*Erysiphe necator*). *Canadian Journal of Plant Pathology*, 31(2):186–194, 2009.

[21] Odile Carisse. Development of grape downy mildew (*Plasmopara viticola*) under northern viticulture conditions: Influence of fall disease incidence. *European Journal of Plant Pathology*, 144:773–783, 2016.

[22] Odile Carisse, Réjean Bacon, and Annie Lefebvre. Grape powdery mildew (*Erysiphe necator*) risk assessment based on airborne conidium concentration. *Crop Protection*, 28(12):1036–1044, 2009.

[23] JE Carroll and WF Wilcox. Effects of humidity on the development of grapevine powdery mildew. *Phytopathology*, 93(9):1137–1144, 2003.

[24] R Caruana, M Elhawary, N Nguyen, and C Smith. Meta clustering. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 107–118. IEEE, 2006.

[25] A Ceglar, M Turco, A Toreti, and FJ Doblas-Reyes. Linking crop yield anomalies to large-scale atmospheric circulation in Europe. *Agric For Meteorol*, 240:35–45, 2017.

[26] Daniel O Chellemi and James J Marois. Development of a demographic growth model for *Uncinula necator* by using a microcomputer spreadsheet program. *Phytopathology*, 81(3):250–254, 1991.

[27] A Chipanshi, Y Zhang, L Kouadio, N Newlands, A Davidson, R Hill, Hand Warren, B Qian, B Daneshfar, and F Bedard. Evaluation of the integrated canadian crop yield forecaster (ICCYF) model for in-season prediction of crop yield across the Canadian agricultural landscape. *Agric For Meteorol*, 206:137–150, 2015.

[28] Nikolaos Christidis, Gareth S Jones, and Peter A Stott. Dramatically increasing chance of extremely hot summers since the 2003 european heatwave. *Nature Climate Change*, 5(1):46, 2015.

[29] Paolo Cortesi, M Bisiach, M Ricciolini, and David M Gadoury. Cleistothecia of uncinula necator—an additional source of inoculum in italian vineyards. *Plant disease*, 81(8):922–926, 1997.

[30] Anaïs Couasnon, Antonia Sebastian, and Oswaldo Morales-Nápoles. A copula-based bayesian network for modeling compound flood hazard from riverine and coastal interactions at the catchment scale: An application to the houston ship channel, texas. *Water*, 10(9):1190, 2018.

[31] Charles J. Delp. Effect of temperature and humidity on the grape powdery mildew fungus. *Phytopathology*, 44(11):615–626, 1954.

[32] J. Dereudre, J.C. Audran, C. Leddet, E. Ait Barka, and O. Brun. Réponse de la vigne (*Vitis vinifera* L.) aux températures inférieures à 0^{o}C; III: Effets d'un refroidissement contrôlé sur des bourgeons en cours de débourrement. *Agronomie*, 13:509–514, 1993.

[33] D Deryng, D Conway, N Ramankutty, J Price, and R Warren. Global crop yield response to extreme heat stress under multiple climate change futures. *Environ Res Lett*, 9(3):034011, 2014.

[34] C Ding and X He. K-means clustering via principal component analysis. In *Proceedings of the Twenty-First International Conference on Machine learning*, page 29. ACM, 2004.

[35] SM Dofine. Growth, phenology, and yield components of barley and wheat grown in Alaska. *Can J Plant Sci*, 72(4):1227–1230, 1992.

[36] Daniele Ehrlich, JE Estes, and A Singh. Applications of noaa-avhrr 1 km data for environmental monitoring. *Remote Sensing*, 15(1):145–161, 1994.

[37] Gal Elidan. Copula bayesian networks. In *Advances in neural information processing systems*, pages 559–567, 2010.

[38] Andrea Ficke, David M Gadoury, and Robert C Seem. Ontogenic resistance and plant disease management: A case study of grape powdery mildew. *Phytopathology*, 92(6):671–675, 2002.

[39] Frank and Rimerman. The Economic Impact of the Wine and Grape Industry in Canada 2015. Frank Rimerman + Co. LPP, The Wine Business Center,

California, USA, http://www.canadianvintners.com/wp-content/uploads/2017/06/Canada-Economic-Impact-Report-2015.pdf, 2017.

[40] David M Gadoury, L. Cadle-Davidson, Wayne F Wilcox, Ian B Dry, Robert C Seem, and Michael G Milgroom. Grapevine powdery mildew (*Erysiphe necator*): a fascinating system for the study of the biology, ecology and epidemiology of an obligate biotroph. *Molecular Plant Pathology*, 13(1):1–16, 2012.

[41] David M. Gadoury and Roger C. Pearson. Initiation, development, dispersal and survival of cleistothecia of *Uncinula necator* in new york vineyards. *Phytopathology*, 78(11):1413–1421, 1988.

[42] David M. Gadoury and Roger C. Pearson. Ascocarp dehiscence and ascospore discharge in *Uncinula necator*. *Phytopathology*, 80(4):393–401, 1990.

[43] David M Gadoury, Roger C Pearson, et al. Germination of ascospores and infection of Vitis by *Uncinula necator*. *Phytopathology*, 80(11):1198–1203, 1990.

[44] David M Gadoury, Robert C Seem, Andrea Ficke, and Wayne F Wilcox. The epidemiology of powdery mildew on concord grapes. *Phytopathology*, 91(10):948–955, 2001.

[45] David M Gadoury, Robert C Seem, Andrea Ficke, and Wayne F Wilcox. Ontogenic resistance to powdery mildew in grape berries. *Phytopathology*, 93(5):547–555, 2003.

[46] David M. Gadoury, Robert C. Seem, Roger C. Pearson, Wayne F. Wilcox, and Richard M. Dunst. Effects of powdery mildew on vine growth, yield, and quality of concord grapes. *Plant Disease*, 85(2):137–140, 2001.

[47] Stuart H Gage. Climate variability in the north central region: characterizing drought severity patterns. *Climate variability and ecosystem responses at long-term ecological research site. Oxford University Press, Oxford*, pages 56–73, 2003.

[48] Stuart H Gage, Julie E Doll, and Gene R Safir. A crop stress index to predict climatic effects on row-crop agriculture in the us north central region. *The ecology of agricultural landscapes: Long-term research on the path to sustainability*, pages 77–103, 2015.

[49] Stuart H Gage and MK Mukerji. A perspective of grasshopper population distribution in saskatchewan and interrelationship with weather. *Environmental Entomology*, 6(3):469–479, 1977.

[50] ER Garnett, ML Khandekar, and JC Babb. On the utility of ENSO and PNA indices for long-lead forecasting of summer weather over the crop-growing region of the Canadian Prairies. *Theor Appl climatology*, 60(1):37–45, 1998.

[51] Cesare Gessler and Ph Blaise. An extended progeny/parent ratio model II. Application to experimental data. *Journal of Phytopathology*, 134(1):53–62, 1992.

[52] AK Gobena and TY Gan. Low-frequency variability in Southwestern Canadian stream flow: links with large-scale climate anomalies. *int J Climatology*, 26(13):1843–1869, 2006.

[53] Thomas M Hamill, Gary T Bates, Jeffrey S Whitaker, Donald R Murray, Michael Fiorino, Thomas J Galarneau Jr, Yuejian Zhu, and William Lapenta. NOAA's second-generation global medium-range ensemble reforecast dataset. *Bulletin of the American Meteorological Society*, 94(10):1553–1565, 2013.

[54] J Graeme Hamilton and Stuart H Gage. Outbreaks of the cotton tipworm, crocidosema plebejana (lepidoptera: Tortricidae), related to weather in southeast queensland, australia. *Environmental entomology*, 15(5):1078–1082, 1986.

[55] P Harris, C Brunsdon, and M Charlton. Geographically weighted principal components analysis. *Int J Geogr Inf Sci*, 25(10):1717–1736, 2011.

[56] JA Hartigan and MA Wong. Algorithm AS 136: A k-means clustering algorithm. *J R Stat Soc Ser C Appl Stat*, 28(1):100–108, 1979.

[57] H. Hoffmann, G. Zhao, S. Asseng, M. Bindi, C. Biernath, J. Constantin, E. Coucheney, R. Dechow, L. Doro, and H. Eckersten. Impact of spatial soil and climate input data aggregation on regional yield simulations. *PloS one*, 11(4):e0151782, 2016.

[58] DP Holzworth, NI Huth, EJ Zurcher, NI Herrmann, G McLean, K Chenu, EJ van Oosterom, Vl Snow, C Murphy, and AD Moore. APSIM–evolution towards a new generation of agricultural systems simulation. *Environ Model Softw*, 62:327–350, 2014.

[59] SM Hsiang and KC Meng. Tropical Economics. *Am Econ Rev Pap Proc*, 105(5):257–261, 2015.

[60] WW Hsieh, B Tang, and ER Garnett. Teleconnections between Pacific sea surface temperatures and Canadian prairie wheat yield. *Agric For Meteorol*, 96(4):209–217, 1999.

[61] SMI Hussein, MC Puri, PD Tonge, M Benevento, AJ Corso, JL Clancy, R Mosbergen, M Li, D-S Lee, and N Cloonan. Genome-wide characterization of the routes to pluripotency. *Nat*, 516(7530):198–206, 2014.

[62] T Iizumi, J-J Luo, AJ Challinor, G Sakurai, M Yokozawa, H Sakuma, ME Brown, and T Yamagata. Impacts of El Niño Southern Oscillation on the global yields of major crops. *Nat Commun*, 5, 2014.

[63] T Iizumi and N Ramankutty. Changes in yield variability of major crops for 1981-2010 explained by climate change. *Environ. Res. Lett.*, 11(3):034003, 2016.

[64] Toshichika Iizumi, Hiroki Takikawa, Yukiko Hirabayashi, Naota Hanasaki, and Motoki Nishimori. Contributions of different bias-correction methods and reference meteorological forcing data sets to uncertainty in projected temperature and precipitation extremes. *Journal of Geophysical Research: Atmospheres*, 122(15):7800–7819, 2017.

[65] F Jailloux, L Willocquet, L Chapuis, and G Froidefond. Effect of weather factors on the release of ascospores of *Uncinula necator*, the cause of grape powdery mildew, in the bordeaux region. *Canadian Journal of Botany*, 77(7):1044–1051, 1999.

[66] Harry Joe. *Multivariate models and multivariate dependence concepts*. CRC Press, 1997.

[67] MD Johnson, WW Hsieh, AJ Cannon, A Davidson, and F Bédard. Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. *Agric For Meteorol*, 218:74–84, 2016.

[68] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[69] Gregory V Jones and Robert E Davis. Climate influences on grapevine phenology, grape composition, and wine production and quality for bordeaux, france. *American Journal of Enology and Viticulture*, 51(3):249–261, 2000.

[70] E Kalnay, M Kanamitsu, R Kistler, W Collins, D Deaven, L Gandin, M Iredell, S Saha, G White, and J Woollen. The NCEP/NCAR 40-year reanalysis project. *Bull Am Meteorol Soc*, 77(3):437–471, 1996.

[71] T Kanungo, DM Mount, NS Netanyahu, CD Piatko, R Silverman, and AY Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans Pattern Anal Mach Intell*, 24(7):881–892, 2002.

[72] Kiran Karra and Lamine Mili. Hybrid copula bayesian networks. In *Conference on Probabilistic Graphical Models*, pages 240–251, 2016.

[73] K-Y Kim, B Hamlington, and H Na. Theoretical foundation of cyclostationary EOF analysis for geophysical and climatic variables: concepts and examples. *Earth Sci Rev*, 150:201–218, 2015.

[74] FSi Koij and J Saba. Using cluster analysis and principal component analysis to group lines and determine important traits in white bean. *Procedia Environ Sci*, 29:38–40, 2015.

[75] Kevin B Korb and Ann E Nicholson. *Bayesian artificial intelligence*. CRC press, 2010.

[76] L. Kouadio, N. Newlands, A. Potgieter, G. McLean, and H. Hill. Exploring the potential impacts of climate variability on spring wheat yield with the APSIM decision support tool. *Agric Sci*, 6(7):686, 2015.

[77] Corey Lesk, Pedram Rowhani, and Navin Ramankutty. Influence of extreme weather disasters on global crop production. *Nature*, 529(7584):84–87, 2016.

[78] David X Li. On default correlation: A copula function approach. *The Journal of Fixed Income*, 9(4):43–54, 2000.

[79] Weixun Lu, David E Atkinson, and Nathaniel K Newlands. Enso climate risk: predicting crop yield variability and coherence using cluster-based pca. *Modeling Earth Systems and Environment*, 3(4):1343–1359, 2017.

[80] Weixun Lu, Nathaniel K Newlands, Odile Carisse, David E Atkinson, and Alex J Cannon. Disease risk forecasting with bayesian learning networks: Application to grape powdery mildew (erysiphe necator) in vineyards. *Agronomy*, 10(5):622, 2020.

[81] M Maadooliat, JZ Huang, and J Hu. Integrating data transformation in principal components analysis. *J Comput Graph Stat*, 24(1):84–103, 2015.

[82] Youcef Mammeri, Jean Baptiste Burie, Michel Langlais, and Agnes Calonnec. How changes in the dynamic of crop susceptibility and cultural practices can be used to better control the spread of a fungal pathogen at the plot scale? *Ecological modelling*, 290:178–191, 2014.

[83] D Maraun, F Wetterhall, AM Ireson, RE Chandler, EJ Kendon, M Widmann, S Brienen, HW Rust, T Sauter, and M Themeßl. Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Rev Geophys*, 48(3), 2010.

[84] Dimitris Margaritis. Learning Bayesian network model structure from data. Technical report, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science, 2003.

[85] Gregory S McMaster and WW Wilhelm. Growing degree-days: one equation, two interpretations. *Agricultural and Forest Meteorology*, 87(4):291–300, 1997.

[86] T Meng, R Carew, WJ Florkowski, and AM Klepacka. Analyzing Temperature and Precipitation Influences on Yield Distributions of Canola and Spring Wheat in Saskatchewan. *J Appl Meteorol Climatology*, 56(4):897–913, 2017.

[87] Perry Miller, Will Lanier, and Stu Brandt. Using growing degree days to predict plant stages. *Ag/Extension Communications Coordinator, Communications Services, Montana State University-Bozeman, Bozeman, MO*, pages 1–2, 2001.

[88] MS Mkhabela, Pl Bullock, S Raj, S Wang, and Y Yang. Crop yield forecasting on the Canadian Prairies using MODIS NDVI data. *Agric For Meteorol*, 151(3):385–393, 2011.

[89] Frances C Moore and David B Lobell. The fingerprint of climate trends on European crop yields. *Proc Natl Acad Sci*, 112(9):2670–2675, 2015.

[90] Malcolm J Morrison and Doug W Stewart. Heat stress during flowering in summer brassica. *Crop science*, 42(3):797–803, 2002.

[91] John R Nairn and Robert JB Fawcett. The excess heat factor: a metric for heatwave intensity and its use in classifying heatwave severity. *International journal of environmental research and public health*, 12(1):227–253, 2015.

[92] Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

[93] Nathaniel K Newlands, Gabriela Espino-Hernández, and R Scott Erickson. Understanding crop response to climate variability with complex agroecosystem models. *Int J of Ecol*, 2012, 2012.

[94] Nathaniel K Newlands, Weixun Lu, and Tracy A Porcelli. Downscaling of regional climate scenarios within agricultural areas across canada with a multivariate, multisite model. In *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science*, pages 335–340. Springer, 2015.

[95] Nathaniel K Newlands, David S Zamar, Louis A Kouadio, Yinsuo Zhang, Aston Chipanshi, Andries Potgieter, Souleymane Toure, and Harvey SJ Hill. An integrated, probabilistic model for improved seasonal forecasting of agricultural crop yield under environmental uncertainty. *Front Environ Sci*, 2:17, 2014.

[96] N.K. Newlands and D.S. Stephens. Increasing confidence in agricultural crop forecasts and climate adaptation decisions with causality analysis. Technical report, University of British Columbia (UBC), https://www.stat.ubc.ca/technical-reports-archive/doc/275.pdf, 2015.

[97] SAE Ouda, Samia Gouda Mohamed, and Fouad Ahmed Khalil. Modeling the effect of different stress conditions on maize productivity using yield-stress model. *International Journal of Natural and Engineering Sciences*, 2(1):57–62, 2008.

[98] Albert J Peters, Elizabeth A Walter-Shea, Lei Ji, Andres Vina, Mlchael Hayes, and Mark D Svoboda. Drought monitoring with ndvi-based standardized vegetation index. *Photogrammetric engineering and remote sensing*, 68(1):71–75, 2002.

[99] Carlotta Pirrello, Chiara Mizzotti, Tiago C Tomazetti, Monica Colombo, Paola Bettinelli, Daniele Prodorutti, Elisa Peressotti, Luca Zulini, Gino Angeli, Marco Stefanini, et al. Emergent ascomycetes in viticulture: an interdisciplinary overview. *Frontiers in Plant Science*, 10:1394, 2019.

[100] Luca Podofillini, Bruno Sudret, Bozidar Stojadinovic, Enrico Zio, and Wolfgang Kröger. *Safety and Reliability of Complex Engineered Systems: ESREL 2015*. CRC press, 2015.

[101] William Pool. Moves towards a Common Market in insurance. *Common Market Law Review*, 21(1):123–147, 1984.

[102] John R Porter and Megan Gawith. Temperatures and the growth and development of wheat: a review. *Eur J Agron*, 10(1):23–36, 1999.

[103] A.B. Potgeiter, G.L. Hammer, and D. Butler. Spatial and temporal patterns in Australian wheat yield and their relationship with ENSO. *Aust J Agric Res*, 53:77–89, 2002.

[104] S. Quiring and D. Blair. *The utility of global teleconnection indices for long-range crop forecasting on the Canadian Prairies*. PhD thesis, University of Winnipeg, 1999.

[105] J.O. Ramsay and B.W. Silverman. *Functional data analysis*. Springer-Verlag New York, 2005.

[106] D.K. Ray, J.S. Gerber, G.K. MacDonald, and P.C. West. Climate variation explains a third of global crop yield variability. *Nat Commun*, 6:5989, 2015.

[107] Siddheswar Ray and Rose H. Turi. Determination of number of clusters in k-means clustering and application in colour segmentation. In *The 4th International Conference on Advances in Pattern Recognition and Digital Techniques*, pages 137–143, 1999.

[108] Bill Rayens. An introduction to copulas. *Technometrics*, 42(3):317, 2000.

[109] Karen E Reid, Niclas Olsson, James Schlosser, Fred Peng, and Steven T Lund. An optimized grapevine RNA isolation procedure and statistical determination of reference genes for real-time RT-PCR during berry development. *BMC Plant Biology*, 6(1):27, 2006.

[110] Belén Rodríguez-Fonseca, Roberto Suárez-Moreno, Blanca Ayarzagüena, Jorge López-Parages, Iñigo Gómara, Julián Villamayor, Elsa Mohino, Teresa Losada, and Antonio Castaño-Tierno. A Review of ENSO Influence on the North Atlantic. A Non-Stationary Signal. *Atmosphere*, 7:87, 2016.

[111] Vittorio Rossi, Tito Caffi, and Sara E Legler. Dynamics of ascospore maturation and discharge in *Erysiphe necator*, the causal agent of grape powdery mildew. *Phytopathology*, 100(12):1321–1329, 2010.

[112] Simone Russo, Alessandro Dosio, Rune G Graversen, Jana Sillmann, Hugo Carrao, Martha B Dunbar, Andrew Singleton, Paolo Montagna, Paulo Barbola, and Jürgen V Vogt. Magnitude of extreme heat waves in present climate and their projection in a warming world. *Journal of Geophysical Research: Atmospheres*, 119(22):12–500, 2014.

[113] Simone Russo, Jana Sillmann, and Erich M Fischer. Top ten european heatwaves since 1950 and their occurrence in the coming decades. *Environmental Research Letters*, 10(12):124003, 2015.

[114] Mahdi-Salim Saib, Julien Caudeville, Maxime Beauchamp, Florence Carré, Olivier Ganry, Alain Trugeon, and Andre Cicolella. Building spatial composite indicators to analyze environmental health inequalities on a regional scale. *Environ Health*, 14(1):68, 2015.

[115] Mary Ann Sall et al. Epidemiology of grape powdery mildew: a model. *Phytopathology*, 70(4):338–342, 1980.

[116] Gianfausto Salvadori, Carlo De Michele, Nathabandu T Kottegoda, and Renzo Rosso. *Extremes in nature: an approach using copulas*, volume 56. Springer Science & Business Media, 2007.

[117] D. Sauchyn and S. Kerr. *Vulnerability and Adaptation: The Canadian Prairies and South America*, chapter Past and Future Drought: Lessons from Climate Science: Canadian Prairies Drought from a Paleoclimate Perspective. University of Calgary Press, Calgary, Alberta, Canada, 2016.

[118] H Scherm and XB Yang. Interannual variations in wheat rust development in China and the United States in relation to the El Niño/Southern Oscillation. *Phytopathology*, 85(9):970–976, 1995.

[119] Antonia Sebastian, EJC Dupuits, and O Morales-Nápoles. Applying a bayesian network based on gaussian copulas to model the hydraulic boundary conditions for hurricane flood risk analysis in a coastal watershed. *Coastal Engineering*, 125:42–50, 2017.

[120] Amir Shabbar, Barrie Bonsal, and Madhav Khandekar. Canadian precipitation patterns associated with the Southern Oscillation. *J Clim*, 10(12):3016–3027, 1997.

[121] Amir Shabbar and Walter Skinner. Summer Drought Patterns in Canada and the Relationship to Global Sea Surface Temperatures. *J Clim*, 17(14):2866–2880, 2004.

[122] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[123] A. Sklar. Fonctions de repartition a n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Universite de Paris*, 8:229–231, 1959.

[124] S.-J. Sohn, C.-Y. Tam, and H.-I. Jeong. How do the strength and type of ENSO affect SST predictability in coupled models. *Sci. Rep.*, 6, 2016.

[125] S. St. George and D. Sauchyn. Paleoenvironmental Perspectives on Drought in Western Canada - Introduction. *Can Water Resour J*, 31(4):197–204, 2006.

[126] R.C. Stone, G.L. Hammer, and T. Marcussen. Prediction of global rainfall probabilities using phases of the Southern Oscillation Index. *Nature*, 384:252–255, 1996.

[127] Liu Sun, Scott W Mitchell, and Andrew Davidson. Multiple drought indices for agricultural drought risk assessment on the Canadian Prairies. *Int J Climatology*, 32(11):1628–1639, 2012.

[128] Else Swinnen, Patrick Claes, Herman Eerens, Walter Heyns, Isabelle Piccard, and Peter Viaene. An integrated long time series of 1km resolution ndvi for europe from the noaa-avhrr and spot-vegetation sensors. In *2007 International Workshop on the Analysis of Multi-temporal Remote Sensing Images*, pages 1–5. IEEE, 2007.

[129] Xuezhi Tan and Thian Yew Gan. Non-stationary analysis of the frequency and intensity of heavy precipitation over canada and their relations to large-scale climate patterns. *Climate Dynamics*, 48(9-10):2983–3001, 2017.

[130] Edmar I Teixeira, Guenther Fischer, Harrij Van Velthuizen, Christof Walter, and Frank Ewert. Global hot-spots of heat stress on agricultural crops due to climate change. *Agricultural and Forest Meteorology*, 170:206–215, 2013.

[131] Bruce Thompson. Canonical correlation analysis. *Encyclopedia Sat Behav Sci*, 2005.

[132] Kevin E Trenberth, Aiguo Dai, Gerard Van Der Schrier, Philip D Jones, Jonathan Barichivich, Keith R Briffa, and Justin Sheffield. Global warming and changes in drought. *Nat Clim Chang*, 4(1):17–22, 2014.

[133] Ioannis Tsamardinos, Constantin F Aliferis, Alexander R Statnikov, and Er Statnikov. Algorithms for Large Scale Markov Blanket Discovery. In *FLAIRS conference*, volume 2, pages 376–380, 2003.

[134] Héctor Valdés-Gómez, Christian Gary, Philippe Cartolaro, Mauricio Lolas-Caneo, and Agnès Calonnec. Powdery mildew development is positively influenced by grapevine vegetative growth induced by different soil management strategies. *Crop Protection*, 30(9):1168–1177, 2011.

[135] Thomas Verma and Judea Pearl. *Equivalence and synthesis of causal models*. UCLA, Computer Science Department, 1991.

[136] K Wagstaff, C Cardie, S Rogers, and S Schrödl. Constrained k-means clustering with background knowledge. In *ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning*, pages 577–584, 2001.

[137] Shanshan Wang, Jianping Huang, Yongli He, and Yuping Guan. Combined effects of the Pacific decadal oscillation and El Nino-southern oscillation on global land dry–wet changes. *Sci Rep*, 4:6651, 2014.

[138] J. Watson, A.J. Challinor, T.E. Fricker, and C.A.T. Ferro. Comparing the effects of calibration and climate errors on a statistical crop model and a process-based crop model. *Clim Change*, 132(1):93–109, 2015.

[139] T.M.L. Wigley and T. Qipu. Crop-climate modeling using spatial patterns of yield and climate. Part 1: Background and an example from Australia. *J Clim Appl Meteorol*, 22(11):1831–1841, 1983.

[140] Laetitia Willocquet, F Berud, L Raoux, and M Clerjeau. Effects of wind, relative humidity, leaf movement and colony age on dispersal of conidia of *Uncinula necator*, causal agent of grape powdery mildew. *Plant Pathology*, 47(3):234–242, 1998.

[141] Laetitia Willocquet and M Clerjeau. An analysis of the effects of environmental factors on conidial dispersal of *Uncinula necator* (grape powdery mildew) in vineyards. *Plant Pathology*, 47(3):227–233, 1998.

[142] Laetitia Willocquet, D Colombet, M Rougier, J Fargues, and M Clerjeau. Effects of radiation, especially ultraviolet b, on conidial germination and mycelial growth of grape powdery mildew. *European Journal of Plant Pathology*, 102(5):441–449, 1996.

[143] Pingping Xie and Phillip A Arkin. Analyses of global monthly precipitation using gauge observations, satellite estimates, and numerical model predictions. *J Clim*, 9(4):840–858, 1996.

[144] Senshan Yang, Joanne Logan, and David L Coffey. Mathematical formulae for calculating the base temperature for growing degree days. *Agricultural and Forest Meteorology*, 74(1-2):61–74, 1995.

[145] Sandeep Yaramakala and Dimitris Margaritis. Speculative markov blanket discovery for optimal feature selection. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 4–pp. IEEE, 2005.

[146] Bin Yu, Xuebin Zhang, Hai Lin, and Jin-Yi Yu. Comparison of Wintertime North American Climate Impacts Associated with Multiple ENSO Indices. *Atmos Ocean*, 53(4):426–445, 2015.

[147] B Zheng, K Chenu, Fernanda DM, and SC Chapman. Breeding for the future: what are the potential impacts of future frost and heat events on sowing and flowering time requirements for Australian bread wheat (Triticum aestivium) varieties? *Glob Chang Biol*, 18(9):2899–2914, 2012.

[148] Aurelius A Zilko, Dorota Kurowicka, Anca M Hanea, and Rob MP Goverde. The copula bayesian network with mixed discrete and continuous nodes to forecast railway disruption lengths. In *6th International conference on Railway Operations Modelling and Analysis, RailTokyo2015, Narashimo, Japan, March 23-26, 2015; Authors version*. Citeseer, 2015.