UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

A MACHINE LEARNING BASED MULTI-MODEL ENSEMBLE APPROACH TO RECONSTRUCT THE HISTORICAL MONTHLY PRECIPITATION OVER OKLAHOMA USING NOAA'S SPEAR DATASET

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

DILIP NEUPANE
Norman, Oklahoma
2024

A MACHINE LEARNING BASED MULTI-MODEL ENSEMBLE APPROACH TO RECONSTRUCT THE HISTORICAL MONTHLY PRECIPITATION OVER OKLAHOMA USING NOAA'S SPEAR DATASET

A THESIS APPROVED FOR THE
SCHOOL OF CIVIL ENGINEERING AND ENVIRONMENTAL SCIENCE

BY THE COMMITTEE CONSISTING OF

Dr. Tiantian Yang, Chair

Dr. Yang Hong

Dr. Pierre E. Kirstetter

Dedication

To my parents who have always supported my decisions and encouraged me throughout my life, I want to dedicate this thesis. I am forever indebted to them for their immense support, numerous sacrifices, and unconditional love. Without them, I would not be the person I am today.

In addition, I dedicate this work to my sibling, cousins, and friends back home, who have been a constant source of support and motivation since my childhood.

Acknowledgments

I am sincerely grateful to Dr. Tiantian Yang for providing me with the opportunity to conduct research under his guidance and supervision during my Master's journey. His constant support and mentorship have been invaluable, contributing significantly to my professional and academic growth. I extend my thanks to my committee members, Dr. Yang Hong, and Dr. Pierre Kirstetter, for their feedback and constructive comments, which helped in refining my research.

Additionally, I am extremely thankful to Dr. Lujun Zhang, who helped me from debugging codes to preparing for conferences and providing comments on my writing. His guidance has been instrumental in shaping me into a better researcher. I express my gratitude to all the members of Dr. Yang's research group, especially Aayush Adhikari, for their support. Lastly, I want to thank all my friends in Norman whose companionship has made my time truly memorable.

Finally, I want to express my gratitude to all the funding sources that supported the research for this thesis. The research works for this thesis is supported by the OU Faculty Investment Program by the Vice President of Research and Partnership (VPRP), the National Science Foundation (NSF) CAREER Award (No. 2236926), and the EPSCoR Track-1 Project under Grant No. OIA-1946093 and its subaward No. EPSCoR-2020-3.

Table of Contents

List of Figures

Abstract

General Circulation Models (GCMs) are important tools in simulating and projecting future precipitation at the decadal scale. However, it is inevitable that simulation and projection errors and uncertainty exist in GCMs, hindering their applications for regional water resources planning. Different post-processing tools are available to address the uncertainty issues associated with GCMs and to utilize these tools better for regional water resources planning. For example, a multi-model ensemble (MME) could reduce uncertainties from different GCMs and help reduce the model biases from a single model. In this study, we employed multiple Machine Learning algorithms (MLs) to combine ensemble members from NOAA’s Seamless System for Prediction and EArth System Research (SPEAR) to reconstruct historical monthly precipitation over Oklahoma during a study period (1981-2014). The employed MLs include Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Support Vector Machines (SVM), and Classification And Regression Trees (CART). The performances of the employed MLs are benchmarked with Simple Model Averaging (SMA), Bayesian Model Averaging (BMA), and Reliability Ensemble Averaging (REA). Our result echoes previous studies where the raw precipitation simulation from SPEAR presents significant simulation bias and marginal simulation skills. Different spatial and seasonal patterns of the simulation bias and skill are also observed over our study region. All the employed multi-model averaging techniques have delivered better performances than any single ensemble member from SPEAR. The employed MLs have outperformed SMA, BMA, and REA, which is evident from the reduction of bias and skill improvement. In general, this study highlights future applications of other data-driven techniques in post-processing the multi-model simulation from GCMs.

1 Introduction

Global climate change alters the climate of precipitation worldwide (Brekke 2009, Trenberth 2011), which could seriously affect water resource operations such as hydropower generation, irrigation planning, proactive flood control, etc. (Handmer et al. 2012, Pielke Sr et al. 2009).

General Circulation Models (GCMs) are important tools to quantify and predict future changes in the climate of precipitation. Available GCMs are coupled with dynamic land surface, oceanic, and atmospheric components, allowing for a comprehensive consideration of the Earth System. To further advance and demonstrate the effectiveness of available GCMs, the World Climate Research Program (WCRP) initiated the Coupled Model Inter-comparison Projects (CMIPs) for an international and multi-agency effort to include and intercompare different GCMs worldwide (Dufresne et al. 2013). The current state-of-the-art CMIP results are known as CMIP6. Compared to previous CMIP results, CMIP6 has advancements made in its simulation resolution and additional inclusion of various earth system processes (O'Neill et al. 2016, Tokarska et al. 2020).

Despite the advancements made in GCMs, the GCMs’ simulation outputs are still subjected to uncertainty arising from various sources, including model structures, parameterization schemes, and boundary conditions (Hawkins and Sutton 2011, Woldemeskel et al. 2012, 2014). Further, it is reported by (Knutti and Sedláček 2013) that errors associated with the GCMs could also arise from the limited computational resources, spatial resolution, and internal variability. As a result, simulation bias as well as low simulation skills of GCM-simulated or -projected precipitation are commonly reported in different study regions over the globe (Aloysius et al. 2016, Kumar et al. 2013, Mehran et al. 2014, Palerme et al. 2017).

To address such simulation uncertainty, various multi-model ensemble techniques are commonly applied to improve the accuracy of GCM simulation/projections. By perturbating the initial states, employing different parameterization schemes, and/or the inclusion of multiple GCM models, an ensemble that contains multiple GCM simulations at the same time could be generated. By adopting such techniques, the uncertainty of GCM-simulated precipitation could be quantified mathematically. But more importantly, it is reported that by combining different ensemble members with the same weight (Simple Model Averaging, i.e., SMA), the simulation performance could be improved in contrast to that of individual simulation members (Ma et al. 2018, Yumnam et al. 2022, Zhang et al. 2021).

Based on the need to combine simulation results of a large ensemble, the Bayesian Model Averaging (BMA) and Reliability Ensemble Averaging (REA) have become more popular. Unlike SMA, which assigns the same weight to different ensemble members, BMA and REA combine estimations from individual simulation members while considering different ensemble members' simulation performance during a training period (Raftery et al. 2005, Raftery et al. 1997). In comparison to SMA, BMA could provide a probability distribution that reflects the prediction uncertainty quantitatively. BMA incorporates the uncertainty of ensemble members based on their simulation performance of the reference dataset. As for REA, it assigns weights to the ensemble members based on their reliability values rather than assigning equal weights to all the ensemble members. The reliability values in REA are calculated based on the calculated bias. Both BMA and REA are extensively studied in the field of hydroclimatology and reported to be effective in improving GCM simulation skills (Jiang et al. 2012, Liu and Merwade 2018, Massoud et al. 2020, Yan et al. 2020).

However, limitations of the BMA and SMA reside when applying them to post-process GCM-simulated precipitation. The SMA does not consider the performance differences between ensemble members when assigning weights to different ensemble members for model averaging. Instead, SMA simply computes the arithmetic mean of all simulations. The BMA assumes that the sum of the likelihood of all candidate ensemble members being the “perfect” model should equal 1, which is not always the case (Ley and Steel 2009). Furthermore, BMA is restricted to static applications, which means that the model might not be suitable for capturing the evolving and changing dynamics of climate (Nonejad 2021). As for REA, it only considers bias and does not consider temporal variabilities (Tegegne et al. 2019). Moreover, both BMA and REA assume that different ensemble members are independent to each other and should follow different distributions, which could be invalid since GCM-simulated ensemble precipitation is normally generated with a shared methodology and data sources (Li et al. 2022). As a result, the limited performance of baseline model averaging techniques from SMA, BMA and REA indicate the need for novel model averaging techniques.

A promising alternative to SMA, BMA and REA could be various Machine Learning (ML) algorithms. MLs can effectively identify the complex relationships between input and target variables, which may not be directly related to each other. MLs can address non-linearity in time-series data, such as precipitation. A good number of previous studies have reported the effectiveness of various MLs in the field of hydroclimatology for post-processing GCM simulations (Balhane et al. 2022, Sachindra et al. 2018, Wang et al. 2017). Ahmed et al. (2020) and Crawford et al. (2019) have found that the application of different ML-based model averaging techniques like Random Forest (RF) and Relevance Vector Machine (RVM) show better simulation performance than MME simulated precipitation. Similarly, other ML algorithms like eXtreme Gradient Boosting (XGB) and Extra Tree Regressor (ETR) have also produced better simulation skills in averaging the MME simulated precipitation (Jose et al. 2022, Shetty et al. 2023).

However, it is also reported that different MLs tend to provide different levels of performance at different geospatial locations over the world. For example, Wang et al. (2018) and Crawford et al. (2019) reported that the RF produces the optimal results for averaging multi-model simulated precipitation in Australia and in parts of North America. Likewise, Jose et al. (2022) found that RF and Long Short-Term Memory (LSTM) to be the most effective for combining multi-model ensembles in India. In addition, Ahmed et al. (2020) concluded that K-Nearest Neighbors (KNN) and RVM produced the optimal results for similar tasks in Pakistan. This implies that different MLs present different performances in simulating the precipitation.

Therefore, the goal of this thesis is to test the performance of different MLs in reconstructing historical monthly precipitation by averaging multiple GCM simulations over Oklahoma during a study period from 1981 to 2014. A total of 4 MLs are employed in this thesis, including the Random Forest (RF), Support Vector Machine (SVM), eXtreme Gradient Boosting (XGBoost), and Classification And Regression Trees (CART). The multi-model ensemble precipitation simulation from NOAA’s Seamless System for Prediction and EArth System Research (SPEAR) is used as the study dataset. The precipitation dataset generated with the Parameter-elevation Relationships on Independent Slopes Model (PRISM) is used as a reference. The performance of the employed MLs is benchmarked with 3 baseline approaches of SMA, BMA and REA. Three different evaluation statistics, including percentage bias, coefficient of determination, and root mean square error, are utilized to conduct the spatial and temporal analysis of the simulated precipitation.

2 Literature Review

2.1 General Circulation Models (GCMs)

General Circulation Models (GCMs) are critical tools in the field of climate science that offer information regarding the physics of the Earth system. Such models simulate interactions between different geophysical processes within the Earth system and are important in understanding the past, present and future climate scenarios (Flato et al. 2014). GCMs are composed of dynamical components that numerically simulate the behavior of the atmosphere, land surface, oceans, sea ice, and many more. Different components of GCMs are dynamically coupled with each other, allowing for a comprehensive representation of the Earth system (Flato et al. 2014, Hunke et al. 2010, Pitman 2003).

The Coupled Model Inter-Comparison Project (CMIP) was initiated to better utilize and advance GCMs in climate science. The CMIP aims to facilitate the comparisons between available GCMs, to propel the advancements of GCMs and climate science (Eyring et al. 2016, Taylor et al. 2012). Over the years, CMIP has grown steadily from its first phase project of CMIP1 to the latest CMIP6. Every phase of CMIP project has served as a major tool for the assessment of climate change all over the globe. CMIP6, being the most recent phase of CMIP, has many notable advancements over its predecessors. CMIP6 has increased the temporal and spatial resolution along with the implementation of a new scenario framework, integrating both shared socio-economic pathways (SSP) and radiative forcing levels (O'neill et al. 2016). Further, it has included the components representing complex processes of the earth system related to biogeochemical cycle, land use, carbon cycle and others to incorporate their corresponding control over climate change (Eyring et al. 2016). As a result, according to Zelinka et al. (2020), CMIP6 models have higher climate sensitivity than the previous models, which will lead to more accurate and skillful simulations/projections of the climate of the Earth system overall.

The Seamless System for Prediction and EArth System Research (SPEAR) from NOAA is one of the contributing models to CMIP6. The SPEAR is developed with the intention of providing seamless climate simulation/projection across different timescales and is used in studying seasonal to decadal climate variability along with the real-time predictions of precipitation (Delworth et al. 2020). There are many advantages of the SPEAR model over previous generation CMIP models, such as the incorporation of advanced parameterization of physical processes (Delworth et al. 2020), the enhancement of the predictive skill for climate forecasts and climatic teleconnections (Xiang et al. 2021), and the reduction of the model's ocean bias (Bushuk et al. 2021). Multiple studies have reported that SPEAR is effective in simulating and projecting future precipitation and other climatic variables (Murakami et al. 2020, Zhang and Cooke 2021). For instance, Pascale et al. (2020) found that the use of SPEAR large ensemble dataset improves the regional-scale simulation of droughts in southern South Africa. More details regarding SPEAR can be found on Delworth et al. (2020).

Despite many advancements made in available GCMs, the GCMs simulated precipitation are still subjected to various errors. The errors in the GCMs occur from various sources: (1) simulation errors due to imperfect model structure, (2) errors due to the resolution of GCMs, and (3) other sources. For source (1), the volume and occurrence of precipitation are jointly determined by intricate feedback mechanisms between cloud, radiation, topography, and soil moisture. But such intricate mechanisms are hard to represent accurately by GCMs, which eventually leads to erroneous simulation and projection of precipitation (Berg et al. 2017, Zhao et al. 2016). As for source (2), the resolution of GCMs prevents the capturing of small-scale physical and chemical processes related to condensation and evaporation (Demory et al. 2014, Ingram and Bushell 2021). As for source (3), factors like topography also produce errors in GCMs. For instance, studies have found that GCMs might not produce accurate projections in areas with complex terrain and topography (Dai 2006, Posada-Marín et al. 2019). Similar limitations persist in SPEAR and CMIP6 projects as well, where these errors in model structures and resolutions have led to projection uncertainties and poor simulation, thereby compromising the performance of the models (Abdelmoaty et al. 2021, Johnson et al. 2022, Li et al. 2021a).

To summarize, GCMs are widely used tools to quantitatively study the Earth System climate. Many advancements have been made in the latest GCMs projects and models, such as CMIP6 and SPEAR. However, accurately simulating historical precipitation remains an extremely challenging task.

2.2 Multi-model ensembles (MME) and model averaging techniques

Efforts have been made to improve the quality of GCM simulations from a post-processing perspective (Duan et al. 2021, Li et al. 2021a, Schepen et al. 2018). Among which, various multi-model ensembles (MME) are widely studied and reported to be effective in quantifying simulation uncertainty and enhancing simulation accuracy (Ahmed et al. 2020, Jiang et al. 2012, Schepen et al. 2018).

The key concept of MME is to create multiple simulation/projection trials at the same time utilizing available GCMs. By doing so, the simulation/projection uncertainty can be quantified numerically. In general, different MME techniques can be categorized into 3 groups: (1) perturbation of the physics of GCMs, (2) considering multiple GCMs, and (3) a combination of (1) and (2) (Murphy et al. 2004). When choosing to adopt (1) to realize MME, multiple simulation/projection outputs are generated using different initial conditions, physics components, parameters or forcings with available GCMs. Whereas if choosing (2) to realize MME, a simulation/projection ensemble is formed with multiple GCM's outputs (Parker 2013, Rowlands et al. 2012). The availability of a large number of ensembles makes the quantification of climatic uncertainties, along with the reduction and detection of errors, easier and efficient (Becker et al. 2022, Jebeile and Crucifix 2020).

With the employment of MME techniques, it is reported that further combining different members of MME leads to overall superior simulations/projections. Figure 2.1 shows the schematic illustration of the model averaging technique using a three-member ensemble and one set of measurements (Vrugt 2016). The right panel of Figure 2.1 shows the probability density functions (PDFs) of the combined forecast (solid black line) and of individual models (solid-colored lines). By combining multiple forecasted values as well as the pdfs of MME, it is expected that the combined forecast shall have a better agreement with the measurement compared to individual predictions. Among all developed multi-model averaging techniques, SMA is the most widely used. SMA considers all the members to be equally informative and present the same level of performance. Miao et al. (2014) reported that SMA delivers superior performance than individual CMIP5 models in simulating the precipitation and temperature over Northern Eurasia. Moreover, several other studies have also found the use of SMA in producing better simulation results than the individual models in different parts of the world (Mitra et al. 2011, Yang et al. 2012, Zhang and Yan 2018).

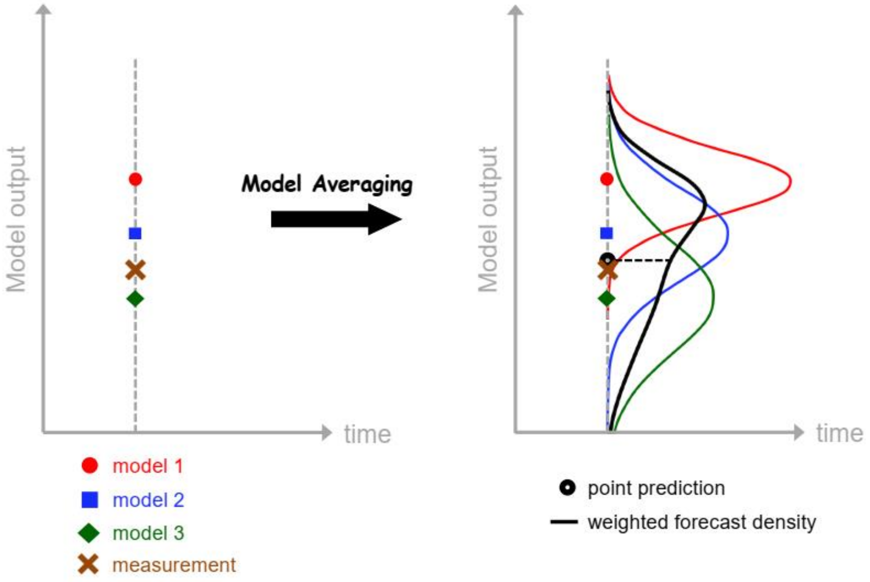

Figure 2.1. Schematic illustration of model averaging (Vrugt, 2016)

However, SMA has its limitations when it comes to model averaging. SMA assigns equal weights to all the ensemble members regardless of their performance. But different ensemble members within MME normally present different levels of simulation skills (Miao et al. 2014, Zhang and Yan 2018). As a result, the improvement associated with SMA is normally limited. Moreover, Lambert and Boer (2001) found that the SMA method produces the best results in comparison to the observations when the models are developed independently to each other. But GCMs extensively share a similar concept of developing parameters and model components along with the duplication of code and sharing of forcing (Sanderson et al. 2015, Wang et al. 2018). This makes different GCM ensemble members normally dependent on each other to some degree, which consequently compromises the performance of the application of SMA in GCMs. Therefore, the use of SMA might not provide appropriate and accurate results in all the scenarios.

Recognizing the limitations of SMA, more advanced model averaging approaches, such as BMA and REA became popular. While SMA assigns equal weights to all the ensemble members, REA and BMA assign variable weights to ensemble members based on their skill in simulating observations during a historical period (i.e., training period). BMA incorporates probabilistic techniques to assign different weights to the models by creating probability density functions (pdf) of weather variables (Sloughter et al. 2007). The pdf value is dependent upon prior distributions, which refers to the initial belief of the BMA model before observing the data. Likewise, REA assigns weight to the individual ensemble members based on their "reliability". The reliability is determined according to the ability of ensemble members to simulate the observations and their degree of convergence compared to other ensemble members (Giorgi and Mearns 2002, Tanveer et al. 2016).

It is reported that REA and BMA produce better simulation skills compared to SMA as the weighted approach accounts for the performance variation of individual models (Ahmed et al. 2020, Leduc et al. 2016, Wang et al. 2018, Yang et al. 2012). Studies performed by Miao et al. (2014) and Tanveer et al. (2016) have demonstrated the applicability and reliability of REA in projecting future climate scenarios. Both BMA and REA have been widely reported to be effective in post-processing predictions and simulations of precipitation in various locations over the world (Ji et al. 2019, Sloughter et al. 2007, Wang et al. 2012). It is also reported that BMA and REA provide results with similar reliability even though they follow different assumptions and technical steps (Duan et al. 2021, Mani and Tsai 2017). In general, past studies reported that the application of weighted ensemble approaches, like BMA and REA, can provide superior performance over the traditional SMA and individual raw ensembles.

However, BMA and REA are associated with many limitations as well. For example, in BMA, the calculation of weights is dependent upon prior distributions representing prior knowledge or assumption regarding the observed data. So, the assignment of prior distributions and other parameters of the models can influence the outcome since the allocation of priors can often be vague (Fragoso et al. 2018, Hinne et al. 2020). On the other hand, BMA assumes that the sum of the likelihood of all candidate ensemble members should equal 1, which is not always the case (Ley and Steel 2009). Furthermore, BMA is restricted to static applications, which means that the model might not be suitable for capturing the evolving and changing dynamics of climate (Nonejad 2021). Similarly, REA model only considers bias in its weighting approach and does not consider the temporal variabilities, which is key in this changing climate (Tegegne et al. 2019).

In summary, Multi-Model Ensembles (MME) that contain multiple members are normally constructed to quantify better and address the simulation/projection uncertainties of GCMs. Moreover, it is reported that combining different members of MME through techniques such as SMA, BMA, and REA further improves the simulation/projection of GCM. However, the weighted approaches, like SMA, BMA and REA, have their own limitations. Therefore, novel MME techniques are critically needed to combine a large number of ensembles.

2.3 Machine Learning (ML) approaches

Given the limitations of conventional model averaging techniques, nowadays very popular ML approaches become promising alternatives for MME model averaging to further improve the quality of GCM simulation/projections (Reichstein et al. 2019).

ML has been a very useful and transformative tool in various fields, and more so in climate science. Specifically, ML has been utilized extensively to enhance the performance of MMEs. ML can identify complex patterns and learn about the features from a large number of datasets, while it can also determine the optimal weights of those datasets for the purpose of model averaging (McGovern et al. 2017, Sloughter et al. 2007). Moreover, the adoption of ML techniques could help quantify the uncertainties of GCM-generated predictions (Ahmed et al. 2020, Song et al.

2020). With these advanced features, ML has the ability to provide better and more accurate results in the projection and reconstruction of various climatic variables, such as precipitation (Li et al. 2021b, Xu et al. 2020).

MLs are considered to have good potential in averaging different GCM-generated precipitation simulations/projections, as ML algorithms can better learn and detect the patterns and trends of precipitation globally. Typically, ML-based model averaging techniques have yielded lower bias and higher skill in the simulation of precipitation across the world (Dey et al. 2022, Xu et al. 2020). Ahmed et al. (2020) used multiple ML-based techniques in simulating the precipitation over Pakistan and found that the application of MLs improved the simulation in all seasons. Further, Li et al. (2021b) and Dey et al. (2022) found that the application of ML preserves regional and temporal patterns of precipitation over different study regions. A number of other studies have also recommended applying ML-based methods to combine MMEs for future climatic projections (Jose et al. 2022, Wang et al. 2023, Xu et al. 2020).

However, different levels of performance have been reported for different ML-based model averaging techniques in terms of simulating precipitation in different geographic locations around the world (Crawford et al. 2019, Jose et al. 2022, Wang et al. 2018). The performance of different ML algorithms can vary due to the particular strengths and weaknesses associated with those algorithms (Osisanwo et al. 2017). Shetty et al. (2023) conducted a study in the Western Ghats of India for combining MME simulated precipitation. This study by Shetty et al. (2023) reported that XGBoost and RF produce better simulation performance, while SVM produces poor performance in the simulation of precipitation. Further, RF produced optimal results in Australia and parts of North America for similar tasks (Crawford et al. 2019, Wang et al. 2018). In addition, KNN and

RVM are reported to present better results in combining MME simulated precipitation in Pakistan (Ahmed et al. 2020).

To summarize, ML algorithms are novel data-driven approaches that have been widely used in the field of hydroclimatology due to their ability to capture and preserve various patterns and trends. But different ML algorithms demonstrate varying simulation accuracies depending upon the geographical location, which emphasizes the need to implement and evaluate multiple ML algorithms' performance in combining the MME simulations. Therefore, applying and evaluating the robustness of different ML algorithms in reconstructing the historical monthly precipitation over Oklahoma is important and remains unexplored.

3 Goals and Motivations

It is evident from the literature review that GCMs are associated with various errors when simulating historical precipitation and projecting future precipitation scenarios. One way to improve the simulation skill of precipitation from GCMs is to adopt model averaging techniques and combine the simulation results from different GCMs. Model averaging can be conducted by using baseline approaches, such as SMA, BMA and REA, and by using novel ML algorithms. Compared to baseline approaches, ML techniques present a superior ability in capturing the nonstationary model error structures. However, when using ML algorithms to enemble GCM outputs, different MLs may have varying performances depending upon the geospatial location of the study areas as well as the specific algorithm being used. Therefore, it is imperative to further test the robustness and effectiveness of different MLs in combining MME preciptiation of GCMs, especially over a region with complex precipitation dynamics.

Given the motivation above, I choose to adopt Oklahoma as the study region of this thesis, considering Oklahoma's significant variability in temporal and spatial patterns of precipitation. The precipitation in Oklahoma is influenced by many physical processes, such as the moisture coming from the Gulf of Mexico, or the deserts from the Southwest US, the Mesoscale Convective Systems (MCS), and others. As a result, the precipitation events in Oklahoma can be attributed to various climate and weather systems, thus making it a suitable study region for testing the robustness of different MLs (Ford et al. 2015b). More details regarding the study region can be found in section 4.2. Specifically, the following research hypotheses are tested in this thesis.

1. The raw monthly GCM-simulated precipitation presents errors and bias over Oklahoma during a historical period.

2. The performance of GCM-simulated monthly precipitation can be improved through baseline MME techniques of SMA, BMA, and REA.
3. Various novel ML algorithms are also effective in combining the GCM-simulated precipitation.
4. The employed ML algorithms present superior simulation skills for precipitation compared to the baseline approaches of SMA, BMA, and REA.

4 Data and Study region

4.1 Data

Two precipitation datasets are used in this thesis to test the performance of different model averaging techniques. The first dataset used in this thesis is the GCM-generated ensemble precipitation simulation from the Seamless System for Prediction and Earth System Research (SPEAR). In this thesis, I use the SPEAR precipitation for model averaging experiments with various algorithms. The second dataset used in this thesis is the reference precipitation from the Parameter-elevation Regressions on Independent Slopes Model (PRISM), which is used to validate the outcome of the models developed using SPEAR.

The SPEAR dataset is developed by the National Oceanic and Atmospheric Administration (NOAA)'s Geophysical Fluid Dynamics Laboratory (GFDL). It contains 30 ensemble members generated through perturbed initial conditions containing simulations of different climatic variables from the period from 1921 to 2100 (Delworth et al. 2020). In the SPEAR dataset, all historical simulations are forced with historical radiative forcing from 1921 to 2014. The historical precipitation from SPEAR is provided with a spatial resolution of 0.625° * 0.5°.

Based on the SPEAR dataset, the verification and validation of all the proposed model averaging techniques are done against the PRISM dataset as a reference. In this thesis, the monthly gridded precipitation observation developed by PRISM Climate group at Oregon State University is used (https://prism.oregonstate.edu/). The PRISM data provides monthly precipitation observations at a spatial resolution of 4 km (~0.04°). The PRISM takes into account the orographic enhancement of precipitation, combining both rain-gauge records and RADAR measurements (Daly and Bryant 2013). PRISM has proven to be a reliable dataset and has been applied by many hydrometeorological related studies (Buban et al. 2020, Prat and Nelson 2015, Zhang et al. 2021).

4.2 Study Area

Oklahoma is located in the Southern Great Plains of the United States constituting plains and gently rolling hills with its elevation decreasing from west to east (Allen and Gichuki 1989). There is a significant variability in the precipitation pattern across Oklahoma, which is evident from its distinct precipitation gradient. The eastern half of Oklahoma receives a considerable amount of rainfall whereas the western part receives comparatively less rainfall (Ford et al. 2015a). Such positive precipitation gradient across the state from the west to the east can be attributed to the moisture brought by the southerly winds from the Gulf of Mexico (Tian and Quiring 2019). Due to the location of Oklahoma being in a midlatitude continental region, the precipitation patterns over Oklahoma are also influenced by multiple climate patterns over a wide range of time scales ranging from diurnal to annual (Fisher 2004). During the warm summer months, the precipitation in Oklahoma is influenced by high levels of moisture activity and Mesoscale Convective Systems (MCS) (Easterling et al. 2017, Hand and Shepherd 2009). This brings more precipitation in the warmer summer months.

5 Methods

A total of seven different model-averaging techniques are used in this thesis. Among these, three of the employed methods are baseline/traditional approaches to benchmark the outcomes generated from the remaining four ML algorithms. The employed baseline approaches include Simple Model Averaging (SMA), Bayesian Model Averaging (BMA), and Reliability Ensemble Averaging (REA). Out of four ML algorithms, three are decision tree-based algorithms of Random Forest (RF), eXtreme Gradient Boosting (XGB), and Classification And Regression Trees (CART). Another employed ML algorithm is a popular non-tree-based algorithm, termed the Support Vector Machine (SVM).

5.1 Data preparation

Considering the common availability of both the SPEAR and PRISM precipitation dataset, the study period is set from 1981 to 2014 for a total of 34 years. Bilinear interpolation is used for both datasets (SPEAR and PRISM) to match their spatial resolution such that the application of subsequent model averaging techniques and the evaluation of results can be consistent. The common spatial resolution is set to be 0.25° * 0.25°. Before the evaluation, the datasets are masked out to contain only the precipitation data over the study area, i.e., the state of Oklahoma. The clipping of data over Oklahoma allows us to analyze the precipitation precisely over Oklahoma, making the evaluation of results more accurate and representative of the study region.

5.2 Training of Model averaging techniques

The training of all employed model averaging approaches follows the same three-fold cross-validation strategy for consistency. To be specific, out of the 34 years of data records (1981-2014), 22 years are used as a training period, and the remaining 12 years of data are used to validate the models. A three-fold cross-validation approach is employed such that the training and validation approach is more robust. For the three-fold cross-validation, the study period (1981-2014) is further divided into three folds or subsets such that the split folds do not overlap with each other. Then in a sequential manner, each fold serves as a validation data set, and the remaining two folds act as the training data set.

All the employed model averaging techniques are trained for different months separately to account for potential seasonal patterns of precipitation. To be more specific, both SPEAR and PRISM precipitation are divided into 12 months for separate training of the employed model averaging techniques. More details about the employed model averaging techniques are described in detail in the subsequent sections.

5.3 Baseline approaches

The baseline model averaging techniques employed in this thesis are SMA, BMA, and REA. The baseline model averaging techniques assign multiple ensemble members with certain weights. More detailed information about SMA, BMA, and REA are described in Sections 5.3.1, 5.3.2, and 5.3.3, respectively.

5.3.1 Simple Model Averaging (SMA)

SMA is the simplest and the most common approach to combine results from different models. The SMA considers all ensemble members equally informative and shall present the same level of performance. SMA combines simulations from different ensemble members by assigning them with equal weights. The computation of SMA can be represented with the following equation (1):

$$P = \frac{1}{n}\sum_{i=1}^{n} M_i \tag{1}$$

Where P is the SMA of an ensemble of simulations, n indicates the total number of ensemble members, and M_i indicates the simulated value of *i*th ensemble member.

5.3.2 Bayesian Model Averaging (BMA)

BMA combines different ensemble members by assigning them with different weights. In BMA, the weights are computed based on the performance of different ensemble members during a training period. With derived weights, BMA generates a probability density function (PDF) that is centered on the averaged simulated values. The weights thus generated reflect the relative contribution of all the ensembles on the obtained multi-model ensemble. Consequently, these weights can be used to determine the usefulness of a particular ensemble member in the combined model. The combined PDF of the simulated values can be represented with the following equation (2):

$$p(y|y^T) \ = \sum_{i=1}^{n} p(y|M_i, y^T)\, p(M_i|y^T) \qquad (2)$$

Where n indicates the total number of ensemble members, $p(y|M_i, y^T)$ is the simulated PDF of a certain ensemble member M_i estimated by using the observations y^T during the training period for BMA, and $p(M_i|y^T)$ is the posterior probability of the model M_i that is corrected using the training data which is computed based on the Bayesian theory with equation (3):

$$p(M_i|y^T) = \frac{p(y^T|M_i\,)p(M_i\,)}{\sum_{j=1}^{n} p(y^T|M_j\,)\, p(M_j|y^T)} \qquad (3)$$

In this thesis, the BMA weights are computed by maximizing the likelihood algorithm according to Raftery et al. (2005). Here, the posterior probabilities of the model are determined by maximizing the equation (3) which are then eventually assigned as weights of the models. More details of the BMA process are provided in Duan and Phillips (2010).

5.3.3 Reliability Ensemble Averaging (REA)

REA is a weighted averaging approach that assigns weight to the individual ensemble members based on their “reliability” (Giorgi and Mearns 2002). The reliability of an individual model is determined primarily from two different criteria: model performance bias and model convergence. The reliability factor of the different ensemble members is calculated using the following equation (4):

$$R_i = \left[(R_{B,i})^m X \, (R_{D,i})^n\right]^{\left[\frac{1}{mxn}\right]} \tag{4}$$

$R_{B,i}$ and $R_{D,i}$, in equation (4) can be calculated as shown in equation (5) and (6) respectively:

$$R_{B,i} = \frac{\epsilon_T}{abs(B_{T,i})} \tag{5}$$

$$R_{D,i} = \frac{\epsilon_T}{abs(D_{T,i})} \tag{6}$$

where R_i indicates the model reliability factor of i^{th} ensemble member, $R_{B,i}$ indicates the model reliability for i^{th} ensemble member, which is a function of model bias $B_{T,i}$, and $R_{D,i}$ indicates the model reliability for i^{th} ensemble member, which is calculated based on distance $D_{T,i}$. The parameter ε is a measure of natural variability in the model. The parameters *m* and *n* are the weights given to the two criteria of model performance bias and model convergence. The values of *m* and *n* are typically assigned to 1.

With the computed reliability factors for all the individual models, the calculation of weights for different ensemble members is represented following equation (7):

$$w_i = \frac{R_i}{\sum R} \tag{7}$$

where, w_i is the weight of an i^{th} model in the combined multi-model ensemble, R_i is the reliability factor of an i^{th} model determined using equation (5), and the denominator is the sum of the reliability factors of all the individual members of an ensemble. The detailed process regarding REA can be found in (Giorgi and Mearns 2002).

5.4 Machine Learning (ML) approaches

Two different types of ML approaches are used in this thesis namely Decision Tree (DT) based ML algorithms and Non-tree-based ML algorithms. The primary motivation behind using these different ML algorithms is to ensure diversity in the employed model-averaging techniques. The Decision Tree (DT) based ML models function by creating a tree-like structure where it continuously partitions training datasets into smaller sub-sets such that regression relationship can be generalized between the training and target variables. The DT is considered to be a "white-box" data driven Machine Learning approach as the running of the models are transparent to the end-users (Yang et al. 2021). While partitioning the data, DT models function by following a simple if-then logical statement. Due to the tree-based structure, the DT models split data more efficiently and intuitively. Since the decision tree models can be visualized graphically, the interpretation and explanation of these models are generally easier than other non-tree-based models (Nourani et al. 2019). However, there are instances of the DT models being overfitted or underfitted in the case of insufficient training data or samples.

In this thesis, three tree-based ML algorithms are employed: CART, RF, and XGB. CART is a traditional decision tree-based model whereas RF and XGB have been developed more recently. CART is a simple decision tree regressor which offers more interpretability as compared to RF and XGB. The decision-making process of the CART model can be visualized clearly with ease. CART is robust to outliers and can neglect irrelevant features without requiring extensive computation. On the other hand, RF and XGB are ensemble learning methods which build multiple base models and come up with a single, more accurate prediction by combining them. Since these ensemble learning methods combine the predictions from multiple base models, they reduce the impact of individual model errors and also overcome the potential overfitting problems. As compared to CART, RF and XGB typically provide more accurate predictions as they can capture non-linear and complex relationship of the precipitation variable. Moreover, RF and XGB algorithms are incorporated with bagging and boosting techniques, respectively. Bagging and Boosting are techniques incorporated to reduce the error and optimize the performance of DT-based MLs RF and XGB.

Unlike DT-based ML models, non-DT-based ML models utilize other algorithms that are not based on the construction of decision trees. Non-DT models differ from DT models in their approach to making predictions. In this thesis, I have employed one non-DT-based ML model, SVM. Being a non-DT-based ML, SVM could complement other employed DT-based MLs by providing a different model-averaging approach. Moreover, SVMs are also effective at capturing non-linear relationships associated with precipitation data and work well with problems dealing with high-dimensional spaces.

5.4.1 Classification And Regression Trees (CART)

The CART is a top-down and greedy approach that recursively conducts binary splitting by creating a tree-like structure (Breiman et al. 1984). The CART predicts a continuous target variable and does not rely on any assumptions in terms of the distribution of input/target data samples. CART is a simple tree-based ML algorithm and has served as a baseline to the later developments of other advanced DTs like RF and XGB. CART is predominantly used in both classification and regression tasks. Specifically, the decision tree regressor from CART is used in this thesis for the precipitation variables. Considering that we assume a training dataset with n samples having x_i as input features and y_i as the corresponding target value. After assigning input variables to a CART model, the corresponding output $\hat{y}$ is expressed in the following equation (8):

$$\hat{y} = \sum_{m=1}^{M} c_m I\ (x \epsilon R_m) \tag{8}$$

Where $\hat{y}$ is partitioned into M groups of $[R_1, R_2, \ldots, R_m]$ and the constant value c_m can be estimated by averaging all output values of the CART in group R_m. The iteration of the splitting process in CART is continued until the error of resulting $\hat{y}$ is minimized with given target values. More details regarding CART can be found on (Yang et al. 2021). In this thesis, the maximum tree depth of employed CART is set to be 5. The minimum number of samples required to split an internal node is set to 15; The minimum number of samples required to be at a leaf node is set to 15.

5.4.2 eXtreme Gradient Boosting (XGB)

XGB is a recently developed DT-based algorithm by Chen and Guestrin (2016) and its development is based on the previously developed gradient boosting algorithm. However, unlike traditional gradient boosting, XGB incorporates advanced regularization techniques in its algorithm by constructing a second-order Taylor approximation of the loss function during the training process. XGB is an ensemble learning process that makes use of a boosting technique. It contains multiple CART models as candidates which are subjected to training. All the individual CART models thus created are trained sequentially one after another where, each time, the new CART model will be trained to correct the errors made by the previously trained corresponding CART model. Out of all the individual CART models used for training, XGB boosts the performance of the "weak learners" through an additive strategy (Boutaba et al. 2018). XGB handles both classification and regression problems, and this thesis employs the XGB regressor to predict the precipitation variable.

The final predicted value from the XGB model can be expressed with the following equation (9):

$$\hat{y} = \sum_{i=1}^{n} f_i\ (x_i) \tag{9}$$

Where x_i indicates the simulation of the *ith* ensemble member, f_i is an individual regression tree among a total of *n* regression trees.

While training an XGB model, the regularization term is defined by an objective function which needs to be minimized. This regularized objective function can be expressed with the following equation (10):

$$\sum_{i=1}^{n} l(y_i, \widehat{y_\imath}) + \gamma T + \frac{1}{2}\lambda\|\omega\| \tag{10}$$

Where y_i is the target variable, $\hat{y}_i$ is output from a certain CART model used in an XGB model and *l* is the loss function between the predicted output from an individual CART model and the target variable. The second and third terms are introduced in the objective function to penalize the model based on their complexity. The parameters in the second and third terms of the objective function can be described as: γ is the complexity of an individual leaf, *T* is the number of leaves in an individual CART model, λ is the trade-off parameter used in scaling the penalty, and ω is the vector of scores on the leaves. In this thesis, I have set the hyperparameter set of the employed XGB to be [subsample:0.7, reg_lambda:7, reg_alpha:0.5, objective: reg:squarederror, n_estimators:80, min_child_weight:9, max_depth:4, learning_rate:0.25, gamma:0.5, colsample_bytree:0.005].

5.4.3 Random Forest (RF)

RF is a powerful and robust supervised DT-based algorithm that randomly creates a forest of decision trees (Breiman 2001). It constructs multiple decision trees and combines their results to come up with a single output. Similar to XGB, RF builds multiple CART models as candidates that are subjected to training using the training dataset. RF uses a bagging approach such that a random sample of input features is selected with replacement. The input features thus selected are then fed into individual CART within the RF model. RF can be used for both classification and regression problems. In this thesis, a Random Forest Regressor is used for regression. Eventually, the final predicted value is determined by combining the predicted values from all the individual CART models that were used in creating the RF model. The aggregation is done by taking the average of all the outputs from the CART models, which can be expressed with the following equation (11):

$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{11}$$

Where n indicates the total number of individual CART models used in creating a RF model, y_i is the predicted value from the i^{th} CART model in RF, and y is the final predicted value which is the average of all the predicted values from individual CART models. The hyperparameters of RF were set to [max_depth:3, n_estimators:500, min_samples_leaf:20, max_features:1] in this thesis.

5.4.4 Support Vector Machine (SVM)

SVM is a supervised non-tree-based DT and was first proposed by Cortes and Vapnik (1995). The key step in SVM is to find an optimal hyperplane that distinctly partitions the data points that are passed as the training input variables. Similar to RF and XGB, SVM can also be used in both classification and regression problems. In particular, the Support Vector Regressor (SVR) is used in this thesis. The regression function of SVM that is used to generate the predictions can be expressed with the following equation (12):

$$f(x) = \omega.\Phi(x) + b \qquad (12)$$

Where ω indicates the normal weight vector, b indicates bias, and *Φ(x)* refers to the nonlinear transformation function that maps the input features into a higher-dimensional feature space.

In SVM, the margin between the optimal hyperplane and training data points are maximized with the aid of the loss function. Considering the optimization of the loss function, the SVM regression function can be expressed as in following equation (13):

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)k(X_i, X) + b \qquad (13)$$

Where $k(X_n, X)$ is known as the kernel function and b indicates the bias. The α terms are a series of Lagrange multipliers used to solve the optimization problem. In this study, I have applied the Radial Basis Function (RBF) Kernel.

5.5 Evaluation Statistics

In this thesis, a total of three evaluation statistics are computed and analyzed to quantify the performance of the employed model averaging techniques. The employed evaluation statistics include Percentage Bias (*pbias*), Coefficient of Determination (R^2) and Normalized Root Mean Squared Error (*NRMSE*).

5.5.1 Percentage Bias (pbias)

Percentage bias measures the percentage difference of long-term climatology between SPEAR and reference PRISM precipitation. The $pbias$ can be computed with the following equation (14):

$$pbias = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)}{\sum_{i=1}^{n} y_i} * 100\% \qquad (14)$$

Where n is the total number of observations (i.e., the total length of the data records), y_i is reference precipitation, and $\hat{y}_i$ is simulated precipitation of SPEAR with the application of different model averaging techniques.

5.5.2 Coefficient of Determination (R^2)

In addition to $pbias$, the R^2 is also computed to quantify the simulation skill of SPEAR after the application of different model averaging techniques. The R^2 is a commonly used metric to quantify the "goodness-of-fit" of simulated and reference sequential datasets. The R^2 can be computed with following equation (15):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \qquad (15)$$

Where n is the total number of reference or simulated precipitation (i.e., the total length of data records), y_i is reference precipitation, $\hat{y}_i$ is simulated precipitation of SPEAR with the application of different model averaging techniques, and $\bar{y}_i$ is the average value of the reference precipitation.

5.5.3 Normalized Root Mean Square Error (NRMSE)

The last evaluation statistic that is employed in this thesis is the Normalized Root Mean Square Error (*NRMSE*). The *NRMSE* quantitively reflects the simulation errors. NRMSE is a dimensionless measure calculated from Root Mean Square Error (RMSE), which measures the average difference between the simulated and reference observation datasets. The $NRMSE$ can be computed with following equation (16):

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\bar{y}_i} \tag{16}$$

Where n indicates the total number of observations (i.e., the total length of data records), y_i is reference precipitation, $\hat{y}_i$ is simulated precipitation of SPEAR with the application of different model averaging techniques, and $\bar{y}_i$ is the average value of the reference precipitation.

6 Results

6.1 Performance of the raw precipitation simulation from SPEAR

Figure 6.1 presents the mean monthly climatology of three randomly selected SPEAR members' precipitation simulation alongside with the climatology of the reference precipitation (i.e., PRISM). The blue colors indicate higher precipitation values, whereas the red colors indicate lower precipitation values. Ideally, the individual ensemble members of SPEAR should replicate the observed precipitation climatology from PRISM. According to Figure 6.1, the observed monthly precipitation climatology shows lower precipitation in the northwestern part of Oklahoma and higher precipitation in the southeastern part of Oklahoma. In general, there is an increasing trend of precipitation gradient going from the west to the east of Oklahoma. The randomly selected individual SPEAR ensemble members have preserved the general pattern of precipitation gradient from west to east. However, SPEAR has failed to present the magnitude of precipitation accurately over Oklahoma. Specifically, SPEAR presents higher precipitation than the observation in the panhandle/western part of Oklahoma. SPEAR also presents lower precipitation compared to the observation in the eastern half of Oklahoma.

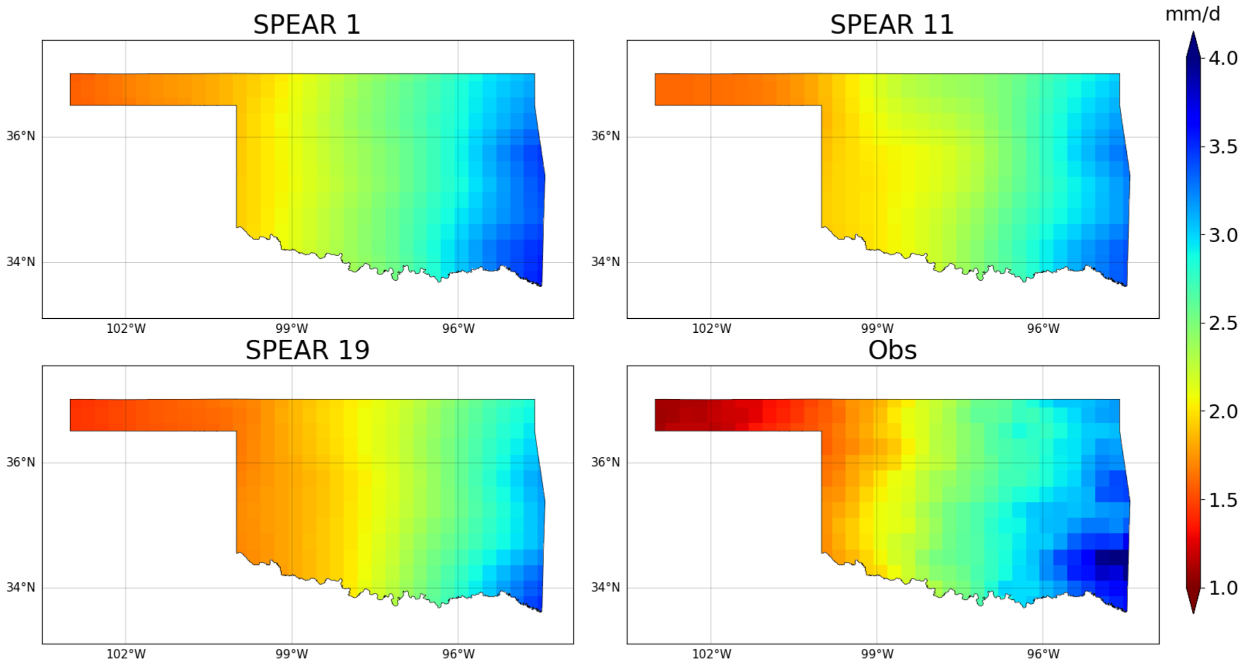

Figure 6.1. Mean monthly precipitation of individual ensemble members from SPEAR (randomly selected the 1st, 11th, and 19th members) and Observations over Oklahoma

Further spatial evaluation of monthly precipitation from randomly selected individual ensemble members of SPEAR is conducted by calculating Percentage Bias for each pixel over Oklahoma, as shown in Figure 6.2. In Figure 6.2, the cooler blue colors indicate a positive bias or overestimation of precipitation, and the warmer red colors indicate a negative bias or underestimation of precipitation. The optimal value of percentage bias is zero which is represented by grey color. In general, all four randomly selected ensemble members show significant bias throughout Oklahoma. Particularly, there is an overestimation of precipitation in the western and panhandle region of Oklahoma, and an underestimation of precipitation in the eastern region of Oklahoma.

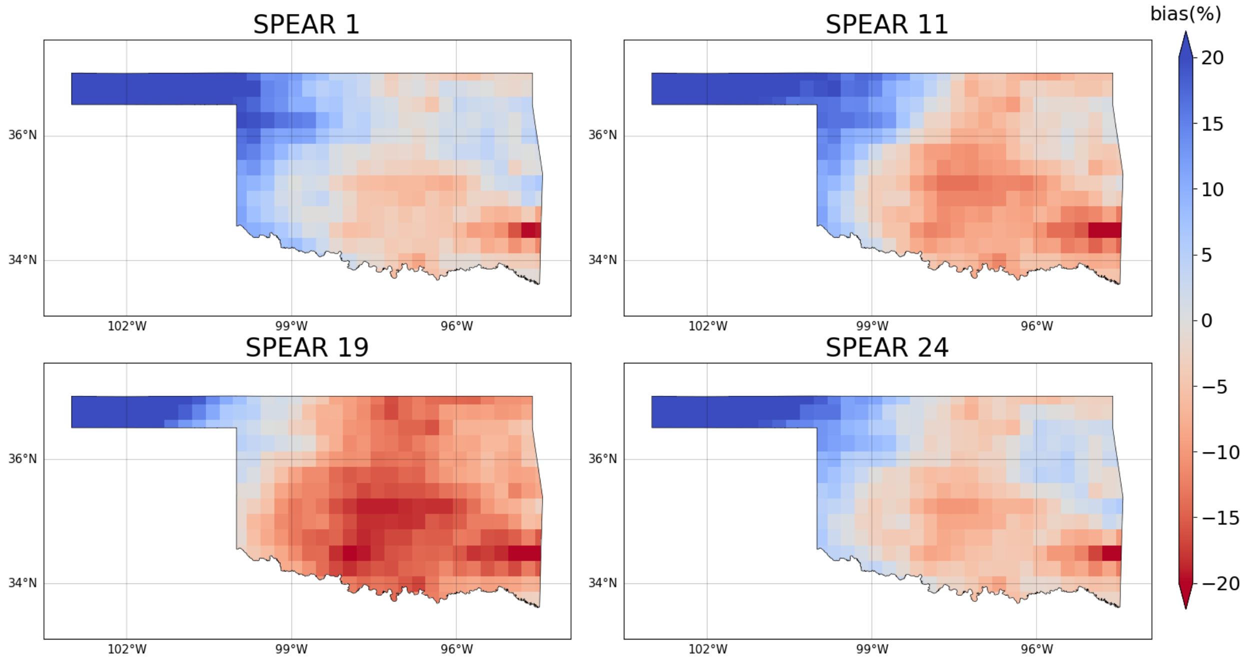

Figure 6.2. Percentage bias of monthly precipitation for individual ensemble members from SPEAR (randomly selected the 1st, 11th, 19th, and 24th members) and Observations over Oklahoma

Figure 6.3 presents the R^2 scores of monthly precipitation for randomly selected individual ensemble members of SPEAR over Oklahoma. In Figure 6.3, the cooler blue colors indicate a positive R^2 score with high simulation skill and the warmer red colors indicate a negative R^2 score with low simulation skill. All the randomly selected individual ensemble members of SPEAR have shown negative R^2 throughout Oklahoma, which represents a complete lack of skill in simulating precipitation.

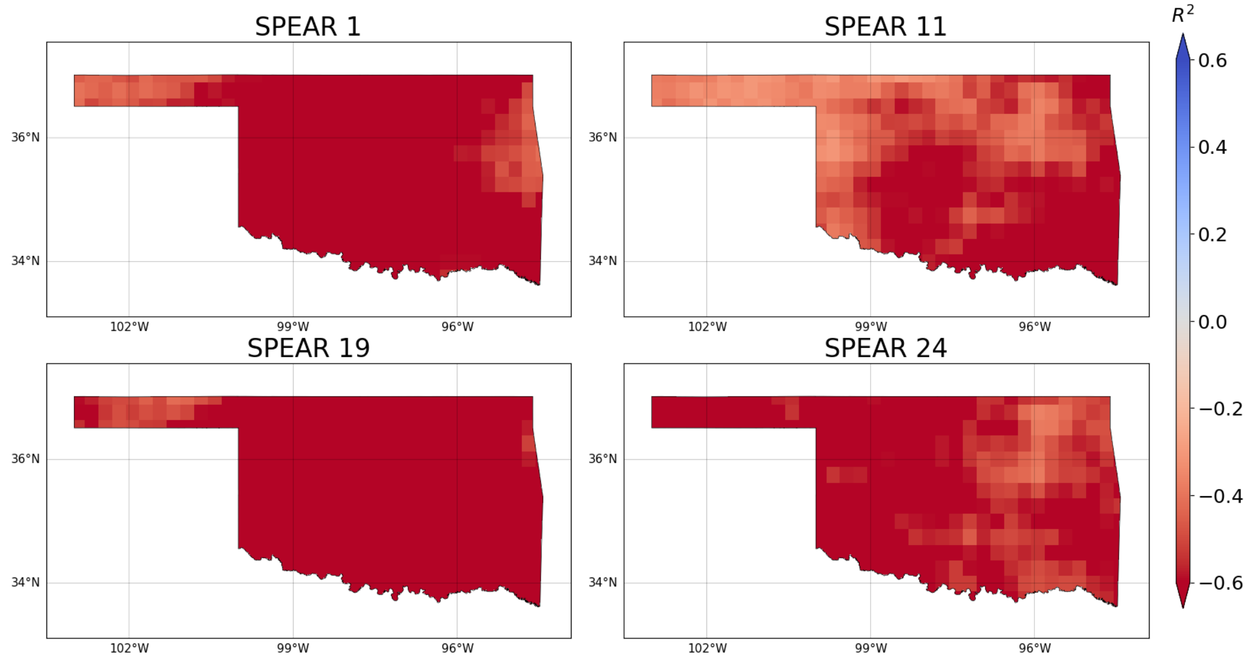

Figure 6.3. Coefficient of Determination (R2) of monthly precipitation for individual ensemble members from SPEAR (randomly selected the 1st, 11th, 19th, and 24th members) and Observations over Oklahoma

6.2 Simulation performance of different model averaging techniques

Figure 6.4 presents the reconstructed monthly mean precipitation over Oklahoma from the period of 1981 to 2014 derived using different model averaging techniques. In Figure 6.4, the blue colors indicate higher precipitation, while the red colors indicate lower precipitation. For an optimal result, the monthly precipitation from different model averaging techniques should match the observed monthly precipitation. The same spatial pattern of the observed monthly precipitation climatology (i.e., west-to-east precipitation gradient) previously observed in Figure 6.1 is also observed in Figure 6.4. From Figure 6.4, it can be observed that the baseline approach of SMA, BMA and REA have preserved the general west-to-east precipitation trend over Oklahoma. Moreover, the baseline approaches of SMA, BMA, and REA have presented higher precipitation than observed precipitation in panhandle or western Oklahoma and lower precipitation than observed precipitation in eastern Oklahoma. So, it is clear that the amount of precipitation is not represented accurately throughout the study region. On the other hand, all the ML-based model averaging techniques have successfully presented the west-to-east precipitation gradient along with the distinct changes in the amount of precipitation across the study region. ML-based techniques show better agreement with the observed monthly mean climatology over Oklahoma as compared to the baseline approaches.

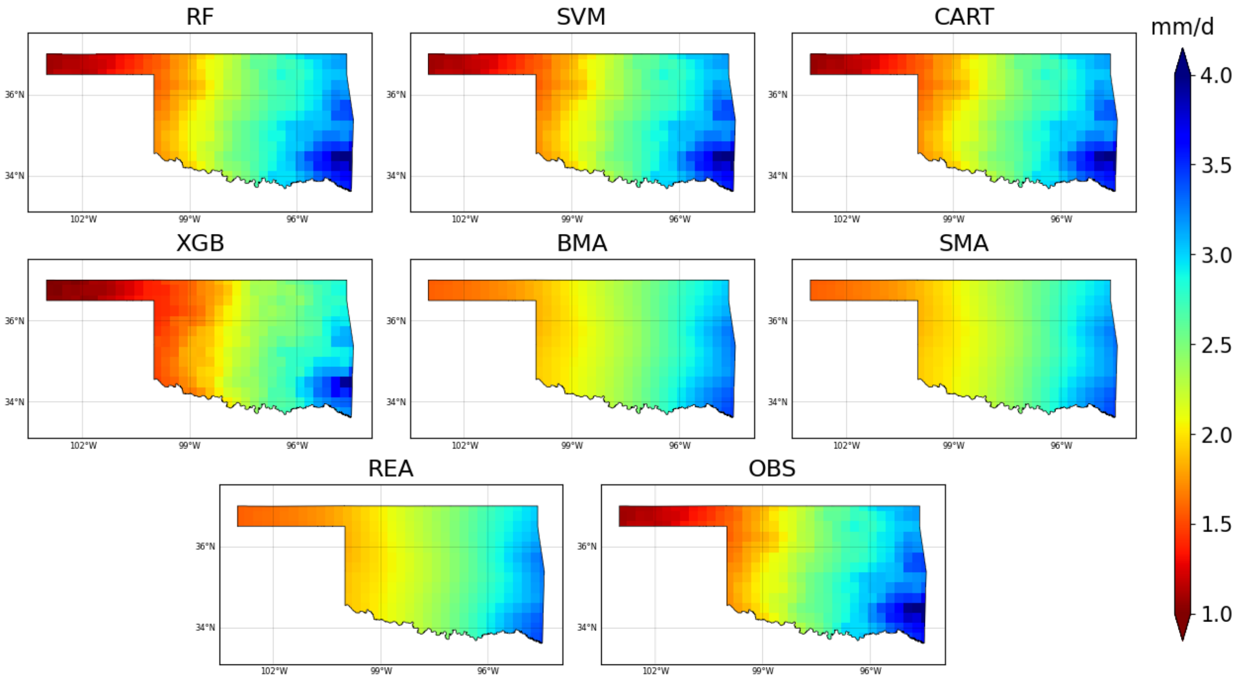

Figure 6.4. Mean monthly precipitation resulted from different model averaging techniques and observations over Oklahoma

Figure 6.5 shows the spatial plot of the percentage bias resulting from the employed model averaging techniques over Oklahoma. In Figure 6.5, the cooler blue colors indicate negative bias, and the warmer red colors indicate positive bias. The optimal value of percentage bias is zero, indicated by grey colors, which implies that there is no bias in the models. From Figure 6.5, It can be observed that the BMA, SMA and REA show very similar patterns over Oklahoma, where they have resulted in positive bias in the northwestern part of Oklahoma and negative bias in the southeastern part of Oklahoma. The ML algorithms (RF, SVM and CART) have delivered superior percentage bias compared to the baseline approaches. In general, MLs have removed the simulation bias almost entirely as compared to benchmark BMA, SMA and REA. Unlike RF, SVM and CART, the application of XGB has presented some degree of negative bias (~10%) throughout the study region.

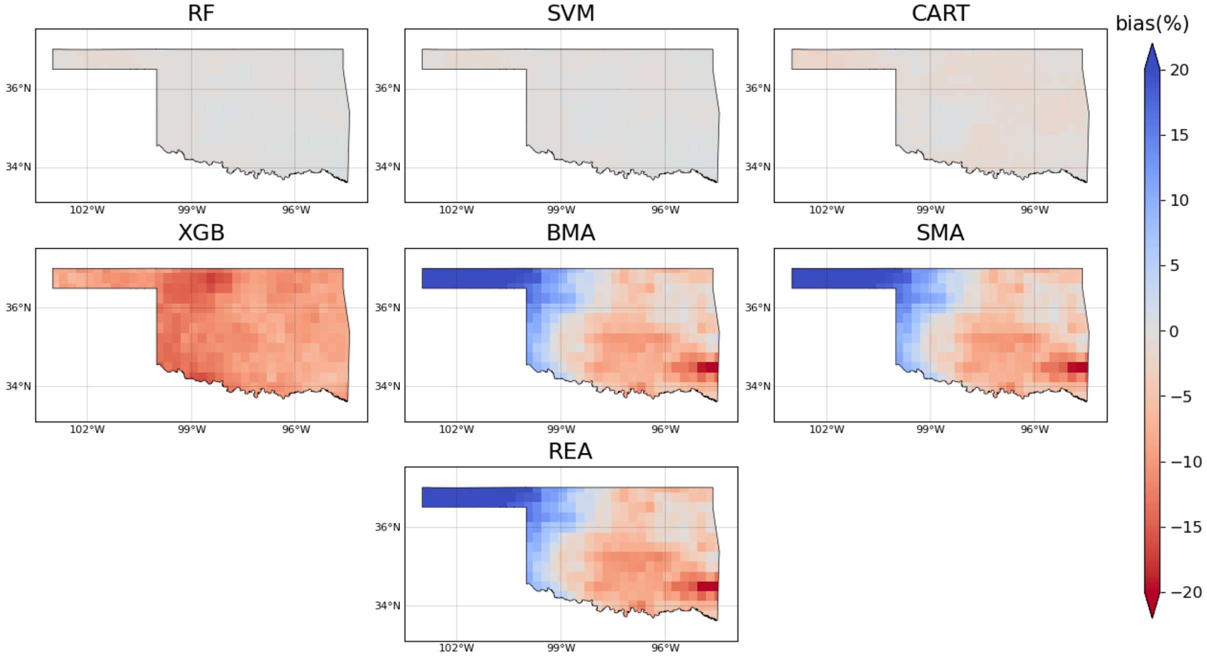

Figure 6.5. Percentage bias of monthly precipitation resulted from different model averaging techniques over Oklahoma

The simulation skill associated with the different model averaging techniques is further quantified and evaluated from the spatial plot of R^2 over Oklahoma as shown in Figure 6.6. The darker green colors in Figure 6.6 represent higher R^2 values (i.e., higher simulation skill), and the lighter green colors represent the lower R^2 values (i.e., lower simulation skills). It is evident that the baseline approaches of SMA, BMA, and REA show lower simulation skills throughout the study region than the ML-based techniques. For instance, SMA, BMA, and REA have nearly zero skills in the northwestern and southeastern regions of Oklahoma. The baseline approaches have presented moderate skill in between the northwestern and southeastern parts, which can be seen in the form of a "stripe". In contrast, the ML-based techniques (RF, SVM, CART, and XGB) displayed comparatively higher simulation skills than most regions of Oklahoma. It should also be noted that, as compared to baseline approaches, the northwestern part of Oklahoma has shown significant improvement in the simulation skill, while the southeastern regions do not show significant improvements. In between the northwestern and southeastern Oklahoma (i.e., the mid-regions of Oklahoma), the MLs have presented moderate improvements in simulation skill. However, these improvements in the mid-regions are not significant as the pixels with low skill are scattered around the area. The employed MLs show a similar level of simulation skill throughout Oklahoma, and no significant differences can be observed in their performance.

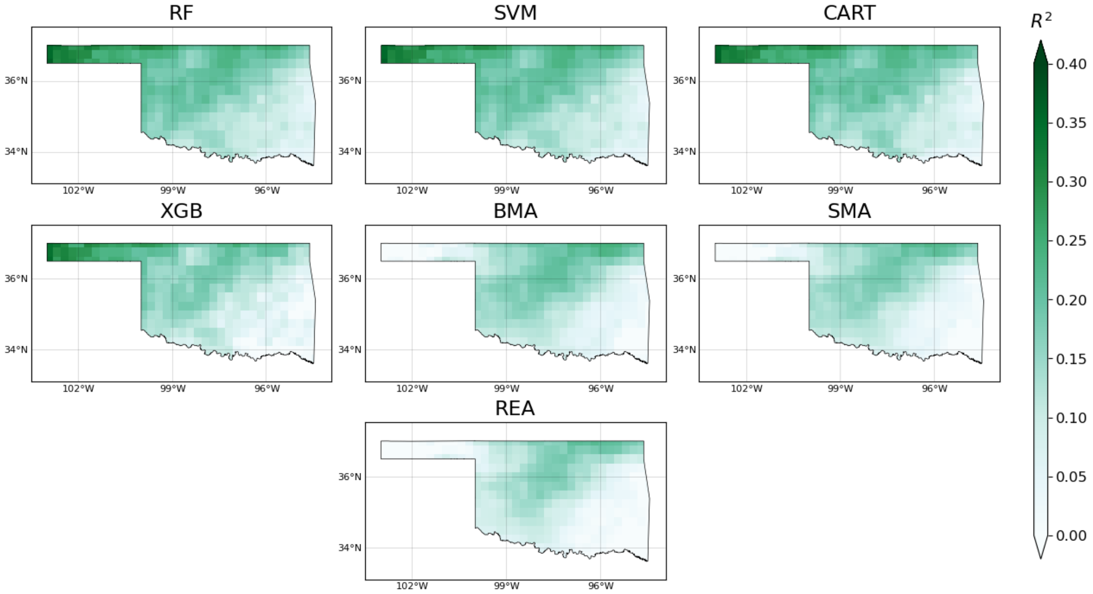

Figure 6.6. Coefficient of Determination (R2) resulted from different model averaging techniques over Oklahoma

6.3 Seasonal evaluation statistics of different model averaging techniques and individual SPEAR ensembles

Figure 6.7 and Figure 6.8 present the R^2 values of the spatially averaged precipitation across different months over Oklahoma. Specifically, Figure 6.7 presents the R^2 values resulting from the employed model averaging techniques, and Figure 6.8 presents the R^2 values of the 30 individual ensemble members of SPEAR. In Figure 6.7, R^2 values from each model averaging technique in each month are labeled within the corresponding boxes, aiding in more detailed comparison. In both Figure 6.7 and Figure 6.8, the cooler blue colors indicate higher simulation skill with higher value of $R^{2,}$ whereas the warmer red colors indicate lower simulation skill with lower value of R^2.

Comparing Figure 6.7 and Figure 6.8, it is evident that all the model averaging techniques have significantly improved the R^2 score over individual ensemble members of SPEAR across all the months. As shown in Figure 6.8, most of the individual ensemble members show negative simulation skills across all the months except for the colder months in December and January where R^2 scores are slightly above zero. However, the R^2 scores obtained after applying model averaging techniques display significant improvements in simulation skills in all the months with a clear seasonal pattern. The employment of different model averaging techniques has presented higher simulation skills in the colder seasons and lower simulation skills in the warmer seasons. Even though there are improvements in both warmer and colder months, the skills in the warmer months of June, July and August are still in the lower spectrum. Moreover, a comparison can also be made between the different model averaging techniques where the MLs have outperformed the baseline approaches of BMA, SMA and REA in all the months. Among the employed MLs, RF, XGB and CART have presented overall better performance than the SVM.

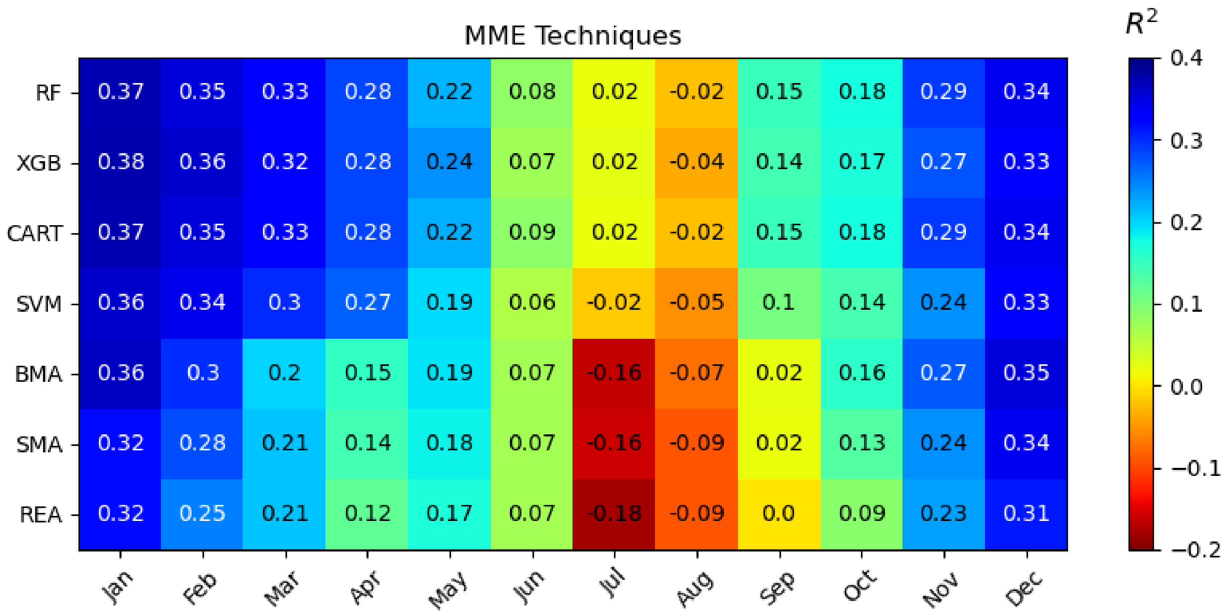

Figure 6.7. Spatially averaged Coefficient of Determination (R2) resulted from different model averaging techniques across the months

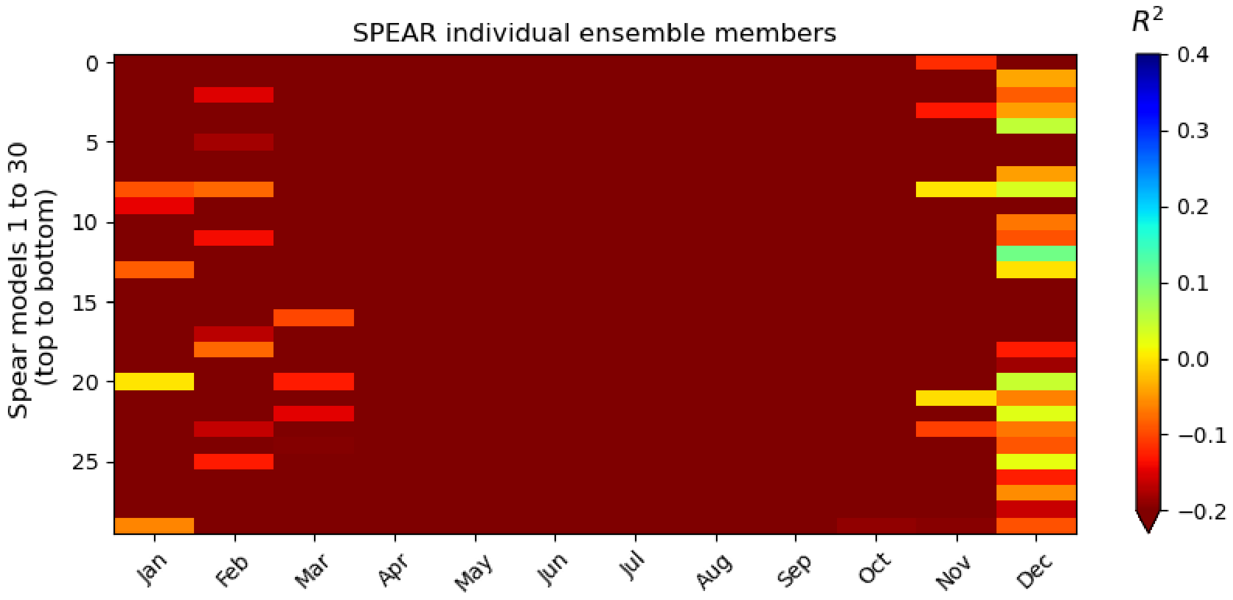

Figure 6.8. Spatially averaged Coefficient of Determination (R2) resulted from individual ensemble members of SPEAR across the months

Similarly, Figure 6.9 and Figure 6.10 present the NRMSE values across all the months for the spatially averaged precipitation over Oklahoma. Figure 6.9 presents the NRMSE values resulting from the employed model averaging techniques and Figure 6.10 presents the NRMSE values of the 30 individual ensemble members of SPEAR. In Figure 6.9, the NRMSE values obtained from each model averaging technique in each month are labeled within the corresponding boxes, aiding in a more detailed comparison. In both Figure 6.9 and Figure 6.10, the cooler blue colors indicate lower simulation error with lower NRMSE values, whereas the warmer red colors indicate higher simulation error with higher NRMSE values.

The NRMSE results obtained in Figure 6.9 and Figure 6.10 are consistent with the R^2 score results observed in Figure 6.7 and Figure 6.8. Significant differences can be observed between the plots showing NRMSE of different model averaging techniques and individual SPEAR ensembles. In Figure 6.10, all the individual ensemble members of SPEAR show higher errors across all the months except for the cooler months of November, December, and January where the errors are comparatively lower. As compared to the individual ensemble members, all the model averaging techniques have presented better simulation performance. Even though there is improvement in all the months after the application of model averaging techniques, the warmer months of July and August are still associated with higher errors. As a result, a clear seasonal pattern can be observed in the performance of different model averaging techniques. Moreover, comparisons can also be made between the different model averaging techniques. All the ML-based techniques have delivered lower NRMSE than the baseline SMA, BMA, and REA in all the months. Among the employed ML techniques, the RF, XGB, and CART present overall better performance than the SVM.

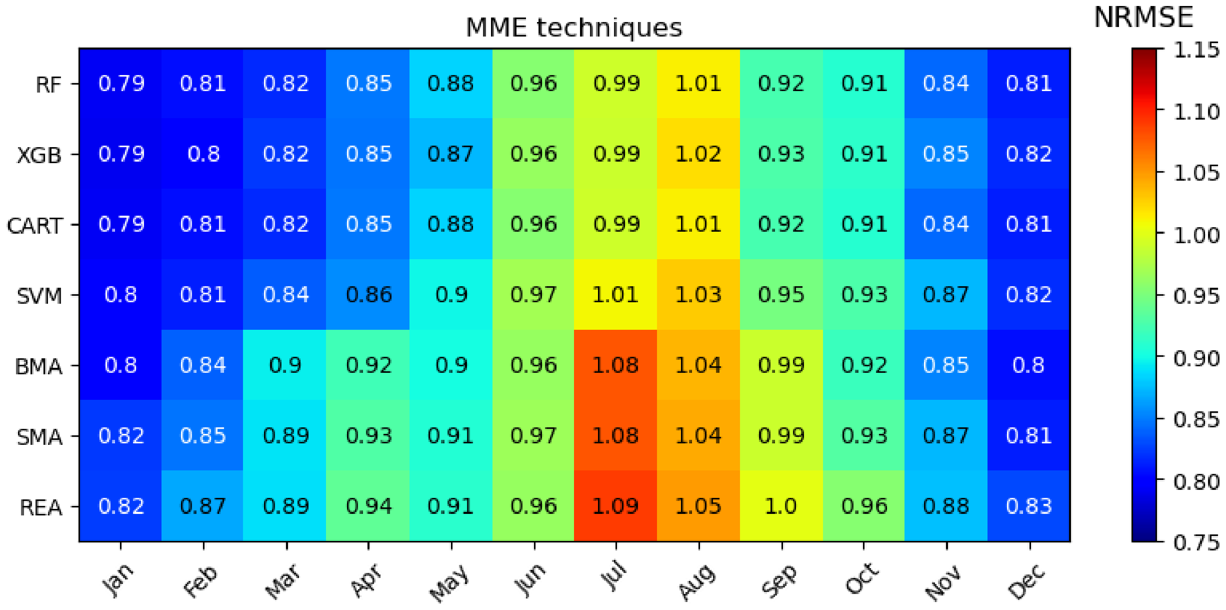

Figure 6.9. Spatially averaged NRMSE resulted from different model averaging techniques across the months

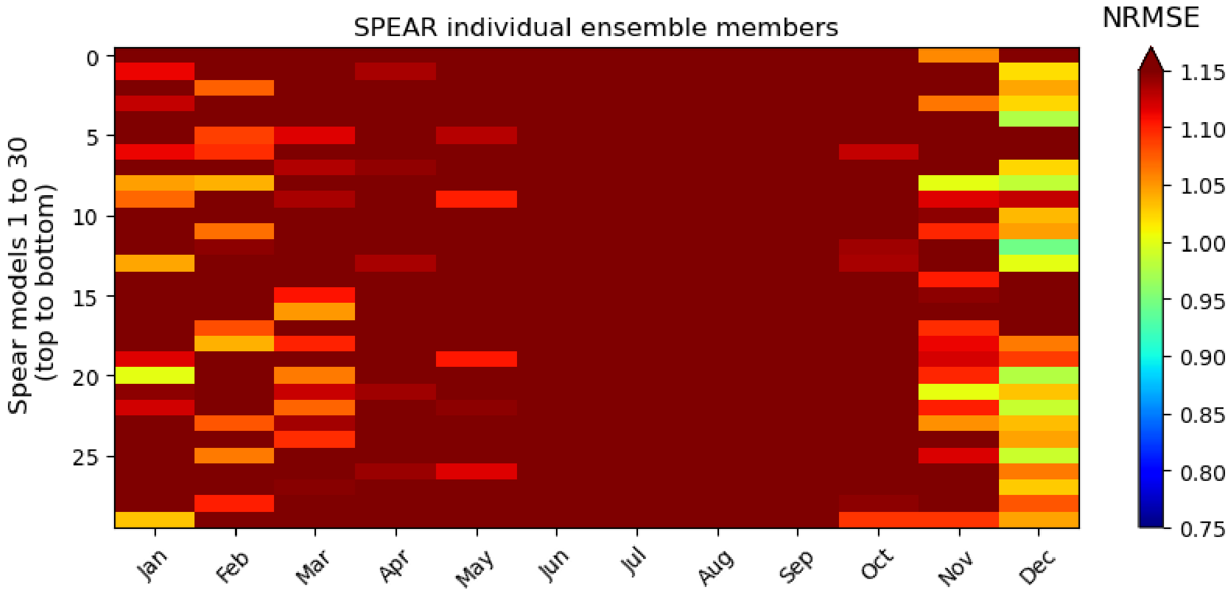

Figure 6.10. Spatially averaged NRMSE resulted from individual ensemble members of SPEAR across the months

7 Discussion

As presented in the results section, the Percentage Bias, R^2 score, and NRMSE values of the individual ensemble members of SPEAR indicate an extremely limited simulation skill of historical monthly precipitation over Oklahoma. After applying different model averaging techniques, the reduced Percentage Bias, improved R^2 score, and lower NRMSE values can be observed. This performance improvement suggests that the different model averaging techniques are effective in reconstructing the historical monthly precipitation over Oklahoma.

Among different model averaging techniques, the employed MLs have shown superior performance to the baseline SMA, BMA, and REA. This is because, as compared to SMA, BMA and REA, the ML-based model averaging techniques can combine the individual ensemble members optimally producing better simulation results (Sloughter et al. 2007). The superior performance of the ML-based techniques can be attributed to MLs' capacity to extract nonlinear, high dimensional and complex patterns between the climatic variables simulated by the climate models (Dey et al. 2022, Li et al. 2021b). Moreover, the DT-based ML algorithms employed in this thesis function by building decision trees. As a result, these DT-based techniques can be used to assign weights to the individual ensemble members by generating feature importance values. The weights thus generated can then be evaluated and compared with the weights generated from the baseline approaches of BMA and REA.

Despite improving the simulation performance, ML-based model averaging techniques do have their caveats which affect the performance of MLs applications. For example, MLs are considered as non-physics-based algorithms which hinder the understanding of mechanisms governing the climate system (Jebeile et al. 2021, Reichstein et al. 2019). While MLs can capture complex non-linear relationships effectively as compared to baseline approaches, MLs might not be able to capture all the nuances and intricacies involved with precipitation since precipitation can be highly non-linear and can have very complex patterns. Other potential limitations of the MLs can be the requirement of a large amount of data to train the ML models, and the issue of overfitting and underfitting of the models. Moreover, MLs assume that the successive values of precipitation in a time series are sequentially independent of each other. However, the precipitation data can exhibit dependencies between successive values (Jose et al., 2022). As a result, the more advanced data-driven techniques like Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN) shall produce better simulation results as they can learn these sequential dependencies of precipitation variables.

On the other hand, even after improving simulation performance with ML-based model averaging techniques, the reconstructed monthly precipitation over Oklahoma still presents performance variations in different seasons and at different geolocations. Given the complex precipitation mechanisms over Oklahoma, I suspect such seasonal and spatial performance patterns of the reconstructed precipitation can be attributed to the inherent limitation of GCMs. To be specific, high rainfall in Oklahoma during summer months is contributed by high levels of moisture and convective instability (Bradley and Smith 1994, Hand and Shepherd 2009). Oklahoma receives much of its precipitation during the warmer summer months, with nearly half of the precipitation being contributed by Mesoscale Convective Systems (MCS) (Easterling et al. 2017, Fritsch et al. 1986). But for GCMs, convective systems are neither perfectly parameterized nor fully resolved (Moncrieff 2019). Moreover, the formation of MCSs occurs within smaller spatial scales that are typically smaller than GCM grids (Eden et al. 2012). Therefore, it is reasonable to suspect that the available GCMs have little skill in simulating such precipitation during warm seasons in Oklahoma. With compromised GCM simulations, it is natural that different model averaging techniques are extremely limited and present comparably inferior performance during warm seasons.

In light of the potential limitation of GCMs during warm seasons, I reckon further improvement of GCMs shall lead to overall more accurate and reliable precipitation simulation/projection. Such improvements of GCMs can be made through three different approaches. The first approach in making advancements in the GCMs can be done by enhancing the resolution of the GCMs. The GCMs with higher resolutions can improve the simulation accuracy of precipitation as it can accurately represent small-scale features, and atmosphere dynamics, and lead to better reproduction of large-scale precipitation patterns (Gao et al. 2008, Mishra et al. 2023). Secondly, GCMs should include and better represent more physical and dynamical forces of the climate system (Chen et al. 2021, Wu et al. 2020). Finally, the lack of inclusion of the physical processes can be better represented through the parameterization schemes of precipitation. With parameterization schemes of precipitation, the sub-grid physical process can be represented more accurately, and the observed climatic variables can be matched better by the GCMs (Demory et al. 2020). I believe that by enhancing the resolution of GCMs, including more physical processes and/or by better parameterization schemes of precipitation, advancements in the GCMs can be made. As a result, these advancements in the GCMs shall lead to more accurate precipitation simulation/projection.

In summary, this thesis has demonstrated the capability of different model averaging techniques in combining the multi-model ensemble from SPEAR to reconstruct the historical monthly precipitation over Oklahoma. The baseline techniques of SMA, BMA and REA presented an advantage over individual ensemble members of SPEAR. The employed ML-based techniques produced superior performance compared to the baseline techniques in all the evaluation statistics for Oklahoma. I believe that the result from this thesis highlights the potential success of other data-driven ML or deep learning techniques in combining multi-model ensembles in the future. On the other hand, I also reckon that newer datasets from more advanced GCMs can be utilized to reconstruct historical precipitation. I believe that the utilization of more accurate and skillful simulation from newer GCMs shall provide a better representation of the observed historical precipitation. And to advance the GCM simulations, the GCMs shall have finer spatial resolutions along with the inclusion of more physical and dynamic processes of the atmosphere.

8 Conclusions

In this thesis, a total of seven different model averaging techniques are utilized to combine the historical monthly precipitation from 30 ensemble members of NOAA's SPEAR. The thesis is conducted over Oklahoma during a study period from 1981 to 2014. Out of seven different model averaging techniques, three baseline approaches of Simple Model Averaging (SMA), Bayesian Model Averaging (BMA) and Reliability Ensemble Averaging (REA), and four Machine Learning (ML) techniques of Classification And Regression Trees (CART), Random Forest (RF), eXtreme Gradient Boosting (XGB) and Support Vector Machine (SVM) are employed. The baseline approaches of SMA, BMA and REA are used as benchmarks to evaluate the performance of the ML algorithms. To validate the performance of precipitation simulation, three different evaluation statistics, namely Percentage Bias, Coefficient of Determination, and Normalized Root Mean Square Error, are used. These evaluation statistics are employed in analyzing and quantifying the spatial and temporal characteristics of precipitation over Oklahoma. The major conclusions from this thesis are listed as follows:

1. All employed model averaging techniques have improved the simulation performance as compared to the individual ensemble members of SPEAR.
2. Among the employed model averaging techniques, the ML approaches lead to better simulation skills and lower simulation bias than baseline BMA, SMA and REA.
3. The decision-tree-based ML algorithms (RF, XGB and CART) slightly outperformed the non-tree-based ML (SVM) in simulating the historical monthly precipitation over Oklahoma.
4. Strong spatial and temporal patterns have been observed in the performance of model averaging techniques while simulating precipitation over Oklahoma.

5. The results from this thesis suggest that novel and more advanced data-driven techniques as well as GCMs, have the potential to improve the performance of GCM-simulated precipitation.

9 References

Abdelmoaty, H.M., Papalexiou, S.M., Rajulapati, C.R. and AghaKouchak, A. (2021) Biases beyond the mean in CMIP6 extreme precipitation: A global investigation. Earth's Future 9(10), e2021EF002196.

Ahmed, K., Sachindra, D., Shahid, S., Iqbal, Z., Nawaz, N. and Khan, N. (2020) Multi-model ensemble predictions of precipitation and temperature using machine learning algorithms. Atmospheric research 236, 104806.

Allen, R.G. and Gichuki, F.N. (1989) Effects of projected CO2-induced climate changes on irrigation water requirements in the Great Plains states (Texas, Oklahoma, Kansas and Nebraska). The Potential Effects of Global Climate Change on the United States 1.

Aloysius, N.R., Sheffield, J., Saiers, J.E., Li, H. and Wood, E.F. (2016) Evaluation of historical and future simulations of precipitation and temperature in central Africa from CMIP5 climate models. Journal of Geophysical Research: Atmospheres 121(1), 130-152.

Balhane, S., Driouech, F., Chafki, O., Manzanas, R., Chehbouni, A. and Moufouma-Okia, W. (2022) Changes in mean and extreme temperature and precipitation events from different weighted multi-model ensembles over the northern half of Morocco. Climate dynamics 58(1-2), 389-404.

Becker, E.J., Kirtman, B.P., L’Heureux, M., Muñoz, Á.G. and Pegion, K. (2022) A decade of the North American Multimodel Ensemble (NMME): Research, application, and future directions. Bulletin of the American meteorological Society 103(3), E973-E995.

Berg, A., Sheffield, J. and Milly, P.C. (2017) Divergent surface and total soil moisture projections under global warming. Geophysical Research Letters 44(1), 236-244.

Boutaba, R., Salahuddin, M.A., Limam, N., Ayoubi, S., Shahriar, N., Estrada-Solano, F. and Caicedo, O.M. (2018) A comprehensive survey on machine learning for networking: evolution, applications and research opportunities. Journal of Internet Services and Applications 9(1), 1-99.

Bradley, A.A. and Smith, J.A. (1994) The hydrometeorological environment of extreme rainstorms in the southern plains of the United States. Journal of Applied Meteorology (1988-2005), 1418-1431.

Brands, S. (2022) Common error patterns in the regional atmospheric circulation simulated by the CMIP multi-model ensemble. Geophysical Research Letters 49(23), e2022GL101446.

Breiman, L. (2001) Random forests. Machine learning 45, 5-32.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984) Cart. Classification and regression trees.

Brekke, L.D. (2009) Climate change and water resources management: A federal perspective, Diane Publishing.

Buban, M.S., Lee, T.R. and Baker, C.B. (2020) A comparison of the US climate reference network precipitation data to the parameter-elevation regressions on independent slopes model (PRISM). Journal of Hydrometeorology 21(10), 2391-2400.

Bushuk, M., Winton, M., Haumann, F.A., Delworth, T., Lu, F., Zhang, Y., Jia, L., Zhang, L., Cooke, W. and Harrison, M. (2021) Seasonal prediction and predictability of regional Antarctic sea ice. Journal of Climate 34(15), 6207-6233.

Chen, J., Arsenault, R., Brissette, F.P. and Zhang, S. (2021) Climate change impact studies: Should we bias correct climate model outputs or post-process impact model outputs? Water Resources Research 57(5), e2020WR028638.

Chen, T. and Guestrin, C. (2016) Xgboost: A scalable tree boosting system, pp. 785-794.

Cortes, C. and Vapnik, V. (1995) Support-vector networks. Machine learning 20, 273-297.

Crawford, J., Venkataraman, K. and Booth, J. (2019) Developing climate model ensembles: A comparative case study. Journal of Hydrology 568, 160-173.

Dai, A. (2006) Precipitation characteristics in eighteen coupled climate models. Journal of Climate 19(18), 4605-4630.

Daly, C. and Bryant, K. (2013) The PRISM climate and weather system—an introduction. Corvallis, OR: PRISM climate group 2.

Delworth, T.L., Cooke, W.F., Adcroft, A., Bushuk, M., Chen, J.H., Dunne, K.A., Ginoux, P., Gudgel, R., Hallberg, R.W. and Harris, L. (2020) SPEAR: The next generation GFDL modeling system for seasonal to multidecadal prediction and projection. Journal of Advances in Modeling Earth Systems 12(3), e2019MS001895.

Demory, M.-E., Berthou, S., Fernández, J., Sørland, S.L., Brogli, R., Roberts, M.J., Beyerle, U., Seddon, J., Haarsma, R. and Schär, C. (2020) European daily precipitation according to EURO-CORDEX regional climate models (RCMs) and high-resolution global climate models (GCMs) from the High-Resolution Model Intercomparison Project (HighResMIP). Geoscientific Model Development 13(11), 5485-5506.

Demory, M.-E., Vidale, P.L., Roberts, M.J., Berrisford, P., Strachan, J., Schiemann, R. and Mizielinski, M.S. (2014) The role of horizontal resolution in simulating drivers of the global hydrological cycle. Climate dynamics 42, 2201-2225.

Dey, A., Sahoo, D.P., Kumar, R. and Remesan, R. (2022) A multimodel ensemble machine learning approach for CMIP6 climate model projections in an Indian River basin. International Journal of Climatology 42(16), 9215-9236.

Duan, K., Wang, X., Liu, B., Zhao, T. and Chen, X. (2021) Comparing Bayesian model averaging and reliability ensemble averaging in post-processing runoff projections under climate change. Water 13(15), 2124.

Duan, Q. and Phillips, T.J. (2010) Bayesian estimation of local signal and noise in multimodel simulations of climate change. Journal of Geophysical Research: Atmospheres 115(D18).

Dufresne, J.-L., Foujols, M.-A., Denvil, S., Caubel, A., Marti, O., Aumont, O., Balkanski, Y., Bekki, S., Bellenger, H. and Benshila, R. (2013) Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5. Climate dynamics 40, 2123-2165.

Easterling, D.R., Arnold, J., Knutson, T., Kunkel, K., LeGrande, A., Leung, L.R., Vose, R., Waliser, D. and Wehner, M. (2017) Precipitation change in the United States.

Eden, J.M., Widmann, M., Grawe, D. and Rast, S. (2012) Skill, correction, and downscaling of GCM-simulated precipitation. Journal of Climate 25(11), 3970-3984.

Eyring, V., Bony, S., Meehl, G.A., Senior, C.A., Stevens, B., Stouffer, R.J. and Taylor, K.E. (2016) Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. Geoscientific Model Development 9(5), 1937-1958.

Fisher, B.L. (2004) Climatological validation of TRMM TMI and PR monthly rain products over Oklahoma. Journal of Applied Meteorology 43(3), 519-535.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S.C., Collins, W., Cox, P., Driouech, F., Emori, S. and Eyring, V. (2014) Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, pp. 741-866, Cambridge University Press.

Ford, T.W., Quiring, S.M., Frauenfeld, O.W. and Rapp, A.D. (2015a) Synoptic conditions related to soil moisture-atmosphere interactions and unorganized convection in Oklahoma. Journal of Geophysical Research: Atmospheres 120(22), 11,519-511,535.

Ford, T.W., Rapp, A.D. and Quiring, S.M. (2015b) Does afternoon precipitation occur preferentially over dry or wet soils in Oklahoma? Journal of Hydrometeorology 16(2), 874-888.

Fragoso, T.M., Bertoli, W. and Louzada, F. (2018) Bayesian model averaging: A systematic review and conceptual classification. International Statistical Review 86(1), 1-28.

Fritsch, J., Kane, R. and Chelius, C. (1986) The contribution of mesoscale convective weather systems to the warm-season precipitation in the United States. Journal of Applied Meteorology and Climatology 25(10), 1333-1345.

Gao, X., Shi, Y., Song, R., Giorgi, F., Wang, Y. and Zhang, D. (2008) Reduction of future monsoon precipitation over China: comparison between a high resolution RCM simulation and the driving GCM. Meteorology and Atmospheric Physics 100, 73-86.

Giorgi, F. and Mearns, L.O. (2002) Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the "reliability ensemble averaging"(REA) method. Journal of Climate 15(10), 1141-1158.

Hand, L.M. and Shepherd, J.M. (2009) An investigation of warm-season spatial rainfall variability in Oklahoma City: Possible linkages to urbanization and prevailing wind. Journal of Applied Meteorology and Climatology 48(2), 251-269.

Handmer, J., Honda, Y., Kundzewicz, Z.W., Arnell, N., Benito, G., Hatfield, J., Mohamed, I.F., Peduzzi, P., Wu, S. and Sherstyukov, B. (2012) Managing the risks of extreme events and disasters to advance climate change adaptation: Special report of the Intergovernmental Panel on Climate Change, pp. 231-290, Cambridge University Press (CUP).

Hawkins, E. and Sutton, R. (2011) The potential to narrow uncertainty in projections of regional precipitation change. Climate dynamics 37, 407-418.

Hinne, M., Gronau, Q.F., van den Bergh, D. and Wagenmakers, E.-J. (2020) A conceptual introduction to Bayesian model averaging. Advances in Methods and Practices in Psychological Science 3(2), 200-215.

Hunke, E.C., Lipscomb, W.H. and Turner, A.K. (2010) Sea-ice models for climate study: retrospective and new directions. Journal of Glaciology 56(200), 1162-1172.

Ingram, W. and Bushell, A.C. (2021) Sensitivity of climate feedbacks to vertical resolution in a general circulation model. Geophysical Research Letters 48(12), e2020GL092268.

Jebeile, J. and Crucifix, M. (2020) Multi-model ensembles in climate science: Mathematical structures and expert judgements. Studies in History and Philosophy of Science Part A 83, 44-52.

Jebeile, J., Lam, V. and Räz, T. (2021) Understanding climate change with statistical downscaling and machine learning. Synthese 199, 1877-1897.

Ji, L., Zhi, X., Zhu, S. and Fraedrich, K. (2019) Probabilistic precipitation forecasting over East Asia using Bayesian model averaging. Weather and Forecasting 34(2), 377-392.

Jiang, S., Ren, L., Hong, Y., Yong, B., Yang, X., Yuan, F. and Ma, M. (2012) Comprehensive evaluation of multi-satellite precipitation products with a dense rain gauge network and optimally merging their simulated hydrological flows using the Bayesian model averaging method. Journal of Hydrology 452, 213-225.

Johnson, N.C., Wittenberg, A.T., Rosati, A.J., Delworth, T.L. and Cooke, W. (2022) Future changes in boreal winter ENSO teleconnections in a large ensemble of high-resolution climate simulations. Frontiers in Climate 4, 941055.

Jose, D.M., Vincent, A.M. and Dwarakish, G.S. (2022) Improving multiple model ensemble predictions of daily precipitation and temperature through machine learning techniques. Scientific Reports 12(1), 4678.

Knutti, R. and Sedláček, J. (2013) Robustness and uncertainties in the new CMIP5 climate model projections. Nature climate change 3(4), 369-373.

Kumar, S., Merwade, V., Kinter III, J.L. and Niyogi, D. (2013) Evaluation of temperature and precipitation trends and long-term persistence in CMIP5 twentieth-century climate simulations. Journal of Climate 26(12), 4168-4185.

Lambert, S.J. and Boer, G.J. (2001) CMIP1 evaluation and intercomparison of coupled climate models. Climate dynamics 17, 83-106.

Leduc, M., Laprise, R., De Elia, R. and Šeparović, L. (2016) Is institutional democracy a good proxy for model independence? Journal of Climate 29(23), 8301-8316.

Ley, E. and Steel, M.F. (2009) On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. Journal of applied econometrics 24(4), 651-674.

Li, D., Marshall, L., Liang, Z. and Sharma, A. (2022) Hydrologic multi-model ensemble predictions using variational Bayesian deep learning. Journal of Hydrology 604, 127221.

Li, J., Huo, R., Chen, H., Zhao, Y. and Zhao, T. (2021a) Comparative assessment and future prediction using CMIP6 and CMIP5 for annual precipitation and extreme precipitation simulation. Frontiers in Earth Science 9, 687976.

Li, T., Jiang, Z., Le Treut, H., Li, L., Zhao, L. and Ge, L. (2021b) Machine learning to optimize climate projection over China with multi-model ensemble simulations. Environmental Research Letters 16(9), 094028.

Liu, Z. and Merwade, V. (2018) Accounting for model structure, parameter and input forcing uncertainty in flood inundation modeling using Bayesian model averaging. Journal of Hydrology 565, 138-149.

Ma, Y., Hong, Y., Chen, Y., Yang, Y., Tang, G., Yao, Y., Long, D., Li, C., Han, Z. and Liu, R. (2018) Performance of optimally merged multisatellite precipitation products using the dynamic Bayesian model averaging scheme over the Tibetan Plateau. Journal of Geophysical Research: Atmospheres 123(2), 814-834.

Mani, A. and Tsai, F.T.-C. (2017) Ensemble averaging methods for quantifying uncertainty sources in modeling climate change impact on runoff projection. Journal of Hydrologic Engineering 22(4), 04016067.

Massoud, E., Lee, H., Gibson, P., Loikith, P. and Waliser, D. (2020) Bayesian model averaging of climate model projections constrained by precipitation observations over the contiguous United States. Journal of Hydrometeorology 21(10), 2401-2418.

McGovern, A., Elmore, K.L., Gagne, D.J., Haupt, S.E., Karstens, C.D., Lagerquist, R., Smith, T. and Williams, J.K. (2017) Using artificial intelligence to improve real-time decision-making for high-impact weather. Bulletin of the American meteorological Society 98(10), 2073-2090.

Mehran, A., AghaKouchak, A. and Phillips, T.J. (2014) Evaluation of CMIP5 continental precipitation simulations relative to satellite-based gauge-adjusted observations. Journal of Geophysical Research: Atmospheres 119(4), 1695-1707.

Miao, C., Duan, Q., Sun, Q., Huang, Y., Kong, D., Yang, T., Ye, A., Di, Z. and Gong, W. (2014) Assessment of CMIP5 climate models and projected temperature changes over Northern Eurasia. Environmental Research Letters 9(5), 055007.

Mishra, A.K., Dubey, A.K. and Dinesh, A.S. (2023) Diagnosing whether the increasing horizontal resolution of regional climate model inevitably capable of adding value: investigation for Indian summer monsoon. Climate dynamics 60(7), 1925-1945.

Mitra, A., Iyengar, G., Durai, V., Sanjay, J., Krishnamurti, T., Mishra, A. and Sikka, D. (2011) Experimental real-time multi-model ensemble (MME) prediction of rainfall during monsoon 2008: Large-scale medium-range aspects. Journal of earth system science 120, 27-52.

Moncrieff, M.W. (2019) Toward a dynamical foundation for organized convection parameterization in GCMs. Geophysical Research Letters 46(23), 14103-14108.

Murakami, H., Delworth, T.L., Cooke, W.F., Zhao, M., Xiang, B. and Hsu, P.-C. (2020) Detected climatic change in global distribution of tropical cyclones. Proceedings of the National Academy of Sciences 117(20), 10706-10714.

Murphy, J.M., Sexton, D.M., Barnett, D.N., Jones, G.S., Webb, M.J., Collins, M. and Stainforth, D.A. (2004) Quantification of modelling uncertainties in a large ensemble of climate change simulations. Nature 430(7001), 768-772.

Nonejad, N. (2021) An overview of dynamic model averaging techniques in time-series econometrics. Journal of Economic Surveys 35(2), 566-614.

Nourani, V., Razzaghzadeh, Z., Baghanam, A.H. and Molajou, A. (2019) ANN-based statistical downscaling of climatic parameters using decision tree predictor screening method. Theoretical and Applied Climatology 137, 1729-1746.

O'Neill, B.C., Tebaldi, C., Van Vuuren, D.P., Eyring, V., Friedlingstein, P., Hurtt, G., Knutti, R., Kriegler, E., Lamarque, J.-F. and Lowe, J. (2016) The scenario model intercomparison project (ScenarioMIP) for CMIP6. Geoscientific Model Development 9(9), 3461-3482.

O’neill, B., Tebaldi, C., Van Vuuren, D., Eyring, V., Friedlingstein, P., Hurtt, G., Knutti, R., Kriegler, E., Lamarque, J. and Lowe, J. (2016) The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6, Geosci. Model Dev., 9, 3461–3482.

Osisanwo, F., Akinsola, J., Awodele, O., Hinmikaiye, J., Olakanmi, O. and Akinjobi, J. (2017) Supervised machine learning algorithms: classification and comparison. International Journal of Computer Trends and Technology (IJCTT) 48(3), 128-138.

Palerme, C., Genthon, C., Claud, C., Kay, J.E., Wood, N.B. and L'Ecuyer, T. (2017) Evaluation of current and projected Antarctic precipitation in CMIP5 models. Climate dynamics 48, 225-239.

Parker, W.S. (2013) Ensemble modeling, uncertainty and robust predictions. Wiley interdisciplinary reviews: Climate change 4(3), 213-223.

Pascale, S., Kapnick, S.B., Delworth, T.L. and Cooke, W.F. (2020) Increasing risk of another Cape Town "Day Zero" drought in the 21st century. Proceedings of the National Academy of Sciences 117(47), 29495-29503.

Pielke Sr, R., Beven, K., Brasseur, G., Calvert, J., Chahine, M., Dickerson, R.R., Entekhabi, D., Foufoula-Georgiou, E., Gupta, H. and Gupta, V. (2009) Climate change: The need to consider human forcings besides greenhouse gases. Eos, Transactions American Geophysical Union 90(45), 413-413.

Pitman, A. (2003) The evolution of, and revolution in, land surface schemes designed for climate models. International Journal of Climatology: A Journal of the Royal Meteorological Society 23(5), 479-510.

Posada-Marín, J.A., Rendón, A.M., Salazar, J.F., Mejía, J.F. and Villegas, J.C. (2019) WRF downscaling improves ERA-Interim representation of precipitation around a tropical Andean valley during El Niño: implications for GCM-scale simulation of precipitation over complex terrain. Climate dynamics 52, 3609-3629.

Prat, O. and Nelson, B. (2015) Evaluation of precipitation estimates over CONUS derived from satellite, radar, and rain gauge data sets at daily to annual scales (2002–2012). Hydrology and Earth System Sciences 19(4), 2037-2056.

Raftery, A.E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005) Using Bayesian model averaging to calibrate forecast ensembles. Monthly weather review 133(5), 1155-1174.

Raftery, A.E., Madigan, D. and Hoeting, J.A. (1997) Bayesian model averaging for linear regression models. Journal of the American Statistical Association 92(437), 179-191.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N. and Prabhat, f. (2019) Deep learning and process understanding for data-driven Earth system science. Nature 566(7743), 195-204.

Rowlands, D.J., Frame, D.J., Ackerley, D., Aina, T., Booth, B.B., Christensen, C., Collins, M., Faull, N., Forest, C.E. and Grandey, B.S. (2012) Broad range of 2050 warming from an observationally constrained large climate model ensemble. Nature Geoscience 5(4), 256-260.

Sachindra, D., Ahmed, K., Rashid, M.M., Shahid, S. and Perera, B. (2018) Statistical downscaling of precipitation using machine learning techniques. Atmospheric research 212, 240-258.

Sanderson, B.M., Knutti, R. and Caldwell, P. (2015) A representative democracy to reduce interdependency in a multimodel ensemble. Journal of Climate 28(13), 5171-5194.

Schepen, A., Zhao, T., Wang, Q.J. and Robertson, D.E. (2018) A Bayesian modelling method for post-processing daily sub-seasonal to seasonal rainfall forecasts from global climate models and evaluation for 12 Australian catchments. Hydrology and Earth System Sciences 22(2), 1615-1628.

Shetty, S., Umesh, P. and Shetty, A. (2023) The effectiveness of machine learning-based multi-model ensemble predictions of CMIP6 in Western Ghats of India. International Journal of Climatology 43(11), 5029-5054.

Sloughter, J.M.L., Raftery, A.E., Gneiting, T. and Fraley, C. (2007) Probabilistic quantitative precipitation forecasting using Bayesian model averaging. Monthly weather review 135(9), 3209-3220.

Song, Y.H., Chung, E.-S. and Shiru, M.S. (2020) Uncertainty analysis of monthly precipitation in GCMs using multiple bias correction methods under different RCPs. Sustainability 12(18), 7508.

Tanveer, M.E., Lee, M.-H. and Bae, D.-H. (2016) Uncertainty and reliability analysis of CMIP5 climate projections in South Korea using REA method. Procedia engineering 154, 650-655.

Taylor, K.E., Stouffer, R.J. and Meehl, G.A. (2012) An overview of CMIP5 and the experiment design. Bulletin of the American meteorological Society 93(4), 485-498.

Tegegne, G., Kim, Y.O. and Lee, J.K. (2019) Spatiotemporal reliability ensemble averaging of multimodel simulations. Geophysical Research Letters 46(21), 12321-12330.

Tian, L. and Quiring, S.M. (2019) Spatial and temporal patterns of drought in Oklahoma (1901–2014). International Journal of Climatology 39(7), 3365-3378.

Tokarska, K.B., Stolpe, M.B., Sippel, S., Fischer, E.M., Smith, C.J., Lehner, F. and Knutti, R. (2020) Past warming trend constrains future warming in CMIP6 models. Science advances 6(12), eaaz9549.

Trenberth, K.E. (2011) Changes in precipitation with climate change. Climate research 47(1-2), 123-138.

Vrugt, J. (2016) MODELAVG: A MATLAB toolbox for postprocessing of model ensembles. Department of Civil and Environmental Engineering, University of California Irvine 4130.

Wang, B., Zheng, L., Liu, D.L., Ji, F., Clark, A. and Yu, Q. (2018) Using multi-model ensembles of CMIP5 global climate models to reproduce observed monthly rainfall and temperature with machine learning methods in Australia. International Journal of Climatology 38(13), 4891-4902.

Wang, D., Liu, J., Luan, Q., Shao, W., Fu, X., Wang, H. and Gu, Y. (2023) Projection of future precipitation change using CMIP6 multimodel ensemble based on fusion of multiple machine learning algorithms: A case in Hanjiang River Basin, China. Meteorological Applications 30(5), e2144.

Wang, Q., Schepen, A. and Robertson, D.E. (2012) Merging seasonal rainfall forecasts from multiple statistical models through Bayesian model averaging. Journal of Climate 25(16), 5524-5537.

Wang, X., Yang, T., Li, X., Shi, P. and Zhou, X. (2017) Spatio-temporal changes of precipitation and temperature over the Pearl River basin based on CMIP5 multi-model ensemble. Stochastic Environmental Research and Risk Assessment 31, 1077-1089.

Woldemeskel, F., Sharma, A., Sivakumar, B. and Mehrotra, R. (2012) An error estimation method for precipitation and temperature projections for future climates. Journal of Geophysical Research: Atmospheres 117(D22).

Woldemeskel, F., Sharma, A., Sivakumar, B. and Mehrotra, R. (2014) A framework to quantify GCM uncertainties for use in impact assessment studies. Journal of Hydrology 519, 1453-1465.

Wu, J., Shi, Y. and Xu, Y. (2020) Evaluation and projection of surface wind speed over China based on CMIP6 GCMs. Journal of Geophysical Research: Atmospheres 125(22), e2020JD033611.

Xiang, B., Harris, L., Delworth, T.L., Wang, B., Chen, G., Chen, J.-H., Clark, S.K., Cooke, W.F., Gao, K. and Huff, J.J. (2021) S2S prediction in GFDL SPEAR: MJO diversity and teleconnections. Bulletin of the American meteorological Society, 1-46.

Xu, R., Chen, N., Chen, Y. and Chen, Z. (2020) Downscaling and projection of multi-cmip5 precipitation using machine learning methods in the upper han river Basin. Advances in Meteorology 2020, 1-17.

Yan, Z., Zhou, Z., Liu, J., Han, Z., Gao, G. and Jiang, X. (2020) Ensemble projection of runoff in a large-scale basin: Modeling with a global BMA approach. Water Resources Research 56(7), e2019WR026134.

Yang, T., Hao, X., Shao, Q., Xu, C.-Y., Zhao, C., Chen, X. and Wang, W. (2012) Multi-model ensemble projections in temperature and precipitation extremes of the Tibetan Plateau in the 21st century. Global and Planetary Change 80, 1-13.

Yang, T., Zhang, L., Kim, T., Hong, Y., Zhang, D. and Peng, Q. (2021) A large-scale comparison of Artificial Intelligence and Data Mining (AI&DM) techniques in simulating reservoir releases over the Upper Colorado Region. Journal of Hydrology 602, 126723.

Yumnam, K., Guntu, R.K., Rathinasamy, M. and Agarwal, A. (2022) Quantile-based Bayesian Model Averaging approach towards merging of precipitation products. Journal of Hydrology 604, 127206.

Zelinka, M.D., Myers, T.A., McCoy, D.T., Po-Chedley, S., Caldwell, P.M., Ceppi, P., Klein, S.A. and Taylor, K.E. (2020) Causes of higher climate sensitivity in CMIP6 models. Geophysical Research Letters 47(1), e2019GL085782.

Zhang, L. and Cooke, W. (2021) Simulated changes of the Southern Ocean air-sea heat flux feedback in a warmer climate. Climate dynamics 56(1-2), 1-16.

Zhang, L., Kim, T., Yang, T., Hong, Y. and Zhu, Q. (2021) Evaluation of Subseasonal-to-Seasonal (S2S) precipitation forecast from the North American Multi-Model ensemble phase II (NMME-2) over the contiguous US. Journal of Hydrology 603, 127058.

Zhang, X. and Yan, X. (2018) Criteria to evaluate the validity of multi-model ensemble methods. International Journal of Climatology 38(8), 3432-3438.

Zhao, M., Golaz, J.-C., Held, I.M., Ramaswamy, V., Lin, S.-J., Ming, Y., Ginoux, P., Wyman, B., Donner, L. and Paynter, D. (2016) Uncertainty in model climate sensitivity traced to representations of cumulus precipitation microphysics. Journal of Climate 29(2), 543-560.