INDIAN AGRICULTURE IN A CHANGING CLIMATE: A STATISTICAL ANALYSIS

by

Balsher Singh Sidhu

M.A.Sc., The University of Toronto, 2016

B.Tech., Indian Institute of Technology Delhi, 2013

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Resources, Environment and Sustainability)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

November 2021

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Indian agriculture in a changing climate: a statistical analysis

submitted by Balsher Singh Sidhu in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

in Resources, Environment and Sustainability

Examining Committee:

Navin Ramankutty, Professor, Institute for Resources, Environment and Sustainability, and School of Public Policy and Global Affairs, UBC Vancouver

Co-supervisor

Milind Kandlikar, Professor, Institute for Resources, Environment and Sustainability, and School of Public Policy and Global Affairs, UBC Vancouver

Co-supervisor

Mark Johnson, Professor, Institute for Resources, Environment and Sustainability, and Department of Earth, Ocean and Atmospheric Sciences, UBC Vancouver

Supervisory Committee Member

M. V. Ramana, Professor, School of Public Policy and Global Affairs, UBC Vancouver

University Examiner

Sean Smukler, Associate Professor, Faculty of Land and Food Systems, UBC Vancouver

University Examiner

Abstract

Predicting crop yield response to climate change is a topic of active research. A popular method involves building statistical models using historical climate and agricultural data, and then applying them on future climate projections for predicting crop yields. Using India as a case study, this dissertation examines these statistical models along two dimensions: the type of climate variables included, and the statistical techniques used. We also employ these models for predicting climate change impact on Indian crop yields till 2100.

First, we examine the role of seasonal (e.g. total seasonal precipitation) versus subseasonal (e.g. precipitation over each crop growing stage) climate variables in explaining crop yields. We observe that even though adding extra climate variables does not always improve overall model accuracy, the proportion of yield variability explained by climate (versus non-climatic variables like geography and time) can increase significantly. This underscores the importance of combining physiological and statistical knowledge while choosing climate variables for statistical crop models. Second, we compare the well-known statistical method of OLS linear regression (LR) to a popular machine learning method called boosted regression trees (BRTs). While LR models were simpler to interpret, BRTs could uncover unexpected non-linear relationships and exhibited better yield prediction accuracy. Compared to LR, BRTs sometimes showed lower sensitivity to temperature variation. Higher flexibility of BRTs allowed them to identify obscure interactions between variables that could be missed by LR.

We then use different climate variables and statistical techniques for building statistical models to predict climate change impact on India's future crop yields. We found that nationally-averaged rice, wheat, and pearl millet yields could reduce by up to 3.4, 4.3, and 5.5 percent (respectively) by 2050 under the intermediate emissions scenario. Some parts of India may benefit from climate change, while other regions could face yield losses of up to 20 percent. Depending on the climate variables or statistical technique employed, we observe high variability in yield change predictions. We therefore suggest combining multiple models for estimating climate change impact on crop yields.

Lay Summary

Agriculture is often counted among the sectors most vulnerable to climate change. To predict the impact of climate change on future crop yields, scientists frequently use statistical models built using historical climate and agricultural production data. We examined these models along two dimensions: the climate variables included in them, and the statistical techniques used. We find that the utility of adding potentially important climate variables may be obscured if the models are ranked using standard statistical metrics; incorporating physiological knowledge during model selection is recommended. Moreover, advanced techniques like machine learning can offer certain advantages over the more commonly used linear regression methods. Finally, we applied these models to India which predicted that nationally-averaged rice, wheat, and pearl millet yields could reduce by up to 3.4, 4.3, and 5.5 percent (respectively) by 2050 under the "middle of the road" climate change scenario.

Preface

I am the primary responsible person for all chapters in this dissertation. The overall objectives and structure of the dissertation were co-designed with my supervisory committee members Dr. Navin Ramankutty, Dr. Milind Kandlikar, and Dr. Mark Johnson. My contributions include developing the methodology, collecting and analyzing data, summarizing results, and preparing the manuscripts.

This dissertation contains three stand-alone chapters (Chapters 2-4) that were written as independent articles. Since each chapter was prepared with the intention of publication in a different peer-reviewed journal, there may be some repetition in the background and introduction sections to maintain the flow of narrative in each.

Versions of chapters 2 and 3 will be submitted for publication with Dr. Navin Ramankutty, Dr. Milind Kandlikar, and Dr. Zia Mehrabi as co-authors. A version of chapter 4 will be submitted for publication with Dr. Navin Ramankutty, Dr. Milind Kandlikar, and Dr. Mark Johnson as co-authors.

This dissertation required no ethics approval because no primary data was collected.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AgMIP	Agricultural Model Intercomparison and Improvement Project
AIC	Akaike Information Criterion
AR5	IPCC Fifth Assessment Report
AR6	IPCC Sixth Assessment Report
BIC	Bayesian Information Criterion
BRT	Boosted Regression Trees
CMIP5	Coupled Model Intercomparison Project Phase 5
CMIP6	Coupled Model Intercomparison Project Phase 6
DTR	Diurnal Temperature Range
DV	Dependent Variable
ET	Evapotranspiration
ET_c	Crop Specific Evapotranspiration
ET_o	Reference Evapotranspiration
FAO	Food and Agriculture Organization
GCM	General Circulation Model
GDD	Growing Degree Days
GR	Green Revolution
ha	Hectare
HPC	High Performance Computing
ICRISAT	International Crop Research Institute for The Semi-Arid Tropics

IPCC	Intergovernmental Panel on Climate Change
IV	Independent Variable
LR	Ordinary Least Squares Linear Regression
ML	Machine Learning
PDP	Partial Dependence Plot
PMKSY	Pradhan Mantri Krishi Sinchayee Yojana
R^2	Coefficient of Variation
RCPs	Representative Concentration Pathway
RMSE	Root Mean Square Error
SSP	Shared Socio-Economic Pathway
WG II	Working Group II

Acknowledgements

Ah, the acknowledgements section, where I am expected to convey my gratitude to everyone that made this dissertation possible. Let me admit outright that it is impossible for me to thank everyone that was a part of my journey, but the best attempt I shall make.

There is a reason why almost all graduate students, in a rather predictable and unoriginal manner, begin the acknowledgements section by thanking their supervisor(s). Comparing my master's and doctoral experience, I have realized beyond doubt that a student's learning experience, skill development, career progress and, most crucially, overall well-being are highly dependent on the support they receive from their supervisor(s). Unfortunately, this is something most graduate students have little control over when they join a research group. I would probably make many peers jealous when I say that I hit a double jackpot with Navin Ramankutty and Milind Kandlikar. Eminent researchers with illustrious careers that they are, they require no introduction here, which shaves off quite a lot of text from this section and gives me more space to truly express my gratitude. Starting from day one, they exceeded all expectations a graduate student may have from their supervisors: granting freedom to chart my own research path while providing invaluable and timely guidance when I was stuck, giving opportunities to expand my skill set beyond the usual graduate school curriculum, or providing honest advice that enabled me to separate the wheat from the chaff and spend time doing stuff that actually matters. Besides that, I am equally grateful for their unwavering belief in my capabilities and work even on the darkest of days when I doubted myself. Equally important was their constructive feedback that has undoubtedly improved my research capabilities and led to this dissertation. Having them

both as supervisors was a truly amazing experience to say the least. Navin's relentless belief in going back to first principles, combined with Milind's love of torturing and questioning all assumptions that others accept as self-evident truth, admittedly made for some gruelling and tiresome coding sessions but led to highly relevant learning experiences and projects.

Graduate school also brings its fair share of mental fatigue and stress. Adding to that the pandemic chaos makes for an even more potent mixture. I too struggled with some professional and personal problems during my years at UBC, and Navin and Milind's support during that time deserves a special mention. In them, I saw not just professional coaches, but also wise mentors with whom I could share my struggles when overwhelmed. They both claim they were just following their assigned duties, but my conversations with colleagues from other groups and departments is a testament to how fortunate I am. Special mention for the extremely funny Twitter banter I have had with them over the years. By putting this all on paper, it is possible that I am digging a hole for my future self if I end up in academia! (My future students, if you are reading this, too bad.)

I am also immensely grateful to my committee member professor Mark Johnson. His invaluable advice when I approached him with technical questions, and the sheer patience in his replies to my incessant volley of emails during the final few weeks of this dissertation preparation are truly appreciated. The submission of this dissertation during the busiest months of the academic year was only possible because of my committee's patience and support.

My research experience was complemented by a wide range of teaching opportunities as the instructor for my institute's summer programs 2018 and 2019, a teaching assistant across multiple faculties, and a climate expert for the UBC Climate Teaching Connector. Special thanks to IRES and UBC for the same. Professors Dowlatabadi, Boyd, Ramankutty, and Kandlikar deserve a special mention for assisting me with planning and delivering the summer program course that added immensely to my pedagogical experience.

I'd also like to acknowledge the wonderful group of people I was fortunate to work alongside in my lab and at IRES. Special thanks to Verena, Vinny, Kalifi, Matt and Larissa for helping me settle down and making my transition into the group as smooth as it could ever be; Ginni and Juan for being perfect peers to learn statistics with; Zia, Christian and Vinny for patiently introducing me to the research methods that this dissertation heavily relies on; Erika, Jumi and Kushank for prodding me to graduate soon so they can take over my desk; Julie for diligently navigating me through the administrative maze; Susanna, Dana, Laura and Bejoy for demonstrating the relevance of examining social aspects of my otherwise mostly technical research; Char, Pedro and Marie-Theres for livening up the office with their stories and laughter; and Sameer for patiently bearing with my slow progress on a wonderful project we got a chance to collaborate on. Working late nights at the Liu institute (sometimes as the lone occupant of the building) also gave me a chance to find a wonderful friend in the ever-smiling Moses; it was nice to know you, my friend. My thanks also go to Gillian, Bonny, Stefanie, Linda and Kelsey for making my experience at IRES so wonderful and easy. I am of course missing a lot of other folks here, to whom I once again convey my sincerest gratitude. This journey would never have been as fun and fulfilling as it was, if it were not for them.

My biggest thanks go to my family. My brother Harsher, who I was lucky enough to be housemates with for most of my doctoral study years. As an immigrant living thousands of kilometers away from the homeland, I had the rare privilege of having a sibling to share my good and bad days with. My parents provided unparalleled advice and encouragement from afar. Their uncanny ability to sense when I was feeling sad and cheer me up over the phone, will forever be a mystery to me. I drew a lot of inspiration from their Facetime calls, which I am also guilty of ignoring when chasing deadlines. And finally, my wife Ardeep. She came to Canada and we started living together just as I was wrapping up my doctoral research, and her contribution during these crucial few months cannot be overstated. She patiently bore my tantrums which became increasingly frequent as my dissertation deadline neared. Her constant reassurances and sagely advice helped me cross the finish line.

I would like to thank the Vanier Canada Graduate Scholarship and the University of British Columbia Four Year Doctoral Fellowship for their financial support. This research was enabled in part by support provided by WestGrid (www.westgrid.ca) and Compute Canada (www.computecanada.ca).

To the farmers of India.

Chapter 1: Introduction

1.1 Overview

Once again, the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC) has underscored the reality of climate change; it has also been proven beyond any doubt that anthropogenic activities and emissions are responsible (IPCC, 2021). Numerous studies have shown the sensitivity of agricultural productivity to both short-term and long-term variability in weather and climate (to cite a few: Davis, Chhatre, Rao, Singh, & Defries, 2019; Lobell, Schlenker, & Costa-Roberts, 2011). Consequently, agriculture occupies a prominent spot in the list of sectors most vulnerable to climate change. Any reduction in agricultural yields not only affects food and commodity production adversely (thereby increasing chances of inflation, food insecurity, malnutrition, and famines), but also puts at risk hundreds of millions of people globally who directly or indirectly depend on agriculture for their livelihood.

Therefore, the importance of understanding agriculture's response to climate and other environmental changes cannot be overstated. Over the past decades, predicting crop yields as a function of climate change has been a topic of extensive research, both at a regional and global scale. The results paint a worrying picture. To cite a few out of numerous examples, Lobell & Field (2007) reported a negative impact of rising global temperature on wheat, maize and barley yields, and pegged the losses at $5 billion per year. Model simulations have also shown a negative impact on maize yields of 8-12 percent globally under 2 degrees Celsius warming scenario (Bassu et al., 2014). Wheat production studies in India have estimated a reduction of 5.9 percent per degree Celsius rise in temperature (R. Gupta, Somanathan, & Dey, 2017). Similarly,

Asseng et al. (2017) analyzed multiple climate change scenarios using mean growing season temperature and reported a decrease of 0.34 tonnes/hectare in India's wheat yield per degree Celsius temperature rise.

Climate change does not affect agriculture only through shifts in mean temperature and precipitation; it is also altering short-term precipitation patterns and the intensity, duration, and frequency of extreme weather events like heat waves and droughts (IPCC, 2021). Past research has shown that such subseasonal (or intra-seasonal) climate variability too can have significant impacts on crop yields (Riha, Wilks, & Simoens, 1996). For instance, a season with drought during the early growing stage followed by heavy rains during harvest may cumulatively be considered a "normal rainfall season", but erratic rainfall can cause yield losses or even crop failure. In recent years, this limitation of seasonal climate-crop analyses has been addressed by some researchers using subseasonal climate data. An agricultural productivity study for Tanzania predicted 4.2, 7.2, and 7.6 percent yield reduction for maize, sorghum, and rice for a 20 percent increase in intra-seasonal precipitation variability (measured in terms of coefficient of variance of monthly precipitation measurements) (Rowhani, Lobell, Linderman, & Ramankutty, 2011). Schlenker & Roberts (2009) conducted an intra-seasonal analysis in the US and reported that temperatures above thresholds of 29^{o}, 30^{o}, and 32^{o} Celsius have adverse and nonlinear impacts on corn, soybeans, and cotton yields. Analysis with weekly soil moisture data from a land surface model for rainfed regions of the US found a significant impact of subseasonal water stress on multiple crops (Ortiz-Bobea, Wang, Carrillo, & Ault, 2019). Thus, there is ample evidence that intra-seasonal climate variations can have strong impacts on crop yields (Yu &

Goh, 2019), and furthering our understanding of the climate-crop relationship at subseasonal level is of utmost importance.

One of the most commonly used methods for predicting future crop yields under various climate change scenarios is to build statistical models with existing/historical data and apply the results to climate data projected into the future. In these models, crop yield is the dependent variable (DV), or the variable of interest, and relevant climate variables like temperature, water availability, or precipitation variability act as independent variables (IVs). The models aim to elicit the relationship between the DV and the IVs, and then use that relationship for future yield predictions. The primary focus of this dissertation is to build, analyze and contrast different types of these models, using rice, wheat, and pearl millet production in India as a case study.

In the next section, I start by describing the different types of models and their advantages and disadvantages. I next provide a brief description of Indian agriculture, primarily from a climate change perspective. Concurrently I identify the three major research gaps that this dissertation aims to fill. The last section in this chapter outlines the structure of this dissertation.

1.2 Typology of climate-crop models

Depending on the experimental design and type of data used, studies involving models for analyzing the impact of climate variability and change on crop yields can be divided into two groups. Many researchers use process-based models which incorporate experimentally-determined plant responses to various factors (temperature, water availability, soil moisture, radiation, carbon dioxide concentration among others) and build empirical mathematical

relationships between them (Roberts, Braun, Sinclair, Lobell, & Schlenker, 2017). The advantage of these process-based models is that the relationships and equations used for building the models are based on biophysics and plant physiology and are backed by clear mechanisms linking weather and crop growth. However, such models are usually built using limited results from lab-controlled experiments and do not necessarily reflect real-life outcomes in farmers' fields because they do not include factors external to the experimental setting like farmer behavior or pest infestation.

The shortcomings of the process-based models can be overcome by building statistical models using observational data collected from real-life agricultural activity. These statistical models find relationships between historical climate and agricultural data; these data can be cross-sectional (multiple measurements across space at a single moment in time), time series (measurements over time from a single geographical unit (district, county, or state)), or a panel (time series data from multiple geographic entities). The strength of statistical models is that they can implicitly incorporate variables not in the control of the observer, like variation in farmer behavior and management. Also, observational statistical models can take full advantage of the large amount of data available, collected over many years across many regions of the world (Roberts et al., 2017). Statistical models are also somewhat easier to build than process-based models and take far less computational resources[1].

[1] Statistical models also have some disadvantages; they may derive mathematical relationships between the observed climate and agricultural data that are not necessarily grounded in crop physiology (Roberts et al., 2017). A possible solution exists in the form of hybrid models. These depend on domain knowledge to identify important climatological metrics for a crop, and then use observational climate and crop data to identify important relationships. A good example of this can be observational models that use soil moisture, based on plant water

In statistical models based on observational data, a variety of different climate variables can be used as input. These include the most intuitive and widely used variables such as average temperature and total precipitation, to variables geared towards identifying and incorporating more complex determinants of yield like number of precipitation days (Fishman, 2016), duration of longest dry-wet spell (Tebaldi, Hayhoe, Arblaster, & Meehl, 2006), growing degree days (Albers, Gornott, & Hüttel, 2017), vapor pressure deficit (Jiang et al., 2021), among others. In addition to climate, statistical models also need data on non-climatic factors that may impact crop yields, such as soil characteristics, irrigation, agrochemicals (fertilizers or pesticides), mechanization, or cultivar varieties. In the absence of some of this data, modelers often include a geographic factor (e.g., lowest geographical entity like district, county, state, or country, as applicable) as a dummy variable to account for spatially-variable but time-invariant drivers of crop yield like soil characteristics, and temporal variables (e.g., year of planting or harvest) to account for factors with temporal trends such as advancement in farming practices, technology adoption, introduction of improved cultivars and others (Lobell & Burke, 2009). These two categories of variables (climatic and non-climatic) together act as input in the statistical models.

The use of time and geography dummies may present a complication for studies using standard statistical metrics like prediction accuracy to estimate the utility of including a specific climatic variable in statistical models. This is because time and geography components in the model

---

demand. In chapter 4, we examine this at greater length by building a crop-specific soil moisture model and combining it with an observational model.

might act as partial proxies for finer scale climatic variables (e.g. subseasonal weather) and mask their impact. **Chapter 2** delves into this issue in greater detail, and suggests methods to uncover potential overlaps between the yield variability explained by climatic and non-climatic portions of statistical models.

Within statistical models using observational data, there are a wide variety of techniques that can be used for building these models. Given the increasing availability of high-quality data and advanced computational facilities, this choice continues to expand, and researchers now have access to a variety of techniques to choose from. Among the most well-known and commonly used is ordinary least squares linear regression (LR hereafter), which minimizes the sum of squares in the differences between observed and predicted values. This popular method has been used by numerous studies (Butler & Huybers, 2013; Davis et al., 2019; Lobell & Field, 2007). At the other end of the spectrum are supervised machine learning methods that do not need a priori specification of the functional forms like LR. Understanding the advantages and disadvantages of these statistical tools, and methods to identify the most appropriate technique, is of utmost importance. In **chapter 3**, I contrast LR approaches to boosted regression trees, a popular machine learning algorithm, using identical training data. This comparison is conducted with the ultimate goal of furthering our understanding of the link between climate variability and crop yields.

1.3 Indian agriculture

One of the worst food disasters of the twentieth century occurred in 1943 in Bengal, a province in British-ruled India. An estimated 3 million people died in the Bengal Famine (Patnaik, 2017)

When India gained independence in 1947, food security was thus on the top of the new government's agenda. Efforts at achieving food self-sufficiency materialized into the Green Revolution (GR) in the 1960s. While expansion of land under cultivation had been ongoing since 1947 and continued to be encouraged during GR, the most striking and successful features of this program related to increasing land productivity through double-cropping and higher crop yields. High yielding crop varieties were fiercely promoted by policymakers and bureaucrats, along with better and more agrochemicals (fertilizers and pesticides) and improved irrigation facilities (Frankel, 2015), which rapidly increased yields of wheat and rice, the two crops that were the major focus of the Revolution. The results have been extremely promising: from 1961 to 2019, national wheat and rice yields increased by 320 and 160 percent (FAOSTAT, 2021), which played a crucial role in transforming the country into a net exporter of food crops. Subsidized inputs and government-assured procurement prices are two prominent tools from the proverbial Green Revolution toolbox that have today made rice-wheat rotation the most prevalent cropping pattern in India (Scott & Sharma, 2009).

The importance of Indian agriculture goes well beyond just producing sufficient food for its own citizens. It also occupies a critical international spot in terms of its contribution to the global food stocks. For example, India produces 70 percent of the world's chickpeas (16 times more than the second largest producer, Turkey), and over one-third of the world's millets and a quarter of the world's rice are grown in India (FAOSTAT, 2021). Global food supply chains are heavily dependent on crop yields in India. During the 2007-2008 food crisis India's ban on rice exports out of concern for domestic food security caused a surge in global rice prices (Menelly, 2016; Reuters, 2008). Besides, many African and Asian countries depend on imports of Indian rice,

closely tying their food security to the success of Indian agriculture. Beyond just the direct value of its output, agriculture in India is also the biggest employment provider: over 42 percent of the national workforce is currently employed in this sector (World Bank, 2021b). Agriculture in India thus plays a prominent role in not only ensuring food security (both within the country and globally) but also by providing livelihood to millions of households.

Indian agriculture is particularly vulnerable to climate change due to its heavy dependence on monsoon rains for water (Sharma, Rao, Vittal, Ramakrishna, & Amarasinghe, 2010). This makes it extremely sensitive to dry spells and short-duration rains which continue to get increasingly frequent with the changing climate (Annamalai, Hafner, Sooraj, & Pillai, 2013; V. Gupta, Singh, & Jain, 2020). This vulnerability is amplified in the rainfed regions which account for approximately half of the net sown area in India (Ministry of Agriculture and Farmers Welfare, 2018). Working group II's (WG II) contribution to the Fifth Assessment Report of the IPCC ranked Indian agriculture among those most vulnerable to climate change (IPCC, 2014). Given the certainty and rapid pace of climate change, it is of utmost importance to build robust crop models for India to not only identify crops and regions most at risk across the country, but also estimate the losses under different possible climate change scenarios. In **chapter 4**, I apply the findings of chapters 2 and 3 to use a range of statistical crop models (varying in both the underlying statistical technique as well as the climatic variables included) to predict future crop yields as a function of climate.

1.4 Dissertation chapters

The overarching objective of this dissertation is to examine and further our understanding of the mechanisms and processes that statistical crop models are based on, and then apply them for predicting climate change impact on crop yields in India. The dissertation chapters are organized as follows:

1.4.1 Chapter 2

Research questions

What is the role of geography and time, as proxies of unobserved non-climate variables, in explaining crop yields across multiple models with different sets of climate variables? What are the implications of comparing and selecting models based solely on generic statistical metrics, in light of the role of non-climatic factors (geography and time) in these models?

Data

Crop production (tonne) and harvested area (ha) data, disaggregated by crop, year, and district, was acquired from the International Crop Research Institute for the Semi-Arid Tropics (ICRISAT) Village Dynamics Studies in South Asia (ICRISAT, 2015). This data is reported for 311 districts from 1966-2011 using 1966 district boundaries as base. District-level daily minimum temperature, daily maximum temperature, and daily precipitation data were acquired from Indian Meteorological Department (Rajeevan, Bhate, Kale, & Lal, 2006). The temperature data (1961-2015) covered 634 districts using current boundaries, while the precipitation data (1961-2015) had 651 districts. I harmonized the climate data to ICRISAT district boundaries by

apportioning data for new districts created after 1966 to their parent districts using area-weighted averaging.

Methodology

I used ordinary least squares (OLS) linear regression to construct multiple statistical models with crop yield as the DV, district dummy and year of harvest as non-climatic IVs, and various combinations of weather parameters (temperature, precipitation, precipitation days, growing degree days among others) as the climatic IVs. I then compared those models using standard statistical metrics like overall variance explained (R^2) and root mean square error (RMSE), in addition to separating the total variance explained into climatic and non-climatic portions using a metric called "relative importance" (Grömping, 2006). The multiple models with different sets of climatic variables were analyzed and contrasted in detail using production data of three major crops from India (rice, wheat, and pearl millet) as a case study.

1.4.2 Chapter 3

Research questions

What are the advantages and disadvantages of using OLS linear regression (LR, hereafter) versus more advanced machine learning methods like boosted regression trees (BRTs) for predicting the relationship between climate variability and crop yields?

Data

I used the same data as chapter 2 described above.

Methodology

I built statistical models using LR and BRTs algorithms. These models were compared in terms of model accuracy (using RMSE as a metric). I analyzed the role various climate variables play in each model using partial dependence analysis, and examined the fundamental methods and assumptions followed by these models using synthetic data. The models were then contrasted in terms of their prediction of historical climate change impacts on India's crop yields for rice, wheat, and pearl millet.

1.4.3 Chapter 4

Research questions

What is the predicted impact of climate change on crop yields in India? How do the estimates vary across various climate change scenarios and statistical model construction methods?

Data

I used the same historical climate and crop data as chapter 2 described above. For the soil moisture model, I relied on crop evapotranspiration coefficients from FAO (Allen, Pereira, Raes, & Smith, 1998). Future climate projections were acquired from Coupled Model Intercomparison Project Phase 6 (CMIP6), the latest framework of climate model experiments. I analyze four shared socioeconomic pathways (SSP1-2.6, SSP2-4.5, SSP3-7.0, and SSP5-8.5) as defined in the Sixth Assessment Report of the IPCC (IPCC, 2021) to cover the range of possible future outcomes.

Methodology

I used the techniques examined in chapter 3 to build my models for predicting yields as a function of climate. The soil moisture model was built using methods detailed in Hargreaves & Allen, (2003), Ramankutty, Foley, Norman, & Mcsweeney (2002), and Saxton & Rawls (2006). The water availability index as calculated from the soil moisture model was input as an IV in the crop statistical models for examining the role of soil moisture variability in determining future crop yields in India.

1.4.4 Chapter 5: Conclusion

In the concluding chapter, I bring together my findings from chapters 2, 3, and 4. I reiterate the impact that choices made while building statistical models (like the statistical techniques, or climate variables to include) can have on predicted crop yields. I also summarize the future of Indian agriculture in a changing climate, and discuss the utility of crop models for stakeholders trying to make crop yields more resilient to the predicted impacts of climate change.

Chapter 2: On the relative importance of climatic and non-climatic factors in crop yield models

2.1 Introduction

Climate is among the major drivers of agricultural output, and predicting crop yields as a function of climate has been a topic of active research for many decades. A common method to predict yields involves building statistical crop models from historical weather and yield data collected over time and/or space. These models can then be used for multiple purposes: estimating the sensitivity of crops to climate variability (Lobell & Field, 2007; Ortiz-Bobea, Wang, Carrillo, & Ault, 2019; Ray, Gerber, Macdonald, & West, 2015; Zachariah, Mondal, Das, Achutarao, & Ghosh, 2020), predicting crop yields under different future climate change scenarios (Birthal, Khan, Negi, & Agarwal, 2014; Ray et al., 2015), assessing benefits of crop switching (Rising & Devineni, 2020), identifying regions where agricultural interventions like irrigation can help mitigate climate change impacts (Zaveri & Lobell, 2019), among others.

A variety of different weather and climate drivers of crop yields have been reported in statistical crop modeling literature. These include the most intuitive and widely used variables like average temperature and total precipitation over the crop growing season, to variables geared towards identifying and incorporating more complex determinants of yield like number of precipitation days (Fishman, 2016), duration of longest dry-wet spell (Tebaldi, Hayhoe, Arblaster, & Meehl, 2006), growing degree days (Albers, Gornott, & Hüttel, 2017), heat or killing degree days (Butler & Huybers, 2013), vapor pressure deficit (Jiang et al., 2021), among others. There are

also studies which indirectly estimate the impact of water availability by including variables related to soil moisture content over different crop growth stages (Ortiz-Bobea et al., 2019).

In addition to weather and climate, statistical models also need data on non-climatic factors that may impact crop yields, such as soil characteristics, irrigation, use of agrochemicals (fertilizers or pesticides), mechanization and technology uptake, cultivar varieties, and others. Since not all data are usually available, this task is often accomplished indirectly by including a geographic factor (e.g., lowest geographical entity like district, county, state, country, as appropriate) as a dummy variable to account for spatially-variable but time-invariant drivers of crop yield like soil characteristics, with temporal variables (often the year of planting or harvest) employed to account for factors with temporal trends such as advancement in farming practices, technology adoption, mechanization, or introduction of improved cultivars (Lobell & Burke, 2009). These two categories of independent variables, climatic and non-climatic, together act as input in statistical crop models, which then model the dependent variable (crop yield in this case) as a function of these independent variables[2].

Crop models are then built, with the data discussed above, using a wide variety of statistical techniques, each with their own strengths and weaknesses. These vary from the most popular Ordinary Least Squares (OLS) linear regression (henceforth, linear regression), to advanced machine learning techniques that can extract more resolved climate-yield relationships that linear

[2] We use the designation "non-climatic" for district dummies and time fixed effect, but this is not strictly true because climatic conditions vary across space, hence district dummies may also carry some climatic signal. This aspect of statistical models is examined in greater detail in this chapter.

regression models may not detect (Beillouin, Schauberger, Bastos, Ciais, & Makowski, 2020). Depending on the statistical method used, a model's accuracy is often measured using statistical parameters like coefficient of variation (R^2), adjusted R^2, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), or root mean square error (RMSE), among others. More details about all these metrics can be found in James, Witten, Hastie, & Tibshirani (2013). Researchers also use these metrics to build multiple models with different climate variables, compare and contrast them, rank them, or identify the most appropriate and accurate one from the mix. The model thus selected may then be employed for accomplishing tasks discussed previously.

This study focuses on the interplay between the two concepts discussed above: the role of non-climatic variables in explaining crop yield, and the use of standard statistical metrics to estimate and compare crop model accuracy. While including geography and time accomplishes the objective of accounting for non-climatic crop yield determinants, comparing models solely on the basis of overall variance explained (R^2) or overall prediction accuracy (RMSE) has a limitation: these metrics only refer to overall model accuracy, but do not quantify the individual contributions of climatic and non-climatic variables in the model. Consequently, when using these metrics to compare two distinct crop models, there is no specific information about the climatic or non-climatic source of any potential differences in model performance. This complicates studies trying to estimate the utility of including specific climatic variables in their statistical models, because a portion of the anticipated improvement in model performance with the inclusion of a new climate variable to an existing model may be subsumed within the non-climatic (geography and time) component. These geography and time variables may therefore

complicate the use of generic model performance metrics like R^2, adjusted R^2, or RMSE for model comparison and selection. This is the primary hypothesis examined in this study. While we focus only on OLS linear regression, the hypothesis warrants examination for other statistical techniques as well.

We start by analyzing the cumulative contributions of climatic and non-climatic (geography and time) factors to a crop model's total predictive worth. We then parse and compare the relative importance of these two groups of variables across an array of models. For this, we use a statistical metric called "relative importance" that has hitherto not been used widely in the crop modeling field (Grömping, 2006). The implications of our findings are then discussed in relation to model utility for predicting the impact of anomalous weather events and long-term climate change on crop yields. Specifically, our study attempts to answer the following two questions:

1. What is the role of geography and time, as proxies of unobserved non-climate variables, in explaining crop yields across multiple models with different sets of climate variables?
2. What are the implications of comparing and selecting models based solely on generic statistical metrics, in light of the role non-climatic factors (geography and time) may play in the models?

2.2 Data and methods

2.2.1 Crop production data

This study was designed as a detailed analysis of statistical crop models using India as a case study. We used data from India because of prior familiarity with this region and easy availability

of long-term crop production and climate data. Although our results are specific to India, we believe they are more generally applicable to studies analyzing statistical relationships between crop yields and climate in other parts of the world.

We focused on India's three major cereal crops for this study: rice and pearl millet grown during the summer monsoon (kharif) season, and wheat, which is primarily grown during the winter (rabi) season. Crop production (tonne) and harvested area (ha) data, disaggregated by crop, year, and district, was acquired from the International Crop Research Institute for the Semi-Arid Tropics (ICRISAT) Village Dynamics Studies in South Asia (ICRISAT, 2015). This data is reported for 311 districts from 1966-2011 using 1966 district boundaries as base. Crop calendar data for crop sowing and harvesting dates at state-level came from Government of India's Agricultural Statistics at a Glance 2016 (Ministry of Agriculture and Farmers Welfare, 2016). Any aggregation of the climate data from daily to seasonal scale was done after masking it for the growing season for each crop-state combination.

2.2.2 Climate data

District-level daily minimum temperature, daily maximum temperature, and daily precipitation data were acquired from Indian Meteorological Department (Rajeevan, Bhate, Kale, & Lal, 2006). The temperature data (1961-2015) covered 634 districts using current boundaries, while the precipitation data (1961-2015) had 651 districts. We extracted climate data for 1966-2011 and harmonized it to ICRISAT district boundaries by apportioning data for new districts created after 1966 to their parent districts using area-weighted averaging. With this daily temperature and precipitation data, we derived multiple climate variables for use in our models (Table 2.1).

Table 2.1 Description of variables included in the crop models.

Variable name	Description
T_avg_mean	Mean daily average temperature during the growing season
T_min_mean	Mean daily minimum temperature during the growing season
T_max_mean	Mean daily maximum temperature during the growing season
gdd_0, gdd_10, gdd_20, gdd_30	Degree day bins, at 10-degree Celsius intervals.
P_sum	Total seasonal precipitation
P_days	Total seasonal precipitation days (precipitation > 0.1 mm (May, 2004)
P_sum_subseasonal_1, P_sum_subseasonal_2, P_sum_subseasonal_3, P_sum_subseasonal_4	Subseasonal precipitation over the four crop growing stages, as defined by FAO (Allen, Pereira, Raes, & Smith, 1998)
P_days_subseasonal_1, P_days_subseasonal_2, P_days_subseasonal_3, P_days_subseasonal_4	Subseasonal precipitation days over the four crop growing stages, as defined by FAO (Allen et al., 1998)

2.2.2.1 Degree day bins

The concept of degree days is very common in crop modeling research; for a clear explanation of the concept with illustrative examples, readers are referred to Roberts, Braun, Sinclair, Lobell, & Schlenker (2017). In our study, instead of using single thresholds for growing or killing degree days, we adopted a more flexible approach and included multiple degree day bins, which the model could then parametrize independently. A continuous temperature function was calculated by fitting a sinusoidal curve to the daily minimum and maximum temperature (University of

California Agriculture & Natural Resources, 2016). Degree days were then calculated for five different bins at 10-degree Celsius intervals: less than 0, 0-10, 10-20, 20-30, greater than 30. The last bin can represent killing degree days as temperatures in this range can have detrimental effects on crop yield. For instance, the upper temperature threshold for wheat's anthesis stage has been reported to be 31 ^{o}C (Porter & Gawith, 1999). Degree days in the [T_{lower}, T_{upper}] bin were calculated using equation (2.1) where T is the continuous temperature function. A schematic for this calculation is shown in Appendix A section A.1.

$$\sum_{growing\ season} \{ \int_{time_{lower\_1}}^{time_{lower\_2}} (T - T_{lower}) dt \quad - \int_{time_{upper\_1}}^{time_{upper\_2}} (T - T_{upper}) dt \} \ , \tag{2.1}$$

where

$time_{lower\_1}$: time of day when temperature first crosses the lower threshold (while increasing),

$time_{lower\_2}$: time of day when temperature next crosses the lower threshold (while decreasing),

$time_{upper\_1}$: time of day when temperature first crosses the upper threshold (while increasing),

$time_{upper\_2}$: time of day when temperature next crosses the upper threshold (while decreasing).

2.2.3 Statistical techniques

All statistical analysis was conducted in R (R core team, 2020); R packages used include tidymodels (Wickham et al., 2019), data.table (Dowle & Srinivasan, 2021), relaimpo (Grömping, 2006), ggthemes (Arnold, 2021), RColorBrewer (Neuwirth, 2014), wesanderson (Ram & Wickham, 2018), gridExtra (Auguie, 2017), doParallel (Microsoft & Weston, 2020a), and foreach (Microsoft & Weston, 2020b).

2.2.3.1 Statistical models

For building and comparing multiple crop models with different sets of climate variables, we used OLS linear regression model specification of the form shown in equation (2.2).

$$y_{it} = \alpha_i + \beta(t) + \gamma_1(clim\_var_1) + \ldots + \gamma_n(clim\_var_n) + \varepsilon_{it} \ , \quad (2.2)$$

where y_{it} is crop yield in district i and year t; α_i is district specific intercept; β is parameter for time (harvest year) trend; γ_n is parameter for the n^{th} climate variable (clim_var) included in the model; ε_{it} is the standard error.

We constructed ten models with varying complexity of climate variables. After a null model with no climate variables (only district ID and year of harvest as variables), the other nine models ranged from a simple model with only mean seasonal temperature and total seasonal precipitation, to the most complex one with mean daily minimum temperature, mean daily maximum temperature, degree day bins, subseasonal precipitation amounts, and subseasonal precipitation days. For clarity, we have divided the climate variables into groups of temperature and precipitation variables, with three sub-groups or levels in each as shown in Table 2.2. The model names and the climate variables included in each are also presented in Table 2.2.

Table 2.2 Model names and climate variables included in the ten models.

Model name	Temperature		Precipitation	
	Sub-group level	Variables	Sub-group level	Variables
Null	0	--	0	--
Tavg_Psum	1	T_avg_mean	1	P_sum
Tavg_Psumday	1	T_avg_mean	2	P_sum, P_days
Tavg_ Psumday_ subseasonal	1	T_avg_mean	3	P_sum_subseasonal_1, P_sum_subseasonal_2, P_sum_subseasonal_3, P_sum_subseasonal_4, P_days_subseasonal_1, P_days_subseasonal_2, P_days_subseasonal_3, P_days_subseasonal_4
Tminmax_ Psum	2	T_min_mean, T_max_mean	1	P_sum
Tminmax_ Psumday	2	T_min_mean, T_max_mean	2	P_sum, P_days
Tminmax_ Psumday_ subseasonal	2	T_min_mean, T_max_mean	3	P_sum_subseasonal_1, P_sum_subseasonal_2, P_sum_subseasonal_3, P_sum_subseasonal_4, P_days_subseasonal_1, P_days_subseasonal_2, P_days_subseasonal_3, P_days_subseasonal_4
Tminmaxgdd_ Psum	3	T_min_mean, T_max_mean, gdd_0, gdd_10, gdd_20, gdd_30	1	P_sum

Model name	Temperature		Precipitation	
	Sub-group level	**Variables**	**Sub-group level**	**Variables**
Tminmaxgdd_ Psumday	3	T_min_mean, T_max_mean, gdd_0, gdd_10, gdd_20, gdd_30	2	P_sum, P_days
Tminmaxgdd_ Psumday_ subseasonal	3	T_min_mean, T_max_mean, gdd_0, gdd_10, gdd_20, gdd_30	3	P_sum_subseasonal_1, P_sum_subseasonal_2, P_sum_subseasonal_3, P_sum_subseasonal_4, P_days_subseasonal_1, P_days_subseasonal_2, P_days_subseasonal_3, P_days_subseasonal_4

Studies estimating impacts of weather and climate usually include irrigation in their analysis. We began by adopting percent area irrigated for each crop-year-district combination as a proxy for irrigation (data on actual water used was unavailable), with interactions between irrigation and each precipitation variable. However, in the "relative importance" section of our analysis (discussed later), irrigation would be deemed a non-climatic variable, but its interaction with climate would complicate the disaggregation of total variance explained into climatic and non-climatic portions. So, we left out irrigation in our analysis presented in the main paper from here on. Nonetheless, we present our results with irrigation included in Appendix A section A.2. The general trends between models' performance are consistent with our results without irrigation.

2.2.3.2 Model performance metrics

We used three popular statistical metrics for computing and comparing model accuracy. The first two were R^2 and adjusted R^2. Both these metrics vary from 0 to 1, and measure how well the model predictions match the actual observed data. A drawback of R^2 is that it always increases (or stays the same at a minimum) with the addition of any variable, regardless of whether that variable has any correlation with the variable of interest (James et al., 2013). Adjusted R^2 fixes this limitation by penalizing the R^2 statistic for the number of variables included in the model. Therefore, adding a variable with little explanatory power can decrease a model's adjusted R^2, unlike R^2 which increases monotonically. Nevertheless, we retained R^2 because of its fundamental relationship with the statistical metric called "relative importance" which we introduce and discuss in the next section.

The third statistic we used was root mean square error (RMSE), the square root of the mean of the squared differences between observed and predicted values. We conducted RMSE analysis using out-of-sample 10-fold cross-validation with random samples stratified over years, a technique commonly used in model comparison and selection studies (Ortiz-Bobea et al., 2019). Out-of-sample predictions prevent overfitting by keeping the model's training and testing datasets separate.

2.2.3.3 Relative importance of variables

To estimate the individual contribution of different variables in explaining the observed yield variance, we used the concept of "relative importance" (Grömping, 2006). This metric refers to the contribution of individual IVs to a multivariable linear regression model. Specifically,

relative importance denotes the portion of total variance explained (or R^2) by a model that can be attributed to a particular variable in the model. For linear regression with uncorrelated data, each IV's contribution is simply the increase in R^2 observed with the addition of that IV to a model with the remaining variables. This, however, is not true in most observational studies like ours where the various IVs are not necessarily independent and usually have some spatial and/or serial correlation.

The various climate variables described earlier are not only correlated with each other, but also with geography and time. Consequently, the increase in R^2 with the addition of a variable is dependent on the variables previously present in the model. To disaggregate total variance explained among the regressors, both climatic and non-climatic, we calculated the average increase in R^2 with the addition of a variable to all possible models with distinct permutations of the remaining variables (Grömping, 2006). For example, for a model with independent variables IV_1, IV_2, and IV_3, the relative importance of IV_2 equaled the average of the increase in R^2 when IV_2 is added to every possible model without IV_2. The resultant relative importances of all variables then add up to total R^2 of the model.

By adding the relative importances of geography and time, we quantified the proportion of variance explained attributable to non-climatic factors, and then added up the relative importances of all climate variables to compute the total variance explained by climate in each model formulation. The trends in non-climatic and climatic variables' relative importance with increasing model complexity were then analyzed to ascertain the overlap, if any, between variance explained by these variables.

2.2.4 Evaluating yield predictions during extreme weather events

In addition to comparing models based on their spatio-temporally averaged accuracy, we also wanted to analyze model performance in different parts of the country during anomalous periods of extreme weather events. For the time period of our study, 1966-2011, we calculated national mean annual temperature and total annual precipitation from area-weighted average of district-level climate data. The years with least total precipitation and highest mean temperature were designated as "drought year" and "hot year", respectively. Individual years with conditions closest to the median temperature and median total precipitation respectively were designated as "normal years" for benchmarking purposes. The performance of all models was then compared for the drought, hot and normal years, in terms of RMSE reduction for a particular year. This analysis was conducted for the whole country as an aggregate, as well as for each state separately.

2.2.5 Simulations of climate change impact

After comparing the models' accuracy and their flexibility to account for anomalous weather patterns, we conducted scenario analysis (as is commonly done in other studies investigating the impact climate change on crop yields (Lobell, Schlenker, & Costa-Roberts, 2011)) to estimate the impact that long-term climate change over the historical period of 1966-2011 has already had on India's crop yields. The daily minimum temperature, daily maximum temperature, and daily precipitation data was linearly detrended to remove time trend at district-scale. This detrended data was then assumed to denote the weather that would have occurred if climate change had not occurred. Using this detrended daily weather data, we used the exact same procedure as we did

with the actual weather data to construct all our climate variables of interest. To obtain district-level estimates of climate change impact on crop yields, we conducted residual bootstrapping (Li & Maddala, 1996) with 500 repetitions to predict crop yields with and without climate change, and then computed the median value and 95 percent confidence intervals of the difference between predictions from the two scenarios. For each crop-district-model combination, the average of the ten yield loss values in the last decade in the dataset (2002-2011) was then presented as the expected impact of climate change that has occurred since 1966, the starting year of this study's time period. While our calculation of the climate change impact on crop yields uses 1966 as the baseline year, anthropogenic climate change has been ongoing since long before that, and therefore our estimated impact of climate change with this simulation is conservative.

2.3 Results

2.3.1 Model performance evaluation using statistical metrics

The performance of the models was first analyzed in terms of adjusted R^2 and RMSE. Top row of Figure 2.1 shows adjusted R^2 (red), and increase in adjusted R^2 (blue) compared to the null model (with only geography and time, no climate variables). Bottom row depicts RMSE (red), and percent decrease in RMSE (blue) compared to the null model. Each of the six panels is divided into four sub-panels using dotted lines, depending on the number of temperature variables in the model. Each sub-panel depicts models that contain the same temperature variables, but three different levels of precipitation variables (simple to complex from left to right).

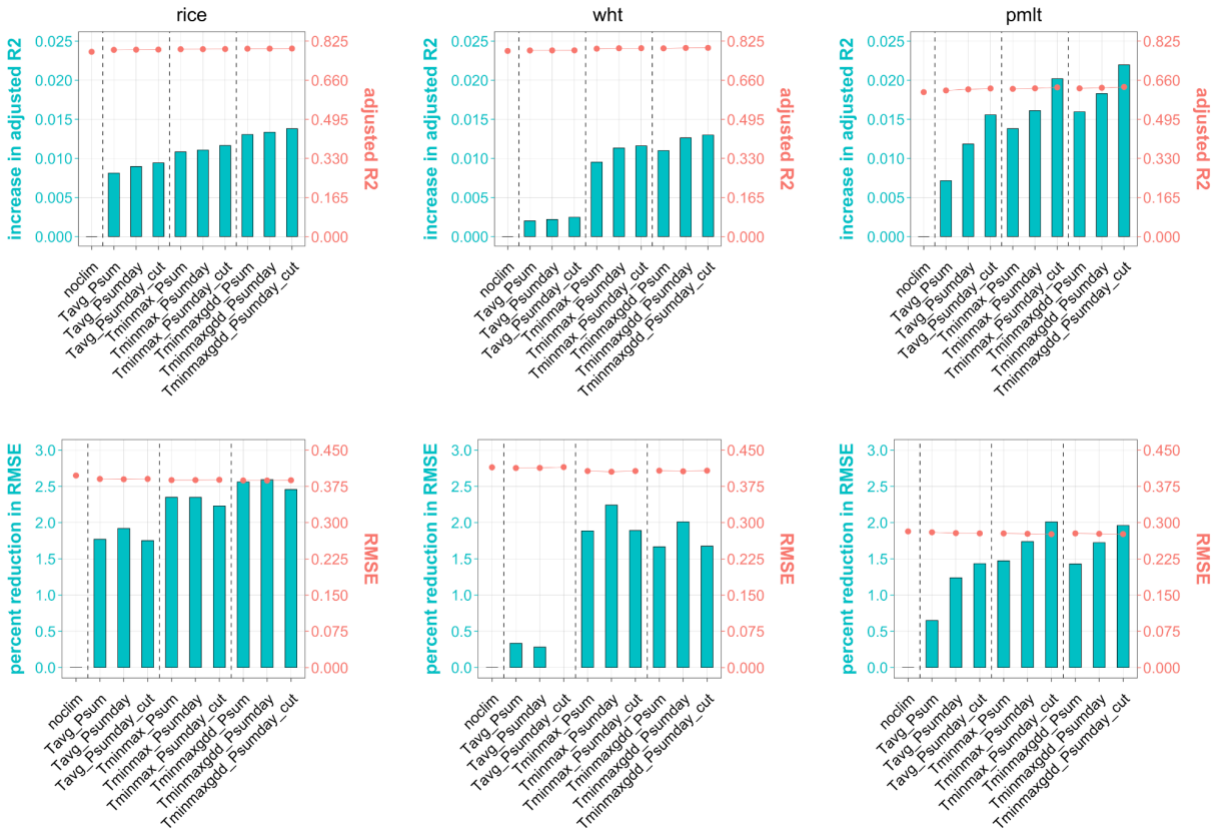

Figure 2.1 Model performance measured in terms of adjusted R^2 (top row; absolute values in red, increase compared to null model in blue) and RMSE (bottom row; absolute values in red, percent increase compared to null model in blue). The three crops are rice (left), wheat (center), and pearl millet (right). Within each panel, models include varying levels of climate data, with three levels each of temperature and precipitation (see Table 2.2 for description of levels). The models are divided into sub-panels with dotted lines and arranged in the following order: null model; temperature level 1 and precipitation levels 1, 2, 3; temperature level 2 and precipitation levels 1, 2, 3; and temperature level 3 and precipitation levels 1, 2, 3.

Adjusted R^2 depicts similar trends for all three crops: while it increases as more climate variables are added to a model, the increase is only marginal. For rice, the advantage of choosing the best performing model with the most climate variables, compared to the null model without any climate variables, is an increase in adjusted R^2 from 0.780 to 0.794. For wheat, the increase is

from 0.784 to 0.797, while pearl millet models outperform the null model by 0.022 units at the most (0.620 to 0.642). This apparently limited utility of climate in our crop models is re-affirmed in the RMSE plots which show that adding more climate variables may even decrease model accuracy, as is visible for both rice and wheat going from level 2 to 3 of precipitation in each sub-panel. In fact, selecting T_avg_Psumday_subseasonal model for wheat (bar 4 in bottom-centre panel in Figure 2.1) provides no benefit over the null model when compared on the basis of RMSE reduction. Pearl millet exhibits a more consistent pattern of improvement in RMSE reduction with more climate variables, even though that trend is broken between levels 2 and 3 of the temperature variables. To summarize Figure 2.1, adjusted R^2 and RMSE show that the accuracy and fit of all models for all three crops are not very different from the null model containing only geography and time as the variables of interest, and that a model's performance does not depend much on what climate variables are included in that model.

2.3.2 Relative importance of variables

We used the previously discussed metric of relative importance to apportion variance explained by a crop model to different explanatory variables included in the model. Unlike with the standard metrics, for all three crops analyzed, the relative importance of geography, and time to a smaller extent, reduces as more climatic variables are added to the models to account for subseasonal climate variability (Figure 2.2). Hence, even though the total variance explained, or R^2, may not increase by the same amount, the portion of the variance explained that can be attributed to climate is increasing disproportionately more compared to the change in model R^2. For all crops, and for all model progressions within each sub-panel, as more climate variables are added to account for precipitation availability (going from total seasonal precipitation to

subseasonal precipitation and precipitation days), the relative importance of climate goes up, while that of time and geography goes down. From a low value of 0.004, 0.078, and 0.005 in the simplest models (seasonal temperature and precipitation) for rice, wheat, and pearl millet, the relative importance of climate goes up to 0.184, 0.162, and 0.142 in the most complex models on the right. While the maximum increase in adjusted R^2 or RMSE over the null model is less than 0.02 units and 3 percent respectively (Figure 2.1), relative importance analysis shows that the contribution of climate can be more than 20 percent of the total variance explained by a model.

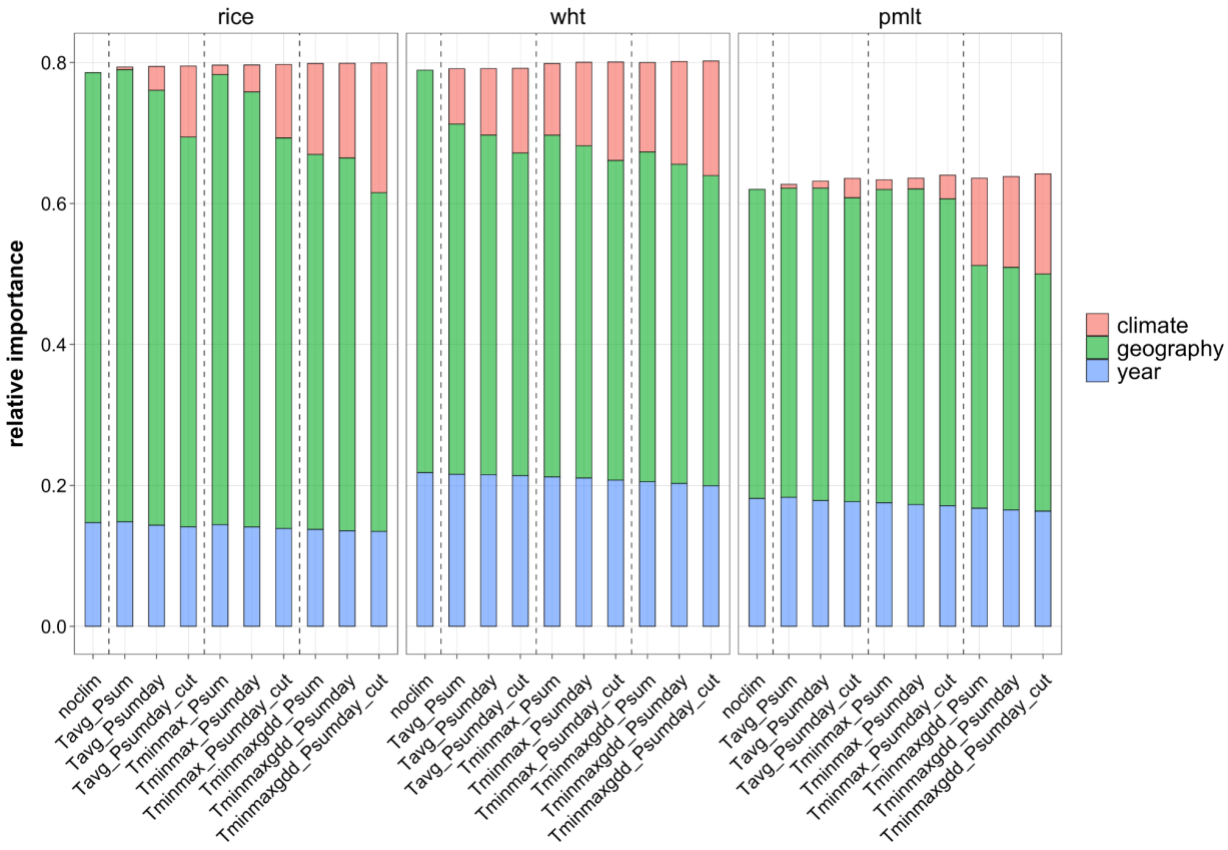

Figure 2.2 Relative importance of time (blue), geography (green), and climate (red) variables across the ten models analyzed for rice (left), wheat (center), and pearl millet (right). The plots follow the same arrangement as Figure 2.1 for direct comparison. Note that the sum of the relative importances of time, geography, and

climate variables equals R^2, which shows minimal improvement in overall model fit from the simplest null model on the left to the most complex model on the right, for each crop. Similar scheme of arranging models by temperature and precipitation variables' complexity is followed as Figure 2.1.

2.3.3 Model sensitivity to extreme weather events

In the timeframe of our study, the least amount of rainfall fell during 2002, which we designated as a "drought year"; 2009 because of its highest mean annual temperature was designated as a "hot year". Our method matches the results of Aadhar & Mishra (2021) who analyzed South Asian climate data from 1951-2016 and found that the worst drought during this period occurred in 2002, affecting more than 65 percent of the region. The years with median precipitation (1993) and median temperature (1996) constituted "normal years". Models' performance in these years was compared by calculating national RMSE of model predictions for each of these years from the 10-fold out-of-sample cross-validation results described previously (Figure 2.3). We also conducted this analysis at a more local-scale by calculating state-level RMSE for each model in a similar manner. Nationally aggregated RMSE reduction for all models and crops (compared to respective null models) is shown in Figure 2.3. Similar plots, but with RMSE aggregated at state-level, are available in Appendix A section A.3.

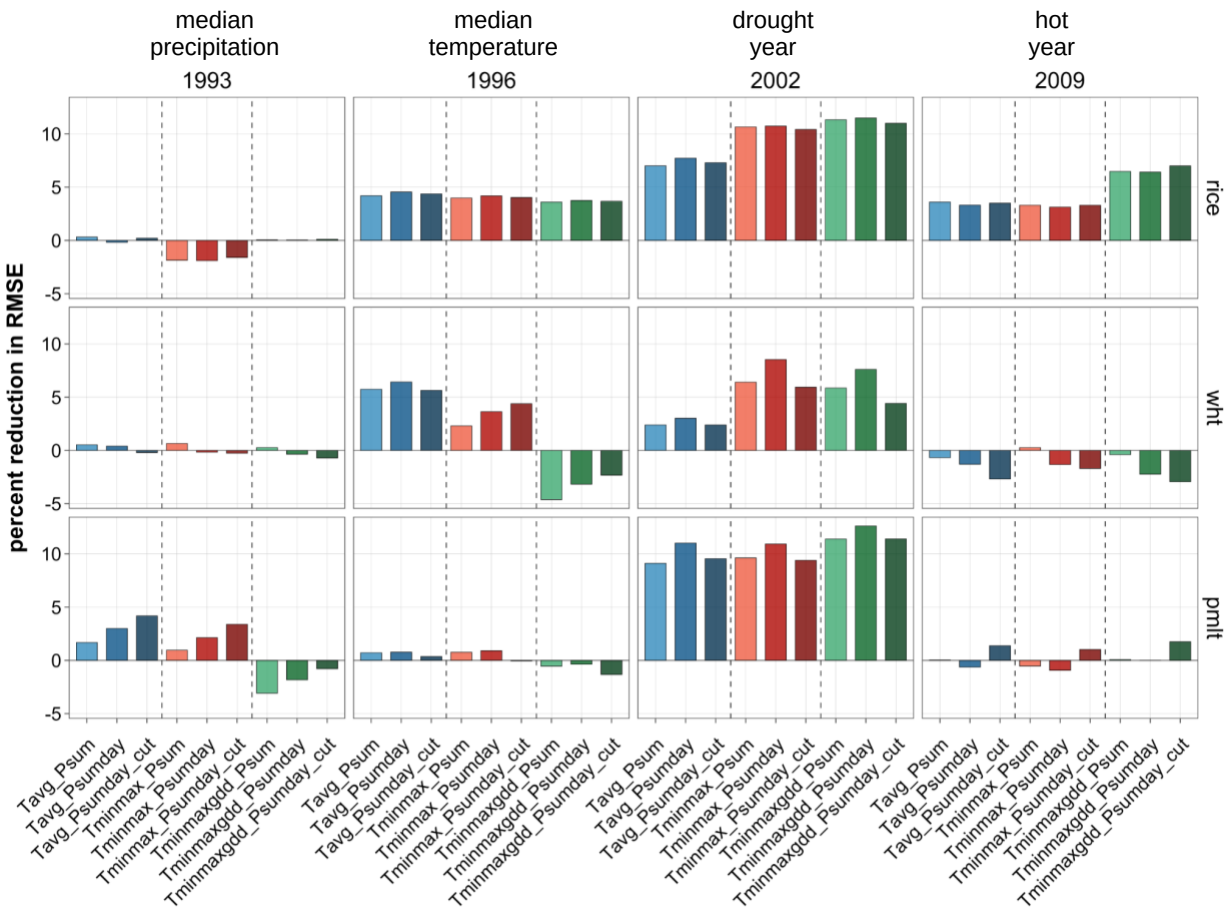

Figure 2.3 Improvement in model performance (in terms of RMSE reduction compared to the null model with no climate variables) for median precipitation (1993), median temperature (1996), drought (2002), and hot (2009) years.

Compared to the null model with only geography and time, the improvement in performance from the simplest model to the most complex models depends a lot on the year in question. In the drought year of 2002, all models exhibit an enhanced performance compared to the other years. There is a general trend of all models performing better in 2002 than the other years, irrespective of the levels of climate variables included in them. While the overall improvement in model performance when measured for the full time period hovers around 2-3 percent reduction in

RMSE compared to null model (Figure 2.1), Figure 2.3 shows that the more complex models exhibit performance improvement in excess of 10 percent during the drought year of 2002.

For rice, the more complex models have a markedly better performance than the simpler models for both the anomalous years (2002 and 2009), in contrast to the normal years where additional climate variables have little impact on model performance. This trend is also exhibited by wheat but only for 2002. Pearl millet shows a more subdued difference between simple and complex models in 2002. In contrast, the performance of the simpler and complex models is similar in 1993 and 1996, leading us to infer that the complex models are often better suited than the simpler models at accounting for anomalous weather patterns.

The difference in the performance of the models in anomalous years is more pronounced when model predictions are analyzed at state-level (Appendix A section A.3). There are some important crop-state combinations, like rice in Madhya Pradesh and Punjab, wheat in Gujarat, Haryana, Maharashtra and Punjab, and pearl millet in Karnataka and Rajasthan, where the RMSE reduction is highest for the more complex models (over 25 percent in some cases) during the drought year of 2002; simpler models are unable to match this accuracy. Figure 2.4 shows the difference between predicted and observed pearl millet yield in the state of Rajasthan, the biggest producer of this crop in India. The results from the simplest (with seasonal temperature and precipitation) and the best performing model are shown in red and blue, respectively. In the anomalous years of 2002 and 2009, the complex model performs better than the simple model (difference between predicted and observed is closer to zero), while there is no discernable difference in models' performance in 1993 and 1996.

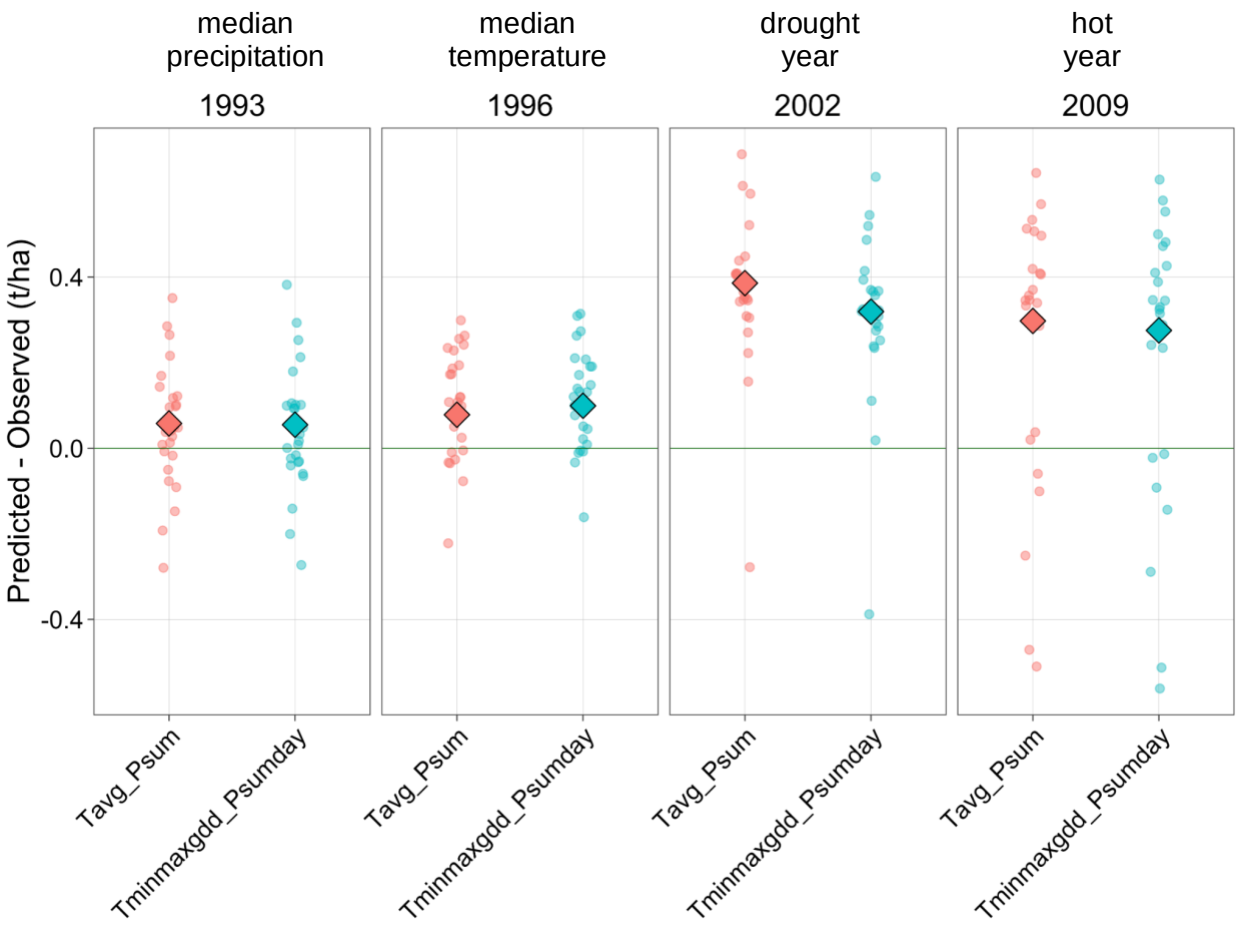

Figure 2.4 Difference between predicted and observed pearl millet yield for all districts of the state of Rajasthan for 1993, 1996, 2002, and 2009. The advantage of the most complex model (blue) over the simplest model (red) is most pronounced in 2002 and to a lesser extent in 2009. Plot shows district-level (round) and average state-level values (district values weighted by crop harvested area; diamond).

While there are instances where the simpler models outperform the complex models in the drought or hot years, or when the complex models outperform the simpler models during normal years too, the trend is biased towards complex models having higher utility than simpler models in 2002 and 2009. For quantitative evidence of this trend, we scored and ranked our models according to the level of climate complexity (Table 2.2), with scores of 1, 2, and 3 for each level

of complexity of temperature and precipitation variables. So, the simplest model (with only mean seasonal temperature and total seasonal precipitation) has a complexity score of 2, and the most complex model has a complexity score of 6. The scores for the best performing model for each crop, year, and state were averaged to get a national score for each crop-year combination. For rice, the average scores of both the normal years are 2.9, while the drought and hot years' scores average 4.2 and 3.2. In other words, more complex models performed better than the simpler models in years when the climate deviated from the normal, especially the drought year of 2002. The trend was visible across the other two crops too, and these scores for wheat and pearl millet were 2.3, 3.1, 4.0, 2.4 and 3.2, 2.6, 4.1, 3.2 respectively (Figure 2.5).

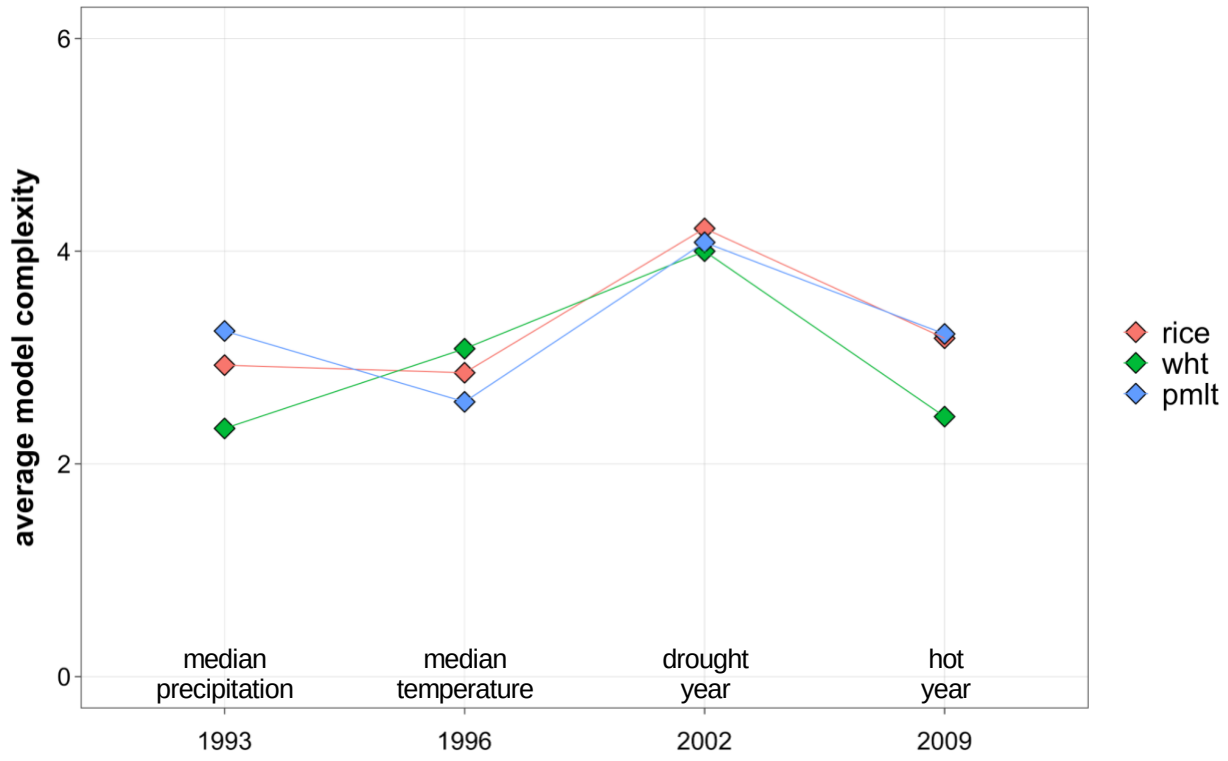

Figure 2.5 Nationally averaged score of the best performing model for each crop-state combination for 1993, 1996, 2002, and 2009. The score (metric of complexity level of climate variables) is markedly higher in the drought year of 2002.

2.3.4 Simulations of climate change impact

In terms of our simulated impact of climate change, nationally averaged yield change results estimate yield losses from all nine models (Figure 2.6). In contrast, more variation is observed for wheat, where the most complex models predict a net gain in nationally averaged yield for the crop. Our estimates with mean seasonal temperatures show that national pearl millet yield has witnessed a reduction due to climate change, although there is no significant change observed from the predictions of models with more granular temperature variables. One common pattern among all three crops, especially rice and pearl millet, is that the estimated impact of mean climate change on crop yields is more dependent on the complexity of temperature variables included as opposed to the level of precipitation variables added.

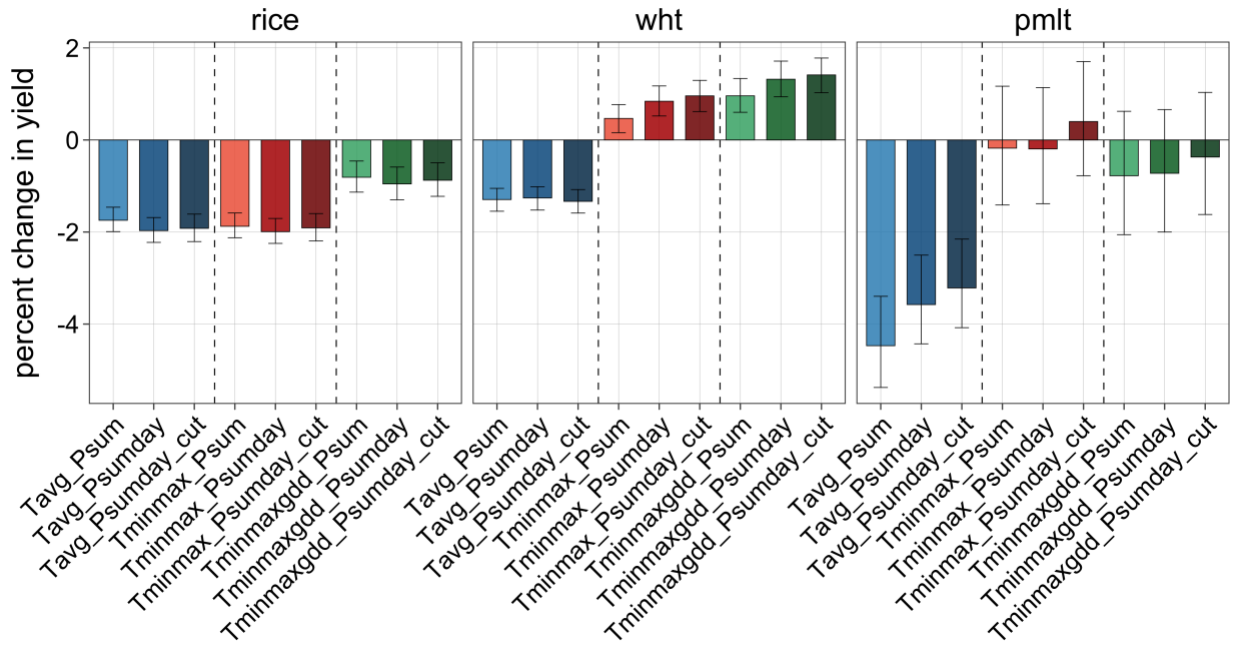

Figure 2.6 Nationally averaged yield change due to historical climate change.

Figure 2.7 shows the simulated impact of climate change on rice yield during the last decade of this study's time period (2002-2011); similar plots for wheat and pearl millet are available in Appendix A section A.4. The null model is not shown because it is climate invariant and predicts zero impact of climate change. The panels from top to bottom depict an increasing number of variables to account for temperature variability, and panels from left to right denote models with increasing levels of precipitation variables. For all three crops, there are significant differences between the predictions by the nine different models we analyzed.

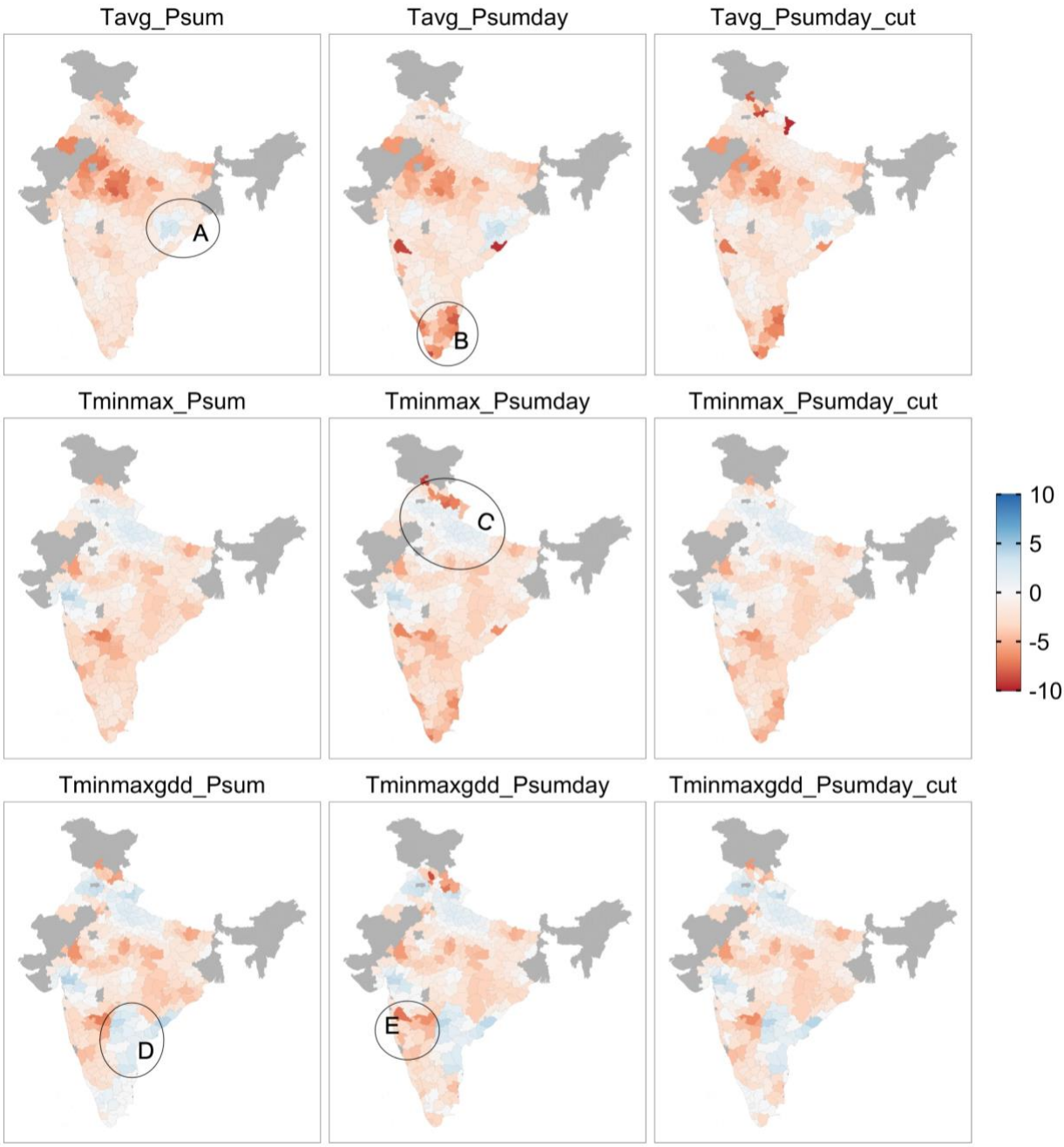

Figure 2.7 Simulated impact of long-term climate change (since 1966) on rice yield (percent change) in the last decade (2002-2011) of the study time period. The climate data was linearly detrended to remove time trend at district-scale. District-level estimates of median value and 95 percent confidence intervals of climate change impact on yield were obtained through residual bootstrapping (n = 500). The average district-level yield loss during the last decade in the dataset (2002-2011) is presented here as the expected impact of climate

change that has occurred since 1966. Only results with 95 percent significance of the confidence intervals are shown; insignificant results are shown in gray.

Rice plots show a negative impact for most of the country for the simpler models containing only mean seasonal temperature (top row), except a small patch in eastern India (region A) where rice yields are predicted to have benefited from climate change. As more precipitation variables are added, there is a region in south India covering the states of Kerala and Tamil Nadu (region B) where the predicted impacts of climate change become more drastic. From top to bottom, as more temperature variables are added to the models in the middle and bottom rows, a bigger range of predicted impacts is visible: compared to the first row, there are larger regions where climate change is predicted to have benefited rice yields. These include the highly mechanized Indo-Gangetic belt comprising the states of Punjab, Haryana, and Uttar Pradesh (region C). In the third row with the most complex models, even the state of Andhra Pradesh (region D), a big rice producer, turns blue from red in the previous panels. Most parts of the country seem to show more drastic impacts of climate change with the simpler models, with the exception of some districts in south-western India (region E) where the more complex models predict a more drastic impact of climate change on rice yields, compared to simpler models in the top row.

For wheat (Figure A.6), while simpler models predict a more consistent impact throughout the country, the more complex models in the middle and bottom rows show more variation; there are regions where climate change has positively impacted wheat yields, and these include major wheat producing states of Punjab, Haryana, Uttar Pradesh (region A). However, there are also districts in eastern and southern India where climate change seems to have had a more

detrimental impact than that predicted by the simpler models in the top row. In fact, certain districts in southern India show a reduction of up to 14 percent in wheat yields because of climate change. The patterns observed in pearl millet panels (Figure A.7) are similar to wheat. The simpler models in the top row predict a more consistent negative impact throughout the country with some blue patches in eastern parts of India (region A), the middle and bottom rows depict a higher contrast in the expected impacts of climate change on pearl millet yields. Huge parts in north and central India that were depicted in red in the top row now seem to show a net positive impact of climate change, while southern India has turned an even darker shade of red, denoting a more serious negative impact of climate change than one would observe if the analysis is limited to simpler models with only mean seasonal temperature.

2.4 Discussion

2.4.1 Role of a priori climate-crop relationship knowledge

Using statistical metrics of adjusted R^2 and RMSE, we observed that the model performance does not vary noticeably between models with various levels of climate variables. This marginal role of climate in improving model performance is consistent with the results of previous studies conducted on Indian agriculture. Fishman (2016) analyzed rice yields and reported an increase in adjusted R^2 from 0.735 for a null model (with no climate data) to 0.758-0.772 with different combinations of climate variables including precipitation, degree days, and rainy days as measures of climate variability. Davis, Chhatre, Rao, Singh, & Defries (2019) similarly reported a decline in their crop model explanatory power (measured using Akaike Information Criteria) with the addition of potential climate variables (number of monsoon dry days, ratio of

precipitation to number of monsoon rain days, and squared terms of temperature and precipitation), and did not include those variables in their final model.

A parsimonious model selection process based purely on examining R^2, RMSE or related measures as above would advocate for the simplest models as being most appropriate. However, relative importance analysis showed that as more climate variables were added, climate started occupying an increasingly important role in predicting crop yields, even if that was not reflected fully in the increase in total variance explained by a model. Crop yield signal that would otherwise be explained by subseasonal climate is subsumed by geography and time in the absence of those climate variables, a trend that gets amplified in periods of anomalous weather.

This result has important implications for the study of climate-crop relationships using statistical models. There are usually multiple climate variables that can be included in a model, and choosing the best model based on generic model performance metrics like R^2 or RMSE may lead to selection of models which downplay the role of climate. This is especially true if model selection were to happen without adequate domain knowledge about important variables that need to be included in a model irrespective of their role in increasing overall model performance. While our analysis was limited to OLS linear regression, this error of omission could easily occur in advanced machine learning based methods as well, where the variables are automatically selected by the algorithm.

Adequate importance needs to be given to fundamental plant physiological understanding of how weather and climate affect crop yields while building crop models even if those climate variables

may initially seem insignificant during model selection. For example, in our study, if models are selected based on statistical metrics like R^2, adjusted R^2, or RMSE, Occam's razor or the principle of parsimonious model selection would dictate that we choose the simplest model with only seasonal average temperature and total precipitation. However, this would ignore the increasingly significant role climate plays in the more complex models. An argument can thus be made for including key a priori variables which are theoretically expected to impact crop yields. For example, field experiments have shown that rice yield can decline by up to 10 percent for every degree Celsius rise in night temperature, but no significant impacts were observed for rising day temperatures (Peng et al., 2004). This is backed up physiologically by evidence of high night temperature adversely impacting movement of carbohydrates and nitrogen within the rice plant (A. Singh, Chaudhuri, & Roychoudhury, 2020). In this case, separately including mean daily minimum temperature and mean daily maximum temperature in statistical crop models makes more sense physiologically, than including just mean daily average temperature. In our study too, while the models accounting for both mean daily minimum and mean daily maximum did not have drastically different adjusted R^2 or RMSE compared to the model containing only mean daily average temperature, climate change simulations showed opposite results for some regions as discussed earlier.

2.4.2 Model performance for extreme weather events and long-term climate change

Even when extra climate variables may not noticeably improve the model performance measured in terms of R^2, adjusted R^2, or RMSE, we showed that in more complex models, climate plays an increasingly crucial role in explaining crop yield variance. Hence, while model performance averaged over time was not significantly impacted by the levels of climate variables included,

more complex models were better able to account for anomalous weather patterns. As discussed earlier, the complex models performed particularly well (in terms of RMSE reduction) for some important crop-state combinations like rice in Punjab, wheat in Haryana and Punjab, and pearl millet in Rajasthan. The importance of this improved performance is underscored by the fact that Punjab and Haryana are among the biggest producers of wheat and rice in India. In 2019-20, they together accounted for 50 percent wheat and 30 percent rice to the central food reserves, a crucial source of subsidized food for the economically weaker sections of society. Meanwhile Rajasthan is the largest pearl millet producer in India.

The accuracy of model predictions is especially critical when the inaccurate predictions are biased towards the positive side of observed yields (predictions are higher than observed values), something RMSE does not factor in since it is insensitive to the direction of change. Models prone to over-predicting crop yields may provide a false sense of security to policymakers when they use these models to predict season-end yields and formulate food policies during extreme weather events. For example, in 2002 in the state of Madhya Pradesh, the simplest model (levels 1 of both temperature and precipitation) over-predicted wheat yields by 15.5 percent, while the model with levels 3 and 2 of temperature and precipitation variables over-predicted by 11.9 percent. Similarly, for national pearl millet production, the difference between predictions from the simplest model and a 3/2 level model was 769 thousand tonnes, or 15.5 percent of the national production in 2002. This susceptibility to over-predict production during anomalous weather events strengthens the case for examining model performance more closely under different conditions instead of making a selection based on standard statistical criteria.

A theoretically grounded (rather than statistically selected) model can also allow researchers to detect patterns of long-term climate change impacts on crop yields that may otherwise not be visible in simpler models. In our study, for all three crops, the simulated impact of climate change on crop yields in the last decade of our dataset's timeline, from 2002-2011, depicts stark differences in the predictions of various models. Simpler models predict more uniform yield losses across the country from climate change, whereas the complex models predict more variegated patterns of both losses and gains depending on the geographic region being analyzed. In addition to these distinct patterns, the more complex wheat models predict yield losses of up to 14 percent in some parts of the country, as opposed to the simpler models in the top row of Figure A.6 where the predicted losses peak at 5 percent. This observation further underscores the importance of considering an ensemble of models for making future yield predictions instead of selecting one solely on the basis of statistical parameters. For example, when a certain constituent of a group of models predicts negative impacts of climate change but is not the most accurate based on standard statistical metrics, the predictions from that model should not be dismissed without adequate examination.

2.4.3 Implications

By breaking up yield variance explained by crop models into climatic and non-climatic components, our study shows the potential pitfalls of building and selecting crop models based only on generic statistical tests without paying adequate attention to physiological processes that may mandate the inclusion or exclusion of specific climate variables. Poor characterization of climate impacts may have negative consequences, which can be amplified during periods of anomalous weather patterns. This is especially significant given the sufficient evidence of

anthropogenic climate change making weather more unpredictable and increasing the frequency and intensity of extreme events in this region. Murari, Ghosh, Patwardhan, Daly, & Salvi (2015) and Rohini, Rajeevan, & Mukhopadhay (2019) independently analyzed Coupled Model Intercomparison Project Phase 5 (CMIP5) future climate projections and report an increase in intensity, duration, and frequency of heat waves across India in the near future; Das & Umamahesh (2021) found similar results using CMIP6 data. Simultaneously, there has already been a significant increase in frequency of dry spells and intensity of wet spells during the monsoon season (D. Singh, Tsiang, Rajaratnam, & Diffenbaugh, 2014), and future predictions estimate a further increase in the frequency and magnitude of hot and dry extreme events (Mishra, Thirumalai, Singh, & Aadhar, 2020).

This study is also important because geography and time are not the only variables that can subsume climate signal; it is possible that some other non-climatic variables which vary across time or space, for example chemical inputs, mechanization, development of roads, atmospheric carbon dioxide concentration and so on, may have correlation with climate leading to subsequent conflation of the non-climatic and climatic signals. This nuance needs to be paid attention to when including such variables in crop models. Our study presents an analytical framework that can be used in such scenarios.

2.4.4 Limitations and future work

There are some caveats and limitations in our study that warrant discussion. One, we only report results for three crops. We did this to focus the discussion on the mechanics of statistical models with three representative crops from the two major growing seasons. With their contrasting

results, these three crops serve as examples of different crop-dependent outcomes of our analysis. Nonetheless, the analysis can be easily extended to other crops. Two, as discussed and rationalized earlier, we excluded irrigation in our analysis, even though it is a big determinant of crop yields. So, the absolute values of R^2 and RMSE results have to be interpreted within the context of this limitation. Three, India is a large country, and national level studies like ours may ignore important trends and patterns that have been reported in more granular studies (Zachariah et al., 2020). This limitation applies to all studies conducted over a large but heterogeneous nation state. A case can therefore be made for building more local models, and assessing variable relative importances in those models.

Some salient questions arose from our study that warrant further research. We observed that in anomalous weather years, our complex models (with most climate variables) had significantly lower RMSE than the simpler models (Figure 2.3). Simultaneously, the overall RMSE analysis shows little difference in model performance over the full time period (Figure 2.1). It is worth investigating if the improvement in model performance is minimal (or zero) in normal years and just amplified during periods of anomalous weather, or if the simple model performs better than the more complex models in normal years and this trend flips in anomalous years. We saw evidence supporting both these possibilities: in 1996, simple rice models outperformed complex ones but all wheat models exhibited similar performance (Figure 2.3), while 2002 saw the complex models perform better than simple ones for both these crops. It may be worthwhile to look into hybrid models that are trained on two distinct datasets: the normal weather years, and anomalous weather years. The predictions from both these models may then be combined with pre-determined probabilities to arrive at more accurate predictions.

2.5 Conclusion

Researchers using crop yield models have a vast array of climate variables to choose from for inclusion in their models. Without adequate domain knowledge about plant physiology or critical climate factors, climate variables are sometimes chosen based solely on overall model performance using common statistical techniques like R^2, adjusted R^2, or RMSE. However, this study demonstrates that obfuscation of the signal between non-climatic and climatic variables may cause the performance thus measured to improve only marginally with the inclusion of new climate variables, even though those omitted climate variables may be explaining important climate-yield relationships. This was seen for the state of Rajasthan in our study, where the seasonal model failed to capture the impact of exceptionally dry or hot weather on pearl millet yield. In contrast the subseasonal model, even though its overall accuracy was similar to the subseasonal model, performed significantly better at capturing yield losses in those anomalous years.

Automatic model selection based on parsimony criteria can seriously fail to parametrize important climate effects and lead to poor predictions of the impact of extreme weather events and long-term climate change. For example, our results showed that the assessment of historical impact of climate change, as measured by the model containing only seasonal variables, may not capture the more drastic impacts predicted at a subnational level by the more complex subseasonal models, as was seen in the case of wheat or pearl millet. Researchers are advised to use statistical metrics in combination with theoretical or process-based knowledge for choosing variables to include in their crop models.

Chapter 3: Statistical versus machine learning methods for estimating the impact of climate variability on Indian crop yields

3.1 Introduction

Statistical models are a popular method for studying the influence of climate variability and climate change on crop yields. These models establish relationships between observed historical weather and crop yield data, and can be used for impact assessment of short-term and long-term climate variability on crop yields. The data used in these models often varies across both time and space, although time series analysis of a single geographical unit and longitudinal analysis of a single time frame over a wider region are also common. The choice of statistical technique(s) is primarily dictated by factors including but not limited to research questions, data and computational resources availability, or statistical expertise. Given the increasing availability of high-quality data and advanced computational facilities, this choice continues to expand, and researchers now have a whole arsenal of techniques and tools to choose from. Consequently, understanding the advantages and disadvantages of these statistical tools, and factors that can assist in identifying the most appropriate technique, is of utmost importance. Our study adds to this field of research by assessing and contrasting two popular statistical techniques for crop yield analysis.

Among the most common techniques in the field of statistical crop yield modeling is ordinary least squares linear regression (LR hereafter) which minimizes the sum of squares in the differences between observed and predicted values. It models the dependent variable (DV) using

linear predictor functions for various terms comprised of relevant independent variables(s) (IV). This popular method has been used by numerous studies modeling crop yields as a function of climate variability (Butler & Huybers, 2013; Davis, Chhatre, Rao, Singh, & Defries, 2019; Lobell & Field, 2007). LR models, due to their relative simplicity and explicitly-specified relationships between variables, are easy to interpret for understanding the role of each IV in explaining the DV[3]. The mathematical equations are easily understood, and the model fitting process is faster and often less computationally-intensive compared to more advanced algorithms (James, Witten, Hastie, & Tibshirani, 2013). Because they have been used so extensively, conditions which influence their performance are also well understood by the greater scientific community. All these reasons place LR among the oldest and most popular algorithms for crop yield analysis.

Conversely, LR has certain shortcomings which require appropriate consideration. One, LR models are prone to oversimplifying relationships that may not always be linear, which penalizes model accuracy and can result in erroneous predictions. This drawback can be especially problematic in crop yield models, where climate and crop yield may not always have a linear relationship. For instance, past research has shown non-linear influences on crop yields of rising seasonal temperature through a disproportionate increase in excess heat days (Schlenker & Roberts, 2009). To account for potential non-linear relationships between crop yield and various climate variables, researchers include polynomial forms of said climate variables, which can

[3] There are certain conditions that need to be met for this to be valid, the most important being that the IVs should not have a strong correlation, because interpretation of a specific IV's regression coefficient assumes all other IVs are held constant. This assumption may be invalid if two or more correlated IVs vary together.

improve model performance compared to a model with only monomials of the climate variables (Fishman, 2016). This technique, however, still relies on the modeler making a priori assumptions about the functional form of each climate variable: while some variables like temperature or rainfall may have physiological justification for including polynomials of higher order, it may be hard to find empirical evidence for all climate variables. Some studies resolve potential non-linearity in crop response to climate by including variables such as growing degree days or moisture stress defined specifically to account for such phenomena (Butler & Huybers, 2013). This approach does require continuous, daily or at least some level of subseasonal weather data, which may not always be available.

A second shortcoming is that LR in its most basic form assumes the same relationship between an IV and DV throughout the data range, which may not always hold true. Segmented LR, which allows independent and piecewise linear relationships between the DV and IV across the data range, is more suitable in such cases. Schlenker & Roberts (2009) used a modified version of this segmented (or piecewise linear) model and reported critical temperature thresholds where the impact of temperature increase on crop yields switched from positive to negative. Segmented LR may, however, need manual specification of the expected number and/or locations of the breakpoint(s) when automatic breakpoint selection fails; this again relies on the modeler making a priori assumptions about the true functional form. Three, LR can be highly sensitive to outliers in data. With crop models built on observational data, this can be a cause of concern because of errors during data collection and reporting. Modified forms of LR such as robust regression is advisable in such cases, although they do have higher computational requirements (Faraway, 2015). Four, climate variables seldom affect crop yield in isolation of each other. LR allows the

specification of interaction between variables of interest, but it has to be specified individually, which leaves room for missing important interactions.

Many of the above-discussed limitations of LR can potentially be addressed using more flexible algorithms that do not need a priori specification of the functional forms and start with fewer assumptions about the relationship between DV and IVs. Machine learning (ML) offers many options to accomplish this, and it is a topic of immediate interest in this field. For instance, Vogel et al. (2019) used "random forests", a popular ML technique, for analyzing the impact of extreme climate on crop yields at a global scale. They found that temperature-related extremes were the biggest determinant of crop yield anomalies.

Here we would like to clarify some terminology used in our study. There is an active debate on the distinction between traditional statistical methods and ML, which has spawned peer-reviewed literature (Breiman, 2001; Bzdok, Altman, & Krzywinski, 2018), blogs, jokes by prominent statisticians, lengthy discussions on popular question and answer websites like stackexchange.com, and even an xkcd comic. The separation between so-called traditional statistical methods and ML is blurry and subjective. Purely for clarity in discussion, this study classifies LR as a standard statistical method, and uses boosted regression trees (BRTs) as an illustration of advanced ML techniques.

BRTs are a form of tree-based regression procedure that create partitions in the predictor space using nested if-else conditions, with the ultimate goal of attaining highest possible prediction accuracy. Tree-based methods are a popular constituent of the ML toolbox and have multiple

advantages: they can handle complex non-linear relationships, require no user input about expected DV-IV relationships, and easily accommodate missing data. Because single-tree models are prone to overfitting and exhibit poor predictive performance (Elith, Leathwick, & Hastie, 2008), BRTs combine multiple regression trees using a popular ensemble method called boosting, wherein a number of weak trees are sequentially trained to improve the performance of the full model with the goal of improving predictive power without overfitting. Compared to LR, the BRT algorithm makes no prior assumptions about the model's functional form, and predictions from tree-based models are comparatively immune to outliers and correlated IVs (James, Witten, Hastie, & Tibshirani, 2013). It can also detect non-linearity and important interactions between IVs without them being explicitly specified by the user, and can handle sharp discontinuities in DV-IV relationships.

On the flipside, like most ML techniques, BRTs are less interpretable, since the output contains no explicit coefficients for each IV that LR models provide. Nonetheless, BRTs can be used to rank IVs in order of their relative contribution to predicting the DV (Elith et al., 2008), making them somewhat interpretable. It would not be wrong to say that many researchers have used LR for years, and are more comfortable interpreting their results; ML methods like BRTs are comparatively novel and therefore not widely understood or used. As with all statistical methods, BRTs can be used for predictions and conditional inference from partial dependence plots (discussed in the next section). BRTs are computationally expensive and working with even medium-sized datasets as ours may need access to high performance computing facilities for tasks like fitting and comparing multiple BRTs, and bootstrapping these models. Nonetheless,

this has become less of an issue in recent years with research institutes and public agencies setting up dedicated high performance computing (HPC) facilities for advanced research.

The primary objective of this study is to compare these two popular techniques, LR and BRT, in terms of their advantages and disadvantages for eliciting the relationship between climate variability and crop yields. To our knowledge, this is among the first studies to explicitly conduct this analysis using identical crop yield and climate data to facilitate a valid comparison. We use India as a case study, and focus on three major crops (rice, wheat, and pearl millet). The article is divided into three major sections. We first analyze model performance in terms of out-of-sample prediction accuracy, interpret the models with reference to various climate variables using partial dependence analysis, and then conduct some historical climate change simulations using both LR and ML models. We conclude by summarizing the pros and cons of both from a crop yield analysis perspective.

3.2 Data and methods

3.2.1 Climate and crop production data

We used the same data as chapter 2, a detailed description of which has already been provided in that chapter.

3.2.2 Statistical software and methods

We conducted our analysis in R (R core team, 2020); R packages used include tidymodels (Wickham et al., 2019), data.table (Dowle & Srinivasan, 2021), ggthemes (Arnold, 2021), RColorBrewer (Neuwirth, 2014), wesanderson (Ram & Wickham, 2018), gridExtra (Auguie,

2017), doParallel (Microsoft & Weston, 2020a), foreach (Microsoft & Weston, 2020b), dismo (Hijmans, Phillips, Leathwick, & Elith, 2020), gbm (B. Greenwell, Boehmke, & Cunningham, 2020), segmented (Muggeo, 2008), and pdp (B. M. Greenwell, 2017).

BRTs were constructed with the gbm.step function in the gbm R package. It allows automated detection of optimum number of trees using k-fold cross-validation. The function needs two user-defined parameters: (i) the learning rate or shrinkage, which determines the contribution of each tree as the model grows, and (ii) the tree complexity, which controls the number of interactions between IVs. Following the recommendations of Elith, Leathwick, & Hastie (2008), we selected five learning rates (0.1, 0.03, 0.01, 0.003, 0.001) and five tree complexity values (2, 4, 6, 8, 10) as possible candidates for our BRT models. We randomly sampled 80 percent of our data (stratified over years), used that to construct a BRT model for each shrinkage/tree complexity/crop/model permutation, and tested on the 20 percent data held out earlier. The shrinkage and tree complexity pair that gave the most accurate results in terms of root mean squared error (RMSE) for the highest number of crop/model combinations was then chosen for all further analysis. In our case, this turned out to be a learning rate of 0.01 and a tree complexity of 6.

3.2.3 Models and climate variables

This study examined three different sets of climate variables as input to the models:

1. *noclim*: no climate variables,

2. *Tavg_Psum*: average seasonal temperature and total seasonal precipitation[4], and
3. *Tavg_Psumday*: average seasonal temperature, total seasonal precipitation and total precipitation days over the growing season.

All three variable sets contained identical geography and time variables to account for non-climatic determinants of crop yield. Detailed description of the process is provided in chapter 2.

Each of the three variable sets were then paired with four different modeling algorithms:

1. *lr_mono*: LR with only monomial terms of time and climate,
2. *lr_quad*: LR containing quadratic forms of time and climate,
3. *lr_sgm*: segmented LR with one knot[5] for year and each climate term, and
4. *brt*: BRT with the same variables used in the LR models[6].

All combinations of the three variable sets and four models were executed for three crops (rice, wheat, and pearl millet). To summarize, we analyzed three climate variable sets, using four different modeling techniques each, for three separate crops, making a total of 36 models. The variable set names, model names, and climate variables included in each run are presented in Table 3.1. Henceforth, for clarity and conciseness, we italicize variable set and model names and use the naming convention of *model:variable set* (e.g., *lr_sgm:Tavg_Psum*) when referring to a

[4] Precipitation as a variable ignores the initial conditions. For example, soil moisture could be low, medium, or high at the start of the season, which is not captured by seasonal precipitation. We address this limitation using a soil moisture model in chapter 4.

[5] The knot determination process is explained in greater detail in a later section.

[6] Note that the *brt* model does not need any functional form as input. Only names of relevant IVs and DV are input into the algorithm.

particular instance of our 36 models. We use the non-italicized capitalized format when referring to LR and BRT in general.

The *lr_mono* model used the following equation:

$$y_{it} = \alpha_i + \beta(t) + \gamma_1(clim\_var_1) + \ldots + \gamma_n(clim\_var_n) + \varepsilon_{it} \ , \quad (3.1)$$

where y_{it} is crop yield in district i and year t; α_i is district specific intercept; β is parameter for time (harvest year) trend; γ_n is parameter for the n^{th} climate variable (clim_var) included in the model; ε_{it} is the standard error. For *lr_quad*, equation (3.1) was modified to include quadratic terms for time and all climate terms. *lr_sgm* again used equation (3.1) as functional form, but with a single knot for each variable (including time), allowing for a more flexible fit. *brt* does not need any formula as input.

Table 3.1 Model specifications. This analysis was conducted separately for three crops (rice, wheat, and pearl millet).

Model type	**Climate variable type**	**Non-climatic variables**	**Climatic variables**
lr_mono	*noclim*	district dummy, year	--
lr_mono	*Tavg_Psum*	district dummy, year	Mean daily temperature during the growing season, Total seasonal precipitation
lr_mono	*Tavg_ Psumday*	district dummy, year	Mean daily temperature during the growing season, Total seasonal precipitation,

Model type	**Climate variable type**	**Non-climatic variables**	**Climatic variables**
			Total seasonal precipitation days (precipitation > 0.1 mm) (May, 2004)
lr_quad	*noclim*	district dummy, year, year^2	--
lr_quad	*Tavg_Psum*	district dummy, year, year^2	Mean daily temperature during the growing season, (Mean daily temperature during the growing season)^2, Total seasonal precipitation, (Total seasonal precipitation)^2
lr_quad	*Tavg_ Psumday*	district dummy, year, year^2	Mean daily temperature during the growing season, (Mean daily temperature during the growing season)^2, Total seasonal precipitation, (Total seasonal precipitation)^2, Total seasonal precipitation days, (Total seasonal precipitation days)^2
lr_sgm	*noclim*	district dummy, year (1 knot)	--
lr_sgm	*Tavg_Psum*	district dummy, year (1 knot)	Mean daily temperature during the growing season (1 knot), Total seasonal precipitation (1 knot)
lr_sgm	*Tavg_ Psumday*	district dummy, year (1 knot)	Mean daily temperature during the growing season (1 knot), Total seasonal precipitation (1 knot), Total seasonal precipitation days (1 knot)
brt	*noclim*	district dummy, year	--

Model type	Climate variable type	Non-climatic variables	Climatic variables
brt	*Tavg_Psum*	district dummy, year	Mean daily temperature during the growing season, Total seasonal precipitation
brt	*Tavg_ Psumday*	district dummy, year	Mean daily temperature during the growing season, Total seasonal precipitation, Total seasonal precipitation days

Similar to chapter 2, we measured model performance in terms of RMSE, the square root of the mean of the squared differences between observed and predicted values. We conducted RMSE analysis using out-of-sample 10-fold cross-validation with random samples stratified over years, a technique commonly used in model comparison and selection studies (Ortiz-Bobea et al., 2019). The exact same procedure was followed for all models. Out-of-sample predictions prevent overfitting by keeping the model's training and testing datasets separate.

3.2.4 Model inference

3.2.4.1 Partial dependence plot

lr_mono outputs regression coefficients for each IV, which can be interpreted as the marginal effect of each IV on the DV. This is true for *lr_quad* and *lr_sgm* as well, although the interpretation is not as straightforward as *lr_mono* because of multiple coefficients for each IV in model output. Meanwhile the *brt* does not fit any functional form to the data. To examine the marginal effect of changes in each IV on the DV for all four model types, we plotted partial dependence plots (PDPs) for each IV. PDPs depict the DV as a function of an IV, holding all other IVs constant at their mean values. Interpretation of partial dependence plots for a specific

IV assumes average value for all other IVs (Hastie, Tibshirani, & Friedman, 2017), so it is conditional on there being low correlation between IVs in the model. There is a high correlation between total seasonal precipitation and number of precipitation days in the season (R^2 of 0.27, 0.68, 0.24 for rice, wheat, pearl millet), compared to a lower correlation (0.07, 0.04, 0.11) between mean seasonal temperature and total seasonal precipitation. Hence, we present and discuss partial dependence plots from only the *Tavg_Psum* models for all three crops.

A limitation of using observational data for model inference is that the true relationship is unknown. So, it is impossible to assess if a statistical model is able to fully unmask the true relationship (unless there is an accompanying controlled experiment for calibration). To extend the ideas from chapter 2 of possible conflation between climatic and non-climatic variables to current analysis, we created synthetic crop yield and climate data with manually specified DV-IV relationships. LR and BRT models were constructed using this data, and their PDPs were examined to assess the model's accuracy with respect to the true relationship. More details about the synthetic data creation and model fitting process, along with the results, are provided in a later section.

3.2.4.2 Specification of segmented LR using BRT partial dependence plots

In addition to interpretation of DV-IV relationships, we also used partial dependence plots for fitting some segmented LR models. The "segmented" R function (Muggeo, 2008) that we used for our *lr_sgm* model, requires an input of either the number of knots for each IV, or the initial location of the knots from where the function can start estimating the breakpoint locations. We first ran the function in the automated knot search mode, and observed that the function failed to

find breakpoints for some crop-model combinations. For example, the *lr_sgm:Tavg_Psum* model for wheat failed for 38 percent of our repeated runs. In such cases, we needed to manually specify the location for the function to start knot search from, for which the partial dependence plots from *brt* were useful. The latter does not require any predetermined functional form specification, so it could identify knots automatically. We used the *brt* partial dependence plots as a diagnostic tool, identified the most probable location of breakpoints visually, and input those as starting location of knots for the "segmented" function. By entering knot locations manually, the proportion of failed runs dropped to 14 percent for *lr_sgm:Tavg_Psum*.

3.2.5 Simulations of climate change impacts

We used the same methodology as described in chapter 2 to estimate the historical impacts of climate change on rice, wheat, and pearl millet yield across India. Predictions from all four modeling procedures (*lr_mono*, *lr_quad*, *lr_sgm*, and *brt*) for the two non-null climate variable sets (*Tavg_Psum*, and *Tavg_Psumday*) were then compared and contrasted.

3.3 Results and discussion

3.3.1 Model accuracy

The models' yield prediction accuracy differs significantly between LR and BRT (Figure B.1). In terms of absolute RMSE values, the accuracy of the three LR models (*lr_mono*, *lr_quad*, *lr_sgm*) is in a similar range, while *brt* outperforms all three LR methods by a substantial margin. For each crop, even the best performing LR models have higher RMSE than the corresponding *noclim* BRT model. So, if our crop models were to be compared solely in terms of their ability to predict absolute yield numbers in the near-term, our results strengthen the case for using ML

models. This is because a BRT is able to fit complex and flexible relationships between yield and geography/time in the *noclim* model, while the LR, even with climate variables, is restricted by the functional form selected by the modeler.

This study, however, primarily focuses on using these statistical techniques for inferring the role of climate variability in determining crop yields. For that, the contribution of climate variables in improving models, over and above geography and time, is of critical importance. Therefore, we also quantified the performance of each model for each crop in terms of percent reduction in RMSE compared to the corresponding model type with no climate variables (*noclim* model) (Figure 3.1). Adding climate variables significantly improved model performance for rice and pearl millet, unlike wheat where climate variables did not add to the model's RMSE reduction capability (both LR and BRT). This pattern was consistently observed for all four model types. Note that the absence of change in RMSE does not necessarily mean that climate does not explain any wheat yield signal: our findings from chapter 2 show that there may be compensatory effects occurring between yield signal attributable to climatic and non-climatic variables, with the result that overall model accuracy stays the same between *noclim*, *Tavg_Psum*, and *Tavg_Psumday* wheat models. Reiterating chapter 2's main conclusion, lack of RMSE reduction does not necessarily nullify the utility of including climate variables in a crop model.

For rice and pearl millet, the percent increase in performance with addition of climate variables was the highest for *brt*. Among all three LR models, the maximum decrease in RMSE compared to the respective *noclim* models was around two percent, while the same metric for BRT models

was over six and five percent for rice and pearl millet, respectively. Interestingly, not only was the performance of *brt:noclim* significantly better than all LR models (Figure B.1), but the jump in yield prediction accuracy with the addition of climate variables too was maximum for the BRT models. This is potential evidence of yield signal attributable to climate that all three LR algorithms are unable to explain with just mean seasonal temperature and total precipitation; the BRT model possibly fits a better function between the same DV and IVs.

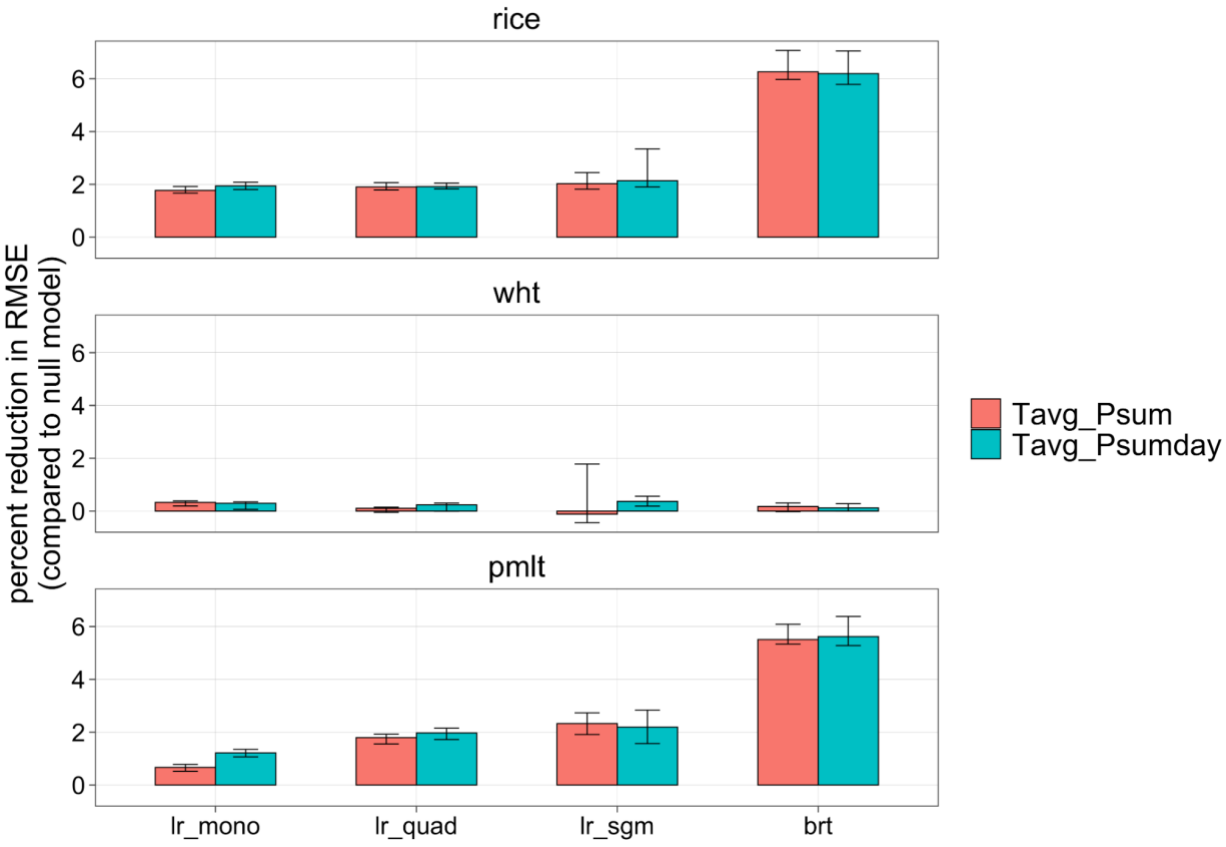

Figure 3.1 Model performance measured in terms of percent decrease in RMSE, compared to the corresponding null model (with no climate data). The three crops are rice (top), wheat (middle), and pearl millet (bottom). Within each panel, the bars are color-coded by climate variables included (red: mean seasonal temperature and total seasonal precipitation; cyan: mean seasonal temperature, total seasonal precipitation and total precipitation days over growing season). From left to right, the various model types

depicted are: (1) *lr_mono*: LR with linear terms; (2) *lr_quad*: LR with quadratic terms for time and all climate variables; (3) *lr_sgm*: LR with single-knot segmented analysis; (4) *brt*: BRT.

The inclusion of precipitation days did not improve model performance significantly in any case except *lr_mono* for pearl millet; this is evident from the overlap in error bars between each red and cyan bar pair in Figure 3.1. We again underscore the possibility that the number of precipitation days, although apparently not important in Figure 3.1, could still be an important variable for explaining crop yield during extreme weather events. Results from chapter 2 also show an increase in "relative importance" of climate with respect to geography and time when adding precipitation days as a variable, so we decided to keep *Tavg_Psumday* models as well.

3.3.2 Partial dependence plots

Partial dependence of pearl millet yield on time is consistent across all four model types (Figure 3.2); a linear approximation seems valid for estimating the increase in pearl millet yield over time due to factors such as advancement in agricultural practices, better technology, mechanization, or availability of improved cultivars (Lobell & Burke, 2009). The climate panels, however, depict a departure from linearity for the response of pearl millet yields to average seasonal temperature (middle) and total seasonal precipitation (right). *lr_quad*, *lr_sgm*, and *brt* predict a decrease in marginal yield benefits with increasing precipitation, before yield reaches a maxima and is predicted to decline with any further increase in precipitation. *lr_mono*, meanwhile, predicts continuously increasing yield as precipitation increases, an artifact of the algorithm forcing a linear relationship between DV and IVs. This obvious flaw shows the necessity of accounting for potential non-linearity in yield-climate response.

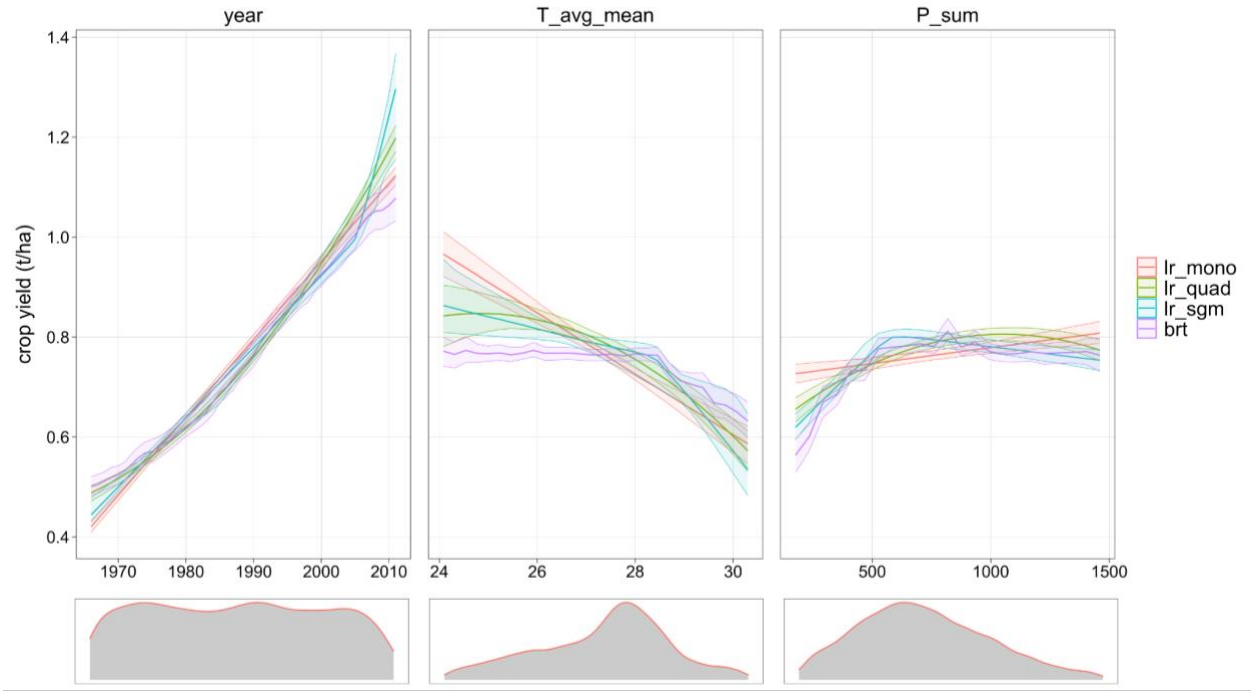

Figure 3.2 Partial dependence plots of the *Tavg_Psum* variable set models for pearl millet (top row), and distribution density of corresponding IV in training data (bottom row). The four model types are color-coded in each panel. Data density plots at the bottom provide an idea of where most of the training data lies for a particular IV. For example, for the temperature partial dependence plot (middle panel), data is heavily concentrated at the higher end of the temperature range, hence the wide confidence intervals towards the left side of the partial dependence plot in that panel.

Some striking patterns are observed in the middle temperature panel. As expected, *lr_mono* fits pearl millet yield as a monotonically decreasing function of average seasonal temperature. brt, however, predicts a non-significant impact of temperature rise till around 28 degrees Celsius, beyond which there is a significant drop in yield with rise in temperature. Fitting such non-linear relationships is a strength of ML algorithms like BRTs; this has also been documented in previous studies for similar algorithms like “random forests” (Vogel et al., 2019). Running

lr_sgm in automatic knot-identification mode sometimes failed to identify any breakpoints. But when 28 degrees Celsius (as identified by *brt*) was input as a starting point for knot estimation, *lr_sgm* ran to completion and identified a breakpoint close to the turning point detected by *brt*. In other words, *brt* acted as a diagnostic tool and helped in detecting the breakpoint for *lr_sgm*. *lr_quad* too predicts an increasingly detrimental effect increase in temperature.

Similar to pearl millet, there is agreement between the partial dependence estimates for precipitation on rice yield as estimated by *lr_sgm*, and *brt*: yield rises with increase in precipitation, before it reaches a threshold beyond 1200 mm (Figure B.2); *lr_mono*'s linear fit deviates a bit, but the difference is not as stark as pearl millet. All four model types estimate no significant impact of precipitation on wheat yield. This may be due to the fact that wheat is the most heavily irrigated crop out of the three analyzed in this study: during the last five year of this study's timeframe, 93 percent of the national area under wheat cultivation had access to irrigation; the corresponding numbers for rice and pearl millet are 64 and 9 percent, respectively. The temperature panels for both rice and wheat present a more interesting trend: all three LR specifications predict a strong and significant negative influence on yield of temperature increase, but the effect size estimates from the *brt* are comparatively small. There are temperature ranges where some LR and ML estimates concur (all four have similar slopes in the high temperature range for rice; *lr_sgm* and *brt* estimate no significant impact of temperature change in the less than 22.5 degrees Celsius region for wheat). However, in general, *brt* predicts a lower sensitivity to temperature compared to the three LR models, and this difference is most stark for rice.

3.3.3 Inference from synthetic data

Here we present a hypothesis and supporting evidence to potentially explain the apparent difference between LR and BRT in terms of rice and wheat sensitivity to climate variability described above. We created three sets of synthetic mean seasonal temperature data for 45 years and 300 districts (to match the size of real data): A) no time trend in temperature (no climate change scenario); B) temperature increasing linearly over time; C) temperature increasing quadratically over time. Random noise was added to these data sets to emulate real conditions and to prevent a perfect fit. The mean and standard deviation of this data matched that of the actual mean seasonal temperature data from our observed climate data.

For each of the three temperature datasets, synthetic crop yield was then calculated, using equation 3.1, as a linear function of temperature with the same coefficient for yield versus temperature, in addition to time and geography as fixed effects. The data thus created was fed back into an LR model with yield as the DV and a monomial term of temperature as the only IV besides geography and time. Exact same process was also repeated with a BRT. Results show that the LR model (red plots in Figure 3.3) is able to fit the expected functions (black dotted line) in all three cases. This is expected since yield was explicitly modelled as a linear function of temperature, and the LR is simply re-discovering that linear relationship. Conversely, as we move from a time-independent to linearly and then quadratically varying temperature data (or, when temperature contains a stronger time trend), the BRT model (blue plot in Figure 3.3) predicts a progressively lower sensitivity of crop yields to temperature, even though yield data was created from temperature using the same coefficient in all three cases. In this case, the BRT, because of how it builds a flexible model with no user-defined restrictions, is prone to conflating

time and temperature when there is correlation between them. Additional tests with synthetic data are available in Appendix B section B.3.

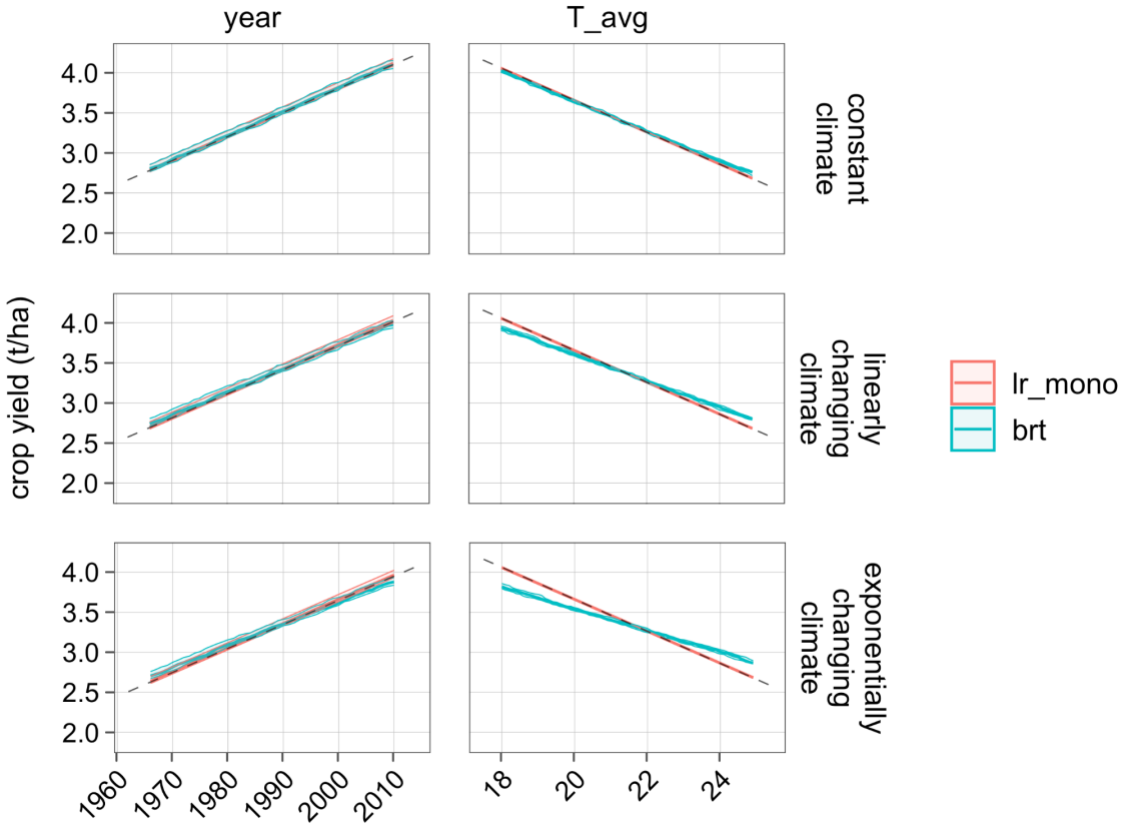

Figure 3.3 Partial dependence plots for LR (red) and BRT (cyan) models fitted on synthetic data using user-defined coefficient for temperature. The LR is able to accurately model the relationship (expected because the underlying data creation used a linear function), but the BRT sensitivity to temperature change goes down (the blue plot becomes flatter) as temperature gets more correlated with time (top to bottom, right column).

Based on these results using synthetic data, we can now better interpret our previous results that showed BRTs having a lower yield sensitivity to temperature. Climate change has introduced a significant time trend in temperature over the period of this study. While rice and wheat yield may have been negatively influenced by rising temperature (Davis et al., 2019; Lobell, Sibley, &

Ivan Ortiz-Monasterio, 2012), it is possible that the *brt* model in our analysis with actual data used the time variable to account for that yield decrease. In other words, the *brt* may be conflating changing climate's impact on rice and wheat yield with the time variable; this could explain the lower yield predictions in later years by *brt* compared to the other three LR model. Indeed, for rice and wheat, the year partial dependence plots (Figures B.2 and B.3) show increasingly declining yield predictions over time by *brt*, in comparison to the three LR models; this is also evident in synthetic data results in Figure 3.3. This is a potential drawback of ML methods that fit non-parametric models without requiring any functional input or a priori information from the user.

Above results notwithstanding, there are cases where the flexible nature of non-parametric models like BRT can lead to better insights, compared to LR with user-specified functional forms. For instance, dataset A from the analysis above was constructed assuming the same relationship between crop yield and temperature across districts (same temperature coefficient was used for calculating yield in each district). At the other end of the spectrum lies the case where crop yield's dependence on temperature (or any other climate variable) varies across regions: for example, a sub-national multi-crop analysis in the Great Plains region of the US reported both positive and negative impacts of temperature increase on crop yields across different counties of the larger region (Kukal & Irmak, 2018). To test LR and BRT in such a scenario, we created another synthetic dataset, with no climate trend in temperature (similar to dataset A above), but a different yield versus climate coefficient for each district. We call it dataset D. This represents a spatially-varying relationship between yield and climate; similar analysis could also be conducted for a temporally-varying relationship, but we skip that for

conciseness. LR and BRT models were then built using dataset D using the exact same procedure as above.

As expected, both LR and BRT are able to elicit the correct relationship for dataset A (Figure 3.4; top row). Since the slope of yield versus temperature is geography-independent, which matches the LR model specification, BRT offers no advantage in terms of accuracy. However, for dataset D, because of contradiction in the way the LR model was specified (with a single coefficient for yield versus temperature in all districts) and the actual geography-dependent relationship between yield and temperature, the BRT model outperforms LR by a big margin (Figure 3.4; bottom row). While it is true that a more flexible LR with district-specific coefficients would have been a better choice here, that needs manual tweaking of the functional form which may not always be possible if there are multiple climate variables. Also, highly parameterized LR models are prone to overfitting and can have detrimental effects on statistical power. BRT, on the other hand, did not need any extra input from the user, and was able to elicit the relationship on its own. In other words, dataset D demonstrates the utility of fitting a more flexible ML technique compared to a more biased LR. To summarize, the advantages and disadvantages of using LR versus BRT are highly context-dependent (Table 3.2).

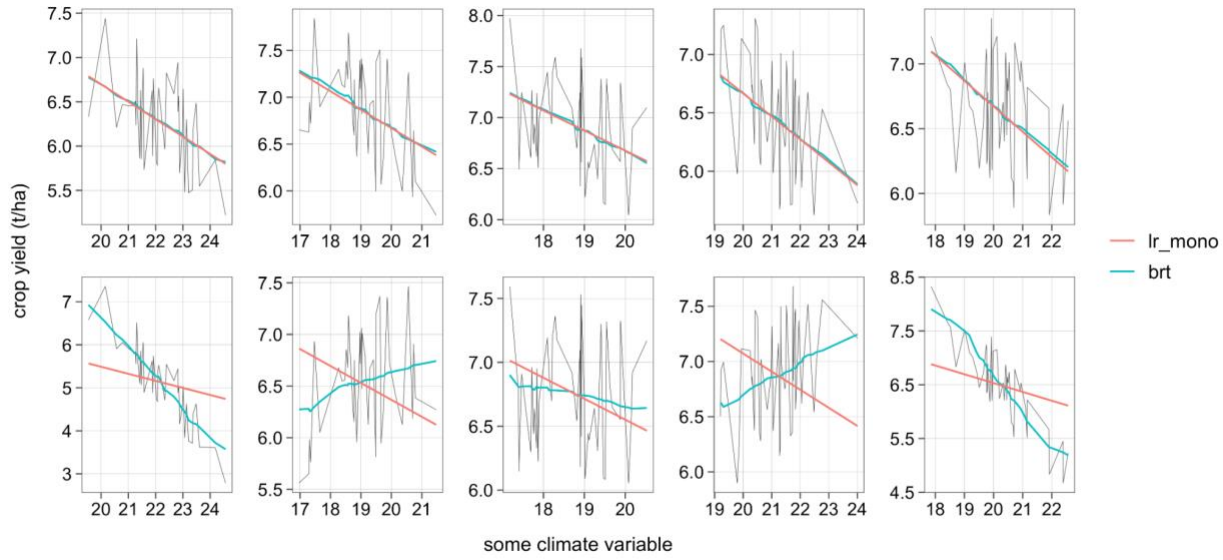

Figure 3.4 Predictions of LR (red) and BRT (cyan) models of crop yield as a function of temperature (proxy for climate) versus actual crop yields in the training data (black). For data A (top row) with the same yield-temperature coefficient in all districts, LR and BRT performance is similar. But when yield-temperature relationship varies across districts (data D; bottom row), BRT predictions match the actual yield data a lot more than LR. For illustration purposes, we show plots for only five districts chosen randomly out of the 300 in the training data.

Table 3.2 Context-dependent advantages and disadvantages of BRT compared to LR. The more advantageous technique for each criterion is italicized.

LR	**BRT**
LR models have been studied, used, and reported extensively; they are relatively simple and easy to interpret.	ML methods like BRTs are comparatively less widely used in the field of crop yield modeling. As opposed to LR, they are also more difficult to make inferences from.

LR	**BRT**
LR needs manual specification of expected functional form, which can get unwieldy when there are multiple DVs. It is also prone to mis-specification in the absence of complete domain knowledge and information about possible limitations to the suggested functional form.	*BRT is able to elicit the relationship between the DV and IVs without any a priori assumption or input from the user.* *This can be especially critical in cases like the one demonstrated in section 3.3, where the relationship between yield and climate varied across space.* *Also, BRTs can account for interactions which the user may not have anticipated (and therefore not included in the LR model). For example, when irrigation is included in models, a term denoting its interaction with precipitation is often added. But studies have shown evidence of irrigation reducing yield sensitivity to extreme heat as well (Zaveri & Lobell, 2019). Unless included explicitly, this relationship will be missed in an LR, but a BRT will be able to include it automatically.*
LR in its most basic form is sensitive to outliers. There are statistical ways to make the analysis robust to outliers, but those again require active intervention by the user.	*BRTs are flexible enough to work with missing data or outliers without adversely affecting model performance.* However, BRTs, like LR, are prone to conflating correlated variables, as shown using our synthetic data analysis.
Simplicity of LR means quicker and less computationally-intensive analysis.	BRTs usually take longer to run, and access to high performance computing facilities may be needed if running a big array of models.

3.3.4 Simulations of climate change impacts

The predicted effects of historical climate change, as predicted by the four model types, vary across three crops analyzed in this study. Figure 3.5 shows the simulated impact of climate change on pearl millet yield during the last decade of this study (2002-2011). Rows depict the results from the four model types (*lr_mono*, *lr_quad*, *lr_sgm*, *brt*) while the two columns correspond to the two climate variable sets discussed in this study (*Tavg_Psum*, *T_avg_Psumday*). The accompanying map on the right shows the 10-year (2002-2011) average of the mean temperature during pearl millet season. For the comparatively cooler regions of the country (central and western districts), *lr_mono* predicts a bigger negative impact of climate change compared to the other three (*lr_quad*, *lr_sgm*, *brt*). This is commensurate with the partial dependence plots in Figure 3.2 where *lr_mono* has the biggest effect size per degree Celsius rise in temperature of all the model types in the lower temperature range (evident from the sharpest negative slope of *lr_mono* amongst the four). But in the northwest, which is the warmest region of India (and Rajasthan being the biggest pearl millet producer in the country), *lr_quad* and *lr_sgm* estimate a bigger influence of climate change than either *lr_mono* or *brt*. This trend too matches the patterns seen in the partial dependence plots in Figure 3.2 (high temperature range towards the right side of the middle panel). Comparing the three LR outputs to *brt*, the latter also predicts the highest impact in the north-western region, although the magnitude of the predicted impact of historical climate change on pearl millet yield is smaller than the LR models.

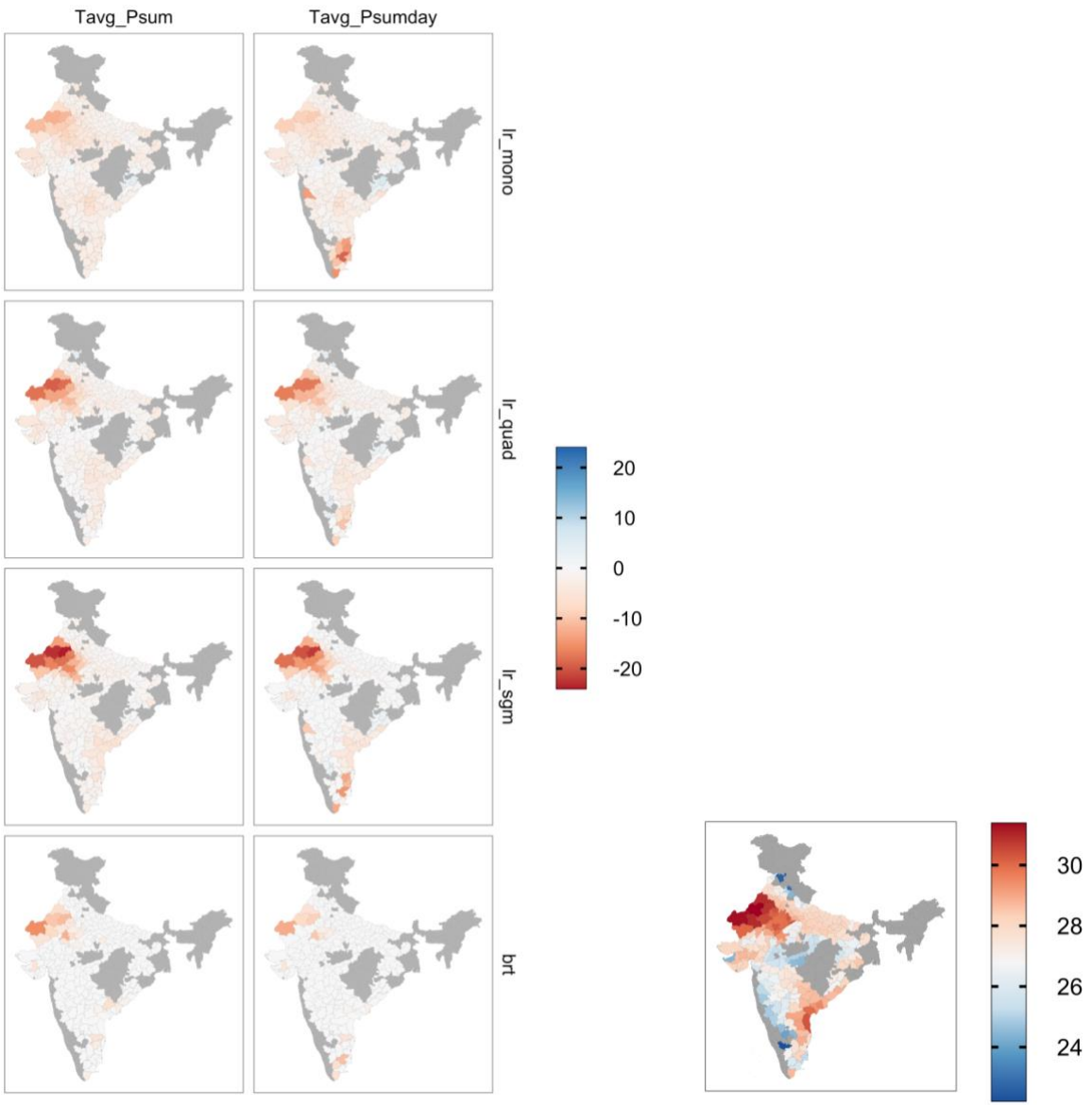

Figure 3.5 Simulated impact of long-term climate change (since 1966) on pearl millet yield (percent change) in the last decade (2002-2011) of the study time period (left); 10-year (2002-2011) average of the mean temperature during pearl millet season (right). The climate data was linearly detrended to remove time trend at district-scale. District-level estimates of median value and 95 percent confidence intervals of climate change impact on yield were obtained through residual bootstrapping (n = 500). The average district-level yield loss during the last decade in the dataset (2002-2011) is presented here as the expected impact of climate change

that has occurred since 1966. Only results with 95 percent significance of the confidence intervals are shown; insignificant results are shown in gray.

For rice (Figure B.5), the climate change impact estimates are similar for all three LR models: rice yield in most regions, except some parts of south-east India, has been negatively influenced by climate change. Some climate change impact hotspots show up in all six LR plots, the biggest of which is the central India region comprising mainly the state of Madhya Pradesh (an important rice producer). The hottest districts in the south show a smaller effect of climate change in lr_quad and lr_sgm compared to the predictions made by lr_mono. This matches the shallower slopes observed for lr_quad and lr_sgm (compared to lr_mono) on the extreme right side of rice partial dependence plots (Figure B.2). However, these differences are not as stark as with pearl millet. Overall, the biggest contrast is between the LR and ML predictions; the latter shows little to no impact of historical climate change across most parts of India. Wheat (Figure B.6) shows similar patterns as rice: all three LR exhibit similar effects of climate change. While *lr_mono* predicts a more consistent impact throughout the country, more variation is observed in *lr_quad* and *lr_sgm* models. There are some districts in the southern state of Karnataka that are predicted to have experienced larger yield losses due to climate change in the *T_avg_Psumday*:*lr_sgm* panel compared to the corresponding predictions in the *T_avg_Psumday*:*lr_mono* panel. Regardless, the trends are more or less similar across the three LR methods, and the biggest difference is observed between LR and ML panels. Just like for rice, *brt* predicts that there has been no wheat yield loss due to climate change in India.

The most important takeaway from this climate change analysis concerns the range of predictions models can make depending on the statistical technique used. Although BRTs have distinctly better accuracy compared to the LR models (in terms of RMSE reduction, Figure 3.1), their predictions of historical climate change on yields of all three crops is considerably smaller. But this lower yield sensitivity shown by BRTs to climate change could be explained by their potential conflation of time with climate change as shown previously.

We therefore caution against using only the most accurate modeling techniques (*brt* in our case) to analyze and predict climate change impacts. That would ignore the significantly higher impacts being predicted by the more sensitive albeit less accurate algorithms. Rather, researchers should consider an ensemble of climate variable sets and models to get a range of predictions that does not ignore any possible outcome. This is even more important considering potential pitfalls in what might otherwise seem like the most appropriate or accurate technique.

3.3.5 Limitations

Some limitations of this study and important considerations while using the described statistical techniques warrant discussion here. Like chapter 2, this analysis too was conducted at a national scale. Trends reported in more granular studies are potentially missed, and the next steps could be replication of this analysis at a more local scale. A very critical point that is sometimes missed in crop yield model studies is the problem with using any statistical method for prediction purposes with new data outside the range of training data. Pushed too far, this can lead to unrealistic predictions. For example, a quadratic functional form for temperature may predict negative yield values when extrapolated to temperatures well beyond the training data range.

Tree-based methods like BRTs are arguably worse than LR at predicting outside the training data range. As a consequence of their core algorithm, BRTs fit a constant value to the DV when predicting for IVs outside the range of the training dataset (as opposed to LR algorithms that just continue to use the modeled function with new data). This may have serious implications for analyzing climate change impact on crop yields since temperatures in some regions are expected to cross the highest temperatures ever recorded.

3.4 Conclusion

The past few years have witnessed a rapid rise in the availability of high quality open-source climate and crop yield data, advanced computational facilities, and increasingly accessible statistical software. The field of crop yield research has benefited a lot from this trend; researchers are using high spatial and/or temporal resolution data with advanced statistical techniques for answering questions and gathering insights that would have been impossible a few years ago. Given the wide range of tools available for current researchers, it is imperative to understand the advantages and disadvantages of each.

This study compared a traditional statistical technique to an advanced machine learning method with the aim of advancing our understanding of the link between climate variability and crop yields. Our analysis found that both methods have their pros and cons, and combining both methods in an ensemble seems like the most optimal choice. While the simpler LR models are faster to fit and simpler to interpret, their accuracy in predicting yields is bested by BRTs by a big margin. In this study, BRTs also helped identify breakpoints in IV-DV relationships that were then used in fitting the segmented LR models.

Conversely, BRTs are comparatively hard to interpret, and users have less control on the functional form fit by the model. We show through an example that ML algorithms like BRTs can conflate two correlated IVs and lead to potentially erroneous results. Nevertheless, as shown using the synthetic data, BRTs can be really powerful at fitting flexible functions where the relationship between yields and climate variables is suspected to change widely over geography and/or time. Also, BRTs can include obscure interactions between IVs that may otherwise be missed if the user employing LR models does not expect them. All benefits that BRTs provide automatically, can also be derived using LR, but it needs manual tweaking of the model which requires a priori knowledge and assumptions about functional form. Even if the latter is available, adding unnecessary complexity to LR may reduce statistical power when there are multiple climate variables being analyzed simultaneously, and the output of such a model gets unwieldy. At the same time, the advantages of the LR family discussed above continue to keep them relevant in this field. Used in conjunction, the two major statistical techniques analyzed in this study can lead to a better understanding of climate-yield relationship and more accurate future yield predictions in the context of a changing climate.

Chapter 4: Indian agriculture in a changing climate: using CMIP6 projections for short-term and long-term crop yield predictions

4.1 Introduction

India is primarily an agrarian society. While the contribution of agriculture to India's economy has gone down over the past many years[7], it continues to be the biggest employment provider: over 42 percent of the national workforce is currently employed in agriculture (World Bank, 2021b). In addition to its importance for the country's economy, India's agricultural sector contributes significantly to global food production as well. For example, India produces 70 percent of the world's chickpeas (16 times more than the second largest producer, Turkey), and over one-third of the world's millets and a quarter of the world's rice are grown in India (FAOSTAT, 2021). Indian agriculture plays a prominent role in ensuring food security (both within the country and globally) and providing livelihood to millions of households.

Agriculture is heavily affected by short-term and long-term climate variability, and Indian agriculture is no exception. The Sixth Assessment Report (AR6) of the Intergovernmental Panel on Climate Change (IPCC) has once again underscored the reality of climate change (IPCC, 2021). The global climate research community has never been more unambiguous that temperatures will continue to rise under every scenario deemed plausible, bringing with them a host of other changes including shifts in intensity and distribution of precipitation, extreme

[7] Agriculture, forestry, and fisheries together accounted for a little over 18 percent of the country's Gross Domestic Product in 2020 (World Bank, 2021a).

weather events like droughts and storms. Heat waves too are projected to increase in intensity, frequency, and duration across India over time (Das & Umamahesh, 2021). Indian agriculture is primarily dependent on monsoon rains for water (Sharma, Rao, Vittal, Ramakrishna, & Amarasinghe, 2010), which makes it extremely vulnerable to dry spells and short-duration rains which have (and will continue to) become increasingly frequent with the changing climate (Annamalai, Hafner, Sooraj, & Pillai, 2013; V. Gupta, Singh, & Jain, 2020). This vulnerability is amplified in the rainfed regions which account for approximately half of the net sown area in India (Ministry of Agriculture and Farmers Welfare, 2018). Working group II's (WG II) contribution to the IPCC Fifth Assessment Report (AR5)[8] has ranked Indian agriculture among those most vulnerable to climate change (IPCC, 2014).

Given the certainty and rapid pace of climate change, it is of utmost importance to build robust crop models for India that operate at subseasonal scale to not only identify crops and regions most at risk across the country, but also estimate the losses under different possible climate change scenarios. This study aims to accomplish this task by using the latest climate change projections, in conjunction with statistical and machine learning (ML) methods. In these models, we use a variety of climate variables including mean seasonal temperature, total seasonal precipitation, subseasonal temperature patterns, number of precipitation days among others. Additionally, we build and run a simplified soil moisture model, whose output serves as input to our crop yield models to help analyze potential benefits of soil moisture models over those with precipitation-based variables.

[8] WG II's contribution to the IPCC Assessment Report 6 will be published in 2022.

For our analysis, we use future climate data projections from Coupled Model Intercomparison Project Phase 6 (CMIP6), the latest framework of climate model experiments which improve over the previous versions by including better representation of global physical, chemical, and biological processes, and running them at higher resolution than in the past (IPCC, 2021). Compared to CMIP5, the CMIP6 framework offers certain advantages: it is less biased, has lower uncertainties, and can better simulate the expected precipitation and temperature patterns over South Asia (Gusain, Ghosh, & Karmakar, 2020; Katzenberger, Schewe, Pongratz, & Levermann, 2020; Zhai et al., 2020).

CMIP6 builds upon the four representative concentration pathways (RCPs) featured in IPCC AR5 by including pathways for them (SSP1-2.6, SSP2-4.5, SSP4-6.0, and SSP5-8.5 to match RCP2.6, RCP4.5, RCP6.0, and RCP8.5, respectively). In addition, CMIP6 also contains new scenarios to provide a wider range of possible futures; these include SSP1-1.9, SSP4-3.4, SSP5-3.4OS, and SSP3-7.0. We follow the IPCC AR6 naming convention of referring to these scenarios as SSPx-y, where "SSPx" refers to the Shared Socio-economic Pathway (SSP) denoting the socio-economic trends underlying the particular scenario, and "y" signifies the level of radiative forcing (W/m^2) expected to result from the scenario in 2100 (IPCC, 2021). IPCC AR6 includes five scenarios: SSP1-1.9 and SSP1-2.6 to illustrate a future with low GHG emissions declining to net zero by 2050 followed by negative emissions up to 2100, SSP2-4.5 with intermediate GHG emissions, and SSP3-7.0 and SSP5-8.5 as examples of high emission scenarios. Here we analyze four scenarios (SSP1-2.6, SSP2-4.5, SSP3-7.0, and SSP5-8.5) to cover the range of possible future outcomes. RCP8.5 and the resultant SSP5-8.5 has been

criticized and termed extremely unlikely by some researchers because of its arguably unrealistic assumptions about coal use over this century (Ritchie & Dowlatabadi, 2017a, 2017b). At the same time, others have also suggested a greater than 35 percent chance of global emissions exceeding those assumed in RCP8.5 (Christensen, Gillingham, & Nordhaus, 2018). Regardless, we primarily focus on the "middle of the road" SSP2-4.5, as being arguably closest to the current trends (Hausfather, 2018).

4.2 Data and methods

4.2.1 Historical climate and crop production data

We used the same historical ICRISAT crop dataset as the previous two chapters, a detailed description of which has already been provided in chapter 2. An additional layer in the current analysis is that of irrigation. Access to irrigation influences crop yield sensitivity to short-term as well as long-term climate variability. The ICRISAT dataset contains irrigated area (ha) disaggregated by year, crop, and district (in addition to previously described crop production and harvested area data). We used the proportion of area irrigated (ratio of irrigated area to harvested area) for each crop-year-district combination as a proxy for irrigation availability in the current analysis; this was necessitated by the lack of historical irrigation water amount data at district-scale.

4.2.2 Soil moisture model

Statistical models discussed so far have one critical weakness. The mathematical relationships they estimate climate and crop yield might not necessarily be grounded in crop physiology (Roberts, Braun, Sinclair, Lobell, & Schlenker, 2017). Hence, statistical models are prone to

errors ranging from unrealistically simplistic relationships, to predicting relationships where none exist. The other family of crop models, called process-based models, incorporate experimentally-determined plant responses to various factors (temperature, water availability, soil moisture, radiation, carbon dioxide concentration ampng others) and build empirical mathematical relationships between them (Roberts et al., 2017). The advantage of such models is that the relationships and equations used for building them are based on plant physiology and backed by clear mechanisms linking weather and crop growth. However, these models rely on lab-controlled experiments and do not necessarily reflect real-life outcomes in farmers' fields because of their inability to include factors external to the experimental setting like farmer behavior, pest infestation among others. So, a good compromise might be to build statistical models driven using physiologically-appropriate variables. This also matches our findings from chapter 2 which recommended using physiologically important climate variables. Specifically, we ran a soil moisture model which incorporated crop-specific water demand across various growing stages as defined as FAO (Allen, Pereira, Raes, & Smith, 1998). The moisture availability factor as derived from this soil moisture model was then used as an input variable in our yield models. Past research too has shown soil moisture to be an important determinant of crop yields (Ortiz-Bobea, Wang, Carrillo, & Ault, 2019). More details about our soil moisture model methodology are available in Appendix C section C.1.

We used our chapter 2 methods to calculate relative importance of climate (as opposed to geography and time) in explaining crop yields for the three models under investigation in the current chapter. Results show crop-specific patterns: for pearl millet, and rice to a smaller extent, the *sub_soil_moisture* model gives equal weightage to climate as the *sub_prec* model (Figure

C.1). For wheat, the relative importance of climate in the *sub_soil_moisture* model lies in between that seen in the *seasonal* and *sub_prec* models. In summary, the similarity between climate's relative importance in the two subseasonal models, coupled with soil moisture's close link to actual crop physiology, makes the moisture model a worthy candidate for further analysis in the rest of this chapter.

4.2.3 Statistical techniques

We conducted our analysis in R (R core team, 2020); R packages used include tidymodels (Wickham et al., 2019), data.table (Dowle & Srinivasan, 2021), ggthemes (Arnold, 2021), RColorBrewer (Neuwirth, 2014), wesanderson (Ram & Wickham, 2018), gridExtra (Auguie, 2017), doParallel (Microsoft & Weston, 2020a), foreach (Microsoft & Weston, 2020b), dismo (Hijmans, Phillips, Leathwick, & Elith, 2020), gbm (B. Greenwell, Boehmke, & Cunningham, 2020), segmented (Muggeo, 2008), and pdp (B. M. Greenwell, 2017).

4.2.3.1 Climate variables

Results from chapter 2 showed that while progressively adding climate variables may not always translate into an increase in model prediction accuracy (measured in terms of common statistical metrics like R^2 or RMSE), the added climate variables may still be important for explaining observed variability in crop yield. Consequently, we ran our models across a range of climate variable sets; summary of results is presented in Appendix C section C.6. For clarity and comprehensibility, we discuss three climate variable sets hereafter: set 1 is the most parsimonious with only mean seasonal temperature, total seasonal precipitation, and total precipitation days, while sets 2 and 3 include more variables to account for potential subseasonal

trends (either through subseasonal precipitation variability, or soil moisture) in climate-crop relationship (Table 4.1). Hereafter, we call them *seasonal*, *sub_prec*, and *sub_soil_moisture*, respectively.

Table 4.1 Climate variable sets analyzed in this study. Details about each variable are available in Table 2.1 of chapter 2.

Climate variable set name	Climate variables included
seasonal	● Mean daily average temperature during the growing season ● Total seasonal precipitation ● Total seasonal precipitation days (precipitation > 0.1 mm) (May, 2004)
sub_prec	● Mean daily minimum temperature during the growing season ● Mean daily maximum temperature during the growing season ● Degree day bins, at 10-degree Celsius intervals (<0, 0-10, 10-20, 20-30, >30 degrees Celsius) ● Subseasonal precipitation over four crop growing stages (Allen et al., 1998) ● Subseasonal precipitation days over four crop growing stages
sub_soil_moisture	● Mean daily minimum temperature during the growing season ● Mean daily maximum temperature during the growing season ● Degree day bins, at 10-degree Celsius intervals (<0, 0-10, 10-20, 20-30, >30 degrees Celsius) ● Soil moisture availability bins: proportion of days spent in particular α bins (=1, 1-0.75, 0.75-0.50, 0.50-0.25, < 0.25)

4.2.3.2 Statistical models

As in chapter 3, we used each variable set (Table 4.1) to build four different types of statistical models: three linear regression (LR) models and one boosted regression trees (BRT) model. The four models have been described in greater detail in chapter 3. We continue to use the naming

convention of *model:variable set* (e.g., *lr_sgm:seasonal*) when referring to any model. All analysis was conducted for three major Indian crops: rice, wheat, and pearl millet. In total, we analyzed three variable sets, four models, and three crops, for a total of 36 model combinations.

District dummies and harvest year accounted for the non-climatic signal in all models. The first LR model, *lr_mono* contained monomial terms of all climate variables. To account for the effects of irrigation, lr_mono also contained an interaction between proportion of crop area irrigated and each climate variable related to water (precipitation, precipitation days, or soil moisture bins). For example, the *lr_mono:seasonal* model used the following equation:

$$y_{it} = \alpha_i + \beta(t) + \gamma(irri\_area) + \delta_1(T\_avg\_mean) + \delta_2(P\_sum) + \delta_3(P\_days) + \delta_4(irri\_area{:}P\_sum) + \delta_5(irri\_area{:}P\_days) + \varepsilon_{it} \quad , \tag{4.1}$$

where y_{it} is crop yield in district i and year t; α_i is district specific intercept; β is parameter for time (harvest year) trend; γ is irrigation area parameter; δ_1 is temperature parameter; δ_2 is precipitation parameter; δ_3 is precipitation days parameter; δ_4 is parameter for interaction between irrigation and precipitation; δ_5 is parameter for interaction between irrigation and precipitation days; ε_{it} is the standard error. Accordingly, models for the other two climate variable sets (*sub_prec* and *sub_soil_moisture*) were built using a modified form of equation (4.1).

With respect to the other two LR model types, *lr_quad* contained quadratic terms for time, irrigation, and climate variables, and *lr_sgm* was the segmented form of equation (4.1) with

knots created as explained in chapter 3. The *brt* models used the same variables as their LR counterparts, including irrigation.

4.2.4 CMIP6 climate projections for future yield prediction

Future climate projections, including those from the CMIP6, are prone to bias errors due to coarse resolution or parametrization (Mishra, Bhatia, & Tiwari, 2020). So, bias correction is critical, especially when using them for local or regional analysis. We used the data product created by Mishra et al. (2020). It contains bias-corrected projections from 2015-2100 of daily precipitation, minimum temperature, and maximum temperature at 0.25^{o} spatial resolution from 13 different General Circulation Models (GCMs) for the four SSP scenarios chosen for our study. It also includes historical climate simulations for 1950-2015. This dataset covers six South Asian nations; we only used the data for India. More details about the bias correction in the final product are available in Mishra et al. (2020).

Since the CMIP6 climate dataset was in gridded format (0.25^{o} spatial resolution), and the ICRISAT crop production data uses political (district) boundaries, we harmonized the gridded CMIP6 historical and future data to ICRISAT district boundaries by apportioning each cell to the district polygon covering it (the polygon in which the centre of the cell lies). Partially covered cells' contribution to a polygon was weighted by area of cells in that polygon. For future crop yield predictions under different climate change scenarios, we divided the CMIP6 climate projections into four 20-year periods: 2021-2040, 2041-2060, 2061-2080, and 2081-2100. We then calculated the median value of each climate variable used in our models for each period, GCM, and SSP. For our study, we designated 2041-2060 as the short-term, and 2081-2100 as the

long-term future climatology. We followed the exact same procedure for the 1951-2000 historical simulation data to compute "reference climatology" (Ortiz-Bobea et al., 2019). We then ran our crop yield models with the short-term and long-term future climatologies and estimated the impact of climate change on crop yields by comparing these future yield predictions to yield estimates from models run using the 1951-2000 reference climatology.

4.3 Results

4.3.1 CMIP6 climate projections

In this section, we present and discuss the modeled trends of climatic variables under the different climate change scenarios to 2100. While we analyzed three crops in this study, we discuss future climate projections for only two in this section for simplicity: rice as a representative of the kharif season, and wheat from the rabi season. For illustration, cumulative plots depicting temporal trends of each climatic variable analyzed in this study, for a single example district, are available in Appendix C section C.2. In the next subsections, we discuss some salient climate parameters in more detail.

4.3.1.1 Temperature and growing degree days

Compared to reference climatology values, the seasonal mean of daily average temperature is expected to increase under all future SSP scenarios, with the degree of increase getting progressively higher for high emission scenarios (Figure 4.1). In the long-term (2081-2100) under SSP5-8.5, seasonal temperature in some parts of the country may increase by almost four and six degrees Celsius during the kharif and rabi season, respectively. Even with strict emissions reduction and carbon capture in SSP1-2.6, some regions will experience a warming of

up to two degrees by 2100. All these estimates are for the district-level median values of all 13 GCMs used in this study; some GCMs predict even higher impacts of climate change on growing season temperature in some areas by the end of this century (outside the range of Figure 4.1).

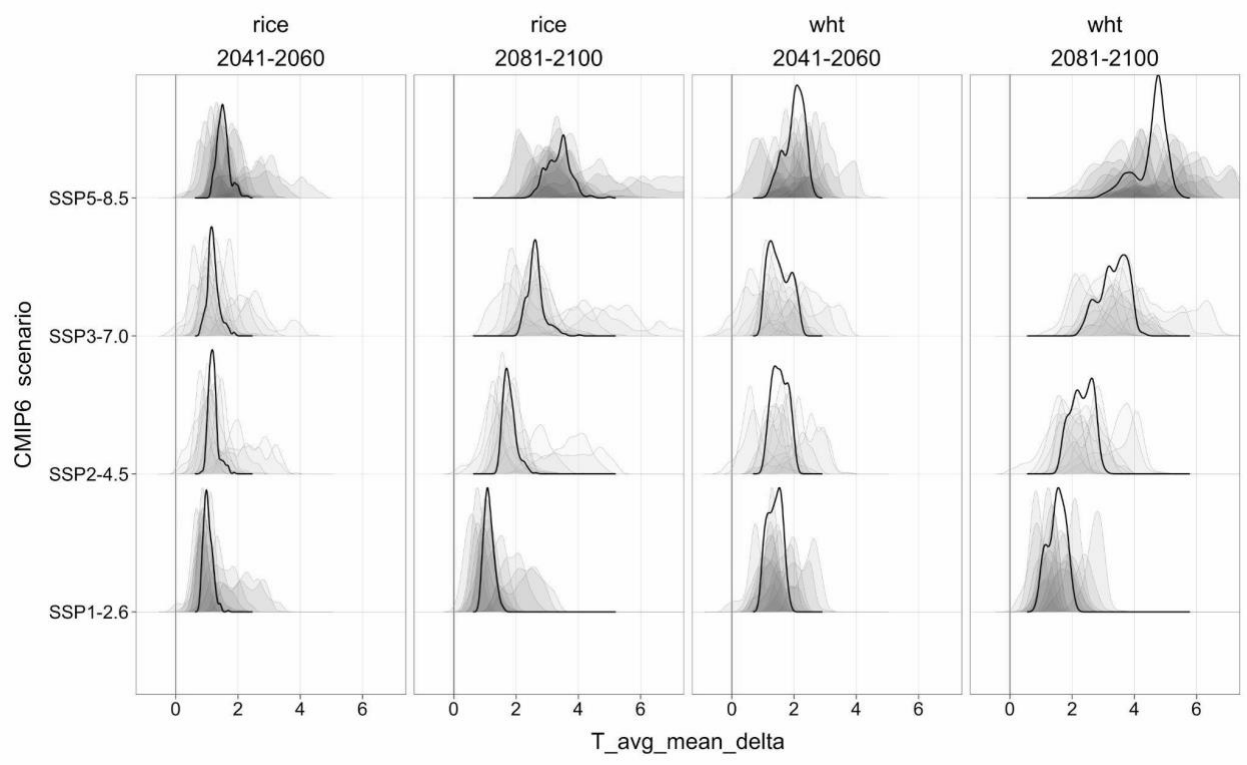

Figure 4.1 Distribution of district-wise increase in mean growing season temperature for representative kharif (rice; columns 1 and 2) and rabi (wheat; columns 3 and 4) crops, both for the short-term (2041-2060; columns 1 and 3) and long-term (2081-2100; columns 2 and 4). Different SSPs are depicted on the y-axis, ordered by intensity of emissions from bottom to top. The semi-transparent density plots depict distributions from each of the 13 GCMs analyzed in this study; the bold black line is the distribution of the median projection of the 13 GCMs for each district.

Along with daily average temperature, we also investigated projected changes in seasonal mean values of daily minimum and daily maximum temperatures (Figure 4.2 and Figure 4.3). The density distributions of the district-level increase in these two parameters show a bigger increase

in daily minimum temperatures compared to daily maximum, a trend seen consistently across time (near-term vs long-term), all CMIP6 scenarios, and the two primary crop seasons (Figure 4.2): the modes of the daily minimum temperature increase plots are always greater than those of the daily maximum temperature increase. Additionally, this relationship is also more or less maintained across the whole country as evident from the SSP2-4.5 and near-term maps in Figure 4.3: all regions in the daily minimum plots (left panels) are darker than their counterparts in the daily maximum plots (middle panels) for both rice (kharif; top) and wheat (rabi; bottom). On a similar note, the increase in seasonal temperature is higher during the winter rabi season (bottom right), compared to the summer kharif season (top right), especially in north and west India (Figure 4.3), a trend reported in past literature too (Almazroui, Saeed, Saeed, Islam, & Ismail, 2020). So not only is the diurnal temperature range (DTR) decreasing, but the temperature range between the two major crop seasons is also getting reduced in a changing climate. Nevertheless, this reduction in seasonal temperature range is less pronounced than the reduction in DTR.

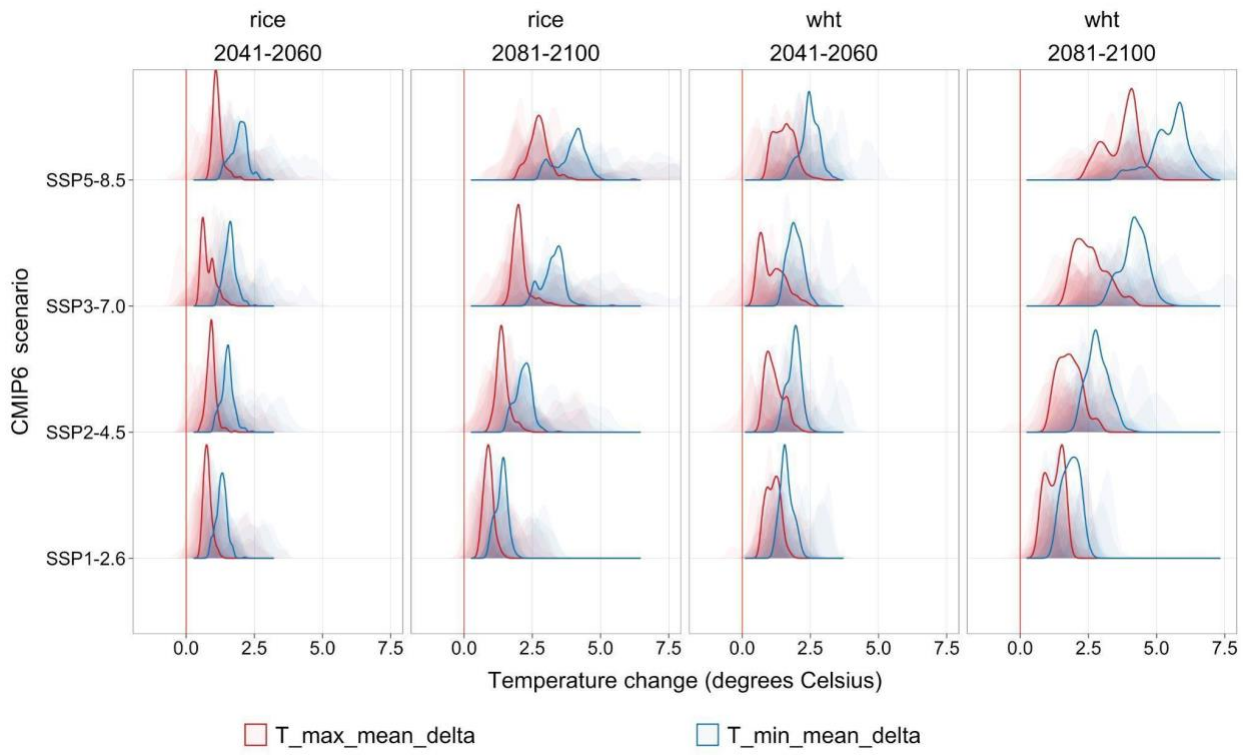

Figure 4.2 Distribution of district-wise increase in mean growing season daily minimum (blue) and daily maximum temperature (red) for representative kharif (rice; columns 1 and 2) and rabi (wheat; columns 3 and 4) crops, both for the short-term (2041-2060; columns 1 and 3) and long-term (2081-2100; columns 2 and 4). Different SSPs are depicted on the y-axis, ordered by intensity of emissions from bottom to top. The semi-transparent density plots depict distributions from each of the 13 GCMs analyzed in this study; the bold line is the distribution of the median projection of the 13 GCMs for each district.

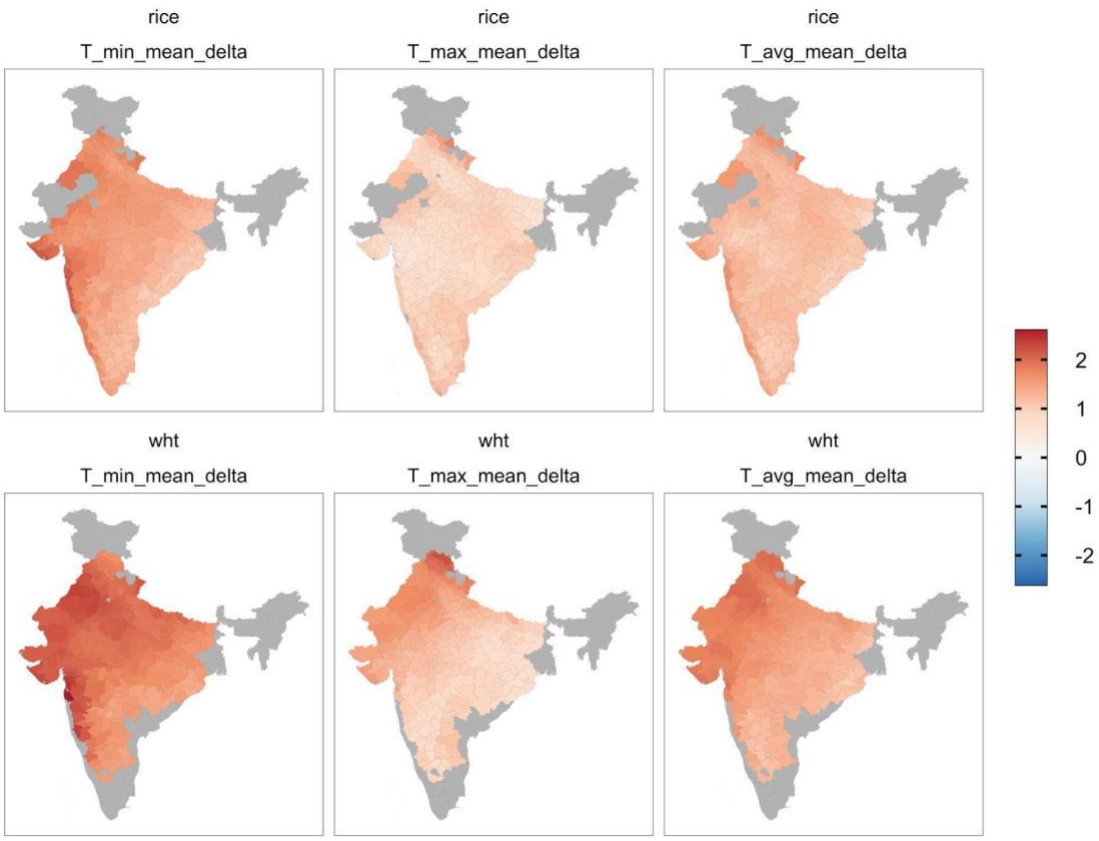

Figure 4.3 Increase (degrees Celsius) in mean growing season minimum daily temperature (left), maximum daily temperature (middle), and average daily temperature (right) for representative kharif (rice; top) and rabi (wheat; bottom) crops. Plots show results for near-term (2041-2060) changes relative to reference climatology for the SSP2-4.5 scenario. Median values of district-wise projections from the 13 GCMs were used to produce these plots.

4.3.1.2 Precipitation amount and variability

The future projections of precipitation amount and frequency (measured in terms of rain days) under various CMIP6 scenarios are less unequivocal compared to the consistent temperature trends discussed in the previous subsection. When measured in terms of the median predictions from all 13 GCMs, climate change is expected to increase the amount and frequency of summer

monsoon precipitation in all SSP scenarios, both in the short and long term (Figures 4.4 and C.6). However, there are some GCMs which predict a decrease in both these variables for some parts of the country in the summer (evident from the density plots lying on the left side of the vertical line in Figures C.6 and C.7). The median values, nonetheless, consistently predict an increase for all parts of the country (Figure 4.4). The trend in winter precipitation is more granular: while an increase in both the amount and frequency is expected in the western and central parts of the country, there are some districts in the eastern half which are expected to experience drier winters with reduced precipitation and fewer rain days over the rabi growing season (bottom row in Figure 4.4; columns 3 and 4 in Figures C.6 and C.7).

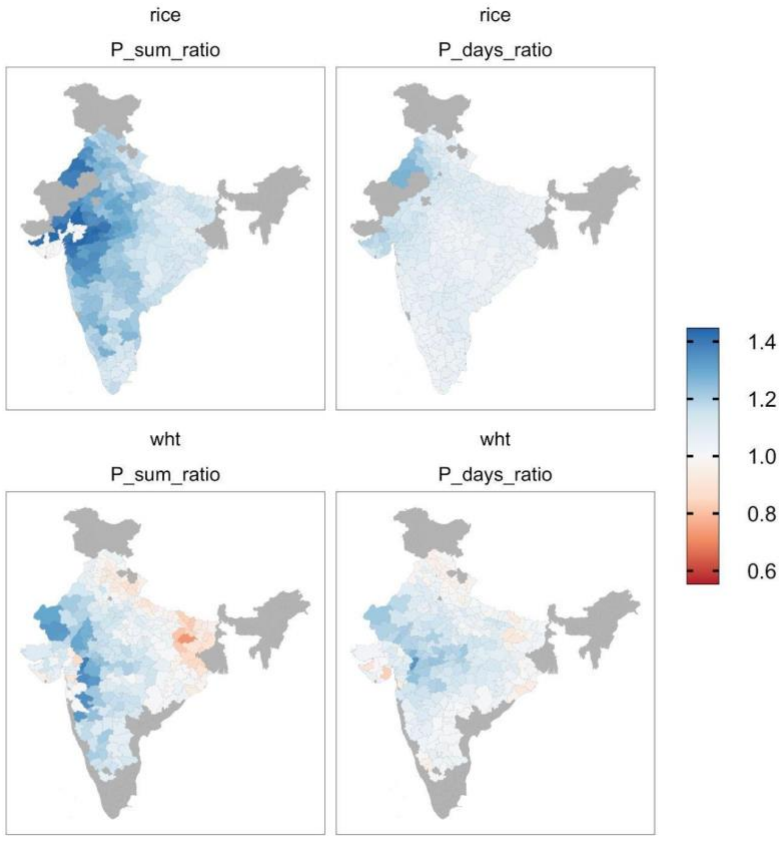

Figure 4.4 Total seasonal precipitation (left) and number of precipitation days (right) for representative kharif (rice; top) and rabi (wheat; bottom) crops. Plots depict the ratio of near-term (2041-2060) values under

SSP2-4.5 scenario to reference climatology. Median values of district-wise projections from the 13 GCMs were used to produce these plots. While there is consistent increase observed for summer monsoon precipitation (both in terms of amount and days), winter precipitation distribution shows both increases and decreases across different parts of the country.

4.3.1.3 Soil moisture variability

Of the five soil moisture availability bins analyzed (Appendix C section C.1), we compared the temporal and spatial trends of the two extremes: days when the crop's water demand is fully met ($\alpha = 1.00$), and days when less than 25 percent of the water demand is met ($\alpha < 0.25$). For the kharif crop season, soil moisture availability for crops is projected to increase with climate change, both in the short and long term for all SSP scenarios (columns 1 and 2 in Figure 4.5): the fraction of days under moisture stress decreases while there is a corresponding increase in fraction spent under sufficient moisture conditions. Rabi season depicts a similar trend, although the mutual proximity of the red and blue plots and their overlap with the zero line in Figure 4.5 shows that the reduction in soil moisture stress will not be as strong as in the summer season. We also analyzed this at a more local scale by picking Patiala, a district in north-west India, as an example, and plotting the change in the above-discussed fractions across the 13 GCMs and four SSP scenarios (Figure C.4). A majority of the 13 GCMs predict a decrease in water-stressed days and an increase in water-sufficient days over the kharif season.

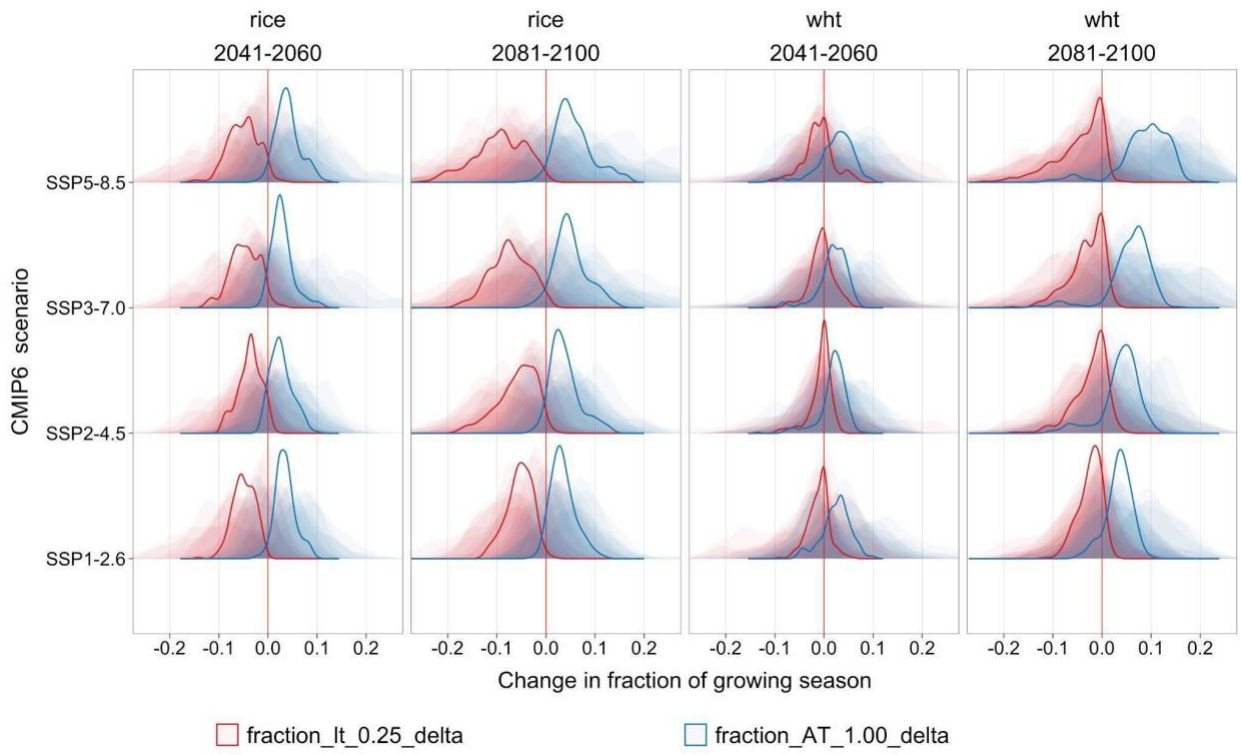

Figure 4.5 Distribution of change in district-wise fraction of growing season days spent at moisture availability (α) of 25 percent or less of actual crop water requirement (red) and at full moisture availability (blue) for representative kharif (rice; columns 1 and 2) and rabi (wheat; columns 3 and 4) crops, both for the short-term (2041-2060; columns 1 and 3) and long-term (2081-2100; columns 2 and 4). Different SSPs are depicted on the y-axis, ordered by intensity of emissions from bottom to top. The semi-transparent density plots depict distributions from each of the 13 GCMs analyzed in this study; the bold line is the distribution of the median projection of the 13 GCMs for each district.

In terms of spatial patterns, kharif crops are slated to benefit in almost all parts of the country from fewer days in the 25 percent bin and more time spent in the full water demand met bin (Figure 4.6). However, rabi crops in eastern parts of the country will witness more days of extreme soil moisture stress. In these regions, both the fraction of growing season spent at 25 percent soil moisture availability and duration spent under sufficient water availability is

expected to increase; the availability of water is expected to get more erratic, with a corresponding reduction in days spent under moderate soil moisture availability.

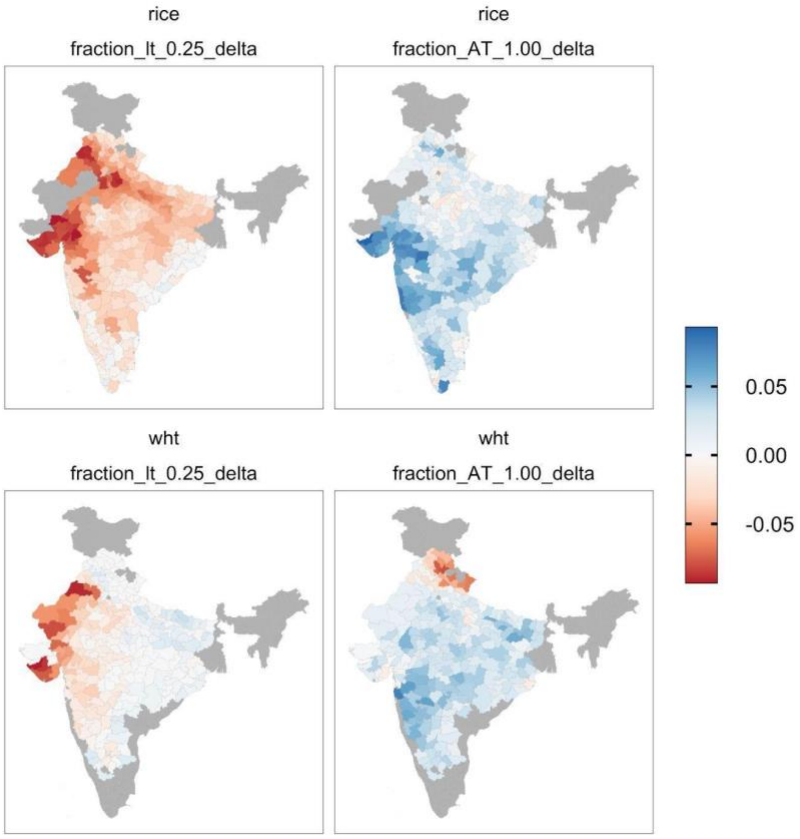

Figure 4.6 Fraction of growing season spent at moisture availability (α) of 25 percent or less of actual crop water requirement (left) and at full moisture availability (right) for representative kharif (rice; top) and rabi (wheat; bottom) crops. Plots depict the increase of near-term (2041-2060) values under SSP2-4.5 scenario to reference climatology. Median values of district-wise projections from the 13 GCMs were used to produce these plots. Note that increase in water availability is depicted by more intense reds in the left column and more intense blues in the right column.

4.3.2 Future crop yield predictions

4.3.2.1 Nationally aggregated results

We ran our statistical models for a range of different climate variable sets, of which we discuss only three here for clarity and comprehensibility: *seasonal*, *sub_prec*, and *sub_soil_moisture* (details available in section 4.2). Nationally aggregated impact of climate change on yield under different SSP scenarios is expected to be negative for almost all combinations of crops, statistical techniques, and climate variable sets (Figure 4.7). Only *sub_soil_moisture* LR models for pearl millet predict a net national positive impact of climate change; nevertheless, the magnitude of that is a lot smaller compared to the drop in pearl millet yield predicted by other models and climate variable sets. As climate continues to change in this century, its negative influence on crop yield is expected to grow continuously, with the biggest yield reductions predicted for the long term under the high emission SSP5-8.5 scenario. Comparing LR to BRT gives interesting results for pearl millet: while adding subseasonal climate variables reduces the predicted crop yield losses for LR models, BRT model predicts higher yield losses with subseasonal climate variables than the seasonal counterpart.

Within each subpanel in Figure 4.7, predicted climate change impacts vary considerably depending on the type of statistical technique or climate variables used. For example, under the moderate emissions SSP2-4.5 scenario in the near term (column 1 in Figure 4.7), the median impact on national pearl millet yield varies from -5.5 to +2.1 percent. Similarly, the variation of yield predictions across the 13 GCMs (depicted by the size of each boxplot in Figure 4.7) shows the utility of using multiple climate prediction datasets in order to cover a wider range of possible future scenarios. For instance, the long whiskers of the *seasonal* model (red) using

quadratic IV terms for pearl millet shows a yield change range of -16.3 to +0.1 percent, depending on the GCM used. This is even more prominent in the long-term predictions, where some boxplots have outliers quite far off from the median values. Consequently, all results and discussions about projected crop yields hereafter refer to the median from all GCMs' predictions of percent yield changes compared to reference climatology, unless otherwise stated.

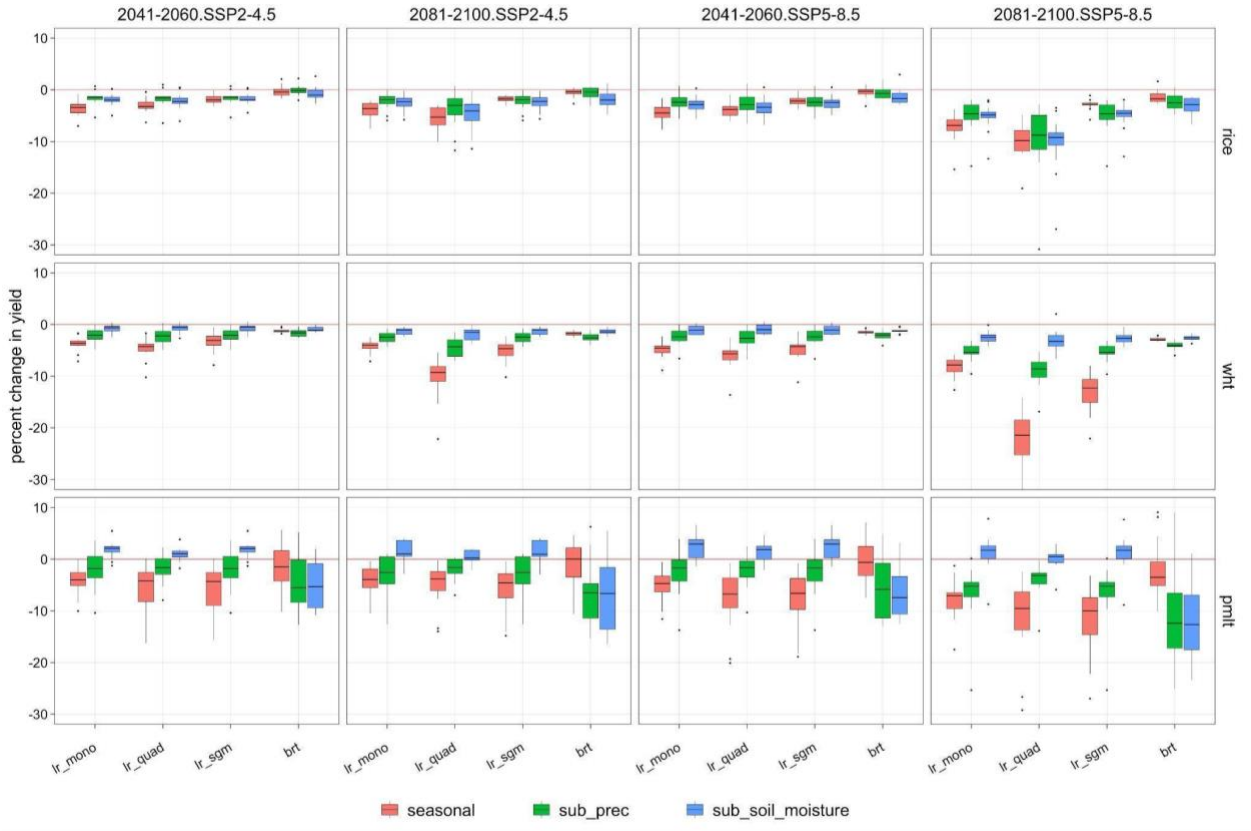

Figure 4.7 Nationally-averaged percent change in yield for rice (top), wheat (middle), and pearl millet (bottom). Columns 1-4 depict SSP2-4.5 near-term, SSP2-4.5 long-term, SSP5-8.5 near-term, and SSP5-8.5 long-term. The plots are color coded by climate variable set: *seasonal* (red), *sub_prec* (green), and *sub_soil_moisture* (blue). Within each panel, boxplots are grouped by model types from left to right: lr_mono, lr_quad, lr_sgm, and brt. Boxplots show median values of estimates from 13 GCMs.

In Figure 4.7, future yield change predictions are dependent on crop model decisions (climate variable sets and statistical technique used) as well as future climate projections (denoted by the 13 GCMs and four CMIP6 scenarios). To compare these drivers in terms of their influence on yield change predictions, we extracted Figure 4.7 data for SSP2-4.5 and ran a relative importance analysis using our chapter 2 methods, separately for the short-term and long-term (Table 4.2). This analysis of relative importance provides useful insights by ranking various factors by their capacity to explain the predicted crop yield changes. Since comparing all four model types would bias results towards LR, we picked segmented LR as a representative of LR models, and combined it with BRT results in this analysis.

For all three crops in the short-term, variability between GCMs is a significantly bigger driver of yield change, compared to the four CMIP6 scenarios; by 2100, this relationship inverts for rice and wheat while the gap narrows for pearl millet (rows 6 and 7 in Table 4.2). This shows that although the short-term effects of climate change on crop yields may be highly uncertain because of inherent variability in climate predictions from multiple GCMs, over time, the climate projections for various scenarios diverge and begin to play a bigger role in determining crop yields. The choice of variable set plays a substantial role in determining wheat yield change, while rice and pearl millet yield changes are less driven by variable choice. When comparing variability due to crop model choices (adding value pairs in rows 4 and 5 in Table 4.2) to that due to climate predictions (adding rows 6 and 7 in Table 4.2), all mid-century yield changes are more dependent on climate projections, than the crop model setup. One important result from Table 4.2 is that in the short-term, impact of CMIP6 scenarios is less prominent than the combined effect of climate variable or statistical technique choices. In other words, mid-century

yield predictions are more dependent on the model assumptions, than the CMIP6 scenario as realized by future emission trajectory.

Table 4.2 Relative importance of crop model choices (climate variable set and model type) and future climate projections (GCMs and CMIP6 scenarios) in determining percent yield changes in the future. Relative importance is measured in terms of percent of variance explained attributable to a particular variable of interest. Sample size of factors is provided in brackets. For model type, we chose segmented LR as a representative of LR models.

Factor of interest	Relative importance (percent of yield change variance explained)					
	Short-term (2041-2060)			Long-term (2081-2100)		
	Rice	Wheat	Pearl millet	Rice	Wheat	Pearl millet
Climate variable set (n = 3)	2	32	7	9	34	5
Model type (n = 2)	33	15	3	15	20	9
GCM (n = 13)	57	47	88	37	13	70
CMIP6 scenario (n = 4)	8	6	2	40	33	16
Total	100	100	100	100	100	100

4.3.2.2 Spatial patterns of climate change impacts on yields

With respect to the district-level rice yield predictions, the contrast between LR and BRT algorithms is interesting: while all three LR models predict a detrimental effect of climate change in most districts, the percent yield change estimates from BRT are more evenly distributed about

the zero line (Figure 4.8). This result holds for all three climate variable sets, although there is more spatial variability in LR predictions with the addition of subseasonal climate variables in *sub_prec* or *sub_soil_moisture* (Figure 4.9). In contrast to LR, the BRT models predict a boost to rice yield from climate change in some districts of India. These regions are mostly located in western and central India, covering large parts of the top five rice producing states (bold outline in Figure 4.9).

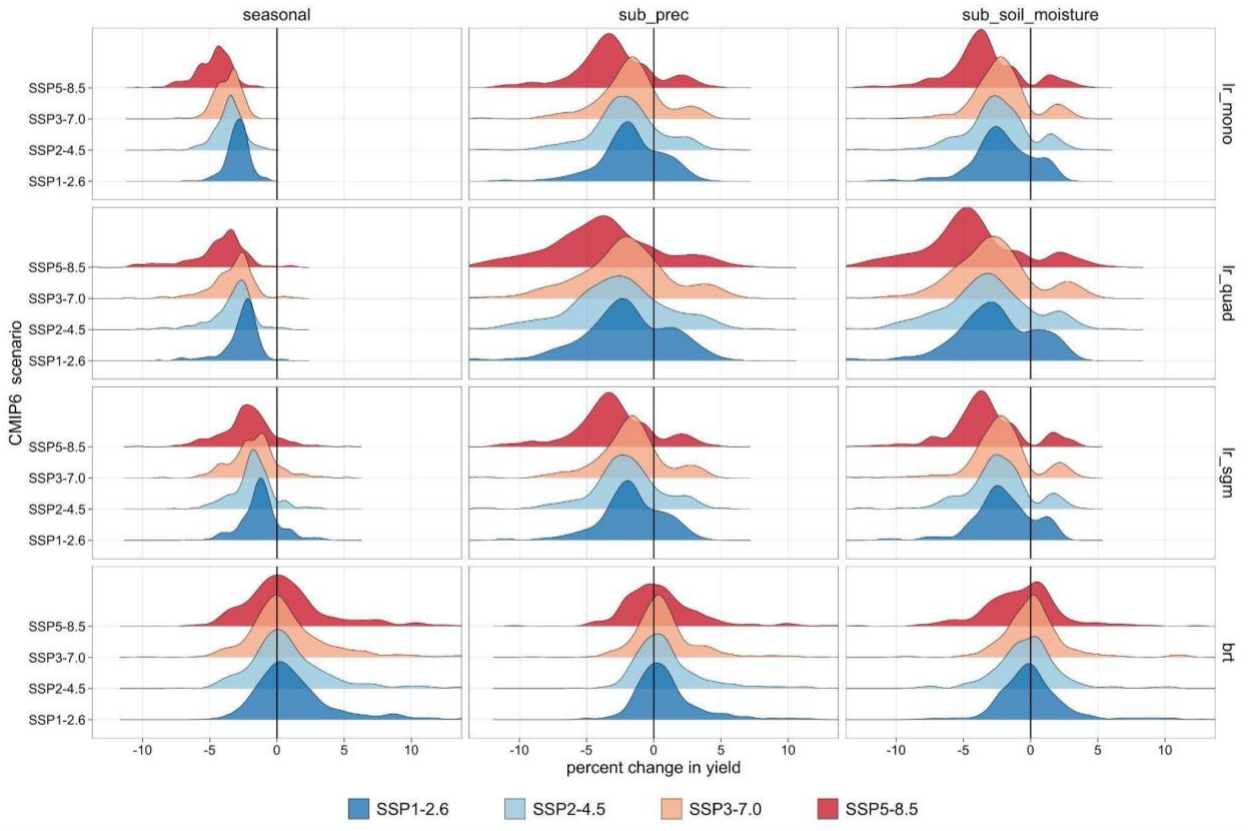

Figure 4.8 Distribution of district-level percent change in yield for rice in the short-term (2041-2060). Rows 1-4 show the four model types: *lr_mono*, *lr_quad*, *lr_sgm*, and *brt*. Columns 1-3 depict the climate variable sets: *seasonal*, *sub_prec*, and *sub_soil_moisture*. SSP scenarios are color-coded within each panel. Plots show median values of estimates from 13 GCMs.

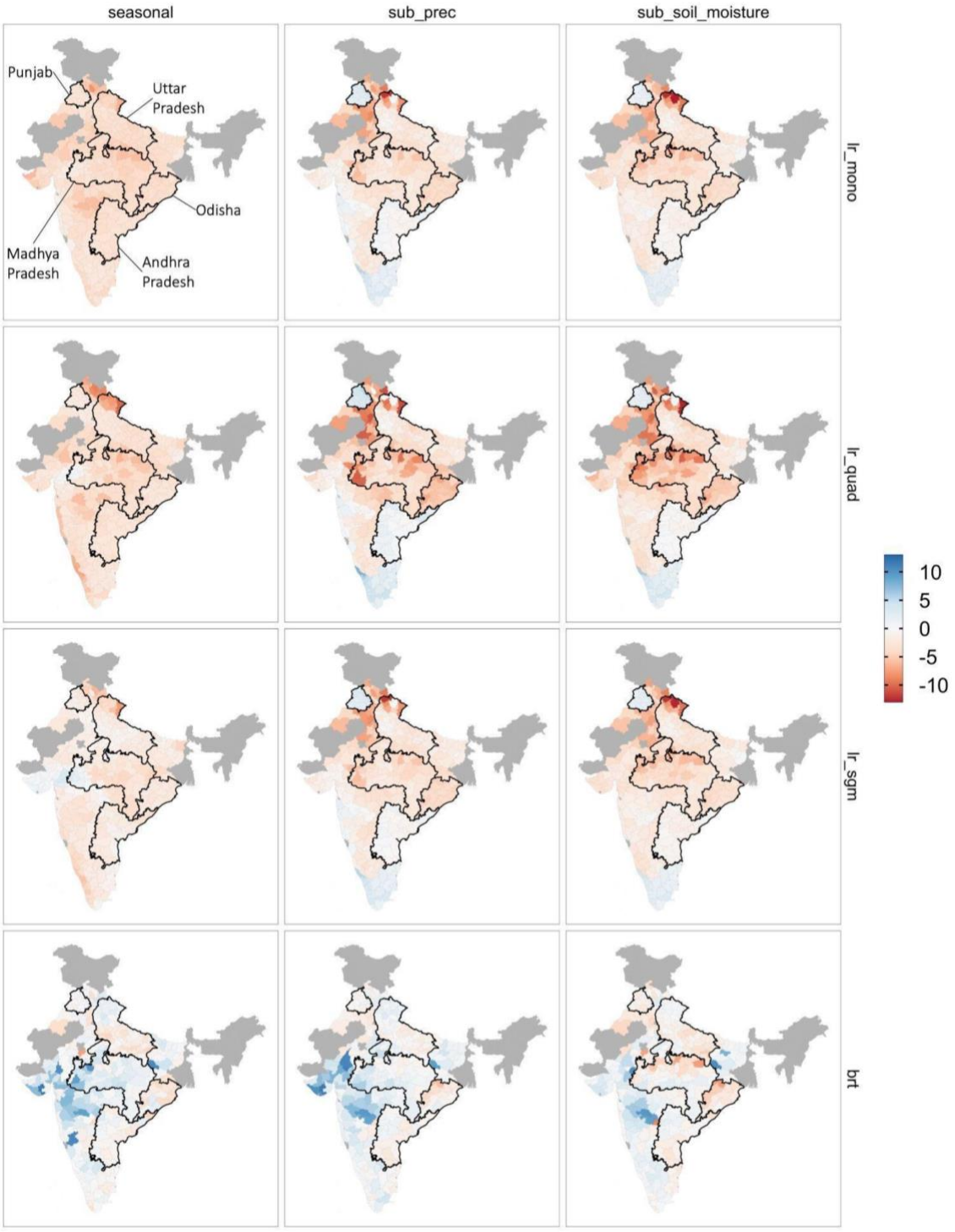

Figure 4.9 District-level percent change in yield for rice in the short-term (2041-2060) for the SSP2-4.5 scenario. Rows 1-4 show the four model types: *lr_mono*, *lr_quad*, *lr_sgm*, and *brt*. Columns 1-3 depict the

climate variable sets: *seasonal, sub_prec,* and *sub_soil_moisture*. Plots show median values of estimates from 13 GCMs. Five biggest rice-producing states are labelled and outlined in black.

Predicted changes in wheat yield are heavily influenced by climate variables included in the statistical model (Figure 4.10). All three *seasonal* LR specifications' forecasts are heavily biased towards predicting yield losses, except a few districts in north India where *lr_sgm* predicts marginally higher yields due to climate change (left column in Figure 4.11). Within these three, *lr_quad* and *lr_sgm* estimate yield losses of more than 10 percent in some districts of central India, including parts of Madhya Pradesh (an important wheat producer). However, LR models containing subseasonal climate variables (middle and right columns in Figures 4.10 and 4.11) estimate a more tempered influence of climate change, with some important wheat-growing regions expected to have comparatively higher yields according to the *sub_soil_moisture* LR model. On the other hand, the BRT model predictions are more agnostic to the choice of climate variables: all three BRT models estimate that most regions will face moderate (compared to LR predictions) wheat yield losses of not more than 5 percent.

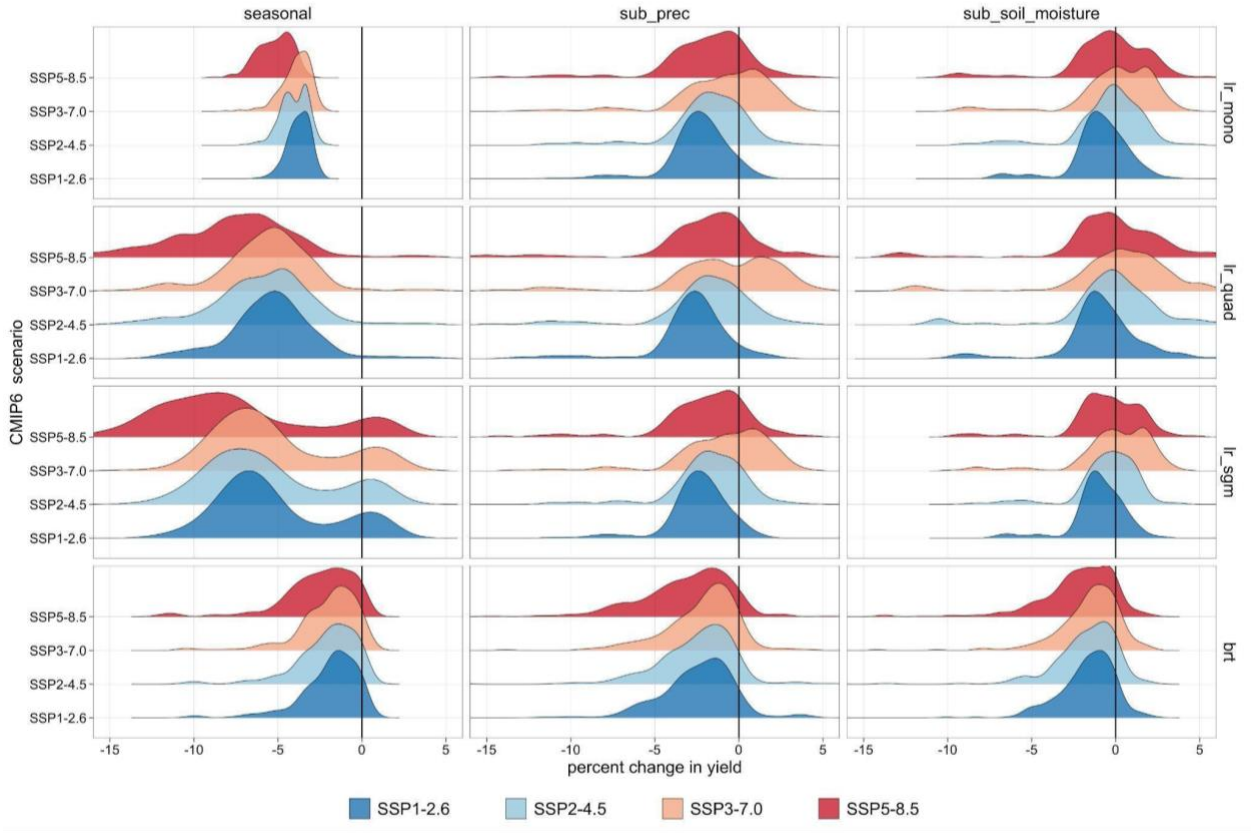

Figure 4.10 Distribution of district-level percent change in yield for wheat in the short-term (2041-2060). Rows 1-4 show the four model types: *lr_mono*, *lr_quad*, *lr_sgm*, and *brt*. Columns 1-3 depict the climate variable sets: *seasonal*, *sub_prec*, and *sub_soil_moisture*. SSP scenarios are color-coded within each panel. Plots show median values of estimates from 13 GCMs.

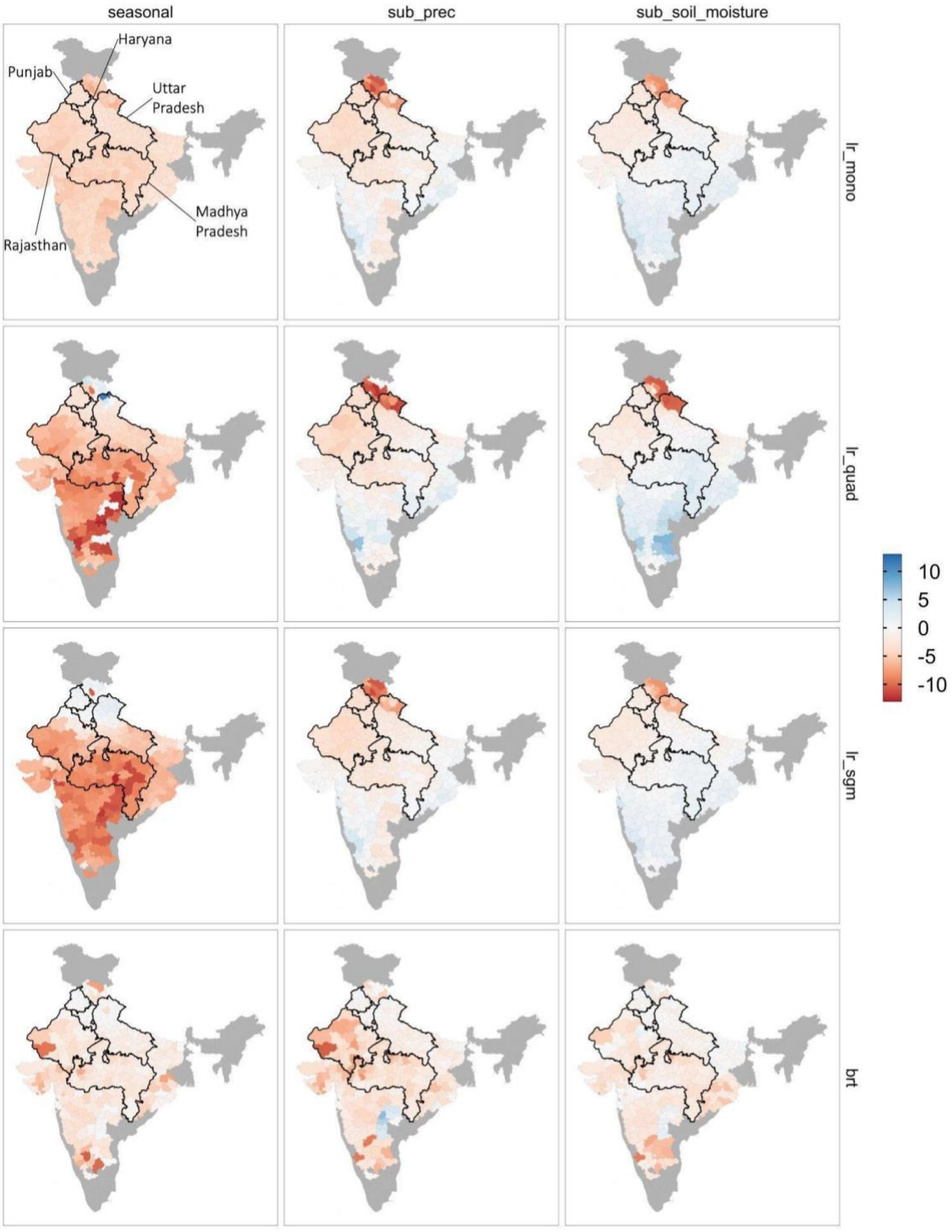

Figure 4.11 District-level percent change in yield for wheat in the short-term (2041-2060) for the SSP2-4.5 scenario. Rows 1-4 show the four model types: *lr_mono*, *lr_quad*, *lr_sgm*, and *brt*. Columns 1-3 depict the

climate variable sets: *seasonal, sub_prec,* and *sub_soil_moisture*. Plots show median values of estimates from 13 GCMs. Five biggest wheat-producing states are labelled and outlined in black.

Like wheat and rice, pearl millet also exhibits higher yield losses in the seasonal LR models compared to their subseasonal counterparts, *sub_prec* or *sub_soil_moisture* (Figure 4.12). Barring the largest producer state Rajasthan, the other four major pearl millet states are all expected to benefit from climate change under the SSP2-4.5 scenario in the short-term, according to all LR model estimates (Figure 4.13). However, the BRT model predicts heavy yield losses, over 20 percent in some regions under SSP2-4.5 by 2050. Geographically, the yield loss hotspot covers large parts of the top five pearl millet-producer states, thereby explaining the potentially catastrophic impact on national pearl millet production (Figure 4.7). It is worth noting that the *sub_prec* and *sub_soil_moisture* BRT models predict higher yield losses than *seasonal* BRT in parts of Rajasthan, Madhya Pradesh, and Uttar Pradesh (last row in Figure 4.13). This underscores the importance of accounting for subseasonal climate variability that may be missed if model selection is based on generic statistical metrics (refer to chapter 2 for more details).

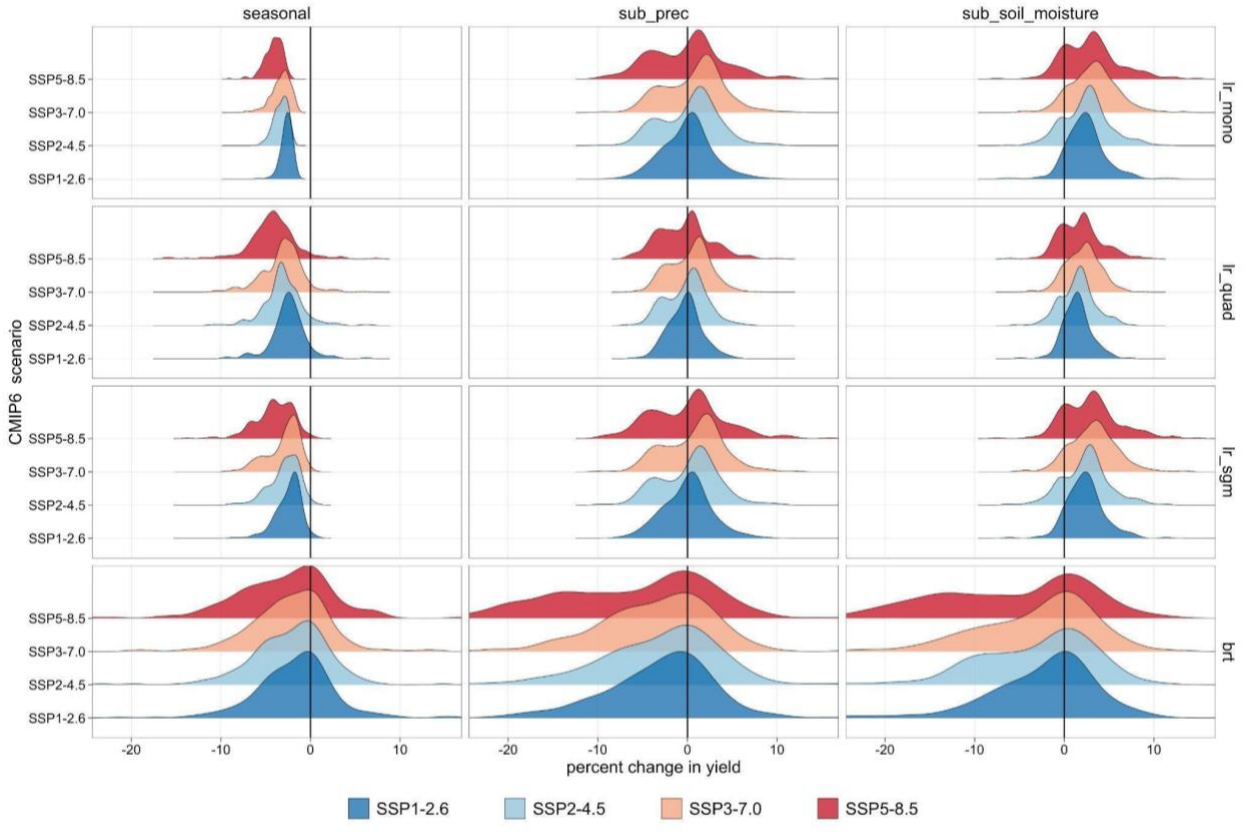

Figure 4.12 Distribution of district-level percent change in yield for pearl millet in the short-term (2041-2060). Rows 1-4 show the four model types: *lr_mono*, *lr_quad*, *lr_sgm*, and *brt*. Columns 1-3 depict the climate variable sets: *seasonal*, *sub_prec*, and *sub_soil_moisture*. SSP scenarios are color-coded within each panel. Plots show median values of estimates from 13 GCMs.

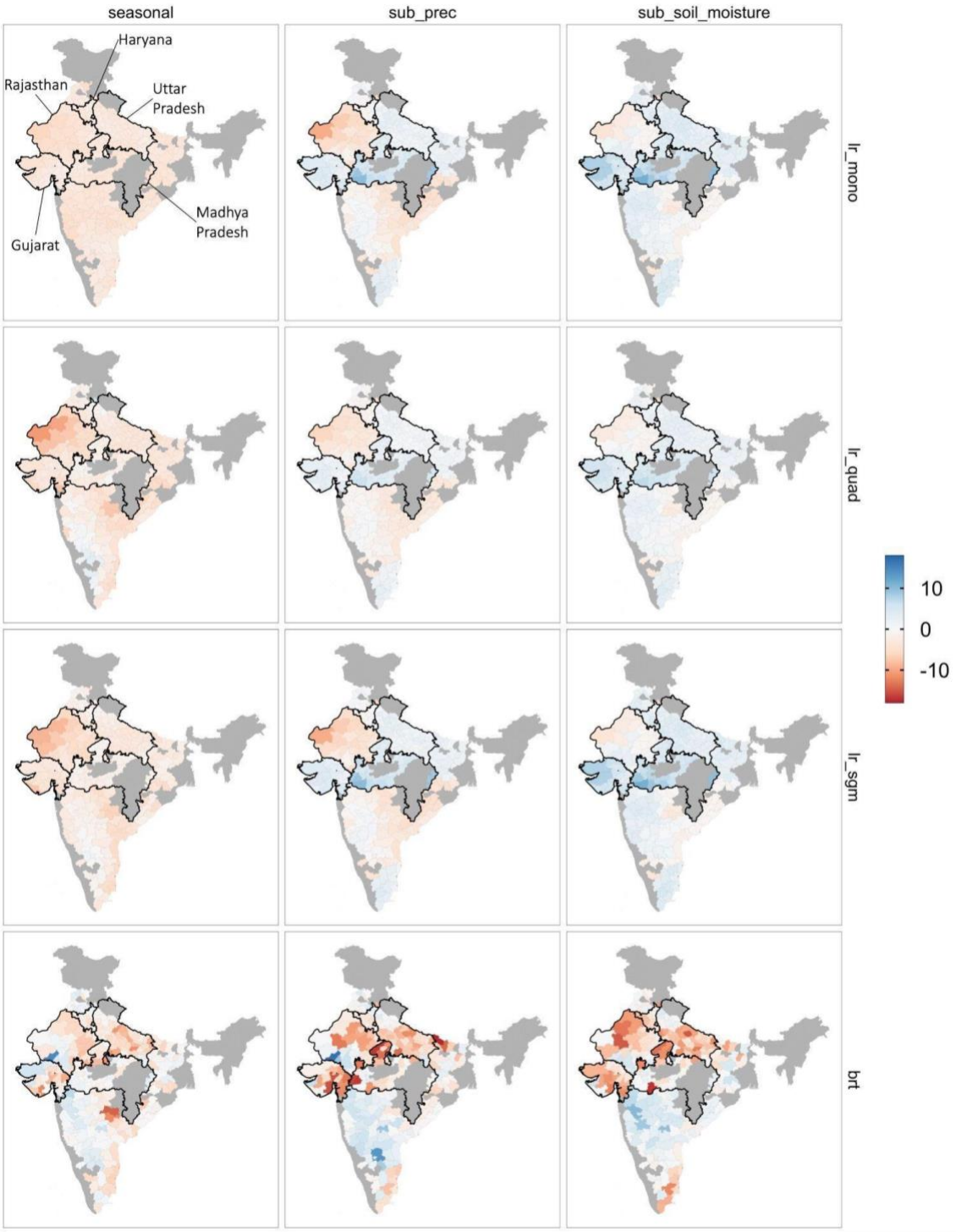

Figure 4.13 District-level percent change in yield for pearl millet in the short-term (2041-2060) for the SSP2-4.5 scenario. Rows 1-4 show the four model types: *lr_mono*, *lr_quad*, *lr_sgm*, and *brt*. Columns 1-3 depict the

climate variable sets: *seasonal*, *sub_prec*, and *sub_soil_moisture*. Plots show median values of estimates from 13 GCMs. Five biggest pearl millet-producing states are labelled and outlined in black.

4.4 Discussion

4.4.1 CMIP6 climate projections

In agreement with the vast climate change literature, our analysis of CMIP6 climate projections reconfirms a consistent and widespread impact of anthropogenic emissions on all climate parameters which have been repeatedly shown to influence crop yields. Compared to the reference climatology, every region of India is expected to get warmer under all four emission scenarios. Numerous studies have found a causal relationship between mean seasonal temperature rise and reduction in crop yields (Asseng et al., 2017; R. Gupta, Somanathan, & Dey, 2017; Lobell & Field, 2007). Simultaneously, the reduction in DTR due to faster rise in nighttime temperature as observed in our analysis has been studied and reported using past climate data at both global (Davy, Esau, Chernokulsky, Outten, & Zilitinkevich, 2017; Sun et al., 2019) and regional scales (Mall et al., 2021); while the latter study claims that DTR reduction across India started in the late 20th century, our results shows that it is expected to continue into the future. A topic of active current research, uneven warming between days and nights has been reported to have potential consequences for vegetation growth and crop yields. For example, Peng et al. (2013) observed a positive (negative) correlation between maximum (minimum) daily temperature and normalized difference vegetation index (NDVI, a vegetation greenness indicator) in most boreal and wet temperate regions; the relationship reversed in the drier temperate regions. Also, field experiments have shown up to 10 percent reduction in rice yields

for every 1ºC rise in nighttime temperature while daytime warming had no significant effect (Peng et al., 2004).

Climate change induced temperature rise will also lead to disproportionate changes in the amount of growing degree days accumulated in various bins discussed in this manuscript. For instance, a simple linear regression with mean temperature during rabi season as the IV, and the growing degrees accumulated in the 20-30 ºC bin as the DV, shows that the DV increases by around 14 percent for every 1 ºC rise in IV. The same analysis conducted with the >30 ºC bin exhibits a 63 percent increase with each 1 ºC increase in mean temperature. This can be extremely critical for crops with temperature thresholds during various growth stages that will begin to get crossed more frequently and for longer durations. For instance, the upper temperature threshold for wheat's anthesis stage has been reported to be 31 ºC (Porter & Gawith, 1999), and excess heat during that stage can reduce grain count leading to lower yields (Wheeler, Craufurd, Ellis, Porter, & Vara Prasad, 2000).

The median of all 13 GCMs' estimates depict an increase in kharif season precipitation (amount as well as precipitation days) in almost every part of India, although there is considerable variation among individual GCM predictions. This pattern has been reported by another study from India which assessed four separate CMIP6 GCMs and found contrasting impacts of climate change (on number of annual rain days) across most of India (V. Gupta et al., 2020). Aadhar & Mishra (2020b) also found high bias while simulating monsoon precipitation in many of the 16 GCMs they used, so we use the median (instead of mean) values of the projections from the 13 GCMs we used in this study. Nevertheless, our results are in agreement with previous studies

that also predict an increase in summer monsoon rainfall with rise in temperature (Almazroui et al., 2020; Katzenberger et al., 2020; Moon & Ha, 2020; Wang, Jin, & Liu, 2020). Winter season precipitation trends are more variable: while most areas are expected to witness more rain as well as rainy days, east Indian regions like the state of Bihar[9] will face precipitation decrease in the future. This can have huge implications for agriculture in this region, as we discuss in section 4.2.

Our soil moisture model uses daily evapotranspiration (ET) (calculated using temperature data) and precipitation as input. With climate change slated to increase both ET and precipitation in almost every part of the country, soil moisture's future trend would depend on which parameter is dominant. Our results show that in general most parts of the country will witness a reduction in the number of days during the growing season that kharif crops spend under critical soil moisture deficient conditions (Figure 4.6), along with an increase in duration of sufficient moisture conditions. Western regions will witness a bigger increase in soil moisture availability than the eastern districts. For the rabi season, on the other hand, there is a contrasting pattern of western India facing fewer soil moisture stress days, and parts of east India including Bihar, southern Uttar Pradesh, and the Himalayan region experiencing increased soil moisture stress. Our patterns for kharif and rabi seasons match the west to east reduction in soil moisture increase estimated in the future in other studies (Aadhar & Mishra, 2020b). This is in contrast to predictions for other regions of the world like southwestern US, eastern Europe, Mediterranean,

[9] We use 1966 state boundaries in our analysis. The state of Bihar, as referred to in this study, was bifurcated into Bihar and Jharkhand in 2000.

and southern Africa where soils are expected to get drier in the future (Green et al., 2019; Grillakis, 2019).

4.4.2 Crop yields in a changing climate

When using models to estimate the impact of climate variability on agricultural production, there are two major processes that need to be analyzed and quantified: the variability and long-term changes expected in certain climatic parameters, and the sensitivity of a crop to those changes. Relying on CMIP6 projections for the first part, this chapter built and ran an array of statistical models to uncover the second process and predict the influence of climate change on future crop yields.

Our results show that future changes in crop yields are dependent more on the type of statistical techniques or the set of climate variables used in the analysis, and less on the future climate scenarios, especially in the short-term (Table 4.2). For example, for pearl millet, the *seasonal* LR models predict a net drop in nationally aggregated yield, but the *sub_soil_moisture* LR models estimate a net gain (Figure 4.7), for both the SSP2-4.5 and SSP5-4.5 in the short as well as long term. When comparing the variability in yield loss prediction between different climate scenarios (the four density plots in each panel of Figures 4.8, 4.10, or 4.12) to variability in yield loss prediction across different statistical models (assessing a particular SSP scenario in panels of Figures 4.8, 4.10, or 4.12), it is again evident that yield predictions are highly sensitive to the statistical method or climate variables of the underlying model, in comparison to the climate scenarios. This bias in sensitivity of yield to crop models has been documented in previous studies too: The Agricultural Model Intercomparison and Improvement Project (AgMIP)

simulated wheat yield using various combinations of five GCMs and five crop models, and attributed 88 percent of variance in simulated yield values to variability between crop models, and only 10 percent to variability in the GCMs (remaining two percent was explained by interaction between crop models and GCMs) (Rosenzweig et al., 2013). In comparison, our results in Table 4.2 are less extreme because they relate to percent changes in yield, instead of actual yield values. When we conducted a similar analysis to Table 4.2 but with actual yield values, our relative importance results too apportioned over 90 percent of the variance in predicted yield values to crop model choices (climate variables and statistical methods), with only a minimal amount going to the climate projections (both inter-GCM variability and CMIP6 variability).

For LR models, accounting for subseasonal climate variability, either through subseasonal precipitation variables or the soil moisture model, predicts smaller reductions in nationally-averaged yields of wheat and pearl millet, compared to the simplest seasonal model (Figure 4.7). Limiting the model choice to a simple seasonal model, a convention still popular in literature, may thus lead to possibly erroneous estimates of climate change impacts. We also know from chapter 2 that subseasonal variables can be important for explaining crop yield models. This, coupled with the highly variable spatial results with subseasonal models discussed below, emphasizes the importance of building and intensively analyzing a suite of models before using them for future yield predictions.

There is a stark difference between the geospatial patterns of yield loss predictions from LR versus those from BRT models. The LR models for rice, for instance, predict a negative impact

of climate change in almost all parts of the country, while BRT predictions are distinctly positive in many parts, especially west India. Here, it is worth reiterating results from chapter 3 where LR models showed a high sensitivity of rice yield to temperature, while the BRT models' partial dependence plots depicted a lower sensitivity to temperature (Figure 4.14). Moreover, the proportional increase in summer precipitation is expected to be the highest in the western regions (Figure 4.4). This potentially explains why the BRT model predicts rice yield increases in west India, unlike the LR models. In other words, while one model (LR) may conclude temperature to be the primary driver of crop yield, the other (BRT) may predict that crop yields are most sensitive to changes in precipitation. The underlying model structure can then lead to different crop yield predictions for the same projected climate data. Spatially disaggregated results show an even greater sensitivity of yield outcomes to modeling choices. For all three crops, the simplest seasonal *lr_mono* model predicts a fairly uniform loss across all parts of the country. Because it contains seasonal variables in a monomial linear formula, the non-linear climate-yield response or subseasonal climate variability that can potentially influence yields is not accounted for. Hence the difference between the first panel and the rest in Figures 4.9, 4.11, and 4.13.

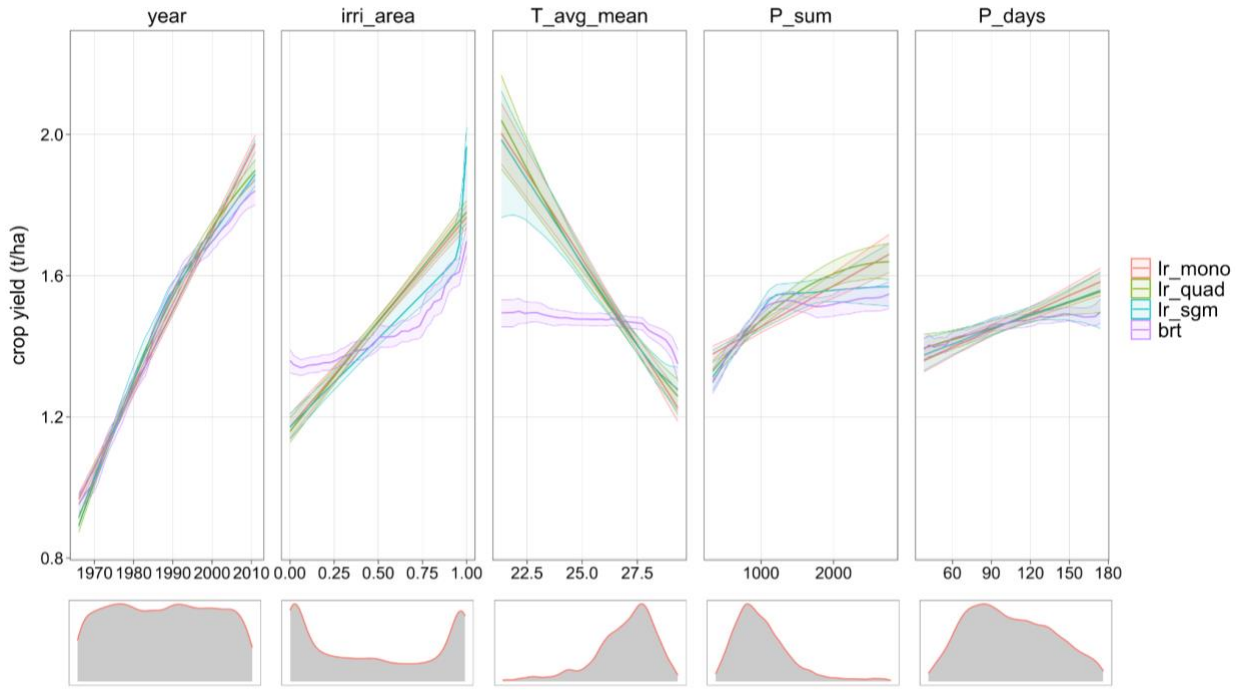

Figure 4.14 Partial dependence plots of the *seasonal* variable set models for rice (top row), and distribution density of corresponding IV in training data (bottom row). The four model types are color-coded in each panel. Data density plots at the bottom provide an idea of where most of the training data lies for a particular IV. For example, for the temperature partial dependence plot (panel 3), data is mostly concentrated at the higher end of the temperature range, hence the wide confidence intervals towards the left side of the partial dependence plot in that panel.

Overall, our results tell a cautionary tale for Indian agriculture. The nationally-average rice, wheat, and pearl millet yield could decrease by up to 3.4, 4.3, and 5.5 percent under SSP2-4.5 by 2050[10], and up to 5.3, 9.3, and 6.6 percent by 2100. At the local scale, there are hotspots that deserve immediate attention. Consider Rajasthan, the largest pearl millet producer state in India:

[10] These are the median of the estimates from all 13 GCMs. When assessing each GCM separately, the highest predicted yield reductions are 7.0, 10.2, and 16.3 percent for rice, wheat, and pearl millet.

all models, regardless of the statistical technique or underlying climate variables, predict reduced yields of the crop; losses could be over 20 percent in some districts by 2050 in the intermediate SSP scenario SSP2-4.5 (Figure 4.13). Similarly, wheat production in Rajasthan and Madhya Pradesh is expected to suffer as climate changes (Figure 4.11). Instead of exact yield predictions, our methodology and results can thus be thought of as a range of possible scenarios that highlight the risk, and which can then be used for formulating interventions.

4.4.3 Limitations and future work

Like all empirical model-based analyses, our study has some limitations that are worth mentioning explicitly. First, our models can only account for variation in IVs that we included. There are many other factors that may affect crop yield, including changes in CO_2 concentration, nutrient availability, farming practices (for example, tillage techniques and mulching), access to technology, among others. Second, we keep crop choices, geographical regions, and crop calendars constant over time. This may not always be true as farmers react to a changing climate or availability of resources like water (Jain et al., 2021). Using models built on historical data for making future predictions also ignores the adoption of improved cultivar varieties better-suited for the changing climate. Finally, a key assumption of our analysis for rice is the exclusion of the state of West Bengal, an important rice producer in India. Lack of irrigation data for the state in the ICRISAT dataset necessitated this.

In the future, within the realm of data availability, it may be worthwhile to extend our analysis to more crops, especially non-grains which have historically not been studied as much. While this study focused exclusively on the influence of long-term climate change on crop yield, the

vulnerability of agriculture to short-term anomalous events like droughts, extreme rainfall, or extreme heat events is an equally important topic because the frequency, intensity, and duration of such events is expected to change in the future (Aadhar & Mishra, 2020b; Das & Umamahesh, 2021; V. Gupta et al., 2020; Ha, Moon, Timmermann, & Kim, 2020). That would be a valuable complementary addition to this analysis.

4.5 Conclusion

Anthropogenic emissions are expected to continue to change global climate through this century. Our analysis of the latest CMIP6 climate projections shows that every region of India will experience a rise in temperature, with faster night-time warming reducing diurnal temperature range. Simultaneously, crops will experience progressively more degree days above threshold temperatures beyond which crop growth is adversely impacted. Precipitation patterns too exhibit important changes; while monsoons are expected to get wetter over time, some parts in eastern India may have to tackle drier rabi seasons in the future. All these changes in climate patterns are expected to have significant impacts on crop yields. Analyzing three major crops (rice, wheat, and pearl millet) from India, we found that nationally-averaged yields could reduce by up to 3.4, 4.3, and 5.5 percent (respectively) by the middle of this century, and by up to 5.3, 9.3, and 6.6 percent by 2100, under the intermediate SSP2-4.5 scenario. These nationally-averaged values hide much stronger spatial patterns; for example, some districts in the biggest pearl millet-producing state Rajasthan could experience up to 20 percent yield losses in the same time period. These estimates pertain to the intermediate “middle of the road” SSP2-4.5 scenario; the losses could be even higher if the higher emission scenarios become a reality.

Importantly, the impact of climate change on agricultural yield depends not only on the degree of change predicted in various climate parameters, but also the influence that climatic shift is expected to have on various crops. Given the complex processes and uncertainty involved in both these components, it is no surprise that predicting crop yields into the future is uncertain. An appropriate way to examine this ambiguity is to build multiple models, using different statistical techniques and assumptions about climate-crop relationships. For example, we found that the seasonal LR model predicts bigger reductions in nationally-averaged yields of wheat and pearl millet (and rice to a smaller extent), compared to the subseasonal precipitation and soil moisture counterpart. When combined with results from chapter 2 that show the value of including subseasonal variables, these patterns show that yield losses may be overestimated in some regions if subseasonal conditions are ignored.

Spatially, climate change influence on crop yields is expected to have high variation. For example, rice LR models predict negative impacts in almost all regions of the country, while BRT predictions are distinctly positive in parts of west India. This is primarily driven by the underlying functions fitted by the models: while LR attaches high sensitivity to rising temperature, the BRT algorithm fits a model that is more sensitive to increasing precipitation (and not temperature) resulting in diverging model predictions. Individually, each model has its unique weaknesses. Unique insights may thus be gained from combining yield predictions from multiple models with their individual strengths and weaknesses for estimating climate change impact on crop yields. While it was outside the scope of our study, complimenting statistical model predictions with results from process-based models or controlled experiments could be

extremely valuable. It is possible that most statistical models misinterpret or discount the impact of a certain climate variable that such an endeavor could help uncover.

Chapter 5: Conclusion

This dissertation is an in-depth examination of the under-the-hood workings of statistical crop models. Statistical models have gained increasing prominence over the past few decades as a tool for detecting and quantifying the response of agricultural productivity to climatic change. Researchers and stakeholders around the world rely heavily on these models for predicting the impact of climate change on crop yields into the future. Predicted crop sensitivity to various environmental factors (e.g., temperature, or water availability) may also be used in the selection of more suitable cultivars for particular growing conditions in specific places. Consequently, improving the accuracy and reliability of these models remains important.

In building statistical models, the researcher has many decisions to make, including the explanatory variables to include and their functional forms, or the structure of the models. In this dissertation, I examined how such modeler choices influence model predictions. The choices are getting wider daily with rapid advancement in modeling tools and the vastly increased availability of data for analysis. I hope my work will provide lessons that can help better inform model choices, instead of simply relying on intuition or adopting commonly used methods or variables.

I chose two major themes for cross-model comparison. In chapter 2, I examined the role various climate variables can play in outcomes of agricultural yield models. I found that the dummy variables that account for spatially-variable but time-invariant yield drivers such as soil quality, or the harvest year added as a proxy temporal yield variability, can capture part of the signal in

yield variability that would otherwise be explained by an omitted climate variable. On the surface, this may not seem like a cause for concern because the accuracy of the two models, measured using standard and popular statistical metrics like R2 or RMSE, may not be significantly different. However, automatic model selection based on parsimony criteria may exclude important climate variables from the model, leading to poor predictions of the impact of climate change. I demonstrated this result using actual historical crop and climate data from India. I also proposed another statistical metric, relative importance, that has hitherto not been used widely in the crop modeling field, and demonstrated its utility in identifying important climate variables that may otherwise be missed had I solely relied on the standard model performance metrics. The chapter's most important recommendation for researchers was to use statistical metrics in combination with theoretical or process-based knowledge for choosing climate variables to include in their crop models; a variable may warrant inclusion even if it exhibits no apparent improvement in model accuracy.

In chapter 3, I conducted a detailed comparison of two distinct statistical techniques for building the aforementioned crop yield models. OLS linear regression, or LR, has the advantage of simplicity and being the most commonly used method in this field. Because they have been used so extensively, conditions which influence LR models' performance are well understood by the greater scientific community. Nevertheless, machine learning (ML) algorithms have some clear advantages; they are often more flexible and make fewer or no prior assumptions about the functional form of the model. The past few years have witnessed a rapid rise in the availability of high quality open-source climate and crop yield data, advanced computational facilities, and

easy-to-use statistical software. The analysis of climate-crop relationships can benefit immensely from ML tools.

In this dissertation, I chose boosted regression trees (BRTs) as an example of ML (James, Witten, Hastie, & Tibshirani, 2013). The results from chapter 3 presented context-dependent advantages and disadvantages of LR and BRT approaches. The mathematical equations of LR are easily understood, and the model fitting process is faster and less computationally-intensive compared to more advanced algorithms. BRTs are significantly more accurate in terms of yield predictions, and can be really powerful at fitting flexible functions where the relationship between yields and climate variables is suspected to vary widely over geography and/or time. BRTs can also capture obscure interactions between IVs that may otherwise be missed by LR model specifications. Some of the benefits that BRTs provide automatically, can also be derived using LR, though this needs manual tweaking of the model, and a priori knowledge of, and assumptions about, functional forms. Even if the latter is available, adding complexity to LR may reduce statistical power when there are multiple climate variables being analyzed simultaneously, and the output of such a model gets unwieldy. However, BRTs too have their Achilles' heel: I show through an example that they may conflate the effect of two correlated variables and can lead to potentially erroneous interpretations.

Chapter 4 built on the findings of chapters 2 and 3. I developed an array of models for three major crops in India (rice, wheat, and pearl millet), and applied them to CMIP6 climate projections for predicting climate change impact on crop yields till 2100. I found that nationally-averaged yields of rice, wheat, and pearl millet could reduce by up to 3.4, 4.3, and 5.5 percent

(respectively) by the middle of this century, and by up to 5.3, 9.3, and 6.6 percent by 2100, under the intermediate SSP2-4.5 scenario. These nationally-averaged values hide much stronger spatial patterns; for example, some districts in the biggest pearl millet-producing state Rajasthan could experience up to 20 percent yield losses in the same time period. The processes governing both the future climate trajectories and their influence on various crops, are complex and fraught with uncertainty. My results show that future changes in crop yields are dependent more on the type of statistical techniques or the set of climate variables used in the analysis, and less on which of the CMIP6 climate scenarios is realized. Therefore, I discourage researchers from trying to predict absolute yield values with a single mechanism (or model). Instead, a more appropriate way is to build multiple models, using different statistical techniques and assumptions about climate-crop relationships. I suggest combining yield predictions from these models using a probabilistic, rather than a deterministic, approach for estimating climate change impact on crop yields.

While chapters 2, 3, and 4 are standalone analyses with independent objectives, methods, and results, there is a significant link between their findings. Essentially, my overall inference is that the application of statistical models for predictions should be accompanied by a more thorough inspection of the assumptions, underlying mechanisms, and possible alternative model specifications. Chapter 2 showed that automated model selection based on principles of parsimony or standard statistical metrics may discount the contribution of potentially important climate variables. Similarly, chapter 3 showed that no one statistical technique can be considered the panacea for predicting crop yields as a function of climate. For example, the possible conflation between time and climate by BRTs may lead researchers to underestimate the

potential losses in crop yields due to climate change. An ensemble of multiple techniques, while admittedly adding more uncertainty to the predictions, may be a more realistic representation of the inherent complexity of the climatic and agricultural processes. Chapter 4 applies lessons from chapters 2 and 3 to develop a more nuanced understanding of Indian agriculture's vulnerability to climate change.

5.1 Application of my models: an irrigation expansion case study

This dissertation furthers our understanding of statistical models, and the factors or parameters that the predictions are contingent on. There are vital lessons to be learnt in terms of not accepting model outputs without a thorough examination of the underlying mechanisms. This aspect of the dissertation's contribution has already been discussed in great detail.

A second contribution of the dissertation is the development of models that can be used by policymakers and stakeholders for assessing the risk from climate change to various crops in their regions of interest, and for designing interventions to reduce the vulnerability of crops to climatic variability. I illustrate this using a short case study below.

Irrigation is among the most important enablers of yield growth across the world and numerous studies show the benefits of irrigation for crop production (Li & Troy, 2018). The critical role irrigation currently plays in agricultural production can be gauged from the fact that while only 20 percent of global cultivated land is under irrigation (Rockström et al., 2007), it accounts for 43 percent of the global cereal production (Siebert & Döll, 2010). Neumann, Verburg, Stehfest, & Müller (2010) combined econometrics with spatial yield analysis to estimate yield gaps of

wheat, maize, and rice. They observed that irrigation intensity not only played a statistically significant role in decreasing yield gaps globally, but also explained spatial yield variation in five out of six world-regions analyzed in their study. A global yield gap analysis reported that just with improved irrigation, 16 percent of the underachieving regions could bring their yields to within 75 percent of the attainable yields (Mueller et al., 2012). According to Lobell, Cassman, & Field (2009), while yields in most irrigated regions are close to maximum yields observed worldwide, rainfed yield gaps are on average 40 percent or higher.

In addition to improving yields, irrigation also plays a key role in building crop resilience to climate variability. Li & Troy (2018) reported that irrigated corn yields in the US exhibit lesser sensitivity to climate variability compared to rainfed yields. Fishman (2018) observed similar results from their regression of India's rice yields against cumulative precipitation and total number of rainy days over the season. The Economic Survey of India 2017-18 (Ministry of Finance, 2018) predicts future declines in crop yields in rainfed regions due to extreme temperature (up to 7.6 percent) and drought events (up to 14.7 percent) to be more severe compared to irrigated areas (up to 3.0 and 6.2 percent, respectively). Hence, there are good reasons to believe that climate change will have drastic consequences for farmers without access to irrigation, and that the expansion of irrigation can play a crucial role in increasing agricultural resilience to climate change. This is all the more relevant for India where the government continues to encourage and invest in the expansion of irrigation facilities to currently rainfed regions, most recently as a part of its flagship irrigation scheme called Pradhan Mantri Krishi Sinchayee Yojana (PMKSY) which has set an ambitious target of increasing net irrigated crop area from the current level of 45 percent to 100 percent (Ministry of Water Resources, 2017).

One option for targeted irrigation expansion is to identify the proverbial low-hanging fruit: regions or crops most vulnerable to climate change and poised to benefit the most from irrigation. These regions or crops could then be prioritized by the decision makers for irrigation development. Here I chose pearl millet for my case study because it is the least irrigated crop of the three analyzed in this dissertation (Figure 5.1). Note that I used the proportion of area irrigated (ratio of irrigated area to harvested area) for each crop-year-district combination as a proxy for irrigation availability; this metric varied from 0 (fully rainfed) to 1 (all crop area irrigated).

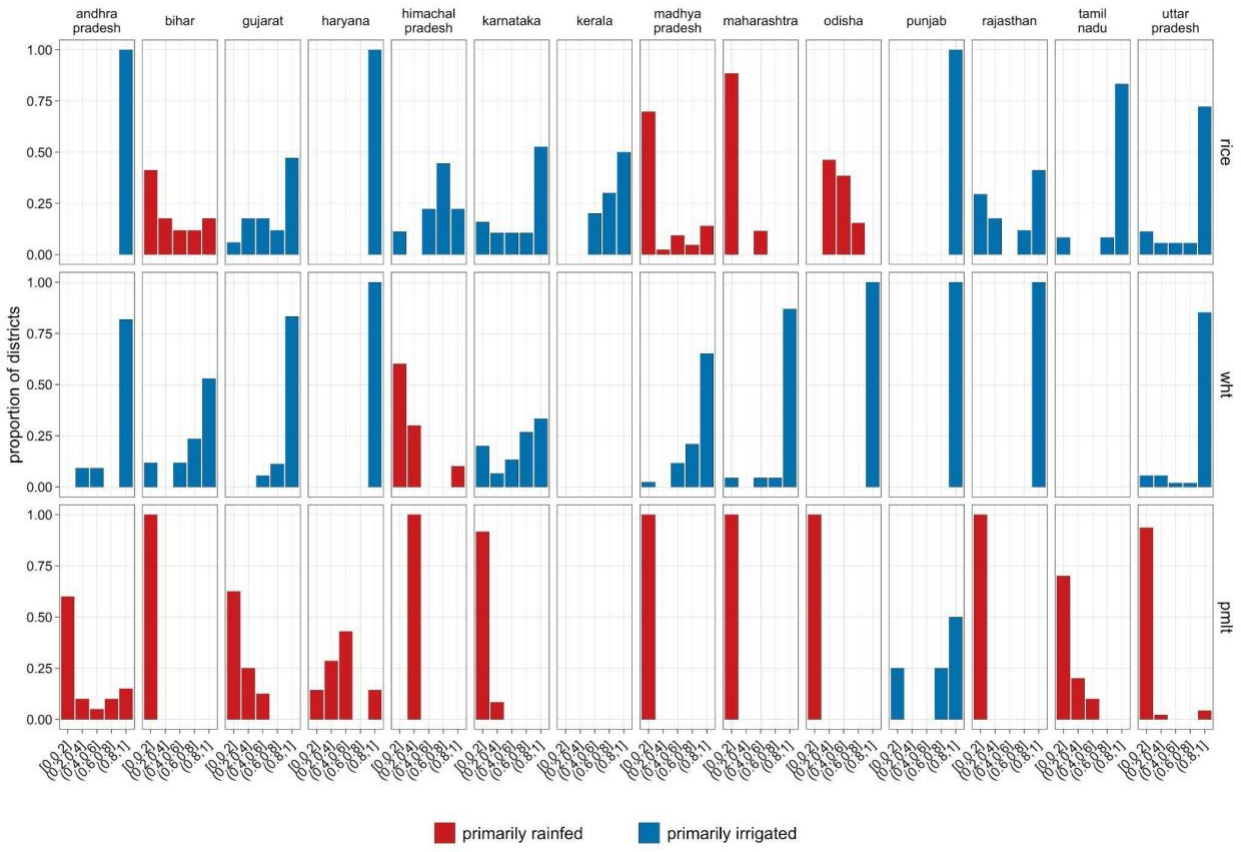

Figure 5.1 Frequency distribution (proportion of total districts) of each state in different irrigation categories. Red bars denote states where irrigation access is available to more than 50 percent of a crop's area; blue denote states where majority (more than 50 percent) of the area is rainfed.

I focus on just the top five pearl millet producer states, and only discuss results from BRTs here for concision. I further filtered districts which met the following criteria:

1. BRTs predicted yield losses for the 2041-2060 period under the "middle of the road" SSP2-4.5 scenario, and
2. median irrigation access during the last 5 years of my data's time range was less than 50 percent.

I re-ran my models with irrigation access increased to 50 percent for all districts, and everything else (climate, crop area, crop calendar) identical to the "no irrigation expansion" scenario. I chose the 50 percent threshold (instead of a higher number like 100 percent irrigation access) to keep my analysis more realistic considering 2050 is just 20 years away and irrigation expansion has multiple practical constraints like water access, setup and maintenance costs, or administrative burden. The latter are obviously extremely important while planning irrigation projects, but my case study is a model experiment to gauge the potential benefits of irrigation as and when the government brings it to farmers currently practising rainfed agriculture.

In my analysis, the difference between the yield loss predicted under the "no change in irrigation" scenario and the "50 percent irrigation access in each district" scenario signifies the yield loss that can be avoided by expanding irrigation to currently rainfed pearl millet production in these districts (Figure 5.2). Madhya Pradesh and Rajasthan stand out as the two states where pearl millet production will gain the most from irrigation expansion. In some districts of these two states, the difference between percent yield losses if irrigation remains constant, and if irrigation access were to increase to 50 percent, is more than five percentage points (for example, the dark blue dot in Rajasthan column in Figure 5.2 denotes a change in yield loss from 30 percent to 20 percent with irrigation; in other words, yield loss there reduced by 33 percent). It can be reasonably argued that within the practical constraints of such an endeavor, providing irrigation to pearl millet farmers in the most vulnerable districts of these two states would provide the biggest return in terms of reducing yield vulnerability to climate change. The other three states of Gujarat, Haryana, and Uttar Pradesh are not expected to benefit as much from

irrigation expansion, mostly because many districts in these states already have some irrigation facilities (Figure 5.1).

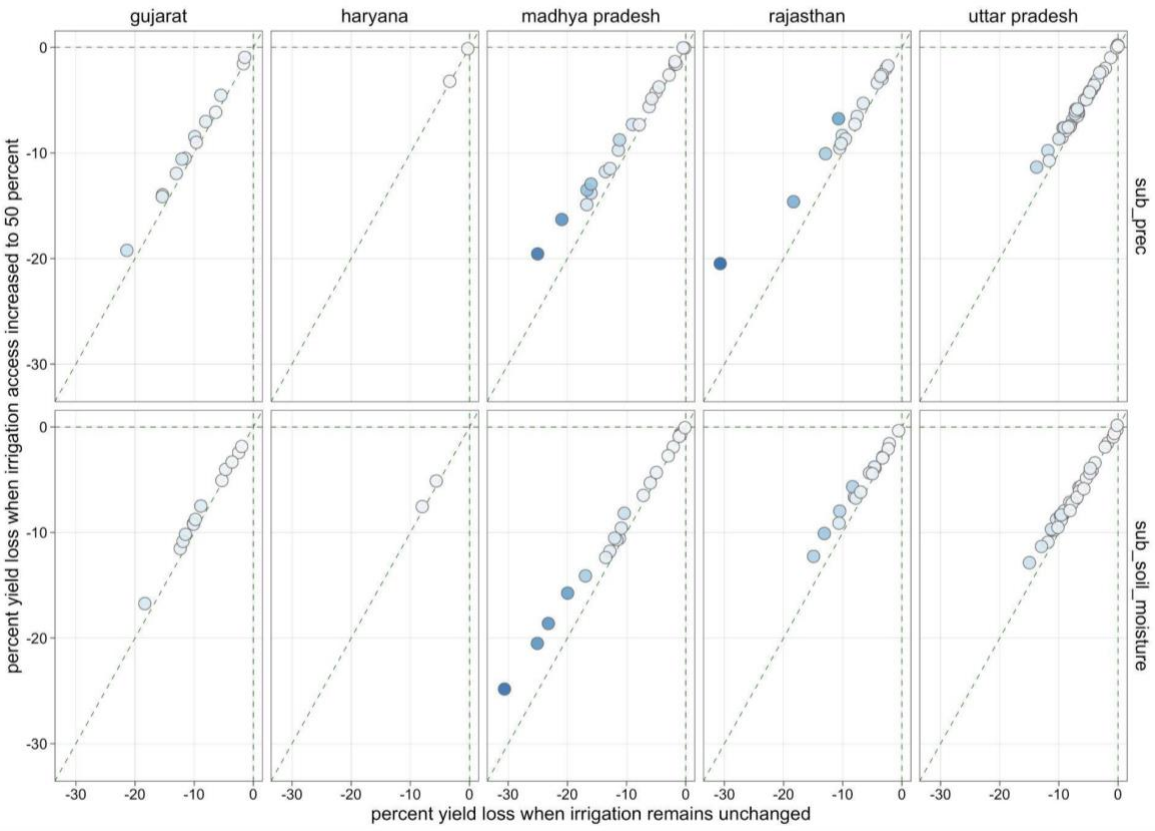

Figure 5.2 District-wise yield loss predictions under "no irrigation expansion" scenario versus those under "50 percent crop area with irrigation access" scenario. Higher the deviation from the y=x line, the bigger the impact of irrigation expansion.

Nevertheless, I reiterate that interventions like the one described above have to be assessed and designed within practical constraints and boundaries. For example, irrigation expansion may not be possible in areas with already stressed water resources. Expanding irrigation into hitherto rainfed regions in India may place additional stress on groundwater reserves of aquifers currently facing depletion (Zaveri et al., 2016). Extreme spatial heterogeneity in groundwater access

across India has led to the development of regional hotspots which have witnessed undesirable outcomes in the form of overexploited aquifers. Currently, annual groundwater extraction exceeds natural recharge in 1186 out of 6881 blocks (groundwater observation units) in India (Central Ground Water Board, 2020). More importantly, most of these areas are concentrated in two regions (Figure 5.2). In the northwestern states of Punjab and Haryana, an epicenter of India's Green Revolution, groundwater use exceeds natural recharge by 66 and 37 percent, respectively (Central Ground Water Board, 2020). Recent studies have reported that parts of India with critically depleted groundwater resources may lose up to 68 percent of their cropped area if they lose access to groundwater; surface irrigation will not be able to fully substitute groundwater in these places (Jain et al., 2021). In these cases, other adaptation strategies like improving irrigation efficiency, tweaking cultivation practices, or adopting drought-resistant varieties may be needed. So, there is a need to understand the socio-economic drivers of groundwater exploitation before the government embarks on an ambitious mission to expand groundwater access lest it manifests in faster depletion of the already stressed resources. In the most critical areas, farmers may even need to be incentivized to shift to less water-intensive crops or compensated for reducing their groundwater use (Sidhu, Kandlikar, & Ramankutty, 2020).

5.2 Limitations and future work

Like all empirical model-based analyses, my study has limitations that warrant discussion. The agricultural data was acquired from ICRISAT, which collected and compiled it from multiple sources. This data has been used extensively in literature (Birthal, Khan, Negi, & Agarwal, 2014; Davis, Chhatre, Rao, Singh, & Defries, 2019; Zaveri & Lobell, 2019), and I do not doubt its

overall quality. Nonetheless, any errors or inconsistencies in the data would affect my results. I noticed some oddities that are worth mentioning here. ICRISAT reports district-level crop production and area; I took the ratio of those to calculate yield, and use it in all my modeling analysis. To investigate yield values for any discrepancies, I binned them at 0.01 t/ha intervals, and plotted the number of observations in each bin. Ideally, this plot should have smooth transitions from one bin to the next; there is no reason for there to be disproportionately more yield values in a certain bin compared to its immediate neighbors. However, I observe spikes at certain values, specifically multiples of either 0.10 or 0.25 t/ha (Figure 5.3). This can only occur if a disproportionate number of reported production and area values are multiples of each other by factors of 0.10 or 0.25. I suspect there is some rounding off and guesstimation occurring when district-level numbers are recorded by on-ground officials. I would like to clarify here that these sampling or data entry errors are not unique to the ICRISAT dataset; they are bound to creep in when real-life data is collected and compiled from multiple sources.

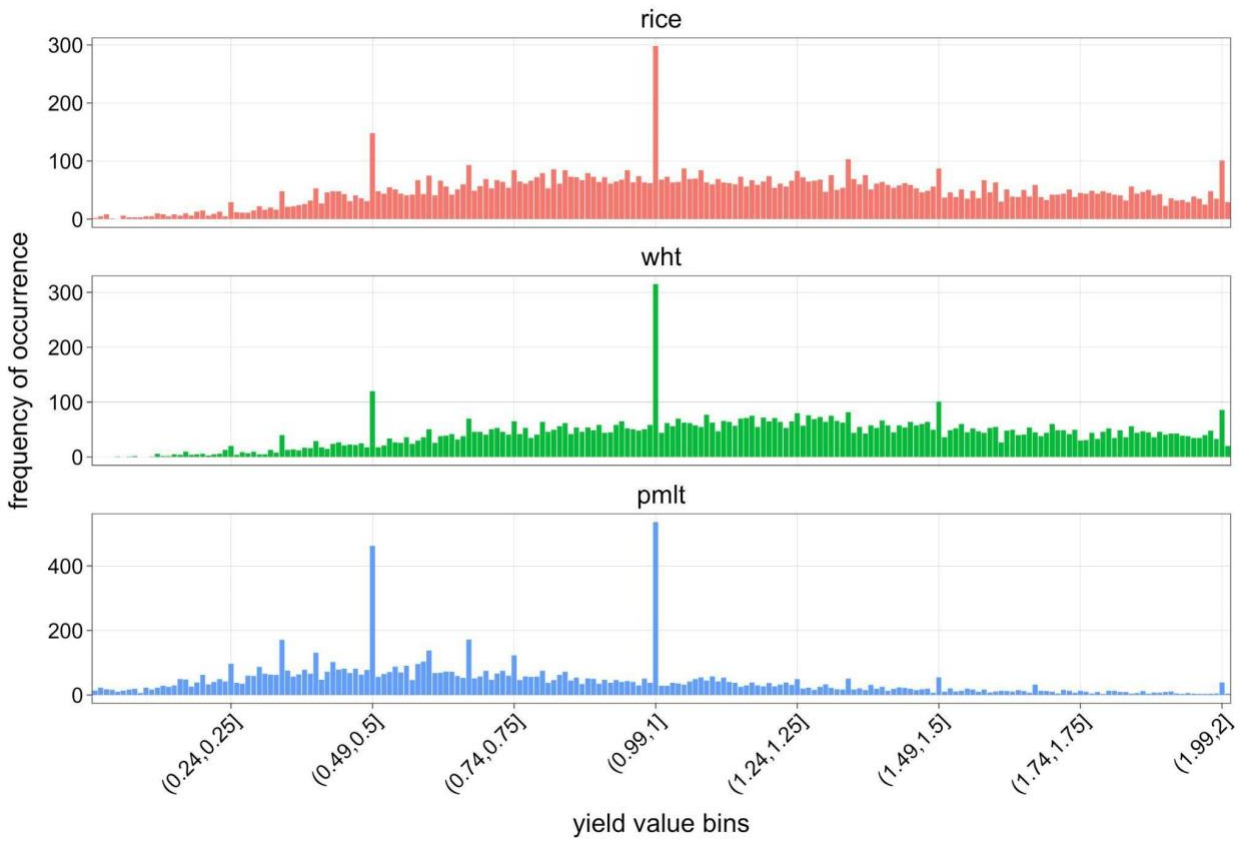

Figure 5.3 District-level yield values for the full dataset binned at 0.01 t/ha intervals. Spikes at multiples of 0.10 and 0.25 t/ha are evidence of rounding off or rough estimates in reported production and area values.

Two, the ICRISAT dataset provides annual crop production values. If a certain crop is harvested multiple times in a year, that data is provided as a single aggregated value. One example is the state of West Bengal where farmers grow rice for two, and sometimes three, seasons in a year (Shah, Chowdhury, & Shah, 2017). The yields calculated from total annual production data in such cases are hard to relate with the climate data because I conducted my analysis using seasonal (as opposed to annual) weather data according to state crop calendars. Also, the ICRISAT dataset had some missing values, such as irrigation data for the states of West Bengal and Assam; I therefore omitted West Bengal and Assam from my analysis.

For future climate change simulations in chapter 4, I assumed that factors like planting dates, length of growing season, or crop choices across regions remain constant over time. Commonly used adaptations like shifting where and when certain crops are grown cannot be addressed using my analysis. My future yield predictions under climate change also ignore the adoption of improved cultivar varieties better-suited for the changing climate. Also, C_3 crops like rice and wheat (unlike C_4 crops like millet) may benefit from elevated CO_2 concentrations in the future (Kukal & Irmak, 2018). CO_2 fertilization cannot be included in statistical studies like mine because CO_2 concentration for observed panel data is a monotonically increasing value and impossible to isolate from the technological trend denoted by the time variable (Schlenker & Roberts, 2009). Furthermore, the soil moisture model used in chapter 4 was based on some assumptions necessitated by (lack of) data availability. I used Hargreaves' equation for calculating daily ET while there are more accurate and physically sound methods such as Penman-Monteith; the former only requires temperature data, as opposed to the latter which needs temperature, humidity, and wind speed data (Allen et al., 1998). Nevertheless, the accuracy of Hargreaves' method has been deemed decent and acceptable in literature (Hargreaves & Allen, 2003; Allen et al., 1998; Aadhar & Mishra, 2020a).

Lastly, India is a large country, and national level studies like mine may ignore important trends and patterns that have been reported in more granular studies (Zachariah, Mondal, Das, Achutarao, & Ghosh, 2020). This limitation applies to all studies conducted over a large but heterogeneous nation state. A case can therefore be made for building more local models, and assessing variable relative importances in those models. Also, this dissertation only focused on

three major crops in India. In the future, my analysis could be extended to more crops for an overall assessment of the impact of climate change on Indian agriculture. Most studies in this field focus on food grains; analysis of non-grain crops would be a welcome contribution. My study was primarily focused on statistical models. As discussed earlier, there are some distinct advantages of process-based models. Future work could include comparisons of my results to process-based modeling results like those from AgMIP (Rosenzweig et al., 2013). And finally, my work only scratches the surface of the advantages of machine learning. I only showed one technique, boosted regression trees, out of numerous popular methods available. The statistical tradeoffs between overfitting and fitting flexible functions could be further explored with such techniques.

5.3 Open questions and concluding thoughts

This dissertation has extensively analyzed and dissected statistical models that are often seen as objectively true and used for precise yield predictions. Few studies have spent time and effort examining the underlying mechanisms. This sometimes, unfortunately, leads to strange results that stretch the limits of credibility, examples of which I will refrain from giving here. Regardless, I believe this dissertation has just touched the tip of the iceberg. Even something as mundane as geography dummy variables and time fixed effects in linear regression turned out to be a lot more complex with important consequences for predicting climate change impact on crop yields.

At the risk of sounding clichéd, I finish with more questions than I started with. Some immediate ones include:

1. What is the most appropriate way to delink climatic and non-climatic determinants drivers of crop yield?
2. What are the potential benefits of building a range of models, and then choosing a subset for near-term yield predictions based on climatic conditions? For example, in chapter 2, there was a noticeable difference between the performance of seasonal and subseasonal model during the drought year of 2002, even though their overall accuracy was similar. Since model accuracy varies under different climatic conditions, it would be worthwhile to understand which models perform best under what conditions, which can then lead to a more informed model selection process for making accurate yield predictions.
3. If the choice of climate variables or statistical techniques has such a significant impact on future yield predictions, should the most accurate model be selected for future yield predictions? Or should multiple models be combined for ensemble predictions? If so, how? Equally weighted? Weighted by prediction accuracy? Based on standard goodness of fit metrics?
4. While combining multiple models, should policymakers remain risk-averse and give extra weightage to models predicting yield losses, compared to those predicting yield gains from climate change?

In conclusion, if I were to summarize the whole dissertation in one sentence, I would do well to borrow the words of the famous statistician, George Edward Pelham Box:

"All models are wrong, but some are useful."

References

1. Aadhar, S., & Mishra, V. (2020a). Increased Drought Risk in South Asia under Warming Climate: Implications of Uncertainty in Potential Evapotranspiration Estimates. *Journal of Hydrometeorology*, *21*(12), 2979–2996. https://doi.org/10.1175/jhm-d-19-0224.1
2. Aadhar, S., & Mishra, V. (2020b). On the Projected Decline in Droughts Over South Asia in CMIP6 Multimodel Ensemble. *Journal of Geophysical Research: Atmospheres*, *125*(20), 1–18. https://doi.org/10.1029/2020JD033587
3. Aadhar, S., & Mishra, V. (2021). On the occurrence of the worst drought in South Asia in the observed and future climate. *Environmental Research Letters*, *16*(2). https://doi.org/10.1088/1748-9326/abd6a6
4. Albers, H., Gornott, C., & Hüttel, S. (2017). How do inputs and weather drive wheat yield volatility? The example of Germany. *Food Policy*, *70*, 50–61. https://doi.org/10.1016/j.foodpol.2017.05.001
5. Allen, R. G., Pereira, L. C., Raes, D., & Smith, M. (1998). *Crop evapotranspiration - Guidelines for computing crop water requirements - FAO irrigation and drainage paper 56* (Vol. 300). Retrieved from http://www.kimberly.uidaho.edu/water/fao56/fao56.pdf
6. Almazroui, M., Saeed, S., Saeed, F., Islam, M. N., & Ismail, M. (2020). Projections of Precipitation and Temperature over the South Asian Countries in CMIP6. *Earth Systems and Environment*, *4*(2), 297–320. https://doi.org/10.1007/s41748-020-00157-7
7. Annamalai, H., Hafner, J., Sooraj, K. P., & Pillai, P. (2013). Global warming shifts the monsoon circulation, drying South Asia. *Journal of Climate*, *26*(9), 2701–2718. https://doi.org/10.1175/JCLI-D-12-00208.1

8. Arnold, J. B. (2021). *ggthemes: Extra Themes, Scales and Geoms for "ggplot2". R package version 4.2.4*.

9. Asseng, S., Cammarano, D., Basso, B., Chung, U., Alderman, P. D., Sonder, K., Reynolds, M., & Lobell, D. (2017). Hot spots of wheat yield decline with rising temperatures. *Global Change Biology*, *23*(6), 2464–2472. https://doi.org/10.1111/gcb.13530

10. Auguie, B. (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3*.

11. Bassu, S., et al. (2014). How do various maize crop models vary in their responses to climate change factors? *Global Change Biology*, *20*(7), 2301–2320. https://doi.org/10.1111/gcb.12520

12. Beillouin, D., Schauberger, B., Bastos, A., Ciais, P., & Makowski, D. (2020). Impact of extreme weather conditions on European crop production in 2018: Random forest - Yield anomalies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1810). https://doi.org/10.1098/rstb.2019.0510

13. Birthal, P. S., Khan, T., Negi, D. S., & Agarwal, S. (2014). Impact of Climate Change on Yields of Major Food Crops in India: Implications for Food Security. *Agricultural Economics Research Review*, *27*(2), 145. https://doi.org/10.5958/0974-0279.2014.00019.6

14. Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–215. https://doi.org/10.1214/ss/1009213726

15. Brück, H., Piro, B., Sattelmacher, B., & Payne, W. A. (2003). Spatial distribution of roots of pearl millet on sandy soils of Niger. *Plant and Soil*, *256*(1), 149–159. https://doi.org/10.1023/A:1026246728095

16. Burney, J., & Ramanathan, V. (2014). Recent climate and air pollution impacts on indian agriculture. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(46), 16319–16324. https://doi.org/10.1073/pnas.1317275111

17. Butler, E. E., & Huybers, P. (2013). Adaptation of US maize to temperature variations. *Nature Climate Change*, *3*(1), 68–72. https://doi.org/10.1038/nclimate1585

18. Bzdok, D., Altman, N., & Krzywinski, M. (2018). Points of Significance: Statistics versus machine learning. *Nature Methods*, *15*(4), 233–234. https://doi.org/10.1038/nmeth.4642

19. Central Ground Water Board. (2020). *Ground Water Year Book - India 2019-20*.

20. Christensen, P., Gillingham, K., & Nordhaus, W. (2018). Uncertainty in forecasts of long-run economic growth. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(21), 5409–5414. https://doi.org/10.1073/pnas.1713628115

21. Das, J., & Umamahesh, N. V. (2021). Heat wave magnitude over India under changing climate: Projections from CMIP5 and CMIP6 experiments. *International Journal of Climatology*, (June), 1–21. https://doi.org/10.1002/joc.7246

22. Davis, K. F., Chhatre, A., Rao, N. D., Singh, D., & Defries, R. (2019). Sensitivity of grain yields to historical climate variability in India. *Environmental Research Letters*, *14*(6). https://doi.org/10.1088/1748-9326/ab22db

23. Davy, R., Esau, I., Chernokulsky, A., Outten, S., & Zilitinkevich, S. (2017). Diurnal asymmetry to the observed global warming. *International Journal of Climatology*, *37*(1), 79–93. https://doi.org/10.1002/joc.4688

24. De Gol, D., Festa, R., & Ratto, C. F. (1987). A simple expression for computing the daily extraterrestrial irradiation on a horizontal surface. *Solar and Wind Technology*, *4*(4), 509–512. https://doi.org/10.1016/0741-983X(87)90028-2

25. Dowle, M., & Srinivasan, A. (2021). *data.table: Extension of "data.frame". R package version 1.14.0*.

26. Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

27. FAOSTAT. (2021). FAOSTAT Database (Rome: Food and Agriculture Organization). Retrieved September 4, 2021, from http://www.fao.org/faostat/en/#home

28. Faraway, J. J. (2015). *Linear models with R* (2nd edition). CRC Press, Taylor and Francis Group.

29. Fishman, R. (2016). More uneven distributions overturn benefits of higher precipitation for crop yields. *Environmental Research Letters*, *11*(2), 24004. https://doi.org/10.1088/1748-9326/11/2/024004

30. Fishman, R. (2018). Groundwater depletion limits the scope for adaptation to increased rainfall variability in India. *Climatic Change*, *147*(1–2), 195–209. https://doi.org/10.1007/s10584-018-2146-x

31. Frankel, F. R. (2015). *India's Green Revolution: Economic Gains and Political Costs*. https://doi.org/doi:10.1515/9781400869022

32. Green, J. K., Seneviratne, S. I., Berg, A. M., Findell, K. L., Hagemann, S., Lawrence, D. M., & Gentine, P. (2019). Large influence of soil moisture on long-term terrestrial carbon uptake. *Nature*, 565(7740), 476–479. https://doi.org/10.1038/s41586-018-0848-x

33. Greenwell, B. M. (2017). pdp: An R Package for Constructing Partial Dependence Plots. *The R Journal*, *9*(1), 421–436.

34. Greenwell, B., Boehmke, B., & Cunningham, J. (2020). *gbm: Generalized Boosted Regression Models. R package version 2.1.8*.

35. Grillakis, M. G. (2019). Increase in severe and extreme soil moisture droughts for Europe under climate change. *Science of the Total Environment*, *660*, 1245–1255. https://doi.org/10.1016/j.scitotenv.2019.01.001

36. Grömping, U. (2006). Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software*, *17*(1), 1–27. https://doi.org/10.18637/jss.v017.i01

37. Gupta, R., Somanathan, E., & Dey, S. (2017). Global warming and local air pollution have reduced wheat yields in India. *Climatic Change*, *140*(3–4), 593–604. https://doi.org/10.1007/s10584-016-1878-8

38. Gupta, V., Singh, V., & Jain, M. K. (2020). Assessment of precipitation extremes in India during the 21st century under SSP1-1.9 mitigation scenarios of CMIP6 GCMs. *Journal of Hydrology*, *590*(August), 125422. https://doi.org/10.1016/j.jhydrol.2020.125422

39. Gusain, A., Ghosh, S., & Karmakar, S. (2020). Added value of CMIP6 over CMIP5 models in simulating Indian summer monsoon rainfall. *Atmospheric Research*, *232*(June 2019), 104680. https://doi.org/10.1016/j.atmosres.2019.104680

40. Ha, K. J., Moon, S., Timmermann, A., & Kim, D. (2020). Future Changes of Summer Monsoon Characteristics and Evaporative Demand Over Asia in CMIP6 Simulations. *Geophysical Research Letters*, *47*(8), 1–10. https://doi.org/10.1029/2020GL087492

41. Hargreaves, G. H., & Allen, R. G. (2003). History and Evaluation of Hargreaves Evapotranspiration Equation. *Journal of Irrigation and Drainage Engineering*, *129*(1), 53–63. https://doi.org/10.1061/(asce)0733-9437(2003)129:1(53)

42. Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning* (Vol. 1). https://doi.org/10.2307/2980421

43. Hausfather, Z. (2018). Explainer: How 'Shared Socioeconomic Pathways' explore future climate change. Retrieved September 4, 2021, from https://www.carbonbrief.org/explainer-how-shared-socioeconomic-pathways-explore-future-climate-change

44. Hengl, T., et al. (2017). SoilGrids250m: Global gridded soil information based on machine learning. In *PLoS ONE* (Vol. 12). https://doi.org/10.1371/journal.pone.0169748

45. Hijmans, R. J., Phillips, S., Leathwick, J., & Elith, J. (2020). *dismo: Species Distribution Modeling. R package version 1.3-3*.

46. ICRISAT. (2015). *Village dynamics in South Asia: meso level data for India: 1966-2011*. Hyderabad.

47. IPCC. (2014). *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part B: Regional Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (V. R. Barros, et al., Eds.). Cambridge University Press.

48. IPCC. (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (V. Masson-Delmotte, et al. Eds.). Cambridge University Press.

49. Jain, M., Fishman, R., Mondal, P., Galford, G. L., Bhattarai, N., Naeem, S., Lall, U., Singh, B., & DeFries, R. S. (2021). Groundwater depletion will reduce cropping intensity in India. *Science Advances*, *7*(9), 1–10. https://doi.org/10.1126/sciadv.abd2849

50. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. https://doi.org/10.1016/j.peva.2007.06.006

51. Jiang, H., Hu, H., Li, B., Zhang, Z., Wang, S., & Lin, T. (2021). Understanding the non-stationary relationships between corn yields and meteorology via a spatiotemporally varying coefficient model. *Agricultural and Forest Meteorology*, *301–302*(February). https://doi.org/10.1016/j.agrformet.2021.108340

52. Katzenberger, A., Schewe, J., Pongratz, J., & Levermann, A. (2020). Robust increase of Indian monsoon rainfall and its variability under future warming in CMIP-6 models. *Earth System Dynamics Discussions*, 1–30. https://doi.org/10.5194/esd-2020-80

53. Kukal, M. S., & Irmak, S. (2018). Climate-Driven Crop Yield and Yield Variability and Climate Change Impacts on the U.S. Great Plains Agricultural Production. *Scientific Reports*, *8*(1), 1–18. https://doi.org/10.1038/s41598-018-21848-2

54. Li, H., & Maddala, G. S. (1996). Bootstrapping time series models. *Econometric Reviews*, *15*(2), 115–158.

55. Li, X., & Troy, T. J. (2018). Changes in rainfed and irrigated crop yield response to climate in the western US. *Environmental Research Letters*, *13*(6). https://doi.org/10.1088/1748-9326/aac4b1

56. Lobell, D., & Burke, M. (2009). *Climate change and food security: Adopting agriculture to a warmer world* (D. Lobell & M. Burke, eds.). Springer Science & Business Media.

57. Lobell, D., & Field, C. (2007). Global scale climate-crop yield relationships and the impacts of recent warming. *Environmental Research Letters*, *2*(1). https://doi.org/10.1088/1748-9326/2/1/014002

58. Lobell, D., Cassman, K. G., & Field, C. B. (2009). Crop yield gaps: Their importance, magnitudes, and causes. *Annual Review of Environment and Resources*, *34*, 179–204. https://doi.org/10.1146/annurev.environ.041008.093740

59. Lobell, D., Schlenker, W., & Costa-Roberts, J. (2011). Climate Trends and Global Crop Production Since 1980. *Science*, *333*.

60. Lobell, D., Sibley, A., & Ivan Ortiz-Monasterio, J. (2012). Extreme heat effects on wheat senescence in India. *Nature Climate Change*, *2*(3), 186–189. https://doi.org/10.1038/nclimate1356

61. Mall, R. K., Chaturvedi, M., Singh, N., Bhatla, R., Singh, R. S., Gupta, A., & Niyogi, D. (2021). Evidence of asymmetric change in diurnal temperature range in recent decades over different agro-climatic zones of India. *International Journal of Climatology*, *41*(4), 2597–2610. https://doi.org/10.1002/joc.6978

62. May, W. (2004). Simulation of the variability and extremes of daily rainfall during the Indian summer monsoon for present and future times in a global time-slice experiment. *Climate Dynamics*, *22*(2–3), 183–204. https://doi.org/10.1007/s00382-003-0373-x

63. Menelly, S. (2016). A lesson in mismanagement: the global rice crisis of 2008. *Harvard International Review*, *37*(3), 44–47.

64. Microsoft, & Weston, S. (2020a). *doParallel: Foreach Parallel Adaptor for the "parallel" Package. R package version 1.0.16*.

65. Microsoft, & Weston, S. (2020b). *foreach: Provides Foreach Looping Construct. R package version 1.5.1*.

66. Ministry of Agriculture and Farmers Welfare. (2016). *Agricultural statistics at a glance*. Retrieved from http://agricoop.nic.in/agristatistics.htm

67. Ministry of Agriculture and Farmers Welfare. (2018). *Annual Report 2018-19*. New Delhi, India.

68. Ministry of Finance. (2018). *Economic Survey 2017-18*.

69. Ministry of Water Resources. (2017). *Report of 5th census of minor irrigation schemes*.

70. Mishra, V., Bhatia, U., & Tiwari, A. D. (2020). Bias-corrected climate projections for South Asia from Coupled Model Intercomparison Project-6. *Nature Scientific Data*, *7*(1), 1–13. https://doi.org/10.1038/s41597-020-00681-1

71. Mishra, V., Thirumalai, K., Singh, D., & Aadhar, S. (2020). Future exacerbation of hot and dry summer monsoon extremes in India. *Npj Climate and Atmospheric Science*, *3*(1). https://doi.org/10.1038/s41612-020-0113-5

72. Moon, S., & Ha, K. J. (2020). Future changes in monsoon duration and precipitation using CMIP6. *Npj Climate and Atmospheric Science*, *3*(1), 1–7. https://doi.org/10.1038/s41612-020-00151-w

73. Mueller, N. D., Gerber, J. S., Johnston, M., Ray, D. K., Ramankutty, N., & Foley, J. A. (2012). Closing yield gaps through nutrient and water management. *Nature*, *490*(7419), 254–257. https://doi.org/10.1038/nature11420

74. Muggeo, V. M. R. (2008). *segmented: an R Package to Fit Regression Models with Broken-Line Relationships*.

75. Murari, K. K., Ghosh, S., Patwardhan, A., Daly, E., & Salvi, K. (2015). Intensification of future severe heat waves in India and their effect on heat stress and mortality. *Regional Environmental Change*, *15*(4), 569–579. https://doi.org/10.1007/s10113-014-0660-6

76. Neumann, K., Verburg, P. H., Stehfest, E., & Müller, C. (2010). The yield gap of global grain production: A spatial analysis. *Agricultural Systems*, *103*(5), 316–326. https://doi.org/10.1016/j.agsy.2010.02.004

77. Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes. R package version 1.1-2*.

78. Ortiz-Bobea, A., Wang, H., Carrillo, C. M., & Ault, T. R. (2019). Unpacking the climatic drivers of US agricultural yields. *Environmental Research Letters*, *14*(6), 64003. https://doi.org/10.1088/1748-9326/ab1e75

79. Patnaik, U. (2017). Mr Keynes and the forgotten holocaust in Bengal, 1943–44: Or, the macroeconomics of extreme demand compression. *Studies in People's History*, *4*(2), 197–210. https://doi.org/10.1177/2348448917725856

80. Peng, S., Huang, J., Sheehy, J. E., Laza, R. C., Visperas, R. M., Zhong, X., Centeno, G. S., Khush, G. S., & Cassman, K. G. (2004). Rice yields decline with higher night temperature from global warming. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(27), 9971–9975. https://doi.org/10.1073/pnas.0403720101

81. Peng, S., et al. (2013). Asymmetric effects of daytime and night-time warming on Northern Hemisphere vegetation. *Nature*, *501*(7465), 88–92. https://doi.org/10.1038/nature12434

82. Porter, J. R., & Gawith, M. (1999). Temperatures and the growth and development of wheat: a review. *European Journal of Agronomy, 10*, 23–36.

83. R core team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria.

84. Rajeevan, M., Bhate, J., Kale, J. D., & Lal, B. (2006). High resolution daily gridded rainfall data for the Indian region: Analysis of break and active monsoon spells. *Current Science, 91*(3), 296–306.

85. Ram, K., & Wickham, H. (2018). *wesanderson: A Wes Anderson Palette Generator. R package version 0.3.6*.

86. Ramankutty, N., Foley, J. A., Norman, J., & McSweeney, K. (2002). The global distribution of cultivable lands: current patterns and sensitivity to possible climate change. *Global Ecology and biogeography*, *11*(5), 377–392.

87. Ray, D. K., Gerber, J. S., Macdonald, G. K., & West, P. C. (2015). Climate variation explains a third of global crop yield variability. *Nature Communications*, *6*, 1–9. https://doi.org/10.1038/ncomms6989

88. Reuters. (2008). Why have rice prices surged to record highs? *Domestic News*. Retrieved September 4, 2021, from https://www.reuters.com/article/us-food-rice/factbox-why-have-rice-prices-surged-to-record-highs-idUSSP9004820080502

89. Riha, S. J., Wilks, D. S., & Simoens, P. (1996). Impact of temperature and precipitation variability on crop model predictions. *Climatic Change*, *32*(3), 293–311. https://doi.org/10.1007/BF00142466

90. Rising, J., & Devineni, N. (2020). Crop switching reduces agricultural losses from climate change in the United States by half under RCP 8.5. *Nature Communications*, *11*(1), 1–7. https://doi.org/10.1038/s41467-020-18725-w

91. Ritchie, J., & Dowlatabadi, H. (2017a). The 1000 GtC coal question: Are cases of vastly expanded future coal combustion still plausible? *Energy Economics*, *65*, 16–31. https://doi.org/10.1016/j.eneco.2017.04.015

92. Ritchie, J., & Dowlatabadi, H. (2017b). Why do climate change scenarios return to coal? *Energy*, *140*, 1276–1291. https://doi.org/10.1016/j.energy.2017.08.083

93. Roberts, M. J., Braun, N. O., Sinclair, T. R., Lobell, D., & Schlenker, W. (2017). Comparing and combining process-based crop models and statistical models with some implications for climate change. *Environmental Research Letters*, *12*(9). https://doi.org/10.1088/1748-9326/aa7f33

94. Rockström, J., Oweis, T. Y., Wani, S., Bruggeman, A., Farahani, J., Karlberg, L., & Qiang, Z. (2007). Managing water in rainfed agriculture. In *Water for Food, Water for Life: A Comprehensive Assessment of Water Management in Agriculture* (pp. 315–352).

95. Rohini, P., Rajeevan, M., & Mukhopadhay, P. (2019). Future projections of heat waves over India from CMIP5 models. *Climate Dynamics*, *53*(1), 975–988. https://doi.org/10.1007/s00382-019-04700-9

96. Rosenzweig, C., et al. (2013). The Agricultural Model Intercomparison and Improvement Project (AgMIP): Protocols and pilot studies. *Agricultural and Forest Meteorology*, *170*, 166–182. https://doi.org/10.1016/j.agrformet.2012.09.011

97. Rowhani, P., Lobell, D., Linderman, M., & Ramankutty, N. (2011). Climate variability and crop production in Tanzania. *Agricultural and Forest Meteorology*, *151*(4), 449–460. https://doi.org/10.1016/j.agrformet.2010.12.002

98. Saxton, K. E., & Rawls, W. J. (2006). Soil Water Characteristic Estimates by Texture and Organic Matter for Hydrologic Solutions. *Soil Science Society of America Journal*, *70*(5), 1569–1578. https://doi.org/10.2136/sssaj2005.0117

99. Schlenker, W., & Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proceedings of the National Academy of Sciences*, *106*(37), 15594–15598. https://doi.org/10.1073/pnas.0906865106

100. Scott, C. A., & Sharma, B. (2009). Energy supply and the expansion of groundwater irrigation in the Indus-Ganges Basin. *International Journal of River Basin Management*, *7*(2), 119–124. https://doi.org/10.1080/15715124.2009.9635374

101. Shah, M., Chowdhury, S. Das, & Shah, T. (2017). *Pro-Poor Farm Power Policy for West Bengal: Analytical Background for a Policy Pilot*.

102. Sharma, B. R., Rao, K. V., Vittal, K. P. R., Ramakrishna, Y. S., & Amarasinghe, U. (2010). Estimating the potential of rainfed agriculture in India: Prospects for water productivity improvements. *Agricultural Water Management*, *97*(1), 23–30. https://doi.org/10.1016/j.agwat.2009.08.002

103. Sidhu, B. S., Kandlikar, M., & Ramankutty, N. (2020). Power tariffs for groundwater irrigation in India: A comparative analysis of the environmental, equity, and economic tradeoffs. *World Development*, *128*, 104836. https://doi.org/10.1016/j.worlddev.2019.104836

104. Siebert, S., & Döll, P. (2010). Quantifying blue and green virtual water contents in global crop production as well as potential production losses without irrigation. *Journal of Hydrology*, *384*(3–4), 198–217. https://doi.org/10.1016/j.jhydrol.2009.07.031

105. Singh, A., Chaudhuri, B., & Roychoudhury, A. (2020). Influence of Night Temperature on Rice Yield and Quality. In A. Roychoudhury (Ed.), *Rice Research for Quality Improvement: Genomics and Genetic Engineering: Volume 1: Breeding Techniques and Abiotic Stress Tolerance* (pp. 579–590). https://doi.org/10.1007/978-981-15-4120-9_24

106. Singh, D., Tsiang, M., Rajaratnam, B., & Diffenbaugh, N. S. (2014). Observed changes in extreme wet and dry spells during the south Asian summer monsoon season. *Nature Climate Change*, *4*(6), 456–461. https://doi.org/10.1038/nclimate2208

107. Sun, X., Ren, G., You, Q., Ren, Y., Xu, W., Xue, X., Zhan, Y., Zhang, S., & Zhang, P. (2019). Global diurnal temperature range (DTR) changes since 1901. *Climate Dynamics*, *52*(5–6), 3343–3356. https://doi.org/10.1007/s00382-018-4329-6

108. Tebaldi, C., Hayhoe, K., Arblaster, J. M., & Meehl, G. A. (2006). Going to the extremes: An intercomparison of model-simulated historical and future changes in extreme events. *Climatic Change*, *79*(3–4), 185–211. https://doi.org/10.1007/s10584-006-9051-4

109. University of California Agriculture & Natural Resources. (2016). Degree days. Retrieved July 5, 2021, from http://ipm.ucanr.edu/WEATHER/ddconcepts.html

110. Vogel, E., Donat, M. G., Alexander, L. V., Meinshausen, M., Ray, D. K., Karoly, D., Meinshausen, N. & Frieler, K. (2019). The effects of climate extremes on global agricultural yields. *Environmental Research Letters*, *14*(5). https://doi.org/10.1088/1748-9326/ab154b

111. Wang, B., Jin, C., & Liu, J. (2020). Understanding Future Change of Global Monsoons Projected by CMIP6 Models. *Journal of Climate*, *33*(15), 6471–6489. https://doi.org/10.1175/JCLI-D-19-0993.1

112. Wheeler, T. R., Craufurd, P. Q., Ellis, R. H., Porter, J. R., & Vara Prasad, P. V. (2000). Temperature variability and the yield of annual crops. *Agriculture, Ecosystems and Environment*, *82*(1–3), 159–167. https://doi.org/10.1016/S0167-8809(00)00224-3

113. Wickham, H., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43). https://doi.org/10.21105/joss.01686

114. Williams, J. R., Dyke, P. T., Fuchs, W. W., Benson, V. W., Rice, O. W., & Taylor, E. D. (1990). *EPIC-Erosion/Productivity Impact Calculator: 2. User Manual* (A. . Sharpley & J. R. Williams, Eds.).

115. World Bank. (2021a). Agriculture, forestry, and fishing, value added (% of GDP) - India. Retrieved September 4, 2021, from https://data.worldbank.org/indicator/NV.AGR.TOTL.ZS?end=2020&locations=IN&start=2010

116. World Bank. (2021b). Employment in agriculture (% of total employment) (modeled ILO estimate) - India. Retrieved September 4, 2021, from https://data.worldbank.org/indicator/SL.AGR.EMPL.ZS?locations=IN

117. Yu, J., & Goh, G. (2019). Estimating non-additive within-season temperature effects on maize yields using Bayesian approaches. *Scientific Reports*, *9*(1), 1–8. https://doi.org/10.1038/s41598-019-55037-6

118. Zachariah, M., Mondal, A., Das, M., Achutarao, K. M., & Ghosh, S. (2020). On the role of rainfall deficits and cropping choices in loss of agricultural yield in Marathwada, India. *Environmental Research Letters*, *15*(9). https://doi.org/10.1088/1748-9326/ab93fc

119. Zaveri, E., & Lobell, D. (2019). The role of irrigation in changing wheat yields and heat sensitivity in India. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-12183-9

120. Zaveri, E., Grogan, D. S., Fisher-Vanden, K., Frolking, S., Lammers, R. B., Wrenn, D. H., Prusevich, A., & Nicholas, R. E. (2016). Invisible water, visible impact: Groundwater use and Indian agriculture under climate change. *Environmental Research Letters*, *11*(8), 1–13.

121. Zhai, J., Mondal, S. K., Fischer, T., Wang, Y., Su, B., Huang, J., Tao, H., Wang, G., Ullah, W., & Uddin, M. J. (2020). Future drought characteristics through a multi-model ensemble from CMIP6 over South Asia. *Atmospheric Research*, *246*(May), 105111.

Appendices

Appendix A Chapter 2

A.1 Schematic of degree day bins calculation

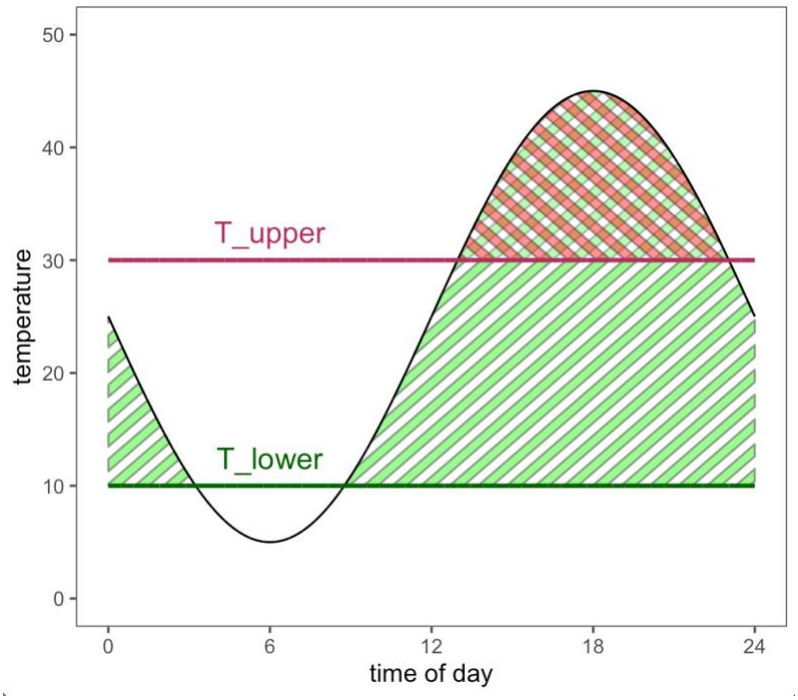

Figure A.1 Degree days accumulated on a particular day in different bins were calculated by subtracting degree days spent above T_{upper} (red) from the number of degree days spent above T_{lower} (green). Appropriate adjustments were made for days when the temperature curve did not cross the T_{lower} and/or T_{upper} limits. This quantity was then summed over the full growing season for each degree day bin and crop-district-year combination.

A.2 Model performance with irrigation

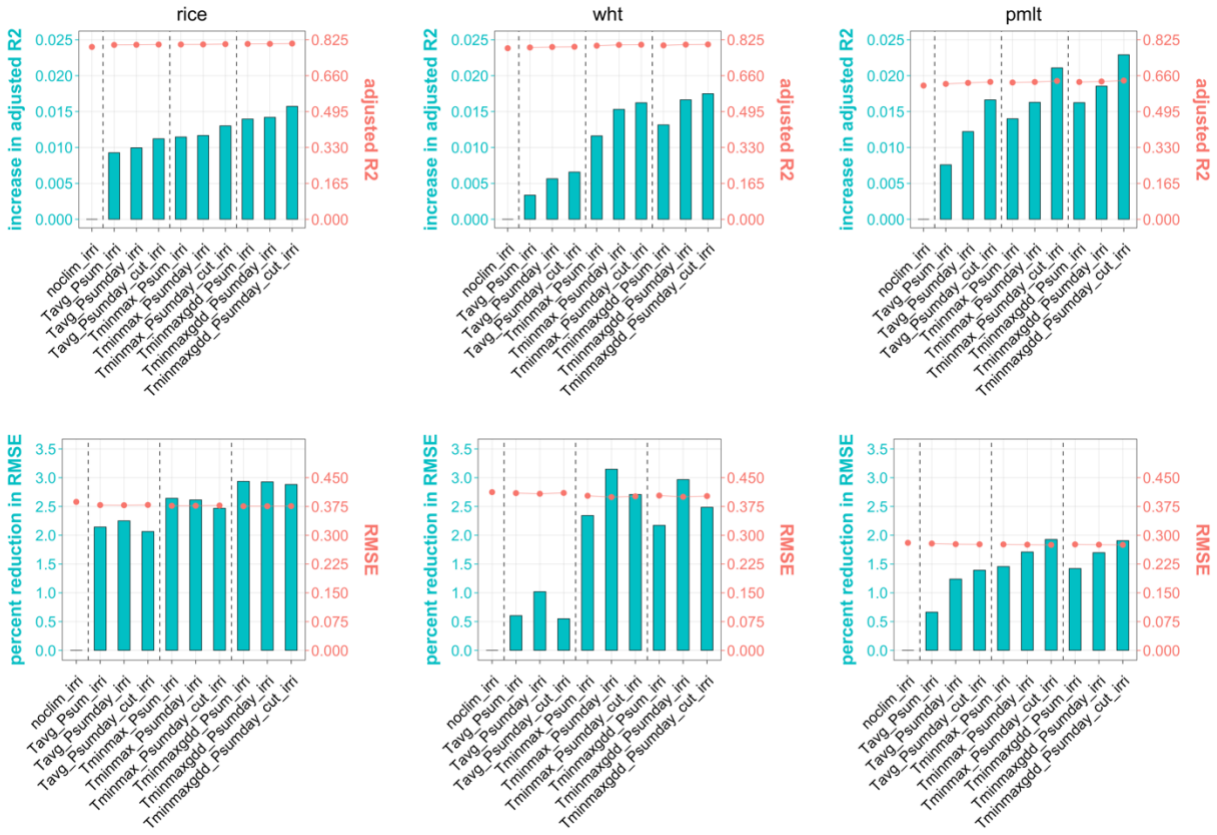

Figure A.2 Model (with irrigation included) performance measured in terms of adjusted R^2 (top row; absolute values in red, increase compared to null model in blue) and RMSE (bottom row; absolute values in red, percent increase compared to null model in blue). The three crops are rice (left), wheat (center), and pearl millet (right). Within each panel, models include varying levels of climate data, with three levels each of temperature and precipitation (see Table 2.2 for description of levels). The models are divided into sub-panels with dotted lines and arranged in the following order: null model; temperature level 1 and precipitation levels 1, 2, 3; temperature level 2 and precipitation levels 1, 2, 3; and temperature level 3 and precipitation levels 1, 2, 3.

A.3 State-level model performance in 1993, 1996, 2002, 2009

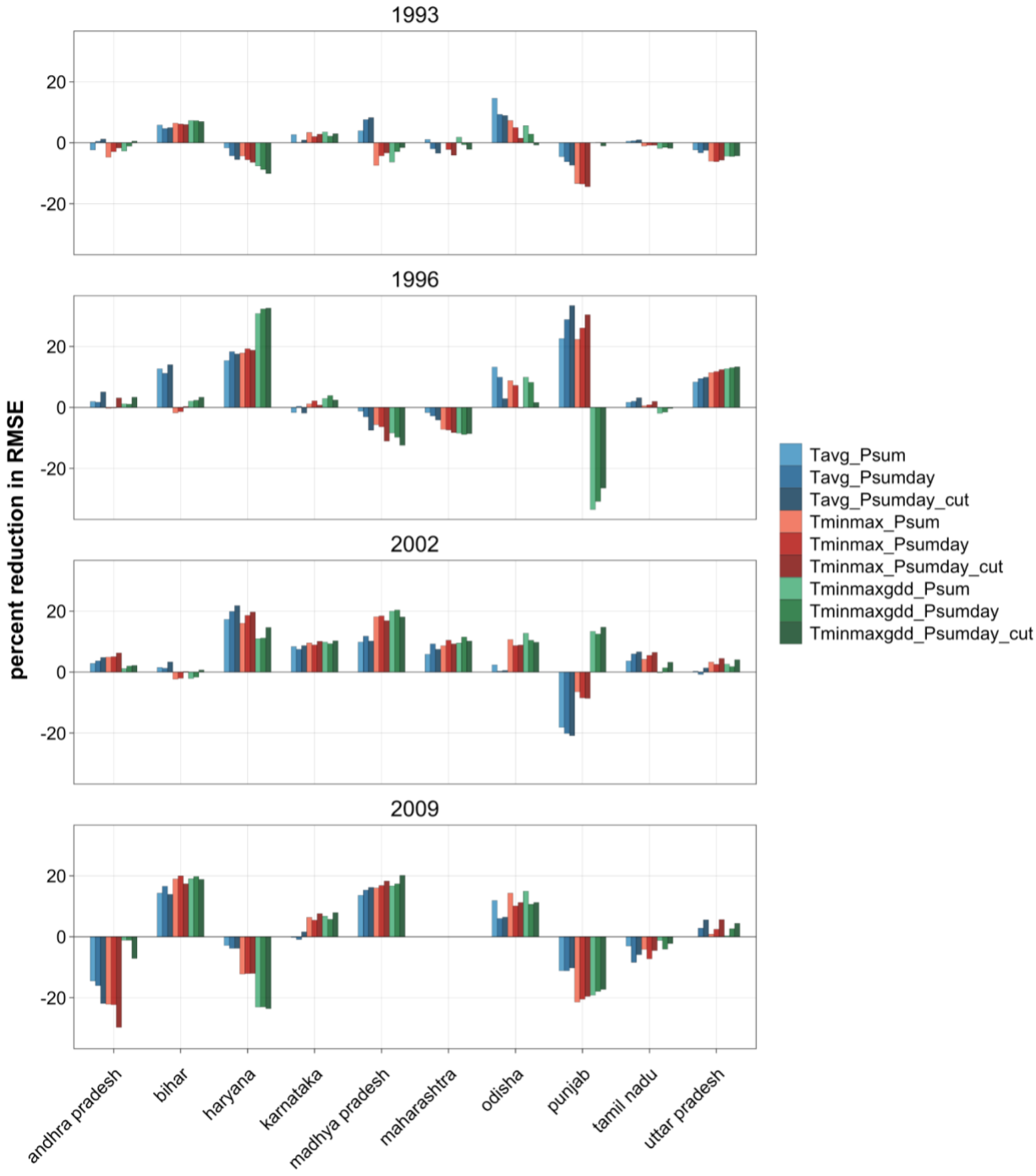

Figure A.3 Improvement in rice model performance (in terms of RMSE reduction compared to the null model with no climate variables), aggregated at state level, for median precipitation (1993), median temperature (1996), drought (2002), and hot (2009) years.

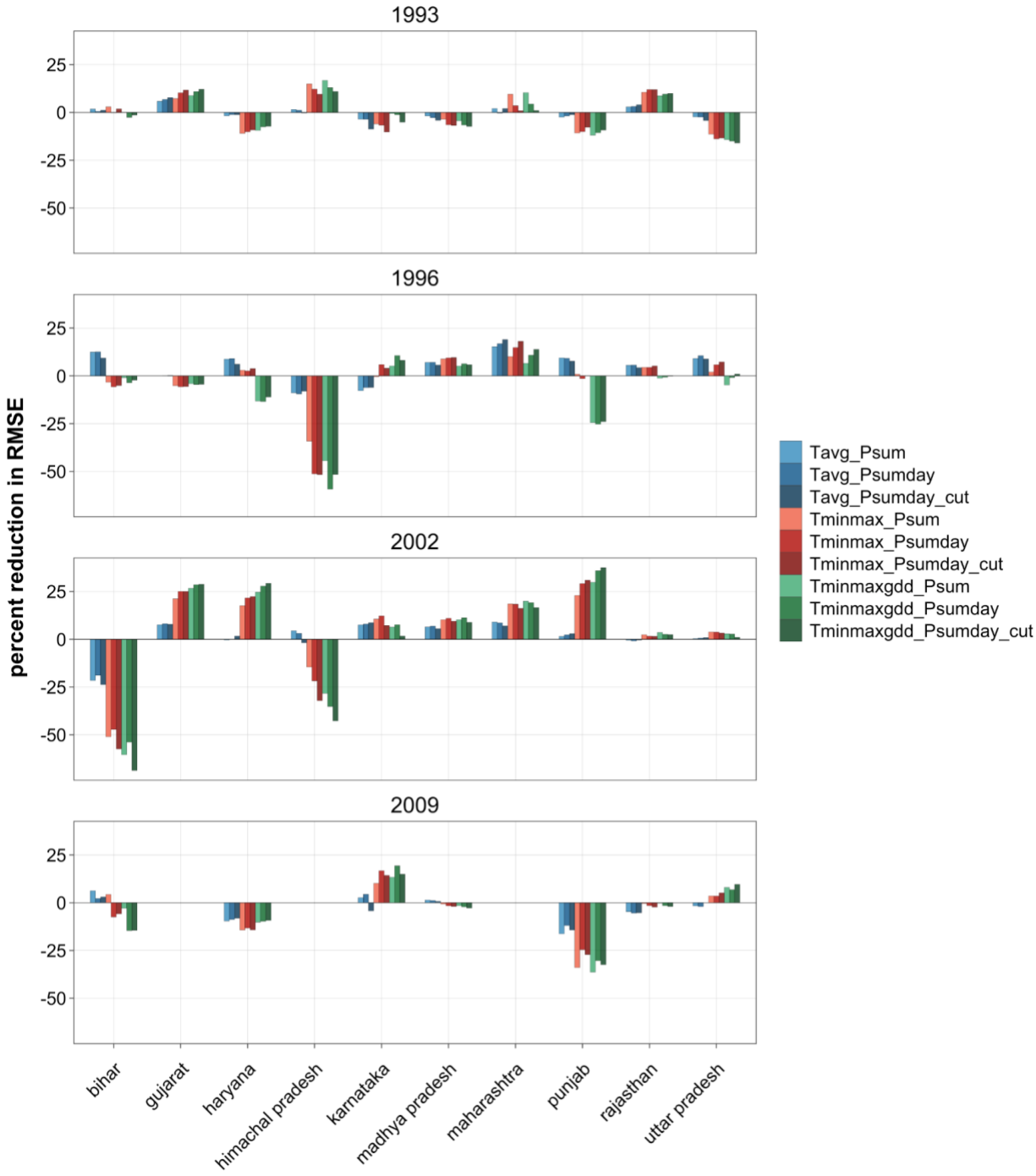

Figure A.4 Improvement in wheat model performance (in terms of RMSE reduction compared to the null model with no climate variables), aggregated at state level, for median precipitation (1993), median temperature (1996), drought (2002), and hot (2009) years.

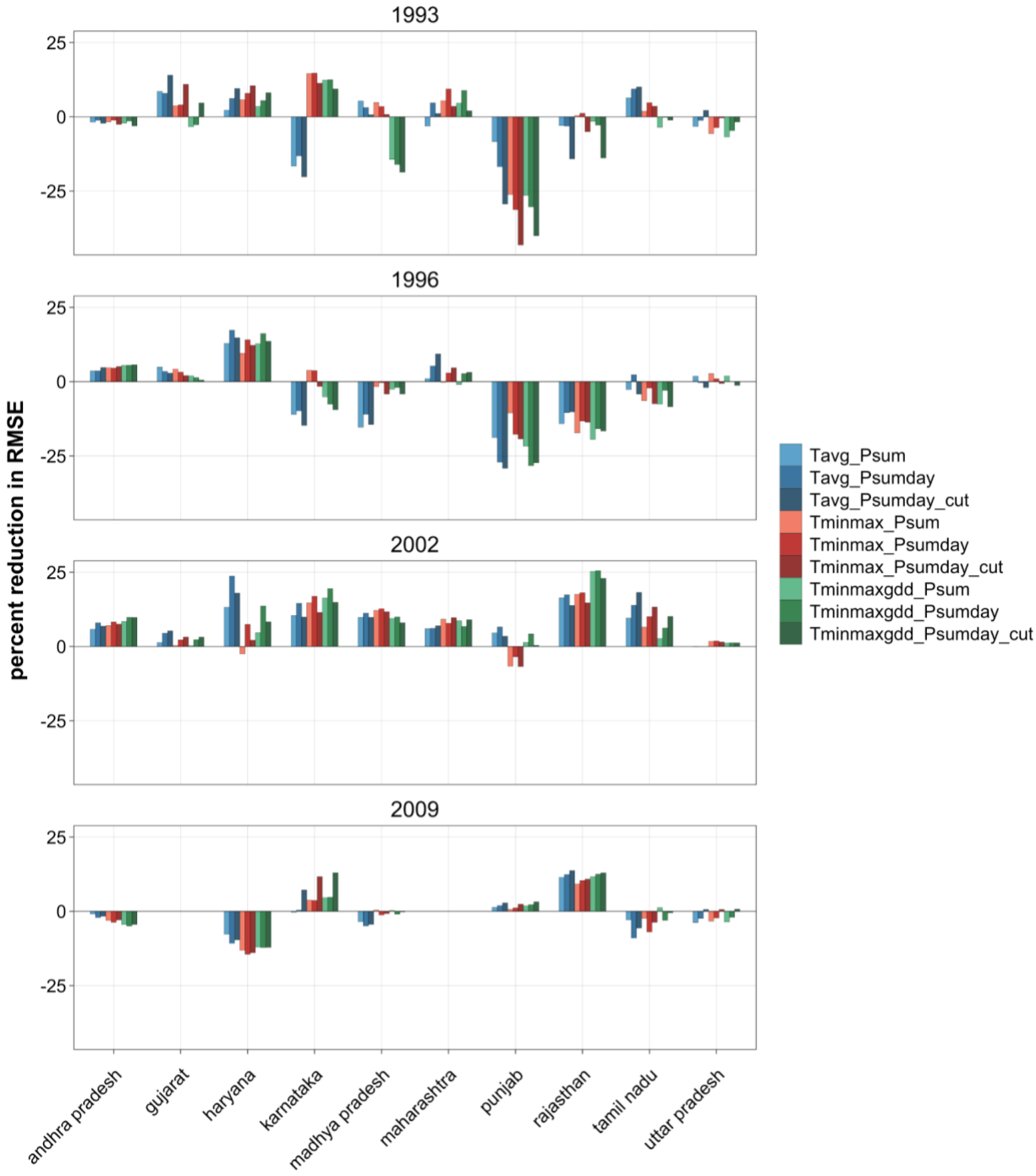

Figure A.5 Improvement in pearl millet model performance (in terms of RMSE reduction compared to the null model with no climate variables), aggregated at state level, for median precipitation (1993), median temperature (1996), drought (2002), and hot (2009) years.

A.4 Simulations of climate change impact (wheat and pearl millet)

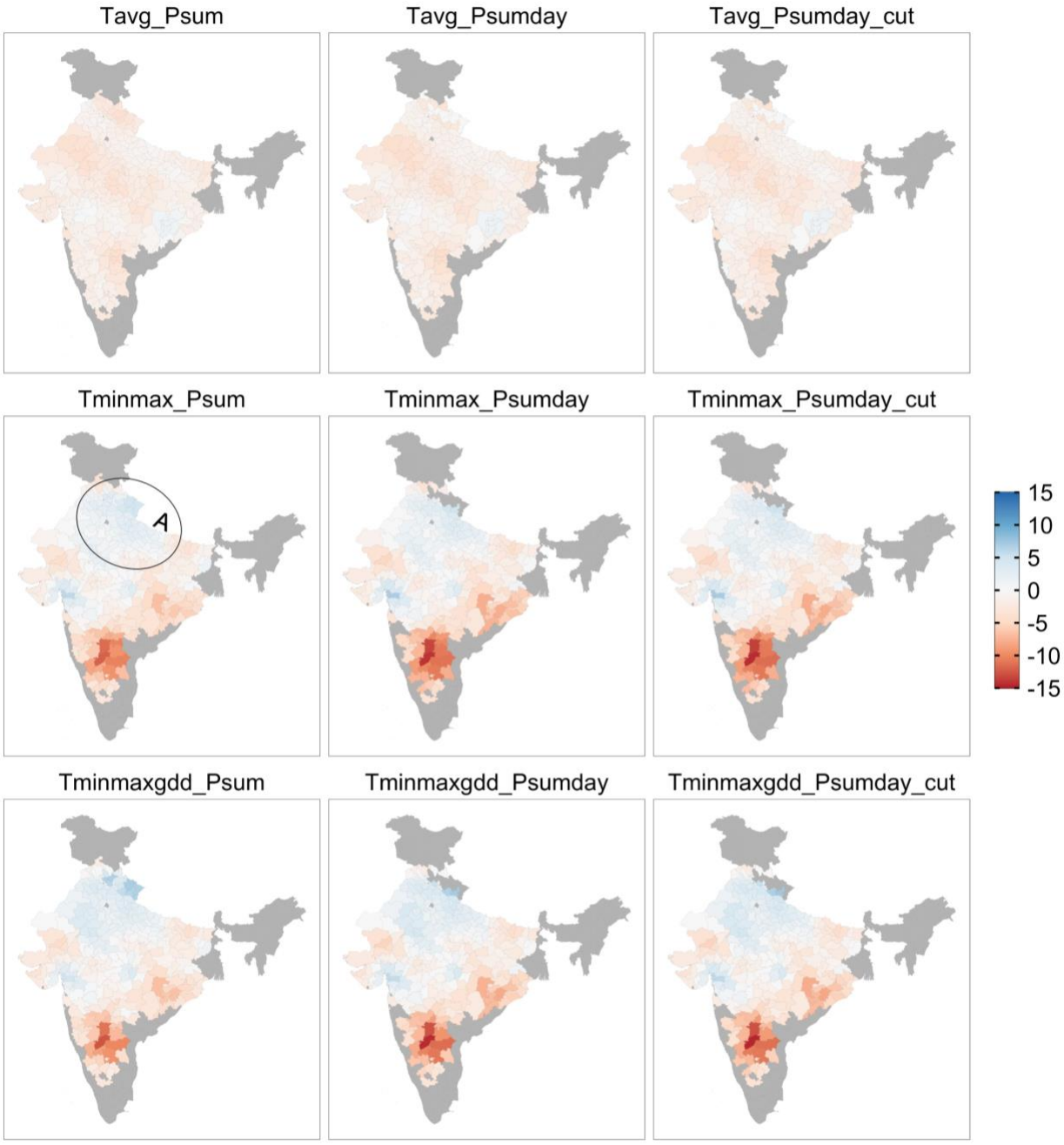

Figure A.6 Simulated impact of long-term climate change (since 1966) on wheat yield in the last decade (2002-2011) of the study time period. The climate data was linearly detrended to remove time trend at district-scale. District-level estimates of median value and 95 percent confidence intervals of climate change impact on yield were obtained through residual bootstrapping (n = 500). The average district-level yield loss during the last

decade in the dataset (2002-2011) is presented here as the expected impact of climate change that has occurred since 1966. Only results with 95 percent significance of the confidence intervals are shown; insignificant results are shown in gray.

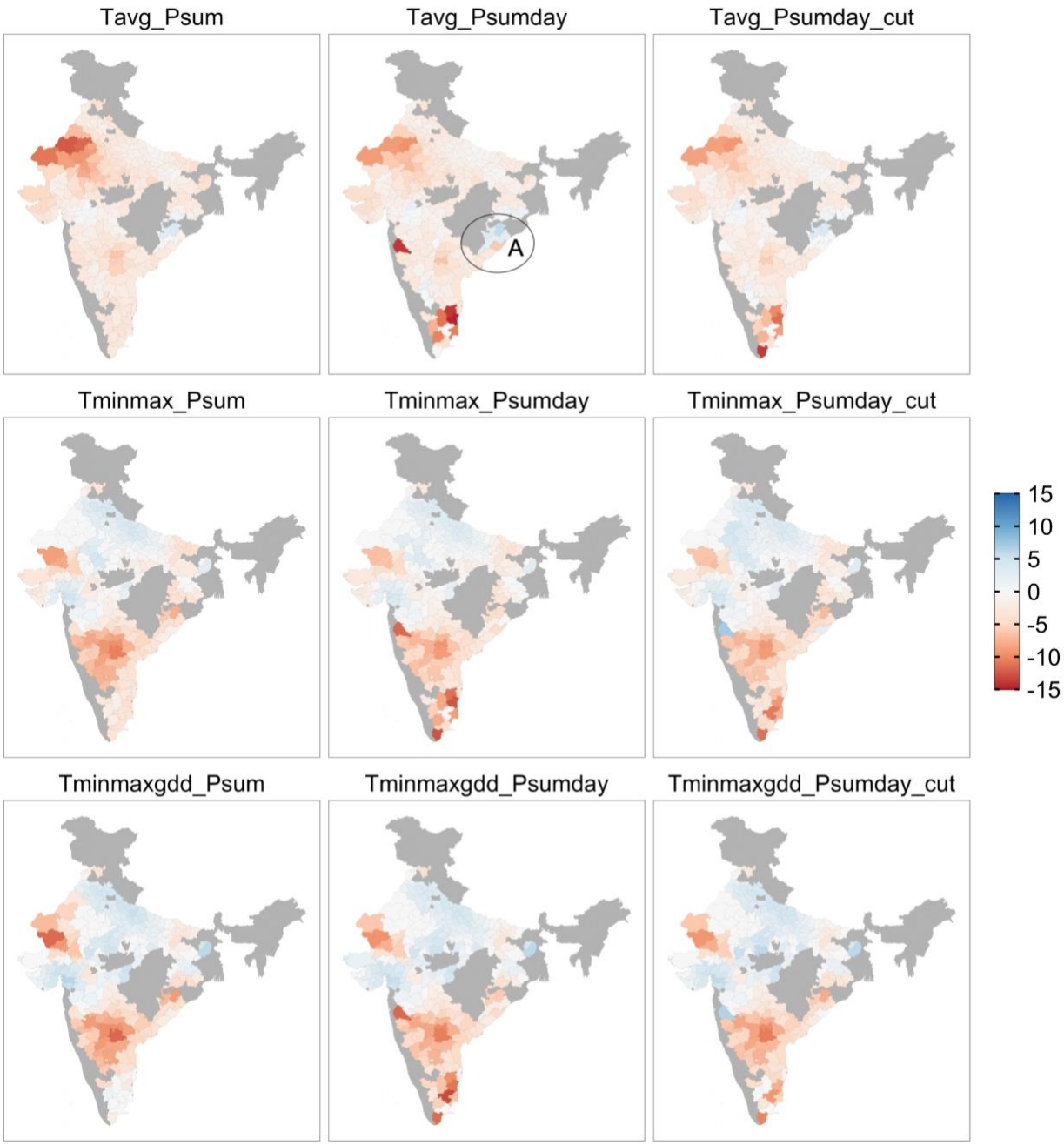

Figure A.7 Simulated impact of long-term climate change (since 1966) on pearl millet yield in the last decade (2002-2011) of the study time period. The climate data was linearly detrended to remove time trend at district-scale. District-level estimates of median value and 95 percent confidence intervals of climate change impact on yield were obtained through residual bootstrapping (n = 500). The average district-level yield loss during the last decade in the dataset (2002-2011) is presented here as the expected impact of climate change

that has occurred since 1966. Only results with 95 percent significance of the confidence intervals are shown; insignificant results are shown in gray.

Appendix B Chapter 3

B.1 Model accuracy for out-of-sample predictions

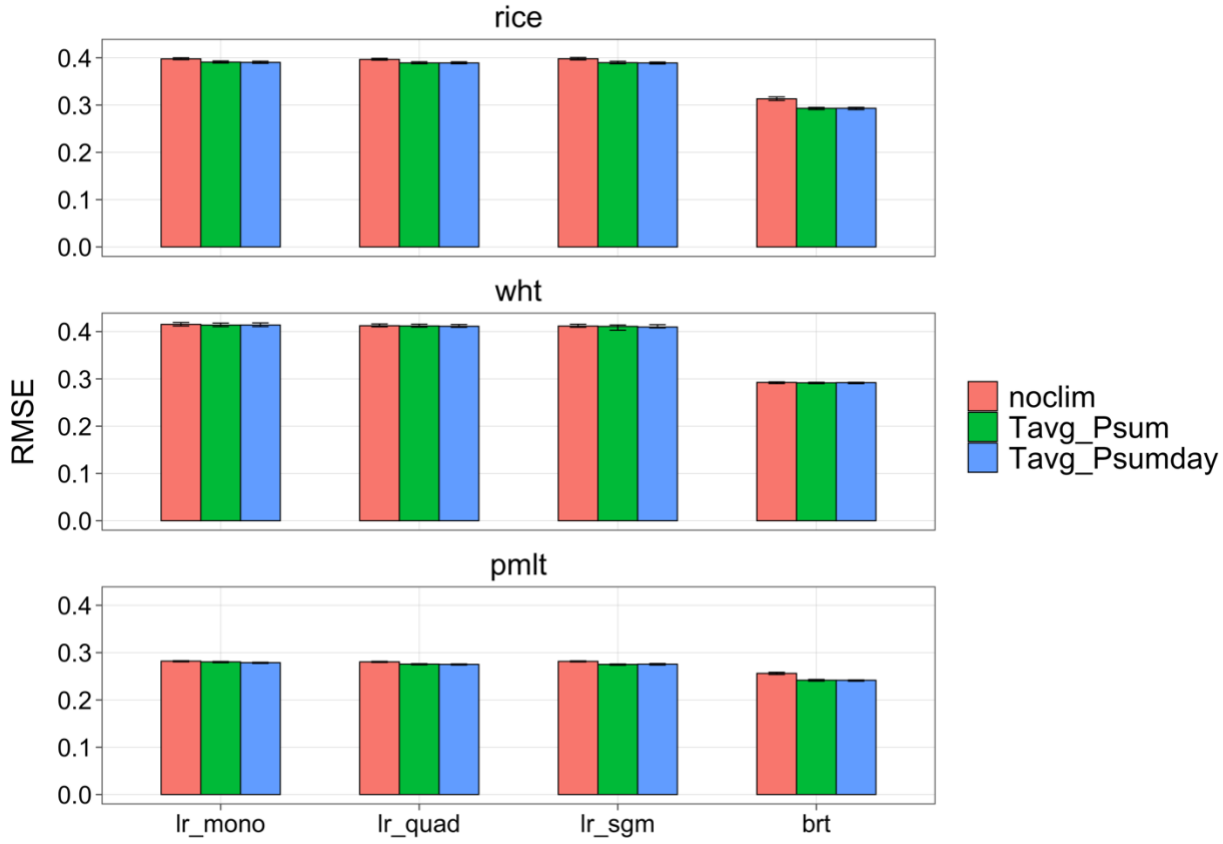

Figure B.1 Model performance in terms of RMSE of out-of-sample predictions (lower is better). The three crops are rice (top), wheat (middle), and pearl millet (bottom). Within each panel, the bars are color-coded by climate variables included (red: no climate; green: mean seasonal temperature and total seasonal precipitation; blue: mean seasonal temperature, total seasonal precipitation and total precipitation days over growing season). From left to right, the various models depicted are: (1) *lr_mono*: LR with linear terms; (2) *lr_quad*: LR with quadratic terms for time and all climate variables; (3) *lr_sgm*: LR with single-knot segmented analysis; (4) *brt*: BRT.

B.2 Partial dependence plots for rice and wheat

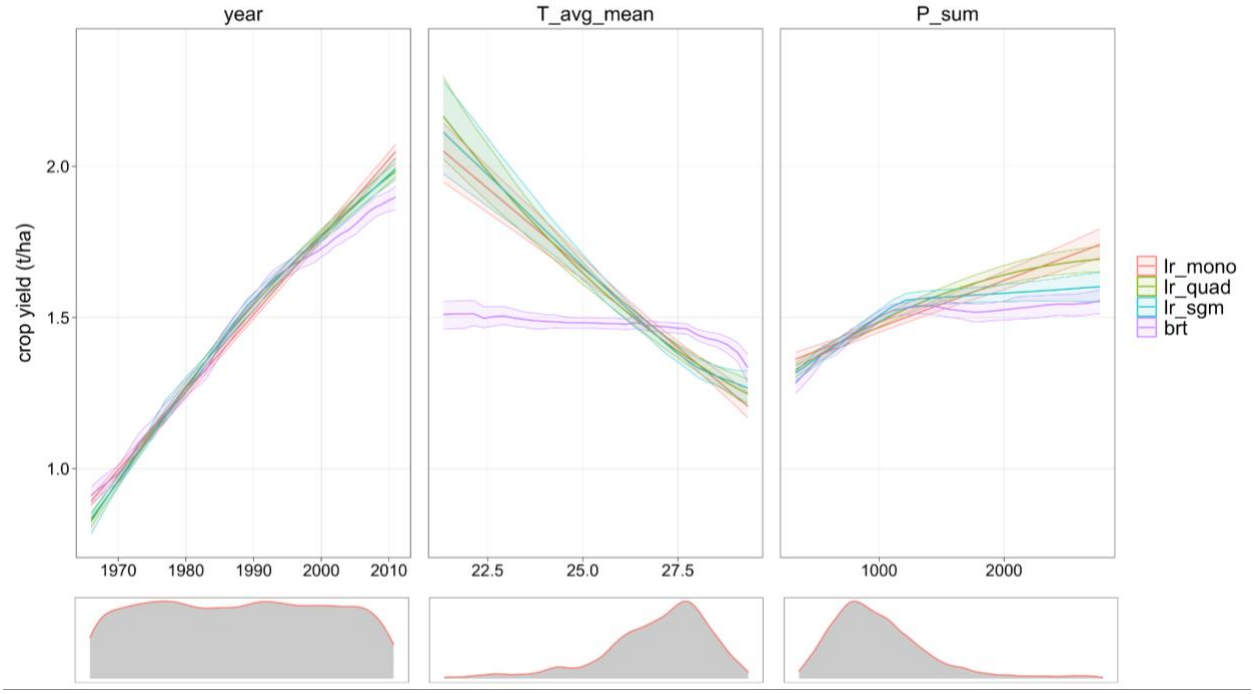

Figure B.2 Partial dependence plots of the *Tavg_Psum* variable set models for rice (top row), and distribution density of corresponding IV in training data (bottom row). The four model types are color-coded in each panel. Data density plots at the bottom provide an idea of where most of the training data lies for a particular IV.

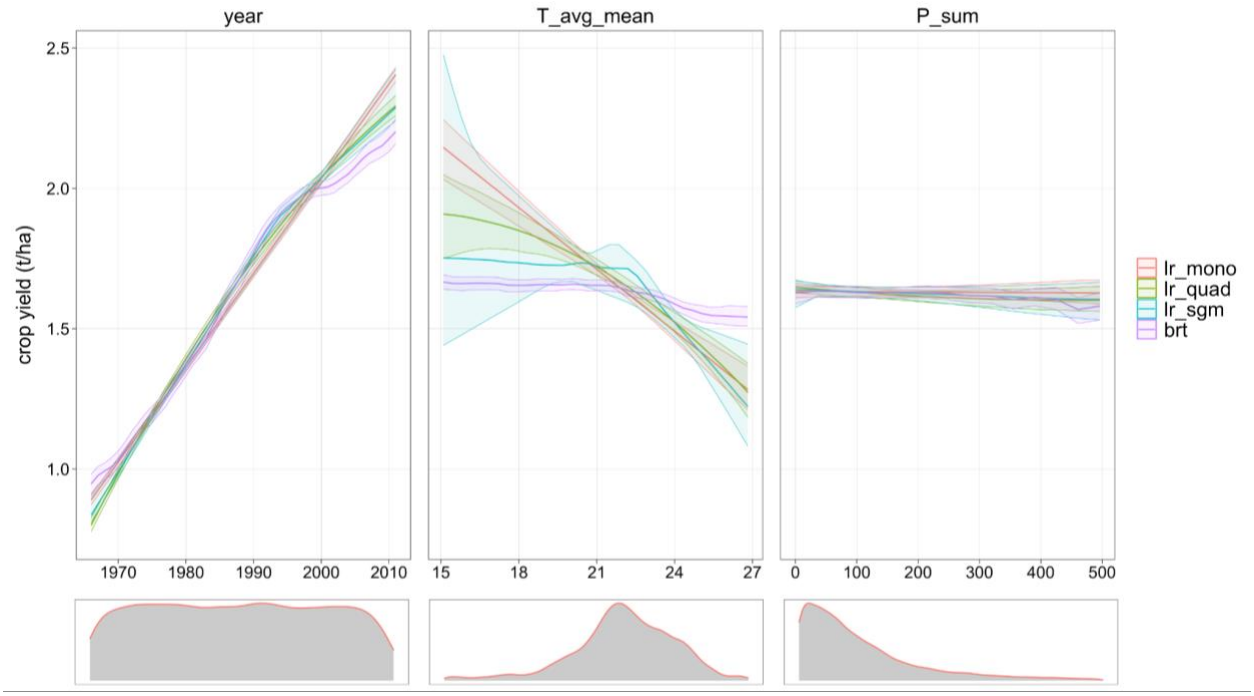

Figure B.3 Partial dependence plots of the *Tavg_Psum* variable set models for wheat (top row), and distribution density of corresponding IV in training data (bottom row). The four model types are color-coded in each panel. Data density plots at the bottom provide an idea of where most of the training data lies for a particular IV.

B.3 Analysis with synthetic data

Similar to Chapter 3 section 3.3.3., we created and analyzed synthetic data, but with an additional knot in the data at 22 degrees Celsius (Figure B.4). With exponentially changing climate, BRT again predicts a smaller impact of temperature on yields and the plot is flatter compared to the dotted line depicting the true functional form.

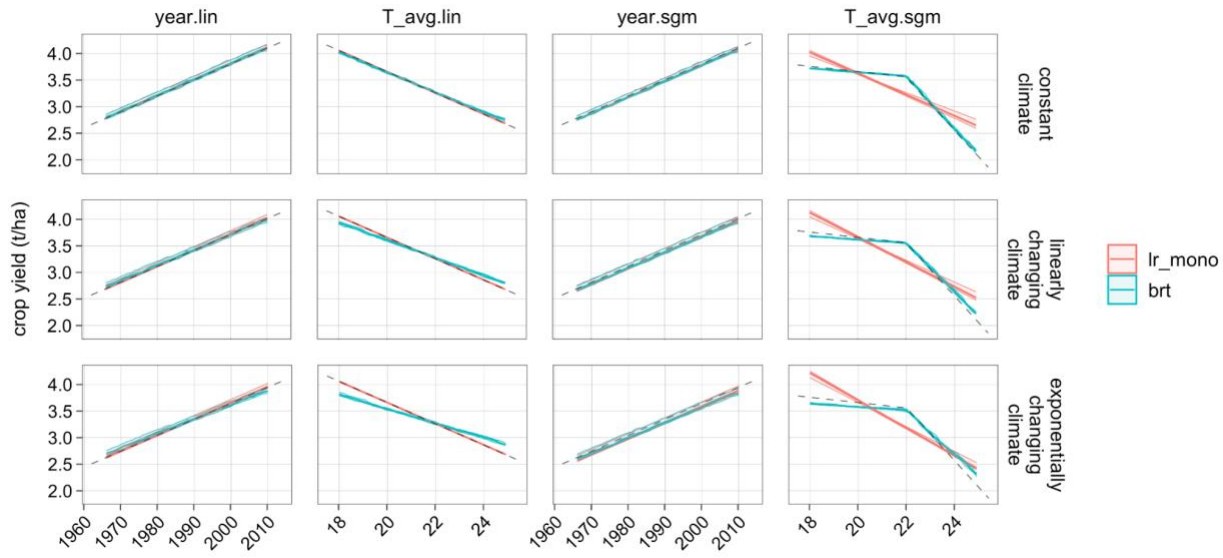

Figure B.4 Partial dependence plots for LR (red) and BRT (cyan) models fitted on synthetic data using user-defined coefficient for temperature. The LR is able to accurately model the relationship when there is no knot present (column 1; expected because the underlying data creation used a linear function), but the BRT sensitivity to temperature change goes down (the blue plot becomes flatter) as temperature gets more correlated with time (top to bottom, columns 2 and 4).

B.4 Climate change simulation for rice and wheat

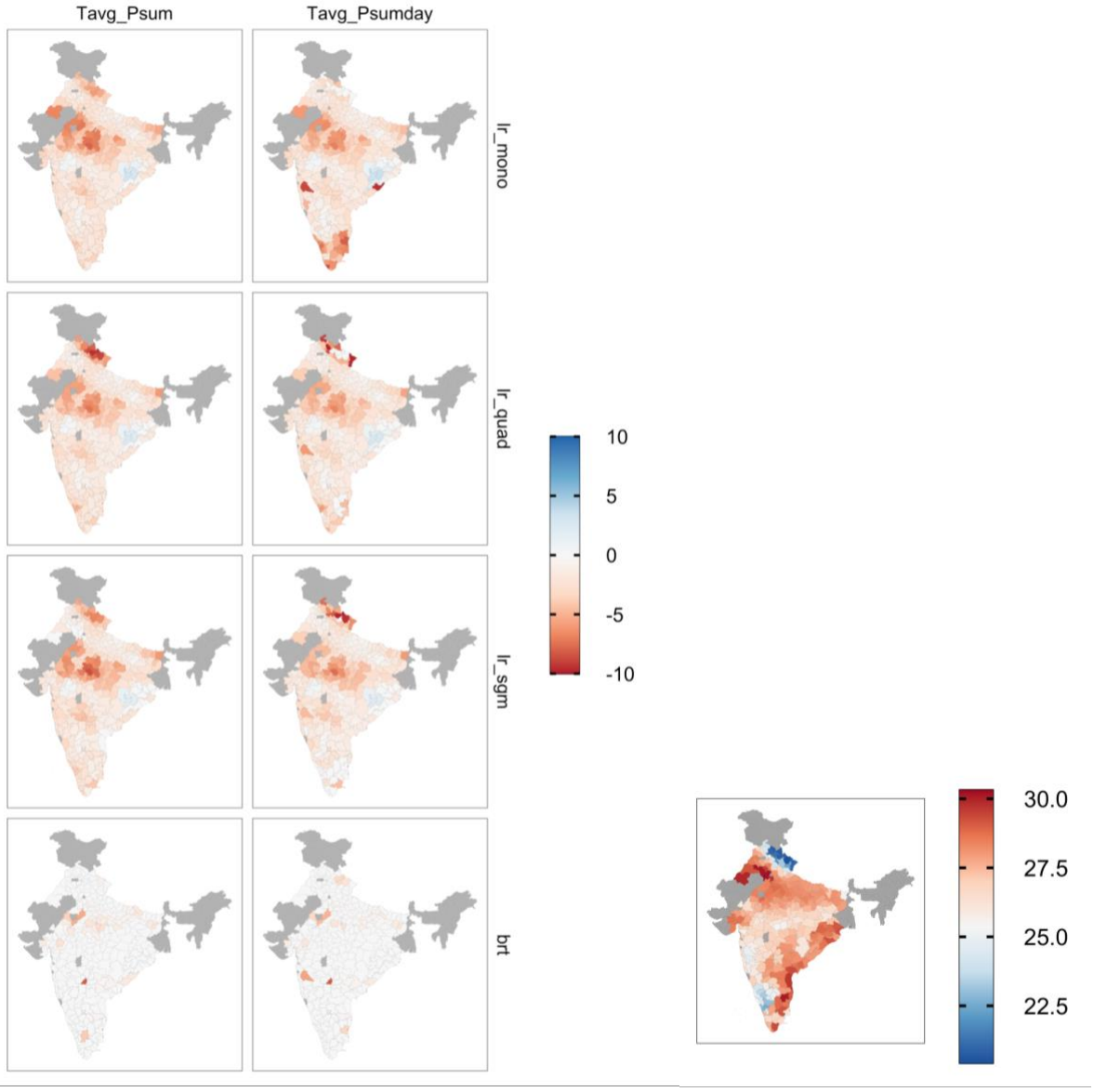

Figure B.5 Simulated impact of long-term climate change (since 1966) on rice yield in the last decade (2002-2011) of the study time period (left); 10-year (2002-2011) average of the mean temperature during rice season (right). The climate data was linearly detrended to remove time trend at district-scale. District-level estimates of median value and 95 percent confidence intervals of climate change impact on yield were obtained through residual bootstrapping (n = 500). The average district-level yield loss during the last decade in the dataset (2002-2011) is presented here as the expected impact of climate change that has occurred since 1966. Only

results with 95 percent significance of the confidence intervals are shown; insignificant results are shown in gray.

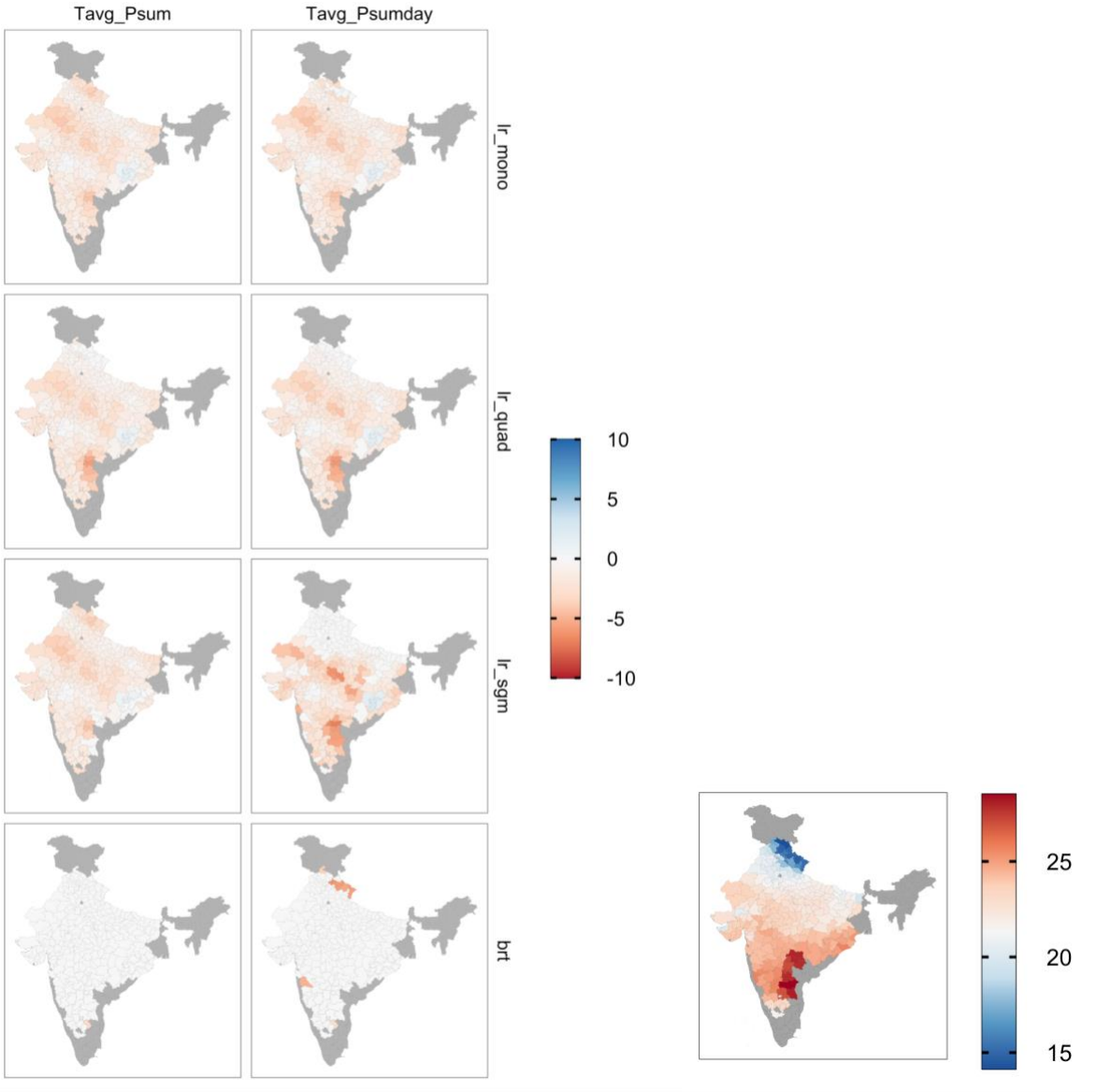

Figure B.6 Simulated impact of long-term climate change (since 1966) on wheat yield in the last decade (2002-2011) of the study time period (left); 10-year (2002-2011) average of the mean temperature during wheat season (right). The climate data was linearly detrended to remove time trend at district-scale. District-level estimates of median value and 95 percent confidence intervals of climate change impact on yield were

obtained through residual bootstrapping (n = 500). The average district-level yield loss during the last decade in the dataset (2002-2011) is presented here as the expected impact of climate change that has occurred since 1966. Only results with 95 percent significance of the confidence intervals are shown; insignificant results are shown in gray.

Appendix C Chapter 4

C.1 Soil moisture model development methodology

Soil field capacity

We acquired gridded soil profile data at 1 km x 1 km spatial resolution from SoilGrids (Hengl et al., 2017). This data contained soil composition (sand fraction, clay fraction, gravel content, and organic matter content) at multiple depths (0, 5, 15, 30, 60, 100, 200 cm). We harmonized this gridded data to the ICRISAT district boundaries, and created a district-wise multi-depth soil property dataset for India. Then we created soil columns or "buckets" for each crop. Total depth of the soil column for calculating field capacity was customized for each crop depending on root depth (Brück, Piro, Sattelmacher, & Payne, 2003; Williams et al., 1990); we assumed that any soil moisture beyond the root depth is inaccessible to a plant, hence total soil field capacity was adjusted accordingly for each crop. This soil column's maximum moisture content, or field capacity, (in mm of water) was calculated by applying the relevant pedotransfer functions (Saxton & Rawls, 2006) on soil characteristics at each depth that SoilGrids data is available for.

Evapotranspiration (ET) calculations

We constructed a simplified water balance model to estimate the influence of soil moisture deficit on crop yield. We first calculated reference evapotranspiration (ET_o) using Hargreaves method (Hargreaves & Allen, 2003). The choice of ET_o calculation method was dictated by historical data availability for our study's time period: Hargreaves' equation only requires temperature data, as opposed to the more popular and physically sound Penman-Monteith technique which needs temperature, humidity, and wind speed data (Allen et al., 1998). Importantly, Hargreaves and Penman-Monteith methods perform similarly for arid and semi-arid

unirrigated regions (Hargreaves & Allen, 2003), and Hargreaves' method is among the few, if not the only, temperature methods that have been shown to give globally valid ET_o estimates without requiring any local calibration (Allen et al., 1998). Even though Hargreaves' method only uses temperature data to calculate ET_o, it has been shown to simulate similar ET_o patterns across South Asia as Penman-Monteith technique (Aadhar & Mishra, 2020a). We calculated ET_o at daily timescale using equation (C.1):

$$ET_o = 0.0023\, R_a\, (TC + 17.8) TR^{0.5} \ , \qquad \text{(C.1)}$$

where R_a is location-specific daily extraterrestrial radiation; TC is average daily temperature; TR is daily temperature range (difference between daily minimum and maximum temperature). R_a was calculated using the methodology outlined in (De Gol, Festa, & Ratto, 1987) and also recommended by FAO (Allen et al., 1998):

$$R_a = 24(3600)/\pi \cdot I_{sc}\, E_o\, [\omega\, sin(\phi)\, sin(\delta) + cos(\phi) cos(\delta) sin(\omega) \ , \qquad \text{(C.2)}$$

$$E_o = 1 + 0.033/cos(2\pi d/365) \ , \qquad \text{(C.3)}$$

$$\omega = cos^{-1}(-tan(\phi)\, tan(\delta)) \ , \quad \text{if } tan(\phi)\, tan(\delta) \in [-1, 1] \qquad \text{(C.4)}$$

$$\omega = \pi \ , \quad \text{if } tan(\phi) tan(\delta) > 1 \qquad \text{(C.5)}$$

$$\omega = 0 \ , \quad \text{if } tan(\phi) tan(\delta) < -1 \qquad \text{(C.6)}$$

$$\phi = \pi/180 \cdot latitude\ (in\ degrees) \ , \qquad \text{(C.7)}$$

$$\delta = 0.4093\, sin(2\pi d/365 - 1.3943) \ , \qquad \text{(C.8)}$$

where R_a is extraterrestrial radiation (MJ m^{-2} day^{-1}); I_{sc} is the solar constant (1366 x 10^{-6} MJ m^{-2} sec^{-1}); E_o is eccentricity correction factor (or inverse relative distance Earth-Sun); ω is the sunrise hour angle (radians); φ is the latitude (radians); δ is the solar declination; d is the day number of the year. R_a was converted to mm/day units using equation (C.9):

$$R_a\ (in\ mm\ day^{-1})\ =\ 0.40756\ R_a\ (in\ MJ\ m^{-2}\ day^{-1})\ , \qquad \text{(C.9)}$$

ET_o thus calculated was subsequently used to calculate crop specific ET (ET_c) for each of our crops using equation (C.10):

$$ET_c\ =\ k_c\ ET_o\ , \qquad \text{(C.10)}$$

where k_c refers to individual crop ET coefficients (Allen et al., 1998). For our analysis, we were also interested in the subseasonal variation in ET_c, so we used subseasonal values of k_c for various growth stages of each crop.

Water balance model

Daily ET_c is the maximum evapotranspiration possible under water-sufficient conditions. We used the technique outlined in Ramankutty, Foley, Norman, & Mcsweeney (2002) to run our daily soil moisture model. The actual ET, ET_{act}, was calculated as the minimum of ET_c and available soil moisture, M:

$$ET_{act} = min(ET_c, M) \quad , \tag{C.11}$$

and the daily change in soil moisture amount was calculated using our bucket model:

$$\Delta M = P - ET_{act} \quad , \qquad \text{while } M \leq M_{max} \tag{C.12}$$

where P is daily precipitation (mm/day); M_{max} is soil field capacity, or maximum moisture held in soil after excess runoff. If $M > M_{max}$, it was reset to M_{max} and all excess water became runoff. The daily availability of water to plants, α, was expressed in terms of a moisture index:

$$\alpha = ET_{act} / ET_c \, . \tag{C.13}$$

α varied from 0 (fully dry soil) to 1 (no moisture stress, ET_{act} equals ET_c). We started our model two years before our study period to allow the soil moisture to come to equilibrium by the first year of our study. We then created five moisture availability categories, and binned each day of the crop growing season depending on daily moisture availability:

1. $\alpha = 1.00$
2. $\alpha < 1.00, \alpha \geq 0.75$
3. $\alpha < 0.75, \alpha \geq 0.50$
4. $\alpha < 0.50, \alpha \geq 0.25$
5. $\alpha < 0.25$

Relative importance of soil moisture models versus precipitation-based models

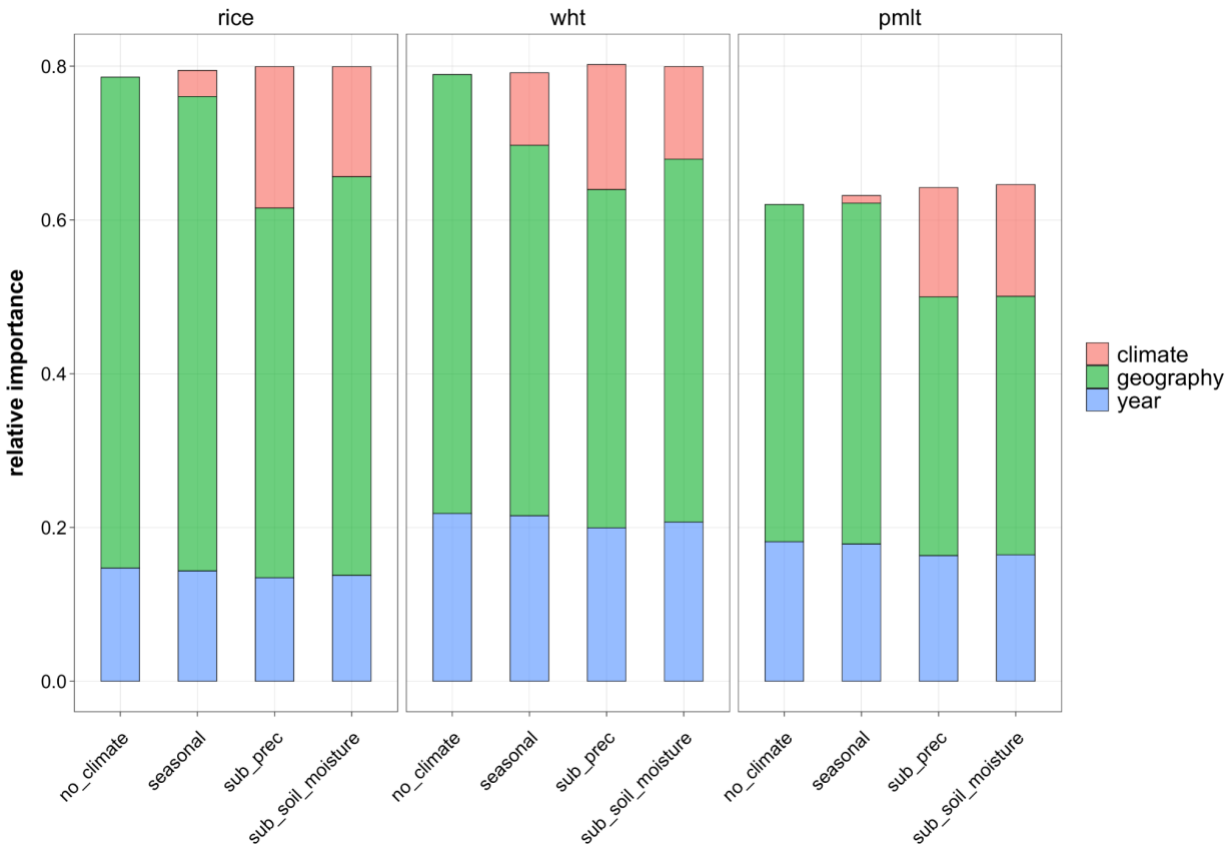

Figure C.1 Relative importance of time (blue), geography (green), and climate (red) variables across the no climate null model, and the three models analyzed for rice (left), wheat (center), and pearl millet (right). Note that the sum of the relative importances of time, geography, and climate variables equals R^2, which shows minimal improvement in overall model fit compared to the simplest null model on the left for each crop.

C.2 Temporal trend in various climate variables for a sample district

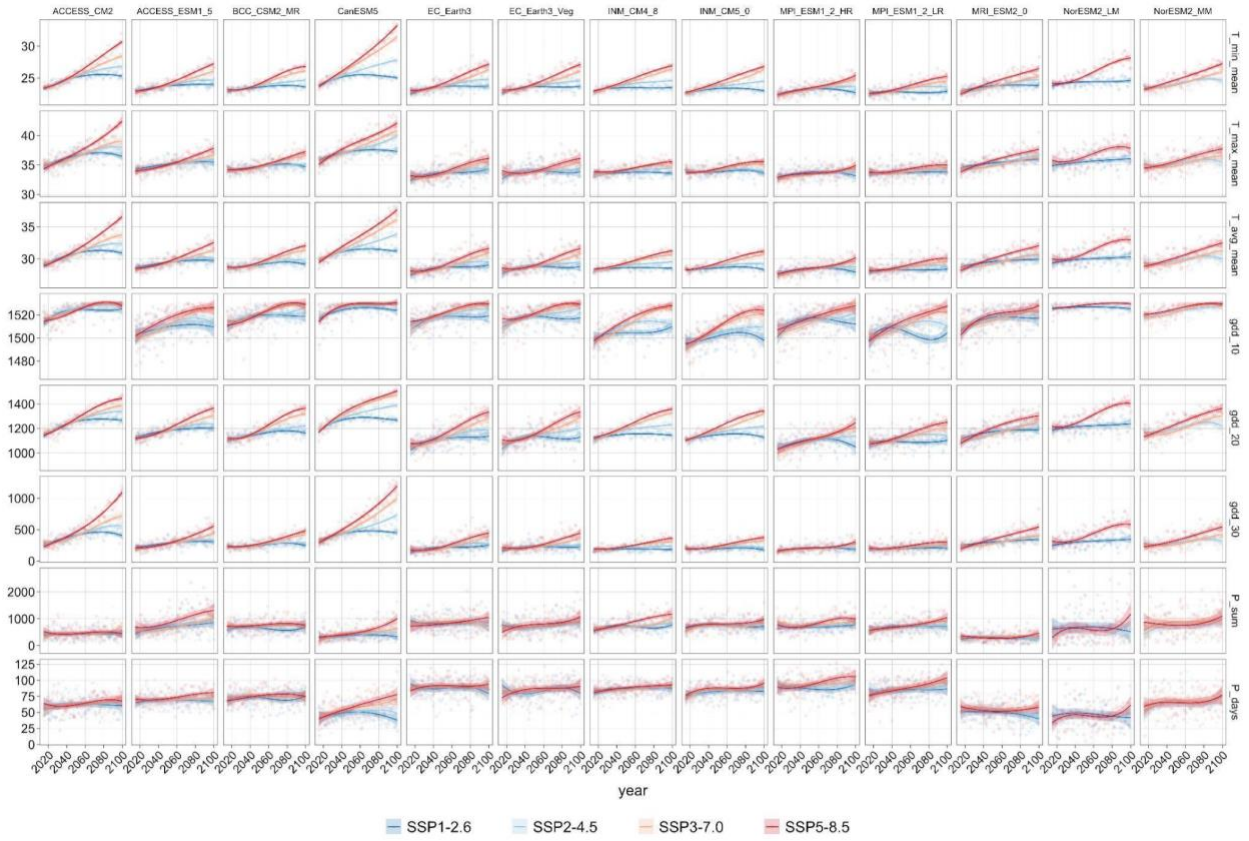

Figure C.2 Temporal trends in various climatic variables from 2020-2100 under different SSP scenarios (coded by color) for a sample district (Patiala (Punjab)) and rice (representative kharif crop). The columns from left to right show different GCMs analyzed in this study, and rows from top to bottom contain various climate variables.

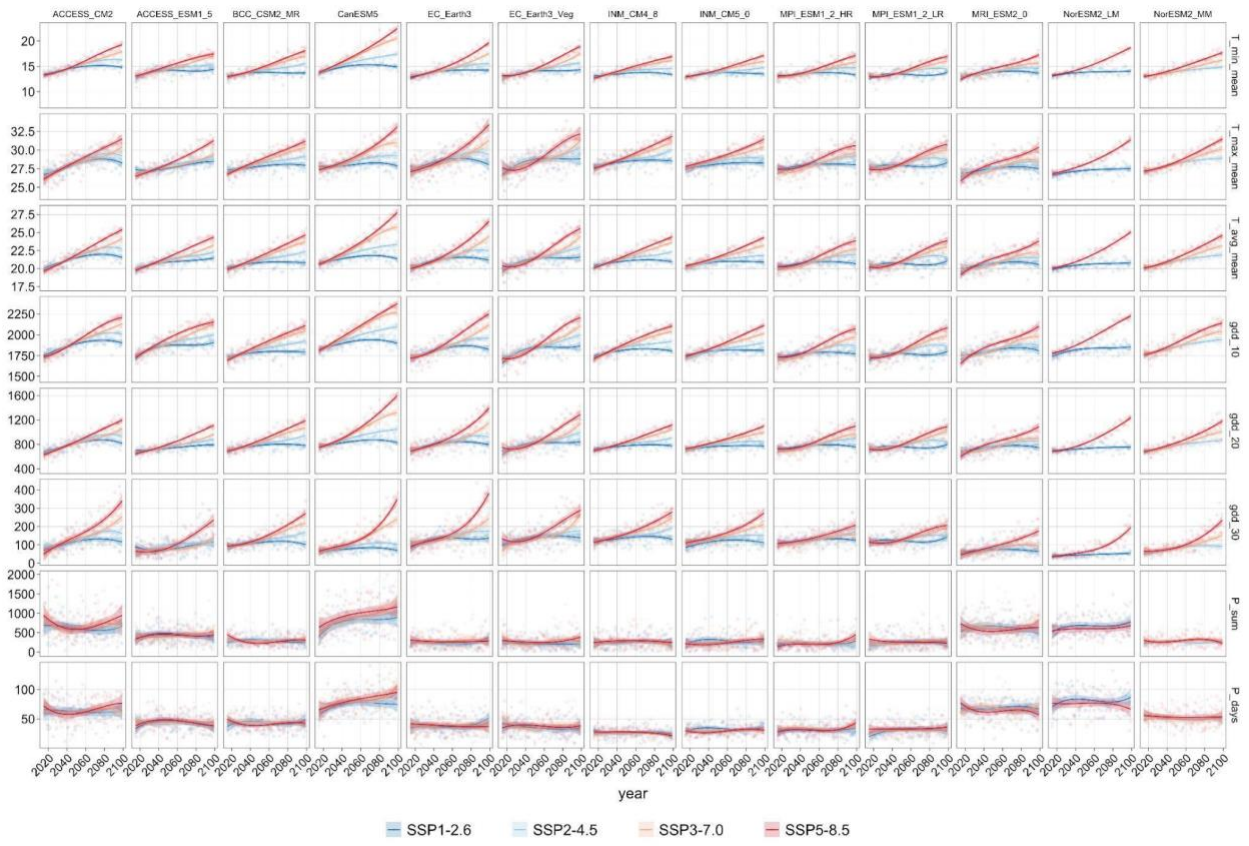

Figure C.3 Temporal trends in various climatic variables from 2020-2100 under different SSP scenarios (coded by color) for a sample district (Patiala (Punjab)) and wheat (representative rabi crop). The columns from left to right show different GCMs analyzed in this study, and rows from top to bottom contain various climate variables.

C.3 Soil moisture trends for a sample district

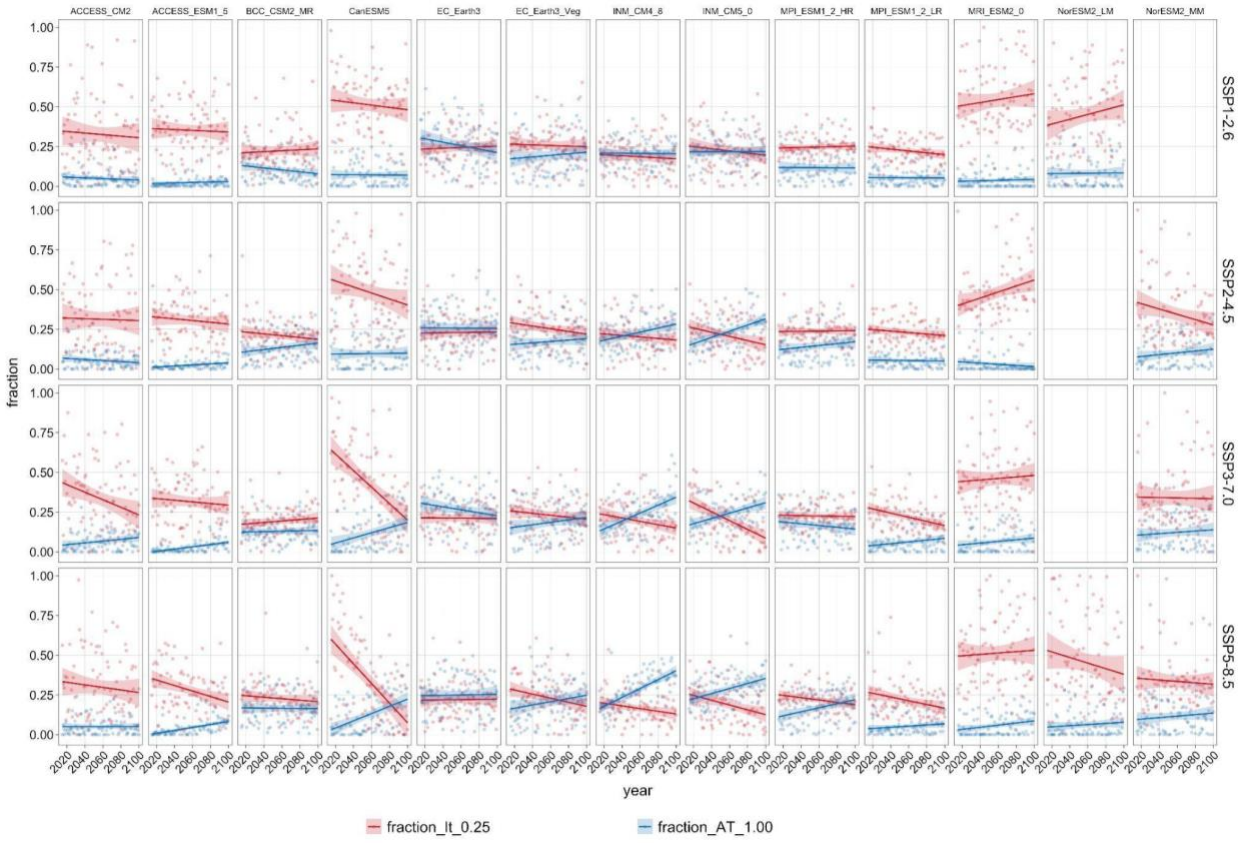

Figure C.4 Temporal trends in soil moisture amount from 2020-2100 under different SSP scenarios (top to bottom) for a sample district (Patiala (Punjab)) and rice (representative kharif crop). The columns from left to right relate to different GCMs analyzed in this study. Red plots show the fraction of growing season spent by the crop in the lowest water availability bin (days with less than 25 percent of daily evapotranspiration demand met), and blue show fraction of growing season spent under zero water stress (sufficient soil moisture to meet full crop evapotranspiration demand). The general trend reveals an increase in soil moisture availability over time (blue plots going up, red plots going down).

C.4 Growing degree days

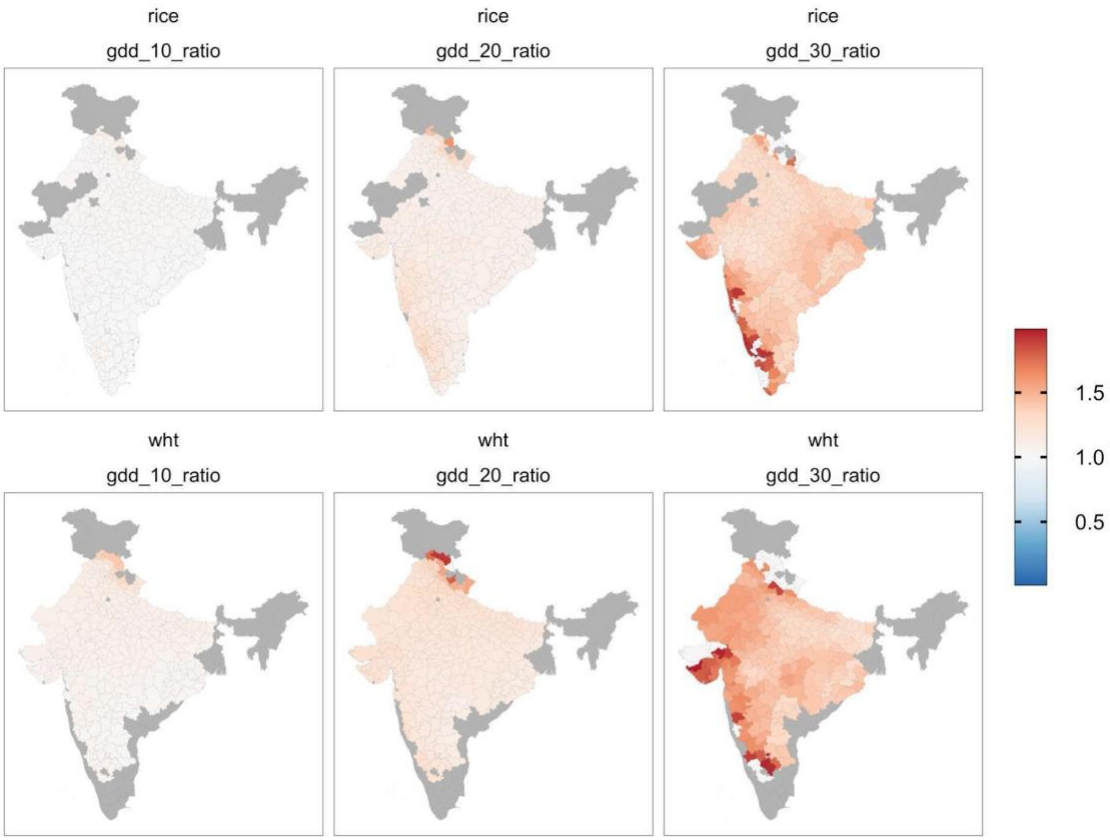

Figure C.5 gdd_10 (left), gdd_20 (middle), and gdd_30 (right) as a ratio of the corresponding variable for the reference climatology (1951-2000). Plots depict representative kharif (rice; top) and rabi (wheat; bottom) crops, for near-term (2041-2060) changes for the SSP2-4.5 scenario. Median values of district-wise projections from the 13 GCMs were used to produce these plots.

C.5 Precipitation amount and precipitation days

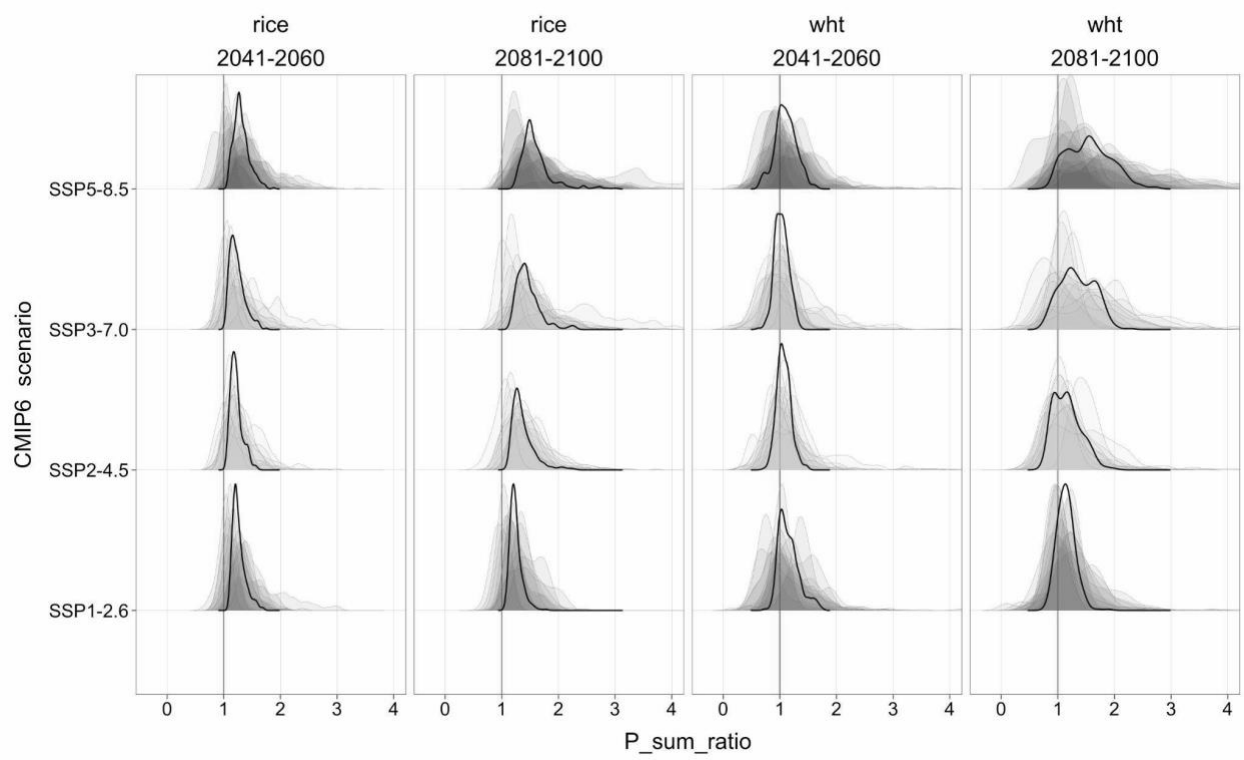

Figure C.6 Distribution of district-wise ratio of total seasonal precipitation (to reference climatology precipitation) for representative kharif (rice; columns 1 and 2) and rabi (wheat; columns 3 and 4) crops, both for the short-term (2041-2060; columns 1 and 3) and long-term (2081-2100; columns 2 and 4). Different SSPs are depicted on the y-axis, ordered by intensity of emissions from bottom to top. The semi-transparent density plots depict distributions from each of the 13 GCMs analyzed in this study; the bold black line is the distribution of the median projection of the 13 GCMs for each district.

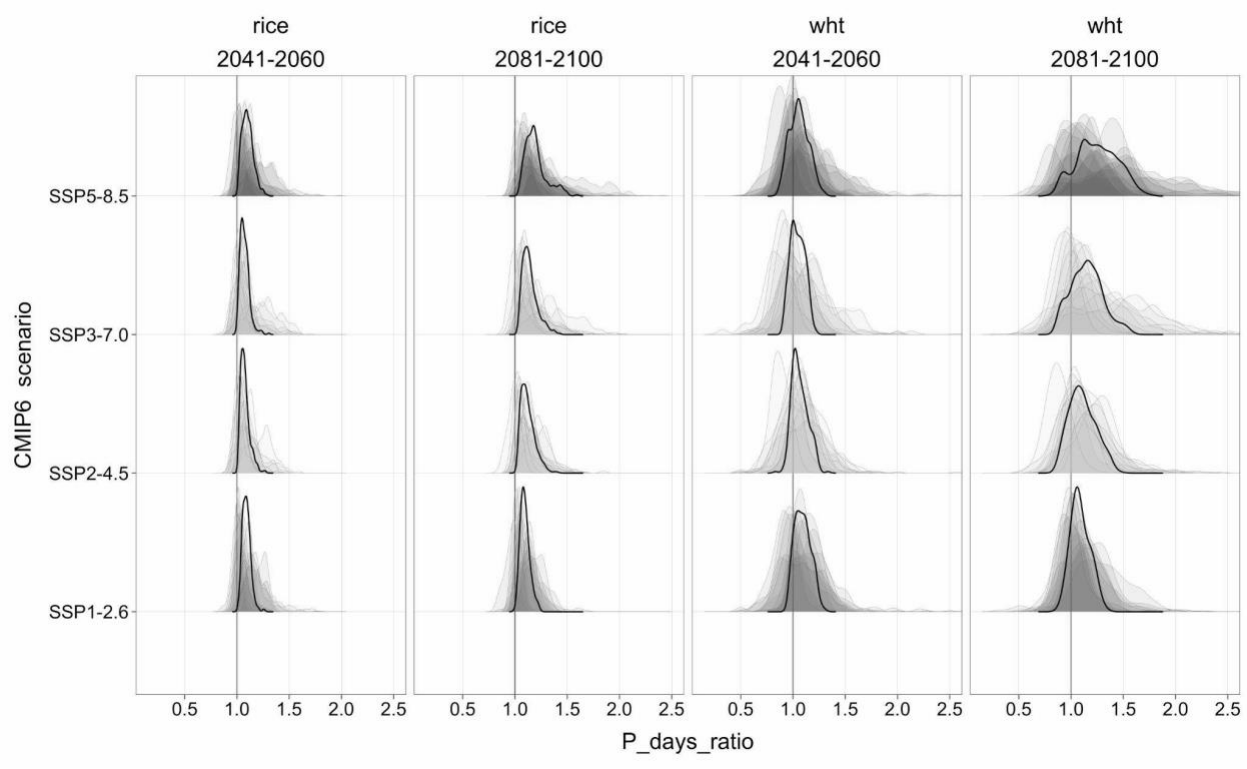

Figure C.7 Distribution of district-wise ratio of total seasonal precipitation days (to reference climatology precipitation days) for representative kharif (rice; columns 1 and 2) and rabi (wheat; columns 3 and 4) crops, both for the short-term (2041-2060; columns 1 and 3) and long-term (2081-2100; columns 2 and 4). Different SSPs are depicted on the y-axis, ordered by intensity of emissions from bottom to top. The semi-transparent density plots depict distributions from each of the 13 GCMs analyzed in this study; the bold black line is the distribution of the median projection of the 13 GCMs for each district.

C.6 National percent loss in crop yield

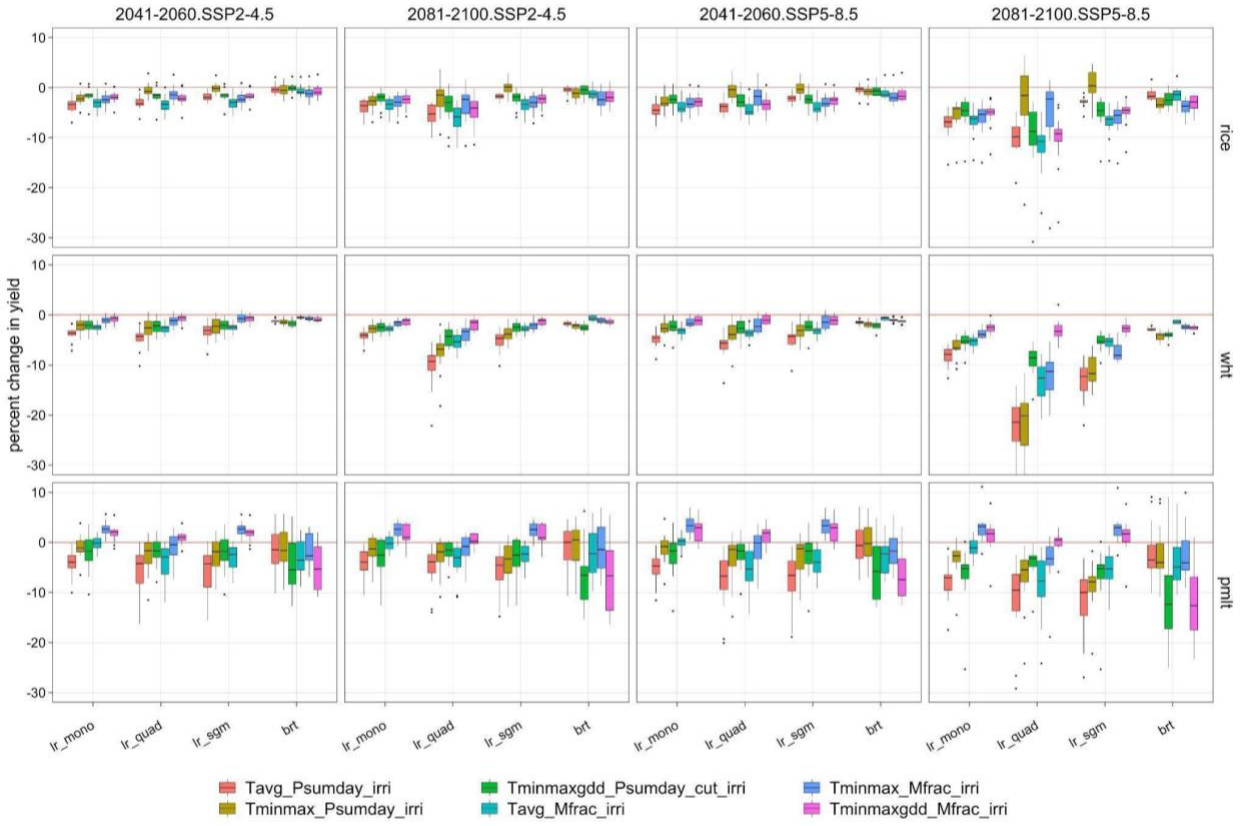

Figure C.8 Nationally-averaged percent change in yield for rice (top), wheat (middle), and pearl millet (bottom). Columns 1-4 depict SSP2-4.5 near-term, SSP2-4.5 long-term, SSP5-8.5 near-term, and SSP5-8.5 long-term. The plots are color coded by climate variable set. Within each panel, boxplots are grouped by model types from left to right: lr_mono, lr_quad, lr_sgm, and brt. Boxplots show median values of estimates from 13 GCMs.

C.7 Predicted reduction in crop yield (all climate variable sets)

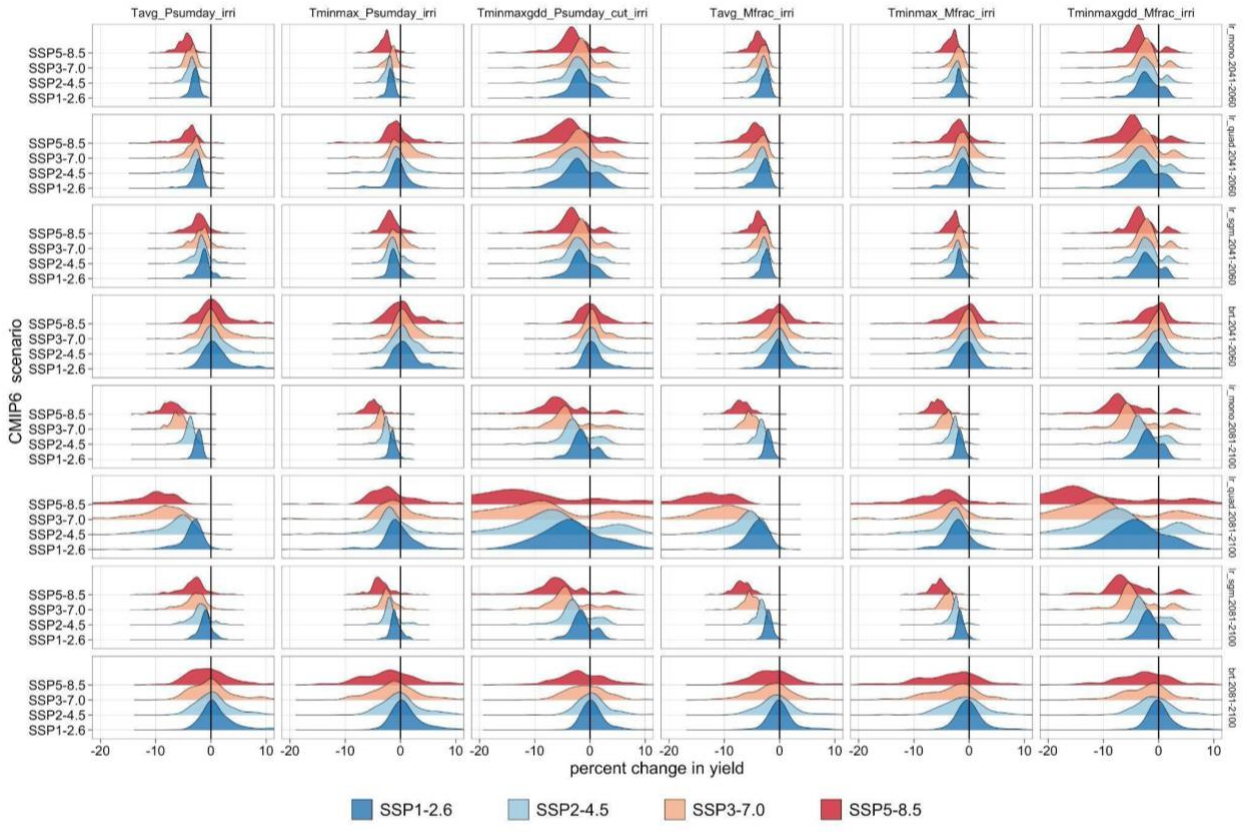

Figure C.9 Distribution of district-level percent change in yield for rice in the short-term (2041-2060). Rows 1-4 show the four model types: lr_mono, lr_quad, lr_sgm, and brt. Columns 1-6 depict all climate variable sets analyzed in this study. SSP scenarios are color-coded within each panel. Plots show median values of estimates from 13 GCMs.

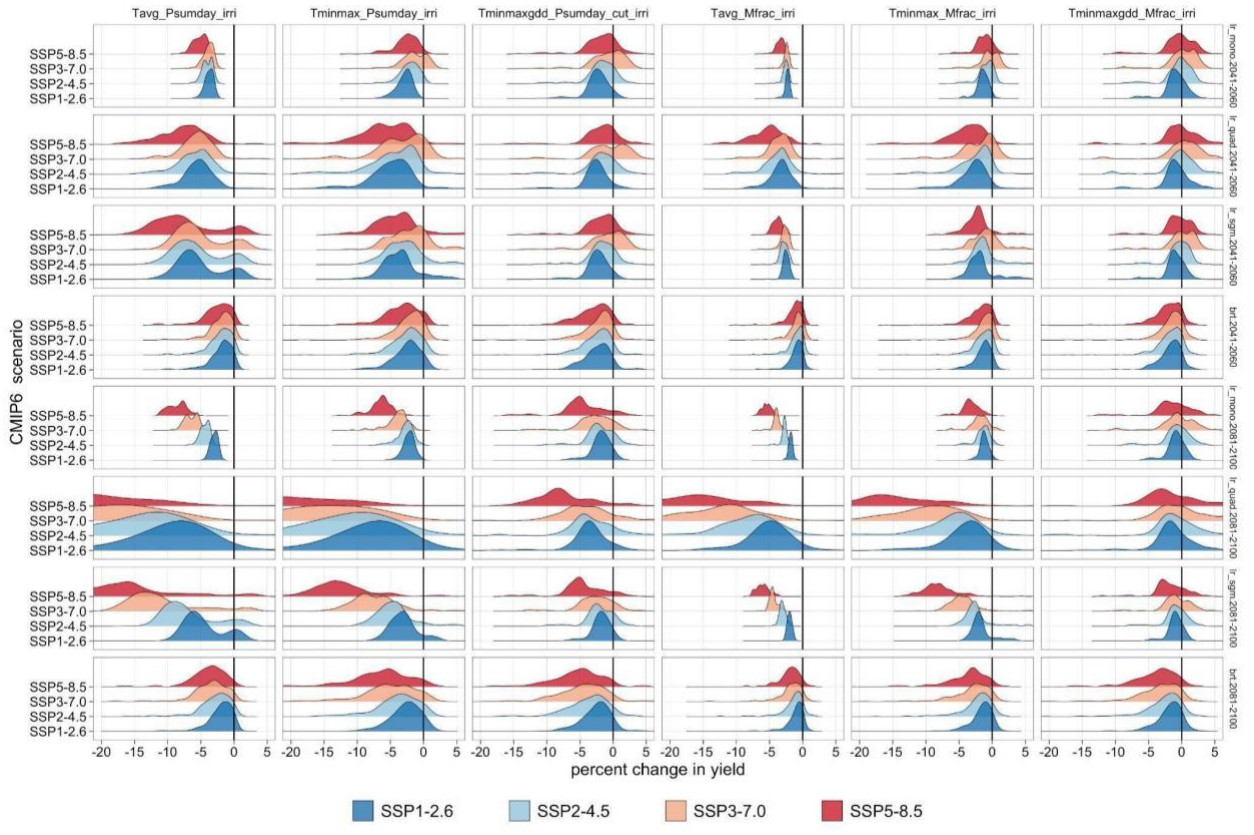

Figure C.10 Distribution of district-level percent change in yield for wheat in the short-term (2041-2060). Rows 1-4 show the four model types: lr_mono, lr_quad, lr_sgm, and brt. Columns 1-6 depict all climate variable sets analyzed in this study. SSP scenarios are color-coded within each panel. Plots show median values of estimates from 13 GCMs.

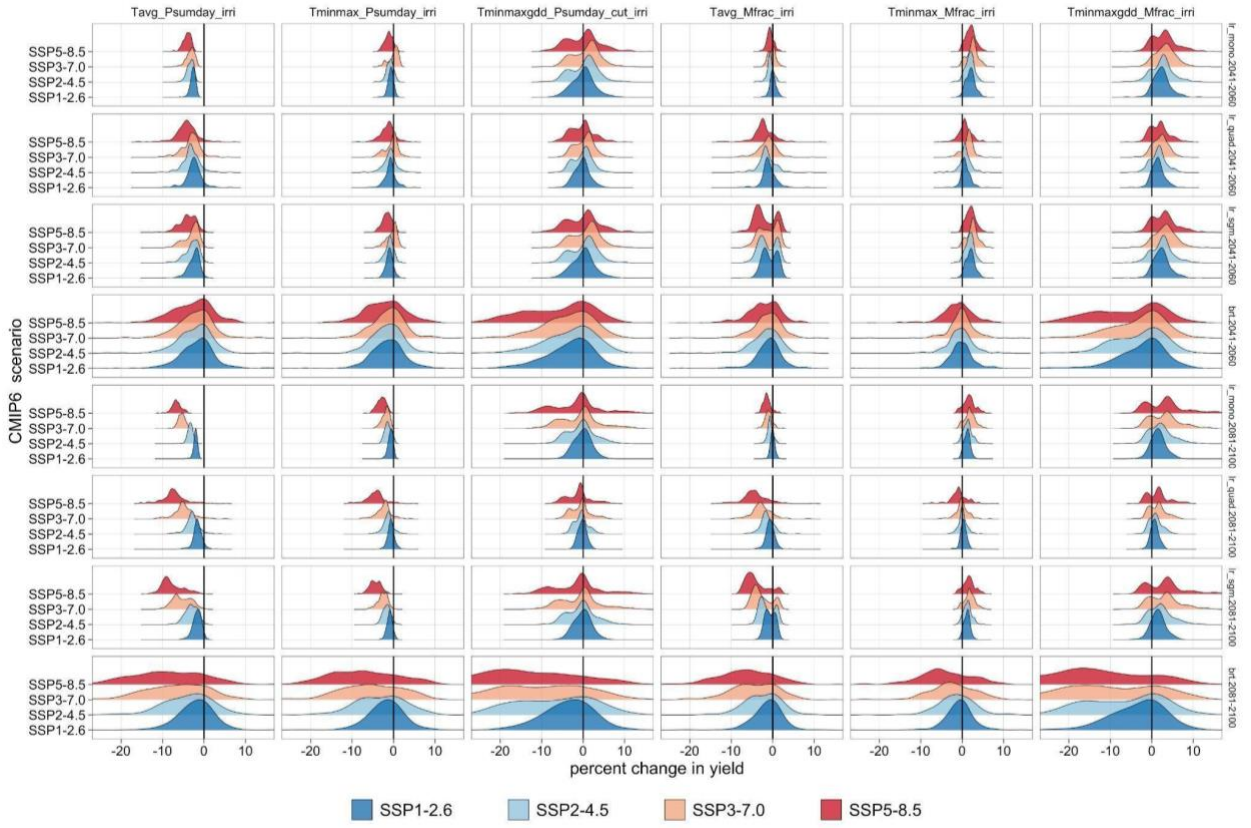

Figure C.11 Distribution of district-level percent change in yield for pearl millet in the short-term (2041-2060). Rows 1-4 show the four model types: lr_mono, lr_quad, lr_sgm, and brt. Columns 1-6 depict all climate variable sets analyzed in this study. SSP scenarios are color-coded within each panel. Plots show median values of estimates from 13 GCMs.