Predicting Monthly Precipitation in Ontario using a Multi-Model Ensemble and the XGBoost Algorithm

PREDICTING MONTHLY PRECIPITATION IN ONTARIO USING A MULTI-MODEL ENSEMBLE AND THE XGBOOST ALGORITHM

By Milena Hadzi-Tosev, B.Sc

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

© Copyright by Milena Hadzi-Tosev, December 2020
All Rights Reserved

Master of Science (2020)

McMaster University (Mathematics and Statistics)

Hamilton, Ontario, Canada

TITLE: Predicting Monthly Precipitation in Ontario using a Multi-Model Ensemble and the XGBoost Algorithm

AUTHOR: Milena Hadzi-Tosev (McMaster University)

SUPERVISORS: Dr. Paul McNicholas, Dr. Zoe Li

NUMBER OF PAGES: ix, 63

Abstract

There is a strong interest in the climate community to improve the ability to accurately predict future trends of climate variables. Recently, machine learning methods have proven their ability to contribute to more accurate predictions of historical data on a variety of climate variables. There is also a strong interest in using statistical downscaling to predict local station data from the output of multi-model ensembles. This project looks at using the machine learning algorithm XGBoost and evaluating its ability to accurately predict historical monthly precipitation, with a focus of applying this method to simulate future precipitation trends.

Acknowledgements

First and foremost, I would like to thank my co-supervisors Dr. Paul McNicholas and Dr. Zoe Li. Their knowledge, guidance, inquiry, and expertise was essential for completing this thesis. I am extremely grateful in having them support me during this journey.

I would also like to thank Dr. Noah Forman, Dr. Zoe Li, and Dr. Paul McNicholas for taking the time to be on my examination committee, and for making an enjoyable defense process. Their comments and questions were incredibly helpful in the pursuit of fine-tuning the final draft of the thesis.

I would like to thank my parents and my grandmother for their love and support. I would like to thank my sisters for reminding me how important it is to take breaks, and to keep the balance of work and life. I am incredibly thankful to have had them all by my side during this journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Climate change is one of the most important challenges facing humanity in the coming decades. There has been many work done detailing the extent of climate change and the future projections that are possible under various future climate scenarios. One of the most commonly used projections for climate data comes from the Coupled Model Intercomparison Project (CMIP). The most recent of those models is the CMIP Phase 6 (CMIP6) (Eyring et al., 2016), which follows the CMIP Phase 5 (CMIP5) (ENES, 2019) and is an extension of the CMIP5 structure. In particular, the CMIP projects include Regional Climate Models (RCMs) and Global Climate Models (GCMs). RCMs are dynamically downscaled using GCMs (Giorgi and Gutowski, 2015), where "the GCM can describe the response of the global circulation to large-scale forcings [...] while the RCM can spatially and temporally refine this large-scale information by accounting for the effects of sub-GCM grid scale forcings and processes" (Giorgi, 2019). Currently, the GCMs are available for CMIP6, but the RCMs have not yet been released. There are many climate variables that are relevant in measuring climate change, and the two

that are of most interest are temperature and precipitation. There are various time scales of interest including hourly, daily, monthly, and yearly data. Previous studies that have found accurate methods for predicting historical temperature in Ontario, Canada, (Wang et al., 2014; Li et al., 2020; Wang et al., 2014) but the methods used in precipitation prediction in the same region are very limited and returned poor predictive results (Wang et al., 2014). In particular, there are a limited number of papers discussing the prediction of precipitation in Ontario; this could in part be due to the difficulty of predicting the pattern of the daily precipitation data. Additionally, there are very few papers that investigate the ability to predict monthly precipitation in Ontario. Recent machine learning methods have improved the ability to predict historical climate data with a moderate degree of accuracy. Some of these methods include neural networks (Nair et al., 2018; Gizaw and Gan, 2016; Bochinski et al., 2017; Zheng et al., 2017; Soares Dos Santos et al., 2016), support vector regression (Gizaw and Gan, 2016; Kisi and Sanikhani, 2015; Wang et al., 2014; Okkan and Kirdemir, 2016), and random forests (Xu et al., 2020). Another machine learning method that has shown an ability to predict data well in the context of climate research and in general is tree boosting (Chen and Guestrin, 2016). In particular, this research will look at extreme gradient boosting (XGBoost), which has demonstrated good predictive ability in recent literature (Zheng et al., 2017). XGBoost is used in this project in combination with a multi-model ensemble of climate models, with an aim of improving the ability to predict monthly precipitation for 12 climate stations in Ontario, Canada.

Chapter 2

Data Introduction

2.1 Physical Climate Models

The climate model data comes from the Coupled Model Intercomparison Project Phase 5 (ENES, 2019) and Phase 6 (PCMDI, 2019) historical and future simulations, where the two phases are denoted by CMIP5 and CMIP6, respectively. CMIP is organized under the World Climate Research Programmes (WCRP) Working Group on Coupled Modelling (WCRP, 2017). According to the WCRP, "the objective of CMIP is to better understand past, present, and future climate changes arising from natural, unforced variability, or in response to changes in radiative forcings in a multi–model context" (WCRP, 2017). The research in this study will be looking at historical simulations, which estimate the actual observed states, and future scenarios, which represents a possible future under different "radiative forcing pathways from greenhouse gas emissions" (Hausfather, 2019).

The CMIP historical simulation is an experiment that simulates the recent past, and the main purpose of this experiment is to evaluate the ability of the climate

models to accurately predict historical climate data. The CMIP6 future simulations are experiments that simulate the near future, under different emissions scenarios (Michaut, 2020).

TABLE 2.1: Locations of the 12 selected stations and their corresponding RCM grid (Li et al., 2020).

Station Name	Short Name	Station Latitude	Station Longitude	RCM Grid Latitude	RCM Grid Longitude	Elevation
Big Trout Lake	BTL	53.83° N	89.87° W	53.76° N	89.84° W	224.1m
London International Airport	LA	43.03° N	81.15° W	42.98° N	81.04° W	278.0m
Moosonee	MUA	51.27° N	80.65° W	51.34° N	80.60° W	9.1m
North Bay Airport	NB	46.36° N	79.42° W	46.28° N	79.50° W	370.3m
Ottawa International Airport	OMIA	45.32° N	75.67° W	45.40° N	75.76° W	222.2m
Sault Ste. Marie Airport	SSMA	46.48° N	84.51° W	46.50° N	84.56° W	192.0m
Sioux Lookout Airport	SLA	50.12° N	91.90° W	50.02° N	91.82° W	294.7m
Timmins Victor Power Airport	TVPA	48.57° N	81.38° W	48.48° N	81.48° W	383.4m
Toronto Island Airport	TIA	43.63° N	79.40° W	43.64° N	79.72° W	173.4m
Toronto Pearson International Airport	TPIA	43.68° N	79.40° W	43.64° N	79.50° W	76.8m
Wiarton Airport	WTA	44.75° N	81.11° W	53.76° N	81.04° W	114.0m
Windsor Airport	WSA	42.28° N	82.96° W	53.76° N	83.02° W	189.6m

FIGURE 2.1: Locations of the twelve selected meteorological stations in Ontario, Canada (Li et al., 2020).

In particular, these scenarios are run as part of Scenario Model Intercomparison Project (ScenarioMIP), which provides "multi–model climate projections based on

alternative scenarios of future emissions and land use changes produced with integrated assessment models" (WCRP, 2017). For this project, the future climate change scenarios are projected under three different emission pathways. The primary focus is on the 12 stations listed in Table 2.1, and Fig. 2.1 contains a map of these stations Ontario, Canada. For the CMIP5 historical prediction, the seven RCMs that were used for this project were downloaded from the North American Coordinated Regional Downscaling Experiment (NA-CORDEX) (Mearns et al., 2020). Data was extracted for monthly observations from 1950–1999. When accessing the RCM files the experiment family is historical, the realm is land, the climate variable of interest is `pr`, and the climate model experiment ensemble is r1i1p1. The ensemble member in the CMIP projects are named in the rip–nomenclature, "r for realization, i for initialization and p for physics, followed by an integer, e.g. r1i1p1" (ENES, 2019). The CMIP simulated data are stored in NetCDF format, containing the variables longitude, latitude, in addition to time and the climate variable. The data is gridded at a 50km resolution and extracted for the Ontario station coordinates, and the RCMs are seen in Table 2.2.

TABLE 2.2: Information on the 7 CMIP5 RCMs and their associated GCMs.

GCM	RCM	Resolution (≈50km)	Modeling Institution	Institution Full Name
CanESM2	CanRCM4	0.44°	CCCma	Canadian Centre for Climate Modelling and Analysis
CanESM2	CRCM5	0.44°	UQAM	Université du Québec à Montréal
CanESM2	RCA4	0.44°	SMHI	Swedish Meteorological and Hydrological Institute
EC–EARTH	HIRHAM5	0.44°	DMI	Danish Meteorological Institute
EC–EARTH	RCA4	0.44°	SMHI	Swedish Meteorological and Hydrological Institute
MPI–ESM–LR	CRCM5	0.44°	UQAM	Université du Québec à Montréal
MPI–ESM–MR	CRCM5	0.44°	UQAM	Université du Québec à Montréal

The RCMs are CanESM2.CanRCM4 (ECCC, 2018; GOC, 2018), CanESM2.–CRCM5 (Takhsha et al., 2018; GOC, 2018), CanESM2.RCA4 (GOC, 2018; Kjellström et al., 2016), EC–EARTH.RCA4 (Kupiainen et al., 2015; EC-Earth, 2020), EC–EARTH.HIRHAM5 (EC-Earth, 2020; Christensen et al., 2007), MPI–ESM–LR.CRCM5 (Giorgetta et al., 2013; Takhsha et al., 2018), and MPI–ESM–MR.–CRCM5 (Climate Workspace TCW, 2020; Takhsha et al., 2018). For the CMIP6 prediction, the fourteen GCMs under historical scenario and future scenario Shared Socioeconomic Pathways (SSP) are available through the WCRP Data Portal (WCRP, 2017). The GCMs are gridded at a 100km resolution, and are extracted for the closest Ontario station coordinates. The historical CMIP6 GCMs have data available for the entire time period 1950–1999, under r1i1p1f1. Information about the CMIP6 GCMs can be accessed in Table 2.3. The 14 GCMs are BCC–CSM2–MR (Wu et al., 2019; Xin et al., 2018), CAMS–CSM1–0 (Rong, 2019), CESM2 (UCAR, 2019), CESM2–WACCM (Danabasoglu, 2019), EC–Earth3 (EC-Earth, 2020), EC–Earth3–Veg (EC-Earth, 2020), FGOALS–f3–L (He et al., 2019), FIO–ESM–2–0 (Song et al., 2019), GFDL–ESM4 (Dunne, 2019), INM–CM4–8 (Volodin et al., 2018), INM–CM5–0 (Volodin and Gritsun, 2018), MPI–ESM1–2–HR (Gutjahr et al., 2019; Botzet, 2020), MRI–ESM2–0 (Yukimoto et al., 2019), and NorESM2–MM (Bethke, 2016). The future climate projections data from the CMIP6 GCMs have five available SSP. Representative Concentration Pathways (RCPs) are "scenarios that include time series of emissions and concentrations of the full suite of greenhouse gases (GHGs) and aerosols and chemically active gases, as well as land use/land cover" (IPCC, 2020). The relationship between the CMIP6 SSP and CMIP5 RCP are seen in Table 2.4. This study will look at the three pathways SSP1–2.6, SSP2–4.5, and SSP5–8.5, and the monthly precipitation

data simulated for the years 2020 to 2099. The pathways can be separated into three greenhouse gas emission levels scenarios: low future emissions (SSP1–2.6), moderate future emissions (SSP2–4.5), and high future emissions (SSP5–8.5).

TABLE 2.3: Information on the 14 CMIP6 GCMs used for this project.

	GCM	Spatial Resolution	Modeling Institution	Source	Institution full name
1	BCC–CSM2–MR	1° 110km	BCC	AOGCM	Beijing Climate Center, China Meteorological Administration
2	CAMS–CSM1–0	1° 100km	CAMS	AOGCM	Chinese Academy of Meterological Sciences
3	CESM2	1° 100km	NCAR	AOGC, BGC	National Center for Atmospheric Research
4	CESM2–WACCM	1° 100km	NCAR	AOGC, BGC, CHEM, AER	National Center for Atmospheric Research
5	EC–Earth3	1° 100km	EC– Earth	AOGCM	European community Earth–consortium
6	EC–Earth3–Veg	1° 100km	EC–Earth	AOGCM	European community Earth–consortium
7	FGOALS–f3–L	1° 100km	CAS	AOGCM	Chinese Academy of Sciences
8	FIO–ESM–2–0	1° 100km	FIO	AOGCM	First Institute of Oceanography, SOA,China
9	GFDL–ESM4	1° 100km	NOAA–GFDL	AOGC, AER, CHEM, BGC	National Oceanic and Atmospheric Administration –Geophysical Fluid Dynamics Laboratory
10	INM–CM4–8	1° 100km	INM	AOGC, AER	Institute for Numerical Mathematics
11	INM–CM5–0	1° 100km	INM	AOGC, AER	Institute of Numerical Mathematics
12	MPI–ESM1–2–HR	1° 100km	MPI–M	AOGCM	Max Planck Institute for Meteorology
13	MRI–ESM2–0	1° 100km	MRI	AOGC, AER, CHEM	Meteorological Research Institute
14	NorESM2–MM	1° 100km	NCC	AOGC, AER, BGC	Norwegian Climate Centre

TABLE 2.4: The relationship between CMIP6 SSP and CMIP5 RCP.

SSP (CMIP6)	RCP (CMIP5)	Explanation of how they are connected
SSP1–2.6	RCP2.6	RCP2.6 has "radiative forcing peaks at approximately $3Wm^{-2}$ before 2100 and then declines" (IPCC, 2020). SSP1–2.6 shows a more "gradual decline in emissions than RCP2.6, and a higher starting point" (Hausfather, 2019).
SSP2–4.5	RCP4.5	RCP4.5 is an intermediate stabilisation pathway, "radiative forcing is stabilised at approximately $4.5Wm^{-2}$ after 2100" (IPCC, 2020). SSP2–4.5 has a "higher starting point, and slightly slower decline than RCP4.5" (Hausfather, 2019).
SSP5–8.5	RCP8.5	RCP8.5 one high emissions pathway where, "radiative forcing reaches greater than $8.5Wm^{-2}$ by 2100 and continues to rise for some amount of time" (IPCC, 2020). SSP5–8.5 has "higher CO2 emissions than RCP8.5, correspondingly larger cuts in non–CO2 emissions" (Hausfather, 2019).

2.2 Observed Weather and Climate Data

The historical weather data is accessed through the Government of Canada historical database under *Past Weather and Climate Data* (ECCC, 2020). The data is available for hourly, daily, and monthly time intervals. This project is looking at monthly data for the years 1950–1999. The historical data for each respective station are extracted in CSV format containing coordinate variables like longitude, latitude, elevation, and time, as well as the observed weather data. The historical observed daily and monthly precipitation data were accessed through the Government of Canada's historical database under *Past weather and Climate data* (ECCC, 2020). The historical data for each respective station is extracted in CSV format containing coordinate variables longitude and latitude, elevation, time, and the climate variables. Table 2.5 contains the average annual precipitation for each station and the number of missing data for each station. It is important to compute the average annual precipitation in order to compare the average annual precipitation to the monthly RMSE values for each station. The

average annual precipitation for each period is calculated by aggregating the average monthly precipitation values, where $m_{i,j}$ is the average monthly precipitation value calculated from the non–missing data for month i and station j. The average annual precipitation for station j is defined in Eq. 2.1.

$$\text{AAP}_j = \sum_{i=1}^{12} m_{i,j} \tag{2.1}$$

TABLE 2.5: Summary of the observed monthly station data.

Station	Average Annual Precipitation (1950–1999)	(1950–1989)	(1990–1999)	Number of Missing Monthly Observations (1950–1989)	(1990–1999)	Notes
BTL	667.887	669.638	608.000	0	3	Obs. end 1992
LA	1029.624	1016.155	1083.500	0	0	
MUA	801.140	804.881	776.469	3	41	No observations: 1994 01– 1996 11
NB	1060.880	1052.725	1093.500	0	0	
OMIA	977.350	969.310	1009.51	0	0	
SSMA	980.253	983.266	971.830	7	0	Obs start in 1961
SLA	798.806	788.733	839.100	0	0	
TVPA	920.648	930.967	884.89	0	5	Obs start in 1955
TIA	853.957	851.258	871.075	8	1	Obs start in 1957, and end in 1994
TPIA	847.770	843.510	864.810	0	0	
WTA	1069.721	1062.072	1099.840	4	0	
WSA	962.314	951.838	1004.22	0	0	

2.3 Data Summary

This project assumes that using the historical data to compile predictions can give an accurate method to simulate future precipitation data. The correlation values between monthly precipitation and the temperature variables for each station are seen in Table A1.1. Most of the stations have a strong correlation between the average monthly precipitation (AMP) and the observed mean monthly minimum temperature, as well as between the AMP and the observed mean monthly maximum temperature. As seen in Table A1.1, the correlation between AMP

and the RCM historical values is the strongest between AMP and the RCMs CanESM2.CanRCM4 and MPI–ESM–LR.CRCM5, and these RCMs are expected to be important in constructing the predictive algorithms. From the correlations between AMP and the 14 GCMs as found in Table A1.2, we expect the GCMs CESM2-WACCM, INM-CM4-8, MRI-ESM2-0, and NorESM2-MM to have the strongest predictive ability for the Ontario stations. It is noted that there is a slight difference in the correlation structure between the training (1950–1989) and full period (1950–1999). There is a large difference in the most highly correlated GCMs between the training (1950–1989) and testing period (1990–1999). It is important to note that in particular the correlation structure of the 1950's is different to the correlation structure of the 1990's, as seen in Table A1.7, which looks at Station 12 as an example. This project looks at the entire available data from 1950-1999, but it is noted that there are limitations in this approach since the correlation structure of the training period, and the correlation structure within the training period by decade, is different than the correlation structure of the testing period. Once available, data from more recent years (the 2000's and 2010's) should be used for historical prediction. The correlation structure of precipitation from more recent years is expected to help produce a stronger predictive algorithm for the near future precipitation.

For each of the 14 GCMs, the spread of the average monthly simulated data is looked at across the time period of 1950–1999, compared to the observed average monthly precipitation. The plots for each respective station are seen in Fig. A1.1 and Fig. A1.2. There is a large spread for the average monthly precipitation values of each of the GCMs, but it varies amongst the stations. For station LA,

the range between the smallest and largest average GCM values is observed to be the highest for the months of January and August, at a 75mm difference in magnitude. Conversely for station MUA, most of the months have a much more tightly packaged spread of average GCM values, with the largest difference in magnitude being 50mm. Overall, the GCMs are much larger than the average observed monthly precipitation values for the months October to March, and the GCMs are much closer to the average observed monthly precipitation values for the months May to September. On average, the observed values peak for the summer months June to September, and the simulated GCM values peak for the winter months November to January. This analysis shows that none of the models on their own result in a perfect simulation of the historical precipitation values, but a combination (or ensemble) of the models might be better equipped to accurately capture the historical values.

Chapter 3

Methodology

3.1 Background

The time frame that was the focus of this study was from 1950 to 1999, and the historical RCMs and GCMs had data available for the full time period of 1950 to 1999. It is noted that there were some observed missing monthly values as seen in Table 2.5, and the complete data used during analysis was the available observed monthly values from 1950 to 1999. The goal was to maximize the total available observed data that overlapped with the available RCM and GCM data. First, the algorithm was trained on an ensemble of CMIP5 simulated historical RCMs and observed values from 1950 to 1989, for a maximum of 480 data points. The methods were tested using an ensemble of CMIP5 simulated historical data RCMs and observed values from 1990 to 1999, for a maximum of 120 data points. Next, the algorithm was trained on an ensemble of CMIP6 simulated historical GCMs and observed values from 1950 to 1989, and tested using an ensemble of CMIP6 simulated historical data GCMs and observed values from 1990 to 1999,

for a maximum of 120 data points. The performance of each RCM and GCM is rooted in its ability to accurately capture the historical monthly precipitation, and its importance for predicting the observed data of a particular station. The 14 GCMs in Table 2.3 have simulated data available for the three SSPs of interest in Table 2.4. Statistical downscaling is a form of regional climate downscaling, and it is the process of first developing statistical relationships between climate variables and predictors, and applying these relationships to simulate local station data (Busuioc et al., 2001).

The algorithm is trained using nested cross–validation (Tashman, 2000), using `caret` and the `timeslice` method. K-Fold cross–validation (McNicholas and Tait, 2019, p. 97) is not appropriate when looking at time series data because of temporal dependencies. The goal of the cross–validation is to estimate the prediction error of the algorithm on the training data (Fushiki, 2009). To do nested cross–validation, we first must set the seeds for each of the nested predictions for each of the parameter combinations, to make the cross–validation results reproducible. The `initialWindow` is set for each individual station so that there are 40 nested validations, such that the values in the `initialWindow` are used to test one time points 12 units away from the end of the training window. In order to improve the efficiency of the nested cross–validation, the parameter `skip` was set to 8.

3.2 Model Evaluation

The statistical downscaling prediction method is evaluated using RMSE and MAE, comparing the predicted values resulting from the ensemble method to the observed monthly historical precipitation values. For historical precipitation from a sample

of size n, where the observed values are denoted by y_i and the predicted values are denoted by $\hat{y}_i$, the RMSE and MAE values are defined by Eq. 3.1 and Eq. 3.2, respectively. In order to produce the best model for each respective station, the available parameters must be selected in a manner such that the RMSE value is minimized.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{3.2}$$

3.3 Extreme Gradient Boosting

The focus of the XGBoost algorithm is to minimize the objective function, which is comprised of a loss function and a regularization term (McNicholas and Tait, 2019). There are two learning task parameters that can be defined, the objective and evaluation metric. The default objective is regression with the goal of minimizing squared loss, and the evaluation metric in this case would be root mean squared error (RMSE), but can also be updated to a metric like MAE (XGBoostDevelopers, 2020). The algorithm can be used on both regression and classification data, and it is a form of supervised learning, where "labelled data are used to make predictions about unlabelled data" (McNicholas and Tait, 2019). XGBoost uses an ensemble of trees and updates the learner based on the previous tree. For this particular research topic, the method will be regression based. XGBoost has the important property of having tune–able parameters, where a set of parameters are chosen by a grid search during the training of the algorithm on a set of data, and make for very flexible prediction models. Another important property of XGBoost is

that its set of parameters can be tuned to focus on reducing overfitting, and in turn create models that can accurately predict the output of data outside of the training set. The model can also use an L1, or L2 regularization term, defined as α and λ respectively. Through increasing the value of the regularization term, the boosting process is made more conservative, and overfitting is further reduced (McNicholas and Tait, 2019). The regularization terms reduce overfitting, allowing a more accurate test set prediction, and more generalized future simulations. The algorithm starts with a dataset that has n observations and p predictor variables. XGBoost uses an ensemble learning approach, where it builds a model of the form, $y_i = f(x_i) + e_i$, and the formula can be re–arranged so that the error can be written as, $e_i = y_i - f(x_i)$ (McNicholas and Tait, 2019). The data is split into a training set with the labelled data, and a test set with unlabelled data. The function $f(x)$ is called the learner and is constructed based on the training set data, and the error is assessed based on the test set data (McNicholas and Tait, 2019). RMSE is used to assess how well the learner performs. RMSE is highly sensitive to outliers, while MAE is not affected by outliers. If we overfit $f(x)$ to the training data, it may not be able to perform well on the test data, even though the training set error may be very small. The prediction of the i^{th} instance at the t^{th} iteration is defined as $\hat{y_i}^t$, and the goal is to minimize the objective function (Chen and Guestrin, 2016), which is defined in Eq. 3.3. The loss function in Eq. 3.3 is $\sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i))$, and the regularization term is denoted by $\Omega(f_t)$. In terms of tree construction algorithm, for a small dataset, the exact greedy tree construction algorithm will be used, which "enumerates all split candidates" (Zheng et al., 2017). The exact greedy algorithm finds the best split by "enumerating over all the possible splits on all the features" (Chen and Guestrin, 2016), and the split is seen in Eq. 3.4.

The greedy algorithm "starts from a single leaf and iteratively adds branches to the tree" (Chen and Guestrin, 2016).

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{3.3}$$

Letting I_L and I_R be the "instance sets of left and right nodes after the split", we then define the loss reduction after the split by L_{split} (Chen and Guestrin, 2016). In the XGBoost algorithm L_{split} is used in evaluating the split candidates (Chen and Guestrin, 2016). L_{split} is defined in Chen and Guestrin (2016) as Eq. 3.4.

$$L_{split} = \frac{1}{2}\left[\frac{(\sum_{i\epsilon I_L} g_i)^2}{\sum_{i\epsilon I_L} h_i + \lambda} + \frac{(\sum_{i\epsilon I_R} g_i)^2}{\sum_{i\epsilon I_R} h_i + \lambda} - \frac{(\sum_{i\epsilon I} g_i)^2}{\sum_{i\epsilon I} h_i + \lambda}\right] - \gamma \tag{3.4}$$

In terms of variable selection, the score of a feature increases as the variable is used more to make key decisions within boosted trees, and the measures of gain, frequency, and cover are used to calculate importance (Zheng et al., 2017). The focus for this study is on the gain measure, where gain is the mean accuracy improvement brought on by creating a split in a tree on a particular variable across the boosting ensemble (McNicholas and Tait, 2019), and each split "tries to find the best feature and splitting point to optimize the objective" (He, 2016). The XGBoost algorithm calculates the "gain on each node", and at the end, "we look into all the trees, and sum up all the contribution for each feature and treat it as the importance" (He, 2016). The degree to which a variable is important in a single decision tree can be written as the score, $w_l^2(T) = \sum_{t=1}^{J-1} \hat{\tau}_t^{\,2}$, where the decision tree has, "`J-1` internal nodes, and partitions the region into two sub–regions at every node t by the prediction feature" (Zheng et al., 2017). The tree algorithm selects the feature that is estimated to provide the strongest improvement $\hat{\tau}_t^{\,2}$, "in the

squared error risk over that for a constant fit over the entire region" (Zheng et al., 2017). Over the additive `M` trees in XGBoost, the variable importance calculation can be written as Eq. 3.5, as in Zheng et al. (2017).

$$w_l^2(T) = \frac{1}{M} \sum_{m=1}^{M} \hat{\tau}_t^2(T_m). \tag{3.5}$$

3.3.1 XGBoost parameters

The number of boosting iterations is denoted by the parameter `nrounds` (McNicholas and Tait, 2019). The learning rate, also known as the shrinkage parameter, is denoted by η (`eta`), (McNicholas and Tait, 2019) and it is used to prevent overfitting. The step size shrinkage works in such a manner where, "after each boosting step we can directly get the weights of new features and η shrinks the feature weights to make the boosting process more conservative" (XGBoostDevelopers, 2020). Additionally, shrinkage "scales newly added weights by a factor eta after each step of tree boosting [...] shrinkage reduces the influence of each individual tree and leaves space for future trees to improve the model" (Chen and Guestrin, 2016). Another parameter that can help with producing a more conservative model is γ (`gamma`), which is "the minimum loss reduction required to make a further partition on a leaf node of the tree" (XGBoostDevelopers, 2020). Another tune–able parameter is the maximum depth of a tree, `max_depth` and "increasing this value will make the model more complex and more likely to over–fit", so this study will focus on smaller values of this parameter (XGBoostDevelopers, 2020). Another optional parameter that can be altered is α (`alpha`), the L1 regularization term on the weights, which if it is increased past its default value of

1 it will return a more conservative model (XGBoostDevelopers, 2020). The L2 regularization term is denoted by the parameter λ (`lambda`), which reduces model complexity and overfitting (XGBoostDevelopers, 2020). An additional parameter that can help create a more conservative algorithm is minimum child weight (`min_child_weight`), such that if a tree partition step results in a leaf node with the "sum of instance weights less than this parameter value, then the building process will give up further partitioning" (XGBoostDevelopers, 2020). Larger values of `min_child_weight` will return more conservative algorithms. A subsample ratio (`sub_sample`) of the training instances is commonly used to prevent overfitting and is taken prior to growing the trees. This subsampling occurs "once in every boosting iteration" (XGBoostDevelopers, 2020), and the default sampling method is uniform. The subsample ratio of columns when constructing each tree (`colsample_bytree`) can be tuned to be different from its default value of 1, which considers all of the predictor variables at each tree split (XGBoostDevelopers, 2020). `Colsample_bytree` values less than 1 ensure that all variables have a better chance of being chosen at each step of the tree construction. It is noted that column subsampling has a higher ability to prevent overfitting, more so than the traditional row sub–sampling (Chen and Guestrin, 2016), and will be included in parameter tuning measures.

3.3.2 Parameter Tuning

The default parameters are seen in Table 3.1. The best tune will be chosen from a grid search, where an appropriate set of parameter values will be assigned, and the lowest RMSE returned by a certain combination of parameters will be determined to be the best tune on the training set. First, Grid Search 1 was ran as seen in

Table 3.2, to show that extracting a personalized set of parameters for each station can improve the predictive performance of the algorithm. The set of parameters that are part of Grid Search 2, the fine tuned search, are seen in Table 3.3. The goal is to use parameters that provide the best predictive model for each respective station, without overfitting the training data. Through reducing overfitting, the model will produce good results for data that does not directly mimic the training set.

TABLE 3.1: Default XGBoost algorithm parameter values and their range (XGBoostDevelopers, 2020).

Parameter	Number of Rounds	Maximum Tree Depth	Eta (η)	Gamma (γ)	Column Sampling by Tree	Sub Sample	Minimum Child Weight
Default Value	100	6	0.3	0	1	1	1
Available Range of Values	[1, ∞)	[1, ∞)	[0, ∞)	[0, ∞)	(0,1]	(0,1]	[0, ∞)

TABLE 3.2: XGBoost Parameter Grid Search 1: Simple tune

Number of Rounds	Maximum Tree Depth	Eta (η)	Gamma (γ)	Column Sampling by Tree	Sub Sample	Minimum Child Weight
40,50,60,70,80	1,2	0.1,0.2,0.3,0.4	0,0.1,0.2,0.3	0.6,0.7,0.8	0.6,0.7,0.8	1

TABLE 3.3: XGBoost Parameter Grid Search 2: Fine tune

Number of Rounds	Maximum Tree Depth	Eta (η)	Gamma (γ)	Column Sampling by Tree	Sub Sample	Minimum Child Weight
40,45,...,80	1,2,...,5	0.1,0.15,...,0.5	0,0.05,...,0.3	0.5,0.6,...,	0.5,0.6,...,0.9	1

3.4 Historical and Future Prediction

3.4.1 Historical evaluation

Statistical downscaling takes place by training the XGBoost algorithm on the data for the period 1950-1999 using the caret function, rolling average cross–validation,

and the seed is set to 1 for reproducible results. Each climate model is used as a predictive variable in determining the historical monthly precipitation. R version 4.0.2 and RStudio version 1.3.959 were used to run the analysis for this project. The R packages used to conduct the analysis were `tidyverse` (Wickham et al., 2019), `caret` (Kuhn, 2020), `XGBoost` (Tianqi Chen et al., 2020), and `ncdf4` (Pierce, 2019). The XGBoost algorithm is run using the full ensemble of variables and using a reduced ensemble of variables. The goal is to reduce the number of variables in the ensemble while keeping a similar degree of prediction accuracy to the full ensemble. The prediction results from using the default parameters are compared to the prediction results from using Grid Search 1 and Grid Search 2, to demonstrate how the results improve in accuracy after parameter tuning. The L2 regularization term λ was increased to 1.2 to further reducing overfitting. Predictor variables like temperature and month label were introduced to better capture the cyclical nature of monthly precipitation. An ensemble method in the context of climate prediction, is where the output from multiple climate models are used to make predictions for each observed data point (Li et al., 2020). For the CMIP5 Method, the full ensemble is the case where the algorithm is predicting the observed historical monthly precipitation using the 7 CMIP5 RCMs, Mean Monthly Minimum Temperature and Mean Monthly Maximum Temperature. The CMIP6 Method 1 full ensemble involves predicting the observed historical monthly precipitation using the 14 CMIP6 GCMs, Mean Monthly Minimum Temperature, and Mean Monthly Maximum Temperature. The CMIP6 Method 2 full ensemble involves predicting the historical monthly precipitation using the 14 CMIP6 GCMs and the month labels. From these three full variable ensemble methods, the top 2 predictors returned by XGBoost variable importance for each of these methods

were extracted. From the full ensemble of CMIP6 Method 2 predictors, the top 2 variables returned by forward/backward variable selection were extracted. Forward and backward variable selection uses a stepwise regression model similar to that in Chowdhury and Turin (2020). For CMIP6 Method 2, both forward and backward directions were used, and the Akaike information criterion (AIC) was used as the selection criteria for the final model. We looked at the top 2 most significant variables returned from the stepwise regression model. CMIP5 Method and CMIP6 Method 1 were trained using the XGBoost algorithm on a reduced ensemble consisting of the top 2 variables returned by XGBoost variable importance. Finally, for CMIP6 Method 2 the performance of the 2-variable ensemble from both variable selection methods were compared, and the method that returned the higher predictive accuracy was retained as the final model. The final predictive models for these three reduced ensemble methods were tested on the period 1990-1999, to evaluate the predictive accuracy of the XGBoost algorithm in combination with a multi-model ensemble.

3.4.2 Future simulations

After the XGBoost algorithm CMIP6 Method 2: 2–variable ensemble best tune is extracted, the personalized algorithm for each station is applied to simulate future monthly precipitation data. The reduced ensemble method is then applied to simulate the precipitation for the future time period 2020 to 2099. The time periods that are often used to separate the simulation results fall into 30 year intervals 2020–2049, 2050–2079, and 2080–2099. There are 9 sets of simulations in total for each of the 12 stations.

Chapter 4

Results

4.1 CMIP5 Historical Prediction

First, an exploratory analysis was conducted on a smaller set of data from CMIP5 RCMs, to test the ability of XGBoost in predicting precipitation data. For CMIP5 historical daily and monthly precipitation, there are a total of 7 RCMs for Ontario. Previous precipitation prediction research often included various temperature variables like mean monthly temperature, minimum monthly temperature, and maximum monthly temperature as a part of the predictor variables (Du et al., 2017). Previous studies such as Cong and Brady (2012) have shown an interdependence between rainfall and temperature. Other weather variables came up in previous methods of precipitation prediction, and various combinations of variables were attempted for the CMIP5 monthly precipitation data that were subsets of the available weather variables, in addition to the 7 RCMs as predictor variables. The temperature variables were very useful in historical prediction, often appearing in the top 2 predictor variables when assessing variable importance using the

XGBoost algorithm, which may be due to their ability to capture the seasonality in the predictions. Other weather variables like mean monthly wind speed and monthly mean temperature did not improve the predictive ability of the algorithm, as they did not appear to drive the creation of the decision trees. Ultimately, mean minimum monthly temperature and mean maximum monthly temperature were chosen due to their strong importance in historical precipitation prediction. The full CMIP5 ensemble included the 7 RCMs and the two weather variables, mean monthly minimum temperature and mean monthly maximum temperature. From this ensemble of 9 predictor variables, the top 2 most important historical predictors for each station's respective set of training data were used to predict the historical monthly precipitation. The results are seen in Table 4.1 for the CMIP5 reduced ensemble method. These historical predictions are used to evaluate and

TABLE 4.1: CMIP5 historical monthly precipitation prediction using Grid Search 1.

	9 Variable Ensemble		2–variable Ensemble		Top 2 variables	
Station	RMSE	MAE	RMSE	MAE	Variable 1	Variable 2
BTL	24.24	18.92	25.20	20.35	Mean Monthly Min Temp	Mean Monthly Max Temp
LA	40.41	29.75	40.79	29.80	Mean Monthly Min Temp	MPI–ESM–MR.CRCM5
MUA	34.79	26.86	31.73	25.63	Mean Monthly Min Temp	Mean Monthly Max Temp
NB	41.75	32.73	43.63	34.18	Mean Monthly Min Temp	Mean Monthly Max Temp
OMIA	38.84	28.79	33.54	26.68	Mean Monthly Min Temp	CanESM2.CanRCA4
SSM	33.87	26.87	33.54	26.26	Mean Monthly Max Temp	CanESM2.CanRCA4
SL	36.18	26.22	35.52	24.98	Mean Monthly Min Temp	MPI–ESM–MR.CRCM5
TVPA	30.31	23.66	32.16	25.71	Mean Monthly Max Temp	EC–Earth.RCA4
TIA	29.04	23.67	31.85	25.22	Mean Monthly Min Temp	CanESM2.CanRCA4
TPIA	29.04	23.67	31.85	25.22	Mean Monthly Min Temp	CanESM2.CRCM5
WTA	36.48	29.08	36.77	28.66	CanESM2.CanRCA4	CanESM2.CanRCM4
WSA	39.60	30.54	38.38	29.98	Mean Monthly Min Temp	MPI–ESM–MR.CRCM5

validate how well XGBoost can accurately predict monthly precipitation data using the RCMs. The 9–variable ensemble method applied to the historical data

for the 12 stations test period produced RMSE values in the range from 24.24–41.75 mm/month, and MAE values in the range from 18.92–32.73 mm/month. The 2–variable ensemble produced RMSE values in the range from 25.20–43.63 mm/month, and MAE values in the range from 20.35–34.18 mm/month. It is noteworthy that the temperature variables often appeared in the top 2 predictor variables when assessing for variable importance.

4.2 CMIP6 Method 1: Historical prediction

At the time of this project, predicting historical precipitation for Ontario, Canada stations using CMIP6 data has not yet been used in any previous literature, and so it is an area of interest. As of the time of this research, the CMIP6 only has GCM data available. This area of research was extending the findings from the previous CMIP5 data which has been extensively used and studied, to the newer CMIP6 data which has limited literature. The CMIP6 GCM historical data prediction looked at a larger time period of evaluating the ability of XGBoost to predict historical data, and using XGBoost to simulate future monthly precipitation. The XGBoost algorithm was used to train the data, using rolling window cross–validation, for each of the 12 stations for the years 1950 to 1989. The resulting best tune of the algorithm parameters were then applied to predict the precipitation for the years 1990 to 1999. First, the default tuning was used for the XGBoost algorithm, and then this prediction was compared to the results from the Grid Search 1 and Grid Search 2. Through adding the two temperature variables mean minimum monthly temperature and mean maximum monthly temperature

to the ensemble of climate models, there was an improvement in the historical prediction. CMIP6 Method 1 is an ensemble of the 14 GCMs and the two temperature variables, for a total of 16 variables, and the results are in Table 4.2. The 16 variable ensemble method applied to the historical data for the 12 stations test period produced RMSE values in the range from 23.73–42.76 mm/month, and MAE values in the range from 18.42–34.05 mm/month. The reduced 2–variable ensemble method, where variable importance was assessed using the XGBoost algorithm, when applied to the historical data for the 12 stations test period produced RMSE values in the range from 26.12–44.61 mm/month, and MAE values in the range from 20.53–35.33 mm/month. It is important to note that CMIP6 Method 1 may not be good for long term prediction, assuming we do not have observations in the near future.

TABLE 4.2: CMIP6 historical monthly precipitation prediction under Method 1 and fine parameter tuning.

	16 Variable Ensemble		2–variable Ensemble		Top 2 variables (XGBoost algorithm)	
Station	RMSE	MAE	RMSE	MAE	1st variable	2nd variable
BTL	23.73	18.42	26.12	20.53	Mean Monthly Min Temp	Mean Monthly Max Temp
LA	42.76	31.65	42.53	30.81	CESM2–WACCM	MPI–ESM1–2–HR
MUA	32.15	25.93	34.89	26.82	Mean Monthly Min Temp	Mean Monthly Max Temp
NB	39.52	31.62	44.61	35.33	Mean Monthly Min Temp	Mean Monthly Max Temp
OMIA	34.63	26.67	35.04	26.80	Mean Monthly Min Temp	MPI–ESM1–2–HR
SSMA	38.71	30.27	34.52	27.21	Mean Monthly Min Temp	GFDL–ESM4
SLA	35.64	26.05	36.89	26.82	Mean Monthly Min Temp	Mean Monthly Max Temp
TVPA	38.19	30.11	34.87	28.83	Mean Monthly Min Temp	BCC–CSM2-MR
TIA	35.49	29.34	32.77	26.08	BCC–CSM2–MR	CAMS–CSM1–0
TPIA	35.49	29.34	32.77	26.08	INM–CM4–8	Mean Monthly Max Temp
WTA	41.05	32.97	35.28	27.45	CESM2	EC–Earth3–Veg
WSA	40.82	34.05	39.61	31.34	Mean Monthly Min Temp	CESM2–WACCM

4.3 CMIP6 Method 2: Historical prediction

CMIP6 Method 2 consisted of the top 2 variables as chosen from the 14 GCMs and the month label. Since it is not possible to use the observed temperature variables

in the future simulations, instead the factor variable monthly label was added to the method to capture the cyclical nature observed in the monthly precipitation for each station. The algorithm was run using a fine tuned grid search, to evaluate if prediction can be improved from extracting a finer set of individual parameters for each station. For CMIP6 Method 2 there were two algorithms used to extract the top 2 most important variables for predicting the monthly precipitation. First, we extracted the top 2 variables that were returned by XGBoost variable importance measures. Next, the top 2 most important variables from stepwise regression were examined with respect to their p–values, treating the month label as one variable, and finding the most important GCM to predict the monthly observed precipitation. The results from the CMIP6 reduced ensemble method 2 are seen in Table 4.3.

TABLE 4.3: CMIP6 Method 2 prediction accuracy under the 15 variable ensemble, the 2–variable ensemble using forward/backward algorithm variable selection, and the 2–variable ensemble using XGBoost algorithm variable importance.

	15 variable ensemble		2–variable ensemble (XGBoost Selected)		2–variable ensemble (F/B Selected)	
Station	RMSE	MAE	RMSE	MAE	RMSE	MAE
BTL	24.27	19.18	27.73	19.41	22.89	19.22
LA	42.88	32.61	42.08	30.83	40.93	30.14
MUA	32.64	25.66	38.94	32.05	30.17	24.74
NB	40.87	33.15	38.46	30.46	38.74	31.21
OMIA	35.14	26.85	36.13	27.69	33.96	26.70
SSMA	37.29	30.77	34.66	27.58	33.14	26.63
SLA	38.11	27.86	48.11	35.35	38.40	27.72
TVPA	35.12	28.14	39.72	31.01	29.60	23.13
TIA	30.60	24.01	32.77	26.08	30.15	24.52
TPIA	30.60	24.01	32.77	26.08	30.15	24.52
WTA	42.66	32.66	35.53	27.79	36.05	27.62
WSA	40.40	32.56	39.20	31.14	36.50	28.47

The 15 variable ensemble method applied to the historical data for the 12 stations test period produced RMSE values in the range from 24.27–42.88 mm/month, and MAE values in the range from 19.18–33.15 mm/month. The 2–variable ensemble method, where variable importance was assessed using the XGBoost algorithm, when applied to the historical data for the 12 stations test period produced RMSE values in the range from 27.73–48.11 mm/month, and MAE values in the range from 19.41–35.35 mm/month. The 2–variable ensemble where variable importance was assessed using the forward/backward selection algorithm, produced RMSE values in the range from 22.89–40.93 mm/month, and MAE values in the range from 19.22–31.21 mm/month.

Table 4.4: CMIP6 Method 2: Top 2 Variables

	Top 2 variables XGBoost selected		Top 2 variables Forward/Backward selected	
Station	1st variable	2nd variable	1st variable	2nd variable
BTL	month label	INM–CM4–8	month label	EC–Earth3
LA	CESM2–WACCM	FIO–ESM–2–0	month label	FIO–ESM–2–0
MUA	CESM2	EC–Earth3–Veg	month label	MRI–ESM2–0
NB	month label	NorESM2–MM	month label	NorESM2–MM
OMIA	MPI–ESM1–2–HR	CAMS–CSM1–0	month label	BCC–CSM2–MR
SSMA	NorESM2–MM	FIO–ESM–2–0	month label	INM–CM4–8
SLA	EC–Earth3	INM–CM4–8	month label	EC–Earth3–Veg
TVPA	CAMS–CSM1–0	EC–Earth3–Veg	month label	CAMS–CSM1–0
TIA	BCC–CSM2–MR	CAMS–CSM1–0	month label	NorESM2–MM
TPIA	BCC–CSM2–MR	EC–Earth3	month label	FIO–ESM–2–0
WTA	CAMS–CSM1–0	INM–CM4–8	month label	FIO–ESM–2–0
WSA	EC–Earth3–Veg	CESM2–WACCM	month label	FIO–ESM–2–0

The CMIP6 reduced ensemble method 2 using the top 2 variables returned by the forward/backward algorithm provided the lowest overall RMSE and MAE values, as seen in Table 4.3. The CMIP6 Method 2 top 2 variables used for the 2–variable ensemble using forward/backward algorithm variable selection, are compared to the top 2 variables used in the 2–variable ensemble using XGBoost

algorithm variable importance in Table 4.4. As seen in Table 4.4, the month label often appeared in the top 2 predictor variables when assessing variable importance. The historical predictions for the 12 climate stations are seen in Fig. 4.1 and Fig. 4.2. The predictive results of the full ensembles are very similar to the reduced ensembles, and using the reduced ensembles improves the efficiency of the predictive algorithm. Through tuning the parameters to be more personalized to each station than the default parameters, the predictive accuracy of the algorithm is improved.

4.4 Method 2: CMIP6 Future Projections

The focus of the future results was to find the average change in precipitation across the 12 stations, to get an overall sense for the future monthly precipitation trends across Ontario. One of the main assumptions we are making is that accurately capturing the historical precipitation using the GCM data will allow us to accurately predict future monthly precipitation data for those same stations. The change in precipitation refers to the difference between the most recent 30 years of historical average monthly observations and the simulated future average monthly data, and the relative change in precipitation refers to the relative difference between the simulated future average monthly data and the most recent 30 years of historical average monthly observations. The most recent 30 years of historical monthly observations are 1963–1992 for Big Trout Lake, 1965–1994 for Toronto Island, and 1970–1999 for the 10 remaining stations. The differences between the historical precipitation and simulated precipitation are calculated to

show the change in precipitation in accordance to each of the three SSP. The difference is defined as the change in precipitation for the simulated period, from the historical period. The relative difference in precipitation takes into account the magnitude of the difference in comparison with the monthly historical average precipitation. Setting the historical average monthly precipitation values be defined as $x_{i,j}$ and future average monthly precipitation values be defined as $y_{i,j,k}$, where `i` = month, `j` = station, and `k` = future time period. The differences are calculated for the twelve months, twelve stations, and three future time periods, with `i`=1–12, `j`=1–12, and `k`=1–3. The difference in precipitation is defined by Eq. 4.1 and the relative difference in precipitation is defined by Eq. 4.2. The average change in precipitation across the 12 stations, between the historical and future time periods are seen in Fig. 4.3.

$$\delta_{i,j,k} = y_{i,j,k} - x_{i,j} \tag{4.1}$$

$$\Delta_{i,j,k} = \frac{y_{i,j,k} - x_{i,j}}{x_{i,j}} = \frac{\delta_{i,j,k}}{x_{i,j}} \tag{4.2}$$

The results are noticeably different if instead the focus is on comparing the simulated future predictions and the simulated historical predictions. Since the future predictions are based on the historical predictions, the difference between the future predictions and historical predictions are smaller than the difference between the future predictions and historical observations. The difference between the simulated historical predictions (1990-1999) and the simulated future predictions are seen in Fig. A1.3. There is an expected increase in precipitation for the winter months, and an overall expected decrease in precipitation for the summer months. These changes are heightened as the time periods move from more recent future (2020-2049) to more the distant future (2080-2099).

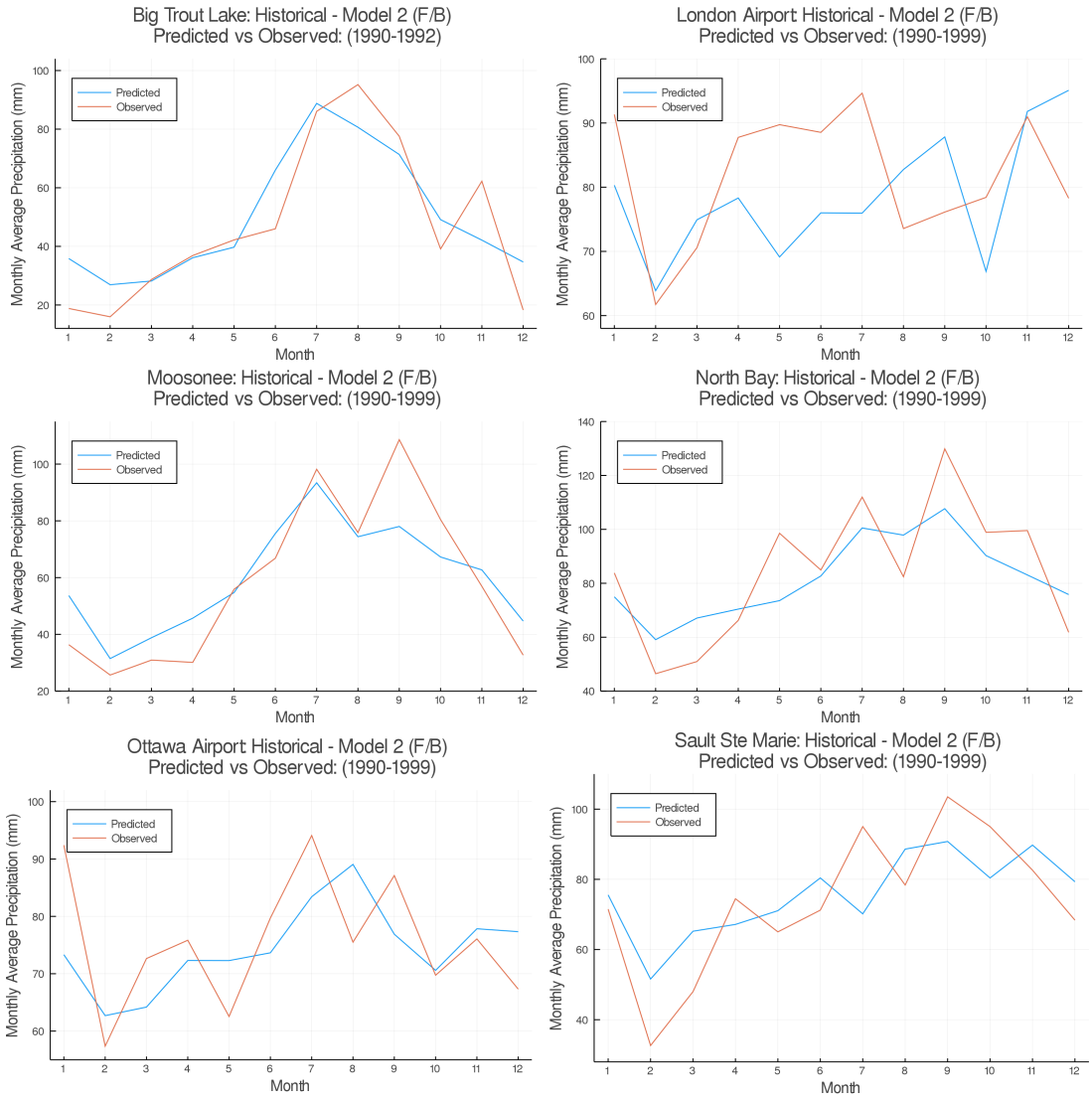

FIGURE 4.1: CMIP6 Method 2 reduced 2–variable ensemble (forward/backward algorithm variable selection). Average monthly predicted precipitation versus average observed monthly precipitation, for stations 1–6.

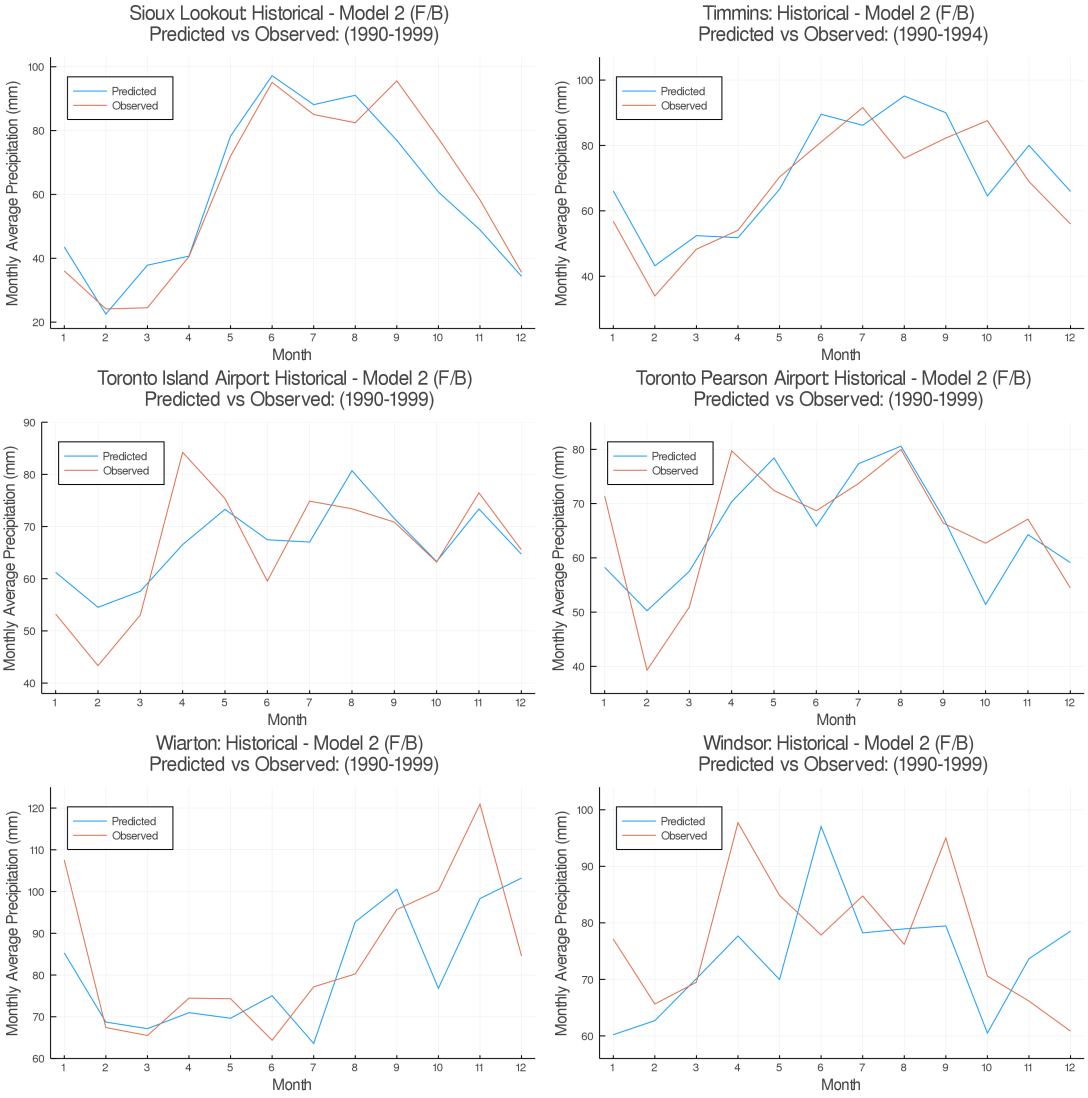

FIGURE 4.2: CMIP6 Method 2 reduced 2–variable ensemble (forward/backward algorithm variable selection). Average monthly predicted precipitation versus average observed monthly precipitation, for stations 7–12.

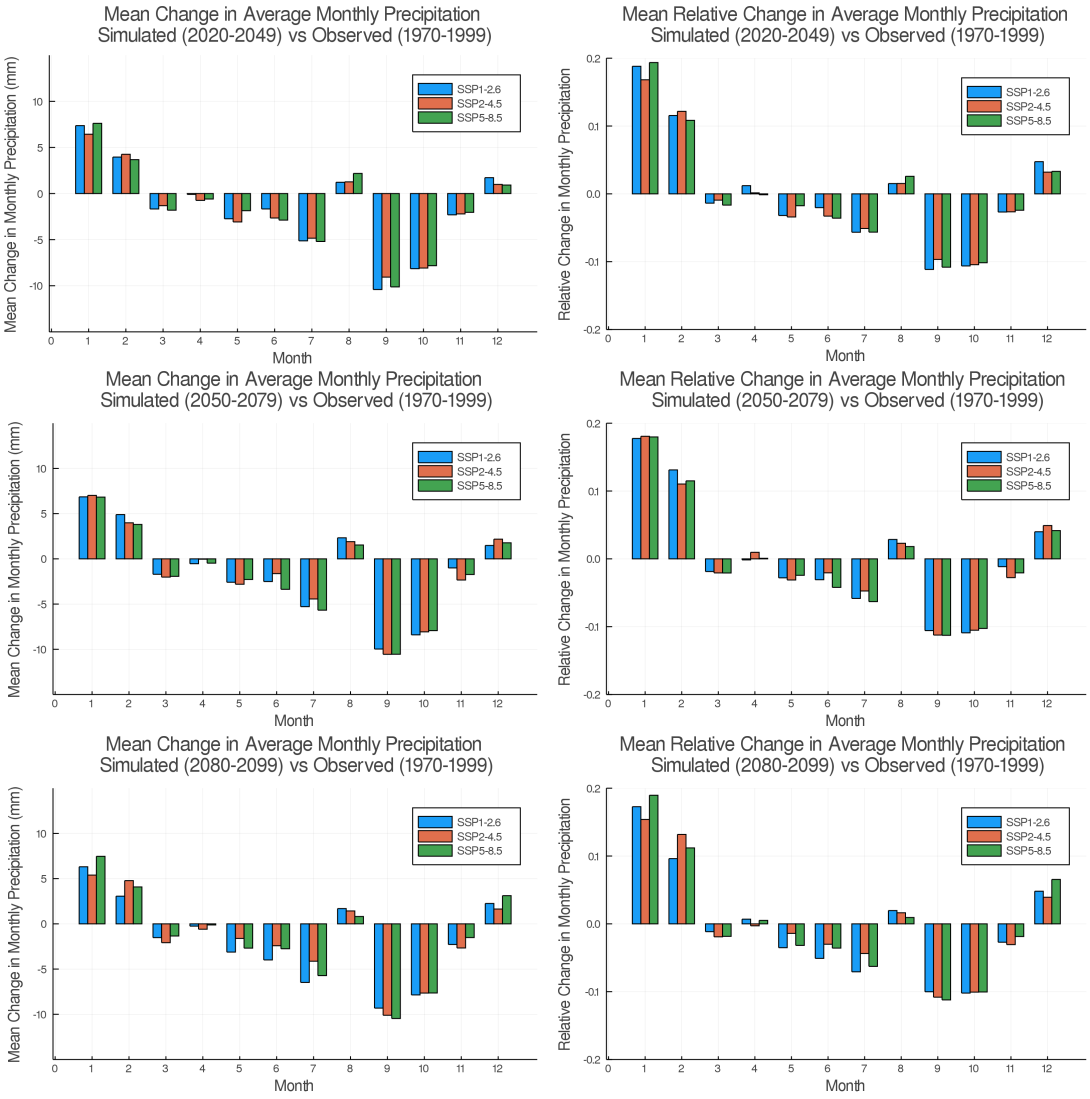

FIGURE 4.3: The average change, and average relative change, in monthly precipitation between the simulated future precipitation and observed historical precipitation.

Chapter 5

Discussion

5.1 Evaluating Model Performance

The historical monthly precipitation prediction was used to evaluate the accuracy of the chosen statistical downscaling method of the XGBoost algorithm and an ensemble of climate models. The prediction accuracy of the downscaling was assessed using the metrics of RMSE and MAE, where the lower value of these metrics equates with a stronger predictive ability. The results from the XGBoost algorithm were compared to other prediction methods like regression, decision trees, and regular boosting. XGBoost returned the lowest RMSE scores, which validated its potential in predicting precipitation data. The first useful finding was that weather variables and the month label, in combination with the climate models, are helpful in predicting historical monthly precipitation for Ontario stations. One important result coming from the historical precipitation prediction is that XGBoost is moderately useful for predicting historical daily precipitation, but it is more useful in constructing a model for historical monthly precipitation.

Extreme daily precipitation is often defined as the top 1% of wet days (Agel et al., 2015). Extreme monthly precipitation does not have one single definition, and for this project it is defined as the top 1%, or the 99th percentile of monthly precipitation for each climate station. The concept of extreme monthly precipitation is very station dependent and varies across different regions. There was an attempt to classify extreme precipitation events and remove these extreme events. However, the results from removing the extreme monthly events were very similar in predictive ability to the results which included all monthly data points, and thus the extreme events were not removed for the CMIP6 predictive analyses moving forward. Using a reduced 2–variable ensemble returned very similar, and often more accurate results than using a reduced 3–variable ensemble, and so the final reduced model for each method included a 2–variable ensemble. It is noted that the months October and December had the weakest predictions, and the largest bias, when applying down–scaling method on the Ontario stations. On average, December had the most over–predicted historical monthly precipitation, and on average October had the most under–predicted historical monthly precipitation. The differences between the predicted and observed monthly precipitation are seen in Table A1.6.

An important result from the historical prediction is that using a multi–model ensemble of climate model outputs, in addition to weather variables like temperature and the factor variable month label, improves the ability of the XGBoost algorithm to capture the cyclical nature of precipitation trends at each climate station. Weather variables like mean minimum monthly temperature and mean

maximum monthly temperature, are valued as important by the XGBoost algorithm in the pursuit of predicting historical monthly precipitation. In the absence of observed weather data, the month label is also very useful in helping predict monthly historical precipitation for Ontario stations. It is noted that XGBoost is able to predict the precipitation at a level which is similar to other previous methods used for predicting historical precipitation (Wang et al., 2014), but it is limited in terms of its accuracy due to its difficulty in capturing the output of more extreme values. Through tuning the XGBoost algorithm parameters with a focus on reducing overfitting, we can further improve the prediction of historical precipitation. In particular, the parameter tuning helps produce lower RMSE output for the test period historical monthly precipitation. The final results are relevant since the predictive ability of XGBoost proved to be superior in performance with respect to its RMSE values when compared to a variety of methods like neural networks, random forests, and linear regression. The XGBoost algorithm was able

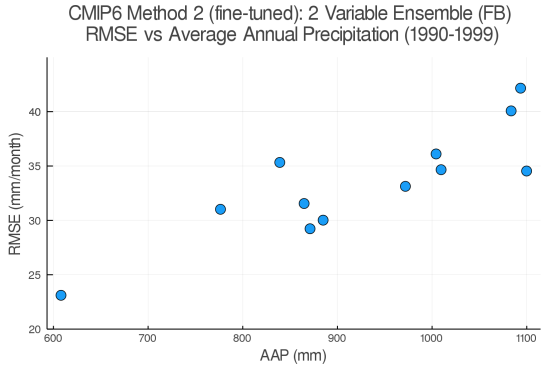

FIGURE 5.1: CMIP6 Method 2 prediction accuracy versus AAP, where prediction accuracy is measured using RMSE.

to accurately capture some of the variability in the monthly average precipitation for the test set data, however, it is important to note that the final predictive results did have a large degree of bias for the larger observed monthly precipitation

values. It is significant to note that the average annual precipitation is highly correlated with the RMSE values returned by CMIP6 Method 2 ($\rho = 0.821$). Fig. 5.1 shows the RMSE values returned by CMIP6 Method 2, plotted against the average annual precipitation, for each of the 12 stations.

5.1.1 CMIP5 Method

Looking at the CMIP5 historical predictions, the top 2 variables often included the temperature variables mean minimum monthly temperature and mean maximum monthly temperature. Additionally, the CMIP5 historical predictions were strongly influenced by the downscaled GCM–RCM pairings of EC–EARTH.RCA4 and CanESM2.RCA4. It is important to note that in Fig. 5.2 the performance of the CMIP5 RCMs are very similar in the predictive ability of the CMIP6 GCMs, for both the full ensemble and the reduced 2–variable ensemble. In particular, the CMIP5 method under Grid Search 1 returns results similar to the CMIP6 Method 2 under Grid Search 2. The CMIP5 method performs better prediction under the full ensemble, than under the reduced ensemble.

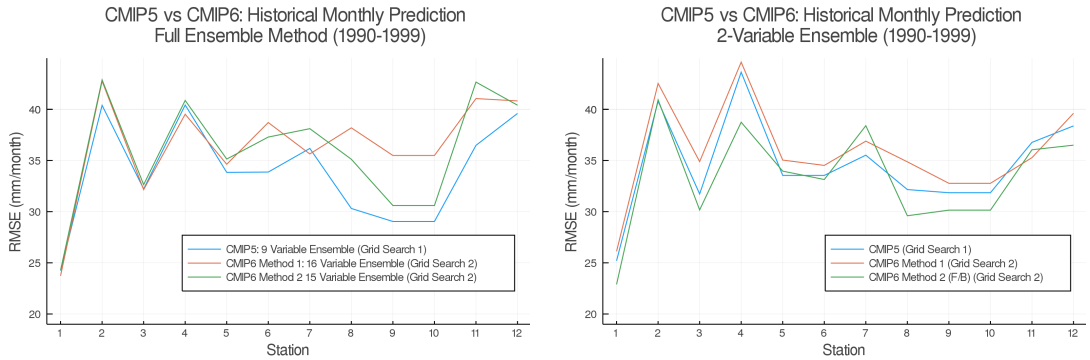

FIGURE 5.2: Comparing the prediction results from the CMIP5 Method to the results from CMIP6 Method 1 and Method 2.

5.1.2 CMIP6 Method 1

The 16 variable ensemble model applied to the historical data is compared to the 2–variable ensemble using the top 2 variables returned by the XGBoost algorithm in Fig. 5.3. It is important to note that the full 16 variable ensemble returned very similar results to the 2–variable ensemble. The two temperature variables often appeared in the top 2 predictor variables when assessing variable importance in for the 2–variable ensemble.

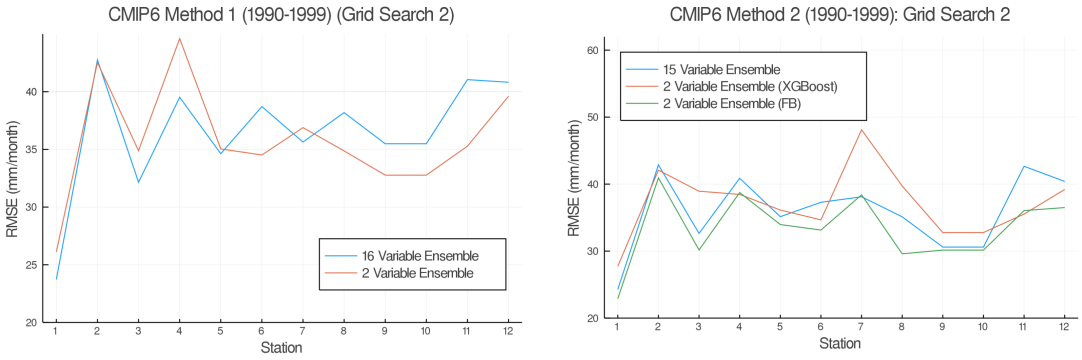

FIGURE 5.3: Comparing the prediction results from CMIP6 Method 1 and CMIP6 Method 2.

5.1.3 CMIP6 Method 2

Fig. 5.3 compares the 15–variable ensemble to the reduced 2–variable ensemble using the top 2 variables returned by the forward/backward algorithm, and the 2–variable ensemble using the top 2 variables returned by the XGBoost algorithm. It is noteworthy that the month label often appeared in the top 2 predictor variables when assessing variable importance in this final reduced model. As well, the full 15–variable ensemble returned very similar results to the reduced 2–variable ensemble. The RMSE values for CMIP6 Method 2 were similar to previous literature

on precipitation prediction including Ontario and global stations. It is noted that there is a limited amount of available literature on the 12 Ontario stations that are the focus of this research, so there are limited results to compare to the historical prediction in this project (Wang et al., 2014; Li et al., 2020). The lowest RMSE values occurred for the more northern stations of BTL, MUA and TVPA, and the more southern stations of WSA, NB and LA had some the largest RMSE values from historical prediction. There were no specific patterns for stations located near water versus the stations that are located more in–land. For the CMIP6 historical predictions using Method 2 (F/B) as seen in Table 4.4, the most important variables across all 12 stations for monthly precipitation prediction were the month label, and the GCMs FIO–ESM–2–0 and NorESM2–MM. In particular, the top predictor was the month label which was the top variable for all 12 of the stations, and the most important GCM was FIO–ESM–2–0, which was in the top 2 variables for 4 of the stations.

It is interesting to note that both of the RCMs that showed up the most in the top 2 variables did not show up as being the most highly correlated with the average monthly observed precipitation. This may have occurred since there is a difference in the correlation structure between the training and testing periods, as well as between each of the decades. The RCM CanESM2.RCA4 may have been useful in accurately predicting the average precipitation for Ontario stations due to its use of the Canadian Model of Ocean Carbon, and the Canadian Terrestrial Ecosystem Model (GOC, 2018). It is interesting to note that the GCM FIO–ESM–2–0 which showed up the most in the top 2 variables was not the most highly correlated GCM with the average monthly observed precipitation for any of the 12 stations. The

GCM FIO–ESM–2–0 may have been useful in accurately predicting precipitation for Ontario stations due to its strong overlap with average observed monthly data, as seen in Fig. A1.2 for stations 11 and 12, for the months June–September. The GCM NorESM2–MM may have been useful in accurately predicting precipitation for Ontario stations due to its low bias for total precipitation rate in the Ontario region, as seen in Fig. 20 of Seland et al. (2020).

5.2 Improving the accuracy and efficiency of the predictive methods

A grid search was used to select parameters that would improve individual station predictive accuracy, using rolling window cross–validation. Parameters like γ and λ were used to reduce overfitting the algorithm to the training data. Variable selection was used to improve the efficiency of the predictive model. Two different variable selection methods were used to extract the top 2 variables that had the strongest predictive importance: variable importance calculated by the XGBoost algorithm, and forward/backward variable selection. The forward/backward variable selection was assessed using the Akaike information criterion, and the XGBoost algorithm variable importance was assessed using the *gain* measure. The two methods returned slightly different top 2 predictor variables, and the top 2 variables as returned by the variable selection method forward/backward selection returned the most accurate downscaled predictions. Using the 2–variable ensemble to perform statistical downscaling returned similarly accurate results to down–scaling using the full ensemble. Temperature data and the factor variable month label were used to capture the cyclical trends over the period of a year. In

particular, the mean monthly minimum temperature had a very strong impact on predicting observed monthly precipitation for the majority of stations, and it was often ranked in the top 2 variables by variable selection algorithms.

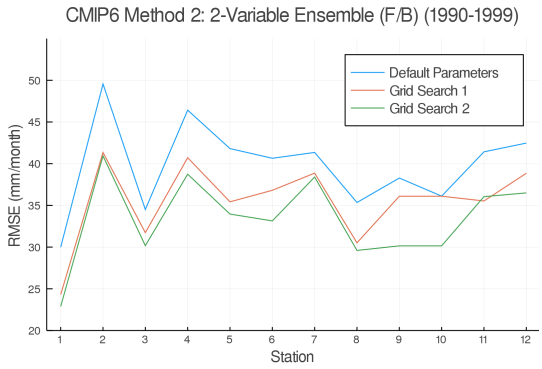

FIGURE 5.4: Comparing the prediction results from various levels of parameter tuning using CMIP6 Method 2 reduced 2–variable ensemble (variables from forward/backward selection algorithm).

There was an improvement in the predictive performance through tuning the XGBoost algorithm parameters using a grid search, as seen in Fig. 5.4. Through tuning the parameters from the default values to the fine tuned grid search, there was a reduction in the RMSE values of up to 10mm/month. It is noted however that the more fine tuned grid searches did not produce remarkably better results, and for some stations they showed only a slight improvement. These results show that a less detailed grid search (Grid Search 1: simple tuning) can perform equally as well as a more fine–tuned parameter grid search (Grid Search 2: fine tuning). A more detailed fine tuned grid search was also attempted, but it only improved the RMSE values by a very small degree. Grid Search 2 returned values similar in accuracy to the more detailed fine tuned grid search, but it is more computationally efficient, taking less time and resources on the server.

5.3 Future Precipitation Simulations

The results from CMIP6 Method 2 are used to simulate future monthly precipitation for the years 2020–2099 in Ontario, under the three SSP as seen in Table 2.4. The goal of the future simulations is to estimate the expected average change in precipitation between the historical and the future time periods. The difference between the most recent 30 years of historical monthly observations and the future simulated monthly data, as well as the relative difference between these two time periods were used to estimate the expected change in average monthly precipitation. The mean relative change is a better representation of the significance of the change in precipitation relative to the observed historical values, while the mean change in average precipitation is a better representation of the magnitude of the change in monthly precipitation.

The smallest change across all the stations, time periods, and SSPs was seen for the month April, and the largest change across all the stations, time periods, and SSPs was seen for the month September. As seen in Fig. 4.3, on average there is an expected increase in precipitation for the months January and February and an expected decrease in precipitation for the months July, September and October, across all time periods and SSPs. For the three different time periods the expected change in monthly precipitation is very similar across the three emission scenarios. The change in precipitation is very similar for future time period 2050–2079 across the three SSPs, and the most different for future time period 2080–2099 across the three SSPs. The relative change is very similar for future time period 2050–2079 across the three SSPs, and the most different for future time period 2080–2099 across the three SSPs. The smallest relative change across all the stations, time

periods, and SSPs occurred for August, and the largest relative change occurred for September. It is notable that the largest relative change in precipitation occurs for the month of January, at a 17–20% increase in precipitation between the two time periods. Comparing the mean change in average precipitation between the future time periods, the months March, April and August are the most similar across the three time periods, and the months January, February and December are the most different across the three time periods. June displayed the largest difference in precipitation between the two time periods 2020–2049 and 2050–2079 of 2.32 mm/month under SSP1–2.6. January shows the largest difference between the two time periods 2050–2079 and 2080–2099 of 1.63mm/month under SSP2–4.5. December displayed the largest difference in precipitation between the two time periods 2020–2049 and 2080–2099 of 2.20 mm/month under SSP5–8.5. Comparing the mean relative change in average precipitation between the future time periods under the three SSPs, the months March, April, and October are the most similar across the three time periods scenarios, and the months January, June and December are the most different across the three time periods. In particular, December shows the largest difference between the two time periods 2020–2049 and 2080–2099 of 3.24% under SSP5–8.5. June shows the largest difference between the two time periods 2020–2049 and 2080–2099 under SSP2–4.5 of 2.01%, and January shows the largest difference between the two time periods 2020–2049 and 2050–2079 under SSP2–4.5 of 2.68%. February shows the largest difference between the two time periods 2020–2049 and 2080–2099 of 1.25%. The differences between the time periods and emissions levels are overall very small, but there is an overall larger decrease in precipitation for the summer and fall months, and increase for the winter months, as the greenhouse gas emissions increase in magnitude.

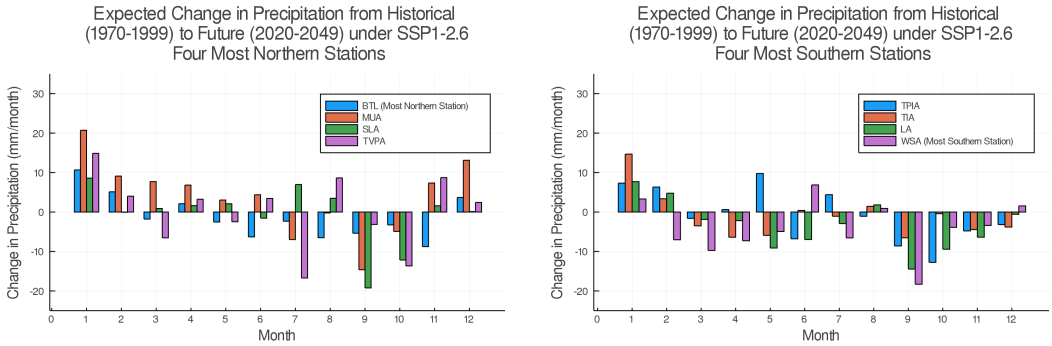

FIGURE 5.5: Comparing simulation results (change in precipitation) from the CMIP6 Method 2 reduced 2–variable ensemble, between the most northern stations and the most southern stations.

Looking at the most northern and the most southern stations with respect to the expected change and relative change in precipitation under SSP1-2.6, from the historical time period to the most near future time period 2020–2049. It is noted in Fig. 5.5 that the change in precipitation between the most northern stations and the most southern is the most different for months July, November, and December. In particular, on average there is a larger decrease in the change in precipitation for July for the more northern stations, compared to the more southern stations. Additionally, there is a larger increase in the change in precipitation for both November and December for the more northern stations. As seen in Fig. 5.6, the overall trend for relative change in precipitation is that on average the more northern stations have a larger increase, and the more southern stations have a larger decrease on average. The relative change in precipitation between the most northern stations and the most southern is the most different for months January–March, and November–December. On average, there is a larger increase in the relative change in precipitation for January–March for the more northern stations, compared to the more southern stations. There is a larger increase in the change in precipitation for November and December for the more northern

stations.

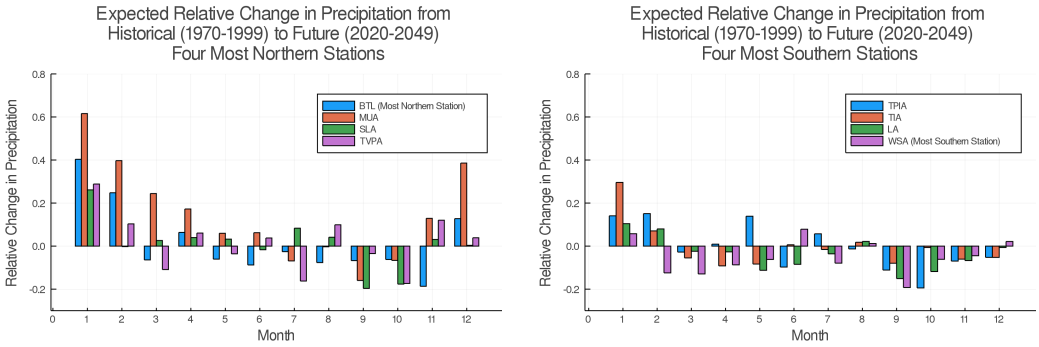

FIGURE 5.6: Comparing simulation results (relative change) from the CMIP6 Method 2 reduced 2–variable ensemble, between the most northern stations and the most southern stations.

Chapter 6

Conclusions and Future Work

This project investigated the ability of XGBoost and an ensemble of climate models in predicting long term precipitation in Ontario. The predictive performance varies across the 12 climate stations, and RMSE increases relative to the average annual precipitation. Variable selection and parameter tuning contributed to improving the efficiency and accuracy of the hindcast prediction. The predictive performance of the algorithm is also improved through adding weather variables and the month label. This may be due to the cyclical nature of these variables which improves their ability to capture the cyclical trend of monthly precipitation. There is an issue with XGBoost being unable to capture the larger magnitudes of precipitation, which is in part due to the focus of the algorithm in minimizing RMSE, which then also minimizes the variance. To accurately capture more near future predictions, the most recent historical data should be used when available, which would contain a correlation structure more similar to that of the near future.

The predictions should be extended to cover more stations across Ontario to confirm trends across various regions in the province. XGBoost is limited in its

ability to accurately capture extreme values, as seen between the maximum observed values, and the maximum predicted values, and this is an area that can be further improved in future work. The prediction of monthly precipitation could be improved by increasing the available training set of data, or adding an extreme classification probability to each observed value, determining the chance of that value being classified as extreme for a given station. Once the historical extreme values are properly captured, the resulting ensemble method can be used to simulate more accurate future monthly predictions. It is also possible to use a combination of XGBoost and another machine learning algorithm, such as using a Gaussian Mixture Model (GMM). The GMM could be used to cluster the training data into `M` clusters, and the XGBoost algorithm could be applied to each of the `M` clusters. Next, the probability of belonging to each of the `M` clusters can be assigned for each of the observations in the test set. Each of the observations in the test set could be predicted using the `M` different XGBoost algorithms. The predicted values returned from each of the `M` algorithms could then be weighted using the probability of each test data point belonging to each of the `M` clusters, similar to the work done in Ni et al. (2020). Another suggestion for future work is upgrading CMIP6 Method 2 to use the monthly temperature GCMs to capture the monthly cyclical trends in precipitation. Temperature simulations *tmax* and *tmin*, similar to the observed mean monthly maximum temperature and mean monthly minimum temperature, might produce useful results as in the case of the historical predictions using CMIP6 Method 1. Finally, the upgraded CMIP6 Method 2 future predictions could be compared to the current CMIP6 future predictions to see if the simulated temperature GCMs would result in very different predictions under the three future climate scenarios.

Appendix A

Appendix

TABLE A1.1: Assessing correlations between the RCM values and observed average monthly precipitation (AMP) (1950–1999), between AMP and Mean Monthly Maximum Temperature, and between AMP and Mean Monthly Minimum Temperature (1950–1999).

Station	RCM: Highest Correlation with AMP	Correlation Values $\rho_{(RCM,AMP)}$	$\rho_{(maxtemp,AMP)}$	$\rho_{(mintemp,AMP)}$
BTL	CanESM2.CanRCM4	-0.739	0.892	0.930
LA	MPI-ESM-LR.CRCM5	0.425	0.244	0.297
MUA	CanESM2.CanRCM4	-0.631	0.882	0.916
NB	CanESM2.CanRCM4	0.515	0.749	0.790
OMIA	CanESM2.CanRCM4	-0.531	0.725	0.760
SSMA	MPI-ESM-LR.CRCM5	0.600	0.479	0.588
SLA	CanESM2.CanRCM4	-0.790	0.925	0.949
TVPA	CanESM2.CanRCM4	-0.605	0.818	0.860
TIA	CanESM2.CanRCM4	-0.468	0.716	0.750
TPIA	CanESM2.CanRCM4	-0.691	-0.691	0.867
WTA	MPI-ESM-LR.CRCM5	0.783	-0.218	-0.124
WSA	CanESM2.CanRCM4	-0.756	0.797	0.794

TABLE A1.2: Assessing correlations between AMP and average monthly GCM values, for three time periods. Readers should refer to Table 2.2 to find the corresponding GCM names.

Time Period	1950–1999		1950–1989		1990–1999	
Station	GCM	$\rho_{GCM,AMP}$	GCM	$\rho_{GCM,AMP}$	GCM	$\rho_{GCM,AMP}$
BTL	4	-0.323	4	-0.343	14	0.431
LA	13	0.435	13	0.536	5	-0.479
MUA	6	-0.275	6	-0.313	4	-0.266
NB	14	-0.333	14	0.349	7	0.290
OMIA	1	-0.333	13	0.360	14	0.590
SSMA	13	0.554	13	0.606	14	0.647
SLA	4	-0.493	6	-0.551	1	-0.382
TVPA	14	0.250	14	0.245	10	-0.386
TIA	10	-0.509	10	-0.453	10	-0.593
TPIA	10	-0.680	10	-0.677	12	-0.612
WTA	13	0.846	11	0.853	13	0.765
WSA	4	-0.751	4	-0.692	12	-0.751

TABLE A1.3: CMIP5 Method reduced 2–Variable Ensemble, where the parameters are tuned using Grid Search 1, parameter tuning results.

Station	Number of Rounds	Maximum Tree Depth	Eta (η)	Gamma (γ)	Column Sampling by Tree	Minimum Child Weight	Sub Sample
BTL	40	1	0.1	0.2	0.7	1	0.8
LA	40	2	0.1	0.3	0.8	1	0.7
MUA	40	1	0.1	0.3	0.7	1	0.6
NB	60	2	0.3	0.3	0.8	1	0.8
OMIA	70	1	0.4	0.3	0.7	1	0.8
SSMA	40	1	0.3	0.1	0.6	1	0.7
SLA	40	1	0.2	0.1	0.6	1	0.6
TVPA	40	1	0.4	0.3	0.7	1	0.7
TIA	40	2	0.4	0.3	0.6	1	0.6
TPIA	40	1	0.1	0.3	0.6	1	0.6
WTA	40	2	0.4	0.3	0.7	1	0.6
WSA	40	1	0.1	0	0.8	1	0.8

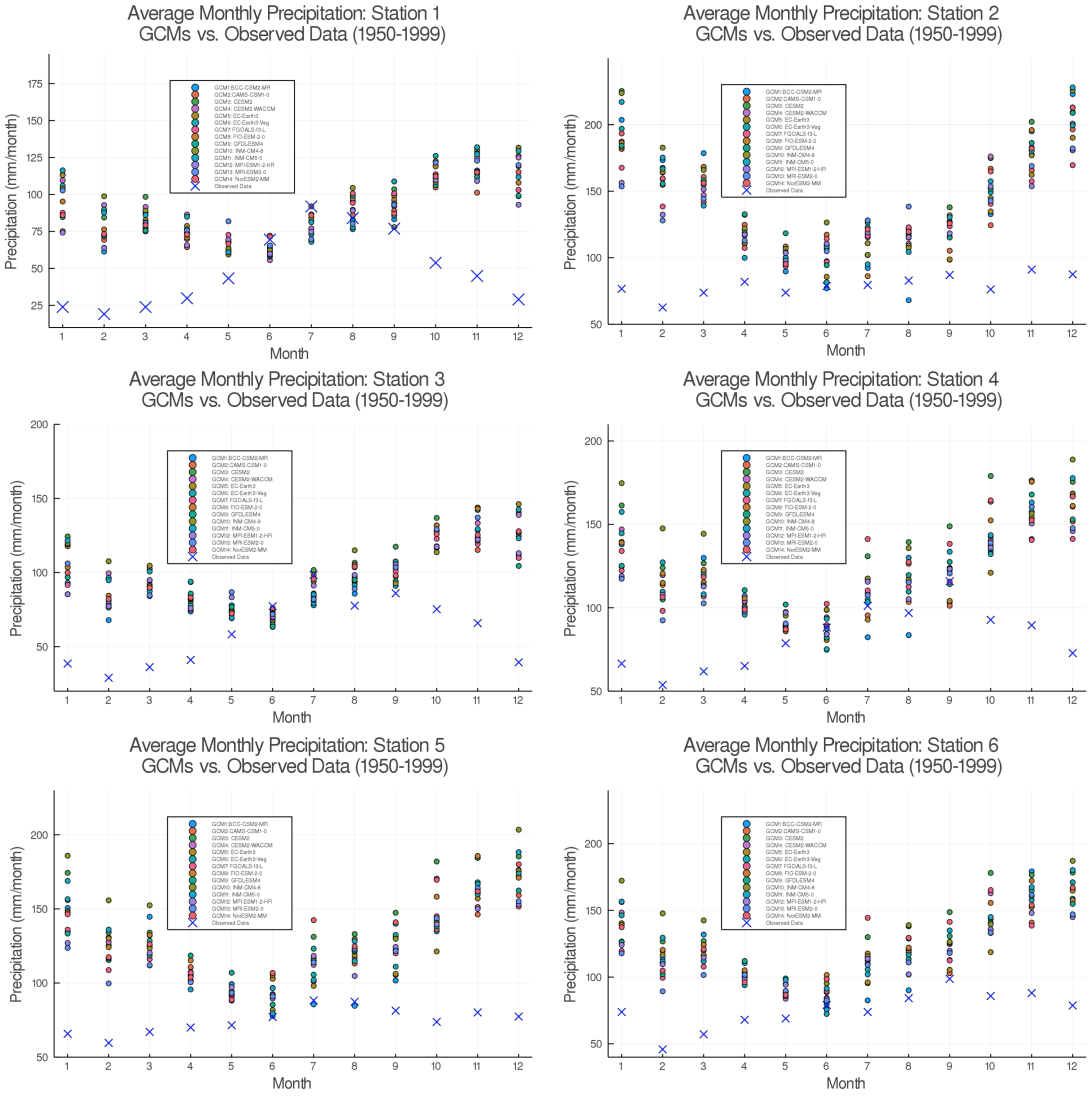

FIGURE A1.1: Average monthly precipitation for the individual GCMs versus AMP, stations 1-6.

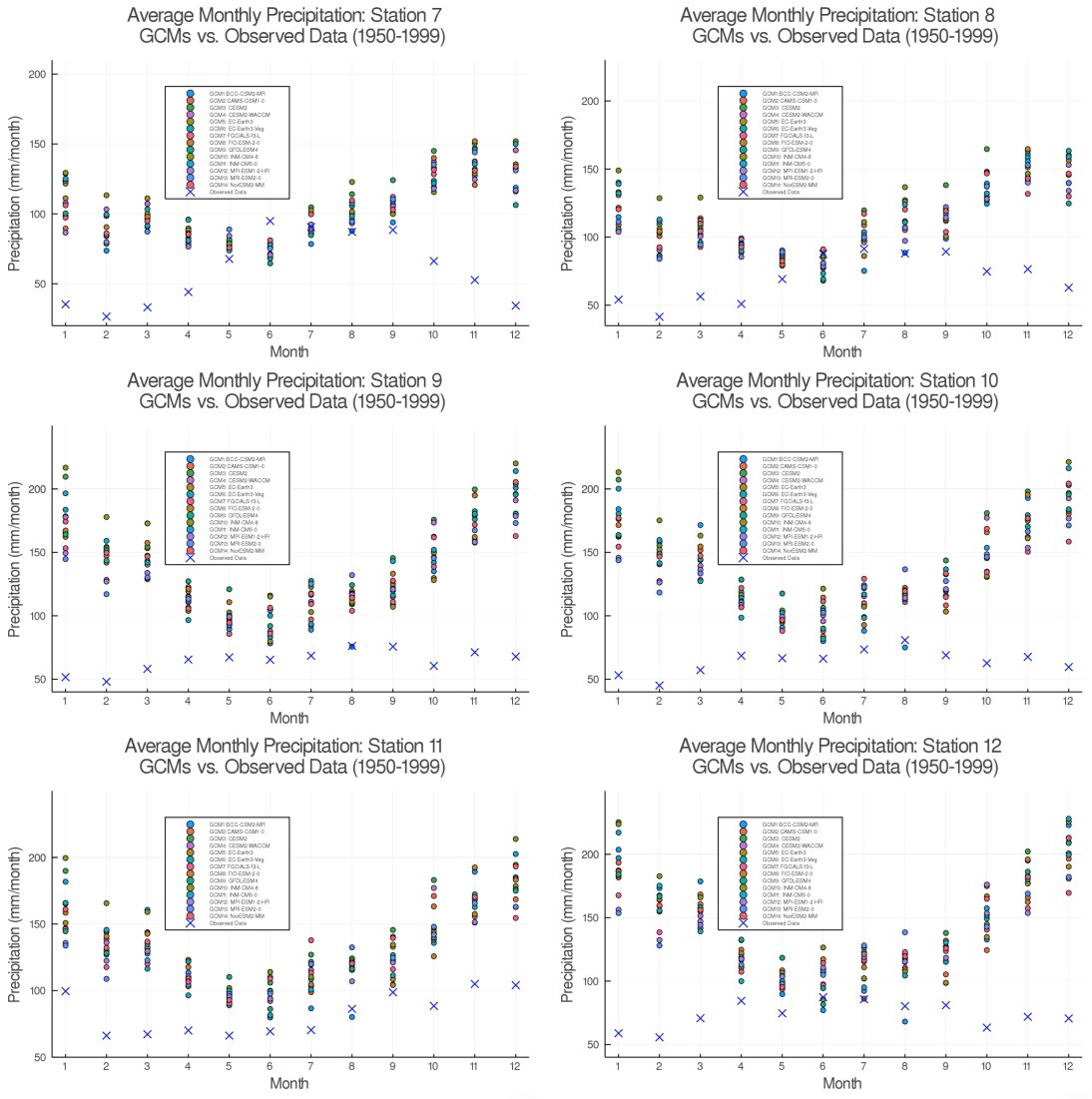

FIGURE A1.2: Average monthly precipitation for the individual GCMs versus AMP, stations 7-12.

TABLE A1.4: CMIP6 Method 1 reduced 2–Variable Ensemble, where the parameters are tuned using Grid Search 2, parameter tuning results.

Station	Number of Rounds	Maximum Tree Depth	Eta (η)	Gamma (γ)	Column Sampling by Tree	Minimum Child Weight	Sub Sample
BTL	45	1	0.45	0.3	0.7	1	0.5
LA	70	1	0.5	0.05	0.9	1	0.5
MUA	60	2	0.5	0	0.5	1	0.6
NB	45	2	0.45	0.05	0.9	1	0.5
OMIA	45	2	0.2	0.3	0.9	1	0.8
SSMA	50	2	0.45	0.3	0.9	1	0.5
SLA	45	3	0.5	0.15	0.5	1	0.5
TVPA	40	2	0.5	0.25	0.7	1	0.5
TIA	45	1	0.5	0.3	0.6	1	0.5
TPIA	75	2	0.3	0.15	0.8	1	0.5
WTA	45	2	0.15	0.1	0.7	1	0.5
WSA	45	1	0.4	0.2	0.6	1	0.5

TABLE A1.5: CMIP6 Method 2 reduced 2–Variable Ensemble (F/B), where the parameters are tuned using Grid Search 2, parameter tuning results.

Station	Number of Rounds	Maximum Tree Depth	Eta (η)	Gamma (γ)	Column Sampling by Tree	Minimum Child Weight	Sub Sample
BTL	40	1	0.3	0.2	0.8	1	0.9
LA	40	2	0.3	0.2	0.8	1	0.9
MUA	50	1	0.3	0.2	0.8	1	0.9
NB	50	1	0.3	0.2	0.8	1	0.9
OMIA	50	1	0.3	0.2	0.8	1	0.9
SSMA	50	1	0.3	0.2	0.8	1	0.9
SLA	40	3	0.45	0.2	0.8	1	0.5
TVPA	40	1	0.45	0	0.8	1	0.5
TIA	45	2	0.5	0.05	0.5	1	0.5
TPIA	50	1	0.5	0.25	0.8	1	0.5
WTA	45	1	0.5	0.15	0.9	1	0.5
WSA	45	3	0.45	0.2	0.5	1	0.6

TABLE A1.6: Difference between predicted and observed monthly precipitation data, using the Method 2 reduced 2–Variable Ensemble where the top 2 variables are chosen from the forward/backward selection algorithm.

Month	Mean Difference	Maximum Difference	Minimum Difference
Jan	-2.34	17.309	-22.268
Feb	6.99	18.944	-2.937
Mar	5.62	17.226	-8.46
Apr	-4.49	15.583	-20.036
May	-3.88	9.753	-24.987
Jun	5.22	20.022	-12.538
July	-7.87	3.534	-24.866
Aug	6.94	19.052	-14.552
Sept	-7.52	11.702	-30.585
Oct	-10.1	9.989	-23.434
Nov	-3.37	11.054	-22.618
Dec	10.8	18.653	-1.272

TABLE A1.7: Correlation structure by decade for Station 12. The correlations between the CMIP6 predictor variables and the observed monthly precipitation (obs).

Time Period	1950-1959	1960-1969	1970-1979	1980-1989	1990-1999
Predictor Variable (x_i)	ρ_{obs,x_i}	ρ_{obs,x_i}	ρ_{obs,x_i}	ρ_{obs,x_i}	ρ_{obs,x_i}
BCC–CSM2–MR	-0.00756	-0.17638	-0.08258	-0.19534	-0.15233
CAMS–CSM1–0	-0.11463	-0.09533	0.06582	-0.09817	-0.11784
CESM2	-0.13253	-0.15824	0.01009	-0.10277	-0.13965
CESM2–WACCM	-0.13285	-0.25552	-0.05912	-0.00630	-0.27150
EC–Earth3	-0.05990	-0.26493	-0.11774	-0.17886	-0.12251
EC–Earth3–Veg	0.12907	-0.16098	-0.22988	-0.11767	-0.03396
FGOALS–f3–L	-0.00682	-0.13453	-0.09062	0.07844	0.01683
FIO–ESM–2–0	-0.12033	-0.10742	-0.02455	-0.02340	-0.04359
GFDL–ESM4	-0.08038	-0.17777	-0.05478	-0.14057	-0.17933
INM–CM4–8	-0.11001	-0.21844	-0.02357	-0.12670	-0.16074
INM–CM5–0	-0.11463	-0.19131	-0.07704	-0.17466	-0.15008
MPI–ESM1–2–HR	0.00662	0.09864	0.00559	-0.18475	-0.11067
MRI–ESM2–0	-0.14695	0.05430	-0.14705	-0.15366	-0.14822
NorESM2–MM	-0.14337	-0.07781	-0.10325	0.01183	-0.08914
Mean Max Temp	0.09801	0.29467	0.15318	0.29800	0.14099
Mean Min Temp	0.11956	0.33221	0.19537	0.33752	0.15671

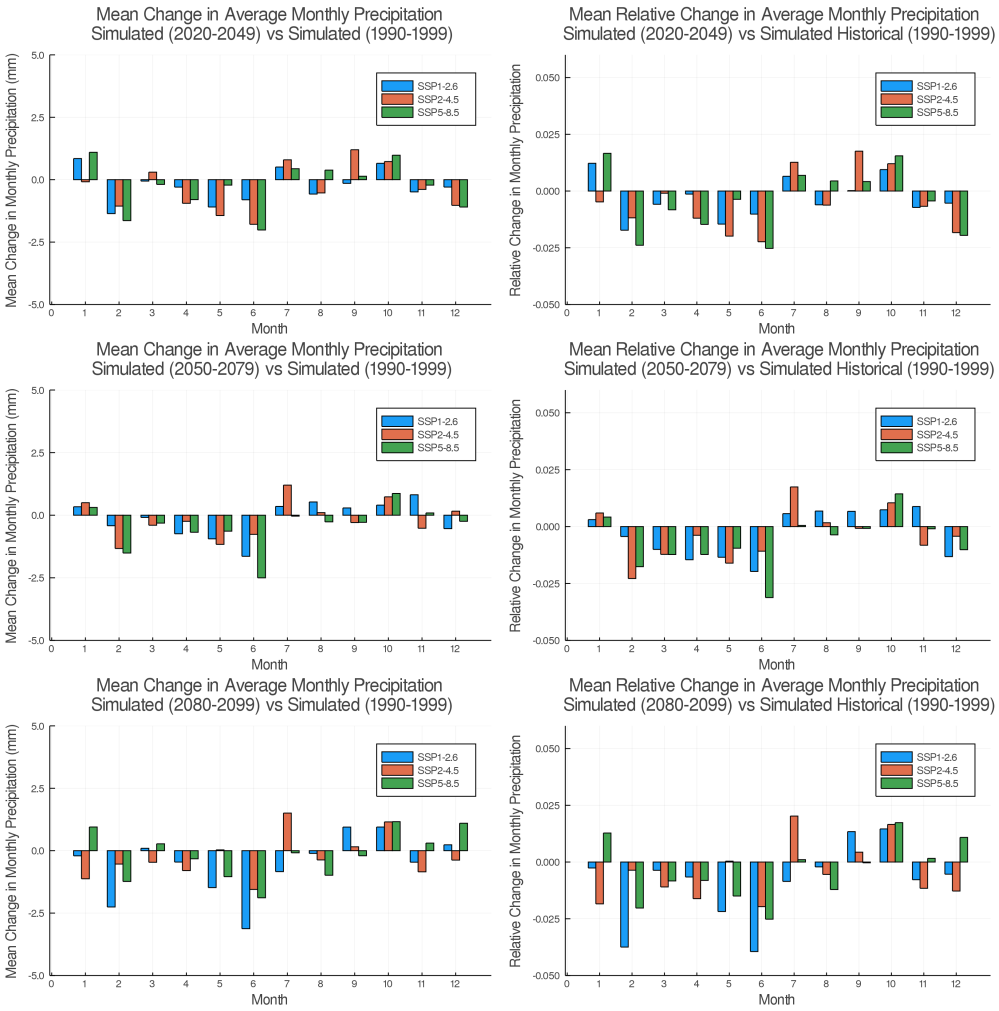

FIGURE A1.3: Change in precipitation calculated using historical and future simulations.

Bibliography

Agel, L., M. Barlow, J. Qian, F. Colby, E. Douglas, and T. Eichler (2015). Climatology of Daily Precipitation and Extreme Precipitation Events in the Northeast United States. *Journal of Hydrometeorology 16*(6), 2537–2557.

Bethke, I. (2016, April). Norwegian Earth System Model (NorESM) preparing for CMIP6. Presented at CCLICS Workshop. https://wiki.met.no/_media/noresm/BethkeEtAl_CCliCS2016_v2.pdf.

Bochinski, E., T. Senst, and T. Sikora (2017). Hyper-parameter optimization for convolutional neural network committees based on evolutionary algorithms. In *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 3924–3928.

Botzet, M. (2020, July 2). MPI-ESM. https://portal.enes.org/models/earthsystem-models/mpi-m-1/mpi-esm. Accessed July 12, 2020.

Busuioc, A., D. Chen, and C. Hellström (2001). Performance of statistical downscaling models in GCM validation and regional climate change estimates: Application for Swedish precipitation. *International Journal of Climatology 21*(5), 557–578.

Chen, T. and C. Guestrin (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge*

Discovery and Data Mining, New York, NY, USA, pp. 785–794. Association for Computing Machinery.

Chowdhury, M. and T. Turin (2020, Feb). Variable selection strategies and its importance in clinical prediction modelling. https://doi.org/10.1136/fmch-2019-000262.

Christensen, O., M. Drews, J. Christensen, K. Dethloff, I. Hebestadt, K. Ketelsen, and A. Rinke (2007, January). The HIRHAM Regional Climate Model version 5 (beta). *DMI Technical Report 06-17*.

Climate Workspace TCW, T. (2020). Max Planck Institute for Meteorology Earth System Model MR. http://www.glisaclimate.org/model-inventory/max-planck-institute-for-meteorology-earth-system-model-mr. Accessed May 6, 2020.

Cong, R.-G. and M. Brady (2012). The interdependence between rainfall and temperature: Copula analyses. *The Scientific World Journal 2012*.

Danabasoglu, G. (2019). NCAR CESM2-WACCM model output prepared for CMIP6 CMIP. Version 20191105. https://doi.org/10.22033/ESGF/CMIP6.10024. Accessed July 2, 2020.

Du, J., Y. Liu, Y. Yu, and W. Yan (2017). A prediction of precipitation data based on support vector machine and particle swarm optimization (PSO-SVM) algorithms. *Algorithms 10*(2), 57.

Dunne, J. P. (2019, October 29-31). GFDL's fourth generation CM4.0 and ESM4.1. https://www.gfdl.noaa.gov/wp-content/uploads/2019/12/7_dunne_Final_CM4_ESM4.pdf. Accessed July 14, 2020.

EC-Earth (2020). EC-Earth - A european community Earth-System model. http://www.ec-earth.org/. Accessed July 9, 2020.

ECCC (2018). The Canadian Regional Climate Model large ensemble. https://open.canada.ca/data/en/dataset/83aa1b18-6616-405e-9bce-af7ef8c2031c. Accessed July 12, 2020.

ECCC (2020). Historical climate data. https://climate.weather.gc.ca/index_e.html. Accessed May 10, 2020.

ENES (2019). CMIP5 data structure. https://portal.enes.org/data/enes-model-data/cmip5/datastructure/datastructure. Accessed July 8, 2020.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development 9*(5), 1937–58.

Fushiki, T. (2009, Oct). Estimation of prediction error by using k-fold cross validation. https://doi.org/10.1007/s11222-009-9153-8.

Giorgetta, M. A., J. Jungclaus, C. H. Reick, S. Legutke, J. Bader, M. Böttinger, V. Brovkin, and et al. (2013, June 28). Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project Phase 5. *Journal of Advances in Modeling Earth Systems 5*(3), 572–97.

Giorgi, F. (2019). Thirty years of regional climate modeling: Where are we and where are we going next? *Journal of Geophysical Research: Atmospheres 124*, 5696–5723.

Giorgi, F. and W. J. Gutowski (2015). Regional dynamical downscaling and the CORDEX initiative. *Annual Review of Environment and Resources 40*(1), 467–490.

Gizaw, M. S. and T. Y. Gan (2016, July). Regional flood frequency analysis using support vector regression under historical and future climate. *Journal of Hydrology 538*, 387–98.

GOC (2018). Climate model: second generation Canadian earth system model. https://www.canada.ca/en/environment-climate-change/services/climate-change/science-research-data/modeling-projections-analysis/centre-modelling-analysis/models/second-generation-earth-system-model.html. Accessed June 8, 2020.

Gutjahr, O., D. Putrasahan, K. Lohmann, J. H. Jungclaus, J.-S. V. Storch, N. Brüggemann, H. Haak, and A. Stössel (2019, July 25). Max Planck Institute Earth System Model (MPI-ESM1.2) for the High-Resolution Model Intercomparison Project (HighResMIP). *Geoscientific Model Development 12*(7), 3241–81.

Hausfather, Z. (2019). CMIP6: the next generation of climate models explained. https://www.carbonbrief.org/cmip6-the-next-generation-of-climate-models-explained. Accessed June 27, 2020.

He, B., Q. Bao, X. Wang, L. Zhou, X. Wu, Y. Liu, G. Wu, and et al. (2019, July 3). CAS FGOALS-f3-L Model datasets for CMIP6 Historical Atmospheric Model Intercomparison Project simulation. *Advances in Atmospheric Sciences 36*(8), 771–78.

He, T. (2016, March 10). An introduction to the XGBoost R package. http://dmlc.github.io/rstats/2016/03/10/xgboost.html. Accessed Jan 5, 2020.

IPCC (2020). Definition of terms used within the DDC pages. https://www.ipcc-data.org/guidelines/pages/glossary/glossary_r.html. Accessed July 6, 2020.

Kisi, O. and H. Sanikhani (2015, Feb 13). Prediction of long–term monthly precipitation using several soft computing methods without climatic data. *International Journal of Climatology 35*(14), 4139–50.

Kjellström, E., L. Bärring, G. Nikulin, C. Nilsson, G. Persson, and G. Strandberg (2016, July 28). Production and use of Regional Climate Model projections: A swedish perspective on building climate services. *Climate Services 2–3*, 15–29.

Kuhn, M. (2020). *caret: Classification and Regression Training*. R package version 6.0-86.

Kupiainen, Marco, C. J., P. Samuelsson, C. Jones, U. Willén, S. Wang, and R. Döscher (2015, Aug 11). Rossby Centre Regional Atmospheric model, RCA4. https://www.smhi.se/en/research/research-departments/climate-research-rossby-centre2-552/rossby-centre-regional-atmospheric-model-rca4-1.16562. Accessed June 9, 2020.

Li, X., Z. Li, W. Huang, and P. Zhou (2020, July 28). Performance of statistical and machine learning ensembles for daily temperature downscaling. *Theoretical and Applied Climatology 140*, 571–88.

McNicholas, P. D. and P. A. Tait (2019). *Data Science with Julia*. Boca Raton, FL: CRC Press, Taylor & Francis Group.

Mearns, L., S. McGinnis, D. Korytina, R. Arritt, S. Biner, M. Bukovsky, H.-I. Chang, O. Christensen, D. Herzmann, Y. Jiao, S. Kharin, M. Lazare, G. Nikulin, M. Qian, J. Scinocca, K. Winger, C. Castro, A. Frigon, and W. Gutowski (2020). The NA-CORDEX dataset, version 1.0. https://doi.org/10.5065/D6SJ1JCH. Accessed May 5, 2020.

Michaut, C. (2020). CMIP Phase 6 (CMIP6). Accessed February 10, 2020.

Nair, A., G. Singh, and U. Mohanty (2018, Jan). Prediction of Monthly Summer Monsoon Rainfall Using Global Climate Models Through Artificial Neural Network Technique. *Pure and Applied Geophysics 175*(1), 403–19.

Ni, L., D. Wang, J. Wu, Y. Wang, Y. Tao, J. Zhang, and J. Liu (2020, July). Streamflow forecasting using extreme gradient boosting model coupled with Gaussian mixture model. https://doi.org/10.1016/j.jhydrol.2020.124901.

Okkan, U. and U. Kirdemir (2016, July 22). Downscaling of Monthly Precipitation Using CMIP5 Climate Models Operated under RCP. *Meteorological Applications 23*(3), 514–28.

PCMDI (2019). CMIP6 Coupled Model Intercomparison Project Phase 6. https://pcmdi.llnl.gov/CMIP6/. Accessed June 27, 2020.

Pierce, D. (2019). *ncdf4: Interface to Unidata netCDF (Version 4 or Earlier) Format Data Files*. R package version 1.17.

Rong, X. (2019). CAMS: CAMS-CSM1-0 model output prepared for CMIP6 CMIP. Version 20190708. https://doi.org/10.22033/ESGF/CMIP6.1399.

Seland, Ø., M. Bentsen, L. Seland Graff, D. Olivié, T. Toniazzo, A. Gjermundsen, J. B. Debernard, A. K. Gupta, Y. He, A. Kirkevåg, J. Schwinger, J. Tjiputra, K. Schancke Aas, I. Bethke, Y. Fan, J. Griesfeller, A. Grini, C. Guo, M. Ilicak, I. H. Hafsahl Karset, O. Landgren, J. Liakka, K. Onsum Moseid, A. Nummelin, C. Spensberger, H. Tang, Z. Zhang, C. Heinze, T. Iverson, and M. Schulz (2020). The Norwegian Earth System Model, NorESM2 – Evaluation of the CMIP6 DECK and historical simulations. *Geoscientific Model Development Discussions 2020*, 1–68.

Soares Dos Santos, T., D. Mendes, and R. R. Torres (2016, Jan 27). Artificial neural networks and multiple linear regression model using principal components to estimate rainfall over South America. *Nonlinear Processes in Geophysics 23*(1), 13–20.

Song, Z.-Y., Y. Bao, and F.-L. Qiao (2019, Sept 3). Introduction of FIO-ESM v2.0 and its participation plan in CMIP6 experiments. *Advances in Climate Change Research 15*(5), 558–65.

Takhsha, M., O. Nikiéma, P. Lucas-Picher, and et al. (2018, July). Dynamical downscaling with the Fifth-generation Canadian regional climate model (CRCM5) over the CORDEX Arctic domain: effect of large-scale spectral nudging and of empirical correction of sea-surface temperature. *Climate Dynamics 51*, 161–86.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting 16*(4), 437–450.

Tianqi Chen, T. H., M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, and Y. Li (2020). *xgboost: Extreme Gradient Boosting*. R package version 1.2.0.1.

UCAR (2019). Community Earth System Model – CESM2. http://www.cesm.ucar.edu/models/cesm2/. Accessed August 4, 2020.

Volodin, E. and A. Gritsun (2018, Oct 25). Simulation of observed climate changes in 1850-2014 with climate model INM-CM5. *Earth System Dynamics 9*(4), 1235–42.

Volodin, E. M., E. V. Mortikov, S. V. Kostrykin, V. Y. Galin, V. N. Lykossov, A. S. Gritsun, N. A. Diansky, and et al. (2018, Dec 17). Simulation of the modern climate using the INM-CM48 Climate Model. *Russian Journal of Numerical Analysis and Mathematical Modelling 33*(6), 367–74.

Wang, X., G. Huang, Q. Lin, and J. Liu (2014, July 10). High-resolution probabilistic projections of temperature changes over Ontario, Canada. https://doi.org/10.1175/jcli-d-13-00717.1.

Wang, X., G. Huang, Q. Lin, X. Nie, and J. Liu (2014, July 9). High-resolution temperature and precipitation projections over Ontario, Canada: A coupled dynamical-statistical approach. *Quarterly Journal of the Royal Meteorological Society 141*(689), 1137–46.

WCRP (2017). WCRP Coupled Model Intercomparison Project (CMIP). https://www.wcrp-climate.org/wgcm-cmip. Accessed February 12, 2020.

Wickham, H., M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen,

E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani (2019). Welcome to the tidyverse. *Journal of Open Source Software 4*(43), 1686.

Wu, T., Y. Lu, Y. Fang, X. Xin, L. Li, W. Li, W. Jie, and et al. (2019, April 24). The Beijing Climate Center Climate System Model (BCC-CSM): the main progress from CMIP5 to CMIP6. https://doi.org/10.5194/gmd-12-1573-2019.

XGBoostDevelopers (2020). Xgboost parameters - xgboost 1.3.0 documentation. https://xgboost.readthedocs.io/en/latest/parameter.html. Accessed February 2, 2020.

Xin, X., J. Zhang, F. Zhang, T. Wu, X. Shi, J. Li, M. Chu, Q. Liu, J. Yan, Q. Ma, and M. Wei (2018). BCC BCC-CSM2MR model output prepared for CMIP6 CMIP. Version 20181128. https://doi.org/10.22033/ESGF/CMIP6.1725.

Xu, R., N. Chen, Y. Chen, and Z. Chen (2020, March 9). Downscaling and projection of multi-CMIP5 precipitation using machine learning methods in the Upper Han River Basin. *Advances in Meteorology 2020*, 1–17.

Yukimoto, S., H. Kawai, T. Koshiro, N. Oshima, K. Yoshida, S. Urakawa, H. Tsujino, M. Deushi, T. Tanaka, M. Hosaka, S. Yabu, H. Yoshimura, E. Shindo, R. Mizuta, A. Obata, Y. Adachi, and M. Ishii (2019). The Meteorological Research Institute Earth System Model Version 2.0, MRI-ESM2.0: Description and basic evaluation of the physical component. *Journal of the Meteorological Society of Japan. Ser. II 97*(5), 931–965.

Zheng, H., J. Yuan, and L. Chen (2017). Short-term load forecasting using EMD-LSTM neural networks with a Xgboost algorithm for feature importance evaluation. *Energies 10*(8).