University of Alberta

Lightning prediction models for the province of Alberta, Canada

by

Karen Blouin

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of requirements for degree of

Master of Science

in

Forest Biology and Management

Department of Renewable Resources

© Karen Blouin

Spring 2014
Edmonton, Alberta

Permission is hereby granted to the University of Alberta Libraries to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only. Where the thesis is converted to, or otherwise made available in digital form, the University of Alberta will advise potential users of the thesis of these terms.

The author reserves all other publication and other rights in association with the copyright in the thesis and, except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatsoever without the author's prior written permission.

ABSTRACT

Lightning is widely acknowledged as a major cause of wildland fires in Canada. On average, 250,000 cloud-to-ground lightning strikes occur in Alberta every year. Lightning-caused wildland fires in remote areas have considerably larger suppression costs and a much greater chance of escaping initial attack. Geographic and temporal covariates were paired with Reanalysis and Radiosonde observations to generate a series of 6-hour and 24-hour lightning prediction models valid from April to October. These models, based on cloud-to-ground lightning from the CLDN, were developed and validated for the province of Alberta, Canada. The ensemble forecasts produced from these models were most accurate in the Rocky Mountain and Foothills Natural Regions achieving hits rates of ~85%. The Showalter index, convective available potential energy, Julian day, and geographic covariates were highly important predictors. Random forest classification is introduced as a viable modeling method to generate lightning forecasts. Limitations and recommendations are also discussed.

ACKNOWLEDGEMENTS

I am sincerely grateful to my Supervisor, Dr. Mike Flannigan. His continual support, encouragement, and genuine concern for all of his students did not go unnoticed. I appreciate the input and recommendation made by my committee members, Bob Kochtubajda, Dr. Scott Nielsen (chair), Dr. Gerhard Reuter, and Dr. Xianli Wang. I am extremely thankful for the numerous hours of programming support provided by Dr. Xianli Wang of the University of Alberta. Without his expertise this could not have been accomplished. I am also thankful to the Western Partnership for Wildland Fire Science and the University of Alberta for providing the opportunity for this research to take place. A big thank you to the Fire Lab Group for being a sounding board and continually providing new and interesting perspectives. I am very thankful for my family and their never ending support and love (even the tough love). Finally, I thank my husband Jerald for being my rock and always encouraging me to chase my dreams.

The funding for this study was generously provided by Alberta Environment and Sustainable Resource Development. Lightning data were provided by Bob Kochtubajda of Environment Canada, and Radiosonde data were provided by Dr. Larry Oolman of the University of Wyoming.

CONTENTS

ACKNOWLEDGEMENTS ..... **II**

LIST OF FIGURES ..... **VIII**

LIST OF ACRONYMS ..... **XII**

CHAPTER 1. INTRODUCTION **1**

1.1 PREAMBLE **1**

1.2 LIGHTNING : BACKGROUND **2**

1.3 CONVECTIVE BASICS **4**

1.4 CLOUD ELECTRIFICATION **8**

1.5 LIGHTNING **10**

1.6 FIRE **14**

1.7 LIGHTNING-FIRE INTERACTIONS **19**

1.8 PREVIOUS WORKS **22**

1.9 OBJECTIVES **25**

CHAPTER 2. DATA AND METHODS **27**

2.1 STUDY AREA **27**

2.2 DATA **28**

2.2.1 PREDICTAND 31
2.2.2 PREDICTORS 32
Geographic and Temporal Covariates 32
Reanalysis II : Pressure Levels 32
Reanalysis II : Surface Grid 33
Reanalysis I : Pressure Levels 34
Radiosonde Observations 34
Calculated Variables 38

2.3 DATA PROCESSING AND MODIFICATIONS **40**

2.3.1 LIGHTNING DATA PROCESSING 42
2.3.2 REANALYSIS DATA PROCESSING 43
2.3.3 RADIOSONDE DATA PROCESSING 44
2.3.4 THIN-PLATE SPLINE 45

2.4 MODELING METHODS **48**

2.4.1 RANDOM FOREST: BACKGROUND 48
2.4.2 CONUNDRUM OF IMBALANCED DATA 49
2.4.3 RANDOM FOREST : MODELING 51
Preliminary Runs 51
Creating Random Forest Models 53

CHAPTER 3. RESULTS 56

3.1 EXPLORATORY RUNS **56**

3.2 DETERMINING TOP PREDICTORS **60**

3.3 LIGHTNING PREDICTION MODELS **64**

3.3.1 2.5° LATITUDE BY 2.5° LONGITUDE 67
Daily Lightning Prediction 67
6-hour Lightning Prediction 68
3.3.2 1.25° LATITUDE BY 2.5° LONGITUDE 70
Daily Lightning Prediction 71
6-hour Lightning Prediction 72
3.3.3 50KM BY 50KM 74
Boreal Forest : Daily Lightning Prediction 74
Boreal Forest : 6-hour Lightning Prediction 76
Parkland and Grassland : Daily Lightning Prediction 79
Parkland and Grassland : 6-hour Lightning Prediction 82
Mountain and Foothills : Daily Lightning Prediction 84
Mountain and Foothills : 6-hour Lightning Prediction 85

CHAPTER 4. DISCUSSION 91

4.1 TOP MODEL SELECTION **91**

4.2 PREDICTOR IMPORTANCE AND CONTRIBUTION **93**

4.2.1 PARTIAL DEPENDENCE PLOTS 93

4.3 RANDOM FOREST MODELING **98**

4.3.1 BENEFITS 99
4.3.2 DATA IMBALANCE 99
4.3.3 OOB ERROR ESTIMATE VS. INDEPENDENT PREDICTIONS 100
4.3.4 EFFECTS OF CORRELATION 101
4.3.5 NUMBER OF TREES 101
4.3.6 DISADVANTAGES 102

4.4 POSSIBLE SOURCES OF ERROR **102**

4.4.1 THIN-PLATE SPLINE 103
4.4.2 LIGHTNING MISCLASSIFICATION 103

4.5 LOOKING FORWARD **104**

CHAPTER 5. CONCLUSIONS..... 106

BIBLIOGRAPHY 109

LIST OF TABLES

***Table 1**: List of input variables. Abbreviations, full names, units of measure, and some additional information about data sources and resolution are given for all of the predictors initially considered.* _____ *30*

***Table 2:** List of the 10 data sets created for the two time frames (daily and 6-hour) and three spatial scales. The finest resolution scale (50km by 50km) is further subdivided into three separate zones. The number of observations remaining after all rows with NAs present were removed are also provided for all data sets. These 10 data sets are the basis for the lightning prediction models.* _____ *47*

***Table 3:** Contingency matrix for model forecasts. The A and D cells (highlighted in grey) show the correctly forecasted events and non-events, respectively.* _____ *55*

***Table 4:** List of the forecast skill criterion used to analyses the lightning prediction models. The inputs for the formulas can be found in the contingency matrix shown in Table 3.* _____ *55*

***Table 5:** Level of imbalance between the positive (events) and negative class (non-events) for each data set. The number of observations are for the training sets, however, the proportions hold true for both the training and validation sets.* _____ *58*

***Table 6:** Variable importance for a 1:1 BRF model with ntree=300 generated for the 6-hour Boreal Forest data set. The higher the MeanDecreaseGini value, the greater the variable importance and thus the rank.* _____ *62*

***Table 7:** List of top 15 predictors for each data set. Based on models built with a Balanced Random Forest (BRF) approach. The various data sets are listed.* _____ *63*

***Table 8:** Six predictions models were generated with the top 15 and top 12 variables from each data set. Let n be the number of observations in the minority class. The second and third columns show how the sampsize arguments for each model were specified to change the number of observations sampled from each class with replacement. The unbalanced model was run under default conditions making the imbalance similar to the original data imbalance outlined in Table 5.*_____ *64*

***Table 9:** Optimum daily ensemble models (0.6:1) produced with 15 variables for the 2.5° latitude by 2.5° longitude scale. The 15 variable balanced models are included for comparison. The three event forecast thresholds of ≥ 0.9, ≥ 0.7, and ≥ 0.5 are shown for each model. As H increases, the FAR and F increase, and the PAG and PC decrease.* _____ *68*

***Table 10:** Comparison of the two top ensemble models (0.6:1) produced with the top 12 and 15 variables for the 2.5° latitude by 2.5° longitude 6-hour forecast. The ensemble variations for the three event forecast thresholds of ≥ 0.9, ≥ 0.7, and ≥ 0.5 are shown for each model.* _____ *69*

***Table 11:** Comparison of the top two daily ensembles lightning forecast models for the 1.25° latitude by 2.5° longitude scale. The three event forecast thresholds of ≥ 0.9, ≥ 0.7, and ≥ 0.5 are shown for each model.* _____ *72*

Table 12: *Forecast skill comparison for the top two 6-hour ensembles lightning forecast models for the 1.25° latitude by 2.5° longitude scale. The three event forecast thresholds of ≥ 0.9, ≥ 0.7, and ≥ 0.5 are shown for each model.* _____ *73*

Table 13: *Ensemble forecast skill comparison of the optimum forecast model (five variable 0.6:1) generated for daily lightning prediction in the Boreal Forest. The optimum model is compared to a model variation often selected as a top forecast model for the coarser spatial scales (12 variable 0.6:1). The three event forecast thresholds of ≥ 0.9, ≥ 0.7, and ≥ 0.5 are shown for each model.* *76*

Table 14: *Ensemble forecast skill comparison of the 0.6:1 BRF models generated for 6-hour lightning prediction in the Boreal Forest. Ensemble models created from the top 12, 10, eight and five variables are included. The three event forecast thresholds of ≥ 0.9, ≥ 0.7, and ≥ 0.5 are shown for each model.* _____ *79*

Table 15: *Comparison of the daily Parkland and Grassland ensemble forecasts generated from the 0.6:1 BRF models for the top 12, 10, eight and five variables. The three event forecast thresholds of ≥ 0.9, ≥ 0.7, and ≥ 0.5 are shown for each model.* _____ *81*

Table 16: *Comparison of the 6-hour Parkland and Grassland ensemble forecasts generated from the 0.6:1 BRF models for the top 12, 10, eight and five variables. The three event forecast thresholds of ≥ 0.9, ≥ 0.7, and ≥ 0.5 are shown for each model.* _____ *83*

Table 17: *Comparison of the Mountain and Foothills ensemble forecasts generated from the 0.6:1 BRF models for the top 12 and 15 variables. The three event forecast thresholds of ≥ 0.9, ≥ 0.7, and ≥ 0.5 are shown for each model.* _____ *85*

Table 18: *Comparison of the Mountain and Foothills ensemble forecasts generated from the 0.6:1 BRF models for the top 12 and 15 variables. The three event forecast thresholds of ≥ 0.9, ≥ 0.7, and ≥ 0.5 are shown for each model.* _____ *86*

Table 19: *The model selected for each of the data sets are ranked from in order of decreasing skill from one to 10 for various forecast skill criteria.*_____ *93*

LIST OF FIGURES

Figure 1: *Vertical temperature profile of the atmosphere. The troposphere extends from the surface to around 12km altitude and has an average temperature lapse rate of 6.5°C km^{-1}. Most convective weather occurs within the troposphere although some very intense storms can breach the capping effect of the tropopause and extend into the stratosphere.* _____ *8*

Figure 2: *Dipole electrical structure of a convective cloud. The upper region and anvil accumulate a net positive charge while the cloud base becomes predominantly negative. Opposite charges attract forming induced charge pools on the ground.*_____ *11*

Figure 3: *Located in in North America, the study area, highlighted in red, covers the province of Alberta, Canada.* _____ *27*

Figure 4: *There are six Nation Regions in the province of Alberta separating the province into regions with similar terrain and ecology.* _____ *29*

Figure 5: *The location of the eight RAOB sounding locations used to provide information about the upper atmosphere conditions over Alberta. General locations are marked by yellow triangle with the corresponding station call names below.* _____ *35*

Figure 6: *Lightning prediction models were created for the province of Alberta at three different spatial scales. From left to right: 2.5 ° latitude by 2.5° longitude grid, 1.25° latitude by 2.5° longitude grid, and a 50km by 50km grid.* _____ *40*

Figure 7: *Three zones were created for the 50km by 50km spatial scale. These zones are based on the Natural Regions of Alberta and include the Boreal Forest Zone, Parkland and Grassland Zone, and the Rocky Mountain and Foothills Zone.*_____ *41*

Figure 8: *An example of a situation where data is only available for six of the eight RAOB stations. When using TPS to interpolate data for Alberta (highlighted), the northern most extrapolated values may be outside of an acceptable range as there is no point to bound the spline.* _____ *47*

Figure 9: *The boxes encompass the 1st and 3rd quartiles with the median indicated by a heavy center line and the mean marked by the red diamond. Whiskers extend to the minimum and maximum values. The Box plots represent the average values for13 consecutive years of lightning data from 1999 to 2011.* _____ *57*

Figure 10: *Comparison of the overall OOB error estimate for a series of balanced and unbalanced random forest models. Models were run on the daily 2.5° latitude by 2.5° longitude data set. The overall error rates are very similar for the two models.*_____ *58*

Figure 11: *Comparison of the class 0 (non-event) and class 1 (event) OOB error estimate for balanced and unbalanced random forest models generated with 12 and 15 variables.*_____ *59*

Figure 12: *The OOB error estimate trend for random forest models with different number of trees (ntree). The OOB estimate or error begins to balance around 100 trees and stabilizes by 200 trees.*_____ *60*

Figure 13: *The OOB error estimate (y-axis) fluctuations as the number of variables included in the random forest model are changed (x-axis). Note the different scales for the x and y axes.* __ *61*
Figure 14: *This plot demonstrates the overall, 1 (event), and 0 (non-event) prediction skills for various ensemble forecasts. All forecasts were generated with the top 15 variables for the 1.25° latitude by 2.5° longitude daily data. The legend represent the BRF (event: non-event) with the number to right specifying the ensemble threshold.* _____ *65*
Figure 15: *As the hit rate (1 skill) increases, the 0 prediction skill falls and the FAR increases. All forecasts were generated with data from the daily 1.25° latitude by 2.5° longitude 15 variable prediction ensembles. The legend represent the BRF (event: non-event) with the number to right specifying the ensemble threshold.* _____ *66*
Figure 16: *The relationship between the False Alarm Ratio (FAR), False Alarm Rate (F) and Equitable Threat Score (ETS). As the ensemble models are tweaked to maximize hit rate, the number of false alarms increase. This decreases the overall skill as measured by the ETS. All forecasts were generated with data from the daily 1.25° latitude by 2.5° longitude 15 variable prediction ensembles. The legend represent the BRF (event: non-event) with the number to right specifying the ensemble threshold.* _____ *66*
Figure 17: *The hit rates (%) for the various 2.5° latitude by 2.5° longitude daily prediction ensemble models. Models built with the top 15 variables are represented by bars, while the red line represents models generated from the top 12 variables. The BRF method applied to each model is shown below the x-axis with the three forecast thresholds (≥ 0.9, ≥ 0.7, and ≥ 0.5) listed on the x-axis. The unbalanced models on the far right were run under default randomForest conditions and are included to highlight the increase in skill by implementing the various BRF approaches.* _____ *68*
Figure 18: *The hit rates for the 2.5° latitude by 2.5° longitude 6-hour prediction ensembles. Models built with the top 15 variables are represented by the bars, while the red line represents the models generated from the top 12 variables. The BRF methods applied to each model are shown below the x-axis with the three forecast thresholds (≥ 0.9, ≥ 0.7, and ≥ 0.5) listed on the x-axis.* _____ *69*
Figure 19: *Variable importance plots for the top models selected for the 2.5° latitude by 2.5° longitude spatial scale. The variable importance for the daily prediction model is shown in Figure 19a while the 6-hour model variable importance is shown in Figure 19b. The MeanDecreaseGini values on the x-axis are relative values of each variables importance. The values cannot be compared between different models however the relative ranking of variables can be compared. A higher value assigned to a variable indicates it has a greater importance.* _____ *70*
Figure 20: *The hit rates for the various 1.25° latitude by 2.5° longitude daily prediction ensembles. Models built with the top 15 variables are represent by the bars, while the red line represents the hit rate of models generated from the top 12 variables. The BRF methods applied to*

each model are shown below the x-axis with the three forecast thresholds (≥ 0.9, ≥ 0.7, and ≥ 0.5) for each BRF ensemble listed on the axis. _____ *71*

Figure 21: *Hit rates of the 1.25° latitude by 2.5° longitude 6-hour prediction ensembles. Models built with the top 15 variables are represented by bars, and the red line represents models generated from the top 12 variables. The BRF methods applied to each model are shown below the x-axis with the three forecast thresholds (≥ 0.9, ≥ 0.7, and ≥ 0.5) listed on the x-axis. The unbalanced models on the far right were run under default randomForest conditions.* _____ *73*

Figure 22: *Variable importance plots of the top models selected for the 1.25° latitude by 2.5° longitude spatial scale. The daily prediction model is shown to the left (Figure 22a) while the 6-hour model is shown on the right (Figure 22b). The MeanDecreaseGini values on the x-axis are relative values of each variables importance. Actual values cannot be compared between different models however the relative ranking of variables can be compared. A higher value indicates greater importance.* _____ *74*

Figure 23: *The hit rates for the various Boreal Forest daily prediction ensembles. Models built with the top 15 variables are represented by bars, and the red line represents models generated from the top 12 variables. The BRF methods applied to each model are shown below the x-axis with the three forecast thresholds (≥ 0.9, ≥ 0.7, and ≥ 0.5) listed on the x-axis.* _____ *75*

Figure 24: *Comparison of the daily ensemble forecast models generated with a 0.6:1 BRF approach for the top 12, 10 and five variables. The hit rates are shown by broken lines while the proportion correct (PC) are given by solid lines.*_____ *76*

Figure 25: *The hit rates for the various Boreal Forest 6-hour prediction ensembles. Models built with the top 15 variables are represented by bars, and the red line represents models generated from the top 12 variables. The BRF methods applied to each model are shown below the x-axis with the three forecast thresholds (≥ 0.9, ≥ 0.7, and ≥ 0.5) listed on the x-axis.* _____ *77*

Figure 26: *Comparison of the 6-hour ensemble forecast models generated with a 0.6:1 BRF approach for the top 12, 10, eight, and five variables. The hit rates are shown by broken lines while the proportion correct (PC) are given by solid lines.* _____ *78*

Figure 27: *The hit rates for the various Parkland and Grassland daily prediction ensembles. Models built with the top 15 variables are represented by bars, and the red line represents models generated from the top 12 variables. The BRF methods applied to each model are shown below the x-axis with the three forecast thresholds (≥ 0.9, ≥ 0.7, and ≥ 0.5) listed on the x-axis.* _____ *80*

Figure 28: *Comparison of the daily ensemble forecast models generated with a 0.6:1 BRF approach for the top 12, 10, eight, and five variables. The hit rates are shown by broken lines while the proportion correct (PC) are given by solid lines.* _____ *81*

Figure 29: *The hit rates for the various Parkland and Grassland 6-hour prediction ensembles. Models built with the top 15 variables are represented by bars, and the red line represents models*

generated from the top 12 variables. The BRF methods applied to each model are shown below the x-axis with the three forecast thresholds (≥ 0.9, ≥ 0.7, and ≥ 0.5) listed on the x-axis. _____ *82*

Figure 30: *Comparison of the 6-hour ensemble forecast models generated with a 0.6:1 BRF approach for the top 12 and top 5 variables. The hit rates are shown by broken lines while the proportion correct (PC) are given by solid lines.*_____ *83*

Figure 31: *The hit rates for the various Mountain and Foothills daily prediction ensembles. Models built with the top 15 variables are represented by bars, and the red line represents models generated from the top 12 variables. The BRF methods applied to each model are shown below the x-axis with the three forecast thresholds (≥ 0.9, ≥ 0.7, and ≥ 0.5) listed on the x-axis.* _____ *84*

Figure 32: *The hit rates for the various Mountain and Foothills 6-hour prediction ensembles. Models built with the top 15 variables are represented by bars, and the red line represents models generated from the top 12 variables. The BRF methods applied to each model are shown below the x-axis with the three forecast thresholds (≥ 0.9, ≥ 0.7, and ≥ 0.5) listed on the x-axis.* _____ *85*

Figure 33: *Variable importance plots for the optimum models selected for each spatial and temporal scale. The models represent the Boreal Forest, Parkland and Grassland, and Mountain and Foothills from top to bottom. The daily models variable importance are displayed on the left and the 6-hour models to the right. The plots were nearly identical for the daily and 6-hour Boreal Forest therefore only the daily plot was included. The MeanDecreaseGini values on the x-axis are relative values of each variables importance. Actual values cannot be compared between different models however the relative ranking of variables can be compared. A higher value assigned to a variable indicates it has greater importance.*_____ *87*

Figure 34: *Ensemble forecast prediction accuracy for a randomly chosen day with lightning (July 13, 2005). The optimum Boreal Forest model was used to generate the ensemble forecast.* ____ *88*

Figure 35: *Ensemble forecast prediction accuracy for a randomly chosen day with lightning (July 23, 2011). The optimum Grassland and Parkland model was used to generate the ensemble forecast.* _____ *89*

Figure 36: *Ensemble forecast prediction accuracy for a randomly chosen day with lightning (July 19, 2005). The optimum Mountain and Foothills model was used to generate the ensemble forecast.* _____ *90*

Figure 37: *Partial dependence plots for Julian day for daily lightning prediction in the Mountain and Foothills (left) and daily lightning prediction at the 2.5° latitude by 2.5° longitude spatial scale (right). The y-axes are the logit of the probability of the lightning occurrence.* _____ *94*

Figure 38: *Partial dependence plots of SHOW00 (left) and CAPE00 (right) for daily lightning occurrence in the Mountain and Foothills. The y-axes are the logit of the probability of the lightning occurrence.* _____ *96*

Figure 39: *Partial dependence plot of elevation for daily lightning occurrence in the Mountains and Foothills. The y-axis is the logit of the probability of the lightning occurrence.*_____ *97*

LIST OF ACRONYMS

BRF	Balanced Random Forest
CAPE	Convective Available Potential Energy
CC	Cloud-to-Cloud
CG	Cloud-to-Ground
CINS	Convective Inhibition
CLDN	Canadian Lightning Detection Network
CSI	Critical Success Index
DEM	Digital Elevation Model
DOE	Department of Environment
EQLV	Equilibrium Level
ETS	Equitable Threat Score
FAR	False Alarm Ratio
F	False Alarm Rate
gpz	Geopotential Height
H	Hit Rate
KINX	K Index
LIFT	Lifted Index
LCC	Long Continuing Current
LCL	Lifted Condensation Level
LCLP	Lifted Condensation Level Pressure
LCLT	Lifted Condensation Level Temperate
MCC	Mesoscale Convective Complexes
MCS	Mesoscale Convective Systems
mslp	Mean Sea Level Pressure
NALDN	North American Lightning Detection Network
NCAR	National Center for Atmospheric Research
NCEP	National Centers for Environmental Protection
NLDN	National Lightning Detection Network
OOB	Out-of-Bag
POD	Probability of Detection
PAG	Post-Agreement
PC	Proportion Correct
PWAT	Precipitable Water

RAOB	Radiosonde Observation
RH	Relative Humidity
SHOW	Showalter Index
SWET	Severe Weather Threat Index
U wind	North-South Wind Component
UTC	Coordinated Universal Time
V wind	East-West Winds Component
WRF	Weighted Random Forest
WUI	Wildland-Urban Interface
Z	Zulu

CHAPTER 1. INTRODUCTION

1.1 PREAMBLE

Since the dawn of civilization people have recorded their interest, fear, and fascination of thunder and lightning (Rakov and Uman, 2003). Thunderstorms play a prominent role in early religious beliefs and ancient mythology as signals of gods' will, displeasure, or anger (Rakov and Uman, 2003; Schonland, 1964; Wåhlin, 1986). We now understand lightning is an atmospheric phenomena that helps maintain and discharge Earth's electric field and thunder is the auditory expression of the process. Even before our recorded interest, lightning was shaping and changing the world by creating ozone and igniting wildfires. Lightning-caused wildland fires have always been, and will most certainly always be, a part of the natural environment of Canada's vast forests (Flannigan and Wotton, 1991), as long as these forests exist.

It is well known that thunderstorms produce lightning, however, other convective systems such as dust storms, wildfire induced pyrocumulus clouds, and volcanic eruptions can also provide the conditions conducive to lightning to occurrence (Orville and Huffines, 1999). From Benjamin Franklin's 1750 experimental design used by Thomas-François Dalibard in 1752 to prove a charge differential exists between clouds and the ground, to current high-tech sensors, our fascination with lightning continually drives us to try to understand and quantify the intense power in the sky that we see, hear, and often fear. This thesis takes another step towards understand lightning and its impacts on wildfires by creating lightning prediction models to forecast the probability of lightning occurrence in Alberta. The models presented predict whether or not lightning will occur for a specific area and time with the intention of providing an additional decision support tool to increase the efficient use of resources for wildland fire management.

In order to begin working towards this goal, one must first understand the basics of what drives lightning and how lightning affects, and often initiates wildfires on the landscape. The introductory chapter walks through some background information including average global lightning occurrence rates with a focus on Alberta, Canada. An introduction to convective dynamics and thunderstorm formation is then presented which sets the stage for the following section on cloud electrification and separation of charges. Different kinds of lightning and their varying characteristics will then be explored,

followed by a look at wildland fire history in Canada and specifically Alberta. Lightning and fire are then considered together and their interactions and impacts are explored. Previous research and studies on lightning occurrence prediction are then considered. The chapter concludes with a section on the research objectives of this study.

1.2 LIGHTNING : BACKGROUND

Lightning is the discharge of static electricity in the sky or between the clouds and the ground (Rakov and Uman, 2003). Thunder is the auditory companion to lightning caused as the air around the discharge channel is heated so rapidly that it expands at a supersonic rate producing a shock wave (Christian et al., 2003). A thunderstorm refers to a convective storm system that produces lightning. At any moment there are ~2,000 thunderstorms occurring worldwide with lightning flashing in the sky and striking the Earth 25-300 times per second (Oliver, 2005). More than 8.6 million strikes occur per day on average; these strikes help to maintain the electrical energy balance, or global electric circuit, of the Earth and its atmosphere (Oliver, 2005).

Over the Great Plains of North America, about one in every six flashes strikes the ground, while over the Gulf Coast the rate of ground strikes is much higher at around one out of every 2.5 (Christian et al., 2003). Globally, the chance of lightning occurring over, or near to land, is ten times greater than over the open ocean and almost 80% of all flashes occur in the tropical zone between 30° N and 30° S (Christian et al., 2003). Rwanda stands out year round as a lightning hot spot with a peak mean annual flash density of 80 flashes km^{-2} (Christian et al., 2003). Lightning is responsible for ~24,000 deaths and ~240,000 injuries per year (Holle, 2008) although a small percentage of these occur in developed countries.

Zooming in to Canada, on average ~2.4 million lightning strikes hit the ground every year with the greatest number of strikes occurring during the month of July (Burrows and Kochtubajda, 2010). The average cloud-to-ground (CG) flash density is between 0.5 to 1.5 flashes km^{-2} $yr^{-1}$for the majority of the country with a significantly higher flash density of 2.8 flashes km^{-2} yr^{-1} occurring in an area of Southwestern Ontario (Burrows and Kochtubajda, 2010). Even the highest CG flash densities experienced here are an order of magnitude less than that of the global hotspot, Rwanda. Lightning activity is influenced by elevation, proximity to large water bodies, length of the warm season and thus intuitively latitude (Burrows and Kochtubajda, 2010). From 1999 to 2008, the average number of CG strikes per year ranged from around 1.98 million to 2.96 million

in Canada (Burrows and Kochtubajda, 2010). Studies of Canadian lightning characteristics show that strike occurrence follows a diurnal pattern with maximum activity in the late afternoon and minimal activity in the early morning hours although mid-continental prairie regions may see considerable nocturnal lightning (Burrows et al., 2002; Burrows et al., 2005).

Lightning occurs throughout the year in Canada, however, the vast majority occurs during the warm months between April and October (Burrows and Kochtubajda, 2010; Wierzchowski et al., 2002). A large percentage of the total strikes can often occur within a single, or a few, dominant storm events (Wierzchowski et al., 2002). Lightning hot spots in Canada include the Foothills of the Rocky Mountains and the Swan Hills in Alberta where a pronounced maximum of more than 30 days per year have cloud-to-ground (CG) lightning occurring (Burrows and Kochtubajda, 2010). A lightning flash occurs once every three seconds during the summer months in Canada (Canadian Safety Council, 2013). These strikes kill ~10 people per year, injure 92 to 164 people (Mills et al., 2008) and ignite around 4,000 forest fires (Stocks et al., 2002). Mills et al. (2010) conclude that lightning is one of the utmost sources of weather-related property damage in Canada.

Research on Canadian lightning climatology and its numerous impacts on human safety, utility services, travel, and ecological disturbance have advanced greatly since the introduction of the Canadian Lightning Detection Network (CLDN) in 1998. Prior to the introduction of this national system, small, short term studies were performed based on provincial, territorial, or private lightning detection data (Burrows and Kochtubajda, 2010). With the introduction of a national data base, more work is being done to quantify and understand lightning and its impacts. Thunderstorms affect almost all aspects of our lives. They can affect our recreational plans and health, cause damage to property, disrupt utilities and their transmission to end users, impact agriculture and range farming, cause large scale ecological disturbances, and even alter our emotional and psychological wellbeing. On the other hand, thunderstorms, and the associated lightning, can provide beneficial services such as maintaining Earth's electrical field (Oliver, 2005), aiding in soil fertilization (Boles and Verbyla, 2000), and providing a beauty that only nature can produce if one takes the time to watch the lightning dance across the sky.

Lightning is widely acknowledged as a major contributor and cause of wildland fire in Canada (Anderson, 2002; Burrows, 2002; Flannigan and Wotton, 1991; Stocks et al., 2002). Thunderstorms are also a significant contributor to blackouts with CG

lightning being recognized as the largest cause of electrical power disruptions in lightning-prone areas (Cummins and Murphy, 2009). Lightning strikes were responsible for the 1977 New York City blackout affecting nine million people, as well as the 1998 blackout that affected the Upper Midwest, Ontario, Manitoba and Saskatchewan for ~19 hour (Mills et al., 2010). A summary of literature relating to lightning impacts and damage, including impacts to air and ground transportation, telecommunications, fire, pipelines and agriculture can be found in Mills et al. (2010).

1.3 CONVECTIVE BASICS

Lightning activity is typically associated with convective systems (Rakov and Uman, 2003). Previous studies on lightning and occurrence prediction have explored many of its fundamentals and the mechanisms of its formation. The exact mechanisms of electrostatic formation and charge separations are not completely understood, yet there is a general consensus that cloud electrification is related to the interactions of the hydrometeors within the cloud (Anderson, 2002; Wåhlin, 1986). Before delving into lightning morphology and dynamics, a basic understanding of convective cloud and storm dynamics is needed. This section provides a brief and general introduction to thunderstorm dynamics and atmospheric convection. For additional information on convective storms and atmospheric charge separation please refer to texts such as Krehbiel (1986), Stolzenburg and Marshall (1998), and Stolzenburg et al. (1998).

Luke Howard developed a cloud naming system in 1803 where Latin names are used to classify cloud types based on their appearance from the ground. His system has since been modified and built upon over the years to create the International Cloud Atlas (World Meteorological Organization., 1987) that is now used all over the world. Some common names include (in the form *Latin* (translation)): *cumulus* (heap), *stratus* (layer), *cirrus* (curl of hair), *incus* (anvil), and *nimbus* (rain) (World Meteorological Organization., 1987). *Cumulonimbus* is used to describe a rain storm cloud implying the clouds shape is heaping and large or mountainous looking with falling precipitation (American Meteorological Society, 2012b).

Before a storm develops, some preliminary atmospheric conditions must first be met. There must be moisture, convective potential, and a mechanism of lift. There are four general convective storm types: single cell or air mass storms, multicell storms, squall lines, and supercell storms. All of these storms are supported by strong up drafts that facilitate the transportation of water droplets from the lower, warm part of the storm to

the upper, cooler region. Moore and Vonnegut (2010) describe thunderclouds as large heat engines running off of solar energy with water vapor as the agent of heat transfer. The convective heart of all thunderstorms has cellular structure(s) with a strong updraft ($\geq$10 m s^{-1}) a few kilometers in diameter (Liu et al., 2010). Thunderstorms typically range from three kilometers to more than 20km in vertical extent (Rakov and Uman, 2003). The horizontal span of active thunderstorms can range greatly from air-mass thunder storms roughly three kilometers to 50km in diameter, to a multicell squall lines, Mesoscale Convective Systems (MCS), and Mesoscale Convective Complexes (MCC) extending hundreds of kilometers (Rakov and Uman, 2003).

All thunderstorms undergo three stages of evolution: (1) developing stage, a towering cumulus cloud forms and develops with strong updrafts as warm moist air is lifted; (2) mature stage: the storm continues to grow in vertical and horizontal extent as warm moist air rises and then spreads out once the equilibrium level is reached creating an anvil shaped cloud, *cumulonimbus incus*. A downdraft is also present. The (3) dissipation stage is reached when the updraft is no longer sufficient to support the storm as the storm becomes dominated by the strong downdraft and begins to collapse (Rakov and Uman, 2003). A supercell storm has a long lived, rotating convective updraft separate from the downdraft but undergoes a similar, but often more intense and long lived evolutionary process and is often associated with high winds, large damaging hail and tornado formation. For simplicity sake, from here on convection and electrification mechanics are discussed from a single cloud (single-cell storm) perspective unless otherwise stated. It should be noted that in multicell storms, MCCs, and MCSs which have a series of cells at varying stages of evolution, these processes can be occurring simultaneously, and in varying stages or evolution.

All of these thunder clouds and thunderstorms are the result of one or more cumulonimbus cloud(s). While not all cumulonimbus clouds produce lightning, a thundercloud is a term that refers to a lightning producing cumulonimbus cloud (Rakov and Uman, 2003). Although sometimes used interchangeably, a thundercloud is a single cell or air-mass storm while a thunderstorm refers to a system of thunderclouds (Rakov and Uman, 2003). Lightning can be produced from other cloud types such as stratus type clouds, pyrocumulus clouds (Orville and Huffines, 1999), and clouds caused by nuclear explosions, volcanic eruptions and large scale sand storms (Rakov and Uman, 2003). Pyrocumulus clouds are convective systems that are induced by the intense heat and winds created by large and intense wildfires (Orville and Huffines, 1999).

One such instance was a pyrocumulus created during the Chisholm fire in May 2001. The induced pyro-cumulonimbus was so intense it reached 2.5km to three kilometers above the tropopause where it introduced large amounts of smoke and ashes into the stratosphere and supported a dense, high-intensity pocket of positive lightning activity (Rosenfeld et al., 2007). This cloud had a high density of intense positive lightning compared to the surrounding clouds which were dominated by negative strikes (Rosenfeld et al., 2007). Observations from the Chisholm fire induced pyrocumulus cloud support Williams et al. (2005) hypotheses that flash rate increases with updraft speed and that positive lightning is related to a high liquid water content in the mixed phase region (Rosenfeld et al., 2007).

Convection of warm, moist, buoyant air occurs most often in the troposphere. The troposphere is located in the lower atmosphere and extends from the surface to the tropopause (American Meteorological Society, 2012d). There is a negative temperature gradient in this region. As you move upward in the troposphere, the surrounding environmental temperature and pressure are decreasing. Air parcels of warm, moist buoyant air rise and cool at the dry-adiabatic lapse rate. As the parcel continues to rise and cool it will eventually reach a level where the humidity of the parcel becomes greater than the saturation point (vapour pressure becomes equal to the saturation vapor pressure). At this point the dewpoint and temperature of the parcel become equal and the parcel is said to have reached the Lifted Condensation Level (LCL). Any further lifting and cooling leads to a phase change where the water vapor begins to condense onto very small airborne particulates creating small water droplets known as hydrometeors[1]. These droplets make up the visible cloud, the bottom of which often represents the height of the lifted condensation level (Rakov and Uman, 2003).

As these particles cool and condense, they undergo a phase change releasing latent heat of condensation. This heat release further fuels the updrafts. If the air parcel remains warmer than the surrounding environment, it continues to rise, cool, condensate, and release latent heat which warms the parcels and further drives the process. Buoyancy is also aided by the fact that air with a high water vapor content is less dense than dry air

[1] Hydrometeors are solid or liquid water particles within the atmosphere that make up the visible clouds. If they are large enough to be influenced by gravity they are called precipitation (Rakov and Uman, 2003).

at the same temperature. The higher the vapor content, the less dense the parcel, and the more buoyant it becomes. This continues as long as the environmental lapse rate is larger than the moist adiabatic lapse rate (0.6ºC/100m) and the atmosphere is considered unstable (Rakov and Uman, 2003). Once the parcel reaches the zero degree isotherm some of the water droplets begin the freeze. The rest of the droplets enter a super-cooled phase where they are between zero and -40°C but remain in a liquid state (Rakov and Uman, 2003). The droplets all turn to ice once the temperature is less than -40°C (Rakov and Uman, 2003). Electrification is thought to occur in the mixed phase region between the zero degree and -40°C isotherms where supercooled liquid, vapor, and ice coexist (Rakov and Uman, 2003; Williams, 1989).

Hail, ice, graupel, and liquid water are transported up by strong updrafts, and down within downdrafts driven by gravity. As the ice, hail and graupel fall into the lower warmer regions they collide with liquid water and the droplets adhere to the frozen surface releasing latent heat of fusion as they undergo the phase change. This heat causes a slight warming of the frozen surface resulting in a mass that is slightly warmer than the environment. This relatively warm mass is known as graupel or soft hail. The graupel grows by accretion of the supercooled cloud droplets while the ice crystals grow as water vapor deposits onto the crystalline structure (Latham, 1991). Some frozen water may also melt and evaporate off of the ice or graupel. This process requires a strong updraft to transport the hail and graupel up within the cloud. Lightning is highly unlikely if the maximum updraft speed does not reach 10 to 12m s^{-1} and maintain an average speed greater than six to seven m s^{-1} (Zipser and Lutz, 1994).

The stratosphere is located above the troposphere and separated by the tropopause (Figure 1). The tropopause is marked by a temperature equalization or inversion where the negative temperature gradient of the troposphere reverses to a positive gradient and temperatures begin to increase with increasing height (American Meteorological Society, 2012c). This inversion and positive temperature gradient acts as a cap on convection and parcel buoyancy preventing most clouds from extending into the stratosphere (Rakov and Uman, 2003). In some cases the updraft is strong enough to overcome the capping force and clouds or storm systems are able to extend into the stratosphere with cloud tops reaching up to 20km in altitude (Rakov and Uman, 2003). The mechanisms of convective cloud formation have been discussed however some additional processes must first set the stage for lightning to occur. A separation of

charges must take place where electrons and positive ions are transported to different regions of the developed cloud to form an electrical dipole (Oliver, 2005).

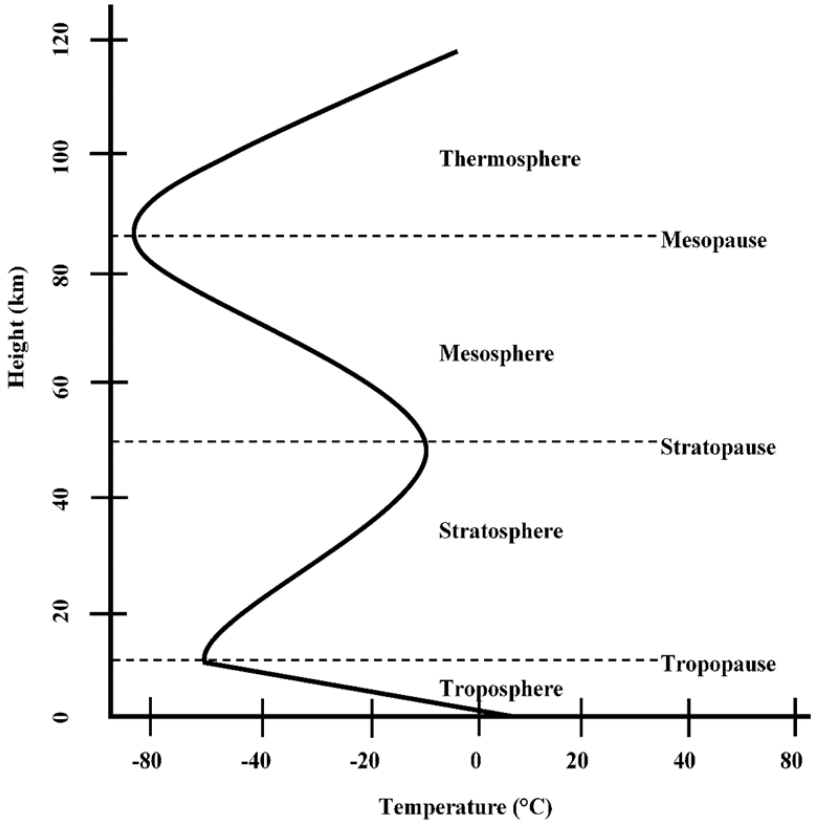

Figure 1: Vertical temperature profile of the atmosphere. The troposphere extends from the surface to around 12km altitude and has an average temperature lapse rate of 6.5°C km^{-1}. Most convective weather occurs within the troposphere although some very intense storms can breach the capping effect of the tropopause and extend into the stratosphere.

1.4 CLOUD ELECTRIFICATION

While we may understand some of the conditions required to produce lightning, the exact mechanisms of formation are still debated. That said, theories tend to focus on the generation of an electric field within a thunder cloud by some mechanism of electric charge separation. The Earth and atmosphere make up a global electric circuit that is constantly changing. In this circuit, the Earth's surface typically holds a net negative charge while the air above tends to be net positive, creating a vertical electric field (Rakov and Uman, 2003). This electric field's surface potential gradient is approximated to be roughly 120V for every meter increase in altitude near the Earth's surface under fair

weather conditions (Harrison, 2004). The global electric circuit is thought to be changed and maintained by the fair weather current, precipitation, corona or point discharge currents, and lightning (Harrison, 2004; Williams and Heckman, 1993). A corona is a point discharge involving a number of individual streamers within the immediate vicinity of a ground object that acts as an electrode but is not self-propagating (Rakov and Uman, 2003). Lightning rods use the characteristics of corona discharge to create a "preferred" path for lightning by concentrating charge at a small point.

Recent studies tend to agree that ice, hail, and graupel play a quintessential role in charge separation within a cloud or storm system and thus are highly important to lightning development (Wåhlin, 1986; Williams et al., 2005; Zipser and Lutz, 1994). Storms that lack a strong updraft typically fail to produce large quantities of ice, hail, and graupel and thus do not typically produce lightning. While a consensus on a quantitative, detailed description of charge separation with in the turbulent environment of a convective cloud is still lacking, most atmospheric physicists agree that the process involves collisions between supercooled water droplets and ice particles, graupel, and hail in the mixed phase region (Oliver, 2005). It is believed that as these collisions occur, electrons are sheared off of the ascending water particles and collected on the descending hail and graupel. This leads to a separation of electrical charges known as the polarization of the cloud.

As the convection continues within the cloud, these collisions continue to occur and a net positive charge pool develops in the upper region of the cloud mirrored by a net negative charge pool at the base creating a positive dipole electric field. The longer the system can sustain a strong updraft, the greater the charge differential potential. At this point the cloud can be thought of as a battery with its positive terminal facing upwards and the negative terminal towards the ground (Figure 2). The terminals are not in direct contact with each other, or any other terminal, rather the atmosphere all around is acting as an insulator. The voltage difference between the two charged centers within the cloud can reach several million volts (Oliver, 2005) The negative charge pool is usually found between the -10°C and -35°C levels where the cloud contains ice and supercooled liquid water. Strong updrafts cause an increased altitude and decreased temperature of the negative cloud charge center which can also be affected by local terrain (Krehbiel, 1986; Livingston and Krider, 1978; Stolzenburg et al., 1998). Krehbiel (1986) also notes that the altitude of the negative charge center appears to remain relatively stable throughout

the storm evolution while the positive charge pool increases in height as the storm develops.

This is a simplified description of cloud charge differentials and it must be noted that these are extremely complex systems that are still not fully understood. This is an introductory look at the most common, generalized charge layout however other charge pool locations can, and do occur (Krehbiel, 1986; Stolzenburg and Marshall, 1998). The charge differential must build up to a point where the charge strength is greater than the insulation strength of the atmosphere. Once the strength of the electron field is greater than the strength of the atmospheres insulation (dielectric strength), the breakdown threshold of the air is met and a transfer of charge occurs. This discharge is known as lightning.

1.5 LIGHTNING

The term "flash" is usually used to describe any lightning discharge whether it is within a cloud, between clouds, or between clouds and the ground (Rakov and Uman, 2003). The word "strike" is a more specific term describing a flash that makes contact with the ground or with a grounded object (Rakov and Uman, 2003). "Component stroke" or "stroke" refers to the part of a cloud-to-ground strike where a downward leader and upward return stroke occur (Rakov and Uman, 2003). The total energy produced by lightning depends on its polarization, duration, and magnitude (Oliver, 2005). For example, a typical negative lightning strike has a current of ~30,000A producing a peak channel temperature of 30,000K while a positive strike can be 10 times stronger (Rakov and Uman, 2003; Rust, 1986).

There are two general categories used to classify lightning flashes: lightning that flashes across the sky, and lightning that strikes the ground. Cloud-to-cloud (CC) lightning is the most common type of lightning accounting for ~70% to 85% of all flashes typically occurring as intracloud or intercloud discharges (Rakov and Uman, 2003). Intracloud lightning are discharges between the lower negative (typically four to eight kilometers altitude) and upper positive (typically eight to 12km altitude) dipoles of a cloud (Cummins and Murphy 2009) while intercloud flashes occur between two different clouds. Intracloud and intercloud lightning occur within or between clouds and therefore do not play a role in lightning-caused fires. Cloud-to-ground (CG) lightning is the special case where a transfer of charge occurs between a cloud and the Earth, or some object on or near the Earth's surface. Cloud-to-cloud lightning occurs more often than

CG lightning due to stronger electrical fields and a lower distance between charge pools. This thesis focuses on CG lightning as strikes interacting with the ground are of concern when dealing with potential lightning-caused fire ignitions. From here on, all references to strikes, strokes, flashes or lightning are referring to CG lightning unless explicitly stated otherwise.

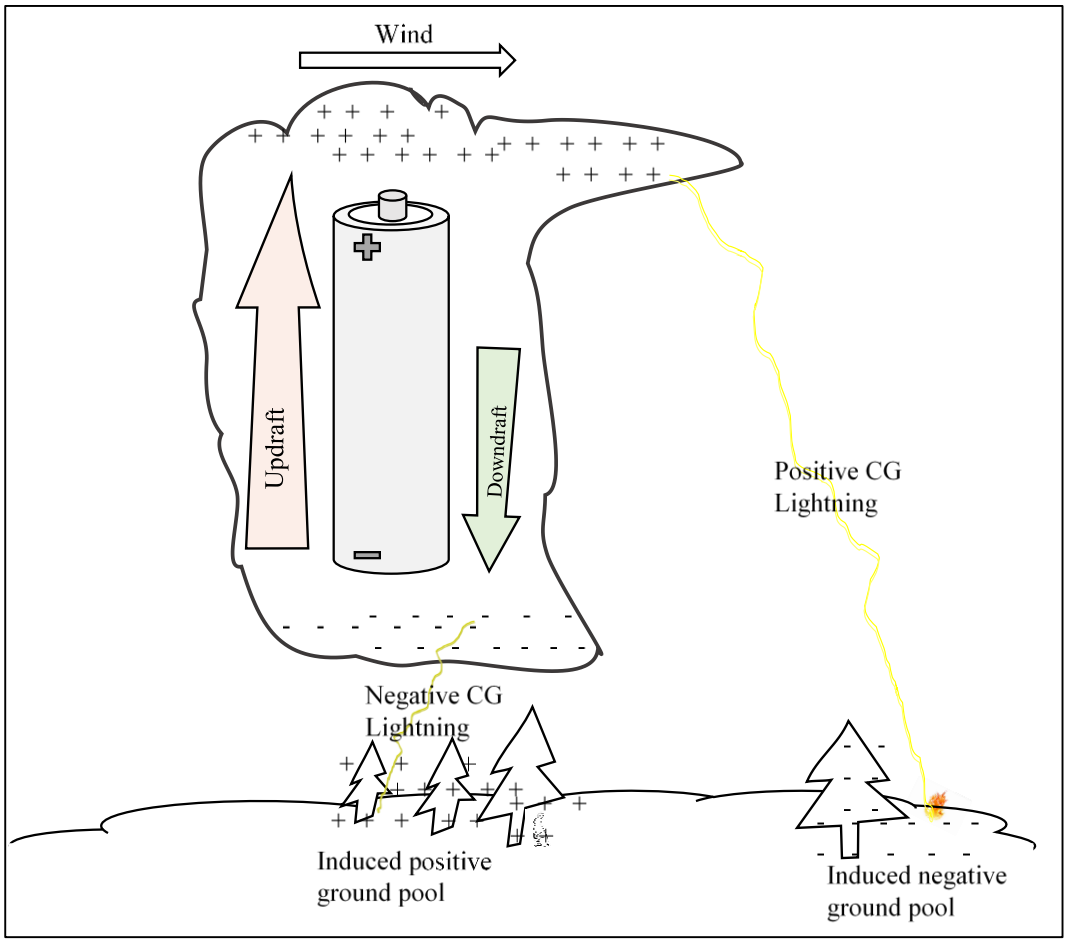

Figure 2: Dipole electrical structure of a convective cloud. The upper region and anvil accumulate a net positive charge while the cloud base becomes predominantly negative. Opposite charges attract forming induced charge pools on the ground.

Cloud-to-ground lightning requires an additional charge differential to occur in addition to the cloud charge differentials. Electrical opposites attract, therefore, the negative cloud base induces a positive charge pool on the ground below (Rakov and Uman, 2003). This pool will follow the cloud until the charge differential is neutralized. A negative charge pool may also develop on the ground below the positively charged anvil of the cloud (Figure 2). The corresponding charge differential between the two regions of the cloud and the two induced pools on the ground make CG lightning possible.

Berger (1967) described four basic types of CG lightning: (1) downward negative, (2) upward negative, (3) downward positive, and (4) upward positive. Globally, downward negative strikes account for almost 90% of all CG strikes (Rakov and Uman, 2003). Less than 10% of strikes are downward positive (Rakov and Uman, 2003). Upward strikes are highly uncommon and are thought to only occur from very tall objects (>100m) or from objects at mountain tops (Rakov and Uman, 2003). This thesis does not attempt to differentiate between downward and upward strokes, all strokes should be considered as downward positive or downward negative.

There are three main modes of charge transfer from clouds to the ground: (a) dart leader return stroke sequence, (b) continuing currents, and (c) M-components which are briefly summarized below from Rakov and Uman (2003). The dart leader return stroke sequences (a) involve a descending leader that creates a channel, or conductive path, towards the ground. A negative charge follows the path to the ground and a return stroke then goes back up the path from ground to cloud, and neutralizes the leader. Continuing currents (b) are a longer lived current (tens to hundreds of milliseconds) that form a quasi-stationary arc between the cloud and ground that persists during the interstroke interval (Cummins and Murphy, 2009) while, M-components (c) are surges that take place in a continuing current (b).

When a discharge occurs between the negative cloud base and the positive ground pool there is a negative transfer of charges from the cloud base to the ground. This is known as negative CG lightning. As the charge differential between the negative cloud base and positive ground pool increases, a negatively charged channel known as a stepped leader begins to descend from the cloud. It descends very rapidly with 20μs to 50μs between steps (Rakov and Uman, 2003). The stepped leader is invisible to the human eye. A "leader" refers to any self-propagating electrical discharge that creates a channel with an electrical conductivity in the order of 10^4S m^{-1} (Rakov and Uman, 2003). As it gets closer to the ground it causes the positive pool to focus and intensify below. A positively charged streamer then reaches up from an object on the ground or directly from the ground. A streamer is typically characterized by its lower electrical conductivity then the leader (Rakov and Uman, 2003). When the stepped leader and the streamer channels connect, a transfer of negative current occurs from the cloud to the ground. This current flow creates the flash of light that we call lightning.

If there is sufficient charge remaining, there may be additional strokes (return strokes) through the same channel (Rakov and Uman, 2003). This is known as

multiplicity and is what we see as the prolonged flickering of a strike. Return strokes have a peak current reaching up to more than 200kA lasting tens of microseconds and may occur in quick succession with 20 to 100ms between strokes (Cummins and Murphy, 2009). A negative lightning strike typically has a multiplicity of two to four return strokes but may have as many as 20 (Cummins and Murphy, 2009). Subsequent flashes may also contact the ground at a different point, up to several kilometers from the initial stroke point (Cummins and Murphy, 2009).

Electricity always takes the path of least resistance. Consequently, taller objects do tend to be hit more often as there is less atmospheric insulation between the negatively charged cloud base and the positive ground. This does not mean that taller objects will always be preferentially hit. Lightning follows the path of least resistance and strikes the closest object once the insulated properties of the atmosphere are overcome, therefore it can strike open ground even if a tree or tall building is nearby. As the stepped leader begins to descend it is "blind" to the ground. Once it is close enough to the ground to be influenced, a relatively small area around the stepped leader is likely to be hit. If a tall object is within that region it may be hit but if the total resistance is less to open ground than to a tall object, the ground will be hit.

Positive CG lightning occurs in a similar fashion to that of negative CG strikes. Positive strike tends to originate from the anvil or upper region of the cloud where the positive charge pool has built up. In this case the stepped leader is positively charged and forms a channel toward to ground where it induces the concentration of a negatively charged streamer that reaches towards it. The net transport of positive charge from cloud to ground results, hence the name "positive cloud-to-ground strike". Positive CG lightning accounts for a small portion of strikes but are of particular importance when considering lightning from the wildland fire perspective.

Positive strikes have a much longer flash duration and higher peak current than negative strikes (Rust, 1986). These characteristics are due in part to the location of origin of the strike from within the cloud. As previously mentioned, positive strikes originate from the upper region of the cloud and therefore must travel much greater distances (10 to 20km) in order to contact the ground. In other words, thinking back to the battery analogy, there is a greater amount of insulation between the two charge terminals. The strength of the charge differentials must therefore be much higher to overcome the insulation threshold. Positive strikes can have a peak current 10 times greater than negative strikes and have been measured as strong as 300,000A or one

billion volts (Rust, 1986). The majority of positive CG strikes have one return stroke which is often followed by a continuing current ranging from a few milliseconds to 250ms (Rust, 1986). As a reminder, continuing currents are longer lived current (tens to hundreds of milliseconds) that form a quasi-stationary arc between the cloud and ground (Rakov and Uman, 2003). It has been found that continuing current in positive return strokes can be in excess of 10kA for up to 10 to 40ms (Brook et al., 1982). In other words, the current transferred is an order of magnitude or larger than that of negative flashes.

While negative strikes typically strike the ground below the cloud or storm, positive lightning usually strikes around the edges and can strike the ground more than 16km away. Positive strikes have been reported to occur most commonly in the latter stage of a storms life and originate from the stratus or anvil (MacGorman and Rust, 1998). This means that a positive strike has a much greater chance of discharging where it has not yet rained and is not currently raining (Rust, 1986). This is known as a dry strike. The ability of positive lightning to strike so far away from the storm adds another dimension to the risks. People may not anticipate danger from the storm and often fail to take necessary precautions. Because the characteristics of positive lightning including higher peak current, longer duration, higher channel temperature, and ability to dry strike, positive strikes are most likely to cause fire starts (Rust, 1986). Additional information on lightning can be found in MacGorman and Rust (1998), Rakov and Uman (2003), and Schonland (1964) among others.

Both lightning and forest fires occur year round in Canada with the majority of the strikes and fires occurring during the warm months (Burrows and Kochtubajda, 2010). Warm season strikes are of particular interest as they often occur when the Canadian forests are warm and dry enough to be ignited by CG lightning. The start of the lightning and fire seasons usually corresponds with the melting of snow, warmer temperatures, and increasing length of daylight hours. The seasons typically end with the southward movement of the Arctic cold front (Kochtubajda et al., 2006).

1.6 FIRE

Wildland fires burn an annual average of 2.5 million ha of Canada's forested area (Stocks et al., 2002) resulting in direct suppression cost ranging from CAD $500 million to one billion dollars per year (Flannigan et al., 2009). Although wildland fire disturbance is a natural and necessary feature of many of Canada's fire driven ecosystems (Rorig and

Ferguson, 1999), human development, valued resources, and health concerns limit our capacity to allow fire to roam freely. Historically wildland fires were able to burn at will over most of Canada. During the late 19th century, human induced suppression measures came into play. In the early 20th century, contemporary fire management in North America became focused around strict suppression and prevention (Martell, 2001). Since then we have learned that suppression is unnatural and has led to many unanticipated and undesirable consequences. Fire exclusion in Canada's National Parks diminished the landscape mosaics which served as the reasoning for protecting the parks in the first place (Weber and Stocks, 1998). Fire suppression also often led to an aging, increasingly homogeneous forested landscape in many regions. The changing conditions resulting from fire suppression have created some areas that are conducive to more severe fire behavior.

Today we are dealing with the consequences of past fire management decisions. To worsen matters, fire and forest managers are also faced with the results of another anthropogenic induced change, climate change. In recent decades, Canada has experienced some worrisome trends including increased area burned and fire season length (Mike Flannigan, personal communication). These trends have been linked to changes in temperature, precipitation and extreme weather events. Climate models suggest that thunderstorm, and thus lightning activity will increase with projected rising temperatures (Price and Rind, 1994) which may translate into more lightning-caused fires (Kochtubajda et al., 2006). Future projections predict local and regional warming with the greatest temperature increases at higher latitudes during the winter months (Weber and Flannigan, 1997). As our climate is shifting, so too are the human dimensions of Canada. Demographic changes, increasing population, and shifts in settlement patterns result in new Wildland-Urban Interfaces (WUI). As urban and wild become intertwined the physical world around us changes and we realize our connection and responsibility to Canada's vast wild lands is paramount.

Of the ~8,000 wildland fires that occur per year in Canada, only three percent (roughly 200 fires) are responsible for 97% of the total area burned (Stocks et al., 2002). These fires often impact water quality, air quality, and can directly influence human health and safety. Although a natural and integral part of forest ecology, when wildland fires come into contact with settlements and infrastructure the major disruptions and damage can carry a large financial burden. An example of the economic costs of fire can be found in the analysis of the economic impact of the Chisholm, Alberta fire (Rittmaster

et al., 2006). Rittmaster et al. (2006) estimated a CAD $20 million loss in timber supply, and $10 million cost of suppression over seven days associated with the 2001 Chisholm fire. They also estimated the one-day health impacts to be in the range of nine million dollars to $12 million. In addition, wildfire impacts on tourism and recreation can cost tens of millions of dollars in lost visitation revenue (Martell and Sun, 2008). In the spirit of unbiased assessment it must also be noted that wildfires also employ numerous people in both seasonal and full time positions directly and indirectly (ex. personal protective equipment manufacturing).

Encompassing ~315 million ha, the boreal forest is the largest forested area in Canada (~315 million ha) containing the majority of the wooded area (Weber and Stocks, 1998) and covering a large swath of continental Canada. This region is distinct due to its short warm summers and long cold winters. Lightning fires are the dominant natural disturbance (Krawchuk et al., 2006) and an underpinning ecosystem process responsible for organizing the physical and biological characteristics of North America's boreal forests with fire regime being the key aspect (Weber and Flannigan, 1997). Canada's largest fires, and greatest fire related problems, occur within the boreal belt where over 97% of total area burned occurs (Pyne, 2007). The greatest annual areas burned on record occur within the past few decades with the exception of almost four million ha burned in 1961 (Weber and Stocks, 1998). Part of this increasing trend in area burned may be attributed to modern technological advances. Large remote areas could have burned undetected prior to extensive resource exploration and the use of remote sensing satellites for resource management and monitoring. Some common Canadian fire trends include a steady increase in the annual number of fires in Canada since the 1930s (possibly due to increasing population/land occupation, byproduct of suppression efforts, and better fire detection), increase in annual area burned over the last three decades, and an increasing recognition of the integral role of lightning.

The boreal forest is a mosaic dominated by trembling aspen (*Populus tremuloides*), white birch (*Betula papyrifera*), black spruce (*Picea mariana*) and jack pine (*Pinus banksiana*) with small clusters of other over story trees (Pyne, 2007). Trembling aspen and paper birch are deciduous trees that typically create fire resistant blocks during much of the year while the coniferous black spruce and jack pine provide clusters apt for intense burning and crown fires (Pyne, 2007). Dense growing jack pine and lodgepole pine (*Pinus contorta*) appear to have co-evolved with fire as they rely on heat to open their otherwise sealed serotinous cones (Burns et al., 1990) and are able to

carry fire throughout most of their life cycle (Pyne, 2007). The majority of the province of Alberta lies outside of the Canadian shield and its abundance of water pockets (Pyne, 2007). Instead, the boreal forest in Alberta lies over the western plains where fire can roam almost unchecked by natural barriers such as lakes and rock outcrops (Pyne, 2007). A changing fire regime will alter boreal forest dynamics resulting in different landscape mosaics, age class distribution, and forest boundaries (Weber and Flannigan, 1997).

Wildfire activity in Canada is increasing (Podur et al., 2002). While climate change is strongly believed to be a leading cause of increased fire activity and severity (Flannigan and Van Wagner, 1991), other factors such as increases in recreational and industrial use of forested land, increased fuel loading due to suppression, and better detection and recording methods have also contributed to this trend (Tymstra et al., 2005). Pre-suppression fire return intervals of ~45 to 69 years have been identified for Northern Alberta through the use of age class distribution analysis of forest inventory data and paleoecological records of charcoal deposits in northern lake sediments (Tymstra et al., 2005). Paleoclimate and paleoecology can be studied to provide insight to the future. In contrast to the law of uniformitarianism (in a geological sense) where the present is the key to the past, reconstruction of past environments may provide a guide to the future when trying to piece together a changing boreal fire regime. Pre-European settlement, fires were the result of indigenous burning and lightning ignition. These fires ran without contemporary suppression efforts likely resulting in a much larger annual area burned than we experience today (Weber and Stocks, 1998).

As climate changes, and the fire season lengthens, fire crews will be struggle to keep up with the changes in fire regime. The boreal forest fire regime encompasses fire intensity, frequency, size, type, severity, and seasonality. While we can characterize a fire regime, it is a statistical combination of individual fire events occurring within a given region and time frame and thus does not necessary represent a current situation in much the same way an intense individual storm day or current weather conditions may not be in sync with an area's climatology (Pyne, 2007). Canada covers a vast and diverse area making it illogical to manage the forests and forecast lightning for the country as a whole. Focusing on Alberta allows consideration of the unique conditions to be taken into account. Site specific conditions are beginning to be taken into account and utilized in Alberta fire management. For example, prescribed burns and natural fire breaks such as marsh lands are being identified and strategically used to aid in decision making allowing for better allocation of resources.

The Wildland-Urban Interface describes the area where humans and human developments meet, mix or intermingle with forest or other types of vegetative fuel (McFarlane, 2006). Population growth and demographic changes have resulted in changes to settlement patterns. The heartlands are becoming more populated and remote areas are becoming increasingly occupied. Resource extraction has resulted in road networks allowing easier access to previously unreachable places. This has both benefited and negatively impacted fire management by providing access to otherwise remote lightning fire locations and increasing the number of human caused remote fires. More subdivisions, towns, and cities are popping up in forest dominated areas of the northern boreal as industry booms and more jobs are created. Recreational properties and vacation developments further contribute to this population shift. As the WUI increases, the chance of wildland fires adversely affecting communities rises.

Fires in the WUI pose a disproportionately large risk to life, health, property and infrastructure when compared to the small size of area burned (McFarlane, 2006). The cost of suppression in the WUI can be 10 times greater than that of wildland fire suppression and often present increased risks to firefighters (McFarlane, 2006). Many initiatives have been undertaken at the community, regional, provincial, and national levels to reduce the impacts of WUI fire disasters. Programs, such as FireSmart® (Vicars, 1999), aim to educate individuals, industry, and communities about protecting values at risk through proper landscaping, use of building materials, and emergency planning (Taylor et al., 2006). FireSmart® largely relies on communities and homeowners for voluntarily implementation (Taylor et al., 2006). Devastating fires in the WUI, such as the Kelowna fires in 2003, and Slave Lake fires in 2011, have resulted in an increased awareness of the risks associated with living in fire prone areas. Despite these disasters and attempts at increasing public awareness, people continue to build further into the forests and fail to take the necessary precautions to make their dwellings less susceptible to wildfire.

Many wildland fires perform a suite of essential ecosystem services. They clean out dead debris, expose soils to solar warming and drying, free up nutrients, promote species richness and biodiversity (depending on the severity and post fire conditions), sustain grasslands, allow regeneration of stands, and promote heterogeneity on the landscape among many other things (Latham and Williams, 2001). While people may view fire as a negative disturbance, the ecosystems that have evolved with fire require it to survive and thrive. Understanding ecosystem dynamics, wildland fire history, and

natural ecosystem processes is necessary for proper management of the boreal forest (Weber and Flannigan, 1997). This includes understanding driving factors such as climate and lightning.

There is an intimate relationship between climate and ecosystem dynamics; therefore understanding climate, and climate change, is paramount to properly managing forests and preparing for fires (Weber and Flannigan, 1997). Fire is highly affected by fuel moisture. Fuel moisture is directly linked to temperature, precipitation, wind speed and relative humidity; thus fire is highly sensitive to changes in weather and climate (Van Wagner, 1974; Weber and Flannigan, 1997). Contemporary management and suppression techniques have difficulties keeping up with extreme fire events and protection of values at risk. Extreme fire-weather conditions, and thus events, are expected to increase in both number and frequency in the near future, making this problem increasingly vital as larger, more severe, intense, and numerous fires occur (Flannigan et al., 2005; Weber and Flannigan, 1997).

1.7 LIGHTNING-FIRE INTERACTIONS

While contemporary Canadian studies show lightning-caused fire is less common (~35%) than human caused (>60%) in most regions, lightning-caused fire are responsible for roughly 80% of the total annual area burned in Canada (Stocks et al., 2002) and 93% of land burned in Alaska (Boles and Verbyla, 2000). Lightning strikes are also reported to start ~80% of wildland fires in the sparsely populated North West Territories (Kochtubajda et al., 2006), and are recognized as the main cause of wildland fires in the western United States forested regions (Rorig and Ferguson, 1999). Some argue that contemporary statistics on lightning-caused fires could well be doubled in order to capture a picture of the magnitude of historical fires and their massive extent (Pyne, 2007). Existing studies show area burned by wildfires is highly variable in Canada and can range from 0.7million to 7.6 million ha per year with an average of 2.5 million ha per year burned (Natural Resources Canada, 2004).

Lightning fires are responsible for such a vast area of burned forest due to two main factors; location, and concentration of strikes. Lightning can occur in remote areas far from human activity. This makes detection and response to ignitions difficult and increases the chance of the fire escaping initial attack (Martell and Sun, 2008; Wierzchowski et al., 2002). Couple remoteness with the reality that lightning has a tendency to occur in large concentrations and at multiple locations simultaneously

(Martell and Sun, 2008), and it is evident that suppression of lightning fires is a difficult feat. In order to adequately plan and react to remote fires, fire management agencies need to know about the danger and risk of fire ignition.

Information regarding fire danger and probability of ignition can be found in the Canadian Forest Fire Danger Rating System (Canadian Forestry Service, 1987). The system provides guidelines about expected fire behavior in varying fuel types and topographies in the Canadian Forest Fire Behaviour Prediction subsystem as well as information about fuel moisture (Forestry Canada Fire Danger Group, 1992). The fuel moisture component is generated by the Canadian Fire Weather Index System (Van Wagner, 1987), a subsystem of the Canadian Forest Fire Danger Rating System. The Canadian Fire Weather Index assesses the potential for fire startup and spread by taking into account past and current weather. For the purpose of this research, human caused fires are not considered, therefore chance of ignition can be thought of as how much CG lightning and where. While lightning may also cause structural fires within city limits, these fires are outside the scope of this paper and are therefore not considered.

While fire danger is well covered in Canada, an accurate medium range lightning prediction model is not currently available. Despite the need, lightning is not included in the Canadian Weather Prediction Model (Burrows et al., 2005) resulting in an information gap that can lead to dire results. Lightning fires that occur in remote areas have drastically larger suppression costs associated with transportation and access (Wierzchowski et al., 2002) and a much greater chance of overrunning initial attack and becoming large fires >200ha (Stocks et al., 2002). Geographically speaking, lightning fires are more random and clustered than human caused fires as their ignition depends on conditions conducive to lightning, the electrical properties of the lightning, local fuel and moisture characteristics (Krawchuk et al., 2006), and whether or not the strikes make contact with something combustible (Pyne, 2007).

Intuitively, fuel flammability and characteristics play a key role in the chance of lightning ignition (Flannigan and Wotton, 1991; Nash and Johnson, 1996). Lightning-caused fires tend to form in clusters around dryer landscape patches which typically correspond with regions most prone to burning as they are subject to relatively dry lightning (Pyne, 2007). Lightning-caused ignition of forest fuels is highly dependent on many variables. Fuel type, moisture, density and structure, along with the characteristics of the lightning strike(s) all play into the likelihood of ignition (Flannigan and Wotton, 1991). Certain kinds of strikes have been identified as having greater potential to start

lightning-caused fires. These characteristics include positive polarity, long continuing current (LCC), and high multiplicity of strikes (Flannigan and Wotton, 1991; Wotton and Martell, 2005).

Strikes with positive polarity have a peak current 10 times greater than negative strikes (Rust, 1986) and a greater chance of having a return stroke lasting more than 40ms known as LCC (Uman, 1987). Long continuing current can transfer twice as much charge to ground compared to a flash without LCC (Uman, 1987). Return strokes with LCC are general accepted as a major cause of lightning-caused wildland fire ignitions (Anderson, 2002; Flannigan and Wotton, 1991; Fuquay et al., 1979; Fuquay et al., 1972). It is estimated that ~85% of positive and ~20% of negative strikes have LCC (Uman, 1987). Despite the fire danger associated with LCC, little is known about it as it has proven very difficult to distinguish and identify strokes with or without LCC (Uman, 1987). The presence or absence of a LCC component is currently not detected by operational lightning detection sensors. Other characteristics such as high strike multiplicity can also increase the chance of ignition (Flannigan and Wotton, 1991) as more than one stroke flows though the same channel striking the same spot on the ground (Rakov and Uman, 2003).

From a fire suppression/prevention perspective, it is fortunate that positive strikes only account for roughly five percent of recorded strikes in the summer months (Orville and Songster, 1987). The percentage of positive strikes is minimum in July and August (Orville and Huffines, 1999) but does increase after October and peaks with around 80% of all strikes in January and February having positive polarity (Orville and Songster, 1987; Orville and Huffines, 1999). Ignition, survival, and spread of fire depends partially on what and where the lightning strikes. If a strike with optimum fire starting characteristics occurs, then fuel flammability and characteristics become the determining factors as to whether or not a fire will ignite. Even though the greatest number of positive strikes, and thus strikes with LCC occur in January and February, the greatest number of lightning-caused fires occur during the warm months as conditions are not favorable to ignition and spread of fire during the cold, snowy, winter months.

Anderson (2002) describes the life cycle of lightning-caused ignition by looking at three stages of fire: ignition, survival and arrival. Ignition is most often the result of a strike with LCC striking the ground and igniting the duff layer. This can smoulder and survive (holdover) for several days (Anderson, 2002). If conditions are right, the smouldering fire can arrive and spread as flaming combustion (Anderson, 2002).

Lightning plays a large role in fire dynamics however it cannot be forgotten that fire can also alter lightning. Pyrocumulus clouds can form under extreme fire conditions and are driven by intense convective updrafts created by large fires (Latham, 1991). Pyrocumulus clouds have a higher proportion of positive CG strikes and create a of positive feedback of sorts where increasing intensity of a fire leads to increased strike occurrence and greater chance of additional lightning ignitions which can further expand the fire and increase its intensity.

The interactions between fire and lightning are complex and intertwined. When looking at ways to improve forest and forest fire management, lightning occurrence must be considered. Lightning is responsible for 80% of area burned in Canada and more than 70% of fires larger than 200ha (Stocks et al., 2002). Fire management costs in Canada range from CAD $500 million to $1 billion per year (Flannigan et al., 2009) and the value of timber lost can well exceed the value of current harvest in extreme fire years (Natural Resources Canada, 2004). Add to the cost of direct suppression an additional CAD $600 million to $1 billion deficit to the Canadian economy due to lost time and disruptions (Stocks et al., 2002) and it becomes clear that additional fire and fire related research is both warranted and necessary.

1.8 PREVIOUS WORKS

Previous studies on lightning, including occurrence prediction, have explored many of the fundamentals of lightning and its formation; however, the exact mechanisms of electrostatic formation are not completely understood (Wåhlin, 1986). Research on North American lightning climatology and its numerous impacts on human safety, utility services, travel, and ecological disturbance have advanced greatly since the introduction of the North American Lightning Detection Network (NALDN) in 1998. The NALDN is a combination of the contiguous United States of America's National Lightning Detection Network (NLDN) and the Canadian Lightning Detection Network (CLDN). Prior to the introduction of the CLDN, Canadian lightning research often involved small, short term studies performed on data from provincial, territorial, or private lightning detection networks (Burrows and Kochtubajda, 2010). Since 1998, large, long term studies have become possible as continuous Canada wide data becomes available.

Lightning is an atmospheric phenomena and as such it is recognized to be highly episodic (Burrows et al., 2002). Large scale predictions such as number of strikes per month or average number of strikes per day are difficult to make as a large fraction of the

yearly lightning can occur within only a few days of the year (Burrows and Kochtubajda, 2010). Although this is the case, there are some areas that are recognized to be "hot spots" or areas of high lightning activity. These areas are typically characterized by major changes in topography and substantial landforms. Lightning is recognized as a major contributor to the ignition of forest fires. Considerable research has been done to study how lightning influences fire regimes and to quantify its occurrence.

Many empirical studies on lightning strike characteristics (Berger, 1967; Burrows and Kochtubajda, 2010; Kochtubajda and Burrows, 2010; Orville et al., 2002) and lightning-caused fire occurrence (Fuquay et al., 1967; Fuquay et al., 1972; Rorig and Ferguson, 1999) have been performed and methods of predicting or classifying the chance of lightning-caused ignition have been proposed (Anderson, 2002; Fuquay et al., 1979). After the initiation of the NALDN in 1998, Orville et al. (2002) provided an overview of continent-wide CG lightning trends as well as some regional differences for three years of continuous lightning detection (1998-2000). Specifically, they looked at peak current (kA), multiplicity, and percentage of strikes with positive polarity. In this study Alberta was identified as a hotspot for maximum mean negative and positive multiplicity (Orville et al., 2002).

Burrows et al. (2002) also used the NALDN data and found that the greatest number of days per year with detected lightning events in Canada and the Northern United States of America occurred over the Alberta Foothills east of the Rocky Mountain continental divide. This suggests that Alberta is indeed a hotspot for lightning activity in Western Canada. Burrows and Kochtubajda (2010) and Kochtubajda and Burrows (2010) provide a fundamental look at the spatial and temporal characteristics of CG lightning in Canada for 10 consecutive years of lightning data. Their analysis explores flash density, occurrence, polarity, multiplicity, and first-stroke peak current (Burrows and Kochtubajda, 2010). This two-part paper provides a wealth of background knowledge and useful benchmarks regarding lightning. Other studies also work towards characterizing lightning in Canada at a provincial or regional scale (Clodman and Chisholm, 1996; Kochtubajda et al., 2002; Morissette and Gauthier, 2008) among others.

The peak lightning season in Canada typically runs from May to October, with July being the most active month in the west (Burrows et al., 2002). Wierzchowski et al. (2002) found that in Alberta, lightning fire occurrence is fairly constant from June to mid-August with a peak in areas burned occurring prior to July (Wierzchowski et al., 2002). The authors propositioned that this could be due to seasonal changes in foliar

moisture content of coniferous trees noting that moisture content decreases in the spring in this region (Wierzchowski et al., 2002). These conditions could lead to spring or early summer lightning fires progressing to crown fires and spreading rapidly, therefore leading to increased area burned when compared to fires started later in the season when foliar moisture content is higher (Wierzchowski et al., 2002). Looking to the future, global lightning activity is expected to increase with rising temperatures (Price and Rind, 1994). The increase in lightning activity and the predicted rate of lightning-caused fire starts, the projected area burned is also anticipated to increase for central eastern Alberta (Krawchuk et al., 2009) and North America as a whole (Price and Rind, 1994).

There exists a pattern in most of Canada's ecozones where lightning activity increases with increasing precipitation (Kochtubajda et al., 2013). Kochtubajda et al. (2013) examined the relationship between CG strikes and convective precipitation across Canada. They found that while a pattern did exist, using the relationship for predicting convective precipitation yielded great uncertainty, especially so in the western region. Steps have also been taken to try to classify lightning and convective days as "wet" or "dry" (Rorig and Ferguson, 1999) with the understanding that lightning-caused ignition risk is much greater when strikes land where significant precipitation has not been received (Rorig and Ferguson, 1999). Such strikes are known as "dry strikes". Rorig and Ferguson (1999) found that dry lightning days occur when there are low moisture levels and instability is high in the lower troposphere, while "wet" lightning days occur when there is variable lower atmospheric moisture and greater atmospheric instability.

Various lightning models have been proposed (Anderson, 1991; Burrows, 2002; Burrows, 2008; Burrows et al., 2005). Occurrence prediction models tend to predict thunderstorm occurrence, and/or mean or total flashes over long time periods (days to months) and over large (province to country wide) spatial scales (Burrows et al., 2005). Other models to quantify or predict the features of lightning and its numerous interactions with the physical world have also been proposed. For example, the computer program Lightning-Caused Fire Occurrence Prediction System (Anderson, 2002) predicts the probability of a lightning strike resulting in a detectable fire based on the physical parameters of the strike, weather and fuels (Anderson, 2002). While work has been done to describe the influence of topography, fuel and weather on lightning-caused fire starts (Wierzchowski et al., 2002), a provincial model to help predict lightning occurrence would greatly advance the cause.

Lightning occurrence models are typically based on categorical and regression tree analysis. Tree-structured regression has been used to make lightning flash probability predictions for 5° latitude by 5° longitude cells across Canada and the northern United States (Burrows et al., 2005). Using decision trees, Burrows (2008) was able to reduce the number of proposed lightning predictors from 189 to 10 for each of the four designated regions of Canada. Top predictors noted in the literature included: mean sea level pressure, Showalter index, Lifted index, geopotential height, and precipitable water (Anderson, 1991; Burrows, 2008). It should be noted that convective indexes were always included among the top predictors. All of the literature encountered to date supports the notions that lightning is episodic in nature and its characteristics vary greatly over time and space. While these studies are useful for finding long term trends and narrowing down the possible significant lightning predictors, a finer temporal and spatial resolution model is needed to further benefit wildland fire management and research communities.

1.9 OBJECTIVES

This study works towards providing lightning occurrence prediction model(s) to aid with wildland fire management in the province of Alberta, Canada. According to the Government of Alberta, forests cover 60% of the province and the forestry industry contributes CAD nine billion dollars to Alberta's economy (Government of Alberta, 2012). Alberta's forests hold great tourist, natural, recreational, cultural, economic, and biodiversity value. While excluding fire from Alberta's landscape is neither economical nor environmentally responsible, the immense values of our forests makes it critical to know where and when fires occur, so responsible decisions can be made to preserve natural ecosystems (including the role of natural disturbances) and insure valued areas can be protected. Forest fire management in Alberta is carried out by Alberta Environment and Sustainable Resource Development as well as Parks Canada. An accurate, medium range lightning prediction model could greatly aid the efforts to responsibly manage Alberta's forests.

The objective of this study was to build models to accurately predict lightning occurrence in Alberta. Ideally these models will require only 12 or 15 input variables. In order to accomplish this, 29 general classes of predictors, and their variations, have been selected based on previous research. These predictors were used to generate diagnostic models to forecast lightning occurrence under a certain set of conditions via random

forest classification. Various ensemble forecasts were then generated. The ensemble model outputs predict daily and 6-hour lightning occurrence in each cell for three different spatial scales: (1) 2.5° latitude by 2.5° longitude, (2) 1.25° latitude by 2.5° longitude, and (3) 50km by 50km. The addition of lightning prediction models to the field of wildland fire science will increase knowledge of lightning ignitions, better fire occurrence models, improve resource allocation, and help increase preparedness of fire management agencies and communities alike.

CHAPTER 2. DATA AND METHODS

2.1 STUDY AREA

Located in the northwestern hemisphere, Alberta is the western most prairie province in Canada (Figure 3). Spanning from the Canada-United States border at the 49th parallel to the Northwest Territories at 60˚, the province reaches west from a shared border with Saskatchewan at the 110th meridian west to 120° at the northwest corner. The western border separating Alberta and British Columbia extends to the south from 60° north and 120° west to the Rocky Mountain Continental Divide, at which point it then follows the Continental Divide southward to the 49th parallel.

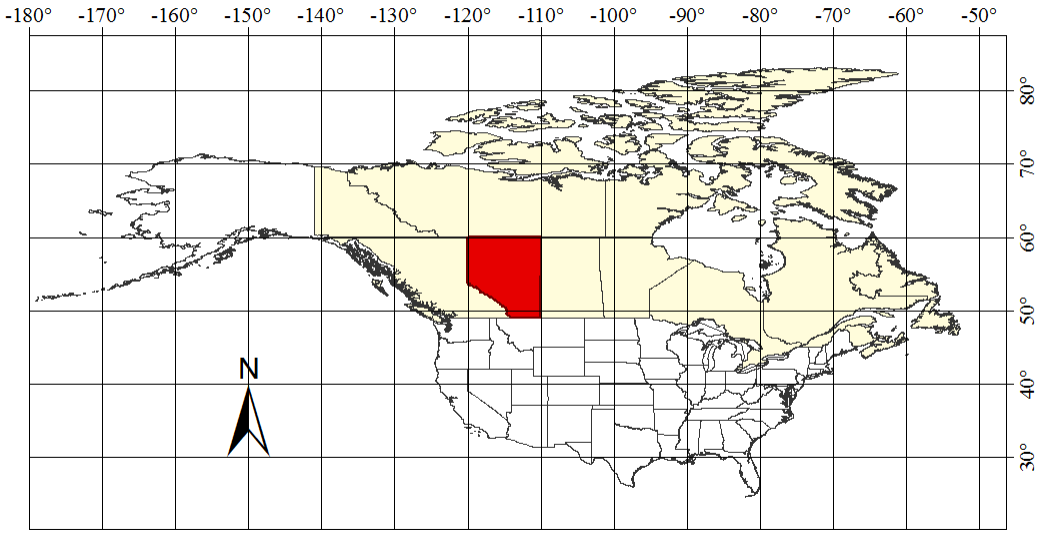

Figure 3: Located in in North America, the study area, highlighted in red, covers the province of Alberta, Canada.

Spanning an area of 661,848 km^2 (Statistics Canada, 2005), Alberta has a population of 3,645,257 people (Statistics Canada, 2013). Alberta is a geographically diverse province (Figure 4) with six major Natural Regions (Natural Regions Committee, 2006). The northern half of the province is primarily covered by the Boreal Forest with a small portion of the Canadian Shield reaching into the northeastern corner. Much of the southern province can be classified as Parkland and Grassland. The Rocky Mountains peak along the southwestern border and gradually lower as you move east into the leeward Foothills. The Foothills transition to Boreal Forest in the north, a central aspen Parkland, and to the Grasslands of the southeast. The Grasslands are scored by the unique

deep canyons of the Alberta badlands; long ago created and continually transformed by the Red Deer River.

Alberta has a dry continental climate with four distinct seasons. Precipitation ranges from 550 to 600mm/year in the Foothills, to 300mm in the southeast, and 400 to 450mm in the north (Stamp 2009). The growing season has a latitudinal gradient lasting around 120 days in the south and 60 days in the north (Stamp, 2009). With warm mild summers and cold snowy winters, Alberta's typical summer daytime highs of 20 to 25°C greatly contrast the normal winter night-time lows of -15 to -25°C; however temperatures can often climb above the mid to upper 30s and drop below -40°C (Government of Alberta, 2013). Lightning occurs year round in Alberta with the majority of strikes taking place during the warm summer months (Burrows and Kochtubajda, 2010). The Foothills and Swan Hills are lightning hotspots experiencing on average more than 30 days per year of CG lightning (Burrows and Kochtubajda, 2010).

2.2 DATA

Thirteen years of data from 1999 to 2011 were collected to perform this study. As implied by the objectives, the predictand is a binary response variable indicating the presence or absence of CG lightning. There are two main groups of predictors: geographic and temporal covariates, and weather data. The geographic and temporal covariates include location and time specific inputs for each data point. Along with the weather data, these two groups make up the input variables used to predict lightning occurrence. Weather data was obtained from four separate sources. NCEP-DOE[2] Reanalysis II Pressure Level data, NCEP-DOE Reanalysis II Surface Data, NCEP/NCAR[3] Reanalysis I Pressure Level data, and Radiosonde parameters and indexes were compiled and evaluated to capture a clear picture of atmospheric conditions and their relationship to CG lightning activity. Some additional variables of interest were calculated from components of the Reanalysis I and Reanalysis II data. A brief description and background of each data set and variable is provided in the following sections and a list of all initial predictors is presented in Table 1. Modifications and data

[2] National Centers for Environmental Protection (NCEP), Department of Environment (DOE)
[3] National Center for Atmospheric Research (NCAR)

quality control measures implemented are discussed in section **2.3 DATA PROCESSING AND MODIFICATIONS.**

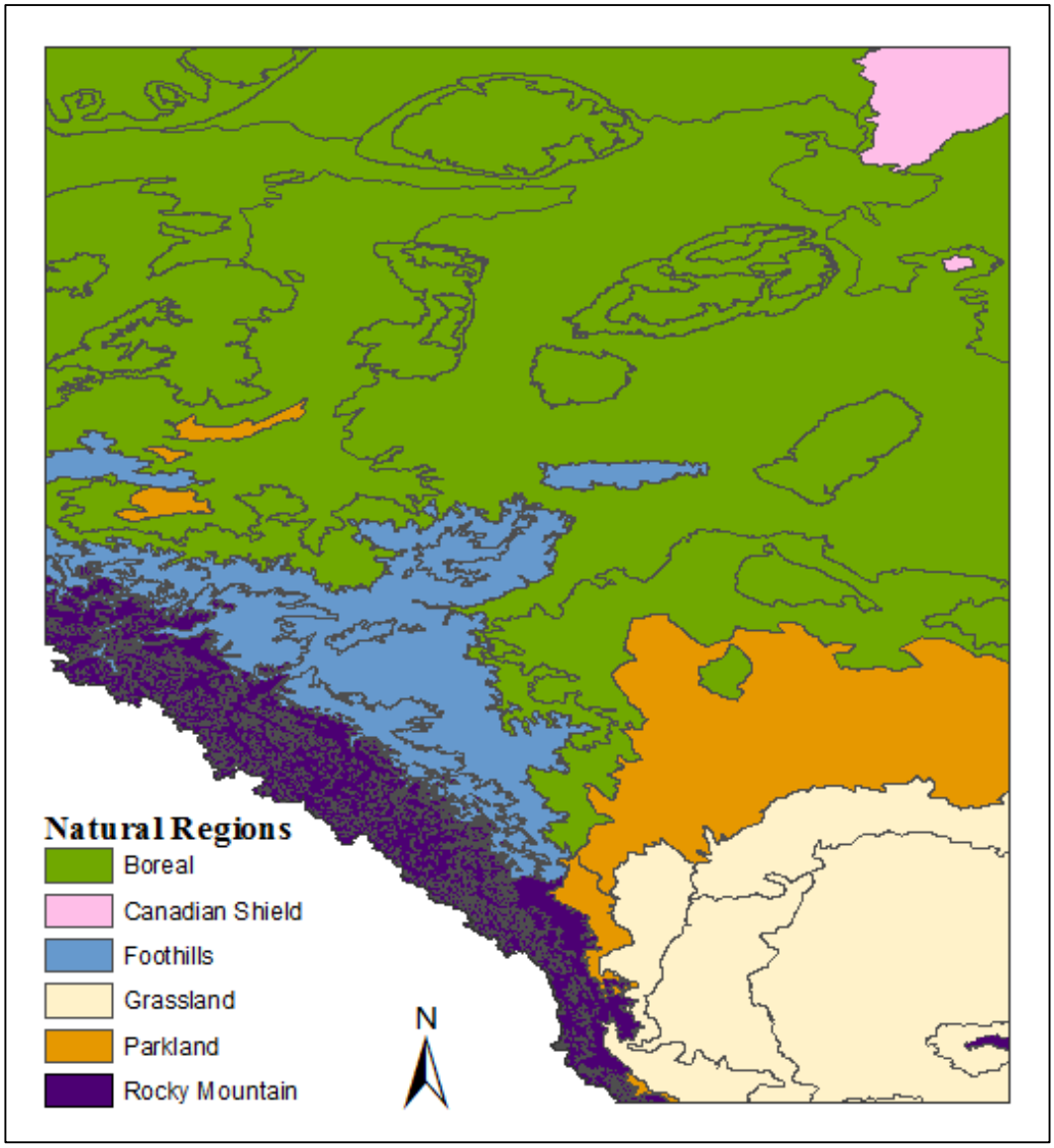

Figure 4: There are six Nation Regions in the province of Alberta separating the province into regions with similar terrain and ecology.

Table 1: List of input variables. Abbreviations, full names, units of measure, and some additional information about data sources and resolution are given for all of the predictors initially considered.

	Predictor Name	Units	Source and Specifications
day	Julian day		
lat	Latitude	Degrees North (°)	
long	Longitude	Degrees West (°)	
time	Time of day	Coordinated Universal Time (Z)	0000, 0600, 1200 and 1800
elv	Elevation[4]	Meters (m)	Digital Elevation Model[5]
Temp	Air temperature	Kelvin (K)	**NCEP-DOE Reanalysis II: Pressure Level** 2.5° Latitude by 2.5° Longitude Grid 17 Pressure Levels (1000 to 10mb)
gpz	Geopotential Height	Meters (m)	
RH	Relative Humidity	Percent (%)	
omega	Omega: Vertical Winds	Meters per Second (m/s)	
U	U Wind: North-South		
V	V Wind: East-West		
PW	Precipitable Water	Kilograms per Square Meter (kg/m^2)	**NCEP-DOE Reanalysis II: Surface Grid** 2.5° Latitude by 2.5° Longitude Grid
sfc_p	Surface Pressure	Pascal (Pa)	
mslp	Mean Sea Level Pressure		
spec_hum	Specific Humidity	Kilograms per Kilogram (kg/kg)	**NCEP/NCAR Reanalysis I: Pressure Level** 2.5° Latitude by 2.5° Longitude Grid Eight Pressure Levels (1000 to 300mb)
CAPE	Convective Available Potential Energy	Joules per Kilogram (J/kg)	**Radiosonde Observations** Fort Smith (YSM), Fort Nelson (YYE), Kelowna (WLW), Edmonton Stony Plain (WSE), Prince George (ZXS), The Pas (YQD), Glasgow (GGW), and Great Falls (TFX)
CINS	Convective Inhibition		
EQLV	Equilibrium Level	Millibars (mb)	
KINX	K Index		
LIFT	Lifted Index		
LCLP	Lifted Condensation Level Pressure	Millibars (mb)	
LCLT	Lifted Condensation Level Temperature	Kelvin (K)	
PWAT	Precipitable Water[6]	Millimeters (mm)	
SWET	Severe Weather Threat Index		
SHOW	Showalter Index		
T_d	Dewpoint Temperature	Degrees Celsius (°C)	Calculated from Reanalysis I and Reanalysis II data
T.Td	Temperature-Dewpoint Spread		
haines	Haines Index		
Vap	Vapour Pressure	Millibars (mb)	

[4] Above mean sea level
[5] Digital Elevation Model (DEM) from Global GIS with 30 arc second spacing (~1km), Elevation accurate to ± 30m (http://webgis.wr.usgs.gov/globalgis/gtopo30/gtopo30.htm)
[6] For entire sounding

2.2.1 PREDICTAND

Yearly lightning flash data for a rectangular polygon encompassing the province of Alberta were obtained for the thirteen year period. The flash data, provided by Environment Canada, include the date (YYYY-MM-DD), Coordinated Universal Time (UTC) in Zulu (Z), latitude (decimal degrees), longitude (decimal degrees), event strength (kA), and cloud or ground status for each flash record. This data was collected and recorded by the Canadian Lightning Detection Network (CLDN), a sub-network of the North American Lightning Detection Network (NALDN). The NALDN provides continuous lightning data of most CG and some CC flashes for Canada and the contiguous United States. The network(s), operated by Vaisala Inc., detect and record lightning flashes year round to about 300km offshore, however the detection efficiency of the system degrades along the northern most reaches and near the periphery.

Locations of the NALDN detectors can be found in Orville et al. (2002) and a more up to date map of the CLDN sensors can be found in Burrows and Kochtubajda (2010). As noted in Burrows and Kochtubajda (2010), the CLDN has experienced many changes since its introduction. In February of 1999 the system became able to differentiate CG and CC discharges. The CLDN CG detection efficiency drops to roughly 70% near the periphery of the network area however it exceeds 80% to 90% in most regions and has a median location accuracy of 500m (Cummins and Murphy, 2009). Alberta's landlocked location places it well within the 70% detection efficiency line as shown in Burrows and Kochtubajda (2010) therefore no modifications were made to correct for differences in detection efficiency.

Additional sensors have been added over the years and starting in 2005 a number of sensors were upgraded from Lightning Positioning and Tracking Sensors (LPATS-IV), which detect and locate lightning via triangulation by measuring the arrival time of radio pulses produced by the flashes, to CG Enhanced Lightning Sensors (LS7000), which combine *time-of-arrival* technology with *magnetic direction finding*. The newer devices only require a strike to be detected by two sensors while the old devices triangulation method requires detection by three sensors. Cummins and Murphy (2009) provide a concise overview on the history of lightning detection, changes to detection technology, and how many of the different sensors work. Information pertaining to field waveforms produced by lightning and specifics on the electromagnetic energies emitted by different

stroke components can be found in Cummins and Murphy (2009), Krider et al. (1980), and Weidman and Krider (1978, 1979) among others.

2.2.2 PREDICTORS

Geographic and Temporal Covariates

Lightning activity is influenced by both geographic and temporal factors. Elevation, slope and aspect, proximity to large water bodies, and length of the warm season are all related to localized rates of lightning occurrence (Burrows and Kochtubajda, 2010). Studies of Canadian lightning characteristics have also shown diurnal and seasonal trends. We typically see maximum activity in the late afternoon and minimal activity in the early morning hours with the greatest number of strikes occurring during the month of July (Burrows et al., 2002; Burrows et al., 2005). In order to try to capture these characteristics, latitude, longitude, elevation, Julian day, and time of day were all included as input variables.

Elevation and longitude provide a gradient for slope, aspect, and height from the Rocky Mountains and leeward Foothills in the west to the relatively flat prairie lands in the east. Latitude and Julian day help capture the length of the warm season as well as typical seasonal conditions. Julian day also captures the seasonal trend of peak lightning occurrence in July. The diurnal variation is represented by time of day as a predictor. These descriptive variables represent the time and geographic characteristic for a specific data point. The geographic covariates represent the physical location and elevation at the center point of a particular spatial bin, while the temporal covariates represent the time frame under consideration. The elevation for each center point was extracted from a Digital Elevation Model (DEM) with a 30 arc second (~one km) spacing and an elevation accuracy of ±30m.

Reanalysis II : Pressure Levels

The first set of weather related variables were provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, and obtained from their web site at http://www.esrl.noaa.gov/psd/. Reanalysis II pressure level (Kanamitsu et al., 2002) data were downloaded for the thirteen year period. Daily mean and four-times daily observations of air temperature, geopotential height (gpz), relative humidity (RH),

vertical (omega) wind speed, north-south (U) wind speed, and east-west (V) wind speed data sets were downloaded for the 17 available pressure levels. This pressure level data provides a vertical profile of the atmosphere with observations at the 1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, 70, 50, 30, 20, and 10mb levels at a 2.5° latitude by 2.5° longitude global grid.

Since more lightning strikes occur during the warm season and during the warm afternoon hours, air temperature may play an important role in lightning prediction. Geopotential height (gpz) is an approximation of the actual height in meters of the pressure level above mean sea level. The gpz contours of various pressure levels including 850, 700, 500, and 300mb, are often analyzed to identify atmospheric ridges and troughs. Troughs can often be found where convection is occurring (low pressure often associated with fronts) and often bring clouds, precipitation or cold air masses (Martell, 2001). Ridges are regions with higher gpz where air is sinking or a warm air mass is present (Martell, 2001).

Ridges often bring warm drier air. Relative humidity (RH) is the ratio of the partial pressure of water vapour to the saturated vapour pressure of water at a given temperature. Relative humidity is relative to the temperature; warm air can hold more water vapour than cooler air therefore the same amount of vapour present in both parcels will produce two different RH values. When RH reaches 100% there is a chance of precipitation if the air is rising at a sufficient rate. There were some data quality issues found in the RH data set (measurement above 100% and below 0%) therefore another humidity metric, specific humidity was also obtained from Reanalysis I: Pressure Level. The vertical and horizontal winds provide insight into the wind shear and how the air is moving at various levels of the atmosphere. Due to the low resolution of the Reanalysis data there is little chance of updrafts or downdrafts being captured. The winds were initially included to see if they could play a role in lightning prediction at this coarse scale.

Reanalysis II : Surface Grid

NCEP-DOE Reanalysis II surface grids provides surface and entire atmosphere weather data at a 2.5 degree latitude by 2.5 degree longitude grid. Daily mean values and four-times daily observations of surface pressure, mean sea level pressure (mslp), and precipitable water *(*entire atmospheric column) were downloaded for the thirteen year period. A falling surface pressure is often associated with increased risk of convective

storm occurrence while consistent pressure can be associated with a stable atmosphere. Precipitable water is a measure of the total atmospheric water vapour (mm) within a vertical column cross-sectional area between the Earth's surface and the top of the atmosphere. Another measure of precipitable water was also obtained from the Radiosonde data.

Reanalysis I : Pressure Levels

A final variable was obtained from NCEP/NCAR Reanalysis I pressure level *(*Kalnay et al., 1996*)* from the website http://www.esrl.noaa.gov/psd/. Specific humidity (kg/kg) is approximately equal to the mixing ratio, ratio of the mass of water vapour in an air parcel to the mass of dry air for the same parcel. Specific humidity was included since unlike relative humidity, it does not vary as the temperature or pressure of an air parcel changes making it useful for calculating additional variables. In addition, the RH data had some quality issues therefore including specific humidity seemed necessary.

Radiosonde Observations

Knowing the physical characteristics of the upper atmosphere is crucial for research, aviation navigation, and weather forecasts including thunderstorm prediction. A "sounding" refers to the process by which observations of temperature, pressure, humidity, and wind speed and direction are made for a vertical column of the atmosphere. There are more than 800 Radiosonde Observations (RAOB) stations worldwide with over 120 stations located in North America and the Pacific islands (NOAA National Weather Service, 2013). Environment Canada operates 31 of these stations across Canada (Environment Canada, 2013), while the United States National Weather Service Upper-air Observations Program operates an additional 92 RAOB stations across the United States and Pacific islands (National Weather Service, 2009). These stations are the launch locations for Radiosondes, small devices suspended below large hydrogen or helium gas balloons.

When launched, the balloons rise at around 300m per minute. During its ascent, the sensors in the Radiosonde take measurements of surrounding temperature, pressure, and humidity, providing an atmospheric profile up to roughly 30km altitude (Dabberdt et al., 2003). Wind speed and direction are recorded by GPS trackers attached to the devices. These observations are then used to compute various upper air indexes and

parameters which are then broadcasted to various organizations and made readily available to the public. Launches occur two times per day (00Z and 12Z) unless severe weather is anticipated, in which case additional launches may occur. Upper air observations from 1973 onward can be accessed online from the University of Wyoming at http://weather.uwyo.edu/upperair/sounding.html.

Eight RAOB stations were chosen in an attempt to provide coverage of the entire province. Only a single station is located within Alberta, station 71119: Edmonton Stony Plain (WSE). The remaining seven stations surround the province. Station 71934: Fort Smith (YSM) is located near the Alberta border in the Northwest Territories, 71945: Fort Nelson (YYE) in northeastern British Columbia, 71203: Kelowna (WLW) in the southern interior of British Columbia, 71908: Prince George (ZXS) roughly between Fort Nelson and Kelowna, 71867: The Pas (YQD) in Manitoba, 72768: Glasgow (GGW) Montana, and 72776: Great Falls (TFX) Montana. Figure 5 shows the locations of the sounding stations used.

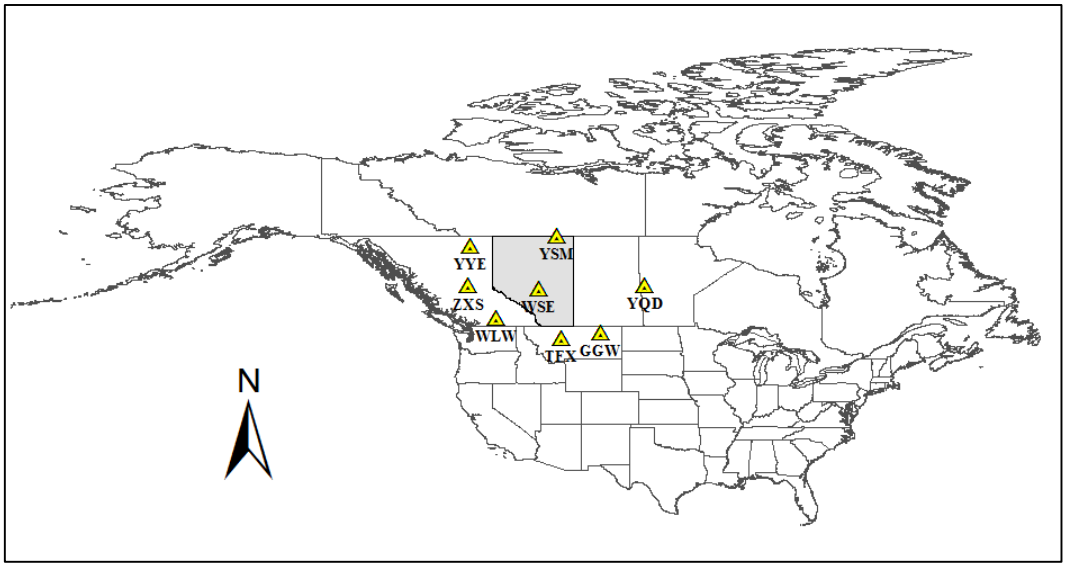

Figure 5: The location of the eight RAOB sounding locations used to provide information about the upper atmosphere conditions over Alberta. General locations are marked by yellow triangle with the corresponding station call names below.

Based on previous literature, including: Burrows (2002, 2008), and Burrows et al. (2005); some of the parameters and indexes available from the sounding data were chosen. The variables chosen are: Convective Available Potential Energy (CAPE), Convective Inhibition (CINS), Equilibrium Level (EQLV), George's K Index (KINX), Lifted Index (LIFT), Lifting Condensation Level Pressure (LCLP), Lifting Condensation Level Temperature (LCLT), Precipitable Water (PWAT), Severe Weather Threat Index

(SWET), and Showalter Index (SHOW). A short review of each Radiosonde variable is provided. For the following equations notation is as follows unless otherwise stated: Temperature in degrees Celsius denoted by T, T_d represents the dewpoint temperature also in degrees Celsius, and numbered subscripts indicate the atmospheric level (mb) at which the variable refers. Acceleration due to gravity is indicated by ***g.***

Convective Available Potential Energy (CAPE): An indicator of atmospheric instability, CAPE is a numerical measure of the amount of energy, in J/kg, available to a parcel of air if lifted through the atmosphere. The positive buoyancy of the parcel can be found by calculating the area on a thermodynamic diagram (Skew-T log-P) between the height of the level of free convection, $\boldsymbol{z_{EQ}}$, and the equilibrium level height, $\boldsymbol{z_{LFC}}$, where the environment temperature profile is cooler than the parcel temperature. CAPE can also be calculated using the virtual temperature (abbreviated as CAPV on soundings) however that is not the case for this study. CAPE values less than 1,000J/kg indicate a relatively stable atmosphere while values in excess of 2,000J/kg and 3,000J/kg indicate sufficient energy for thunderstorm and severe thunderstorms respectively.

$$CAPE = \int_{Z_{LFC}}^{Z_{EQ}} g\left(\frac{T_{parcel} - T_{env}}{T_{env}}\right) dz$$

Convective Inhibition (CINS): The opposite of CAPE, CINS is the amount of energy in J/kg needed to overcome the negative buoyant energy exerted by the environment on an air parcel. Like CAPE, CINS can be determined from a Skew-T log-P diagram by taking the area between the cooler parcel temperature and warmer environment temperature profile beginning at the surface, $\boldsymbol{Z_{sfc}}$, and going up to the level of free convection, $\boldsymbol{Z_{LFC}}$. CINS can also be calculated using the virtual temperature (abbreviates as CINV on soundings) however that is not the case for this study. Given as a negative value, CINS of zero to -50J/kg represents a weak capping effect, -50 to -200J/kg represent a moderate cap, and less than -200J/kg represents a strong capping force and thus likely a stable atmosphere.

$$CINS = \int_{Z_{sfc}}^{Z_{LFC}} g\left(\frac{T_{parcel} - T_{env}}{T_{env}}\right) dz$$

Equilibrium Level (EQLV): The height of the level of neutral buoyancy where a parcel is no longer buoyant. This level is often near the tropopause. Due to numerous missing values in the data set, Equilibrium Level was removed from the list of variables prior to analysis.

Georges K-Index (KINX): Taking the lapse rate, dewpoint temperature at 850mb, and the 700mb temperature-dewpoint spread into account, KINX provides an estimate of

the likelihood of thunderstorms. A value of less than 20 indicates zero likelihood, but as the value increases to 20 or greater the chance of precipitation and thunderstorms is expected to increase. A value greater than 35 indicates a good chance of numerous thunderstorms (George, 1960).

$$K = (T_{850} - T_{500}) + T_{d\,850} - (T_{700} - T_{d\,700})$$

Lifted Index (LIFT): While the formula is the same as that for SHOW, when calculating LIFT the temperature of the parcel (***T*****$_{parcel}$**) is the 500mb temperature of a lifted parcel with average pressure, temperature, and dewpoint of the layer 500m above the surface (Morales et al., 2007). As LIFT decreases, the atmosphere becomes more unstable. A LIFT of 10 or more indicates stable weather. As LIFT falls below zero thunderstorms become possible, while values less than or equal to -4 indicate severe thunderstorm potential.

$$LIFT = (T_{500} - T_{parcel})$$

Lifted Condensation Level Pressure (LCLP): The pressure (mb) level at which a parcel of air lifted[7] from the surface dry-adiabatically would become saturated. The lifted condensation level can often be observed as the cloud base and is easily found on a Skew-T log-P by lifting a near surface temperature and dewpoint value. Where ***T*** is in Kelvin and KAPPA is Poisson's constant, 2/7.

$$LCLP = p * (\frac{LCLT}{T}))\text{^}(1/KAPPA)$$

Lifted Condensation Level Temperature (LCLT): Temperature in Kelvin at the lifted condensation level (LCL). The LCL is the level at which a parcel of air lifted from the surface dry-adiabatically would become saturated. Note: for this equation temperate and dewpoint are in Kelvin.

$$LCLT = [1/(\frac{1}{(T_d - 56)} + \frac{\ln(\frac{T}{Td})}{800})] + 56$$

Precipitable Water (PWAT): Total atmospheric water vapour (mm) within a vertical column cross-sectional area between the Earth's surface and the top of the atmosphere. Provides an idea of how much precipitation could fall as the result of a low pressure system or storm.

---

[7] Unless otherwise stated, when a parcel is lifted it is done so dry-adiabatically until saturated at which point if it continues to lift it will do so moist-adiabatically.

Severe Weather Threat Index (SWET): More commonly referred to as SWEAT, the severe weather threat index (Williams et al., 2008) is used to analyze thunderstorm potential. Values greater than 300 indicate an increased risk of severe thunderstorm, while values of 400 and greater represent considerable risk of tornadoes. In the formula, ***TT*** is the Total Totals Index (if less than 49, set to zero), ***V*** refers to wind speed in knots and $\boldsymbol{\Delta V_{500\text{-}850}}$ is the change in wind direction (degrees) between the 500mb and 850mb levels.

$$SWET = 12T_{d\,850} + 20(TT + 49) + 2V_{850} + V_{500} + 125[SIN(\Delta V_{500-850}) + 0.2]$$

Showalter Index (SHOW): Another atmospheric instability index that evaluates the potential for convective storm activity. The Showalter (1947) Index is the difference between the 500mb environmental temperature, $\mathbf{T_{500}}$, and the temperature of a parcel lifted from 850mb to 500mb, $\mathbf{T_{parcel}}$. As the Showalter Index decreases below zero the chance of convective activity, including precipitation and thunderstorms, is expected to increase.

$$SHOW = (T_{500} - T_{parcel})$$

Calculated Variables

Additional variables were calculated to supplement those obtained from the previous data sources. Dewpoint temperature, vapour pressure (mb), temperature-dewpoint spread, and Haines index (Haines, 1988) were calculated from the Reanalysis I and Reanalysis II data. In order to obtain the dewpoint temperature (°C), vapour pressure (mb) was calculated from the specific humidity obtained from Reanalysis I. The following two formulas (Bolton, 1980) for specific humidity (kg/kg), ***q***, and vapour pressure (mb), ***e***, were modified in order to calculate the dewpoint temperature. ***p***=surface pressure in mb.

$$q = \frac{0.622 * e}{p - (0.378 * e)}$$

$$e = 6.112 * \exp(\frac{17.67 * T_d}{T_d + 243.5})$$

By solving for vapour pressure, ***e***, the dewpoint temperature can then be calculated.

$$e = p * \frac{q}{0.622 + 0.378 * q}$$

$$T_d = (243.5 * \log(\frac{e}{6.112}))/(17.67 - \log\left(\frac{e}{6.112}\right))$$

Vapour pressure and dewpoint temperature were calculated for the 1,000mb level, 850mb level, and 700mb levels. Quality controls were set in the codes preventing a dewpoint temperature from being greater than the corresponding air temperature. If such an event

occurred then the dewpoint temperature was set equal to the air temperature. The temperature dewpoint spread, a simple subtraction of dewpoint temperature from air temperature, was then calculated for the three levels.

The Haines Index (Haines, 1988) is a lower atmosphere stability index often used to help understand and describe fire weather. Although typically used as an indication of the potential of wildfire growth and risk of extreme fire behaviour, this index was included due its representation of moisture and stability in the lower atmosphere. Wildland fire organizations typically calculate the Haines index from the 12Z morning sounding in North America. Ranging from values of two to six. A Haines index of two represents a moist and stable lower atmosphere with very low fire growth potential while a value of six indicated a dry unstable lower atmosphere with a high risk of fire growth and extreme fire behaviour (Haines, 1988).

The Haines Index can be calculated for low (950 to 850mb), mid (850 to 700mb) and high (700 to 500mb) elevations. These variables were calculated from the Reanalysis II data which does not have a value for the 950mb level therefore the low elevation Haines Index was not calculated. Due to the general elevation of the province (excluding the mountainous region) the mid-level Haines index was considered sufficient. Calculated in a series of steps based on various thresholds the mid-level Haines Index was calculated as follows:

The Stability Term is based on the difference between the 850mb and 700mb air temperatures,

$$T_{diff} = T_{850} - T_{700}$$

where if $\boldsymbol{T_{diff}} \leq 5°$, Stability Term = 1

$6° \leq \boldsymbol{T_{diff}} \leq 10°$, Stability Term = 2

$\boldsymbol{T_{diff}} \geq 11°$, Stability Term =3.

The Moisture Term represents the temperature dewpoint spread at the 850mb level,

$$T_{spread} = T_{850} - T_{d\,850}$$

where if $\boldsymbol{T_{spread}} \leq 5°$, Moisture Term = 1

$6° \leq \boldsymbol{T_{spread}} \leq 12°$, Moisture Term = 2

$\boldsymbol{T_{spread}} \geq 13°$, Moisture Term =3

The Haines index can then be calculated as the sum of the Stability Term and Moisture Term. Finally, 24-hour change, was also calculated for each of the existing Reanalysis variables and each of the newly calculated variables.

2.3 DATA PROCESSING AND MODIFICATIONS

From a wildland fire management perspective, the primary interest and application of lightning predictions models lies in accurately predicting lightning events in remote forested areas. Nonetheless, lightning prediction models were created for the entire province. It is believed the benefits of lightning prediction models reach far beyond the proposed fire management objectives of improving resource allocation and preparedness. Models were created for two time frames (daily and 6-hour) and three different spatial scales (Figure 6): 2.5 ° latitude by 2.5° longitude grid, 1.25° latitude by 2.5° longitude grid, and a 50km by 50km grid. The province as a whole was considered for the first two spatial scales. For the third scale, 50km by 50km, the province was subdivided into three separate zones based on the Natural Regions of Alberta (Natural Regions Committee, 2006). Figure 7 shows the three zones which separate the province into large geographically and ecologically similar regions. Daily and four-times daily (6-hour) lightning predictions models were developed for each of the spatial scales. No predictions were made for the Canadian Shield Region as it is a small (9,719km^2), sparsely vegetated area of transition between the tundra and forest and therefore is not of particular concern for wildland fire occurrence. In addition, even at the 50km scale there would only be a couple of points within this area.

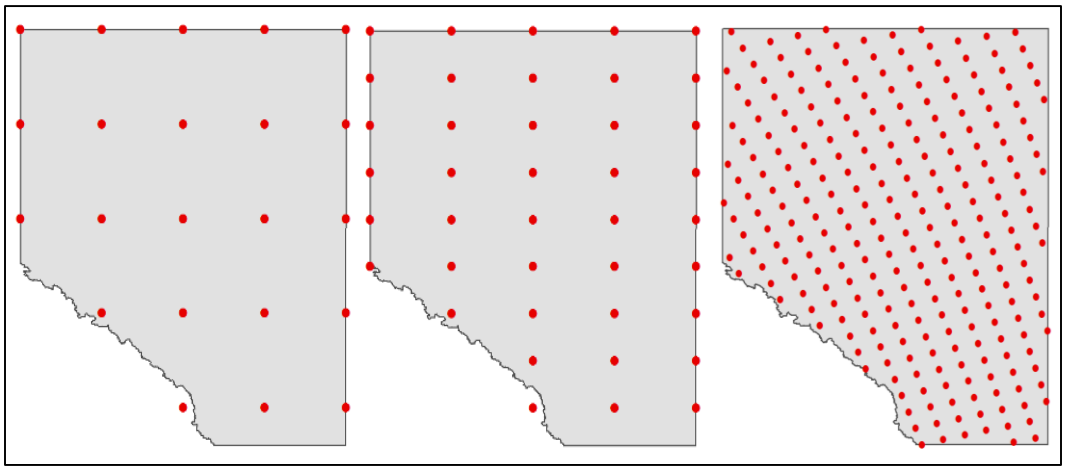

Figure 6: Lightning prediction models were created for the province of Alberta at three different spatial scales. From left to right: 2.5 ° latitude by 2.5° longitude grid, 1.25° latitude by 2.5° longitude grid, and a 50km by 50km grid.

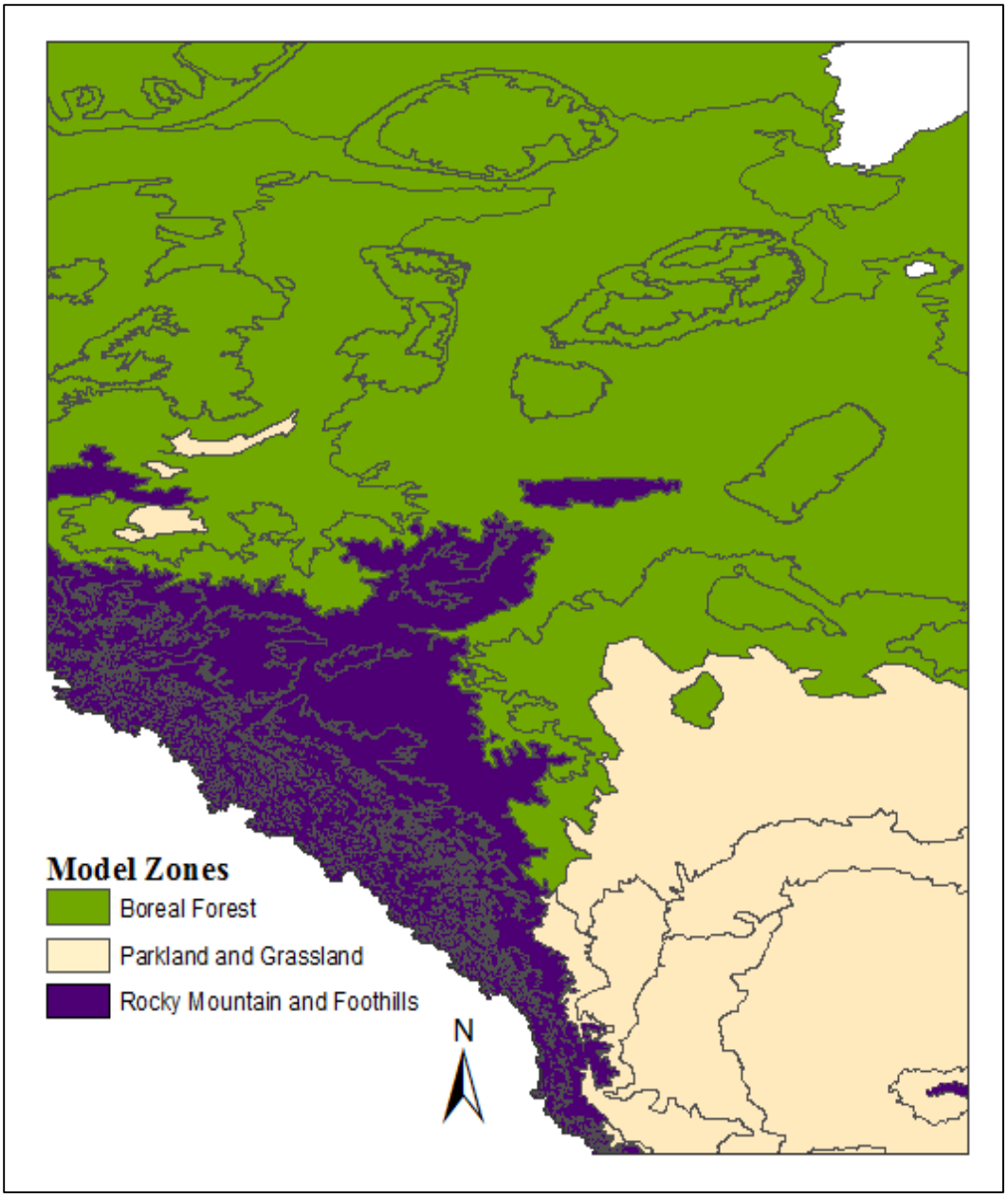

Figure 7: Three zones were created for the 50km by 50km spatial scale. These zones are based on the Natural Regions of Alberta and include the Boreal Forest Zone, Parkland and Grassland Zone, and the Rocky Mountain and Foothills Zone.

Before the prediction models were created, a suite of data processing, interpolation, quality control, and modifications were needed. This section walks through the methods used to process, generate, and modify the lightning and predictor data. A description of the interpolation methods and quality control measures applied to the Reanalysis and Radiosonde data is then provided along with an overview of the interpolation technique. Finally the methods for the modeling component are outlined

along with a basic background of the statistics and algorithms used. All of the data processing, quality control, analysis, and modeling was performed in the R programming environment (R Core Team, 2013).

2.3.1 LIGHTNING DATA PROCESSING

The objective of this study is to create models that predict warm season CG lightning occurrence for multiple spatial and temporal scales within the Province of Alberta. In order to model lightning occurrence, the raw lightning data needed to be processed and reorganized to meet the objectives. The CLDN data includes all lightning flashes recorded within a rectangular polygon encompassing the province of Alberta. Using the *maptools* (Bivard and Lewin-Koh, 2013) package and a shape file of Alberta, a clip was performed to remove all flash records outside of the Alberta boundary. The data set was also subset to only include lightning flash data from April 1st to October 31st for each of the thirteen years.

Quality control measures were then implemented on the remaining data. The same methods of reclassification used in Burrows and Kochtubajda (2010) were implemented to correct for misclassification of CG status. Any positive CG strikes recorded with a peak current strength less than 15kA were reclassified as CC flashes, and all positive flashes classified as CC with a strength greater than 20kA were reclassified as CG. Misclassifications occurred in less than 7% of the data. The majority of the 7% classification error were flashes incorrectly classified as CG (less than 15kA).

In contrast to Burrows et al. (2005), who aimed to predict all detected lightning, a decision was made to remove the CC strikes and only build models to predict CG strikes as these are the flashes of interest when predicting lightning from a wildland fire perspective. After the reclassification process, all CC strikes were removed from the data set. Once the data was subset to the proper spatial range, time frame, and lightning type, the data needed to be processed to generate a series of individual data sets with various spatial and temporal resolutions. Six distinct lightning data sets were needed. Each of the three spatial scales required two separate sets, one with the total daily occurrence, and the other with lightning strikes binned into four separate six-hour time bins for each day.

In order to create the different spatial grids, a series of center points were identified for each scale. A series of rectangular bins were then created surrounding each center point. Figure 6 shows the three spatial scales and the approximate location of the center points. The center points and corresponding spatial bins for the first two scales

were created with ArcGIS with a World Geodetic System 84 (WGS 84) projection and are measured out in degrees latitude and longitude. The Lambert Conformal Conic projection was used to generate the 50km by 50km grid. Any strikes falling within the boundary of particular bin were assigned a "POINT_ID" corresponding to the unique spatial bin. Four distinct time bins were also created.

The Reanalysis data is available in daily mean and four-times daily observations occurring at 00Z, 06Z, 12Z, and 18Z. The upper air sounding data are typically available for 00Z and 12Z observations. Four time bins were created for the lightning data based on the temporal resolution of the predictor data sets. All lightning strikes occurring within the given time frame were classified with corresponding "tbin" value:

- tbin = 1 from 21Z to 03Z (15:00 to 21:00 MST)
- tbin = 2 from 03Z to 09Z (21:00 to 03:00 MST)
- tbin = 3 from 09Z to 15Z (03:00 to 09:00 MST)
- tbin = 4 from 15Z to 21Z (09:00 to 15:00 MST)

Since the tbin=1 time frame spans over a two day period (from 9pm to 3am UTC), a one day adjustment is needed to properly place the six-hour consecutive time frame together. In order to do this a fifth temporary time bin was created. The fifth time bin was applied to all strikes occurring after 21:00:00Z (hh:mm:ss). This temporary time bin, tbin=5, allowed for one Julian day to be added to each strike occurring between 21Z and 24Z, effectively coupling these strikes with the following days 00Z to 03Z observations. The spatial bins and time bins created for the lightning data were also used to classify the predictor variables allowing the predictors and the predictand to be merged together into six master data sets.

2.3.2 REANALYSIS DATA PROCESSING

Each of the daily mean and four-times daily Reanalysis variables downloaded contained a year of observations for the entire globe. The variables were clipped to the spatial and temporal ranges as outlined in 2.3.1 LIGHTNING DATA PROCESSING. The basic variables from the Reanalysis II: Pressure Level data have measurements for 17 vertical levels. Since a unique variable variation is created for each of the six basic variables at each of the 17 levels at each time (4+1, four-times daily and daily mean), the number of variables stemming from the Reanalysis II: Pressure Level data alone was 510. In order to narrow down the amount of variables, some of the levels were removed prior

to processing. Removing even a few of the levels resulted in a significant decrease in data size and reduced the number of inputs for the models.

All levels above 250mb were removed effectively decreasing the number of layers from 17 to nine. Once some of the levels were removed, a data quality overview was performed where the maximum and minimum values for each variable were found. While almost all variables had their extreme values fall within an acceptable range, RH stood out as having some quality issues. Physically impossible values of greater than 100% and less than 0% were found throughout the data set. The less than 0% invalid entries appeared to propagate across time and space indicating a possible systematic upload error in the RH data set. A decision was made to replace any invalid data with the closest acceptable value and a note was made of each invalid data point.

The specific humidity data available through Reanalysis I: Pressure Level had only eight levels of data available versus the 17 available for the other pressure level data. The eight levels follow the same pressure level spacing but only extend up to the 300mb level. All eight levels were kept. The Reanalysis data came from the source as 2.5° latitude by 2.5° longitude gridded data. In order to interpolate the Reanalysis data for the 1.25° latitude by 2.5° longitude and the 50km by 50km spatial scales a thin-plate spline was used. The interpolated variables were then subject to quality control as outlined in section 2.3.4 THIN-PLATE SPLINE. No additional processing was needed for geographic and temporal covariates as these variables are explanatory characteristics of the data points. The elevation was extracted from the DEM for each of the three spatial scales.

2.3.3 RADIOSONDE DATA PROCESSING

The Radiosonde data set includes point data for eight stations in and surrounding Alberta. An initial data quality assessment turned up some entries where -9999 were reported for missing data. This can occur if the sounding has a malfunction or if a certain level or index cannot be found or computed. The missing data code, -9999, showed up so often for the Equilibrium Level (EQLV) that a decision was made to remove the variable prior to any processing, reducing the number of Radiosonde variables to nine. For the remaining sounding variables, if a value of -9999 was reported it was replaced with NA in order to prevent future errors if -9999 was considered as a numeric value. Sometimes soundings are not performed at the 00Z and 12Z preset time. They may be performed

early, be delayed due to malfunction, or more than one may be done if severe weather is anticipated.

Codes were written to classify all soundings occurring before 03Z as 00Z, after 21Z as 24Z, and between 09Z and 15Z as 12Z. Those with a 24Z classification had one Julian day added to the date column and then were reclassified as 00Z. The average was computed for each of the variables if more than one sounding was available for a particular RAOB station, date, and time. This averaging resulted in one or fewer (if no sounding was available) rows of data for each station at a given date and time. Once this was complete, the data were subset to include only data with the 00Z and 12Z time classification. In order to interpolate values for Alberta at the same spatial grids as the lightning and Reanalysis data, three separate thin-plate splines were used to interpolate and smooth the newly generated Radiosonde data.

2.3.4 THIN-PLATE SPLINE

A Thin-plate Spline (TPS) regression was performed using the *Tps* function in the *fields* package (Furrer et al., 2013). The *Tps* function fits a TPS surface to irregularly spaced data. A TPS can be thought of as trying to fit a semi ridged sheet over an uneven surface terrain. The resistance of the sheet to bending is analogous to the penalties applied by the TPS for roughness. Now imagine that instead of the underlying terrain you only have an incomplete version with ***x*** number of irregularly spaced points. Given only ***x*** number of points, you want to best interpolate the missing data. A TPS is an interpolating and smoothing technique that tries to minimize the residual sum of squares. A TPS is restricted in that the function must have a certain level of smoothness (Green and Silverman, 1994).

Thin-plate splines were used on all of the weather predictor data sets in order to interpolate data points for the different spatial scales. For the initial 2.5° latitude by 2.5° longitude scale, TPS were applied to the sounding data only as the Reanalysis data came from the source with a 2.5° latitude by 2.5° longitude grid spacing. For the second spatial scale of 1.25° latitude by 2.5° longitude, TPS were applied to the sounding and Reanalysis data. The Reanalysis and sounding data were then subject to the third set of TPS for the 50km by 50km spatial scale.

When performing the TPS for the sounding data, an additional condition was set. Only eight data points (RAOB locations) are available for the sounding data. A condition was placed on the TPS such that a minimum threshold of five stations must have valid

observations of the interpolated variable in order for the TPS to run. If for a particular time (00Z or 12Z) and day there were less than five stations with valid observations available, no interpolations were made, instead the interpolated values for the variable were set to NA.

Quality control measures were put in place for all interpolated values. Acceptable ranges were set for each variable according to their physical limits or original range of values. Variables with physical limits were given a range with hard values. For example RH must be between zero and 100%, CAPE must be ≥ zero, and CINS must be ≤ zero etc. When a hard value was not physically, or theoretically supported, values were set based on the range of values for that variable in its original scale. The maximum and minimum values for each variable falling into this category were calculated and a buffer of ±10% was applied. If an interpolated value fell outside of the acceptable range, the value was replaced with the nearest acceptable value.

The presence of unacceptable values does not necessarily indicate there are issues with the TPS model. An unacceptable value can be generated for a point external to the reference points. For example, in Figure 8, six stations of data are available however none of the stations are located in the northern region. Without a station to bound the TPS it may continue on the current trajectory extrapolating values outside of the normal range. This is common when attempting to extrapolate values outside or at the edge of the reference points and is discussed in further detail in the Discussion Chapter in section 4.4.1 THIN-PLATE SPLINE. Once all of the data were processed, interpolated, and cleaned, they were merged together into six master files.

Each daily data set and each 6-hour data set contained the geographic and temporal covariates, daily mean Reanalysis data, four-time daily Reanalysis data and the 00Z and 12Z sounding observations for that spatial scale. The lightning observations were also merged into each data set. The 50km by 50km data were further subdivided into three zones shown in Figure 7 leading to the 10 data sets listed in Table 2. Before moving on to the modeling phase, all data points with NAs present were removed. This prevented the missing data from creating future problems when running correlation analyses, building the models, and using the models for predictions.

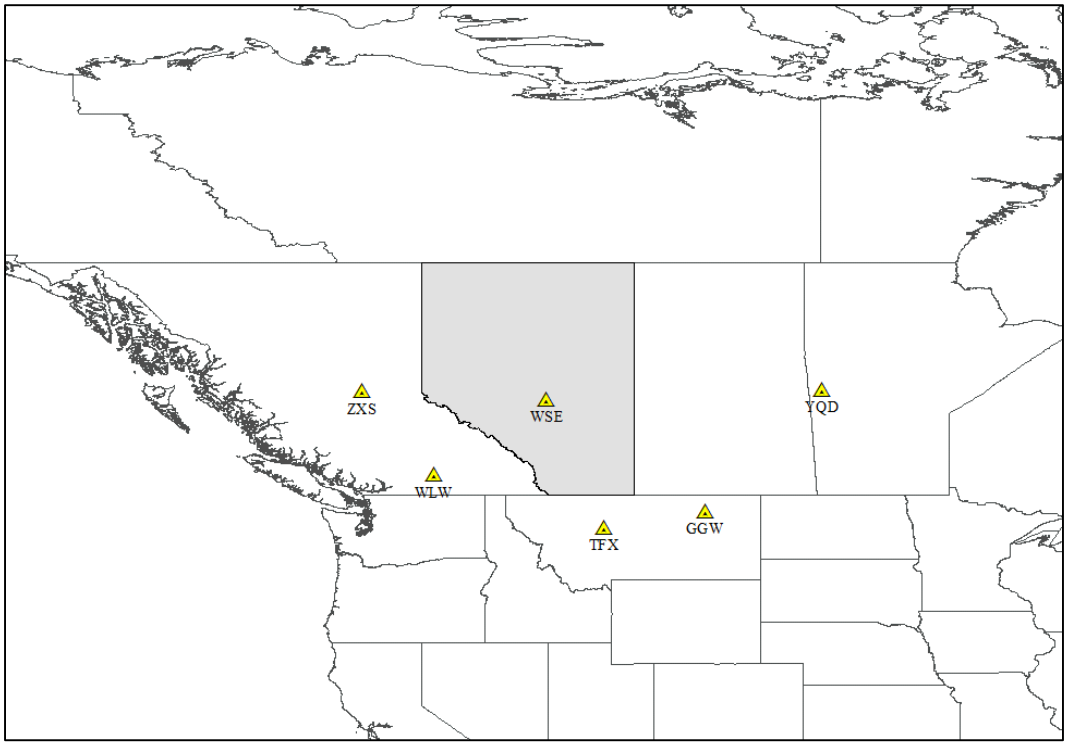

Figure 8: An example of a situation where data is only available for six of the eight RAOB stations. When using TPS to interpolate data for Alberta (highlighted), the northern most extrapolated values may be outside of an acceptable range as there is no point to bound the spline.

Table 2: List of the 10 data sets created for the two time frames (daily and 6-hour) and three spatial scales. The finest resolution scale (50km by 50km) is further subdivided into three separate zones. The number of observations remaining after all rows with NAs present were removed are also provided for all data sets. These 10 data sets are the basis for the lightning prediction models.

Lightning Prediction Model Data Sets	Number of Observations
2.5° Latitude by 2.5° Longitude Daily	64,674
2.5° Latitude by 2.5° Longitude 6-hour	282,896
1.25° Latitude by 2.5° Longitude Daily	106,875
1.25° Latitude by 2.5° Longitude 6-hour	499,680
50km by 50km Boreal Forest Daily	290,394
50km by 50km Boreal Forest 6-hour	1,083,904
50km by 50km Parkland and Grassland Daily	93,483
50km by 50km Parkland and Grassland 6-hour	348,928
50km by 50km Mountain and Foothills Daily	119,340
50km by 50km Mountain and Foothills 6-hour	445,440

2.4 MODELING METHODS

A series of random forest models were created for each spatial scale and time frame. The *randomForest* package (Liaw and Wiener, 2002) was used. An iterative process was used where the variable importance from each random forest model was used to decrease the number of predictors and generate a new random forest model until only the top five variables remained. This section provides a basic introduction to random forests followed by the various techniques used to generate the random forest models.

2.4.1 RANDOM FOREST: BACKGROUND

Random forest (Breiman, 2001) is a supervised machine learning algorithm that constructs a forest by growing multiple classification trees. For a training set with ***n*** number of rows (number of entries) and ***m*** number of columns (number of input variables), each tree, and thus the forest, is grown as follows (modified from Breiman (2001)). Unless otherwise specified by the argument *sampsize*, a random bootstrap sample (with replacement) of ***n*** entries creates a new training set which will be used to grow a new tree. A subsample of the ***m*** input variables are randomly selected and the best split is selected to split the node. The best split for the node is often a strong variable with a clear division of values (or categories) between the categorical outputs. No pruning takes place therefore every tree is grown to the largest possible extent. Any number of trees can be grown to construct the forest (argument *ntree*).

The error rate depends primarily on if any two trees are correlated (and the level of correlation) as well as the individual strength of the trees (Breiman, 2001). Correlation between trees increases the error rate while high tree strength improves the skill and decreases the error rate. The number of input variables has a direct impact on the error rate. Increasing the number of input variables has a positive relationship with between tree correlation, increasing the correlation and thus error (Breiman, 2001). Increasing the number of inputs also increases the individual tree strength and thus decreases error (Breiman, 2001). Due to these relationships, it is common practice to look for an optimum number of variables by comparing the Out-of-Bag (OOB) error rate for different forests generated with different numbers of input variables.

When the training set for the current tree is being sampled with replacement under default conditions, roughly two-thirds of the data set is selected. This leaves one-third as the OOB data. Once the tree is grown the OOB data is used to find the OOB error

estimate and to estimate the importance of each input variable. Due to the retention of OOB data, Breiman (2001) argues there is no need for cross-validation. Despite this recommendation a decision was made to manually separate the original data sets into separate training and validation sets. It was felt that a measure of truly independent validation could be accomplished in this way. Each *randomForest* generates an output of the input variables relative importance. The *importance* function displays the *MeanDecreaseGini* variable importance for the forest.

The *MeanDecreaseGini* output provides a relative ranking of variable importance. If a variable is important it plays a large role making accurate predictions (Liaw, 2009). Likewise, randomly altering the variables values would have a strong effect on the skill of the predictions. The *MeanDecreaseGini* output tries to capture this importance by assigning relative values to all of the input variables (Liaw, 2009). The highest values are given to the most important variables while lower values are given to variables with less importance. The random forest algorithm was chosen for this study due to its efficiency with large data sets, ability to handle large number of input variables, and ability to rank input variables by importance. Both categorical and quantitative variables are used to predict lightning. This is not a problem for *randomForest*. In addition, the *randomForest* package has some built in arguments that can be modified to deal with imbalanced data.

2.4.2 CONUNDRUM OF IMBALANCED DATA

Many practical classification problems involve imbalanced data. A data imbalance occurs when one or more of the classes account for a small proportion of the data while another accounts for a large percentage. In other words there is one large class and another small or rare class. Often our interests rest in the rare class, also known as the "positive class", while there is little interest is the large class "negative class". Many examples of imbalanced data can be found in research dealing with rare disease diagnosis and fraud detection (Chen et al., 2004). Focusing back on lightning observations, we see that lightning events account for a small minority of the data while non-occurrences make up the majority of the observations.

Machine learning classification algorithms can often fall short when dealing with imbalanced data as they are formulated to minimize overall error rates. This focus on overall accuracy favours the *negative class* producing a model with little information about the object(s) of interest. Researchers from many different fields and disciplines

have worked towards addressing this problem employing different statistical methods. One such method is to modify the arguments within the *randomForest* algorithm for classification and regression. There are three different typical approaches to deal with increasing the accuracy of the positive class prediction skill when working with random forests.

The first two approaches involve resampling of the data to help balance out the disproportion between the classes while the third takes a different approach by implementing cost-sensitive learning (Chen et al., 2004). Up-sampling (boosting) the minority class, and down-sampling the majority class are the two general resampling methods known as a Balanced Random Forest (BRF) approach (Chen et al., 2004). Up-sampling involves random resampling of the *positive class* with replacement. This boosting of the minority class adds repeated data points thereby increasing the overall number of positive class measurements. In contrast, down-sampling of the majority class involves a random, or strategic, sub-sampling of the majority class to omit data points and decrease the number of measurement in the negative class. While both approaches increase the skill of the model with respect to the positive class, up-sampling leads to increased computational time while the down-sampling approach can result in a loss of information as not all data points from the negative class may be used.

Weighted Random Forest (WRF) is the third method suggested by previous research and involves a cost-sensitive approach where class weights are implemented into the random forest model such that more weight is applied to the object(s) of interest (Pazzani et al., 1994). The performance of the three approaches was assessed by Chen et al. (2004) by comparing the resulting confusion matrixes from six unique data set runs. It was found that while both WRF and BRF show about equal improvements, BRF has a better false negative rate while WRF has a slightly better true positive rate (Chen et al., 2004).

The BRF method was chosen as it decreases the overall computation time compared to WRF (Chen et al., 2004). A decision was made to perform the BRF with down-sampling only as up-sampling would increase computational time and expense. Furthermore, down-sampling can be performed without a loss of information if implemented correctly. The initial BRF method used involves the down-sampling of the majority class such that the sample sizes are equal for the minority and majority classes. The number of trees grown are at minimum greater than the number needed to include all of the training data available for the majority class resulting in little or no lost

information. This method is more computationally efficient than the others as a smaller, balanced, sample of data is used for each tree.

The following approach was used to build the series of balanced random forests used to determine the top predictors. Let ***n*** be the number of observations in the minority class. Using the *sampsize* argument in the *randomForest* package, a random bootstrap sample with replacement of ***n*** number of data points were drawn from the minority and the majority classes. This effectively balances the number of observations from each class that will be used to build each tree. The key is to limit the negative examples and keep all positive examples even if they are noisy as they are too rare to waste (Kubat and Matwin, 1997).

2.4.3 RANDOM FOREST : MODELING

A suite of random forests models were run on the 10 data sets. Before an attempt was made to run the models for all of the data sets, a handful of preliminary models were run on subsets of the various sets to get a feel for the general trends and relationships between different variables and lightning occurrence. These exploratory runs were also used to narrow down the number of variables and thus the size of the data sets. The large size of the data files made computation time and computer memory a major concern. The majority of the processing and model building was able to be done in steps with 16GB of RAM however some runs required a 32GB machine. Following the preliminary models the newly refined data sets were run through a suite of random forest algorithms to produce multiple models for each spatial and temporal scale.

Preliminary Runs

At this point the data sets included all of the Reanalysis variables daily mean and four-time daily observations, two-times daily sounding observations, and the geographic and temporal covariates. In addition, 24-hour change is included for each weather variable. The data sets also contained a binary (**0**/**1**) predictand where **0** indicates no lightning and **1** indicates lightning observed. The exploratory runs began with the smallest of the data sets, the daily 2.5° latitude by 2.5° longitude. This data set contained ~740 columns of input variables (including all variable variations) and ~65,000 rows corresponding the location and time of each observation. Before any random forest modeling was attempted, a correlation analysis was performed.

The *findCorrelation* function from the *caret* package (Kuhn et al., 2013) was used to remove correlated variables. Correlation thresholds of $r \geq 0.7$ ($r^2 = 0.49$) and $r \geq 0.9$ ($r^2 = 0.81$) were implemented in the code. The r^2 is often thought of as the amount of variation in one variable explained by the other. A correlation coefficient between 0.7 and one (-0.7 and -1) indicates a strong linear relationship between the two variables. A threshold of 0.7 allowed for removal of one of the strongly correlated variables. When two variables are found to be correlated, the function removes the variable with largest mean absolute correlation (Kuhn et al., 2013). The correlation analysis resulted in a much smaller data set with only 111 variables remaining when the 0.7 threshold was implemented. Some preliminary random forests were then run on the remaining daily 2.5° latitude by 2.5° longitude data.

A random subsample of seven years made up the training set which was used to build the model (1999, 2000, 2002, 2004, 2006, 2007, and 2009). For consistency these same seven years were used to build all future training data sets. The *randomForest* codes were run under default conditions (shown below) unless otherwise stated. Where ***x*** = input variables, ***y*** = response vector, and *ntree* specifies the number of trees in the forest (default 500).

randomForest(x, y, ntree=500,
replace=TRUE, classwt=NULL,
sampsize = nrow(x))

The bootstrap sample of training data taken will be of size ***n*** = number of rows in ***x*** and occurs with replacement by default. No class weightings are applied unless specified. The response vector, ***y***, is a binary (**0/1**) predictand which is classified as factor therefore the *randomForest* runs as a classification algorithm. If the predictand were numeric, regression would be assumed (Liaw and Wiener, 2002).

The exploratory runs resulted in a decision to remove all of the daily mean Reanalysis variables as well the RH, U wind and V wind variables from the data sets. A new correlation analysis was run on the remaining variables (~380) using the same method described above. The resulting trimmed down data set contained 75 input variables, far smaller than the initial ~740. Some additional preliminary models were run on subsets of other data sets to insure the removal of the daily mean variables, RH, U wind and V winds would not negatively impact the models skill. The removal of these variables had no significant impact on the models skill and the decision was made to

remove the variables in question and all of their variations from all of the data sets prior to the correlation analysis.

An additional predictor variable, elevation, was added to the 50km by 50km data set. Due to the large size of the 50km by 50km data sets, an attempt was made to decrease the total number of data point thereby reducing computational time and memory requirements by changing the study time frame from April 1st to October 31st to May 1st to September 30th. An exploratory run was performed on the 6-hour and daily Mountain and Foothills data sets to compare the influence the change of dates would have on the model. The shorter date range altered the model accuracy and therefore was not continued despite the computational benefits.

Creating Random Forest Models

The newly trimmed down data sets were used to begin building the predictive models for each of the data sets. For each of the spatial and temporal scales the following methods were implemented. An assessment of data imbalance for each data set was performed. The *findCorrelation* function (Kuhn et al., 2013) was used to remove variables with correlation coefficient greater or equal to 0.7. The data classification of the latitude, longitude, time (if 6-hour model), and Haines index inputs were set to factor while the rest remained as numeric. An exception occurred for the 50km by 50km models where the latitude and longitude could not be classified as factors due to a limitation of *randomFores*t only permitting factors with 32 or fewer classes. For these models the latitude and longitude were left as numeric.

The same seven years randomly chosen to create the training set were used to build the models (1999, 2002, 2004, 2009, 2000, 2007, and 2006). A series of random forests were then run. The first random forest was run with all of the variables remaining post correlation analysis. Due to the imbalanced nature of the data the random forests were modified with a BRF sub-sampling approach. The *sampsize* argument was set as follows: For the binary predictand (**0**/**1**), let ***n*** be the number of rows in which **1** occurs (number of lightning events). The random bootstrap sample with replacement was forced to sample ***n*** entries with lightning events and ***n*** entries with non-events effectively balancing the training set for each tree.

For the first run, seven balanced random forests were run. The number of trees was set to 10 and for the first run with subsequent runs performed for *ntree* of 25, 50, 100, 200, 300, and 400. The range of trees were used to determine the number of trees

needed to stabilize the model. The codes written were programmed to automatically select and save the best random forest model and use that model to rank the predictors by *MeanDecreaseGini* score. The top 75 predictors were then selected and a new set of *randomForest*s were run with *ntree*= 100, 150, 200, 250, 300, 350, and 400. The 2.5° latitude by 2.5° longitude data only had 75 predictors remaining post correlation analysis. These predictors were used in the previous runs therefore this step was skipped for data at this scale. In the same manner, the variables were ranked and a smaller number of variables were used for the next model. This process was repeated as the number of predictor variables were decreased from 75 to 35, 25, 20, 15, 21, 10, eight, and five.

For comparison, an additional set of models were created for the 2.5° latitude by 2.5° longitude data. These models were created with an imbalanced approach where the *sampsize* argument was left in its default form. Once all of the prediction models were created, the various models for each scale were compared to determine the optimum number of input variables. Predictions were then made with the top 12 variables and top 15 variables. The prediction were made on the independent validation data set with the six unused years (2001, 2003, 2005, 2008, 2010, and 2011).

The predictions were done in an ensemble like fashion where the top 12 and top 15 variables, as determined in the previous stages, were used to regenerate a series of new random forests from the training set. Each newly generated forest was then used to make predictions for the validation data set. A model output of **1** indicates the models forecasts a lightning event and a forecast of **0** represents a forecasted non-event. This was repeated 10 times. The 10 predictions were then averaged to create ensemble forecast ranging from zero to one. This process was done for a series of *sampsize* variations in an attempt to maximize the **1** (event) prediction skill.

The skills of the models were then analyzed. Different thresholds were used to separate lightning predictions and non-lightning predictions. Lightning prediction thresholds of ≥ 0.5, ≥ 0.7, and ≥ 0.9 were applied to the ensemble forecasts. Meteorological forecast skill criteria were then used to analyze the performance of the different models. The models skill for correctly forecasting the lightning events were measured by the Post-Agreement (PAG), Hit Rate (H), and Proportion Correct (PC). The False Alarm Ratio (FAR) and False Alarm Rate (F) were used to assess the skill of the models with respect to false event forecasting. The Equitable Threat Score (ETS) was also calculated for each model to provide an overall picture of model skill. The Critical Success Index (CSI), or Threat Score, is sensitive to imbalanced data and often gives

poor scores for rare events therefore this metric was calculated but is not included. A list of the forecast skill measures, formulas and general information are shown in Table 4. The formula inputs for Table 4 are shown in the contingency table provided in Table 3.

Table 3: Contingency matrix for model forecasts. The A and D cells (highlighted in grey) show the correctly forecasted events and non-events, respectively.

		Event Observed	
		Yes (1)	**No (0)**
Event Forecast	**Yes (1)**	***A*** *(hit)*	***B*** *(false alarm)*
	No (0)	***C*** *(miss)*	***D*** *(correct non-event)*

Table 4: List of the forecast skill criterion used to analyses the lightning prediction models. The inputs for the formulas can be found in the contingency matrix shown in Table 3.

Abbreviation/Name		**Formula**	**Description**
CSI	Critical Success Index (Threat Score)	$CSI = \frac{A}{A+B+C}$	Ranging from zero to one, a one indicates a perfect forecast. Taking into account both false alarms and missed events, the CSI is sensitive to imbalanced data, often giving poor scores for rare events. This index was not used but is included in the table for comparison to the ETS.
ETS	Equitable Threat Score	$ETS = \frac{A - A_r}{A+B+C-A_r}$ where, $A_r = \frac{(A+B)(A+C)}{A+B+C+D}$	Provides a more balanced threat score when dealing with rare events. The ETS has a range from -1/3 to one. A higher score indicates increased forecast skill while a score below zero indicates an unskilled forecast.
F	False Alarm Rate	$F = \frac{B}{B+D}$	Also known as the probability of false detection, F shows the fraction of observed non-events that were forecasted as false alarms.
FAR	False Alarm Ratio	$FAR = \frac{B}{A+B}$	Represents the fraction of forecasted events that are false alarms.
H	Hit Rate (1 skill)	$H = \frac{A}{A+C}$	Also known as the probability of detection, the Hit Rate is a score from zero to one where one is a perfect forecast. This measure of skill is sensitive only to misses and does not take false alarms into account.
PAG	Post-Agreement	$PAG = \frac{A}{A+B}$	The complement of FAR (1-FAR), PAG is the fraction of the forecasted events which are correct.
PC	Proportion Correct	$PC = \frac{A+D}{A+B+C+D}$	Provides the overall skill of the model but is not a good measure of skill for forecasting rare events of interest.

CHAPTER 3. RESULTS

The results are split into three separate sections. The various findings from the exploratory runs that were instrumental in shaping the random forest modeling are discussed first in this chapter. Next, the results from the top predictor selection process for each data set are provided. Finally, the results from the various lightning ensemble forecasts for each spatial and temporal scale are presented.

3.1 EXPLORATORY RUNS

The exploratory runs provided valuable insight, highlighting many data and model trends. Basic statistics were generated for the lightning data. Figure 9 shows the diurnal and seasonal trends. Some variables were rarely, if ever, found to be highly important as determined by the *MeanDecreaseGini* output. RH, U winds, V winds, and the daily mean Reanalysis variables for all surface and vertical levels were rarely selected by the models as top predictors. Removal of these variables in the preliminary runs also found that the model skill was not negatively affected. In fact, removal of these variables prior to the correlation analysis allowed other variables previously removed by the correlation analysis to make it through to the model runs. Some of the new variables were found to be important contributors to the models' skill. Due to the large number of input variables and the lack of contribution to the models skill, a decision was made to remove all of the daily mean variables as well as the 6-hour RH, U wind, and V wind variables prior to the correlation analysis.

Two different Pearson's correlation coefficient ($\boldsymbol{r}$) thresholds, $|\boldsymbol{r}| \geq 0.7$ and $|\boldsymbol{r}| \geq 0.9$, were explored for the correlation analysis. A threshold of $|\boldsymbol{r}| \geq 0.9$ resulted in many strongly correlated variables remaining in the data set. The computational expense was increased as a greater number of variables made it through the correlation analysis and into the initial model runs. Despite the increase in the number of variables, no increase in model skill was observed, therefore a decision was made to use the $|\boldsymbol{r}| \geq 0.7$ threshold for all future models. The findings of Dormann et al. (2013) support the decision to remove variables with $|\boldsymbol{r}| \geq 0.7$. The authors found that when variables with correlation coefficients greater than 0.7 were included in various multiple regression and machine learning approaches (including random forest), the collinearity began to severely distort the models and thus degrade the predictive skill (Dormann et al., 2013).

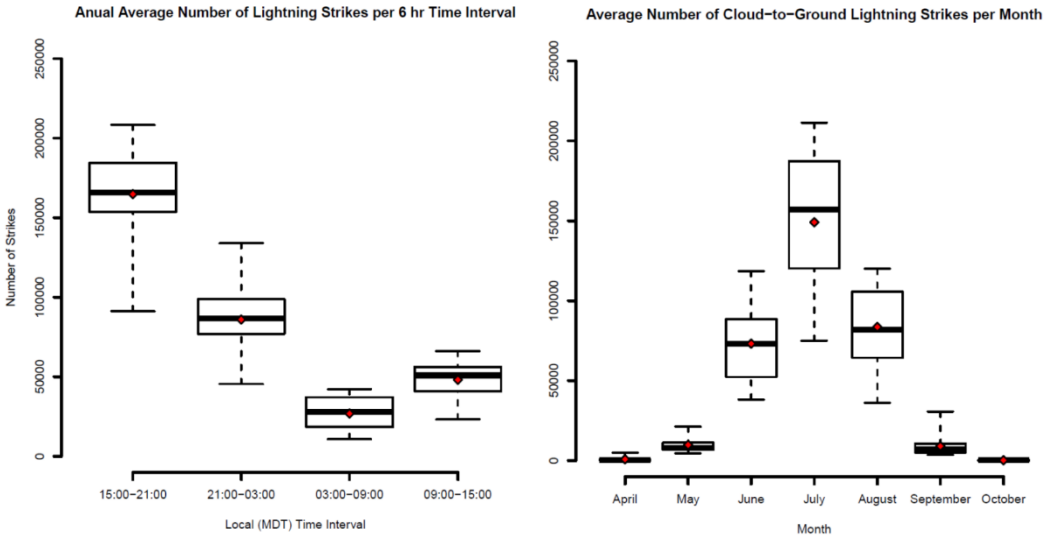

Figure 9: The boxes encompass the 1st and 3rd quartiles with the median indicated by a heavy center line and the mean marked by the red diamond. Whiskers extend to the minimum and maximum values. The Box plots represent the average values for13 consecutive years of lightning data from 1999 to 2011.

In an attempt to further decrease the computational expense of the models, a trial set of random forest runs were performed with a shorter subset of days on the 50km by 50km data. Observations from May 1st to September 30th were used instead of the initial April 1st to October 31st date range. While the models built on the smaller data set had a slightly better (~1%) prediction skill for event occurrence and non-occurrence, the FAR increased by over 6%. A decision was made to stick with the initial date range.

The various data imbalances were also identified in this preliminary phase. The imbalances (Table 5) range from a ~30/70 (1/0) split for the daily lightning data at 2.5° latitude by 2.5° longitude spatial scale, to a ~4/96 (1/0) split for the 6-hour 50km by 50km Parkland and Grassland data set. The 6-hour lightning data sets have a higher imbalance than the corresponding daily data sets. The predictand imbalance also increases with increasing spatial resolution. An initial assessment was performed on the imbalanced data at the 2.5° latitude by 2.5° longitude with no balancing modifications. While the overall error rate was good (<14% with 15 variables), the positive class error rate was quite dismal (>30% for 15 variables). Forcing equal sample sizes through the *randomForest* model via the *sampsize* operation produced superior results for the objectives by minimizing the positive class error rates while still maintaining a near equal overall error rate.

Table 5: Level of imbalance between the *positive* (events) and *negative class* (non-events) for each data set. The number of observations are for the training sets, however, the proportions hold true for both the training and validation sets.

Data Sets	Lightning Events		Non-Lightning Events	
	#	%	#	%
2.5° Latitude by 2.5° Longitude Daily	19,435	30.05%	45,239	69.95%
2.5° Latitude by 2.5° Longitude 6-hour	34,351	12.14%	248,545	87.86%
1.25° Latitude by 2.5° Longitude Daily	25,830	24.17%	81,045	75.83%
1.25° Latitude by 2.5° Longitude 6-hour	46,872	10.12%	416,449	89.88%
50km by 50km Boreal Forest Daily	38,626	14.25%	232,350	85.67%
50km by 50km Boreal Forest 6-hour	54,414	5.02%	1,029,490	94.98%
50km by 50km Parkland and Grassland Daily	14,740	13.23%	96,620	86.77%
50km by 50km Parkland and Grassland 6-hour	16,079	3.65%	424,714	96.35%
50km by 50km Mountain and Foothills Daily	14,996	17.19%	72,236	82.80%
50km by 50km Mountain and Foothills 6-hour	22,249	6.38%	326,679	93.62%

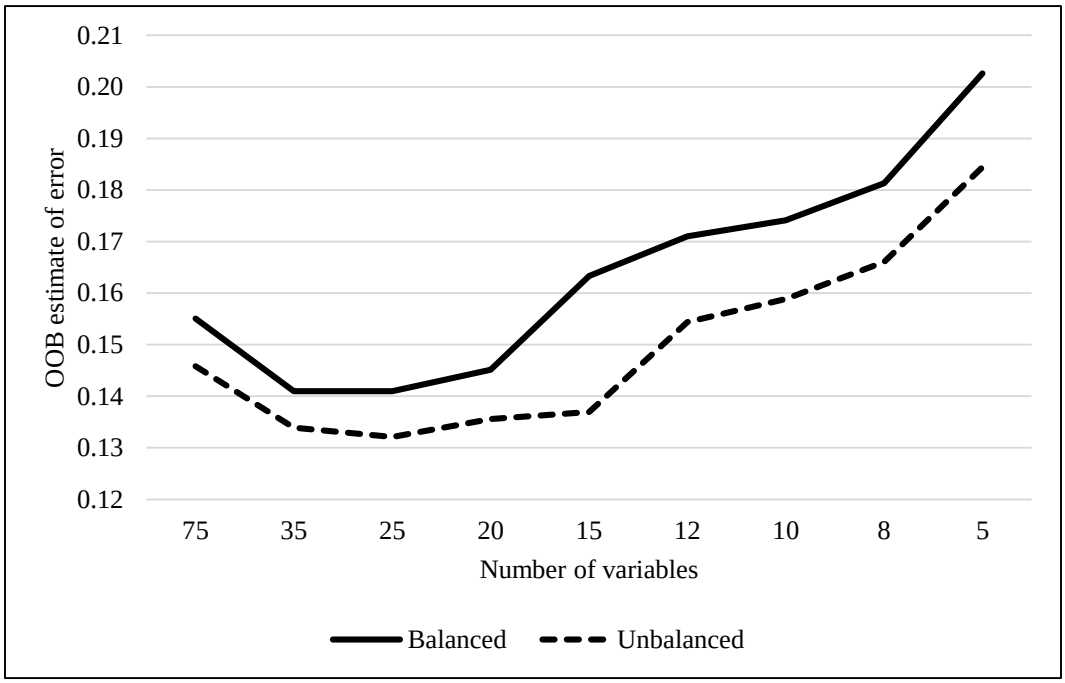

Figure 10: Comparison of the overall OOB error estimate for a series of balanced and unbalanced random forest models. Models were run on the daily 2.5° latitude by 2.5° longitude data set. The overall error rates are very similar for the two models.

Figure 10 provides a comparison of the overall OOB error estimate for models generated with balanced and unbalanced approaches. The overall error rates are quite similar for the two model approaches with the unbalanced model consistently having a slightly lower overall error rate. A BRF approach does result in a slight (1 to 2%) overall increase in error rate however the increased skill to the positive class is necessary to help meet the objectives of predicting lightning events. Breaking down the OOB error estimate by class shows that the slight decrease in PC is due to an increase in the non-event prediction error and an increase in event prediction skill (hit rate). In Figure 11 we see the non-event prediction error rate increase by ~7% while the event error rate is nearly halved. Due to these results, a BRF approach was implement to build all of the random forests.

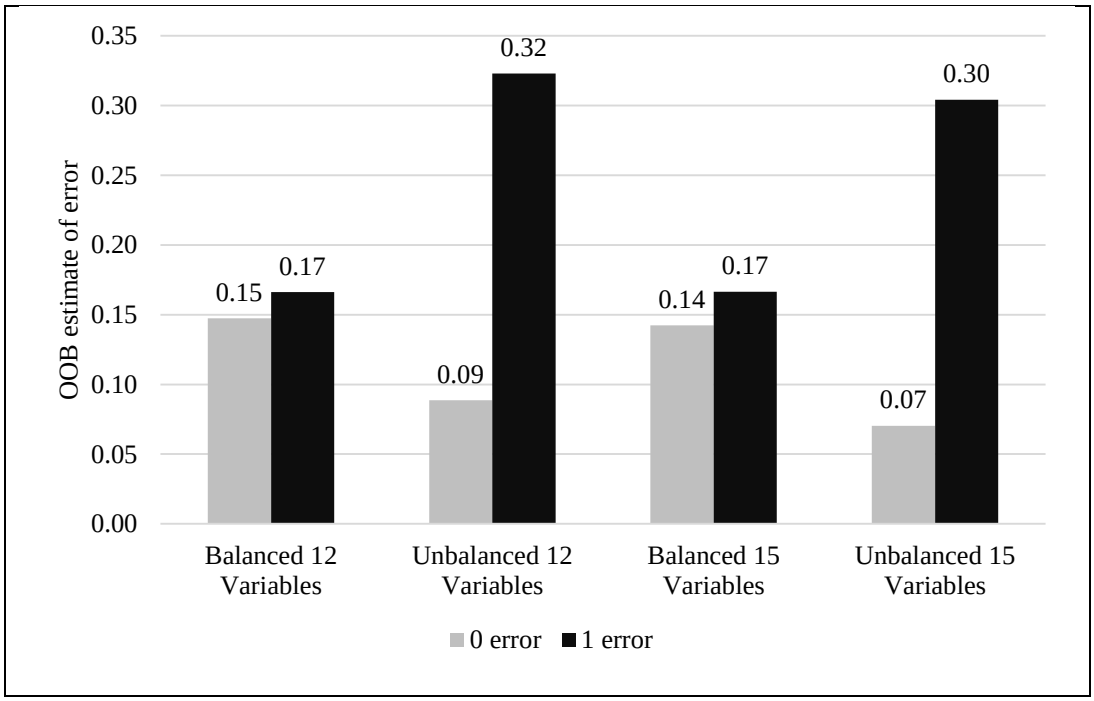

Figure 11: Comparison of the class **0** (non-event) and class **1** (event) OOB error estimate for balanced and unbalanced random forest models generated with 12 and 15 variables.

A series of trial random forests were run to find the number of trees required to obtain stable model outputs. A sufficient number of trees was determined by generating a simple plot of the OOB error estimate for the series of random forests. The random forests were built under default conditions and under BRF conditions. The number of trees (*ntree*) were specified at seven different values (Figure 12). Various runs were performed on each of the data sets. It was found that the same general trend shown in

Figure 12 emerged for all of the data sets. The plot shows the OOB error estimate beginning to balance out around 100 trees and stabilizing by 200 trees. Based on these runs, all future random forest runs were performed in seven steps starting with *ntree*=100 and increased at increments of 50 until *ntree*= 400.

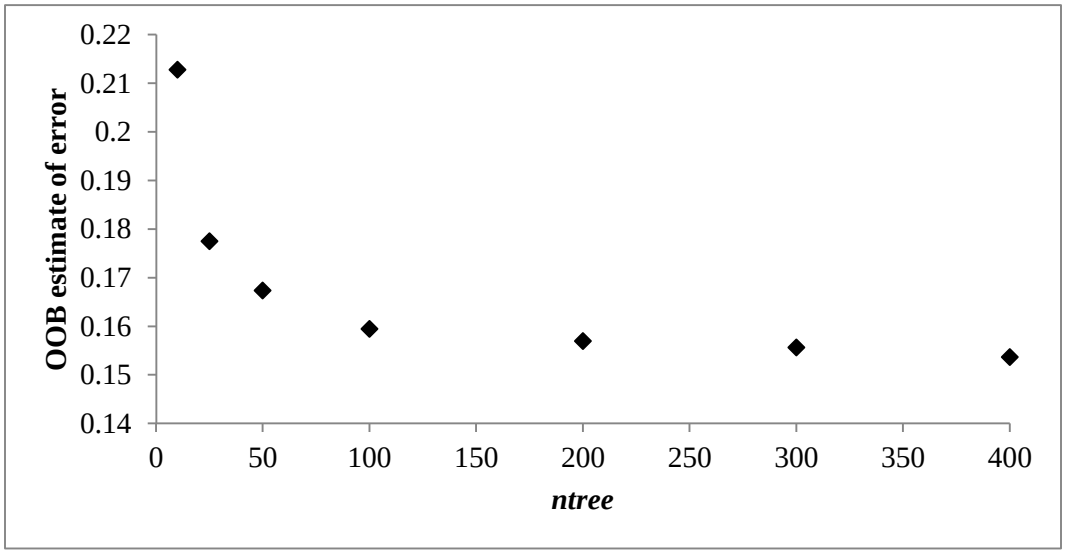

Figure 12: The OOB error estimate trend for random forest models with different number of trees (*ntree*). The OOB estimate or error begins to balance around 100 trees and stabilizes by 200 trees.

3.2 DETERMINING TOP PREDICTORS

Following the TPS and quality control measurers described in 2.3.4 THIN-PLATE SPLINE, all data points with missing data (NA) were removed. The removal of NAs resulted in a loss of ~7% of the points on average. The percentage of points removed from the positive and negative predictand classes were nearly equal. Once all NAs were removed, a series of balanced random forest runs were performed on each data set as outlined in 2.4.3 RANDOM FOREST : ODELLING. For each data set, the model with the lowers OOB in each run (with ***x*** number of variables) was selected and used to narrow down the number of variables for the following run. The OOB error estimate for the optimum models from each random forest run with ***x*** variables are show in Figure 13a-e.

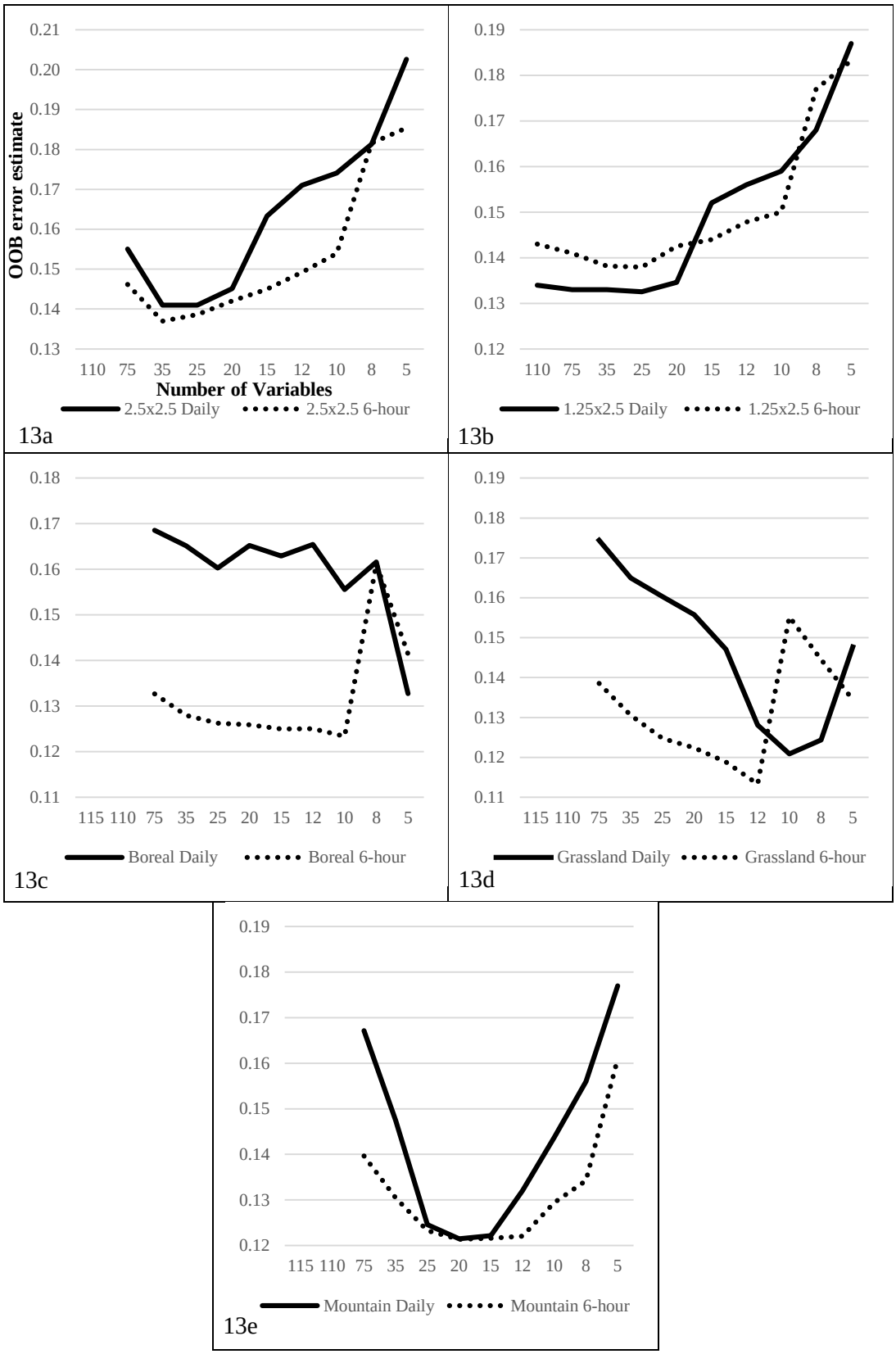

Figure 13: The OOB error estimate (y-axis) fluctuations as the number of variables included in the random forest model are changed (x-axis). Note the different scales for the x and y axes.

The first random forest run for each data set includes all the variables remaining after the correlation analysis. There are 75 variables for the initial daily and 6-hour 2.5° latitude by 2.5° longitude runs, 110 variables for the 1.25° latitude by 2.5° longitude runs and 115 variables for each of the three 50km by 50 km runs. According to the overall OOB error estimate, the optimum number of variables for the 2.5° latitude by 2.5° longitude and 1.25° latitude by 2.5° longitude scales occurs around the 25 to 35 variables mark. The Boreal Forest models optimize around 10 to 12 variables with a second optimization occurring when only 5 variables are present. The Parkland and Grassland models also have optimum error rates when 10 to 12 variables are used with the overall error rate decreasing when only 5 variables are used for 6-hour Parkland and Grassland models. Finally, the Mountain and Foothills have optimal error rates when the top 12 to 20 variables are included in the random forest.

Ten separate sets of top predictors were created based on the *MeanDecreaseGini* values. Table 6 provides an example of variable importance as generated by the *importance( )* function. The output provides the *MeanDecreaseGini* which is a relative value of the variables' importance. Since the values are relative, they cannot be compared between models. The top 12 and 15 predictors from each data set are shown in Table 7. The top predictors are similar among many of the models. The 00Z Showalter Index (SHOW00), Convective Available Potential Energy (CAPE00), Lifted Condensation Level Temperature (LCLT00), Julian day , 00Z and 12Z Convective Inhibition (CINS00, CINS12), and time of day (for 6-hour models) are some of the highest ranked predictors for the 2.5° latitude by 2.5° longitude and 1.25° latitude by 2.5° longitude models. A few of the top predictors for the 50km by 50km scale varied from those identified for the coarser scales. Latitude (lat), mean sea level pressure (mslp), elevation (elv), and Severe Weather Threat Index (SWET00, SWET12) show up as top predictors alongside SHOW00, CAPE00, Julian day, and time. The top 12 and 15 predictors shown in Table 7 were then used to generate lightning forecast ensembles for each of the independent validation data sets.

Table 6: Variable importance for a 1:1 BRF model with *ntree=300* generated for the 6-hour Boreal Forest data set. The higher the *MeanDecreaseGini* value, the greater the variable importance and thus the rank.

Boreal Forest 6-hour		
Predictor	***MeanDecreaseGini***	**Rank**
SHOW00	4132.005	1
CAPE00	2718.889	2
time	2300.323	3
mslp_00	2296.990	4
lat	2279.292	5
day	2231.839	6
CINS00	2053.856	7
SWET12	2029.684	8
Temp_8_18	2000.488	9
omega_9_12	1979.637	10
SWET00	1968.168	11
CINS12	1827.819	12

Table 7: List of top 15 predictors for each data set. Based on models built with a Balanced Random Forest (BRF) approach. The various data sets are listed across the top with daily lightning data sets highlighted in white and 6-hour lightning lowlighted with a medium grey background.

Top Predictors	2.5° Latitude by 2.5° Longitude Daily	2.5° Latitude by 2.5° Longitude 6-hour	1.25° Latitude by 2.5° Longitude Daily	1.25° Latitude by 2.5° Longitude 6-hour	50km by 50km Boreal Forest Daily	50km by 50km Boreal Forest 6-hour	50km by 50km Boreal and Parkland Daily	50km by 50km Grassland and Parkland 6-hour	50km by 50km Grassland and Foothills Daily	50km by 50km Mountains and Foothills 6-hour
1	SHOW00	SHOW00	SHOW00	SHOW00	SHOW00	SHOW00	lat	SHOW00	SHOW00	SHOW00
2	CAPE00	LCLT00	CAPE00	CAPE00	lat	time	SHOW00	lat	lat	time
3	LCLT00	CAPE00	LCLT00	LCLT00	CAPE00	CAPE00	day	day	elv	CAPE00
4	day	day	day	day	mslp_00	mslp_00	elv	CAPE00	long	lat
5	CINS12	time	CINS00	time	day	day	CAPE00	SWET12	CAPE00	elv
6	CINS00	CINS00	mslp_00	CINS00	SWET12	CINS00	mslp_00	mslp_00	day	day
7	mslp_00	CINS12	CINS12	CINS12	CINS00	lat	SWET12	SWET00	SWET12	CINS00
8	T.Td_4_18	mslp_00	SWET00	mslp_00	SWET00	SWET12	SWET00	time	CINS00	long
9	SWET12	SWET00	SWET12	SWET12	CINS12	CINS12	omega_9_00	omega_9_18	mslp_00	SWET12
10	SWET00	SWET12	T.Td_1_12	temp_8_18	temp_8_18	temp_8_18	omega_9_18	temp_8_18	sfc_p_00	CINS12
11	T.Td_1_12	T.Td_1_12	T.Td_4_18	T.Td_4_18	omega_1_00	SWET00	temp_8_18	CAPE12	omega_9_00	omega_9_12
12	LCLP00	T.Td_4_18	T.Td_4_00	omega_9_12	T.Td_1_06	omega_9_12	LCLP00	elv	SWET00	mslp_00
13	omega_9_12	omega_9_12	omega_9_00	SWET00	omega_9_12	T.Td_1_06	T.Td_1_12	CINS12	temp_8_18	temp_8_18
14	T.Td_3_00	T.Td_1_18	CAPE12	LCLP00	T.Td_1_12	omega_9_18	T.Td_1_00	T.Td_4_00	LCLP12	omega_1_12
15	long	CAPE12	lat	CAPE12	elev	T.Td_1_12	CAPE12	CINS00	CINS12	SWET00

3.3 LIGHTNING PREDICTION MODELS

Models for the prediction ensemble were generated from the results of the 1:1 BRF runs performed on each data set. Six different approaches were taken to building the prediction models for the 10 data set (Table 8). Each of the six models used a different bootstrap sampling method to generate the random forest models. A BRF approach was used for five of the six models. For the final unbalanced model approach, the random forests were built with default *sampsize* making the imbalance for each model roughly equal to that of the original data set (Table 5). The predictions were made in an ensemble like fashion where the top 12 and top 15 variables, as determined in the previous stages, were used to regenerate a series of 10 new random forests from the training set with *ntree*=251. Each newly generated forest was then used to make predictions for the corresponding validation data set. A model output of **1** indicates the model forecasts a lightning event while an output of **0** represents a forecasted non-event. This was repeated 10 times. The 10 predictions were then averaged to create ensemble forecasts ranging from zero to one.

Table 8: Six predictions models were generated with the top 15 and top 12 variables from each data set. Let ***n*** be the number of observations in the minority class. The second and third columns show how the *sampsize* arguments for each model were specified to change the number of observations sampled from each class with replacement. The unbalanced model was run under default conditions making the imbalance similar to the original data imbalance outlined in Table 5.

Prediction Model Name	Number of Non-Events (0) Sampled	Number of Lightning Events (1) Sampled
Balanced (1)	***n***	***n***
Balanced (0.8)	0.8* ***n***	0.8****n***
Balanced (0.6)	0.6****n***	0.6****n***
0.8:1	0.8****n***	***n***
0.6:1	0.6* ***n***	***n***
Unbalanced	Proportion roughly equal to original class imbalance	

Lightning prediction thresholds of ≥ 0.5, ≥ 0.7, and ≥ 0.9 were applied to the resulting ensemble forecasts. Contingency tables were generated for each ensemble and the skill was analyzed with meteorological forecast skill criteria (Table 4). The changes in overall forecast skill, event forecast skill, and non-event forecast skill between the various prediction ensemble model are demonstrated in Figure 14. The overall forecast skill of the models are around 80% for all of the data sets. As the event forecast skill increases, the non-event skill and the overall forecast skill decreases. Intuitively, setting the forecast threshold to ≥ 0.5 produces a superior hit rate while also increasing the FAR

and F. The threshold of ≥ 0.5 implies that only 50% of the models must predict lightning for the ensemble to forecast a lightning event. The three thresholds for the 0.6:1 prediction models generate the highest skilled ensemble forecasts when measured by hit rate while also generating the least skilled forecast for non-events. The balanced (1:1) ensemble with a ≥0.9 threshold produce the lowest hit rates for all of the data sets while the 0.6:1 ensemble with a ≥0.5 threshold produces the highest hit rate.

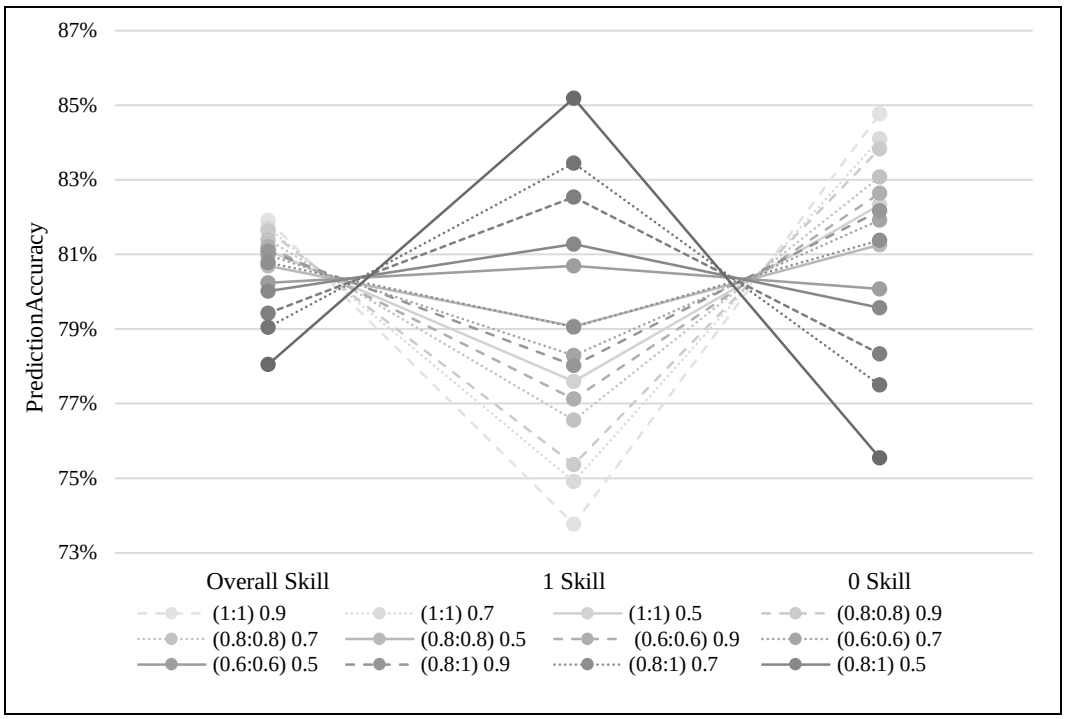

Figure 14: This plot demonstrates the overall, **1** (event), and **0** (non-event) prediction skills for various ensemble forecasts. All forecasts were generated with the top 15 variables for the 1.25° latitude by 2.5° longitude daily data. The legend represent the BRF (event: non-event) with the number to right specifying the ensemble threshold.

As the hit rates increases, the number of false alarms also increase producing an increased FAR. The conservative ensembles with the 0.9 threshold and 1:1 BRF approach, tend to have lower FAR and higher non-event prediction skill but lower hit rates (Figure 15). An example of the relationships between the FAR, F, and Equitable Threat Score for various models are shown in Figure 16. The F is consistently lower than the FAR due to the imbalanced nature of the data. FAR represents the fraction of forecasted events that are false alarms by taking the hits and misses into account while the F takes looks at the fraction of observed non-events that were forecasted as false alarms. The following sections break down the results for each spatial and temporal scale. Plots of the various models created for each scale are provided. The scale of the y-axis varies from plot to plot.

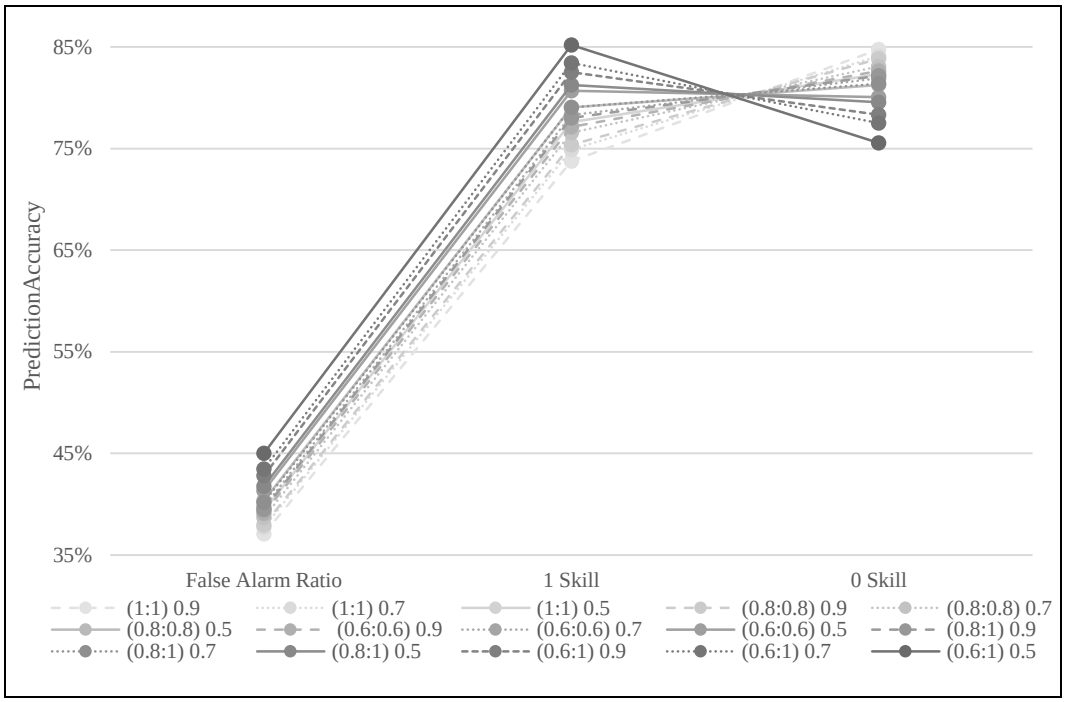

Figure 15: As the hit rate (1 skill) increases, the **0** prediction skill falls and the FAR increases. All forecasts were generated with data from the daily 1.25° latitude by 2.5° longitude 15 variable prediction ensembles. The legend represent the BRF (event: non-event) with the number to right specifying the ensemble threshold.

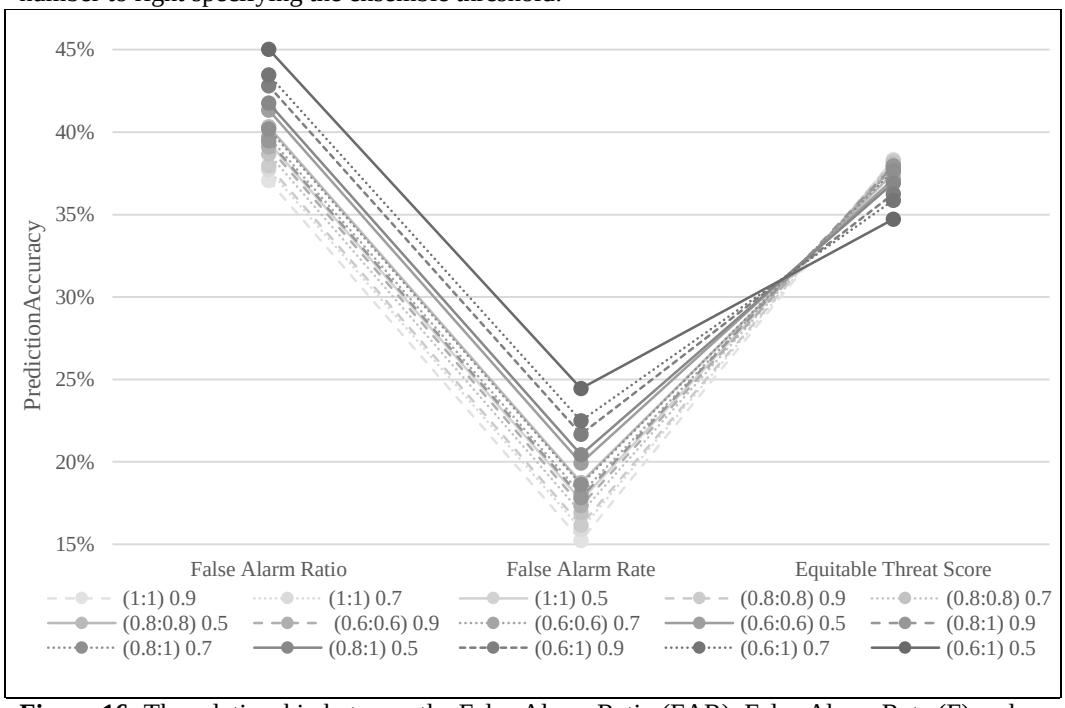

Figure 16: The relationship between the False Alarm Ratio (FAR), False Alarm Rate (F) and Equitable Threat Score (ETS). As the ensemble models are tweaked to maximize hit rate, the number of false alarms increase. This decreases the overall skill as measured by the ETS. All forecasts were generated with data from the daily 1.25° latitude by 2.5° longitude 15 variable prediction ensembles. The legend represent the BRF (event: non-event) with the number to right specifying the ensemble threshold.

3.3.1 2.5° LATITUDE BY 2.5° LONGITUDE

The optimum hit rates for lightning prediction at the 2.5° latitude by 2.5° longitude spatial scale are achieved with the 0.6:1 BRF model approach. The highest ensemble hit rates for both daily and 6-hour lightning prediction are produced by 0.6:1 models generated with the top 15 predictors and an event threshold of ≥ 0.5. The 0.6:1 models created with 12 variables for the 6-hour time scale have nearly equal measure of skill compared to the 15 variable model. The top predictors for each temporal prediction scale are similar with SHOW00, LCLT00, CAPE00, Julian day, and CINS00 showing up in the top 5 variables of the two optimum models chosen for the 2.5° latitude by 2.5° longitude spatial scale.

Daily Lightning Prediction

The hit rates for the 2.5° latitude by 2.5° longitude daily prediction ensemble models built with the top 15 and top 12 variables are shown in Figure 17. The unbalanced *randomForest* ensemble forecasts have a hit rate around 15-20% lower than the BRF models with the same event thresholds. The unbalanced model results highlight the improved event forecast skills achieved by using a BRF method. Since the unbalanced models do not produce skillful event forecasts relative to the BRF models, their results are excluded for the majority of the following spatial scales. The optimum hit rates for daily lightning prediction at the 2.5° latitude by 2.5° longitude scale are achieved with the 0.6:1 BRF models.

The 15 variable 0.6:1 ensemble forecast with an event threshold of 0.5 has the highest hit rate at just over 85% and an overall PC of ~78% (Table 9). The proportion of forecasted events that are correct (PAG) is over 61% while the proportion of observed non-events forecasted as false alarms (F) is ~24.5%. The proportion of incorrectly forecasted events (FAR) is~38.5%. In contrast, the 1:1 balanced model with a 0.5 threshold has a hit rate of ~6.5% lower and a PC ~ 2% lower. The PAG is roughly 5% better while the F and FAR are ~6.5% and 5% lower. The 15 variable 0.6:1 model with an ensemble threshold of ≥ 0.5 is the optimum model for daily lightning event forecasting at the 2.5° latitude by 2.5° longitude. SHOW00, LCLT00, Julian day, CAPE00 and 00Z mean sea level pressure make up the top five variables in order of importance, for the optimum model (Figure 19a).

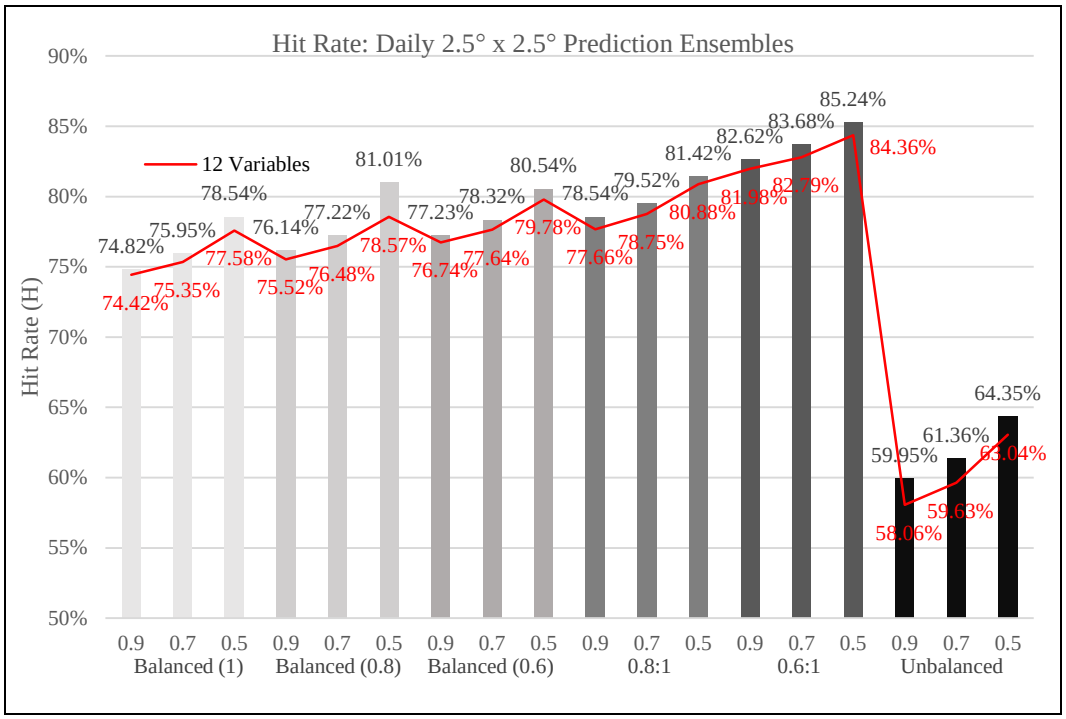

Figure 17: The hit rates (%) for the various 2.5° latitude by 2.5° longitude daily prediction ensemble models. Models built with the top 15 variables are represented by bars, while the red line represents models generated from the top 12 variables. The BRF method applied to each model is shown below the x-axis with the three forecast thresholds (≥ 0.9, ≥ 0.7, and ≥ 0.5) listed on the x-axis. The unbalanced models on the far right were run under default *randomForest* conditions and are included to highlight the increase in skill by implementing the various BRF approaches.

Table 9: Optimum daily ensemble models (0.6:1) produced with 15 variables for the 2.5° latitude by 2.5° longitude scale. The 15 variable balanced models are included for comparison. The three event forecast thresholds of ≥ 0.9, ≥ 0.7, and ≥ 0.5 are shown for each model. As H increases, the FAR and F increase, and the PAG and PC decrease.

2.5x2.5 Daily	Balanced (1)			0.6 : 1		
Skill Measure	0.9	0.7	0.5	0.9	0.7	0.5
Hit Rate (H)	0.7482	0.7595	0.7854	0.8262	0.8368	0.8524
Post Agreement (PAG)	0.6864	0.6789	0.6668	0.6340	0.6295	0.6155
Proportion Correct (PC)	0.8130	0.8111	0.8088	0.7950	0.7934	0.7823
False Alarm Ratio (FAR)	0.3136	0.3211	0.3332	0.3660	0.3705	0.3845
False Alarm Rate (F)	0.1572	0.1651	0.1805	0.2193	0.2265	0.2449

6-hour Lightning Prediction

The hit rates for the 2.5° latitude by 2.5° longitude 6-hour prediction ensemble models built with the top 15 variables and top 12 variables are shown in Figure 18. The optimum hit rates are achieved with the 0.6:1 BRF models. The 0.6:1 ensemble model with 15 variables produces the best forecast model in terms of probability of detection. This model is compared to the 0.6:1 BRF model generated with the top 12 variables in

Table 10. The two models produce very similar results for all measures of skill (within less than 0.5%). The 15 variable 0.6:1 model with an event threshold of 0.5 produces the best forecast in terms of probability of lightning detection (H) with a 84% hit rate. The 12 variable 0.6:1 model with an event threshold of 0.5 produced a comparable forecast skill with a hit rate of 83.5%. The 12 variable 0.6:1 model with an event threshold of 0.5 is chosen as the optimum model as it produces roughly the same results with three fewer variables and is thus more computationally efficient.

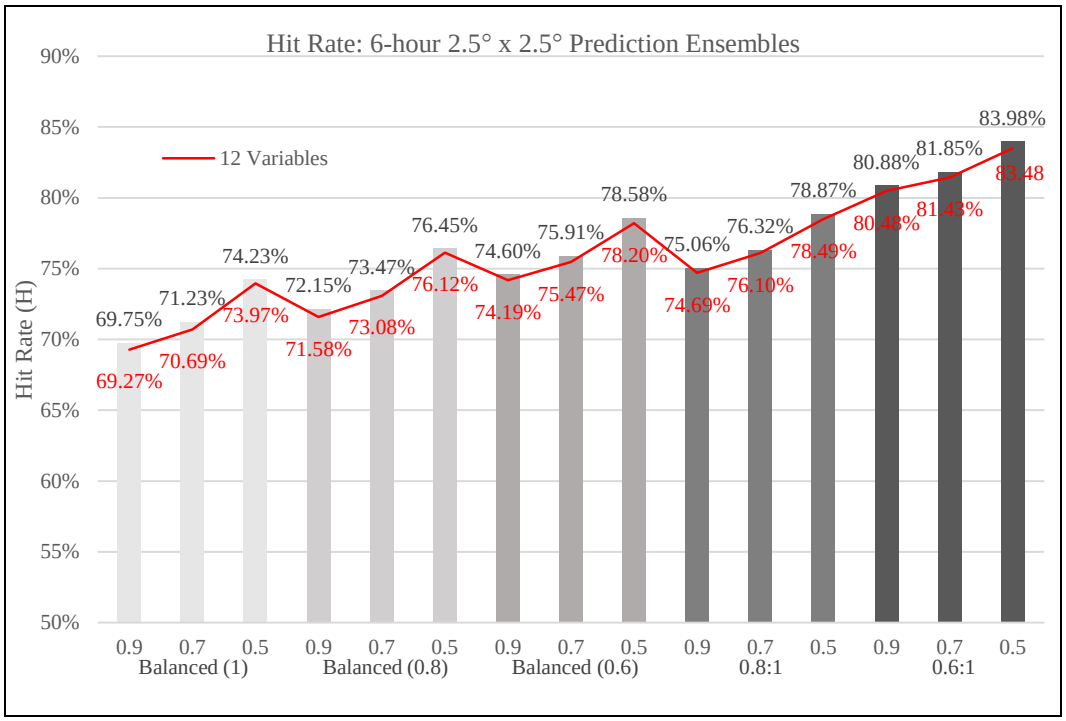

Figure 18: The hit rates for the 2.5° latitude by 2.5° longitude 6-hour prediction ensembles. Models built with the top 15 variables are represented by the bars, while the red line represents the models generated from the top 12 variables. The BRF methods applied to each model are shown below the x-axis with the three forecast thresholds (≥ 0.9, ≥ 0.7, and ≥ 0.5) listed on the x-axis.

Table 10: Comparison of the two top ensemble models (0.6:1) produced with the top 12 and 15 variables for the 2.5° latitude by 2.5° longitude 6-hour forecast. The ensemble variations for the three event forecast thresholds of ≥ 0.9, ≥ 0.7, and ≥ 0.5 are shown for each model.

2.5x2.5 6-hour	**12 Variables 0.6 : 1**			**15 Variables 0.6 : 1**		
Skill Measure	0.9	0.7	0.5	0.9	0.7	0.5
Hit Rate (H)	0.8485	0.8143	0.8348*	0.8088	0.8185	0.8398*
Post Agreement (PAG)	0.3767	0.3705	0.3563	0.3807	0.3737	0.3586
Proportion Correct (PC)	0.7860	0.7800	0.7655	0.7888	0.7823	0.7670
False Alarm Ratio (FAR)	0.6233	0.6295	0.6437	0.6193	0.6263	0.6414
False Alarm Rate (F)	0.2171	0.2256	0.2458	0.2145	0.2236	0.2448

The 12 variable 0.6:1 ensemble forecast with an event threshold of 0.5 has an overall PC of over 76.5%. The proportion of forecasted events that are correct (PAG) is ~36% while the proportion of observed non-events forecasted as false alarms (F) is ~24.6%. The proportion of incorrectly forecasted events (FAR) is~64.4%. The 12 variables and their rank in terms of importance are shown in Figure 19b. The same five variables identified in the optimum 2.5° latitude by 2.5° longitude daily prediction model are present in the top five for the 6-hour model. In order of importance, SHOW00, LCLT00, CAPE00, Julian day, and CINS00 make up the top five variables for 6-hour lightning prediction.

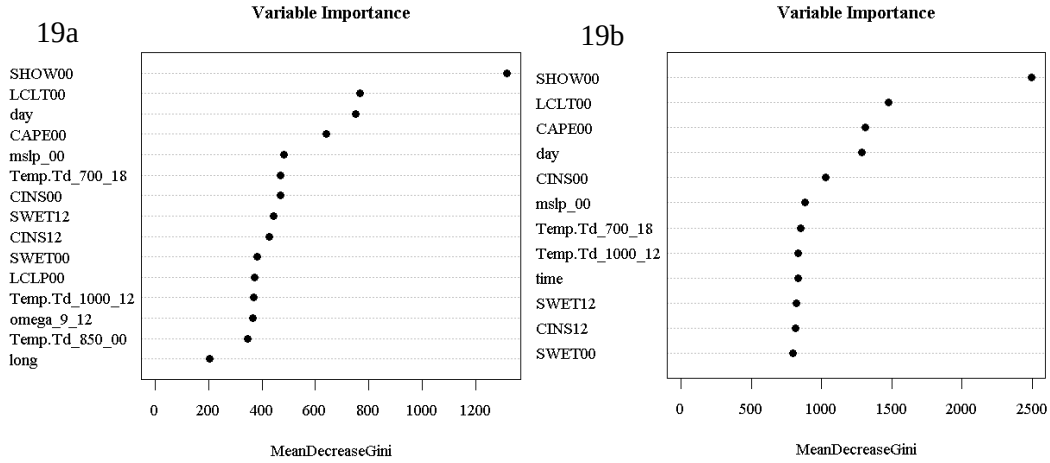

Figure 19: Variable importance plots for the top models selected for the 2.5° latitude by 2.5° longitude spatial scale. The variable importance for the daily prediction model is shown in Figure 19a while the 6-hour model variable importance is shown in Figure 19b. The *MeanDecreaseGini* values on the x-axis are relative values of each variables importance. The values cannot be compared between different models however the relative ranking of variables can be compared. A higher value assigned to a variable indicates it has a greater importance.

3.3.2 1.25° LATITUDE BY 2.5° LONGITUDE

The prediction models for this spatial scale produce similar results to the models generated for the 2.5° latitude by 2.5° longitude spatial scale. Both scales produce better hit rates for daily lightning forecasts than for the 6-hour forecasts. Similar to the 2.5° latitude by 2.5° longitude forecasts, the optimum hit rates for lightning prediction at the 1.25° latitude by 2.5° longitude spatial scale are achieved with the 0.6:1 BRF approach. The top predictors are similar for optimum models selected for each temporal scale.

Daily Lightning Prediction

The hit rates for the daily 1.25° latitude by 2.5° longitude prediction ensemble models built with the top 15 and top 12 variables are shown in Figure 20. The 12 and 15 variable 0.6:1 ensemble models produce similar results with all measures of skill within 1% of each other (Table 11). The 15 variable model produces superior measure of skill for hit rate, PAG, PC, FAR and F compared to the 12 variable model. With a hit rate of ~85%, an overall PC score of 78% and a FAR and F or 45% and 24% respectively, the 15 variable 0.6:1 model with an event threshold of 0.5 is the preferred model for daily lightning prediction at this scale. The models performance is similar to the top model chosen for the daily 2.5° latitude by 2.5° longitude scale. The hit rates, PC, and F skills are nearly equal, for the two models, however, the daily 2.5° latitude by 2.5° longitude model has a better FAR (~38.5) than the 1.25° latitude by 2.5° longitude (~45.0%).

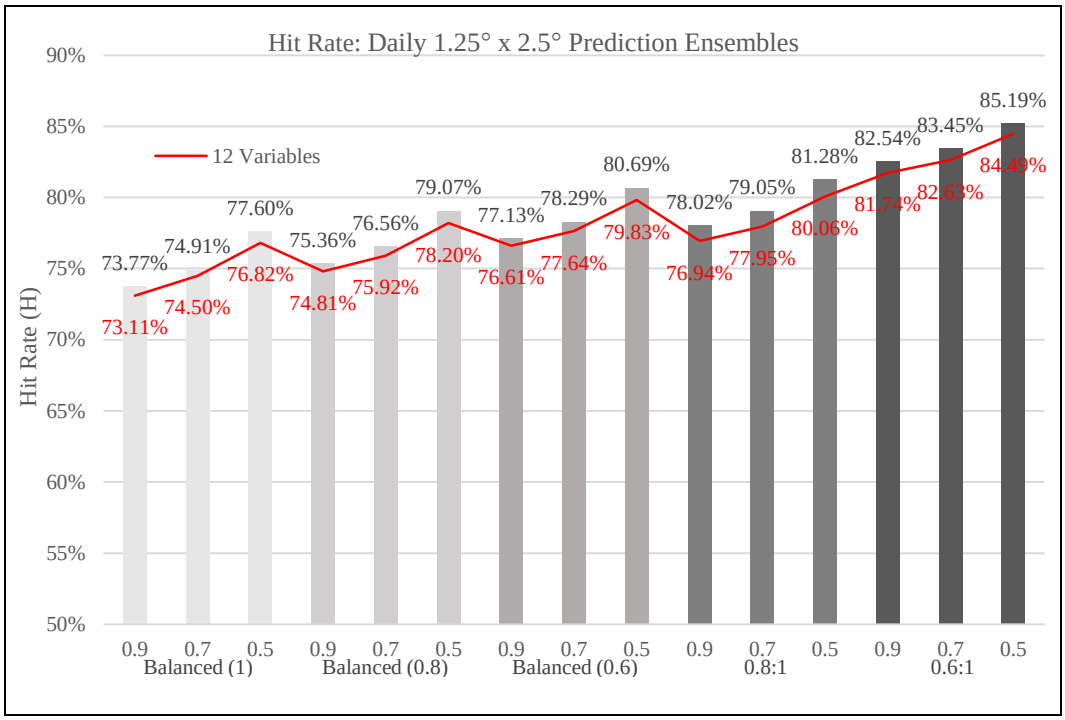

Figure 20: The hit rates for the various 1.25° latitude by 2.5° longitude daily prediction ensembles. Models built with the top 15 variables are represent by the bars, while the red line represents the hit rate of models generated from the top 12 variables. The BRF methods applied to each model are shown below the x-axis with the three forecast thresholds (≥ 0.9, ≥ 0.7, and ≥ 0.5) for each BRF ensemble listed on the axis.

Table 11: Comparison of the top two daily ensembles lightning forecast models for the 1.25° latitude by 2.5° longitude scale. The three event forecast thresholds of ≥ 0.9, ≥ 0.7, and ≥ 0.5 are shown for each model.

1.25x2.5 Daily	12 Variables 0.6 : 1			15 Variables 0.6 : 1		
Skill Measure	0.9	0.7	0.5	0.9	0.7	0.5
Hit Rate (H)	0.8174	0.8263	0.8449	0.8254	0.8345	0.8519
Post Agreement (PAG)	0.5601	0.5538	0.5414	0.5718	0.5652	0.5498
Proportion Correct (PC)	0.7860	0.7821	0.7740	0.7942	0.7905	0.7805
False Alarm Ratio (FAR)	0.4399	0.4462	0.4586	0.4282	0.4348	0.4502
False Alarm Rate (F)	0.2251	0.2333	0.2509	0.2167	0.2250	0.2445

6-hour Lightning Prediction

The unbalanced ensemble forecasts for the 6-hour 1.25° latitude by 2.5° longitude scale generate hit rates of only ~7-12% while all other proposed models generated hit rates above 64% (Figure 21). Similar to the previous models, the 0.6:1 models generate maximum hit rates compared to all other BRF models. The 12 and 15 variable 0.6:1 models produce similar forecast skills with the hit rates typically within 1% of each other. A closer comparison of the two models 0.5 event threshold forecast skill (Table 12) show that the two models have nearly equal hit rates (81.54%, 81.58%). The 12 variable model has a slightly better PAG (29.24%), PC (76.89%), FAR (70.76%), and F (23.67%). Although the 12 variable hit rate is 0.04% less than the 15 variable hit rate, all other measure of skill are slightly better therefore the 12 variables 0.6:1 model with a 0.5 event threshold is chosen as the optimum lightning prediction model for the 6-hour 1.25° latitude by 2.5° longitude scale. The top ranked predictors in this model are similar to those selected for daily lightning prediction (Figure 22).

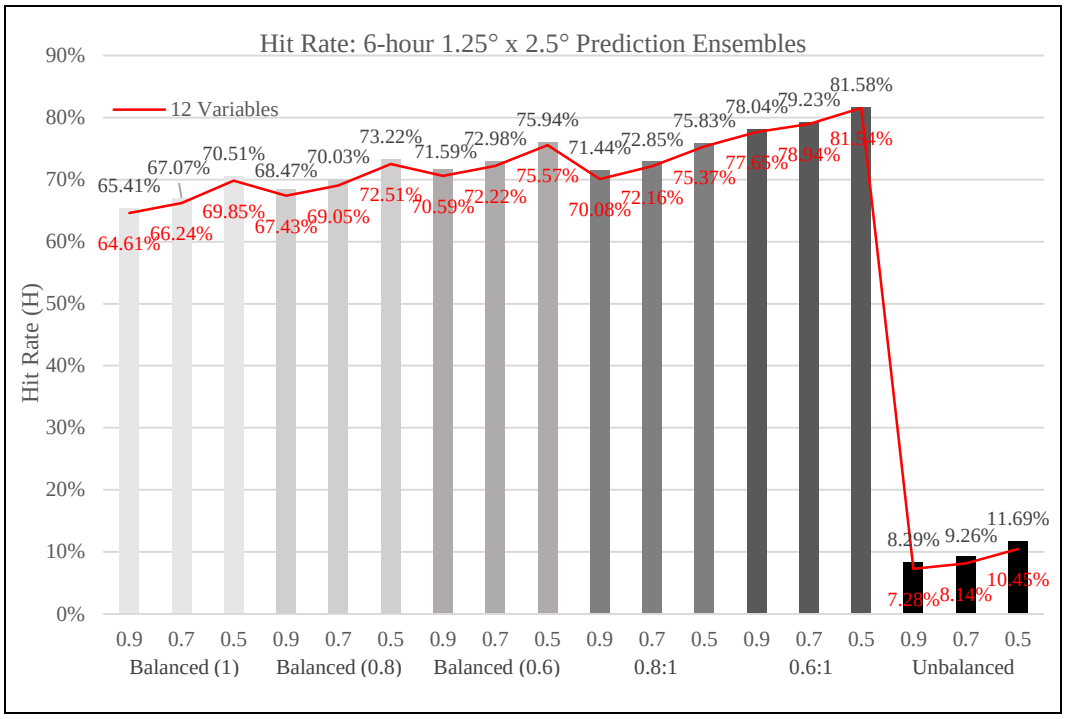

Figure 21: Hit rates of the 1.25° latitude by 2.5° longitude 6-hour prediction ensembles. Models built with the top 15 variables are represented by bars, and the red line represents models generated from the top 12 variables. The BRF methods applied to each model are shown below the x-axis with the three forecast thresholds (≥ 0.9, ≥ 0.7, and ≥ 0.5) listed on the x-axis. The unbalanced models on the far right were run under default *randomForest* conditions.

Table 12: Forecast skill comparison for the top two 6-hour ensembles lightning forecast models for the 1.25° latitude by 2.5° longitude scale. The three event forecast thresholds of ≥ 0.9, ≥ 0.7, and ≥ 0.5 are shown for each model.

1.25x2.5 6-hour	12 Variables 0.6 : 1			15 Variables 0.6 : 1		
Skill Measure	0.9	0.7	0.5	0.9	0.7	0.5
Hit Rate (H)	0.7765	0.7894	0.8154	0.7804	0.7923	0.8158
Post Agreement (PAG)	0.3112	0.3056	0.2924	0.3086	0.3033	0.2911
Proportion Correct (PC)	0.7920	0.7853	0.7689	0.7892	0.7828	0.7675
False Alarm Ratio (FAR)	0.6888	0.6944	0.7076	0.6914	0.6967	0.7089
False Alarm Rate (F)	0.2062	0.2152	0.2367	0.2097	0.2183	0.2383

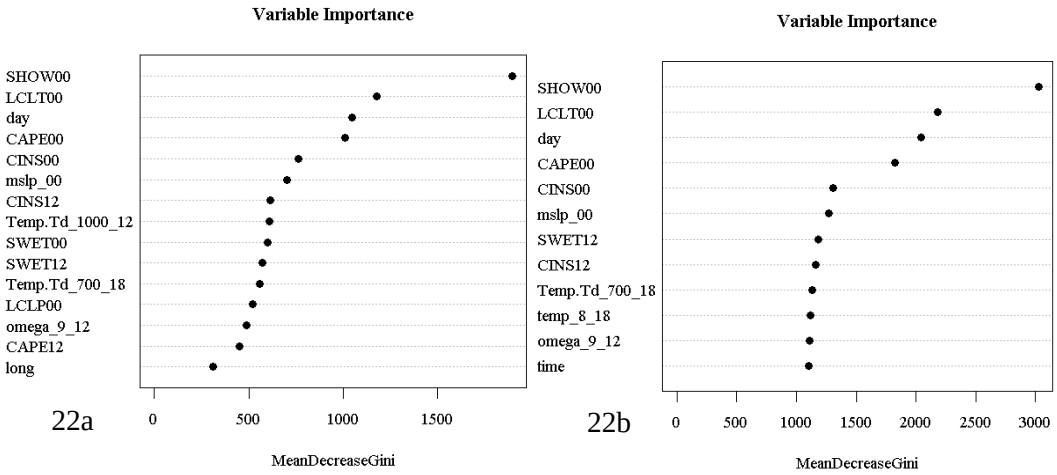

Figure 22: Variable importance plots of the top models selected for the 1.25° latitude by 2.5° longitude spatial scale. The daily prediction model is shown to the left (Figure 22a) while the 6-hour model is shown on the right (Figure 22b). The *MeanDecreaseGini* values on the x-axis are relative values of each variables importance. Actual values cannot be compared between different models however the relative ranking of variables can be compared. A higher value indicates greater importance.

3.3.3 50KM BY 50KM

The 50km by 50km ensemble models predict lightning at a much finer spatial scale. As the spatial resolution increases, so too does the data imbalance between the positive (event) and negative (non-event) classes. The range of hit rates between the various BRF models generated for the 50km by 50km scale is larger than that of the 2.5° latitude by 2.5° longitude and 1.25° latitude by 2.5° longitude scales. For the 50km by 50km prediction models, the province was split into three geographically and ecologically similar zones. The daily and 6-hour prediction results for each of the three zones are provided in the following six sections. The variable importance plots for the various optimum models are provided at the end of the 50km by 50km results in Figure 33.

Boreal Forest : Daily Lightning Prediction

The five BRF models generated with the top 12 and 15 variables for the Boreal Forest zone produced daily ensemble forecasts with hit rates ranging from ~48-72% (Figure 23). The daily Boreal Forest data set is moderately imbalanced with lightning events make up around 14% of the total observations. Similar to the previous two scales, the probability of detection is maximized with the 0.6:1 BRF approach. The skills of the

ensemble forecasts generated with models containing the top 15 variables are quite similar to those generated with the top 12 variables. When determining the top predictors in section **3.2 DETERMINING TOP PREDICTORS**, it was found that the OOB error decreased as the number of variables included in the models were reduced (Figure 13c). The lowest OOB error estimate occurred when only five variables were included in the model. Additional ensemble forcasts were created to explore this trend.

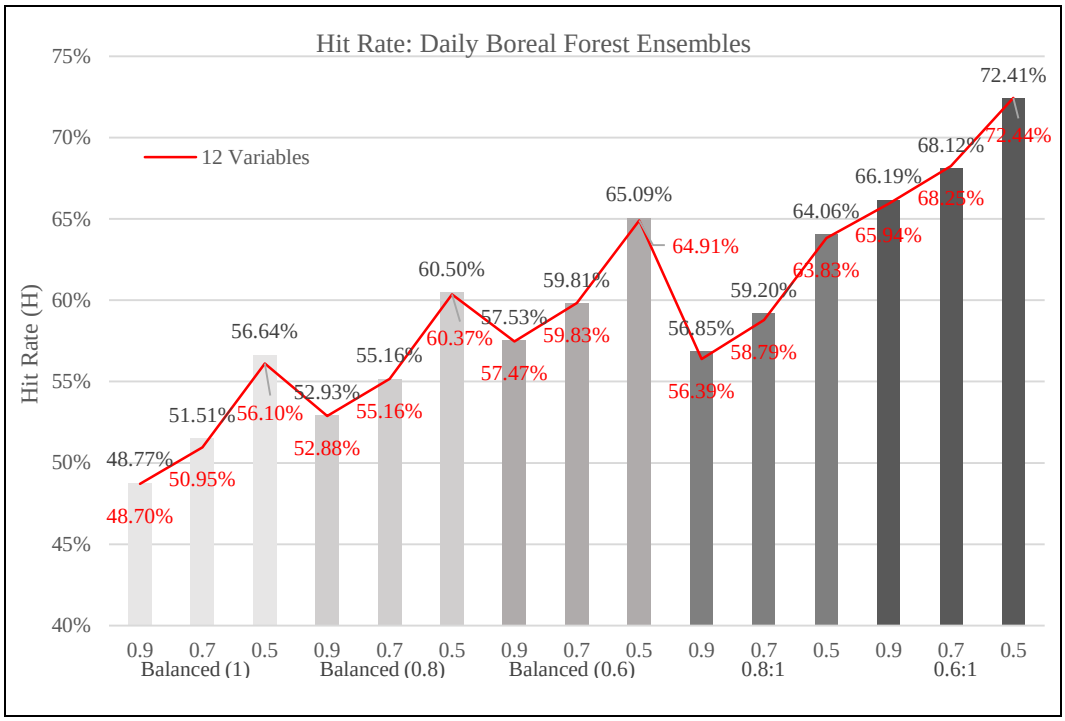

Figure 23: The hit rates for the various Boreal Forest daily prediction ensembles. Models built with the top 15 variables are represented by bars, and the red line represents models generated from the top 12 variables. The BRF methods applied to each model are shown below the x-axis with the three forecast thresholds (≥ 0.9, ≥ 0.7, and ≥ 0.5) listed on the x-axis.

Ensemble models for the top 10, eight, and five variables were created with the same BRF methods and event threshold used for the top 12 and 15 variables. The ensembles created with the models containing the top eight variables produced similar results to those generated from the top 12 and 15 variables. The top 10 and top five variable ensembles produced higher hit rates (~74-78%) than all other models (Figure 24). The five variable 0.6:1 models generated the highest hit rates and PC scores while also having lower FAR and F than the 10, 12 and 15 variables models. The ensemble predictions made with the five variable 0.6:1 models and 0.5 event threshold, have a ~6% higher hit rate, 5% higher PC, 7% lower FAR and a 5% lower F when compared to the

optimum 12 variable model (Table 13). The ensemble created with the 0.6:1 five variable model and a 0.5 threshold is selected as the optimum forecast model.

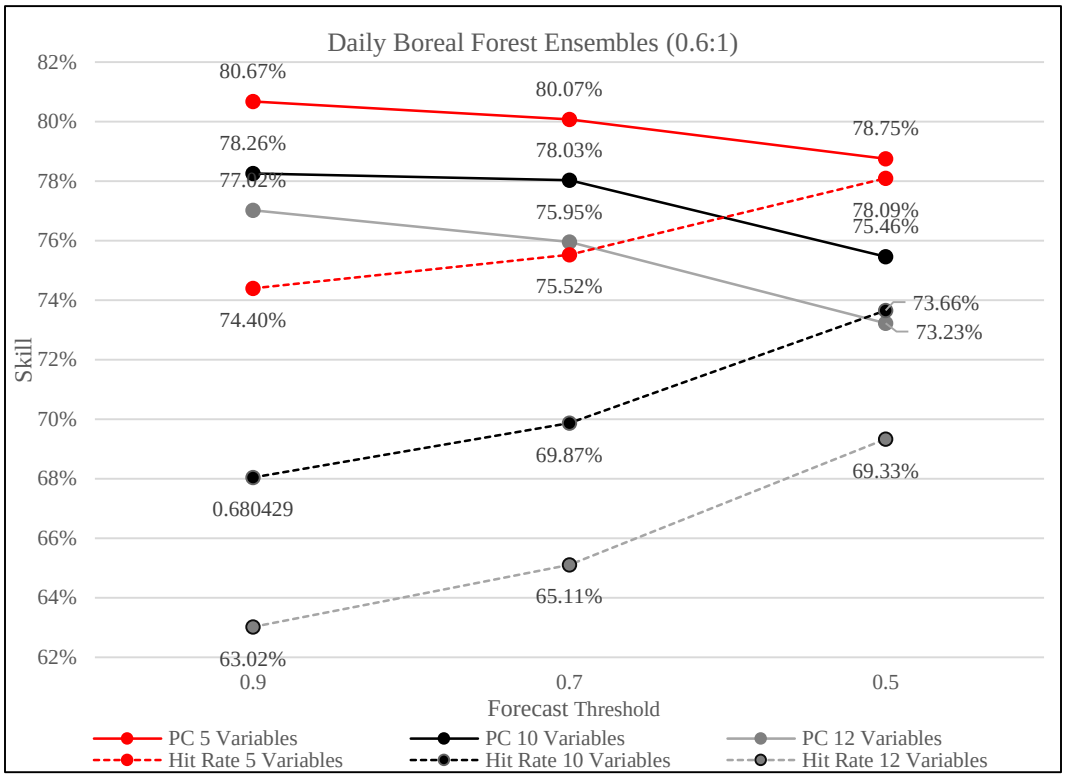

Figure 24: Comparison of the daily ensemble forecast models generated with a 0.6:1 BRF approach for the top 12, 10 and five variables. The hit rates are shown by broken lines while the proportion correct (PC) are given by solid lines.

Table 13: Ensemble forecast skill comparison of the optimum forecast model (five variable 0.6:1) generated for daily lightning prediction in the Boreal Forest. The optimum model is compared to a model variation often selected as a top forecast model for the coarser spatial scales (12 variable 0.6:1). The three event forecast thresholds of ≥ 0.9, ≥ 0.7, and ≥ 0.5 are shown for each model.

Boreal Forest Daily	**12 Variables 0.6 : 1**			**5 Variables 0.6 : 1**		
Skill Measure	0.9	0.7	0.5	0.9	0.7	0.5
Hit Rate (H)	0.6594	0.6825	0.7244	0.7440	0.75522	0.7809
Post Agreement (PAG)	0.3672	0.3573	0.3339	0.4265	0.4183	0.4020
Proportion Correct (PC)	0.7702	0.7595	0.7323	0.8067	0.8007	0.7875
False Alarm Ratio (FAR)	0.6328	0.6427	0.6661	0.5735	0.5817	0.5980
False Alarm Rate (F)	0.2094	0.2262	0.2663	0.1820	0.1910	0.2113

Boreal Forest : 6-hour Lightning Prediction

The data imbalance of the 6-hour Boreal Forest data sets is roughly three times greater than that of the daily Boreal Forest set. Lightning events making up ~5% of the total observations. The five BRF models generated with the top 12 and 15 variables

produce ensemble forecasts with hit rates ranging from ~43-72% (Figure 25). The probability of detection is maximized with the 0.6:1 BRF approach. The 15 variable ensemble have a lower lightning prediction rate than that of the 12 variable models. In section **3.2 DETERMINING TOP PREDICTORS**, the 6-hour Boreal Forest random forests produced the best OOB error estimate when the top 10 or top five variables were included in the model (Figure 13c). Additional ensemble forcasts were created to explore this trend.

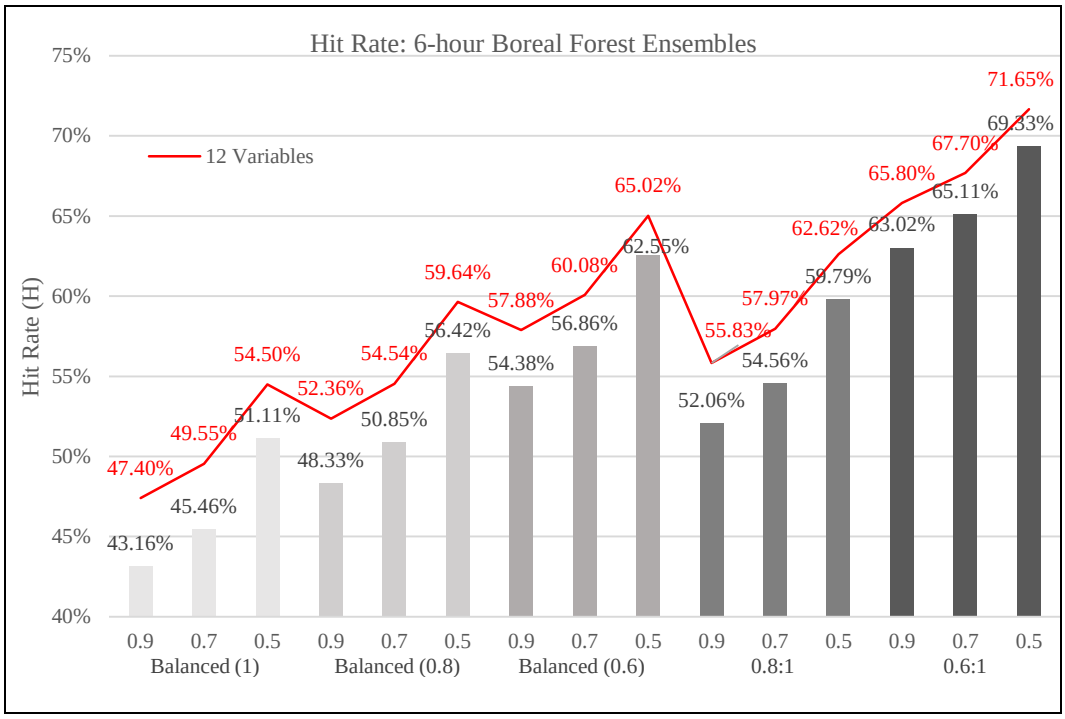

Figure 25: The hit rates for the various Boreal Forest 6-hour prediction ensembles. Models built with the top 15 variables are represented by bars, and the red line represents models generated from the top 12 variables. The BRF methods applied to each model are shown below the x-axis with the three forecast thresholds (≥ 0.9, ≥ 0.7, and ≥ 0.5) listed on the x-axis.

None of the ensemble forecast models generated for the Boreal Forest at a 6-hour time had a PAG greater than 23%. The forecasts predicted false alarm often with the 15 variable balanced (1) BRF producing the lowest FAR of ~77%. The majority of the models were fairly skilled at predicting non-events even with the BRF efforts consistently producing non-event prediction skills of around 80% and higher. The 15 variable balanced (1) model ensemble with a 0.5 threshold had a hit rate of only 51% but was able to accurately forecast non-events with ~89% accuracy.

The ensemble forecasts generated with the 0.6:1 BRF method with the top 12, 10, eight and five variables are compared in Table 14 and the event forecast skills are displayed in Figure 26 . In terms of optimal hit rates, the five variable 0.6:1 model is the most skilled producing hit rates of just over 68% to ~74% (Figure 26). The five variable model is selected as the optimum forecast model when a threshold of 0.5 is implemented. Under these settings, the ensemble predictions have a hit rate of ~74%, roughly 2% higher than the 10 variable ensemble, while the PAG, PC, FAR and F are within less than 1% (Table 14). These top two models exhibit similar skills however the computation efficiency of removing half the variables makes the five variable model more desirable. The top predictors are similar for the optimum models selected for each temporal scale (Figure 33).

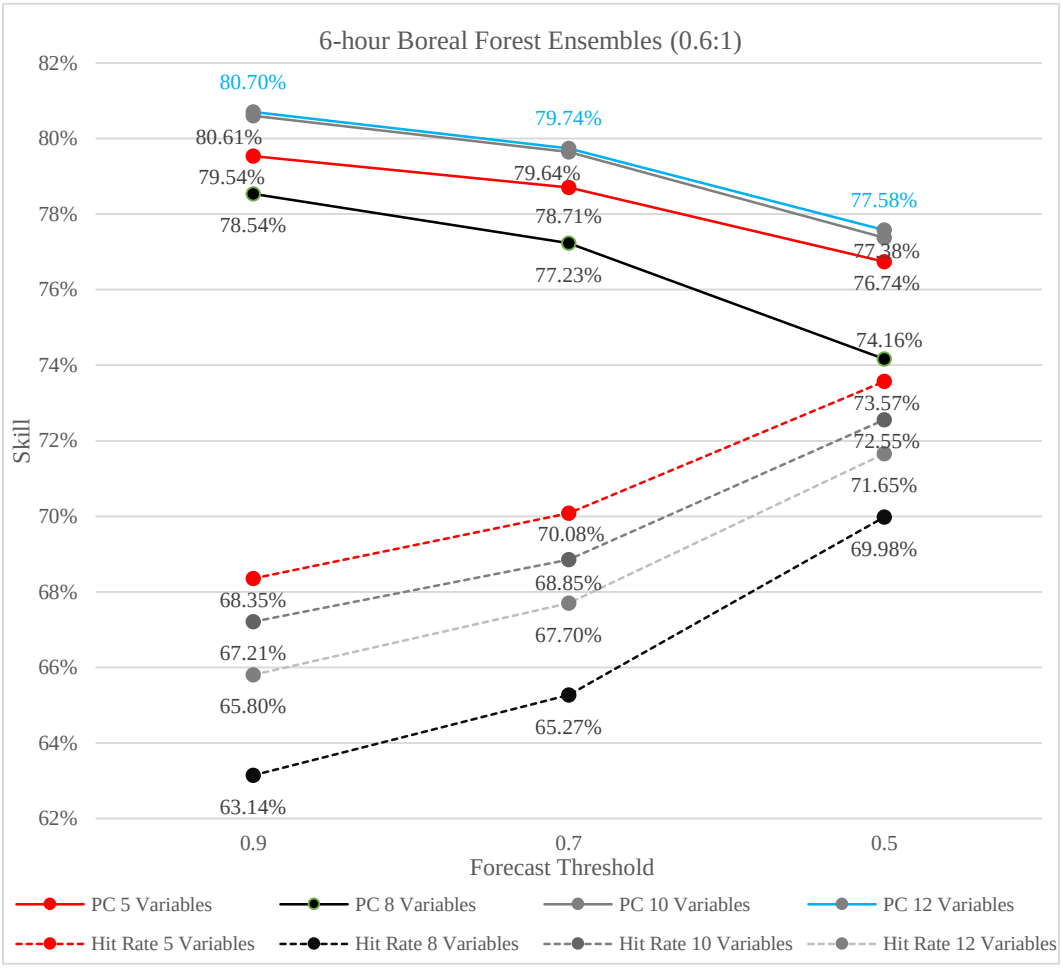

Figure 26: Comparison of the 6-hour ensemble forecast models generated with a 0.6:1 BRF approach for the top 12, 10, eight, and five variables. The hit rates are shown by broken lines while the proportion correct (PC) are given by solid lines.

Table 14: Ensemble forecast skill comparison of the 0.6:1 BRF models generated for 6-hour lightning prediction in the Boreal Forest. Ensemble models created from the top 12, 10, eight and five variables are included. The three event forecast thresholds of ≥ 0.9, ≥ 0.7, and ≥ 0.5 are shown for each model.

Boreal Forest 6-hour	12 Variables 0.6 : 1			10 Variables 0.6 : 1		
Skill Measure	0.9	0.7	0.5	0.9	0.7	0.5
Hit Rate (H)	0.6580	0.6770	0.7165	0.6721	0.6885	0.72554
Post Agreement (PAG)	0.1721	0.1677	0.1589	0.1737	0.1689	0.1590
Proportion Correct (PC)	0.8070	0.7974	0.7758	0.8061	0.7964	0.7738
False Alarm Ratio (FAR)	0.8279	0.8323	0.8411	0.8263	0.8311	0.8410
False Alarm Rate (F)	0.1843	0.1956	0.2208	0.1861	0.1973	0.2234
	8 Variables 0.6 : 1			**5 Variables 0.6 : 1**		
	0.9	0.7	0.5	0.9	0.7	0.5
Hit Rate (H)	0.6314	0.6527	0.6998	0.6835	0.7008	0.7357
Post Agreement (PAG)	0.1516	0.1469	0.1373	0.1673	0.1370	0.1565
Proportion Correct (PC)	0.7854	0.7723	0.7416	0.7954	0.7871	0.7674
False Alarm Ratio (FAR)	0.8484	0.8531	0.8627	0.8327	0.8360	0.8435
False Alarm Rate (F)	0.2056	0.2207	0.2559	0.1981	0.2079	0.2308

Parkland and Grassland : Daily Lightning Prediction

The daily Parkland and Grassland data set has a similar data imbalance to that of the Boreal Forest with lightning events making up around 13% of the total observations. The five BRF models generated with the top 12 and 15 variables produced daily ensemble forecasts with a hit rates varying between ~55-78% (Figure 27). The ensemble forecasts generated with the top 15 variables are less skillful at predicting lightning occurrence than the 12 variable models (Figure 27). When determining the top predictors for the Parkland and Grassland zone by examining the changes in OOB error estimate, it was found that as the number of variables decreases, so too does the OOB error estimate. The random forest with the top 10 variables included had the lowest OOB error estimate (Figure 13d). Additional ensemble forcasts were created to explore this trend.

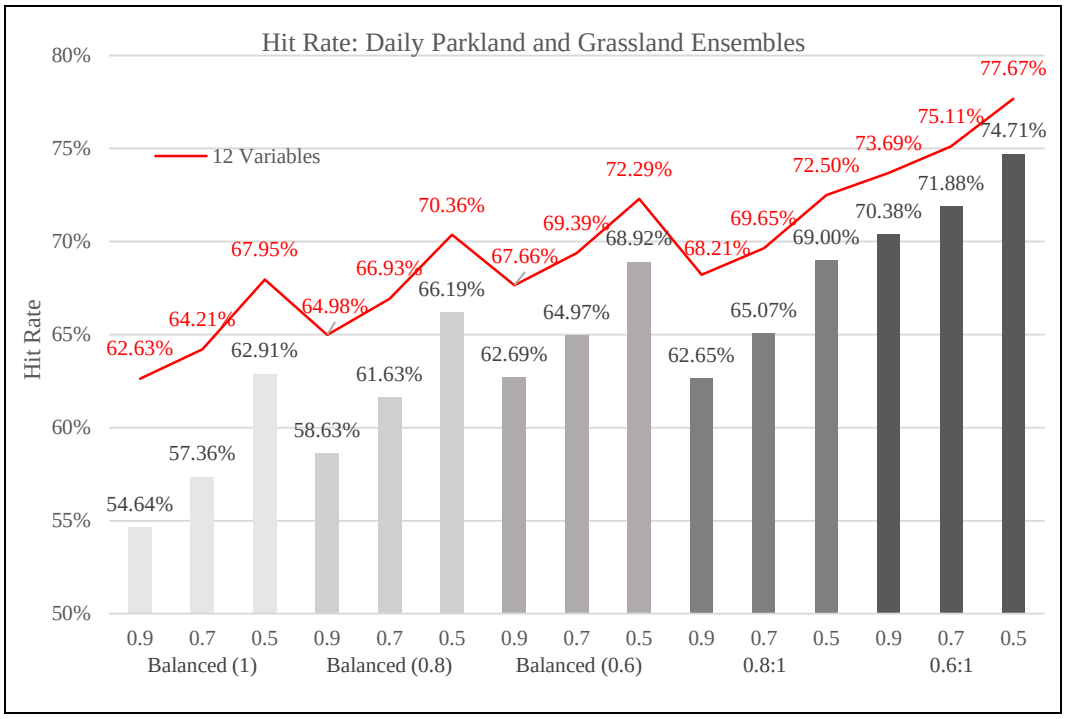

Figure 27: The hit rates for the various Parkland and Grassland daily prediction ensembles. Models built with the top 15 variables are represented by bars, and the red line represents models generated from the top 12 variables. The BRF methods applied to each model are shown below the x-axis with the three forecast thresholds (≥ 0.9, ≥ 0.7, and ≥ 0.5) listed on the x-axis.

Ensemble predictions were made with the 0.6:1 BRF method for models including the top 10, eight and five variable. The newly generated ensemble forecasts produced superior hit rates compared to the 12 and 15 variable models (Figure 28). As the number of input variables decreases, the hit rate and FAR rise (Table 15). The PAG skill for the daily Parkland and Grassland ensembles are between 36% and 43%. The F rates are fairly low for the 12 to five variable ensembles (~15-23%) while the FAR are quite high (~57-64%). The ensemble forecasts based on the five variable model with an event threshold of 0.5 produces a maximum hit rate of 86.4%. An overall PC of ~79% is achieved however the PAG is only 36% for this ensemble. The five variable model was selected as the optimal model.

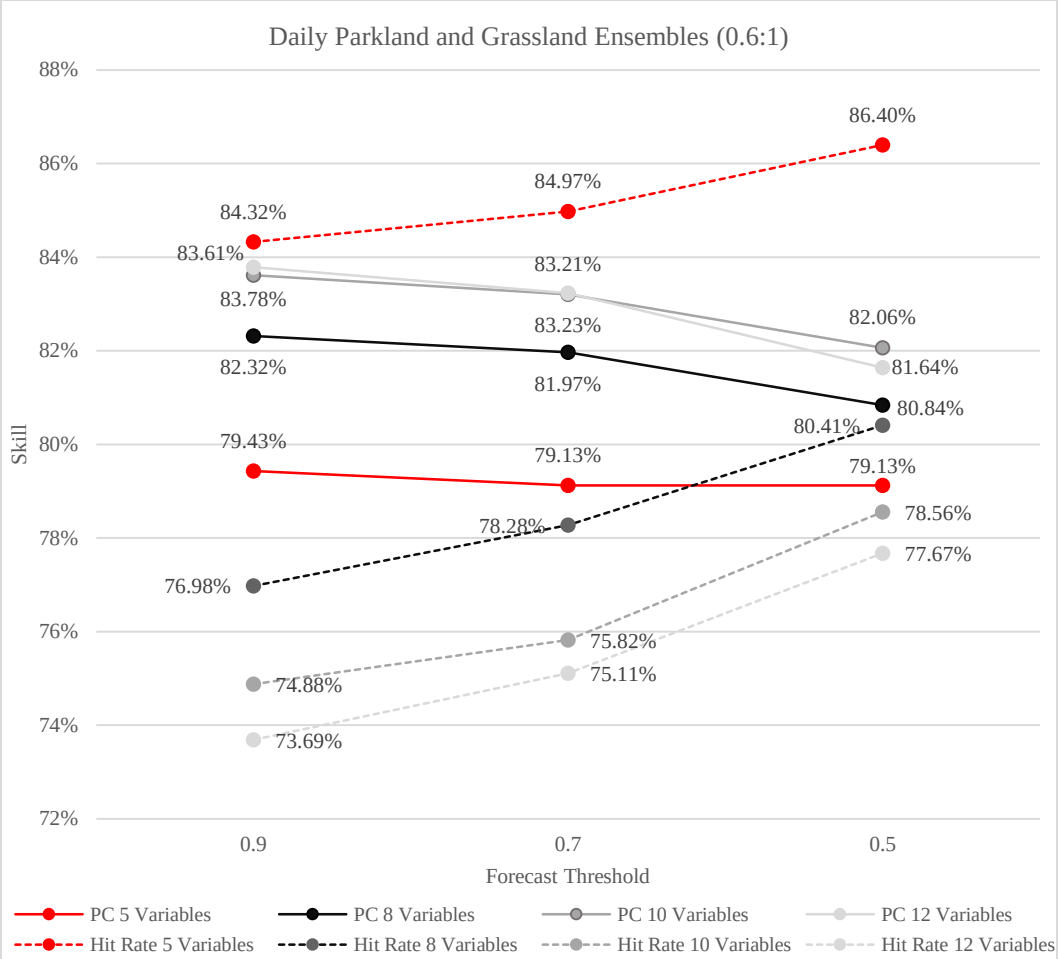

Figure 28: Comparison of the daily ensemble forecast models generated with a 0.6:1 BRF approach for the top 12, 10, eight, and five variables. The hit rates are shown by broken lines while the proportion correct (PC) are given by solid lines.

Table 15: Comparison of the daily Parkland and Grassland ensemble forecasts generated from the 0.6:1 BRF models for the top 12, 10, eight and five variables. The three event forecast thresholds of ≥ 0.9, ≥ 0.7, and ≥ 0.5 are shown for each model.

Parkland Daily Skill Measure	12 Variables 0.6 : 1			10 Variables 0.6 : 1		
	0.9	0.7	0.5	0.9	0.7	0.5
Hit Rate (H)	0.7369	0.7511	0.7767	0.7488	0.7582	0.7856
Post Agreement (PAG)	0.4312	0.4221	0.3979	0.4290	0.4223	0.4055
Proportion Correct (PC)	0.8378	0.8323	0.8164	0.8361	0.8321	0.8206
False Alarm Ratio (FAR)	0.5688	0.5779	0.6021	0.5710	0.5777	0.5945
False Alarm Rate (F)	0.1469	0.1554	0.1777	0.1507	0.1568	0.1741
	8 Variables Balanced (1)			**5 Variables 0.6 : 1**		
	0.9	0.7	0.5	0.9	0.7	0.5
Hit Rate (H)	0.7698	0.7828	0.8041	0.8432	0.8497	0.8640
Post Agreement (PAG)	0.4081	0.4037	0.3889	0.3736	0.3705	0.3629
Proportion Correct (PC)	0.8232	0.8197	0.8084	0.7943	0.7913	0.7913
False Alarm Ratio (FAR)	0.5919	0.5963	0.6111	0.6264	0.6295	0.6371
False Alarm Rate (F)	0.1687	0.1748	0.1910	0.2131	0.2175	0.2286

Parkland and Grassland : 6-hour Lightning Prediction

The data imbalance for the 6-hour Parkland and Grassland data sets is roughly four times greater than that of the daily lightning occurrence data set. Lightning events making up less than~4% of the total observations. The five BRF models generated with the top 12 and 15 variables produce ensemble forecasts with hit rates ranging from ~39-70% (Figure 29). The 15 variable models ensemble prediction skills are less than that of predictions made with the 12 variables models. Due to the increased skill with the 12 variable models and the trend in Figure 13d, an additional set of ensemble predictions were run for the top 10, eight and five variables.

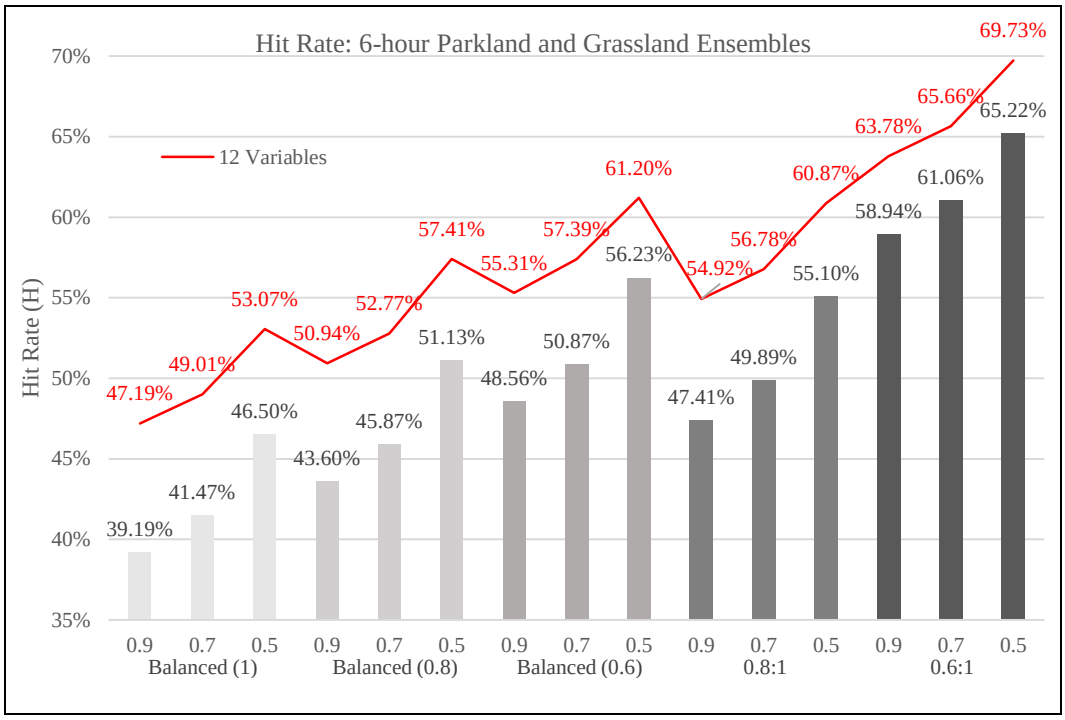

Figure 29: The hit rates for the various Parkland and Grassland 6-hour prediction ensembles. Models built with the top 15 variables are represented by bars, and the red line represents models generated from the top 12 variables. The BRF methods applied to each model are shown below the x-axis with the three forecast thresholds (≥ 0.9, ≥ 0.7, and ≥ 0.5) listed on the x-axis.

For the 10 variable 0.6:1 ensemble, the hit rates of 62%, 64%, and 69% for the thresholds of 0.9, 0.7 and 0.5 respectively, were higher than the 15 variable hit rates but lower than those generated with the top 12 variables (Table 16). The FAR and F were also higher when only 10 variables were included. Similar trends were found for models generated with the top eight variables. The hit rate is highest when only five variables are

included in the models, however the PAG is low (less than 16%). The PAG is below ~24% for all of the models and well below 20% for the 0.6:1 BRF ensembles. The five variable 0.6:1 model with an event threshold of 0.5 produces the highest hit rate (78.7%). The PC is more than 2% lower than the comparable 12 variable models, yet the FAR is within less than 0.5%. Again, the 5 variable model is chosen due to its comparable skill and decreased computational expense.

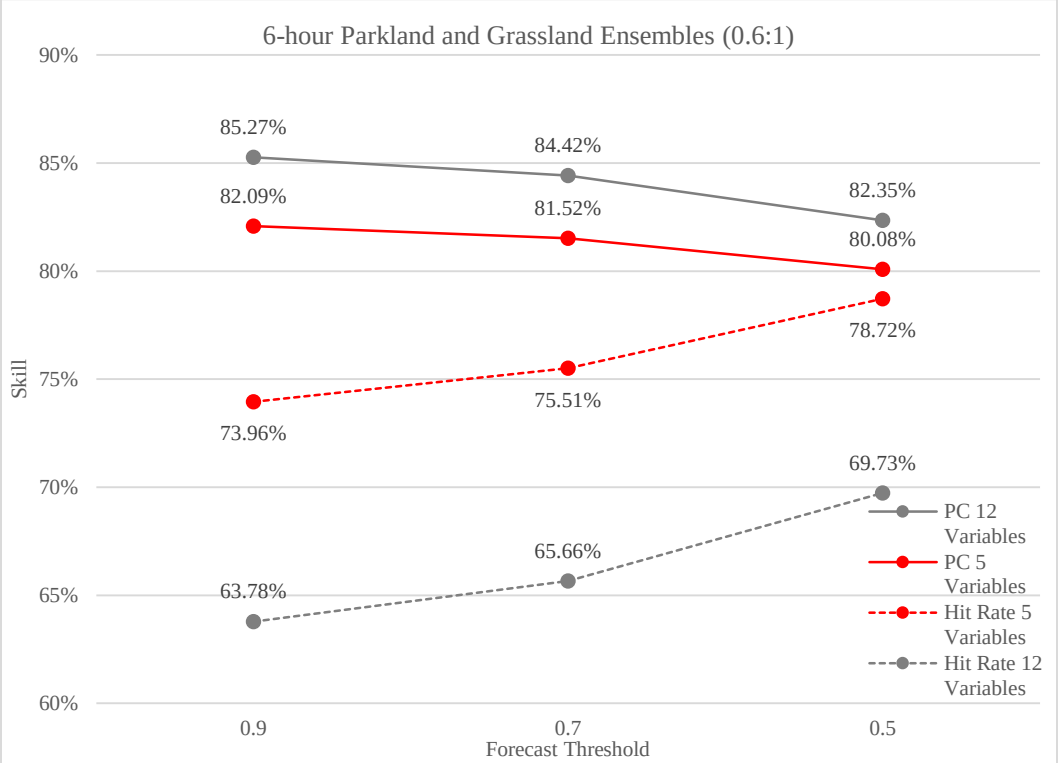

Figure 30: Comparison of the 6-hour ensemble forecast models generated with a 0.6:1 BRF approach for the top 12 and top 5 variables. The hit rates are shown by broken lines while the proportion correct (PC) are given by solid lines.

Table 16: Comparison of the 6-hour Parkland and Grassland ensemble forecasts generated from the 0.6:1 BRF models for the top 12, 10, eight and five variables. The three event forecast thresholds of ≥ 0.9, ≥ 0.7, and ≥ 0.5 are shown for each model.

Skill Measure	12 Variables 0.6 : 1			10 Variables 0.6 : 1			5 Variables 0.6 : 1		
	0.9	0.7	0.5	0.9	0.7	0.5	0.9	0.7	0.5
H	0.6378	0.6566	0.6973	0.6209	0.6426	0.6888	0.7396	0.7551	0.7872
PAG	0.1797	0.1739	0.1615	0.1531	0.1492	0.1408	0.1659	0.1635	0.1574
PC	0.8527	0.8442	0.8235	0.8284	0.8190	0.7969	0.8209	0.8152	0.8008
FAR	0.8203	0.8261	0.8385	0.8469	0.8508	0.8592	0.8341	0.8365	0.8426
F	0.1372	0.1470	0.1706	0.1618	0.1727	0.1980	0.1752	0.1820	0.1985

Mountain and Foothills : Daily Lightning Prediction

The daily Mountain and Foothills data set has the lowest data imbalance of all the 50km by 50km zones. Roughly 17% of the total observations have lightning occurrence. The five BRF models generated with the top 12 and 15 variables produce ensemble forecasts with hit rates ranging from ~70% to over 85% (Figure 31). The prediction skills of the15 variable model ensembles are less than the skills of the 12 variable models (Table 17). Due to the relatively high prediction skill, low FAR, and trend in Figure 13e, no additional models were run for the Mountain and Foothills zone.

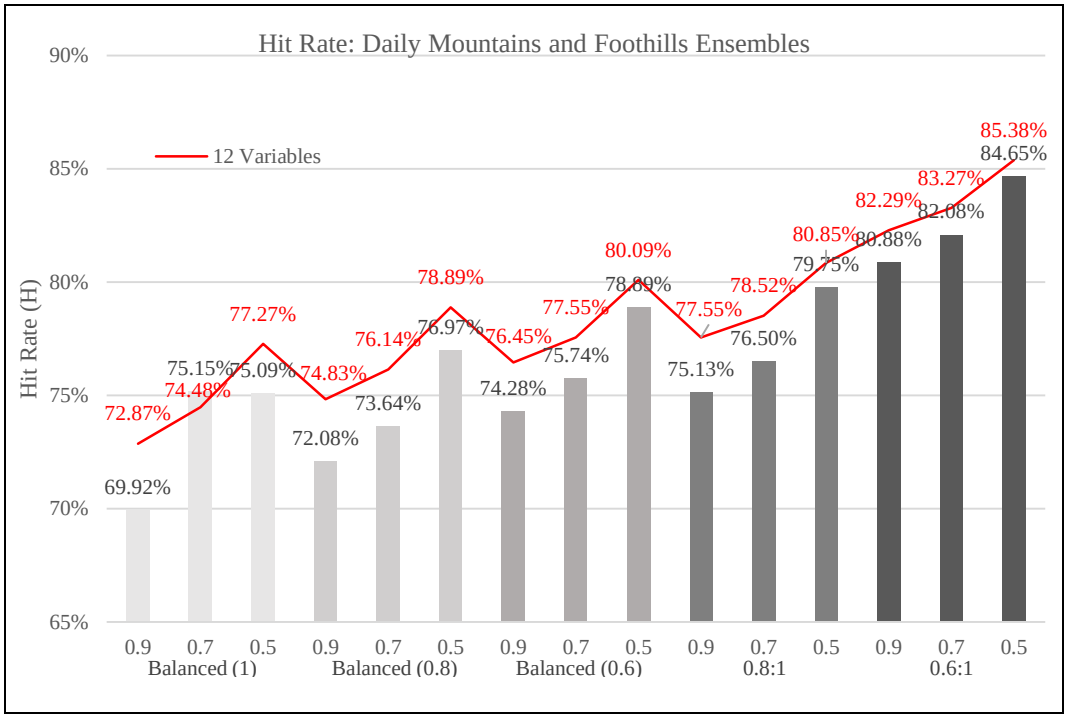

Figure 31: The hit rates for the various Mountain and Foothills daily prediction ensembles. Models built with the top 15 variables are represented by bars, and the red line represents models generated from the top 12 variables. The BRF methods applied to each model are shown below the x-axis with the three forecast thresholds (≥ 0.9, ≥ 0.7, and ≥ 0.5) listed on the x-axis.

Overall, the 15 and 12 variable ensembles are quite comparable in terms of skill. The 15 variable 0.6:1 ensemble produced slightly lower hit rates but higher PC than the corresponding 12 variable models. The FAR is also lower for the 15 variable ensembles although it is within less than 2% of the 12 variable ensembles. The F skills are within less than 2% of each other with the 15 variable model having a slightly lower number of false alarms. The two models show some of the highest hit rates (~82-85%) and lowest FAR (~50-53%) compared to the other 50km by 50km models. The 12 variable 0.6:1 model with an event threshold of 0.5 is chosen as the optimum model as it produces

roughly the same results with three fewer variables and is thus more computationally efficient.

Table 17: Comparison of the Mountain and Foothills ensemble forecasts generated from the 0.6:1 BRF models for the top 12 and 15 variables. The three event forecast thresholds of ≥ 0.9, ≥ 0.7, and ≥ 0.5 are shown for each model.

Mountains Daily Skill Measure	12 Variables 0.6 : 1			15 Variables (0.6:1)		
	0.9	0.7	0.5	0.9	0.7	0.5
Hit Rate (H)	0.8229	0.8327	0.8538	0.8088	0.8208	0.8465
Post Agreement (PAG)	0.4826	0.4769	0.4655	0.5013	0.4950	0.4814
Proportion Correct (PC)	0.8152	0.8115	0.8035	0.8263	0.8227	0.8142
False Alarm Ratio (FAR)	0.5174	0.5231	0.5345	0.4987	0.5050	0.5186
False Alarm Rate (F)	0.1864	0.1930	0.2071	0.1700	0.1770	0.1927

Mountain and Foothills : 6-hour Lightning Prediction

The data imbalance for the 6-hour Mountain and Foothills data sets is around three times greater than the imbalance for daily lightning occurrence. Roughly 6% of the total observations in the data set have lightning occurrence. The BRF models generated with the top 12 and top 15 variables produce ensemble forecasts with hit rates ranging from ~63-82% (Figure 32). The 15 variable models ensemble hit rate skills are less than

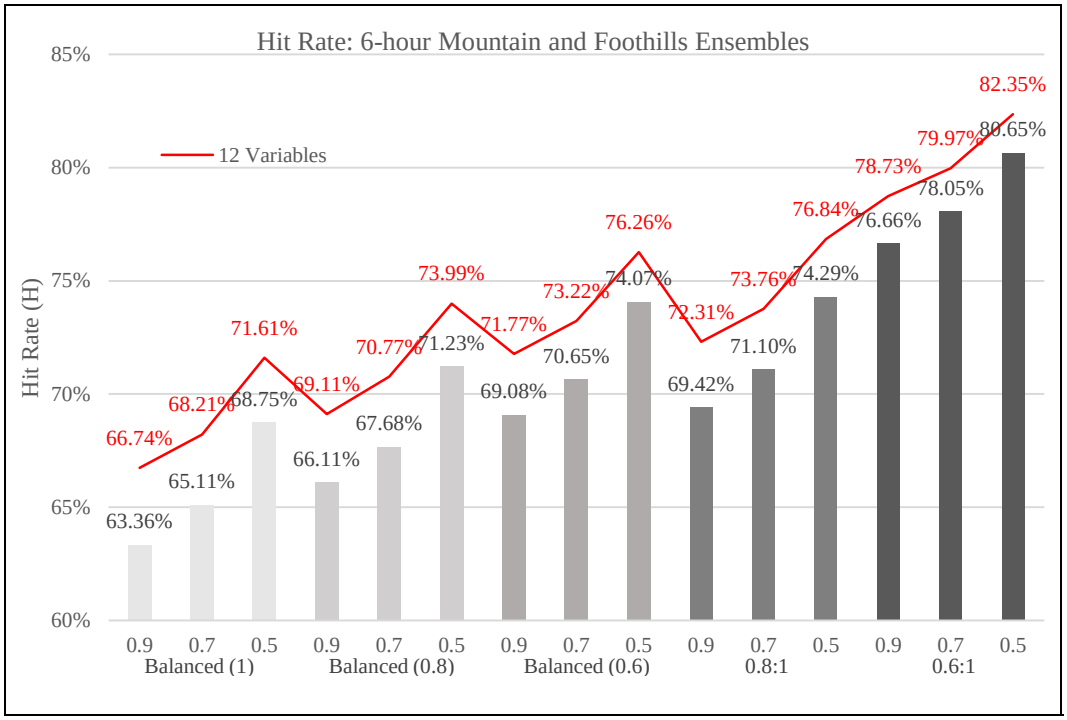

Figure 32: The hit rates for the various Mountain and Foothills 6-hour prediction ensembles. Models built with the top 15 variables are represented by bars, and the red line represents models generated from the top 12 variables. The BRF methods applied to each model are shown below the x-axis with the three forecast thresholds (≥ 0.9, ≥ 0.7, and ≥ 0.5) listed on the x-axis.

that of predictions made with the 12 variable models. Due to the relatively high hit rate, trend in Figure 13e, and the already relatively high FAR of the 12 and 15 variable models, no additional models were run for the Mountain and Foothills zone.

Overall, the 15 and 12 variable ensembles are quite comparable in terms of skill. The 12 variable 0.6:1 ensembles produced hit rates of ~79-82% compared to the ~77-81% of the 15 variable models (Table 18). The 12 variable ensemble with a 0.5 threshold also produced nearly equal PAG and PC skills. The FAR is slightly lower for the 12 variable 0.5 threshold ensemble although it is within ~0.1% of the 15 variable ensemble. The F skills are within less than 0.3% of each other with the 15 variable model having a slightly lower value. The 12 variable 0.6:1 model with an event threshold of 0.5 is chosen as the optimum model as it produces comparable measures of skill and a slightly better hit rate (80.7%) with three fewer variables and is thus more computationally efficient.

Table 18: Comparison of the Mountain and Foothills ensemble forecasts generated from the 0.6:1 BRF models for the top 12 and 15 variables. The three event forecast thresholds of ≥ 0.9, ≥ 0.7, and ≥ 0.5 are shown for each model.

Mountains 6-hour	12 Variables 0.6 : 1			15 Variables (0.6:1)		
Skill Measure	0.9	0.7	0.5	0.9	0.7	0.5
Hit Rate (H)	0.7873	0.7997	0.8235	0.7666	0.7805	0.8065
Post Agreement (PAG)	0.2529	0.2482	0.2352	0.2588	0.2508	0.2341
Proportion Correct (PC)	0.8359	0.8305	0.8155	0.8403	0.8351	0.8169
False Alarm Ratio (FAR)	0.7471	0.7518	0.7648	0.7412	0.7492	0.7659
False Alarm Rate (F)	0.1607	0.1674	0.1850	0.1517	0.1612	0.1824

The variable importance plots for the top models selected for each 50km by 50km zones are illustrated in Figure 33. The variable importance plots for the daily and 6-hour optimum models were nearly identical, therefore only the daily plot was included. SHOW00, CAPE00, and Julian day show up in nearly all the top models for every spatial scale. SHOW00, latitude (lat), and Julian day were in the top five predictors for all of the 50km by 50km models. CAPE00 appears in the top five variables for both Boreal Forest models, the 6-hour Parkland and Grassland, and the 6-hour Mountain and Foothills models. Elevation (elv) and longitude (long) are highly important for the daily Mountain and Foothills models. Sample forecast skill maps are also provided for each the three zones (Figure 34, Figure 35, and Figure 36).

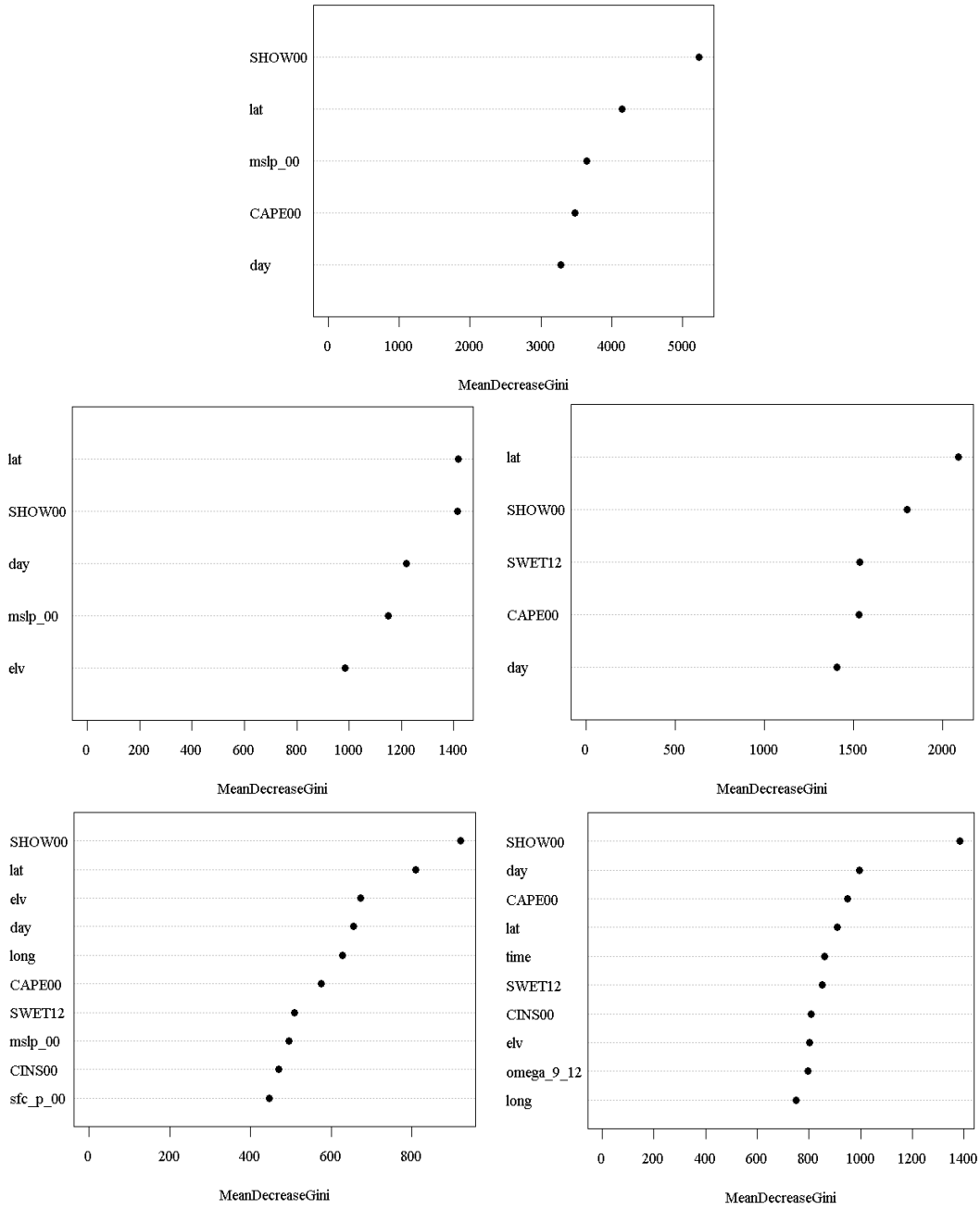

Figure 33: Variable importance plots for the optimum models selected for each spatial and temporal scale. The models represent the Boreal Forest, Parkland and Grassland, and Mountain and Foothills from top to bottom. The daily models variable importance are displayed on the left and the 6-hour models to the right. The plots were nearly identical for the daily and 6-hour Boreal Forest therefore only the daily plot was included. The *MeanDecreaseGini* values on the x-axis are relative values of each variables importance. Actual values cannot be compared between different models however the relative ranking of variables can be compared. A higher value assigned to a variable indicates it has greater importance.

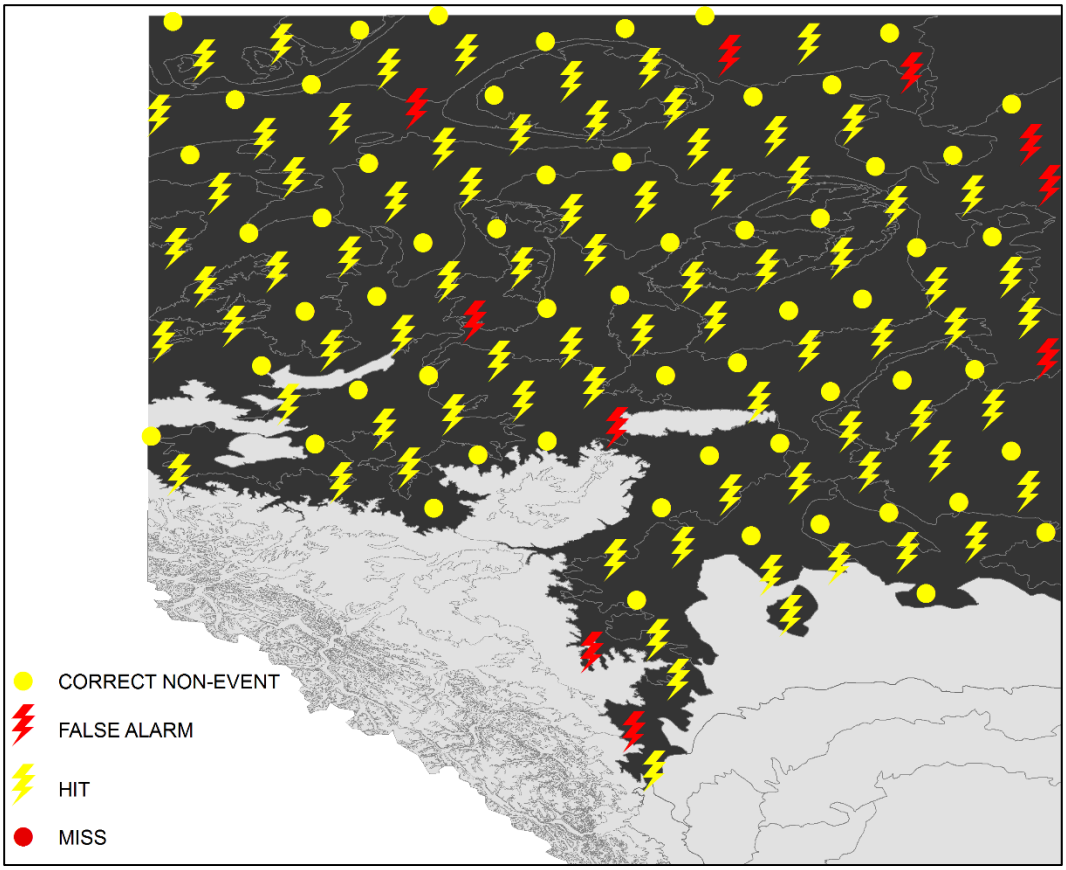

Figure 34: Ensemble forecast prediction accuracy for a randomly chosen day with lightning (July 13, 2005). The optimum Boreal Forest model was used to generate the ensemble forecast.

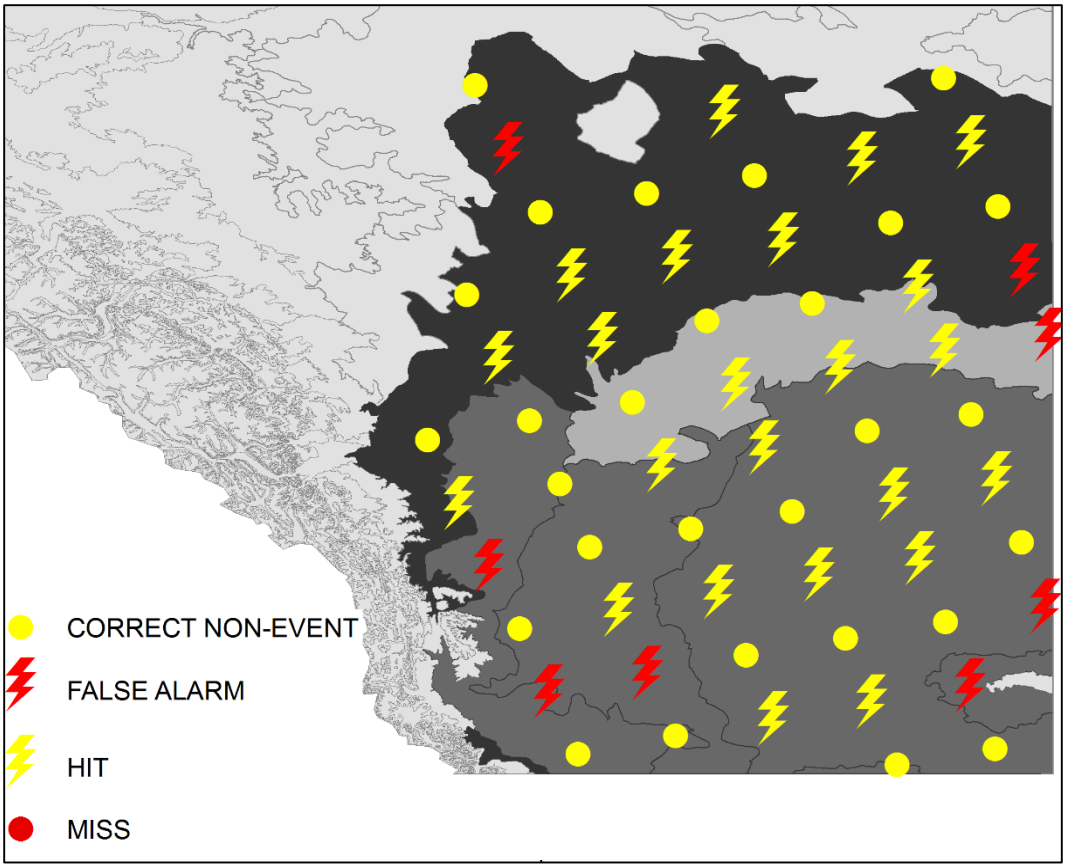

Figure 35: Ensemble forecast prediction accuracy for a randomly chosen day with lightning (July 23, 2011). The optimum Grassland and Parkland model was used to generate the ensemble forecast.

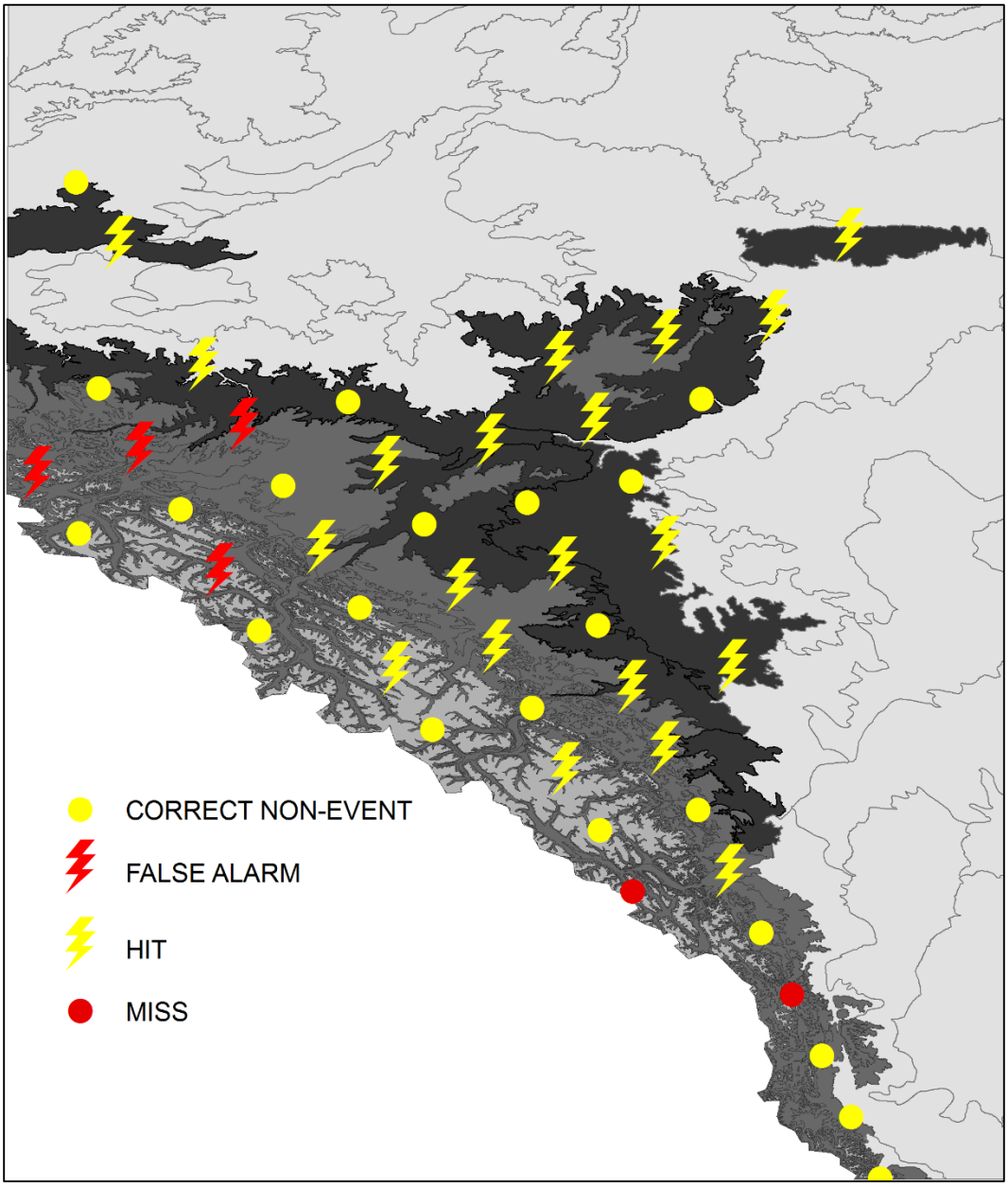

Figure 36: Ensemble forecast prediction accuracy for a randomly chosen day with lightning (July 19, 2005). The optimum Mountain and Foothills model was used to generate the ensemble forecast.

CHAPTER 4. DISCUSSION

The top ten models created through implementation of the random forest classification algorithm are compared and discussed followed by an in-depth look at variable importance and the predictor-predictand relationships. Random forest modeling is then discussed in greater detail as the approach is fairly new to the fields of meteorology and wildland fire science. The advantages and disadvantages of the modeling approach are discussed along with the various modifications implemented. To conclude the chapter, possible sources of model error are discussed followed by some suggestions for future research.

4.1 TOP MODEL SELECTION

The top 10 models selected were compared and rank from one to10 in order of their various measures of skill (Table 19). All of the top models selected were generated with the 0.6:1 BRF approach and interpreted with a 0.5 event threshold. Of the top 10 models selected, the daily prediction models for the Parkland and Grassland, Mountain and Foothills, 2.5° latitude by 2.5° longitude, and 1.25° latitude by 2.5° longitude have the top four hits rates of 86.40%, 85.38%, 85.24% and 85.19% respectively. The lowest FAR, and therefore best PAG skills, are achieved by the daily models for the 2.5° latitude by 2.5° longitude, 1.25° latitude by 2.5° longitude, Mountain and Foothills and Boreal Forest. The range of PC is small with a maximum value of 81.55% for the 6-hour Mountain and Foothills and the lowest value of 76.55% for the 6-hour 1.25° latitude by 2.5° longitude model.

The ETS were calculated for each ensemble. There was little variation between the models ETS therefore this measure didn't provide much insight for picking an optimum model. It should be noted that all ETS calculated were positive. This indicates that none of the models generated were "unskilled". A model is "unskilled" if the chance forecast is preferred to the generated forecast. The 2.5° latitude by 2.5° longitude, and 1.25° latitude by 2.5° longitude models had ETS values of ~35% while the 50km by 50km has values around 15-18%.

A PAG of less than 50% indicates that the model forecasts false alarms more often than hits. The daily 2.5° latitude by 2.5° longitude, and 1.25° latitude by 2.5° longitude, are the only two models that generate a PAG greater than 50%. This is partially due to selecting top models based on optimal hit rates, as hit rates are often

maximized at the expense of the false alarms. The 0.6:1 BRF approach contributes to the low PAG skill as the sample size specification create a model that is biased to the positive class. Although not to the same extent, the 0.6:1 BRF approach essentially flips the data imbalance such that the majority class sample is now the positive class. The daily Mountain and Foothills (47%) and daily Boreal Forest (~40%) forecasts have the highest PAG out of the 50km by 50km models.

From a wildland fire perspective, accurate lightning prediction is most valuable in the Boreal Forest and part of the Mountain and Foothills zones. Unfortunately the 6-hour Boreal Forest ensemble has one of the worst forecast performances in terms of hit rate, PC and FAR. The daily Boreal Forest model was ranked 4/10 for PAG, FAR and F and 5/10 for PC but ranked 9/10 for hit rate. The 6-hour Mountain and Foothills model ranked 1st for F and PC, 8/10 for PAG and FAR, and 6/10 for hit rate. The daily Mountain and Foothills model has the best overall performance ranking 2/10 for hit rate and PC and 3/10 for all other measures. This ensemble model was generated with the top 12 variables and a 0.6:1 BRF approach. Four of the top variables from the Mountain and Foothills model include Julian day, elevation, and SHOW00, and CAPE00. The individual contribution of these four variables to the random forest model are interpreted from the partial dependence plots and discussed in 4.2.1 PARTIAL DEPENDENCE PLOTS.

With minor modifications for data input discontinuities (such as scale differences), and a small amount of training, these models could easily be introduced into fire management and operations. The majority of the variables in the final model selections are geographic covariates or from the Radiosonde data. The radiosonde data are available almost immediately after the sounding. A different source is necessary to replace the reanalysis variables as these are not immediately available. The Reanalysis variables are mean sea level pressure, 250mb vertical winds and surface pressure. These variables could be obtained from weather station data, forecasts, numerical weather prediction models, or calculated from the soundings.

Table 19: The model selected for each of the data sets are ranked from in order of decreasing skill from one to 10 for various forecast skill criteria.

Rank	Hit Rate (H)	Post-Agreement (PAG)	Proportion Correct (PC)	False Alarm Ratio (FAR)	False Alarm Rate (F)
1	Daily Parkland	Daily 2.5°x2.5°	6-hr Mountain	Daily 2.5°x2.5°	6-hr Mountain
2	Daily Mountain	Daily 1.25°x2.5°	Daily Mountain	Daily 1.25°x2.5°	6-hr Parkland
3	Daily 2.5°x2.5°	Daily Mountain	6-hr Parkland	Daily Mountain	Daily Mountain
4	Daily 1.25°x2.5°	Daily Boreal	Daily Parkland	Daily Boreal	Daily Boreal
5	6-hr 2.5°x2.5°	Daily Parkland	Daily Boreal	Daily Parkland	Daily Parkland
6	6-hr Mountain	6-hr 2.5°x2.5°	Daily 2.5°x2.5°	6-hr 2.5°x2.5°	6-hr Boreal
7	6-hr 1.25°x2.5°	6-hr 1.25°x2.5°	Daily 1.25x2.5°	6-hr 1.25°x2.5°	6-hr 1.25°x2.5°
8	6-hr Parkland	6-hr Mountain	6-hr 1.25°x2.5°	6-hr Mountain	Daily 1.25°x2.5°
9	Daily Boreal	6-hr Parkland	6-hr Boreal	6-hr Parkland	Daily 2.5°x2.5°
10	6-hr Boreal	6-hr Boreal	6-hr 2.5°x2.5°	6-hr Boreal	6-hr 2.5°x2.5°

4.2 PREDICTOR IMPORTANCE AND CONTRIBUTION

Variable importance plots for the top 10 models selected are shown in sections: 3.3.1 2.5° LATITUDE BY 2.5° LONGITUDE, 3.3.2 1.25° LATITUDE BY 2.5° LONGITUDE, and 3.3.3 50KM BY 50KM . The daily Mountain and Foothills model has the best overall performance of the 10 models and is thus selected as the optimum model produced. This section discusses the variable importance and predictor-predictand relationship with respect to this specific model unless otherwise stated.

4.2.1 PARTIAL DEPENDENCE PLOTS

Partial dependence plots can be used to help understand the contribution a variable makes to a given random forest model (Hastie et al., 2009). The partial dependence plots provide a visualization of the relationship between a selected individual predictor variable and the binary (**0/1**) output from the random forest. In essence, the plots show the average trend of a selected variable integrating all other variables out (Liaw and Wiener, 2002). By default, the plots focus on the first class (0); therefore the *which.class* argument was used to generate the plots with a focus on lightning events (1). The y-axis on the partial dependence plots provided is the logit of the probability of

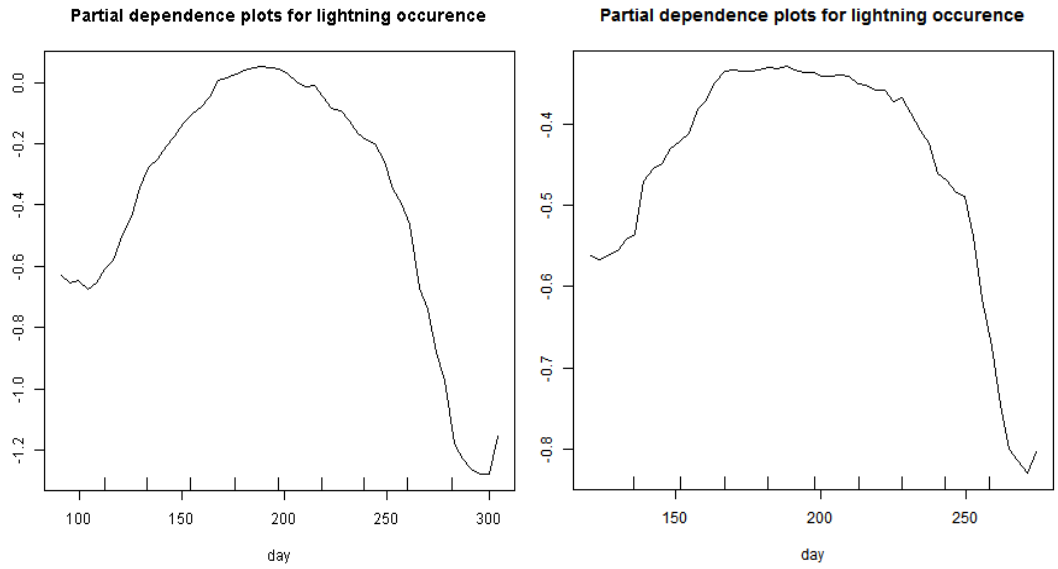

Figure 37: Partial dependence plots for Julian day for daily lightning prediction in the Mountain and Foothills (left) and daily lightning prediction at the 2.5° latitude by 2.5° longitude spatial scale (right). The y-axes are the logit of the probability of the lightning occurrence.

lightning occurrence. The ticks inside the x-axis show the deciles[8] of the data distribution for the corresponding variables (Hastie et al., 2009). A lower density of ticks in a certain region of a plot means that the data density is low in that range; therefore the curves are less determined (Hastie et al., 2009). While the numbers on the y-axis vary from plot to plot, the general trends can be compared between variables by comparing the shape and range of the plots.

Figure 37 shows the partial dependence plots generated for Julian day for daily lightning prediction in the Mountain and Foothills and daily lightning prediction at the 2.5° latitude by 2.5° longitude spatial scale. The plots have a unimodal distribution with the logit of predicted probability of lightning being highest when Julian day is ~150-230, and lowest for the days occurring at the tail ends (April and October).This trend was anticipated and clearly shows the seasonal variation with peak lightning occurrence in the months of June, July and August. Peak lightning occurrence during these months in Alberta is well documented (Burrows et al., 2002). Both plots show similar trends however the trend is more clearly defined in the daily lightning 2.5° latitude by 2.5°

[8] A variation of quantile, the deciles are the nine values that split the data into ten equal parts.

longitude plot. This is partially due to the larger spatial scale encompassing all of Alberta and thus having more unique thunderstorm events over a larger range of days, as shown by the x-axis and decile tick marks.

The partial dependence plots for SHOW00, and CAPE00 are shown in Figure 38. The 00Z Showalter index has a fairly linear negative relationship with the logit of the probability of lightning occurrence. Lower SHOW00 values correspond to a higher probability of lightning. This trend was anticipated as a negative SHOW value indicates that the air in the upper planetary boundary layers is unstable when compared to the middle troposphere therefore convection may occur. A positive Showalter index indicates stable air while values below zero indicate increasing instability as the value decreases (Showalter, 1947). Burrows et al., (2005) also found the Showalter index to be a good predictor of lightning, and identified it as the top overall predictor for Canada and the northern United States. The Showalter Index is also recognized as an important instability measure for severe storms and tornadic activity (Dupilka and Reuter, 2011).

The CAPE00 partial dependence plot has a fairly linear positive relationship from zero to ~600J/kg at which point the curve plateaus. The positive relationship with the logit of the probability of lightning occurrence indicates that as higher values of CAPE00 correspond to higher probabilities of lightning occurrence. Again, this trend was anticipated as CAPE is a measure of the convective potential of the lower atmosphere. CAPE values less than 1000J/kg indicate a relatively stable atmosphere while values in excess of 3000J/kg indicate sufficient energy for severe thunderstorms. The partial dependence plot doesn't seem to capture the upper end variations in CAPE. This could be due to the low number of observations available in the upper range. The thin-plate spline (TPS) interpolation of CAPE may have also failed to capture the solar heating of the south facing slopes of the mountains. The slopes facing the sun heat at a faster rate than the relatively flat areas to the east due to the angle of inclination. As the warm air rises into cooler ambient air, it has a positive buoyancy. This process may encourage strong updrafts that may become sufficient to support storm activity.

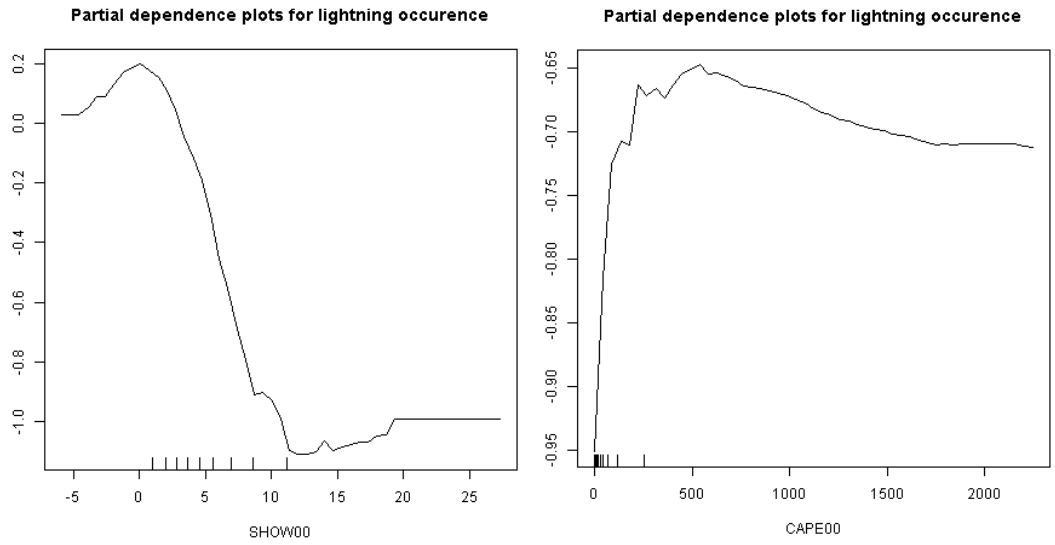

Figure 38: Partial dependence plots of SHOW00 (left) and CAPE00 (right) for daily lightning occurrence in the Mountain and Foothills. The y-axes are the logit of the probability of the lightning occurrence.

Elevation was selected as a top predictor for the both of the Mountain and Foothills models. Elevation (center of cell) was only included as an input at the 50km by 50km scale. The two other scales are too coarse for center point or average elevation to provide valuable information. The partial dependence plot of elevation is provided in Figure 39. Burrows et al. (2002) identified areas of high flash density with elevation of ~1000m along the eastern slopes of the foothills. A band of high flash density was also identified between the 1200m and 2000m contour lines from west of Brazeau County to north of Canmore. These findings correspond fairly well with the partial dependence plot. The sharp rise and peak at around 2800m in the partial dependence plots was not anticipated. The increase at ~2800m occurs well outside of the deciles indicating there were not a large number of events.

Upon exploring the possibility that a small number of lightning events may have been captured around a single mountain (or a few neighbouring peaks), it was found that all strikes records with an elevation above 2500m occurred in a single 50km by 50km spatial cell. The cell had a center point location of 82.57916°N and 117.2654°W. These coordinates are located in Jasper National Park on the NE side of Cornucopia Peak. Looking back at the training data, it was found that lightning strikes occur in this specific bin between ~20 to 30 times per year. Cornucopia Peak is located roughly 5km to the north east of Mt. Brazeau and is within the band of high flash density identified by Burrows et al. (2002). When creating decisions trees for warm season lightning

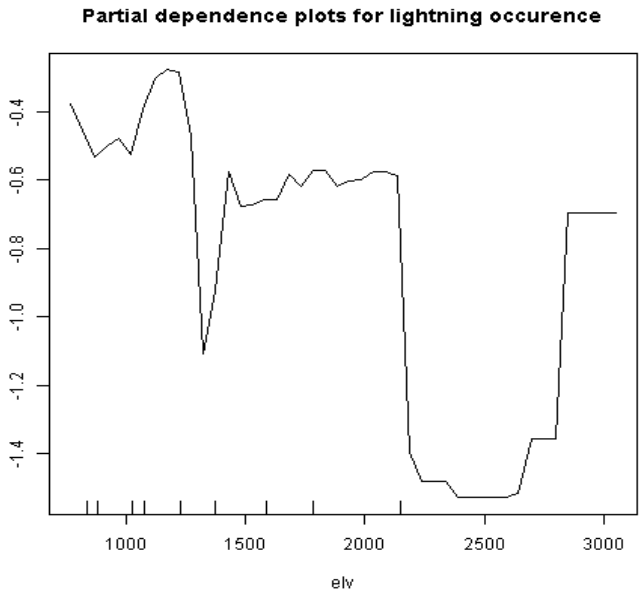

Figure 39: Partial dependence plot of elevation for daily lightning occurrence in the Mountains and Foothills. The y-axis is the logit of the probability of the lightning occurrence.

prediction Burrows et al. (2005) discussed that while elevation ranked fairly low overall in their models, it was a highly important variable for locations with significant elevation and gradients such as the Mountain and Foothills zones.

Partial plots were not generated for the remaining top variables as the trends became very difficult to interpret and were obscured by noise. This is a typical response as only the highly relevant predictors for each models are likely to produce informative partial dependence plots (Hastie et al., 2009). In addition, variables with an additive effect also provide the clearest plots (Hastie et al., 2009). The remaining variables include latitude, longitude, SWET12, mslp, CINS00 and surface pressure at 00Z. Mean sea level pressure and Severe Weather Threat index were also identified as top predictors ranking second and fourth overall by Burrows et al. (2005)

The mslp and surface pressure at 00Z would likely provide information to the model about the location of low pressure systems. A decrease change in surface pressure can occur due to changing in the mass of air in the vertical column above a given point or by causing an air mass near the surface to rise to fall. As explained section **1.3 CONVECTIVE BASICS,** an air mass that is warmer than the surrounding environment will become buoyant and rise. This can occur due to solar heating and warm air advection (Aguado and Burt, 2010). The contributions of these variables to the model cannot be discussed with certainty, however, the general relationship between lower surface pressure and convective activity is well known. The severe weather threat index is used to analyze thunderstorm (and thus lightning) potential therefore it may be considered intuitive that it is selected by the models. It is postulated that this variable would have a positive relationship with probability of occurrence.

Convective inhibition is the amount of energy in J/kg needed to overcome the negative buoyant energy exerted by the environment on an air parcel. Although CINS was ranked fairly low by Burrows et al. (2005), its importance in the Mountain and Foothills models is not unexpected as CINS00 showed in the top 15 predictors for all of the spatial scales (Table 7). Again, the exact impact of CINS00 on the model cannot be stated however it is proposed that CINS had a positive relationship with the probability of lightning. Given as a negative value, CINS of zero to -50J/kg represent a weak capping effect, -50 to -200J/kg represent a moderate capping effect, and values lower than -200J/kg represents a strong capping force, and thus likely a stable lower atmosphere.

Latitude and longitude were also selected by the Mountain and Foothills models. While the trends were unclear in the partial plots, there are documented variations in lightning occurrence with both variables (Burrows and Kochtubajda, 2010). Elevation and longitude can provide proxy information to the models about the slope and aspect as the cells move east from the mountains into the leeward Foothills. Latitude and Julian day help capture the length of the warm season as well as typical seasonal conditions. The prevalence of variables with the 00Z time frame may speak to the importance of diurnal trends and daily solar heating for thunderstorm production. The standard local time in Alberta is Mountain Standard Time (MST) and is equal to UTC-7. Therefore, the 00Z observations correspond to 5:00pm MST (and 6:00pm MDT).

4.3 RANDOM FOREST MODELING

Previously created lightning prediction models have focused on classification and regression tree analysis. While tree-structured regression (Burrows et al., 2005) and decision tree analysis (Burrows, 2008) have been used to predict lightning, no attempts of creating prediction models via random forest modeling were found to date. Although many aspects of this study were built off of previous research, the random forest classification approach to prediction of cloud-to-ground lightning is novel. Lightning is episodic in nature and its characteristics vary greatly over time and space making creation of prediction models difficult. Random forest classification is well suited for lightning prediction modeling due to its ability to model complex interactions, efficiency with large data sets, and very high classification accuracy compared to other classification and regression models (Chen et al., 2004; Cutler et al., 2007).

4.3.1 BENEFITS

There are many characteristics of random forest classification that make it well suited for lightning occurrence prediction. The ability to model complex interactions efficiency with large data sets, and high classification accuracy compared to other classification and regression models are some of the benefits of random forest (Chen et al., 2004; Cutler et al., 2007). Single decision trees can often be biased or have high levels of variance. Random forests attempt to mitigate this by producing numerous decision trees whose output votes are averaged, theoretically producing some level of balance (Breiman, 2001). The ability to modify the model parameters to compensate for data imbalances and balance the class specific errors is another strong benefit. For data sets with multiple interactions between the variables, random forests tend to outperform linear models (Cutler et al., 2007).

4.3.2 DATA IMBALANCE

At first sight, the initial assessment of the data sets performed with no balancing modifications appeared to produce promising results. The overall OOB error estimates indicated models accuracies of ~ 80-90% for the coarser spatial scales and well over 90% for the 50km by 50km scale. Upon further analysis it was found that while the overall skill was quite high, this is a typical response of random forest when dealing with imbalanced data sets. While the data imbalances were by no means extreme, they were sufficient to significantly bias the prediction skill of the models to the negative class.

This conundrum of imbalanced data classification is well documented in the research pertaining to bioinformatics (Chen et al., 2004; Kubat and Matwin, 1997). The important need for accurate rare event predictions is also prevalent in meteorology and wildfire science. Forecasting severe storms, issuing weather alerts, and trying to predict extreme fire behaviour are three such examples. When trying to accurately predict rare events, modeling becomes a balancing act between maximizing hits without drastically over predicting false alarms. The five BRF methods applied to generate the ensemble forecasts play off this balancing act by forcing various proportions of each predictand class into the random forest models in an attempt to generate optimal hit rate forecasts.

4.3.3 OOB ERROR ESTIMATE VS. INDEPENDENT PREDICTIONS

The unbiased OOB error estimate is a beneficial tool allowing general model skill to be assessed without running a cross-validation. While the OOB error estimate may be sufficient for generating an unbiased estimate of error for the test set, separate training and validation data sets were used. If the goal was simply to identify top predictors of lightning then the entire data set could be used and the OOB-estimate of error would provide a sufficient overview of the skill. However, since the objective is to use the generated models to produce a series of predictions in order to create ensemble forecasts, an independent data set is required. Since forecast ensembles were generated from a series of predictions from 10 separate random forest models, it was necessary to have a test set to build the models and a separate validation set on which to make the predictions. Using the entire data set to build the model would mean re-running the test set (or a portion of it) through the model to create predictions. This would have generated a deceptively low estimate of error. Although each tree only uses a portion of the data set to grow, and retains the remainder to find the OOB error estimate, once the entire forest is grown each data point will have been included in multiple trees. As Liaw (2010) explains, due to the processes used for tree and forest building, near perfect prediction on the training data set is basically, by design, a self-fulfilling prophecy.

As described in section 2.4.1 RANDOM FOREST: BACKGROUND, the OOB error estimate of a random forest model depends primarily on two conditions, the level of correlation between trees and the strength of the individual trees (Breiman, 2001). The number of input variables included in the model have a direct impact on the two conditions, and thus the error rate. Theoretically, as the number of input variables is decreased, the between tree correlation and individual tree strength also decrease. The Mountain and Foothills plot in Figure 13e demonstrates this trend. Since the primary interest of this research lies in correctly predicting the positive class (1), this trend is not necessary upheld when changing the number of variables as a BRF approach is applied. By implementing this approach, the overall skill of the models are decreased in order to optimize the event forecast skill. This deviation from the default conditions of *randomForest* helps explains why even though the optimum number of variables was expected to be 15 to 25, when the various BRF models were run and the ensemble forecasts were generated, the optimum models often had a smaller number of predictors.

4.3.4 EFFECTS OF CORRELATION

It was anticipated that there would be a high level of correlation between many of the input variables as many of the predictors are directly or indirectly related. Each of the pressure level reanalysis variables have a series of vertical observations of daily mean measurements and four-times daily observations. Adding to the possible correlation, 24-hour change was calculated for each variable. Removal of the very highly correlated variables is essential to preserving the skill of the model and thus its predictive capabilities (Dormann et al., 2013; Tolosi and Lengauer, 2011). If multiple sets of variables that are highly correlated are included in the model the chance of between tree correlations may be higher. Strobl et al. (2008) observed that if multiple highly correlated variables are included in the random forest they are often used interchangeably. This can lead to less relevant (correlated) variables replacing others with higher predictive capability.

The *party* package (Hothorn et al., 2013) was developed to mitigate the effect of correlated variables. The package uses conditional inference trees to help alleviate the bias that random forests have towards highly correlated variables. Attempts were made to produce the models with this package however some of the necessary features were deemed too memory and CPU-intensive. In order to mitigate the possible introduction of model error from highly correlated variables a correlation analysis was performed as outlined in section **3.1 EXPLORATORY RUNS**. Variables with $|r| \geq 0.7$ were removed prior to generating the random forests. The findings of Dormann et al. (2013) supported this decision. The authors found that when variables with correlation coefficient greater than 0.7 were included in various multiple regression and machine learning approaches (including random forest), the collinearity began to severely distort the models and thus degrade the predictive skill (Dormann et al., 2013).

4.3.5 NUMBER OF TREES

The number of trees necessary to produce stable model outputs was determined in the preliminary phase. To ensure the number of trees included in the random forests remained sufficient for the various spatial and temporal scales, OOB error estimate plots, similar to Figure 12, were generated and checked for each series of runs. A relatively small number of trees were needed to generate a stable model output (~200 trees). Since the number of observations is much greater than the number of input variables, a smaller

number of trees is sufficient compared to high dimensional data where number of observations is less than number of predictors (Genuer et al., 2008). For each run, forests with 100 to 400 trees, at 50 tree intervals were created. Once the top predictors were selected, *ntree* was set to 251 for each of the models created for the ensemble forecasts. An odd number of trees were used to prevent ties from occurring within the model. If an even number of trees are used, there is a possibility the forest will have an equal number of votes. If this occurs, the tie is randomly broken (Liaw and Wiener, 2002). Using an odd number of trees avoids this issue all together.

4.3.6 DISADVANTAGES

Some disadvantages of random forests include possible over fitting of noisy classification and regression tasks and poor skill when large number of irrelevant features are included. Fortunately, it is believed that neither of these disadvantages affected the models generated. Over fitting occurs in only extreme scenarios where either the sample size is very small or there are a large number of highly noisy variables (Segal, 2004), which is not the case for this study. The vast majority of variables included in the random forest models were chosen because they were previously found to be useful predictors for lightning occurrence. In addition, the removal of correlated variables and the iterative process used to narrow down the number of variables would have further prevented the chance of a few, let alone a large number of irrelevant features being included.

Another, and more prominent, disadvantage of random forests are the difficulties associated with interpreting the models. Due to the highly computational procedures of the random forests, the predictor-predictand relationships are complex and the full classification functions cannot be represented by formulas or graphs (Cutler et al., 2007). This makes interpreting the models difficult. Partial dependence plots can be generated for one or two specific variables, however if multiple variables are highly importance to the classification or if there are high-level interactions the plots provide little information (Cutler et al., 2007). Building the models is also fairly memory intensive.

4.4 POSSIBLE SOURCES OF ERROR

The removal of data points with NAs present should not have affected the skill of the models. Roughly the same proportion (7%) of data point were removed from each

year and spatial scale. The proportion of data points removed from the two predictand classes were nearly equal.

4.4.1 THIN-PLATE SPLINE

The primary deficiency of data interpolated with the TPS is due to the low density of the RAOB stations. Tait et al. (2006) also experienced the drawbacks of low station density when interpolating daily rainfall in New Zealand from climate station data. Interpolation methods, such as TPS, are designed to interpolate, not extrapolate therefore inaccurate and physically impossible results can occur when such extrapolations are made (Xiao et al., 1996). Of the values generated by the TPS, ~2-4% were values outside of the normal range and ~3-5% of the outputs were NAs. If fewer than 5 stations had valid measurements for a particular variable a TPS was not performed, instead a value of NA was assigned. When unacceptable values were encountered, the value was corrected by replacing it with the nearest acceptable value and a note was made.

In addition to the problem of poor extrapolation skill, the small number of stations can also produce highly generalized interpolations. Entire thunderstorm events in Alberta can be missed if the storm system is not located within close proximity of the nearest RAOB station. In contrast, if independent thunderstorm events are occurring near multiple RAOB locations the interpolated conditions may be highly skewed providing a false set of conditions with strong convective potential when in actuality only a few small scattered or even no thunderstorm events in Alberta have actually occurred. This could cause model confusion if some of the storm conditions were captured by the reanalysis data but not picked up by the sounding variables.

4.4.2 LIGHTNING MISCLASSIFICATION

Misclassifications of CG or CC lightning occurred in less than 7% of the data. Of the 7% classification error, the majority were CC flashes incorrectly classified as CG (less than 15kA). The CLDN claims a CG detection efficiency exceeding 80-90% in most regions and a median location accuracy of 500m (Cummins and Murphy, 2009). The claims of detection efficiency and location accuracy are cautiously accepted as it is not uncommon for accuracy to be overstated or reported for best case scenarios. At best, the CG lightning data used to generate and validate the models is ~90% accurate. This implies that there is a possible error of 10% or greater introduced to the models. Areas

with low station density or down sensors can also contribute to error. Additional error is introduced by binning the strikes into predetermined cells, especially at the 2.5° latitude by 2.5° longitude and 1.25° latitude by 2.5° longitude scales.

Environmental conditions may vary greatly within each cell and between neighbouring cells. Strikes occurring near the periphery of the spatial bins may be misplaced into a neighbouring spatial bins and thus be separated for the corresponding atmospheric conditions. In general, any model is limited by the spatial and temporal resolution, and level of accuracy of each input variable.

4.5 LOOKING FORWARD

Alberta is a geographically diverse province with distinct landforms and ecosystems. This makes modeling for the entire province a challenge. In order to try to capture this variation, the province was subdivided into three separate zones based on the Natural Regions of Alberta for the 50km by 50km scale. It was believed that splitting the province into these zones would allow for region specific models to be built that could capture unique local conditions. For example, the weather systems developing leeward of the Rocky Mountains are not the same as storm development along the north eastern border. In addition, the Foothills are a lightning hotspot, experiencing on average more than 30 days per year of CG lightning (Burrows and Kochtubajda, 2010) and are therefore highly important to model accurately. Looking forward, a further breakdown of the province into geographically unique zones could prove beneficial. Separating the Mountains from the Foothills may produce some interesting results for the lower foothills region. The Boreal Forest could also be further divided into its Natural Subregions, or the Alberta Wildfire Management Areas designated by Alberta Environment and Sustainable Resource Development[9].

The models generated to predict lightning were generated from a binary predictand. Perhaps implementing a measure of the number of lightning strikes observed instead of a straight **0/1** predictand would have produced superior results. If the number of strikes were used, lightning density could also be predicted. The number of strikes can

[9] Alberta Wildfire Management Areas map available at: http://esrd.alberta.ca/wildfire/wildfire-maps/documents/WildfireManagementAreas-2013.pdf

be easily demined from the CLDN data set by summing the number of occurrences over a given time frame and spatial scale. It would also be interesting to see if the polarity of lightning strikes could be predicted. Naturally, better quality, higher resolution data would improve the capabilities of the models.

The predictors were narrowed down from all variables reaming post-correlation analysis to the top 12 by running a series of random forest models with a 1:1 BRF approach. This uniform approach was applied to all of the data sets regardless of the data imbalance. Perhaps selecting and implementing the optimum balancing method (BRF or WRF) from the initial random forest run (post-correlation) for each data set would have produced superior results. Finally, due to the highly correlated nature of the variables, a principal component analysis may prove beneficial for capturing the trends of related variables. The *party* package, or a similar approach using conditional inference trees to reduce the bias that random forests have towards highly correlated variables may also prove useful for future prediction models generated with random forest.

CHAPTER 5. CONCLUSIONS

Cloud-to-ground lightning is a major contributor to wildland fires in Canada. Despite the need, lightning is not included in the Canadian Weather Prediction Model resulting in an information gap that can lead to dire results. The primary objective of this study was to generate skillful lightning prediction models for the province of Alberta. Before the data and methods were introduced an overview of convective dynamics, thunderstorm formation, cloud electrification, types of lightning, and wildland fire in Canada was provided in the introductory chapter. A literature review including previous research on generating lightning occurrence prediction models was also provided.

Thirteen years (1999-2011) of weather and lightning data were collected. Upper air indexes and parameters from Radiosonde observations, surface and pressure level Reanalysis data, and a few additional calculated variables comprised the atmospheric inputs for the models. Geographic and temporal covariates such as latitude, longitude, elevation, and Julian day were also included as predictors. Random forest classifications were used to generate a series of lightning prediction models and forecast ensembles for Alberta.

A series of 6-hour and 24-hour lightning prediction models valid April-October were developed at three different spatial scales. The entire province was included in the first two spatial scales of 2.5° latitude by 2.5° longitude and 1.25° latitude by 2.5° longitude. For the third spatial scale of 50km by 50km, the province was divided into three separate zones based on the Natural Regions of Alberta. The first zone corresponds to the Boreal Forest Region, the second zone includes the Parkland and Grassland Regions, and the third zone encompasses the Rocky Mountain and Foothills Regions. Thin-plate splines were used to interpolate the input variables for the various spatial scales. Variables with Pearson's correlation coefficients greater or equal to 0.7 were excluded from the models.

A Balanced Random Forest (BRF) approach was used to generate a set of top predictors for each spatial scale and time frame. The BRF approach provides a level of model bias correction that was introduced by the imbalanced lightning occurrence data. The thirteen years of data were split into a seven year training set and a six year validation set. Random forest models were grown on the training set and then used to make ensemble predictions on the validation set. This provided an independent measure of forecast skill.

A wide range of models were generated for each spatial and temporal scale in an attempt to maximize hit rate. Optimum models were then selected based on the generated forecast skill measures. It was found that hit rate was maximized when the predictand (**0:1**) class sample size used to generated each tree in the forest was specified at a proportion of 0.6:1, where **0** indicates a non-event and **1** indicates a lightning event. Ensemble forecasts generated with this 0.6:1 BRF approach were interpreted with three different event thresholds (≥ 0.5, ≥ 0.7, and ≥ 0.9). Intuitively, setting the forecast threshold to ≥ 0.5 produced a superior hit rate while also increasing the number of false alarms as only 50% of the models included in the ensemble must predict lightning for the ensemble to forecast a lightning event.

In order to maximize the hit rate, a larger number of variables were required for the 2.5° latitude by 2.5° longitude, 1.25° latitude by 2.5° longitude, and the Mountain and Foothills zone than for the Boreal Forest zone and Parkland and Grassland zones. Showalter index (00Z), Julian day, convective available potential energy (00Z), and time of day (for 6-hour models) were commonly identified top predictors. The geographic covariates of latitude, longitude and elevation were highly important for the Mountain and Foothills models.

The daily lightning prediction model for the Mountain and Foothills zone was selected as the model with the best overall performance as determined from the independent prediction ensembles. The ensemble forecast had a hit rate of over 85%, an overall proportion correct of ~80%, false alarm ratio of 53%, and a false alarm rate of 19%. The variable importance and predictor-predictand relationships were discussed for this model. Showalter index (00Z), latitude, elevation, Julian day, and longitude were the top five variables in order of importance selected by this model.

Possible sources of error and limitations of the methods and models were discussed. Although many aspects of this study incorporated findings from previous research, the random forest classification approach to prediction of cloud-to-ground lightning is novel. A discussion regarding random forest modeling methods with respect to lightning prediction was provided and the advantages of this method were discussed. Recommendations for future research on lightning occurrence prediction modeling are also provided. While the models generated had good hit rates, the high rate of false alarms is a major drawback.

The addition of lightning prediction models to the field of wildland fire science will increase knowledge of lightning ignitions, better fire occurrence models, improve

resource allocation, and help increase preparedness of fire management agencies and communities alike. With a few weather input modifications and a small amount of operator training, the suggested models could easily be introduced into current wildland fire management and operations to aid with decision making.

BIBLIOGRAPHY

Aguado, E., Burt, J. 2010. Understanding Weather and Climate. 5th Edition. Pearson Education, Inc., Upper Saddle River, New Jersey. 586pp.

American Meteorological Society. 2012a. def: adiabatic process. http://glossary.ametsoc.org/wiki/Adiabatic_process. July 5, 2013.

American Meteorological Society. 2012b. def: cumulonimbus. http://glossary.ametsoc.org/wiki/Cumulonimbus. July 5, 2013.

American Meteorological Society. 2012c. def: stratosphere. http://glossary.ametsoc.org/wiki/Stratosphere. July 5, 2013.

American Meteorological Society. 2012d. def: troposphere. http://glossary.ametsoc.org/wiki/Troposphere. July 5, 2013.

Anderson, K. 1991. Models to predict lightning occurrence and frequency over Alberta. MSc. Thesis, University of Alberta, Edmonton.

Anderson, K. 2002. A model to predict lightning-caused fire occurrences. International Journal of Wildland Fire. 11:163-172.

Berger, K. 1967. Novel observations on lightning discharges - Results of research on Mount San Salvatore. Journal of the Franklin Institute. 283:478-525.

Bivard, R., Lewin-Koh, N. 2013. Maptools: Tools for reading and handling spatial objects. R package version 0.8-27. http://CRAN.R-project.org/package = maptools.

Boles, S. H., Verbyla, D. L. 2000. Comparison of three AVHRR-based fire detection algorithms for interior Alaska. Remote Sensing of Environment. 72:1-16.

Bolton, D. 1980. The computation of equivalent potential temperature. Monthly Weather Review. 108:1046-1053.

Breiman, L. 2001. Random forests. Machine Learning. 45:5-32.

Brook, M., Nakano, M., Krehbiel, P., Takeuti, T. 1982. The electrical structure of the Hokuriku winter thunderstorms. Journal of Geophysical Research- Oceans and Atmosphere. 87:1207-1215.

Burns, R. M., Honkala, B. H. (tech. Coords.) 1990. Silvics of North America: 1. Conifers; 2. Hardwoods. Agriculture Handbook 654, United States Department of Agriculture, Forest Service, Washington, D.C. volume 2.

Burrows, W. R. 2002. Statistical models for lightning prediction using Canadian Lightning Detection Network observations. Presented at the 16th Conference on Probability and Statistics in the Atmospheric Sciences, Orlando, Florida, USA.

Burrows, W. R. 2008. Dynamical-statistical models for lightning prediction to 48-hr over Canada and the United States. Presented at the 20th International Lightning Detection Conference, Tucson, Arizona, USA.

Burrows, W. R., King P, Lewis, P. J., Kochtubajda, B., Snyder, B., Turcotte, V. 2002. Lightning occurrence patterns over Canada and adjacent United States from Lightning Detection Network observations. Atmosphere-Ocean. 40:59-80.

Burrows, W. R., Kochtubajda, B. 2010. A decade of cloud-to-ground lightning in Canada: 1999–2008. Part 1: Flash density and occurrence. Atmosphere-Ocean. 48:177-194.

Burrows, W. R., Price, C., Wilson, L. J. 2005. Warm season lightning probability prediction for Canada and the northern United States. Weather and Forecasting. 20:971-988.

Canadian Forestry Service. 1987. Canadian forest fire danger rating system - Users' guide. Canadian Forestry Service, Fire Danger Group. Ottawa, Ontario, Canada. Three-ring binder (unnumbered publication).

Canadian Safety Council. 2013. Keep safe when lightning strikes. https://canadasafetycouncil.org/community-safety/keep-safe-when-lightning-strikes. May 5, 2013.

Chen, C., Liaw, A., Breiman, L. 2004. Using random forest to learn imbalance data. Department of Statistics, University of California, Berkeley, California, USA.

Christian, H. J., et al. 2003. Global frequency and distribution of lightning as observed from space by the Optical Transient Detector. Journal of Geophysical Research-Atmosphere. 108:D1.

Clodman, S., Chisholm, W. 1996. Lightning flash climatology in the southern Great Lakes region. Atmosphere and Ocean. 34:345-377.

Cummins, K. L., Murphy, M. J. 2009. An overview of lightning locating systems: History, techniques, and data uses, with an in-depth look at the US NLDN. IEEE Transactions on Electromagnetic Compatibility. 51:499-518.

Cutler, R. D., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., Lawler, J. J. 2007. Random forests for classification in ecology. Ecology. 88:2783-2792.

Dabberdt, W. F., Shellhorn, R., Cole, H., Paukkunen, A., Hörhammer, J., Antikainen, V. 2003. Radiosondes. in: Holton, J. R., Curry, J. A., Pyle, J. A. (eds.), Encyclopedia of atmospheric sciences. Boston: Academic Press. Amsterdam. 1900-1903.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., García Marquéz, J. R., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., Lautenbach, S.2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography. 36:27-46.

Dupilka, M. L., Reuter, G. W. 2011. Composite soundings associated with severe and tornadic thunderstorms in central Alberta, Atmosphere-Ocean, 49:269-278, DOI: 10.1080/07055900.2011.607146

Environment Canada. 2013. Aerological network. http://www.ec.gc.ca/-default.asp?lang=En&n=592AB94B-1&news=06F87D0A-4EC0-41F2-99EE-729855FCEA65. December 1, 2013.

Flannigan, M. D., Amiro, B. D., Logan, K. A., Stocks, B. J., Wotton, B. M. 2005. Forest fires and climate change in the 21st century. Mitigation and Adaptation Strategies for Global Change. 11:847-859.

Flannigan, M. D., Stocks, B. J., Turetsky, M., Wotton, B. M. 2009. Impacts of climate change on fire activity and fire management in the circumboreal forest. Global Change Biology. 15:549-560.

Flannigan, M. D., Van Wagner, C. E. 1991. Climate change and wildfire in Canada. Canadian Journal of Forest Research. 21:66–72.

Flannigan, M. D., Wotton, B. M. 1991. Lightning-ignited forest fires in northeastern Ontario. Canadian Journal of Forest Research. 21:277-287.

Forestry Canada Fire Danger Group. 1992. Development and structure of the Canadian Forest Fire Behavior Prediction System. Ottawa, Ontario, Canada. 64pp.

Fuquay, D. M., Baughman, R. G., Latham, D. J. 1979. A model for predicting lightning fire ignition in wildland fuels. USDA Forest Service, Research Paper INT-217, 22p. Intermountain Forest and Range Experiment Station. Ogden, Utah, USA.

Fuquay, D. M., Baughman, R. G., Taylor, A. R., Hawe, R. G. 1967. Characteristics of 7 lightning discharges that caused forest fires. Journal of Geophysical Research. 72:6371-6373.

Fuquay, D. M., Taylor, A. R., Hawe, R. G., Schmid, C. W. 1972. Lightning discharges that caused forest fires. Journal of Geophysical Research. 77:2156-2158.

Furrer, R., Nychka, D., Sain, S. 2013. fields: Tools for spatial data. R package version 6.9.1. http:/CRAN.R-project.org/package=fields.

Genuer, R., Poggi, J. M., Tuleau, C. 2008. Random forests: Some methodological insights. Institut National de Recherche en Informatique et en Automatique. 32.

George, J. J. 1960. Weather forecasting for aeronautics. Academic Press., New York, New York, USA.

Government of Alberta. 2012. Industry and economy: Forestry. http://alberta.ca/home/181.cfm#Forestry. Jan 20, 2012.

Government of Alberta. 2013. Climate and geography. http://www.albertacanada.com/immigration/choosing/province-climate-geography.aspx. Dec 5, 2013.

Green, P. J., Silverman, B. W. 1994. Nonparametric regression and generalized linear models : a roughness penalty approach. 1st edition. Chapman & Hall., London and New York. 175pp.

Haines, D. 1988. A lower-atmosphere severity index for wildland fires. National Weather Digest. 13:23-27.

Harrison, R. G. 2004. The global atmospheric electrical circuit and climate. Surveys in Geophysics. 25:441-484.

Hastie, T., Tibshirani, R., Friedman, J. H. 2009. The elements of statistical learning : data mining, inference, and prediction. 2nd edition. Springer., New York, New York, USA. 737pp.

Holle, R. L. 2008. Annual rates of lightning fatalities by country. Presented at the 20th International Lightning Detection Conference, Tucson, Arizona, USA.

Hothorn, T., Hornik, K., Strobl, C., Zeileis, A. 2013. party: a laboratory for recursive partytioning. http://cran.r-project.org/web/packages/party/vignettes/party.pdf.

Kalnay, E., et al. 1996. The NCEP/NCAR 40-year reanalysis project. Bulletin of the American Meteorological Society. 77:437-471.

Kochtubajda, B., Burrows, W. R. 2010. A decade of cloud-to-ground lightning in Canada: 1999–2008. Part 2: Polarity, multiplicity and first-stroke peak current. Atmosphere-Ocean. 48:195-209.

Kochtubajda, B., Burrows, W. R., Liu, A., Patten, J. K. 2013. Surface rainfall and cloud-to-ground lightning relationships in Canada. Atmosphere-Ocean. 51:2:226-238.

Kochtubajda, B., Flannigan, M. D., Gyakum, J. R., Stewart, R. E., Logan, K. A., Nguyen, T-V. 2006. Lightning and fires in the Northwest Territories and responses to future climate change. Arctic. 59:211-221.

Kochtubajda, B., Stewart, R. E., Gyakum, J. R., Flannigan, M. D. 2002. Summer convection and lightning over the Mackenzie River Basin and their impacts during 1994 and 1995. Atmosphere-Ocean. 40:199-220.

Krawchuk, M. A., Cumming, S. G., Flannigan, M. D. 2009. Predicted changes in fire weather suggest increases in lightning fire initiation and future area burned in the mixedwood boreal forest. Climatic Change. 92:83-97.

Krawchuk, M. A., Cumming, S. G., Flannigan, M. D., Wein, R. W. 2006. Biotic and abiotic regulation of lightning fire initiation in the mixedwood boreal forest. Ecology. 87:458-468.

Krehbiel, P. R. 1986. The electrical structure of thunderstorms. in: Krider, E. P., Roble, R. G., eds. The Earth's electrical environment. National Academy Press., Washington, D.C., USA. 90-113.

Krider, E. P., Noggle, R. C., Pifer, A. E., Vance, D. L. 1980. Lightning direction-finding systems for forest fire detection. Bulletin of the American Meteorological Society. 61:980-986.

Kubat, M., Matwin, S. 1997. Addressing the curse of imbalanced training sets: one-sided selection. in: Fisher, D. (ed.), Machine learning: Proceedings of the fourteenth international conference (ICML'97), San Francisco, California. Morgan Kaufmann Publishers. 179-186.

Kuhn, M., et al. 2013. caret: Classification and regression training. R package version 6.0-21. http://CRAN.R-project.org/package=caret.

Latham, D. 1991. Lightning flashes from a prescribed fire-induced cloud. Journal of Geophysical Research- Atmosphere. 96:17151-17157.

Latham, D., Williams, E. 2001. Lightning and forest fires. in: Johnson, E. A., Miyanishi, K. (eds.), Forest fires : Behavior and ecological effects. Academic Press, Inc., San Diego, California, USA. 375-418.

Liaw, A. 2009. Random forests variable importance question. R help. http://r.789695.n4.nabble.com/ Random-Forests-Variable-Importance-Question-td884339.html. Dec 3, 2013.

Liaw, A. 2010. Random Forest AUC. R help. http://r.789695.n4.nabble.com/Random-Forest-AUC-td3006649.html. Dec 2, 2013.

Liaw, A., Wiener, M. 2002. Classification and regression by randomForest. R News. 2/3:18-22.

Liu, C. T., Williams, E. R., Zipser, E. J., Burns, G. 2010. Diurnal variations of global thunderstorms and electrified shower clouds and their contribution to the global electrical circuit. Journal of the Atmospheric Sciences. 67:309-323.

Livingston, J. M., Krider, E. P. 1978. Electric-fields produced by Florida thunderstorms. Journal of Geophysical Research- Oceans and Atmosphere (1978) 83:385-401.

MacGorman, D. R., Rust, W. D. 1998. The electrical nature of storms. Oxford University Press. New York, New York, USA.

Martell, D. 2001. Forest fire management. in: Johnson, E. A., Miyanishi, K. (eds.), Forest fires : behavior and ecological effects. Academic Press, Inc., San Diego, California, USA. 527-583.

Martell, D. L., Sun, H. 2008. The impact of fire suppression, vegetation, and weather on the area burned by lightning-caused forest fires in Ontario. Canadian Journal of Forest Research. 38:1547-1563.

McFarlane, B. L. 2006. Human dimensions of fire management in the wildland-urban interface: a literature review. in: Hirsch, K., Fuglem, P. (eds.), Canadian wildland fire strategy: Background synthesis, analysis, and perspectives. Canadian Council of Forest Ministers. Natural Resources Canada, Canadian Forest Service, Northern Forestry Centre. Edmonton, Alberta, Canada. 27-34.

McNutt, S. R., Williams, E. R. 2010. Volcanic lightning: global observations and constraints on source mechanisms. Bulletin of Volcanology. 72:1153-1167.

Mills, B., Unrau, D., Parkinson, C., Jones, B., Yessis, J., Spring, K., Pentelow, L. 2008. Assessment of lightning-related fatality and injury risk in Canada. Natural Hazards. 47: 157-183.

Mills, B., Unrau, D., Pentelow, L, Spring, K. 2010. Assessment of lightning-related damage and disruption in Canada. Natural Hazards. 52: 481-499.

Morales, C. A., Anagnostou, E. N., Williams, E., Kriz, J. S. 2007. Evaluation of peak current polarity retrieved by the ZEUS long-range lightning monitoring system. IEEE Transactions on Geoscience and Remote Sensing. 4:32-36.

Morissette, J., Gauthier, S. 2008. Study of cloud-to-ground lightning in Quebec: 1996-2005. Atmosphere-Ocean. 46:443-454.

Nash, C. H., Johnson, E. A. 1996. Synoptic climatology of lightning-caused forest fires in subalpine and boreal forests. Canadian Journal of Forest Research. 26:1859-1874.

National Weather Service. 2009. Upper-air observations program. http://www.nws.noaa.gov/ops2/ua/. Dec 1, 2013.

Natural Regions Committee. 2006. Natural regions and subregions of Alberta. Compiled by DJ Downing and WW Pettapiece. Edmonton, Alberta, Canada.

Natural Resources Canada. 2004. The state of Canada's forests 2003-2004. Headquarters, Policy, Planning and Internal Affairs Branch, Ottawa, Ontario, Canada.

NOAA National Weather Service. Radiosonde observations. 2013. http://www.ua.nws.noaa.gov/factsheet.htm. Dec 15, 2013.

Oliver, J. E. 2005. Encyclopedia of world climatology. Springer, Dordrecht, The Netherlands.

Orville, R. E., Songster, H. 1987. The east-coast lightning detection network. IEEE Transactions on Power Delivery 2:899-907.

Orville, R.E., Huffines, G. R. 1999. Lightning ground flash measurements over the contiguous United States: 1995-97. Monthly Weather Review. 127:2693-2703.

Orville, R. E., Huffines, G. R., Burrows, W. R., Holle, R.L., Cummins, K. L. 2002. The North American Lightning Detection Network (NALDN) - First results: 1998-2000. Monthly Weather Review. 130:2098-2109.

Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., Brunk, C. 1994. Reducing misclassification costs. in: Proceedings of the 11th Intentional Conference on Machine Learning (ICML'94). New Brunswick, New Jersey, USA.

Pejovic, T., Williams, V. A., Noland, R. B., Toumi, R. 2009. Factors affecting the frequency and severity of airport weather delays and the implications of climate change for future delays. Transportation Research Record: Journal of the Transportation Research Board. 2009:97-106.

Podur, J., Martell, D. L., Knight, K. 2002. Statistical quality control analysis of forest fire activity in Canada. Canadian Journal of Forest Research. 32:195-205.

Price, C., Rind, D. 1994. The impact of a 2-X-Co2 climate on lightning-caused fires. Journal of Climate. 7:1484-1494.

Pyne, S. J. 2007. Awful splendour – a history of fire in Canada. University of British Columbia Press. Vancouver, British Columbia, Canada. 584pp.

R Core Team.2013. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rakov, V. A., Uman, M. A. 2003. Lightning : Physics and effects. Cambridge University Press. Cambridge, U.K. ; New York, New York, USA.

Rittmaster, R., Adamowicz, W. L., Amiro, B., Pelletier, R.T. 2006. Economic analysis of health effects from forest fires. Canadian Journal of Forest Research. 36:868-877.

Rorig. M. L., Ferguson, S.A. 1999. Characteristics of lightning and wildland fire ignition in the Pacific Northwest. Journal of Applied Meteorology. 38:1565-1575.

Rosenfeld, D., Fromm, M., Trentmann, J., Luderer, G., Andreae, M. O., Servranckx, R. 2007. The Chisholm firestorm: Observed microstructure, precipitation and lightning activity of a pyro-cumulonimbus. Atmospheric Chemistry and Physics. 7:645-659.

Rust, W. D. 1986. Positive cloud-to-ground lightning. in: The Earth's electrical environment. National Academy Press, Washington, DC, USA. 41-45.

Schonland, B. 1964. The flight of thunderbolts. 2d edition. Clarendon Press, Oxford.

Segal, M. 2004. Machine learning benchmarks and random forest regression. Center for Bioinformatics and Molecular Biostatistics, University of California. http://escholarship.org/uc/item/35x3v9t4.

Showalter. A. K.1949. A stability index for forecasting thunderstorms. Bulletin of the American Meteorological Society. 34:250-252.

Stamp, R. M. 2009. Alberta: climate. http://www.thecanadianencyclopedia.com/en/article/alberta/. Dec 11, 2013.

Statistics Canada. 2005. Land and freshwater area, by province and territory. http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/phys01-eng.htm. Dec 11, 2013.

Statistics Canada. 2013. Population and dwelling counts, for Canada, provinces and territories, 2011 and 2006 censuses. http://www12.statcan.gc.ca/census recensement/2011/dp-pd/hlt-fst/pd-pl/Table-Tableau.cfm?T=101&S=50&O=A. Dec 11, 2013.

Stocks, B. J., Mason, J. A., Todd, J. B., Bosch, E. M., Wotton, B. M., Amiro, B. D., Flannigan, M. D., Hirsch, K. G., Logan, K. A., Martell, D. L., Skinner, W. R. 2002. Large forest fires in Canada, 1959-1997. Journal of Geophysical Research. DOI: 107/108:8149.8141-8149.8112.

Stolzenburg, M., Marshall, T. C. 1998. Charged precipitation and electric field in two thunderstorms. Journal of Geophysical Research- Atmosphere. 103:19777-19790.

Stolzenburg, M., Rust, W. D., Smull, B. F., Marshall, T. C. 1998. Electrical structure in thunderstorm convective regions - 1. Mesoscale convective systems. Journal of Geophysical Research- Atmosphere. 103:14059-14078.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., Zeileis, A. 2008. Conditional variable importance for random forests. BioMed Central Bioinformatics. 9:307.

Tait, A., Henderson, R., Turner, R., Zheng, X. G. (2006. Smoothing spline interpolation of daily rainfall for New Zealand using a climatological rainfall surface. International Journal of Climatology. 26:2097-2115.

Taylor, S. W., Stennes, B., Wang, S., Taudin-Chabot, P. 2006. Integrating Canadian wildland fire management policy and institutions: Sustaining natural resources, communities and ecosystems. in: : Hirsch, K., Fuglem, P. (eds.), Canadian wildland fire strategy: Background synthesis, analysis, and perspectives. Canadian Council of Forest Ministers. Natural Resources Canada, Canadian Forest Service, Northern Forestry Centre. Edmonton, Alberta, Canada.

Tolosi, L., Lengauer, T. 2011. Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics. 27:1986-1994.

Tymstra, C., Wang, D., Rogeau, M-P. 2005. Alberta wildfire regime analysis. Alberta Department of Sustainable Resource Development, Forest Protection Division, Wildfire Policy and Business Planning Branch. Edmonton, Alberta, Canada. 171.

Uman, M. A. 1987. The lightning discharge. Academic Press, Orlando, USA. 377pp.

Van Wagner, C. E. 1974. Structure of the Canadian Forest Fire Weather Index. Department of Environment, Canadian Forestry Service Publication, 1333 Ottawa, Ontario, Canada.

Van Wagner, C. E. 1987. Development and structure of the Canadian Forest Fire Weather Index System. Department of Environment, Canadian Forestry Service Technical Report, 35. Ottawa, Canada.

Vicars, M. 1999. Firesmart: protecting your community from wildfire. Partners in Protection. Edmonton, Alberta.

Wåhlin, L. 1986. The thundercloud. in: Atmospheric electrostatics. Research Studies Press, Letchworth, Hertfordshire, England. 57-78.

Weber, M. G., Flannigan, M. D. 1997. Canadian boreal forest ecosystem structure and function in a changing climate: impact on fire regimes. Environmental Reviews. 5:145-166.

Weber, M. G., Stocks, B. J. 1998. Forest fires and sustainability in the boreal forests of Canada. Ambio. 27:545-550.

Weidman, C. D., Krider, E. P. 1978. Fine-structure of lightning return stroke wave forms. Journal of Geophysical Research- Ocean and Atmosphere. 83:6239-6247.

Weidman, C. D., Krider, E. P. 1979. Radiation-field wave forms produced by intracloud lightning discharge processes. Journal of Geophysical Research- Ocean and Atmosphere. 84:3159-3164.

Wierzchowski, J., Heathcott, M., Flannigan, M. D. 2002. Lightning and lightning fire, central cordillera, Canada. International Journal of Wildland Fire. 11:41-51.

Williams, E. R., Mushtak, V., Rosenfeld, D., Goodman, S., Boccippio, D. 2005. Thermodynamic conditions favorable to superlative thunderstorm updraft, mixed phase microphysics and lightning flash rate. Atmospheric Research. 76:288-306.

Williams, E. R. 1989. The tripole structure of thunderstorms. Journal of Geophysical Research- Atmosphere. 94:13151-13167.

Williams, E. R., Heckman, S. J. 1993. The local diurnal-variation of cloud electrification and the global diurnal-variation of negative charge on the earth. Journal of Geophysical Research- Atmosphere. 98:5221-5234.

Williams, J. K., Sharman, R., Craig, J., Blackburn, G. 2008. Remote detection and diagnosis of thunderstorm turbulence. Proceedings of the Society of Photo-Optical Instrumentation Engineers. 7088.

World Meteorological Organization. 1987. International cloud atlas. Revised edition. Secretariat of the World Meteorological Organization, Geneva.

Wotton, B. M., Martell, D. L. 2005. A lightning fire occurrence model for Ontario. Canadian Journal of Forest Research. 35:1389-1401.

Xiao, Y. C., Ziebarth, J. P., Woodbury, C., Bayer, E., Rundell, B., VanderZijp, J. 1996. The challenges of visualizing and modeling environmental data. IEEE Visualization ’96 Proceedings. 413-416.

Zipser, E. J., Lutz, K. R. 1994. The vertical profile of radar reflectivity of convective cells - a strong indicator of storm intensity and lightning probability. Monthly Weather Review. 122:1751-1759.