Uncertainty and Predictability of Seasonal-to-Centennial Climate Variability

Nathan J. L. Lenssen

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2022

Abstract

Uncertainty and Predictability of Seasonal-to-Centennial Climate Variability

Nathan J. L. Lenssen

The work presented in this dissertation is driven by three fundamental questions in climate science: (1) What is the natural variability of our climate system? (2) What components of this variability are predictable? (3) How does climate change affect variability and predictability? Determining the variability and predictability of the chaotic and nonlinear climate system is an inherently challenging problem. Climate scientists face the additional complications from limited and error-filled observational data of the true climate system and imperfect dynamical climate models used to simulate the climate system. This dissertation contains four chapters, each of which explores at least one of the three fundamental questions by providing novel approaches to address the complications.

Chapter 1 examines the uncertainty in the observational record. As surface temperature data is among the highest quality historical records of the Earth's climate, it is a critical source of information about the natural variability and forced response of the climate system. However, there is still uncertainty in global and regional mean temperature series due to limited and inaccurate measurements. This chapter provides an assessment of the global and regional uncertainty in temperature from 1880-present in the NASA Goddard Institute for Space Studies (GISS) Surface Temperature Analysis (GISTEMP).

Chapter 2 extends the work of Chapter 1 to the regional spatial scale and monthly time scale. An observational uncertainty ensemble of historical global surface temperature is provided for

easy use in future studies. Two applications of this uncertainty ensemble are discussed. First, an analysis of recent global and Arctic warming shows that the Arctic is warming four times faster than the rest of the global, updating the oft-provided statistic that Arctic warming is double that of the global rate. Second, the regional uncertainty product is used to provide uncertainty on country-level temperature change estimates from 1950-present.

Chapter 3 investigates the impacts of the El Niño-Southern Oscillation (ENSO) on seasonal precipitation globally. In this study, novel methodology is developed to detect ENSO-precipitation teleconnections while accounting for missing data in the CRU TS historical precipitation dataset. In addition, the predictability of seasonal precipitation is assessed through simple empirical forecasts derived from the historical impacts. These simple forecasts provide significant skill over climatological forecasts for much of the globe, suggesting accurate predictions of ENSO immediately provide skillful forecasts of precipitation for many regions.

Chapter 4 explores the role of initialization shock in long-lead ENSO forecasts. Initialized predictions from the CMIP6 decadal prediction project and uninitialized predictions using an analogue prediction method are compared to assess the role of model biases in climatology and variability on long-lead ENSO predictability. Comparable probabilistic skill is found in the first year between the model-analogs and the initialized dynamical forecasts, but the initialized dynamical forecasts generally show higher skill. The presence of skill in the initialized dynamical forecasts in spite of large initialization shocks suggest that initialization of the subsurface ocean may be a key component of multi-year ENSO skill.

Chapter 5 brings together ideas from the previous chapters through an attribution of historical temperature variability to various anthropogenic and natural sources of variability. The radiative forcing due to greenhouse gas emissions is necessary to explain the observed variability in temperature nearly everywhere on the land surface. Regional fingerprints of anthropogenic aerosols are detected as well as the impact of major sources of natural variability such as ENSO and Atlantic Multidecadal Variability (AMV).

Table of Contents

List of Figures

List of Tables

Acknowledgements

Thank you to the many, many people in my life who have supported me and made this accomplishment possible. In particular, to my wife Stephanie, my parents Maureen and Nick, my brothers Michael and Kieran, and many other family and friends for their constant faith in me and encouragement over this program and throughout my life. Thanks to Doug Nychka, and later Dorit Hammerling, for giving me the opportunity to experience climate research as a high school student and undergraduate, to Tian Zheng for supporting me in finding my path to DEES, and to the numerous DEES/APAM professors and students who convinced me how fun and rewarding the DEES Ph.D. would be. Particularly, thanks to Gavin Schmidt and Lisa Goddard for giving me the chance to work and study as a climate scientist and providing me with invaluable mentorship on science and life over the past five plus years. Thanks to Simon Mason, Yochanan Kushnir, and Mingfang Ting for serving on my past and present committees and guiding me through this endeavor. A big thanks to all of my professors, classmates, staff, and friends who made the process of learning exciting over the past years, even through the dark days of COVID. A thank you to my students and mentees who provided constant joy and inspiration. This dissertation would not be possible without many funders including the National Science Foundation through the Graduate Research Fellowship Program, the Food and Agriculture Organization of the United Nations support of Chapters 1 and 2, and Columbia University's World Project ACToday for supporting Chapters 3–5.

Dedication

To Lisa Goddard, who I miss dearly. May she live on through the innumerable lives she has enriched through her teaching, research, and advocacy.

Introduction

Over the past century, the study of the Earth's climate has evolved from a primarily empirical field devoted to observing the atmosphere, oceans, and land surface to a science tasked with understanding the physics, statistics, and societal consequences of variations and changes in the climate system. The effect of human behaviors, primarily the release of greenhouse gases from burning of fossil fuels, has added urgency to the study of the climate system. In this dissertation, three fundamental questions in climate science are explored with each of the chapters detailing a novel contribution the scientific community's understanding of these questions. The three questions are:

Q1) What is the natural variability of the Earth's climate system?

Q2) What components of the Earth's natural variability are predictable?

Q3) How does climate change affect variability and predictability?

Improving quantitative estimates of the natural variability and predictability of our climate system is critical for understanding the past, present, and future climate of Earth. That is, the community must improve and validate the estimates of key statistics of the climate as well as properly quantify the uncertainty of these estimates and connect such estimates back to physical processes.

These three fundamental questions are incredibly broad, and have been central questions in climate science and meteorology for decades. When investigating aspects of these questions, the community has two primary sources of data: direct historical observations of the climate system and simulations of the Earth's climate from dynamical climate models. Attempting to summarize

the behavior of the chaotic, coupled climate system is an immensely challenging problem in even the most idealized problems, but the scientific community community faces two major limitations due to the available data:

L1) Observations of the Earth's climate and weather are limited in spatial and temporal extent and often contain errors

L2) Simulations of the Earth's climate with dynamical climate models are biased representations of the true dynamical system

In this dissertation, the consequences of these data challenges are explored and new methodology is presented to quantify these limitations as well as assess the impact of these limitations on results.

The natural variability of the climate system refers to changes in the mean climate state that would be present even without changes in the climate system due to external forcing (Leith 1973; Madden 1976). Natural variability is traditionally defined through the standard deviation or similar statistics of time-averaged weather, or more generally, through the power spectral density (Madden 1976). The generality of the power spectrum is useful as the climate varies on timescales from hours due to weather systems to millions of years due to changes in the Earth's tectonics (Raymo & Ruddiman 1992; Ghil 2002). In the work presented in this dissertation, natural variability is investigated on seasonal timescales as defined by the statistics of monthly or seasonal means and decadal and multi-decadal timescales as defined by the statistics of annual mean values.

It is critical to properly quantify the relevant natural variability for nearly every question in climate science. A non-exhaustive list of major sub-fields that necessitate characterizing natural variability are: determining the sensitivity of the global mean temperature to changes in external forcing (Sherwood *et al.* 2020), the detection and attribution of human activities on long-term climate trends (Hegerl *et al.* 2006) and extreme events (Trenberth *et al.* 2015), determining the predictability limits of the climate system at various timescales from weeks to decades (Merryfield *et al.* 2020; Meehl *et al.* 2021), assessing the fidelity of dynamical climate model simulations of historical and future Earth climate (Deser *et al.* 2020), and assessing the short-term impacts of

climate change on human systems (Schwarzwald & Lenssen *in review*).

Although natural variability is often presented as climatic noise (Madden 1976), some components of natural variability are predictable, providing knowledge of the climate in future months or years. Since predictability of the atmosphere is limited to about 14 days in the future due to chaotic and fast-evolving nature of the system (Lorenz 1969b), climate predictability at longer lead times arises from predictable components of slowly-evolving components of the Earth system such as ocean, ice, and land surface processes (Goddard 2012). As an example, the air-sea coupled process in the tropical Pacific known as the El Niño-Southern Oscillation (ENSO) can be predicted a few months (Cane *et al.* 1986; Barnston *et al.* 2019) to a few years in advance (Gonzalez & Goddard 2016) and has global impacts on precipitation and temperature (Mason & Goddard 2001; Lenssen *et al.* 2020). Understanding if, how, and when these slowly-evolving components can be predicted and how they affect the weather and climate in regions important to humans is useful to support activities critical to society such as agriculture (Rahman *et al.* 2016), water management (Crochemore *et al.* 2016), and public health (Borbor-Mendoza 2016).

In addition to natural variability, the climate has been changing due to human activities, primarily the burning of fossil fuels. The resulting release of carbon dioxide (CO_2), methane, and other radiatively active greenhouse gases have resulted in changes to the climate including an increase in global (Hansen *et al.* 1981; Lenssen *et al.* 2019) and regional mean temperature (Gulev *et al.* 2021), changes to the hydroclimate including increased likelihood of drought in subtropical regions such as southwestern North America (Held & Soden 2006; Williams *et al.* 2022), and more extreme precipitation events (Fischer & Knutti 2016). The regular assessment reports by the Intergovernmental Panel on Climate Change (IPCC) provide an exhaustive review of the science of climate change (Masson-Delmotte *et al.* 2021) and the impacts of climate change on human systems (Pörtner *et al.* 2022).

Chapters 1 and 2 of this dissertation present an investigation of uncertainty in the observed global surface temperature record from 1880-present. Chapter 1 quantifies the uncertainty in the global, annual mean as well as other large-scale annual mean series from 1880-2016 (Lenssen *et*

al. 2019). Chapter 2 expands on this methodology to generate an ensemble of gridded monthly temperature fields that properly characterize the spatial and temporal statistics of observational uncertainty in gridded monthly temperature anomalies from 1880-2020. The observed surface temperature record is a critical data source for investigating the natural variability and forced response of the climate system, with the investigations presented in these chapters contributing to fundamental questions (Q1) and (Q3). In addition, this work addresses the limitation (L1) by better quantifying uncertainty and providing numerous example applications that highlight how accounting for uncertainty in observed climate data is critical for properly assessing the impacts of climate change.

Chapters 3 and 4 of this dissertation investigate the ENSO-driven predictability of global precipitation at lead times of months and years. Chapter 3 determines the global impact of ENSO on seasonal precipitation using historical observations and compares the skill of predictions made with these historical impacts to state-of-the-art climate forecasts (Lenssen *et al.* 2020). The significant predictive skill of the historical impact forecasts suggests that useful precipitation forecasts could be issued further in advance if ENSO could be predicted further in advance. Thus, Chapter 4 explores the potential predictability of ENSO multiple years in advance using methods that leverage existing global dynamical climate simulations. These studies provide insight into the ENSO-driven variability and predictability of the climate system, addressing fundamental questions (Q1) and (Q2). Chapter 3 also presents novel methodology for approaching missing data in the precipitation record, addressing limitation (L1) while Chapter 4 investigates biases in climate models addressing limitation (L2). Together, these chapters suggest exciting opportunities for extending climate prediction from the current nine months forecasts to two or three years into the future.

Chapter 5 presents a decomposition of observed global and regional temperature into variability associated with natural variability as well as anthropogenic and naturally occurring radiative forcings. This work provides insight towards answering all three fundamental questions as the results highlight the roles of natural variability and external forcings in explaining temperature variations

over much of the land surface. In addition, this investigation provides important information for the development of decadal prediction by identifying regions where temperature variability has historically been associated with variations in predictable modes of natural variability.

The remainder of this introduction provides deeper history and background relevant to the work presented in the chapters. Section 1 provides an introduction to global temperature observations and analyses, providing context for Chapters 1 and 2. Section 2 provides a deeper introduction to climate variability and prediction providing context for Chapters 3 - 5.

0.1 Historical Observed Global Temperature

The global mean surface temperature is a critical statistic for climate science as it is directly linked to the energy balance of the Earth system. The mean surface temperature is controlled by the intensity of incoming solar radiation, the reflectivity of the Earth to solar radiation, and the emissivity of the atmosphere to radiation emitted by the Earth. The reflectivity is commonly referred to as the Earth's albedo. The emissivity of the atmosphere to the infrared radiation emitted by the Earth is colloquially referred to as the greenhouse effect and is caused by radiatively active gases including water vapor, CO_2, methane, and ozone that absorb and reemit outgoing radiation. The relationship between global temperature, albedo, and emissivity have been long studied with early 0-d and 1-d models of the Earth, providing important results to explain the effect of solar radiation, albedo, and greenhouse gas concentrations on the Earth's equilibrium temperature (Manabe & Wetherald 1967; Budyko 1969; Sellers 1969). In addition, foundational work in linking greenhouse gas emissions to global climate change critically included direct observations of the historical global surface temperature that linked the increase in atmospheric CO_2 to an increase in global mean temperature (Callendar 1938; Hansen *et al.* 1981).

The earliest study that used measurements of temperature from metrological stations to support claims that increased greenhouse gas emissions resulted in a substantial increase in global mean temperature was the work of Callendar (1938). Here, he used data from 147 weather stations

between 60°S and 60°N to make an estimate of global temperature change from 1880-1935 suggesting an increase of around 0.25°C. He extended his analysis study was extended to 1870-1950 in Callendar (1961) with over 400 stations. These studies remarkably match modern estimates of the 60°S - 60°N mean temperature, despite using many less station records(Hawkins & Jones 2013). In addition to using data and idealized models to support the proposed link between greenhouse gas emissions and global temperature, Callendar (1961) also discussed the effect of other sources of variability in global mean temperature including the solar radiative forcing due to sunspot activity as is further explored in Chapter 5 of this dissertation. The work of Callendar in these two studies set the stage for modern climate science, including the three fundamental questions discussed here and the challenge of working with limited observational data.

Improved idealized models of the Earth's energy balance (Manabe & Wetherald 1967; Budyko 1969; Sellers 1969), as well as the first general circulation models (GCMs) of the Earth's atmosphere (Phillips 1956; Manabe & Bryan 1969), provided further evidence for the link between greenhouse gases and global mean temperature. This work was synthesized in a report by the US National Research Council in 1979 (now known as the "Charney Report") that proposed a link between greenhouse gases and global temperature increase in no uncertain terms, suggesting a climate sensitivity of 2 - 4.5°C (National Research Council 1979), a range that is in line with modern estimates (Sherwood *et al.* 2020). However, these analyses were conducted primarily with physics-based models (Manabe & Wetherald 1967; Manabe & Bryan 1969) of the Earth system or used observations primarily from the northern hemisphere (Callendar 1938; Callendar 1961; Mitchell Jr. 1961; Budyko 1969).

The first fully-global analysis of mean surface temperature was conducted at National Aeronautics and Space Administration (NASA) Goddard Institute for Space Studies (GISS) and used around 1,000 stations globally to validate results from a 1-d energy balance model (Hansen *et al.* 1981). This method was refined and the number of stations was increased to more than 2,000 a few years later (Hansen & Lebedeff 1987). The method presented in Hansen & Lebedeff (1987) is still the backbone of the NASA GISS Surface Temperature Analysis (GISTEMP) and is summarized

in Chapter 1 of this dissertation. Simultaneously, the Climatic Research Unit (CRU) at the University of East Anglia was developing an independent estimate of global temperature that would become the British Met Office Hadley Centre/Climatic Research Unit global surface temperature data set (HadCRUT). They published their first hemispheric products and global analyses in 1986 (Jones *et al.* 1986a; Jones *et al.* 1986b; Jones *et al.* 1986c). Both GISTEMP and HadCRUT became operational data products with monthly updates to their estimates of global and regional mean temperature, providing critical near-realtime information about the temperature of the Earth. Over the following decades, additional independent operational global surface temperature analyses were developed by the Japanese Metrological Agency (Ishihara 2006), the U.S. National Ocean and Atmospheric Administration (NOAA) (Vose *et al.* 2012) and Berkeley Earth (Rohde *et al.* 2013a).

Estimates of global temperature are made from a limited number of error-filled station and ship records. Thus, it is important to quantify the uncertainty in estimates made by these operational products. Initially, the HadCRUT product led the way in uncertainty quantification with numerous studies on the topic published in the 1990s and 2000s (See Jones (2016) for a review). The GISTEMP analysis also continued to be improved with various adjustments to the averaging method to improve estimates in regions with limited station coverage and iterative improvements to remove non-climatic biases due to changes in station location and urban heating effects (Hansen *et al.* 1999; Hansen *et al.* 2001; Hansen *et al.* 2010). However, a comprehensive study of the uncertainty in GISTEMP-estimated global mean temperature was not conducted until Lenssen *et al.* (2019), which is presented as Chapter 1 of this dissertation.

The 95% confidence intervals used to quantify mean annual global temperature uncertainty in Lenssen *et al.* (2019) are very useful for qualitatively understanding the uncertainty in the global mean temperature. However, these confidence intervals have limited use as this uncertainty contains temporal structure not well modeled by traditional statistical time series methods such as autoregressive processes (Menne *et al.* 2009; Kennedy 2014). The temporal structure primarily arises from persistent biases in the data arising from inhomogeneities in the land (Menne *et al.* 2009; Menne *et al.* 2018) and sea surface (Kennedy *et al.* 2011a; Kennedy *et al.* 2011b; Kennedy

2014; Huang *et al.* 2017) records due to changing instrumentation, collection methods, and/or station location There has been substantial progress in correcting these biases in the most recent station record analyses from the NOAA Global (GHCN) (Menne *et al.* 2018) and sea surface temperature analyses from NOAA (Huang *et al.* 2020) and the Hadley Center (Morice *et al.* 2021). However, the correction of these biases is done statistically which introduces temporally persistent biases in global and regional temperature series that cannot be sufficiently represented by time series models.

Thus, the state-of-the-art method for quantifying and communicating uncertainty in historical global and regional temperature is through an uncertainty ensemble where each member of the ensemble contains an equally likely record of the Earth's temperature. The benefits to this approach are twofold. First, persistent biases are well represented with potential corrections to station and ship records reflected in different ensemble members. Second, it is nearly trivial to apply these uncertainty estimates in subsequent analyses by repeating any analysis with each ensemble member. This approach has become the standard in the latest HadCRUT4 (Morice *et al.* 2012) and HadCRUT5 datasets (Morice *et al.* 2021) as well as the NOAA GlobalTemp analysis (Huang *et al.* 2020). In Chapter 2 of this dissertation, a GISTEMP uncertainty ensemble is presented and analyzed, providing the most comprehensive study of uncertainty in the GISTEMP estimate of global and regional mean temperature to date.

Direct measurements of the Earth are critical to advancing climate science. The historical surface temperature record plays a central role in understanding the change and variability of the climate system. Quantifying the uncertainty of this record provides necessary information to researchers seeking to understand how our climate has changed and may change in the future. The communication and distribution of the temperature record uncertainty through an ensemble lowers the barrier to properly implementing observational uncertainty in subsequent studies.

0.2 Climate Variability and Prediction

The mean state of the climate system varies naturally on timescales from hours to millions of years. Many of this variability is due to chaotic, unpredictable behaviors in the Earth system, but some physical processes allow variations in mean climate to be predicted months, years, and decades in advance. As the initial condition predictability of the atmosphere is limited to around two weeks (Lorenz 1969b), predictability further in the future is possible through the prediction of slowly evolving processes in the climate system. In this section, two related timescales will be discussed: seasonal-to-interannual predictability or the prediction of changes in the climate that occur over a couple months to about a year and decadal predictability or the prediction of changes in the climate that occur over a period of a year to tens of years.

Before diving into the specifics of each of these timescales, it is important to outline a common approach to climate prediction that has inspired many of the questions pursued in this dissertation (Goddard 2012). First, a climate variable of interest such as mean temperature or total precipitation is identified for a specific region and time scale, often due to the societal importance of this climate variable. Then, statistical and dynamical experiments are used to link variations in the climate variable to some combination of physical climate processes, often referred to as modes of variability. The remote influence of a mode of climate variability on regional climate is commonly referred to as a teleconnection. Finally, the potential predictability of each of these modes of variability is determined. If a mode of variability is both found to influence a regional climate as well as be reliably predicted, a climate prediction has the potential to be skillful and ultimately useful.

Such an approach is evident in seasonal climate prediction where the high predictability of ENSO and the global impact of ENSO on climate and weather enable skillful seasonal forecasts over much of the Earth. More specifically, seasonal variability in temperature and precipitation are linked to variations in ENSO (Ropelewski & Halpert 1987; Mason & Goddard 2001; Lenssen *et al.* 2020), whose evolution can be predicted at lead times of months (Zebiak & Cane 1987; Zebiak 1989; Barnston *et al.* 2019) to years (Gonzalez & Goddard 2016; DiNezio *et al.* 2017). This

approach of identifying predictable modes of variability that impact climate and weather in regions of interest has been adopted successfully in subseasonal-to-seasonal forecasting by leveraging the predictability and teleconnections of the Madden-Julian Oscillation (MJO) (Vitart & Robertson 2018; Meehl *et al.* 2021) as well as in seasonal-to-decadal forecasting where predictability is primarily linked to Atlantic Multidecadal Variability (AMV) (Smith *et al.* 2019; Meehl *et al.* 2021).

0.2.1 Seasonal Prediction and ENSO

The philosophy of climate prediction presented above was developed as the community was focused on predicting seasonal-to-interannual variability (hereafter seasonal variability). The dominant mode of seasonal variability in temperature and precipitation globally is associated with often predictable variations in ENSO (Goddard *et al.* 2001). The ENSO phenomenon is an atmospheric-oceanic coupled dynamical process occurring in the equatorial Pacific where coupled feedbacks drive changes in zonal winds, the distribution of oceanic heat content, sea surface temperature, and the location of deep convection. The heating due to the intense convection in the tropical Pacific drives planetary waves (Gill 1980) causing teleconnections or changes in climate and weather worldwide (Mason & Goddard 2001; Lenssen *et al.* 2020).

ENSO has two active phases that are departures from climatology, which includes easterly trade winds, an equatorial gradient in sea surface temperature with warm surface waters in the west and cool surface waters in the east, and a tilted thermocline which is deep in the western Pacific and shallow in the eastern Pacific. Warm El Niño events exhibit a relaxing of the trade winds, a shift of the western warm pool to the central Pacific, and a corresponding flattening of the thermocline. Importantly, El Niño events result in the release of an enormous amount of energy stored in the western Pacific warm pool into the atmosphere (Goddard & Philander 2000). Cool La Niña events are approximately, but not exactly, dynamically opposite with a strengthening of the trade winds, a compacting of the western warm pool westward, and a steeping of the thermocline. The ENSO system oscillates between the phases with El Niño events occurring approximately every 2-7 years. A recent review (Timmermann *et al.* 2018) and monograph (McPhaden *et al.*

2020) provide a detailed description of the current state of ENSO understanding past the scope of this introduction.

The oceanic component of ENSO has been known by the people of Peru for centuries. They referred to the warmer ocean and change in currents that brought excessive rainfall around Christmas as El Niño. The first climate forecasting study following the philosophy outlined above was conducted by Sir Gilbert Walker in the 1920s. Walker linked boreal winter changes in tropical pressure and related zonal winds to drought in India during the following monsoon (Walker 1924). This variability in pressure and wind was named the Southern Oscillation. Almost fifty years later, it was shown that the Southern Oscillation was the atmospheric manifestation of ENSO and was coupled with the El Niño sea surface temperature variability observed in Peru (Bjerknes 1969). The theory of Bjerknes (1969) described the necessary positive feedbacks for the growth of El Niño and La Niña events, but more work over the ensuing decades was needed to determine that transport and dispersion of subsurface heat and mass content was necessary in describing the eventual decay of El Niño events (Wyrtki 1975).

With the core theory in place, the first simple dynamical model of ENSO followed in the 1980s, leveraging the positive and negative feedbacks to capture the oscillatory behavior (Zebiak & Cane 1987). Critically this model could be initialized with the current state of the ocean and atmosphere to make skillful predictions of the future evolution of ENSO (Cane *et al.* 1986), opening the door for initialized seasonal prediction of the climate for months into the future. The extreme El Niño event and resulting widespread impacts in 1997-1998 highlighted the need for and potential utility of climate prediction (Changnon 2000). Over the following decades, the usefulness of ENSO for explaining and predicting seasonal climate variability months in advance was explored in detail (Goddard *et al.* 2001), with operational climate prediction centers formed at metrological agencies around the world.

Understanding the variability and predictability of ENSO is of critical importance to climate science, but study of the phenomenon is difficult due to the lack of long data records and the difficulty of properly simulating the process in state-of-the-art GCMs. There are limited observations

of the sea surface temperature in the region prior to the 1950s and critical subsurface observations only began in the mid-1990s with the Tropical Ocean-Global Atmosphere (TOGA) Tropical Atmosphere Ocean (TAO) Array (Hayes *et al.* 1991). Additionally, GCMs have difficulty in representing the tropical Pacific climatology and variability (Bellenger *et al.* 2014). Thus, it is of great importance to continue collecting and expanding observational data sources of ENSO as well as continue development of statistical and dynamical models for the study of ENSO and seasonal prediction.

The work presented in Chapter 3 of this dissertation presents new methods for working with the limited global precipitation and sea surface temperature records to link ENSO and seasonal precipitation variability globally. In addition, this chapter shows the skill of precipitation forecasts made with these empirical relationships, motivating improved and longer-lead ENSO forecasts (Lenssen *et al.* 2020). Chapter 4 takes on this problem of extending the prediction of ENSO past the currently operational nine month lead times and proposes methods that reduce the degradation of forecast skill due to GCM biases. Together, these chapters advance the field's understanding of seasonal variability and predictability while addressing both of the limitations in observational data and in dynamical models discussed in this introduction.

0.2.2 Decadal Prediction

The prediction of the climate in the coming 1-10 years, or decadal prediction, is a relatively new field with initial studies occurring in the late 2000s (Smith *et al.* 2007; Keenlyside *et al.* 2008; Meehl *et al.* 2009) and continuing over the following decade (Cassou *et al.* 2018; Kushnir *et al.* 2019; Meehl *et al.* 2021). Like seasonal prediction, some of the potential predictability in the climate at decadal timescales comes from the initial state of the climate system. However, the longer timescales in decadal prediction necessitate incoporating predicted changes in climate due to natural and anthropogenic radiative forcings. Thus, decadal prediction falls into a middle ground between traditional initialized climate prediction where the predictability is entirely derived from the initial state, and uninitialized climate projections where the predictability is entirely due to the

radiative forcing (Meehl *et al.* 2009; Goddard 2012).

There are two major decadal modes of natural variability in the climate system: the Pacific Decadal Oscillation (PDO) (Mantua *et al.* 1997; Newman *et al.* 2016) and AMV (Kushnir 1994; Zhang *et al.* 2019). The PDO has not been shown to be reliably predictable, but AMV appears to be a potential driver of decadal predictability worldwide, with changes in north Atlantic sea surface temperatures associated with AMV having been shown to drive temperature and precipitation in regions worldwide (Ting *et al.* 2011; Smith *et al.* 2019) including predictability of the North Atlantic Oscillation (NAO) over the next decade (Smith *et al.* 2019; Smith *et al.* 2020). However, climate models underestimate the relative predictable signal of these key decadal processes by an order of magnitude, requiring extremely large sample sizes and therefore computational costs, to achieve skillful forecasts (Scaife & Smith 2018; Smith *et al.* 2020).

A key problem in decadal prediction is determining what portion of the predictable signal is due to the initial conditions and therefore the evolution of predictable climate variability. The remaining predictable signal is then due to the external forcing (Goddard *et al.* 2013). Decomposing the decadal climate variability into these two sources is critical for directing future research in decadal prediction as well as the design and implementation of forecast systems (Hawkins *et al.* 2011; Suckling *et al.* 2017). If external forcing drives all of the decadal variability in a region, costly initialized predictions in that region will be unnecessary. Likewise, if a region is shown to have initialization-driven predictability in empirical and model-based experiments, more research should be devoted to determine the dynamics of this predictability as well as any relevant teleconnections. Chapter 5 of this dissertation preforms such a study, decomposing the observed climate variability into terms associated with changes in natural and anthropogenic radiative forcings and potentially predictable modes of natural climate variability.

Chapter 1: Uncertainty in Observed Global Annual Mean Temperature

This first-author work is published as Lenssen et al. *(2019) in the Journal of Geophysical Research: Atmospheres.*

Attempts to seriously estimate the changes in temperature at the hemispheric and global scale date back at least to (Callendar 1938) who used 147 land-based weather stations to track near global trends from 1880 to 1935 (Hawkins & Jones 2013). Subsequent efforts used substantially more data (180 stations in (Mitchell Jr. 1961), 400 stations in (Callendar 1961), "several hundred" in (Hansen *et al.* 1981) etc.), and with a greater global reach. While efforts were made to estimate the uncertainty associated with these products, efforts were more suggestive than comprehensive.

As the data sets have grown in recent years (through digitization and synthesis of previously separate data streams) (Rennie *et al.* 2014; Freeman *et al.* 2016; Thorne *et al.* 2018), and efforts have been made to improve data homogenization, bias corrections and interpolation schemes, the sophistication of the uncertainty models has also grown. Notably, with the introduction of the Hadley Centre SST analysis HadSST3 (Kennedy *et al.* 2011a; Kennedy *et al.* 2011b), Berkeley Earth (Rohde *et al.* 2013a), and the joint Hadley Centre and University of East Anglia's Climatic Research Unit HadCRUT4 (Morice *et al.* 2012), Monte Carlo methodologies have been applied to generate observational ensembles that quantify uncertainties more comprehensively than was previously possible.

GISTEMP is a widely-used data product that tracks global climate change over the instrumental era. However, the existing uncertainty analysis currently contains only rough estimates of uncertainty on the land surface air temperature (LSAT) mean and no estimates of the sea surface temperature (SST) or total (land and sea surface combined) global mean. This chapter describes a novel end-to-end assessment of all the known uncertainties associated with the current GISTEMP

analysis (nominally based in the methodology described in (Hansen *et al.* 2010), but with changes to data sources as documented on the GISTEMP website and outlined below), denoted version 4. The study uses independently derived uncertainty models for the land station homogenization (Menne *et al.* 2010; Menne *et al.* 2018) and ocean temperature products (Huang *et al.* 2015b; Huang *et al.* 2017), combined with an assessment of spatial interpolation and coverage uncertainties, as well as parametric uncertainty in the GISTEMP methodology itself.

The analysis was performed in the open source language R (R Core Team 2020) and the data, code, and intermediate steps needed to generate all figures in this report are available on the GISTEMP website (`https://data.giss.nasa.gov/gistemp/uncertainty`).

1.1 Overview of Surface Temperature Products

All of the most commonly cited surface temperature analyses split up the calculation of global anomaly fields into separate LSAT and SST anomaly analyses. These independent LSAT and SST analyses are combined into a total (LSAT and SST) global surface temperature index from which spatially averaged global and regional time series can be computed (note this is not strictly equal to the true surface air temperature anomaly (Cowtan *et al.* 2015)). Likewise, the uncertainty analyses for the LSAT and SST are performed separately, then combined into total global uncertainty.

Semi-operational surface temperature analyses have been available since the first products by NASA/GISS and joint work from the Hadley Centre and Climatic Research Unit in the UK in the late 1970s. There are now multiple updated and peer-reviewed surface temperature products available, notably produced by NASA/GISS (GISTEMP), NOAA National Centers for Environmental Information (NCEI) with the Merged Land–Ocean Surface Temperature Analysis (MLOST), the Hadley Centre/Climatic Research Unit (HadCRUT), an analysis from the Japanese Meteorological Agency (JMA) (Ishihara 2006) and a reanalysis-based product from ECMWF. These analyses use considerably different methods for the calculation of historical global and regional mean time series, but broadly agree on the trends and interannual variations in the global annual mean time series (fig. 1.1), though they differ at more regional scales as a function of data coverage and

interpolation method (Rao *et al.* 2018). However, interpreting the comparisons across surface temperature products has to be nuanced since the raw data and intermediate product sources are often shared and not completely independent. Of the six major products that are currently being updated in real-time, GISTEMP was notable in not having rigorous confidence intervals on the global and regional mean time series.

The treatment of missing land surface data is a major distinction between products. Since monthly temperature anomalies are strongly correlated in space, spatial interpolation methods can be used to infill sections of missing data. However, smoothing due to interpolation obscures spatial variability as grid box estimates are some weighted combination of many stations. HadCRUT4 performs the least interpolation. If a $5^\circ \times 5^\circ$ grid box does not have any station data, this grid box is reported as missing (Morice *et al.* 2012). The HadCRUT method has the major advantage of clarity in that every grid box is the simple average of the station anomaly values contained in the grid box, but suffers in coverage, particularly in the critical Arctic region. At the other extreme, GISTEMP performs the most interpolation by giving stations a 1,200 km radius of influence, regardless of latitude (Hansen *et al.* 2010). The interpolation allows for infilling during the data-poor early years (pre-1960), but makes it more complex to determine how stations contribute to grid box values. The GISTEMP method is explained in detail in the following section. The NOAA method performs an intermediate amount of interpolation by aggregating a $5^\circ \times 5^\circ$ grid up to a $15^\circ \times 15^\circ$ grid before modeling the fine-scale variability using an empirical orthogonal function teleconnection analysis as described in Appendix A of (Smith & Reynolds 2005). The JMA method is similar to that of HadCRUT4. Comparisons to reanalyses products suggest that the interpolated products have less overall bias compared to the true global mean (Simmons *et al.* 2016) because the missing data areas are predicted (and seen) to be changing more than the global mean.

Recently, the Berkeley Earth group (Rohde *et al.* 2013a) and (Cowtan & Way 2014) have released more statistically sophisticated products that confirm the observed warming in the NASA, NOAA, and HadCRUT products and provide a more natural uncertainty quantification. Berkeley Earth used an additive Kriging model for the LSAT analysis to estimate interpolated LSAT fields

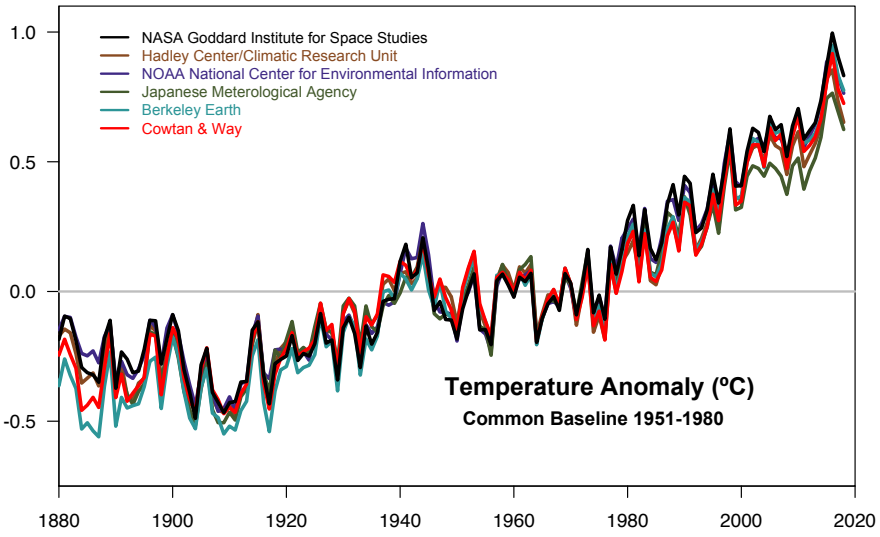

Figure 1.1: Comparison of six analyses of the annual global surface temperature anomaly through 2018.

rigorously. Cowtan and Way took this approach a step further and used methods to interpolate both SST and LSAT fields used in HadCRUT. The results of Cowtan and Way suggest that the inclusion of interpolation is necessary to capture the global effect of the higher rate of warming in the Arctic.

1.2 Operational GISTEMP

The current operational method used in GISTEMP to compute the mean land surface temperature anomaly is an extended version of the process outlined by (Hansen & Lebedeff 1987). The analysis contains two major steps: interpolation of individual station data and averaging of interpolated fields. Preliminary to the two core steps, the monthly station data are processed following (Hansen *et al.* 2010). The publicly available code, written in Python, has been updated to modern standards (Barnes & Jones 2011).

GISTEMP uses the equal-area grid developed in (Hansen & Lebedeff 1987). The Earth is divided into 80 equal-area boxes arranged in bands of constant latitude. By constraining each box to cover the same area, the bands have unequal numbers of grid-boxes resulting in an irregular grid. There are four bands in each hemisphere representing the polar region, mid-latitudes, subtropics, and tropics which respectively contain 4, 8, 12, and 16 equal area boxes. Therefore the bands account for 10, 20, 30, and 40 percent of the area of the hemisphere. Each of the 80 boxes are divided into 100 equal-area sub-boxes resulting in an equal-area grid of 8000 grid-boxes covering the Earth.

1.2.1 Interpolation Step

Calculating the values of the sub-boxes in the equal area grid from the station anomaly record is referred to as the interpolation step in this study. For a single sub-box, all stations within a given distance are successively combined starting with the longest record. A new station is averaged in if there is at least a 20-year overlap, and an offset is applied to leave the mean over that common period unchanged individually for each calendar month. The weight W for a station d km away from the sub-box center within a given radius r is determined using a linear radial basis function of the form

$$W_r(d) = \max\left(\frac{r-d}{r}, 0\right) \tag{1.1}$$

The value of the radius, r = 1,200 km, was estimated based on an investigation of the correlation of the annual mean series of pairs of stations as a function of their spatial separation (Hansen & Lebedeff 1987); this simple device turned out to be quite similar to the form of the estimated covariance function in the modified Kriging method used by the Berkeley Earth analysis (Rohde *et al.* 2013b). If there are no stations within 1200km of a sub-box center, it is given a missing value.

1.2.2 Averaging Step

The averaging step calculates the regional and global time series from the interpolated sub-box records. In this context, regional refers to hemispheric and the 8 latitudinal bands in the equal area grid. First, an average series is computed for each of the 80 equal area boxes by the method described in the interpolation step section, except that equal weight is given to each equal area sub-box series. The LSAT and SST data are combined when each of the 80 box series are created. In each sub-box, either a pure SST series or a pure LSAT series is selected. SST data are used only for ocean sub-boxes that contain no sea ice and whose center is more than 100 km off the nearest land station. Everywhere else the LSAT data is used.

The averages for the eight latitudinal zonal bands are then computed from the box series weighted by the number of sub-boxes with data. The three extra-tropical bands in each hemisphere are combined in the same way into a single series. These two series and the two tropical series are converted to anomaly series with respect to the 1951–1980 period. Global and hemispheric anomalies are computed as weighted averages of these four band means, weighted by the full area of these bands.

1.2.3 Changes to Operational GISTEMP 2010–2018

The only difference in methodology since (Hansen *et al.* 2010) not caused by changes in the available input data, was combining into single polar boxes the 40 sub-boxes reaching the North and the South poles (starting September 2016). This only insignificantly affected the results, but produced more natural looking images near the poles.

All other changes relate solely to the input data. In 2010, GISTEMP was using GHCN-Monthly version 2 (GHCNv2), the U.S. Historical Climatology Network version 2.0 (USHCN2) and the Scientific Committee on Antarctic Research (SCAR) temperature data over land, with Hadley Centre Sea Ice and Sea Surface Temperature data set (HadISST) and Optimum Interpolation Sea Surface Temperature (OISSTv3) for the ocean. With the upgrade to GHCNv3 in December 2011 (and then to v3.2 in September 2012, and now to v4), the need for USHCN2 was obviated. In GHCNv3 as

in GHCNv4, the various data series from different sources for a location, that were available in GHCNv2, are merged into a single series, and the resulting inhomogeneities are resolved in the adjustment procedure. Hence GISTEMP is using the adjusted GHCNv3 and v4 data. Whereas combining different sources at a location and manual corrections are no longer needed, the GISS urban adjustment scheme is still being applied. For the ocean data, the ocean temperature product was replaced with the more homogeneous Extended Reconstructed Sea Surface Temperature (ERSST) v3b in January 2013, which was updated to ERSSTv4 in July 2015, and to ERSSTv5 in August 2017. The impacts over time of these changes are recorded and maintained on the GISTEMP History page `https://data.giss.nasa.gov/gistemp/history`.

Analyses subsequent to (Hansen *et al.* 2010) that use GHCNv3 are now being denoted GISTEMP v3. The integration of GHCNv4 into the GISTEMP code in January 2019 is denoted as GISTEMP v4.0; this version does not use the SCAR data except as far as they are part of GHCNv4. Going forward, a more rigorous version numbering scheme will be adopted to better track methodological and input data variations. GISTEMP v3 will nonetheless be maintained for the time being for legacy purposes. The uncertainty analysis presented here is strictly valid for GISTEMP v4.0, but the differences with it applied to v3 are insignificant and primarily arise from differences in GHCN homogenization.

1.2.4 Prior Uncertainty Estimates

GISTEMP has previously presented uncertainties due to incomplete spatial coverage of the station record (Hansen & Lebedeff 1987). Most recently, (Hansen *et al.* 2010) reported estimates of this uncertainty for three large time periods: 1880–1900, 1900–1950, and 1960–2008. The analysis sub-sampled a long run of the GISS-ER climate model (Hansen *et al.* 2007) according to the coverage of the station network on the Earth during these three time periods. This model had a $4^\circ \times 5^\circ$ latitude by longitude grid. Global annual land-only means of the sub-sampled model were compared with global annual land-only means using all of the grid-boxes.

Since the global mean calculation in GISTEMP aggregates from small sub-boxes to the 80

equal-area boxes, the coarse model grid approach has considerable value in quantifying the large-scale sampling uncertainty in the approach assuming that the model is capturing sufficient statistical structure of the underlying fine-scale global temperature anomaly field. The uncertainty calculation also roughly captures large-scale spatial and temporal sparsity. An equal-area box that has no data within 1200km is "missing" in the GISTEMP global and regional mean calculation and is on the approximate scale of the model grid. Furthermore, the large grid-box size of that model serves as a rough approximation of the interpolation step of the GISTEMP procedure.

This chapter addresses a number of deficiencies in the legacy GISTEMP LSAT sampling uncertainty analysis. The first goal is increasing the temporal resolution of uncertainty from around 50 years to decadal estimates of LSAT uncertainty. Further refinements to the annual or even monthly timescale do not make a substantive difference. Second, uncertainty in the interpolation step of GISTEMP is better captured. The coarse resolution of the previously used model grid does not describe the fine scale behavior of the true temperature anomaly field and does not allow us to replicate the interpolation step. Using a reanalysis product with a much finer grid to replicate the entire GISTEMP global and regional mean calculation, as is detailed in the following section. Thus, the resulting uncertainties will better reflect the actual analysis method used. Finally, the global mean uncertainty of the GISTEMP band averaging scheme and a simple latitude-weighted mean are compared.

The previously reported GISTEMP uncertainties do not include parametric uncertainties due to homogenization of the station record or uncertainties associated with the SST reconstruction. A holistic estimate of the full uncertainty in the GISTEMP product is made by adding in the homogenization uncertainty from the GHCN dataset to the LSAT analysis as well as propagating the total SST uncertainty from the ERSSTv5 dataset through the full GISTEMP procedure.

1.3 Sources of Uncertainty

1.3.1 Statistical Formulation of Uncertainty

Before outlining the sources of uncertainty in the land and ocean reconstructions, it is useful to step back and discuss the underlying statistics in general terms. Letting $\mu(t)$ be the true (latent) global anomaly for a year t, the calculated (observed) annual mean temperature anomaly $A(t)$ can be decompsed as

$$A(t) = \mu(t) + \epsilon(t). \tag{1.2}$$

In this formulation, $\epsilon(t)$ is a random variable that represents the total uncertainty in the estimate of the annual mean temperature anomaly. Assuming that the estimation procedure is unbiased (an assumption that will be revisited in the discussion of the results), the expected value $\mathbb{E}[\epsilon(t)] = 0$ for all years t. The uncertainty in the calculation of the global mean is then defined as

$$\mathcal{E}(t) = \mathrm{Var}(\epsilon(t)). \tag{1.3}$$

This study breaks down the uncertainty into two components: the uncertainty in the global mean due to uncertainties in the land calculation $\epsilon_L(t)$ and uncertainty in the global mean due to uncertainties in the sea surface calculation $\epsilon_S(t)$. The total uncertainty is decomposed as

$$\epsilon(t) = \epsilon_L(t) + \epsilon_S(t). \tag{1.4}$$

If these uncertainties are independent, the calculation of the uncertainty is the sum of the individual variances

$$\mathcal{E}(t) = \mathrm{Var}(\epsilon(t)) = \mathrm{Var}(\epsilon_L(t)) + \mathrm{Var}(\epsilon_S(t)) \tag{1.5}$$

This study makes the assumption that the land and ocean uncertainties are independent. However, there is potentially correlation between the uncertainty due to the land calculation and the uncertainty due to the ocean calculation. In addition to correlation between the land and ocean

uncertainties, there is also some amount of correlation of the uncertainties in time, particularly at the monthly time scale. Not accounting for this positive correlation of uncertainties in time will lead to underestimation of the uncertainty. To reduce the impact of this autocorrelation, this study focuses on the annual mean temperature anomalies which exhibit much lower autocorrelation.

1.3.2 Land Surface Temperature Uncertainty

Quantifying the uncertainties that arise from using the land station record to calculate regional and global land-only mean temperatures has been an active field for many years. In particular, NOAA (Vose *et al.* 2012) and HadCRUT (Morice *et al.* 2012) groups have developed sophisticated uncertainty models for this portion of the analysis. It is generally assumed that there are three major independent sources of uncertainty in the land record that add uncertainty to global temperature calculations: station uncertainty, bias uncertainty, and sampling uncertainty. These three sources are summarized below (though see Brohan *et al.* (2006) for a detailed discussion). As with the operational GISTEMP, the land surface is any grid-box that is classified as either land or sea ice.

Station Uncertainty

Station uncertainty encompasses the systematic and random uncertainties that occur in the record of a single station and include measurement uncertainties, transcription errors, and uncertainties introduced by station record adjustments and missed adjustments in post-processing. The random uncertainties can be significant for a single station, but comprise a very small amount of the global LSAT uncertainty to the extent that they are independent and randomly distributed. Their impact is reduced when looking at the average of thousands of stations.

The major source of station uncertainty is due to systematic, artificial changes in the mean of station time series due to changes in observational methodologies. These station records need to be homogenized, or corrected to better reflect the evolution of temperature. The homogenization process is a difficult, but necessary statistical problem that corrects for important issues albeit with significant uncertainty for both global and local temperature estimates.

Bias Uncertainty

Bias uncertainty refers to the biases in a single station record due to non-climatic sources. Thermometer exposure change bias (Parker 1994) refers to biases introduced to the station record by the evolution of temperature measurement techniques, such as the switch to Stevenson screens in the 19th Century or the change to Max-Min Temperature Sensor (MMTS) automated recorders in recent decades in the United States (Menne *et al.* 2009). Urban biases are not due to systematic biases in the instrumentation, but rather due the local warming effect of urban centers through land surface changes, reductions in evapotranspiration, and local heat sources. These urban biases are corrected for in global temperature studies, since the goal is to understand the changes in the global climate system, not the localized effect of urban heat islands. An urban bias correction was added to GISTEMP in 1998 (Hansen *et al.* 1999); it confirmed that its impact on global temperature anomalies is small. As shown in (Hansen *et al.* 2010), the effect of the urban adjustment on global temperature change is on the order of 0.01°C.

Sampling Uncertainty

Sampling uncertainty is an umbrella term for uncertainties introduced into global and regional annual means by incomplete spatial and temporal coverage. Whereas the station uncertainties are observed to mostly cancel out in modern-era global annual means, as many of the uncertainties are independent from station to station, the sampling uncertainties remain significant. Understanding the sampling uncertainty of GISTEMP is crucial because, unlike HadCRUT, GISTEMP extrapolates out the anomaly field into regions without station data. Quantifying the sampling uncertainty will provide a measure of confidence in the extrapolation. Since reduction in bias in the global mean due to interpolation comes with an uncertainty variance increase, it is important that interpolation does not drastically inflate the sampling uncertainty.

Quantifying the sampling uncertainty is critical to providing uncertainties for the mean temperatures for two reasons. First, the HadCRUT analysis has shown that the sampling uncertainty is a significant component of the uncertainty in the global annual means in the modern instrumental

era (Morice *et al.* 2012). Second, updating the sampling uncertainty model provides transparent continuity in the GISTEMP analysis for numerous researchers that rely on the data product for their own analyses. As detailed in the following section, GISTEMP has historically made only rough estimates of the sampling uncertainty. The update provided here provides a transition from the original GISTEMP uncertainty model towards a more modern statistical approach.

1.3.3 Sea Surface Temperature Uncertainty

The current production versions of GISTEMP use the ERSSTv5 product provided by NOAA/NCEI (Huang *et al.* 2017) for ocean temperatures. ERSSTv5 uses the same underlying method (Huang *et al.* 2015a) and uncertainty quantification method (Liu *et al.* 2015a; Huang *et al.* 2016a) as the previous generation ERSSTv4. The major upgrade in v5 is a more sophisticated parameter tuning, resulting in more realistic spatiotemporal patterns in the reconstructed SST fields. In addition, v5 incorporates new data sources from the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) 3.0 (Freeman *et al.* 2016) and the Argo float network of near-surface readings.

The uncertainty calculation in ERSSTv4/v5 breaks down the ocean uncertainty into two independent components: parametric uncertainty and reconstruction uncertainty (Liu *et al.* 2015a; Huang *et al.* 2016a). Parametric uncertainty quantifies the internal statistical variability of the ERSST procedure and is defined by the standard deviation of a perturbed parameter ensemble. The ensemble has been constructed such that the parametric uncertainty contains both the bias and sampling uncertainty (Huang *et al.* 2016a). Reconstruction uncertainty represents the information lost in using a finite number of empirical orthogonal teleconnection (EOT) functions to model the high-frequency component. Reconstruction uncertainty can be large at small spatial scales, but averages out to nearly zero at global scales as seen in Figure 2c of (Huang *et al.* 2016a). As this study is focused on global and hemispheric mean uncertainty, it is reasonable to ignore the reconstruction uncertainty and focus only on the parametric uncertainty.

1.4 Update to GISTEMP's Uncertainty Analysis: Methods

1.4.1 Updated Land Surface Temperature Uncertainty Methodology

Data Sources

GHCN: The primary data source for LSAT data in GISTEMP v4.0 is the GHCN product from NOAA/NCEI. As mentioned in the discussion of the updates to operational GISTEMP in Section 1.2.3, GHCNv4 replaced the combined GHCNv3 and SCAR as of January 2019. Thus, the LSAT uncertainty analysis is conducted primarily using GHCNv4, but will also evaluate and discuss how the results apply to GISTEMP v3. GHCNv4 contains significantly more stations than GHCNv3/SCAR, though many of the additional time-series are short. In general, the added stations in GHCNv4 do not significantly alter spatial coverage after interpolation and so will not effect the spatial uncertainty significantly, though it does slightly reduce some homogenization uncertainty. Coverage is quantified by the number of grid boxes in the Modern-Era Retrospective Analysis for Research and Applications (MERRA) grid that contained a station with decadal coverage within the 1,200 km interpolation radius of influence (Figure 1.2) and shows nearly no difference between versions. The increased quantity and quality of stations in GHCNv4 will be most useful for more localized analyses.

Reanalyses: Three distinct reanalysis products are used as globally complete 'ground truth' temperature fields to quantify the contribution of the incomplete spatial and temporal coverage of the station record to the uncertainty in the global temperature anomaly. They are the fifth generation European Centre for Medium-range Weather Forecasting (ECMWF) atmospheric reanalysis (ERA5), the JMA JRA-55 analysis (hereafter JRA), and MERRA-2 (hereafter MERRA). Since the legacy approach inherently aggregates spatially due to large grid-box size it cannot utilize GISTEMP's interpolation method for the uncertainty analysis. This study takes a similar methodological approach, using a high-resolution reanalysis product in place of the climate model output. The rough idea is the same: total coverage global means are compared with realistic (reduced) coverage global means and the uncertainty is described by summary statistics. The finer spatial resolution

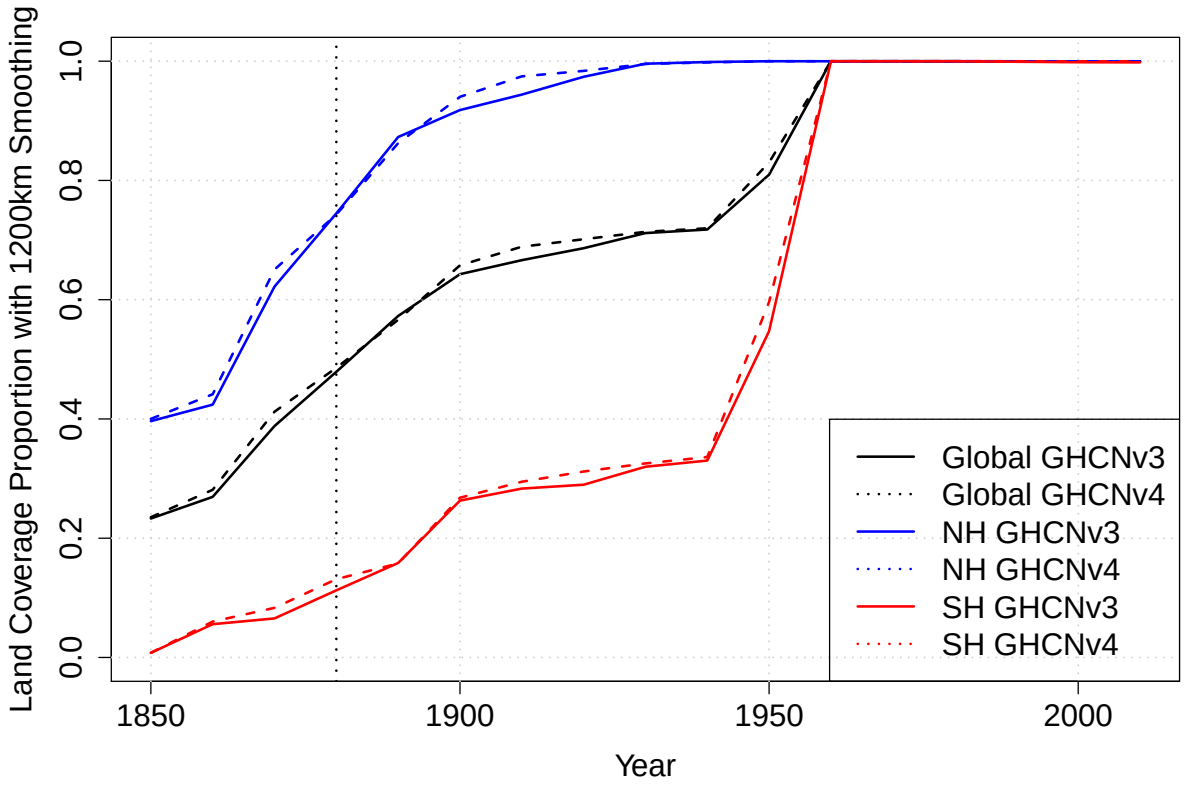

Figure 1.2: A comparison of the decadal land area coverage proportion in GHCNv3 and v4. A location is said to be covered if it is within 1200 km of a station with decadal coverage and will be included in the production GISTEMP analysis.

of the reanalyses than the climate model used previously allows us to treat single grid-box temperature anomaly values as station anomalies. The combination of improved spatial resolution and an analysis more closely mirroring the production GISTEMP procedure will give us more robust calculations of the coverage uncertainty for the global mean.

The primary reanalysis used in the study is ERA5 from 1979–2018 (Copernicus Climate Change Service (C3S) 2017). The 2m monthly average temperature anomalies are averaged to the 0.5° × 0.625° MERRA grid to facilitate comparison and speed up computation. The analysis was also conducted on the native 31 km grid and resulted in no significant changes to the results. ERA5 was used as the primary reanalysis as it best replicates the observed global mean over its record. Furthermore, ERA55 and JRA55 produce generally consistent results while results found using

MERRA often deviate.

MERRA provides monthly temperature means for the entire Earth from 1980–2018 at a 0.5° × 0.625° resolution (Gelaro *et al.* 2017). The addition of the MERRA reanalysis is also due to the observational data sources used in assimilation. Since the goal is to determine the uncertainty that arises from the incomplete coverage of the GHCN station record, it is ideal to use a reanalysis that does not incorporate any GHCN information. Over land, MERRA only assimilates surface temperature data from the surface reading of the radiosonde network, ensuring that the statistical model for GHCN sampling uncertainty is fit with an independent data source (McCarty *et al.* 2016). All results are also verified with the JRA55 reanalysis over 1979–2013 (Kobayashi *et al.* 2015) to provide clarity when the results from MERRA and ERA5 disagree.

LSAT Sampling Uncertainty Method

A grid-box is determined to be land for the purpose of this study if its land area proportion is greater than 0% on the MERRA grid, approximately replicating the 100 km influence of land stations onto ocean grid cells in operational GISTEMP. As in operational GISTEMP, sea ice extent is determined for each month by the maximal extent of sea ice in the MERRA reanalysis. Grid-boxes that are not classified as land are classified as ocean with uncertainty quantified by the SST uncertainty analysis.

Monthly temperature anomalies are computed for the entire reanalysis grid for each of the 12 months by removing the single month mean for each grid-box time series. The full monthly temperature anomaly fields are used to calculate the baseline global and zonal annual means. A modified version of the GISTEMP averaging step is used with the same zonal bands and 80 equal area grid boxes as operational GISTEMP and the reanalysis grid used instead of the sub-boxes. The baseline global mean represents the true global anomaly $\mu(t)$ which will be compared with the means calculated with reduced coverage $A(t)$.

The spatial sub-sampling of the anomaly field is determined at a decadal temporal resolution. A station has temporal coverage in a decade if it has coverage for at least 5 of the 10 years. To have

coverage for a year it must have coverage for at least 3 seasons which requires at least two months in the season. A grid box is said to have coverage in a decade if it contains at least one station with coverage as defined above. Using these definitions, 14 decadal coverage masks are created on the reanalysis grid, one for each decade from the 1880s to the 2010s. That is, a constant mask that contains the coverage of the observing network is generated for each decade.

Reduced coverage global annual means, $A_k(t)$ are calculated for each of the 14 decadal time periods using a modified GISTEMP procedure. In the notation $A_k(t)$, k represents the decade used and t represents the year in the reanalysis record. The interpolation step is performed on the grid using a radius of 1200km. Then the averaging step is performed as described in the baseline global mean calculation with the sub-boxes taken to be the area-weighted gridboxes. Thus, the baseline global mean is an annual time series indexed by t spanning 1980–2017. There are $k = 1, \ldots, 14$ reduced coverage global means for the 14 decades of the study, each annual time series spanning from 1980–2017.

As the sampling uncertainty in ocean regions is quantified as part of the SST uncertainty analysis for ERSSTv5, only land area is included in the LSAT sampling uncertainty calculation. The global and reduced coverage global land means are taken over land and sea ice regions following the GISTEMP procedure. Sea ice regions are defined using MERRA as the maximum extent of ice for each month over the reanalysis record.

The variance of the sampling uncertainty in GISTEMP $\mathcal{E}_L(t)$ is determined as the sample variance of the difference between the baseline global mean and each of the mask means. Rearranging equation 1.2, the difference series $D_k(t)$ is defined for decade k as

$$D_k(t) \equiv \mu(t) - A_k(t) = \epsilon_k(t). \tag{1.6}$$

Then the uncertainty is $\mathrm{Var}(D_k(t))$. Note that this method assumes that the method for calculating the global mean does not have any systematic mean bias.

LSAT Sampling Extensions

The sampling uncertainty analysis allows investigation of other properties of the GISTEMP LSAT method. Here, three experiments addressed in this study are described. First, the assumption that the land surface mean temperature estimate is an unbiased estimate is investigated. Then, the minimum achievable sampling uncertainty from the GISTEMP interpolation is calculated by assuming full global station coverage. Finally, the value of the GISTEMP averaging method is investigated by comparing with a naive averaging scheme.

Sampling Bias: Recent studies have shown the likely presence of bias in surface temperature products compared to the true global mean (Simmons *et al.* 2010; Cowtan & Way 2014; Karl *et al.* 2015; Jones 2016; Simmons *et al.* 2016). In addition, recent evidence from remote sensed temperature analyses suggest production GISTEMP may be underestimating Arctic warming (Susskind *et al.* 2019). To quantify the potential sampling biases due to limited station coverage, potential systematic additive bias α_k and multiplicative bias β_k are introduced. Then, determining the variance of ϵ_k can be formulated as the univariate regression

$$\mu(t) = \alpha_k + \beta_k A_k(t) + \epsilon_k(t) \tag{1.7}$$

Since the analysis is conducted with anomalies that are standardized over the entire time period of ERA5 (1979–2018), the additive bias $\alpha_k = 0$ for all decades as all of the grid-box time series are mean zero. However, the full linear regression is fit as a sanity check as it will have practically no effect on the estimation of β_k or the uncertainty. Since the ERA5 reanalysis currently spans 1979-2018, only the estimates for the 1980s though 2010s are representative of potential bias in operational GISTEMP. The estimates for decades pre-1980 do not reflect the actual bias in GISTEMP during their periods as the underlying climate variability is not properly accounted for. However, the estimates of bias due to limited coverage in early decade are useful for understanding the importance of station coverage for capturing the current pattern of global temperature change.

Limiting Uncertainty: Running the sampling uncertainty analysis in Section 1.4.1 with the as-

sumption that there is station coverage for every land grid box provides a lower bound of the sampling uncertainty. This limiting uncertainty will be greater than zero as the smoothing arising from interpolation increases the uncertainty in the global mean. Calculation of the limiting uncertainty is important to determine the relative values of increased data availability and methodological improvements for lowering the uncertainty of the global mean estimate. In addition to quantifying the lower uncertainty bound, the sampling bias analysis is repeated with the simulated full coverage to determine if the GISTEMP method has any systematic bias in an idealized case over the 1979–2018 period.

Comparison of Averaging Methods: In addition to using the GISTEMP band-average method, the sampling uncertainty analysis in Section 1.4.1 is repeated using a simple latitude-weighted mean. Comparison of the resulting LSAT sampling uncertainties shows the difference between the two averaging methods in accounting for missing data.

GHCN Homogenization Uncertainties

Station uncertainty due to homogenization of station series is quantified in the GHCNv4 analysis and incorporated in the GISTEMP uncertainty analysis with no modification(Menne *et al.* 2018). The GHCNv4 method divides the total homogenization uncertainty for land stations into two independent components: the parametric uncertainty associated with the Pairwise Homogenization Algorithm (PHA; (Menne & Williams 2009)) used to homogenize the GHCNv4 monthly data and incomplete homogenization caused by artificial shifts in the data that remain undetected by the PHA.

The PHA detects artificial time series mean shifts due to changes in observing practice by comparing a station series with neighboring stations (Menne & Williams 2009). Various parameters, such as the minimum number of neighboring stations, are set in implementing the PHA and affect the sensitivity and accuracy of the method. Parametric uncertainty is quantified by running the PHA as an ensemble whose members have randomly varying parameter settings from a set of configurations that produced the best results when run on realistic benchmark datasets (Williams *et al.*

2012). For GHCNv4 monthly, 100 different versions of the PHA were used to homogenize the GHCNv4 data, yielding 100 different homogenized versions of each GHCN station record (Menne *et al.* 2018). The parameter uncertainty is determined by the sample standard deviation of the 100 feasible records.

While the PHA detects large (>0.2°C) breaks in time series, it (and other break-point detection methods) is unable to detect small shifts. This uncertainty associated with incomplete homogenization is estimated by adding small adjustments to the homogenization ensemble members at random dates and with random magnitudes. The frequency and magnitude of the added adjustments were determined by estimating the distribution of the missed (mostly small) breaks from the distribution of actual breaks detected by the PHA. Detected breaks in GHCNv4 have a bimodal distribution with peaks around ±0.5°C. In between these peaks is the so-called "missing middle" of the distribution, which (Menne *et al.* 2018) estimated as having a mean of about −0.01°C and a standard deviation of 0.2 with an average frequency of occurrence of about 1 in 50 years. The number of missed adjustments for each station record in the ensemble was determined by sampling from a Poisson distribution with an average frequency of 1 in 59 years, and their magnitude was selected by a random draw from a normal distribution $\mathcal{N}(-0.01, 0.2)$.

Total Land Surface Temperature Uncertainty Methodology

As introduced in the discussion of sources of land uncertainty in Section 1.3.2, land surface temperature uncertainty arises due to station, bias, and sampling uncertainties. Since the three sources are independent and the bias uncertainty can be ignored for means of large spatial scale, the total uncertainty is simply the sum of the station homogenization and sampling uncertainties. As these uncertainties are expressed as variances, it is critical that the variance for the homogenization and sampling uncertainties are added rather than the standard deviations or confidence intervals.

1.4.2 Sea Surface Temperature Uncertainty Methodology

The uncertainty analysis from ERSSTv4 is used to quantify the uncertainty in the ocean temperature in the GISTEMP analysis as ERSSTv5 did not make any changes to the underlying reconstruction or uncertainty methods (Liu *et al.* 2015b). ERSSTv4 quantified their uncertainty through an ensemble of feasible SST fields rather than a single uncertainty field. The largest ensemble simulation contains 1,000 members and was constructed to quantify the parametric uncertainty in their prediction (Huang *et al.* 2016a). This analysis utilizes this 1,000 member large ensemble to understand how the uncertainty in the ERSST product impacts the GISTEMP uncertainty.

The parametric SST global and hemispheric uncertainty calculation closely follows the analysis performed by the ERSST team (Huang *et al.* 2016a). The GISTEMP averaging step is performed with no land data for each of the 1,000 ensemble members resulting in 1,000 possible global and hemispheric time series. That is, the global mean with an ocean-only mask is calculated for each of the ERSST ensemble members. The 95% confidence interval for the parametric uncertainty of the SST model are then calculated for each time point using the empirical 95% confidence interval of possible global mean sea surface temperature.

The assumption made in this calculation is that the ERSST large ensemble is symmetric about the median for global and hemispheric means and that ERSSTv5 is the median value of the ensemble. Both of these assumptions are not perfect, but reasonable for these large scale means. The mean and median of the global sea surface temperature mean ensemble are nearly identical. Furthermore, the strong agreement between the operational and ensemble global mean (and thus global median from the result presented here) in Figure 12 of Huang *et al.* (2016a) supports the assumption that the global uncertainty is symmetric.

1.4.3 Total Global Uncertainty Methodology

The final step in the global uncertainty analysis is the combination of the separate land and ocean uncertainties into a total global uncertainty. If $\overline{A}(t)$ is the annual global mean anomaly for a year t and given an estimate of the global mean anomaly $\tilde{A}(t)$, the uncertainty of the global annual

mean temperature is

$$\mathcal{E}(t) = \mathrm{Var}\left(\tilde{A}(t)\right) \tag{1.8}$$

The land-only uncertainty is comprised of the sampling uncertainty calculated using the method described in Section 1.4.1 with missing values for all of the ocean grid cells and the homogenization uncertainty according to the GHCNv4 analysis. Likewise, ocean-only uncertainty is calculated using the method described in Section 1.4.2 with missing values for all of the land grid cells. The resulting uncertainties then describe the uncertainty over a subset of the area of the Earth.

To calculate the global uncertainty, the different regions and therefore area-proportions of Earth that the land and ocean cover are calculated. The LSAT uncertainty estimate is defined as $\mathcal{E}_L$ and the SST uncertainty estimate from the ERSST ensemble analysis is $\mathcal{E}_S$. Using a_L and a_S (the area of the land and ocean on the Earth respectively) and assuming that the land and ocean uncertainty components are independent, the total global uncertainty variance is then

$$\mathcal{E}(t) = \left(\frac{a_L}{a_L + a_S}\right)^2 \mathcal{E}_L(t) + \left(\frac{a_S}{a_L + a_S}\right)^2 \mathcal{E}_S(t) \tag{1.9}$$

Hemispheric and other regional combined land and ocean uncertainties are calculated similarly.

Uncertainty values from products that are not operational are assumed constant for time periods after the end of their record. For the SST uncertainty, the ERSST ensemble was only issued through 2014. Thus, the 2014 value is used for years 2015–2018 and will update the analysis as more data become available. Likewise, the GHCNv4 homogenization was conducted through 2016 resulting in the 2017 and 2018 homogenization uncertainties being set to the 2016 value.

1.5 Results

1.5.1 LSAT Uncertainty Results

The sampling and total uncertainty in the global annual land surface mean temperature as calculated by each of the three reanalyses is shown in Figure 1.3a. As expected, increased number

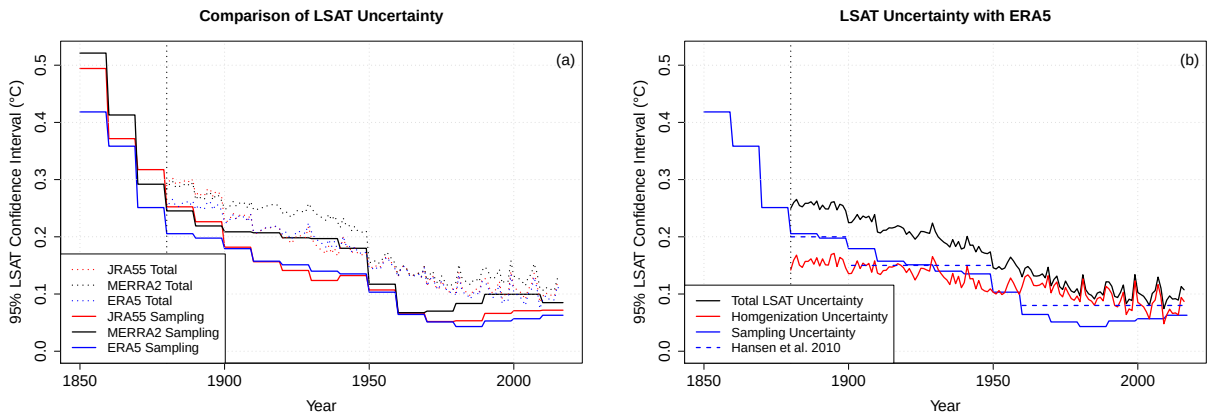

Figure 1.3: The total uncertainty (2σ) in the global annual mean land surface temperature decomposed into the sampling and homogenization uncertainty components where the homogenization uncertainty is found in an independent analysis and is currently limited to 1880 (Menne *et al.* 2018). (a) The sampling and resulting total LSAT uncertainty calculations using the three reanalyses. (b) The LSAT uncertainty as calculated with ERA5, the reanalysis selected for the analysis. The LSAT sampling uncertainty estimate from Hansen *et al.* (2010) is shown for comparison.

of stations and coverage of stations as time progresses results in decreasing sampling uncertainty over time. The three reanalyses are in general agreement with any differences in the sampling uncertainty shrinking in the total LSAT uncertainty. In the early decades of the study period, sampling uncertainty and homogenization uncertainty are of similar magnitude.

Figure 1.3b shows the LSAT as found with the ERA5 sampling uncertainty analysis. The ERA5 analysis is used for all LSAT estimates in the remainder of the study. The homogenization component includes both the parametric uncertainty as well as uncertainties due to missed breaks. Approaching the present, the global sampling uncertainty decreases as the majority of the land has some station coverage, but the global homogenization uncertainty remains high. In particular, the major drop in sampling uncertainty in 1950–1970 occurs due to the inclusion of Antarctica. The relative lack of decrease in uncertainty in the global mean due to homogenization results from correction uncertainties in station records propagating forward in time (Menne *et al.* 2018). The minor contribution of sampling uncertainty to the total modern LSAT uncertainty illustrates how increasing coverage of temperature monitoring will not fix the uncertainty issue in the land surface temperature record.

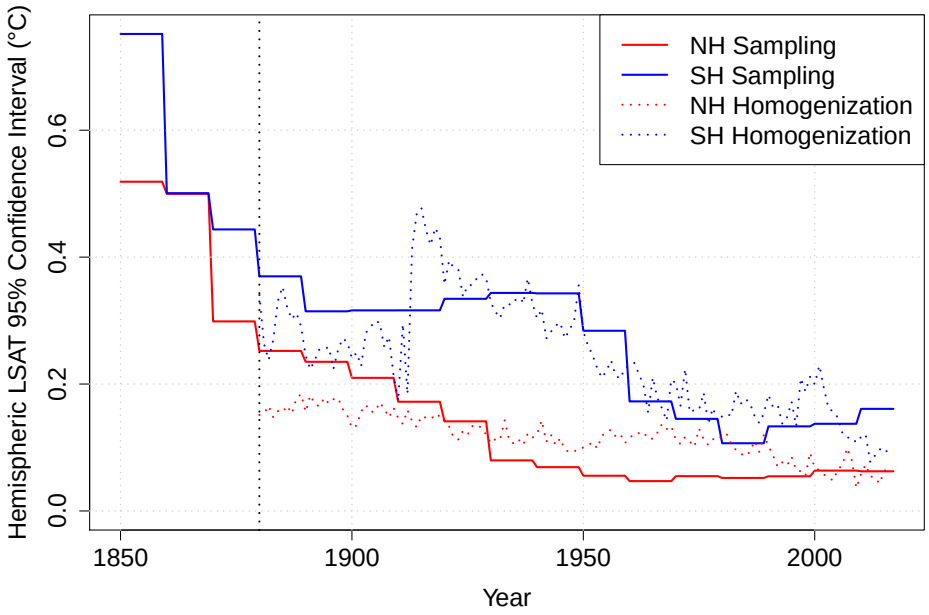

Figure 1.4: Annual land surface temperature anomaly sampling (solid) and homogenization (dotted) uncertainty (2σ) per hemisphere. As expected, the uncertainty in the southern hemisphere is greater in all decades, but reduces greatly to near the northern hemisphere uncertainty post-1960.

The ERA5 analysis shows that the uncertainties in (Hansen *et al.* 2010) were quite good for the early record, but overestimate the sampling uncertainty post-1950. In particular, there is nearly exact agreement over 1880–1900. The sampling uncertainty analysis also suggests that the GISTEMP annual mean time series may be extended to dates earlier than 1880 as is done in HadCRUT4 and Berkeley Earth, but not without suffering a large increase in sampling uncertainty, particularly if including data prior to 1870.

Separating the land uncertainty by hemisphere, it is shown that the southern hemisphere has greater sampling uncertainty post-1920 coinciding with improved northern hemispheric coverage of the Arctic land and sea ice region (Figure 1.4). The effect of Antarctica on the southern hemisphere is shown through the reduction in sampling uncertainty from 1950–1970. The hemispheric homogenization uncertainties are slowly decreasing as is the global mean with the exception of the large jump in southern hemisphere uncertainty in the mid 1920s, which can be explained by

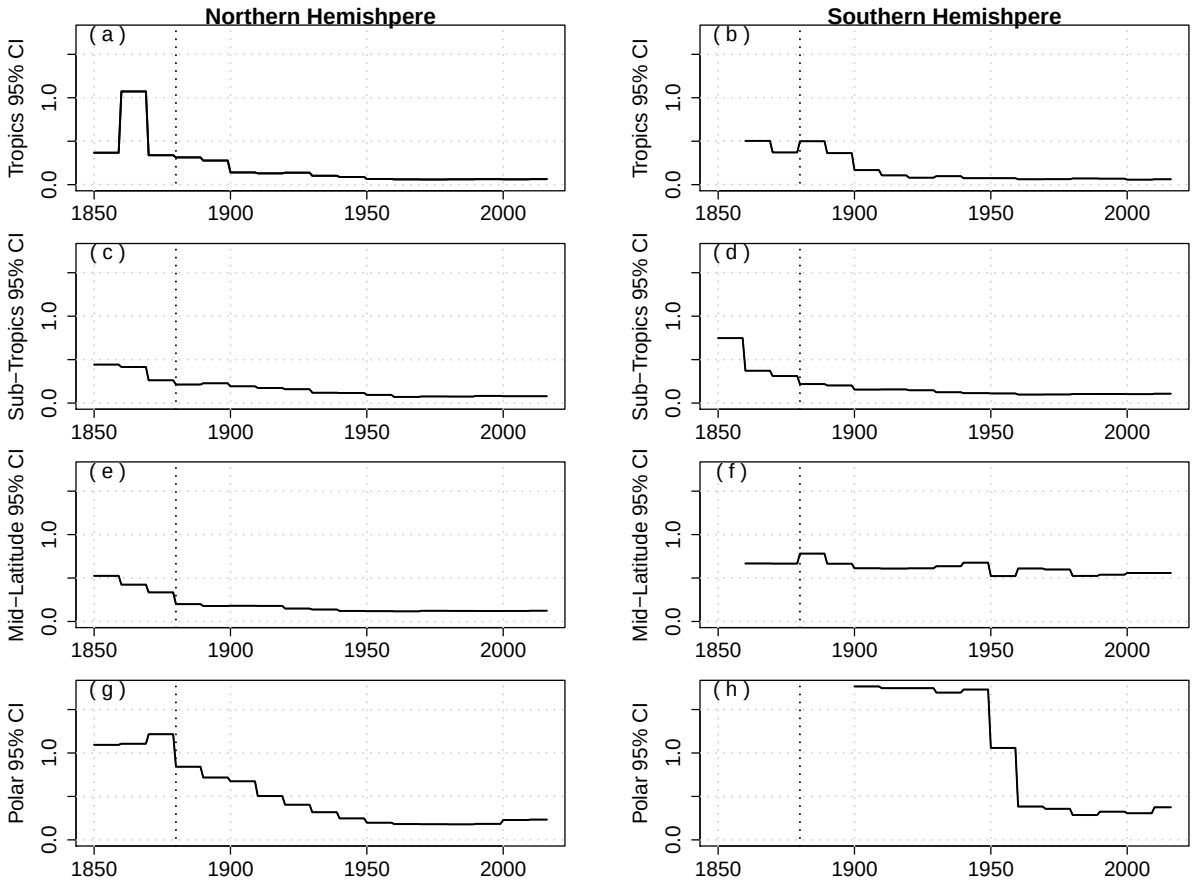

Figure 1.5: Annual land surface temperature anomaly (°C) uncertainty (2σ) per latitudinal band on the GISTEMP grid. The tropics (a)/(b) are 0–23.6°, the sub-tropics (c)/(d) are 23.6–44.4°, the mid-latitudes (e)/(f) are 44.4–64.2°, and the polar regions (g)/(h) are 64.2–90°. The dotted line marks 1880, the current start date of production GISTEMP.

limited number of stations in the southern hemisphere available for comparison.

Further decomposing the sampling uncertainty analysis to the GISTEMP band level highlights latitudinal regions where the station record may be less reliable. Figure 1.5 shows the time series for each of the 8 latitudinal bands used in the GISTEMP analysis. The polar series confirm that these regions are driving the decrease in sampling uncertainty for both hemisphere.

Combining the improved total LSAT uncertainty with the GISTEMP land surface temperature time series gives a intuitive description of the certainty of the land warming trend over the modern record period. Figure 1.6 shows the LSAT time series from the operational GISTEMP analysis with confidence intervals according to the sampling and homogenization uncertainties. The magnitude

in the trend is many times greater than the uncertainty at any period. Additionally, the uncertainty is much lower in the 1960–present period in which much of the warming has occurred.

1.5.2 LSAT Extensions Results

Sampling Bias Results

Since the results of the sampling bias assessment were not robust among reanalyses,the results for all three reanalyses are shown in Figure 1.7. In general, the JRA55 and ERA5 products agree, with MERRA being an outlier. There is no evidence for sampling bias for the in-sample 1980–Present time period when using the ERA analysis. The smallest confidence intervals of the three analyses are from ERA5 demonstrating that the non-significance of the bias is a robust result.

As mentioned, the major caveat in the bias calculation is that the climate has been highly non-stationary over the past 150 years and this analysis calculates the bias due to a particular incomplete sampling using the climate changes over the ERA period of 1979–2018. That is, the analysis determines how good of a job a particular station arrangement could do at observing the climate change that has occurred from 1979–2018; a period in which the Arctic is warming faster than the rest of the land. In addition, it is assumed that the Arctic temperature is changing at a fixed multiple of the global average. This assumption is reasonable as model studies have shown that modeling the amplification trend linearly is a reasonable choice over recent decades (Serreze & Barry 2011; Cohen *et al.* 2014).

The large and significant cool biases in the ERA and JRA reanalyses in the early record describe how undersampling the observed 1979–Present temperature change would lead to a biased calculation in the global mean. The approach of the estimates to unbiased mirrors the global coverage shown in Figure 1.2. The relationship between coverage and bias in estimating the 1979–present warming makes sense, particularly because station coverage in polar regions was limited or nonexistent in the early record and Arctic temperature changed more rapidly over the past few decades.

Limiting Uncertainty Results

Running the sampling uncertainty analysis assuming perfect coverage suggests that 0.034 is the limiting sampling 95% confidence interval for the GISTEMP method. In other words, adding additional station observations will not reduce the sampling uncertainty below this level. The current coverage is already quite close to this value as shown in Figure 1.8 implying that the GISTEMP model is close to the limiting coveragel. Roughly speaking, the limiting uncertainty decreases with the amount of smoothing in the interpolation. As station coverage continues to improve, the choice of interpolation in GISTEMP should be revisited.

The limiting sampling bias is found to be significant, albeit small. The GISTEMP procedure overestimates the true global mean LSAT over the ERA5 record by 1.5% with a 95% confidence interval of (1.0%, 2.0%). A small limiting bias again suggests a reduction in the smoothing radius as full coverage is approached. In the context of the results in the previous section, production GISTEMP is nearly unbiased, even in the pathological limiting case.

Averaging Method Comparison Results

Figure 1.8 compares the LSAT sampling uncertainty from the simple latitude-weighted mean and GISTEMP band mean methods. The GISTEMP method almost always outperforms the simple method with the 1890s and 1900s being the only exception. Furthermore, the GISTEMP method outperforms the simple method by up to 50% in the 1930s and 1940s, primarily due to the added Arctic coverage providing better NH polar band estimates. The results demonstrate the value added by the GISTEMP averaging scheme by leveraging the zonal correlation of temperature anomalies.

1.5.3 Ocean

The global uncertainty from the ERSST large ensemble using the GISTEMP averaging scheme resembles the global uncertainty calculated by the ERSST team. Similar uncertainty is expected as the GISTEMP averaging scheme converges to a latitudinal-weighted grid cell average as missing data approaches zero and the ERSST large ensemble has complete coverage of the oceans. The

GISTEMP operational global annual average SST time series is shown in Figure 1.9. As in the LSAT global time series, the magnitude of the warming trend dominates the uncertainty of the calculation.

Looking at the hemispheric uncertainty in the annual SST anomaly, there are minor differences between the two hemispheres (Figure 1.10). The larger uncertainty in the southern hemisphere post 1945 drives the global uncertainty as the southern hemisphere has double the area occupied by ocean compared to the northern hemisphere.

1.5.4 Total Global Uncertainty

Figure 1.11 shows the production GISTEMP global time series with the 95% confidence interval calculated in this study. The confidence interval has been added to the distributed GISTEMP time series facilitating uncertainty quantification in studies that utilize the GISTEMP product. As in both the SST and LSAT time series, the warming signal is greater than the underlying uncertainty. The possible uncertainty of the warming signal is discussed in the following section.

As in the land and ocean analyses, the global uncertainty is decomposed into northern and southern hemisphere uncertainties (Figure 1.12). Following the larger land uncertainty and comparable ocean uncertainty, the total uncertainty on the annual hemispheric mean is almost always larger in the southern hemisphere.

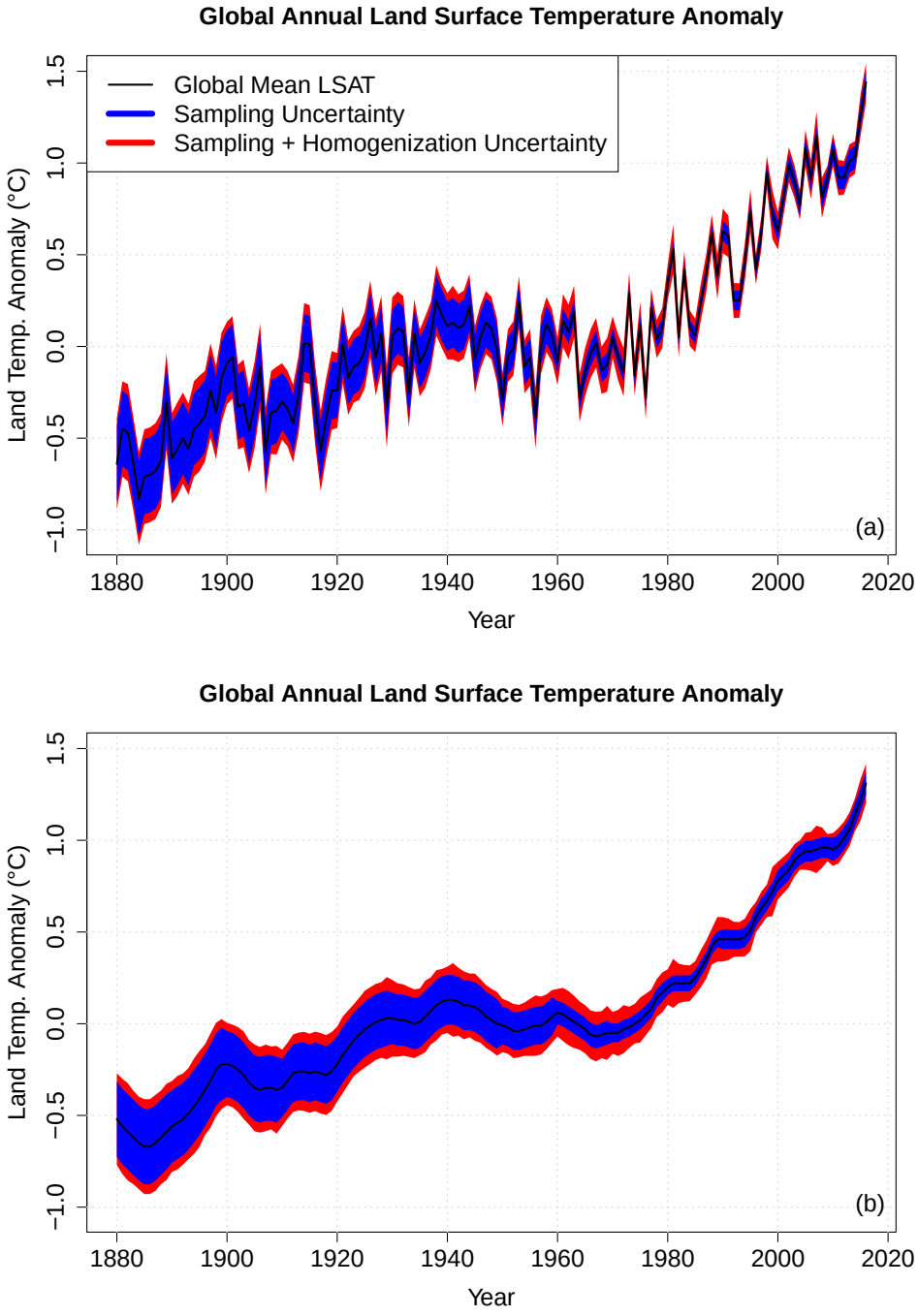

Figure 1.6: The GISTEMP land-only mean with 95% confidence intervals for (a) annual mean and (b) annual mean smoothed by LOWESS with 5-year bandwidth. For both plots, the envelopes show the annual uncertainty of the sampling uncertainty alone as well as the total uncertainty when including the homogenization. Anomalies are calculated with respect to the 1951-1980 climatology. The annual uncertainty on the 5-year smoothed series is presented to illustrate that the trend has much larger magnitude than the uncertainty.

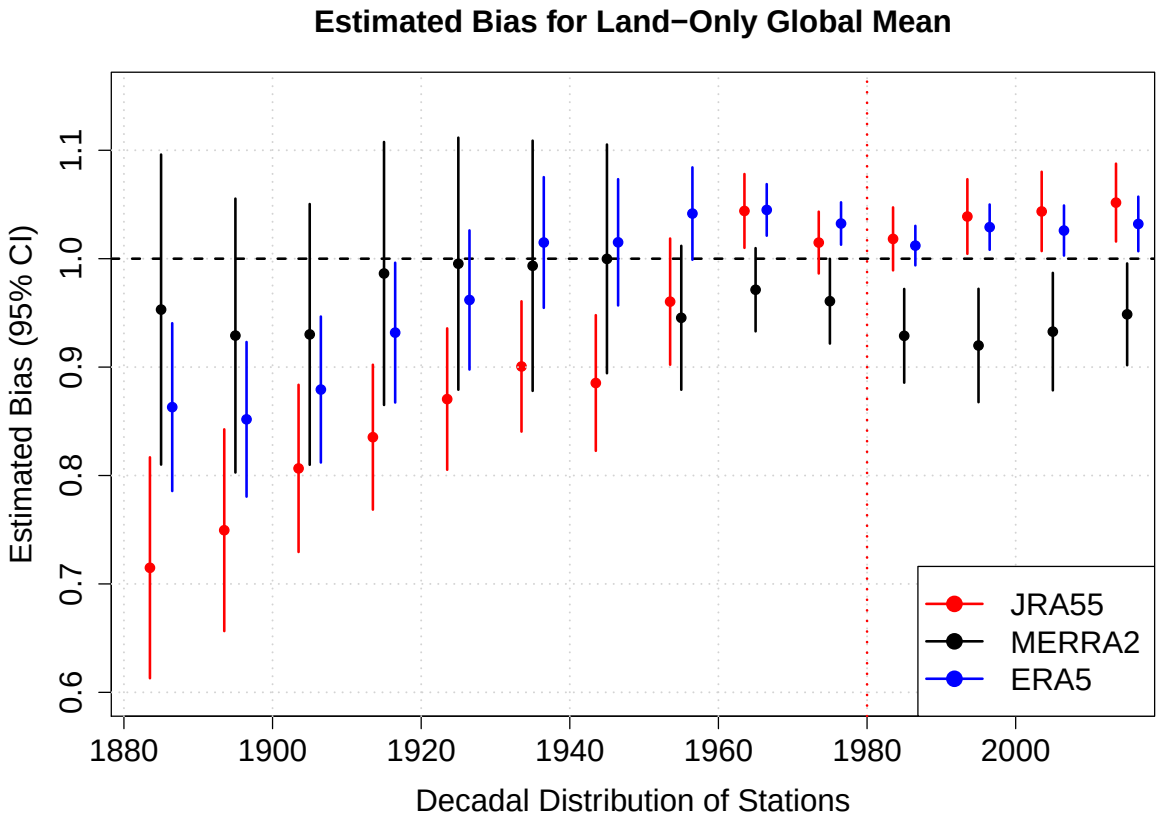

Figure 1.7: Estimates of the scaling bias on the global mean anomaly due to the decadal incomplete sampling in the LSAT for each of the three reanalyses. The line at 1.0 signifies an unbiased estimate and confidence intervals larger or smaller than this value signify statistically significant bias. The red line signifies the start date of the products; decades after this point can be interpreted as a measure of the bias in the global mean of GISTEMP.

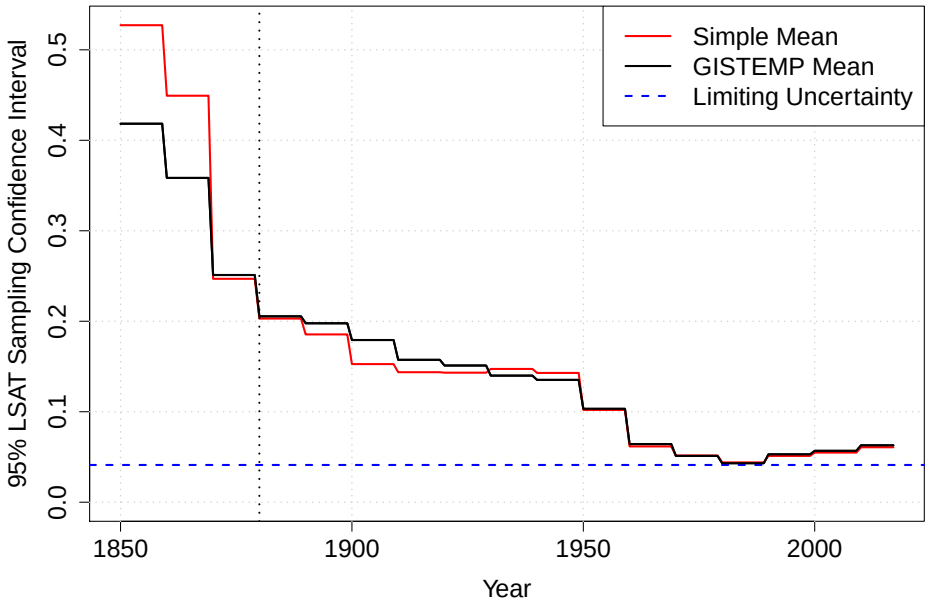

Figure 1.8: A comparison of the sampling uncertainty in the global land-only annual mean temperature anomaly when using the GISTEMP averaging scheme and a simple cosine-weighted mean. The limiting mean is sampling uncertainty found in the ERA5 sampling analysis assuming that there is a station at every grid point and represented the uncertainty introduced into the estimate by the interpolation.

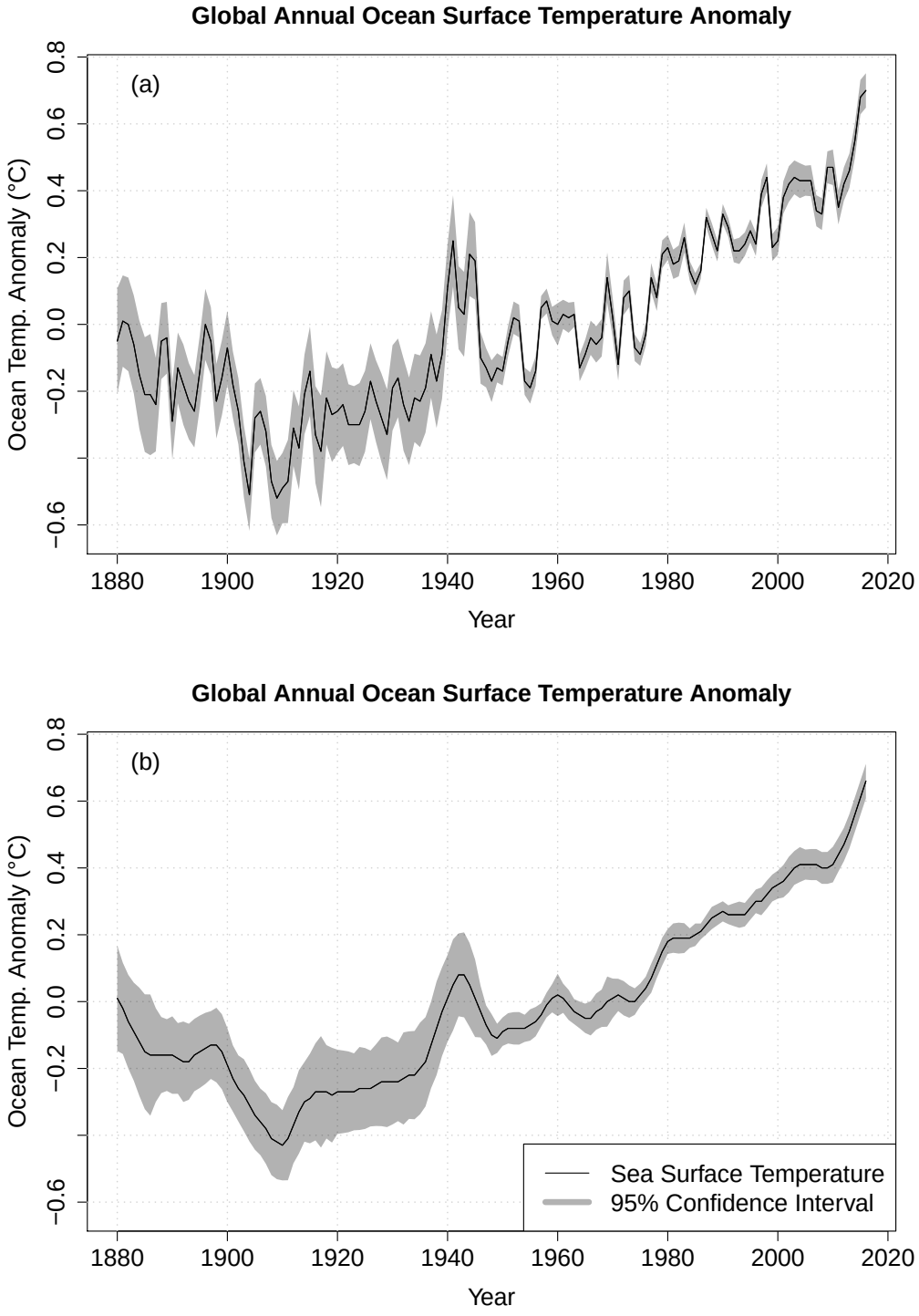

Figure 1.9: The GISTEMP ocean-only mean with 95% confidence intervals for (a) annual mean and (b) annual mean smoothed by LOWESS with 5-year bandwidth. The envelopes show the annual SST parametric uncertainty as calculated from the ERSSTv4 large ensemble. Anomalies are calculated with respect to the 1951-1980 climatology. The annual uncertainty on the 5-year smoothed series is shown to illustrate that the trend has much larger magnitude than the uncertainty.

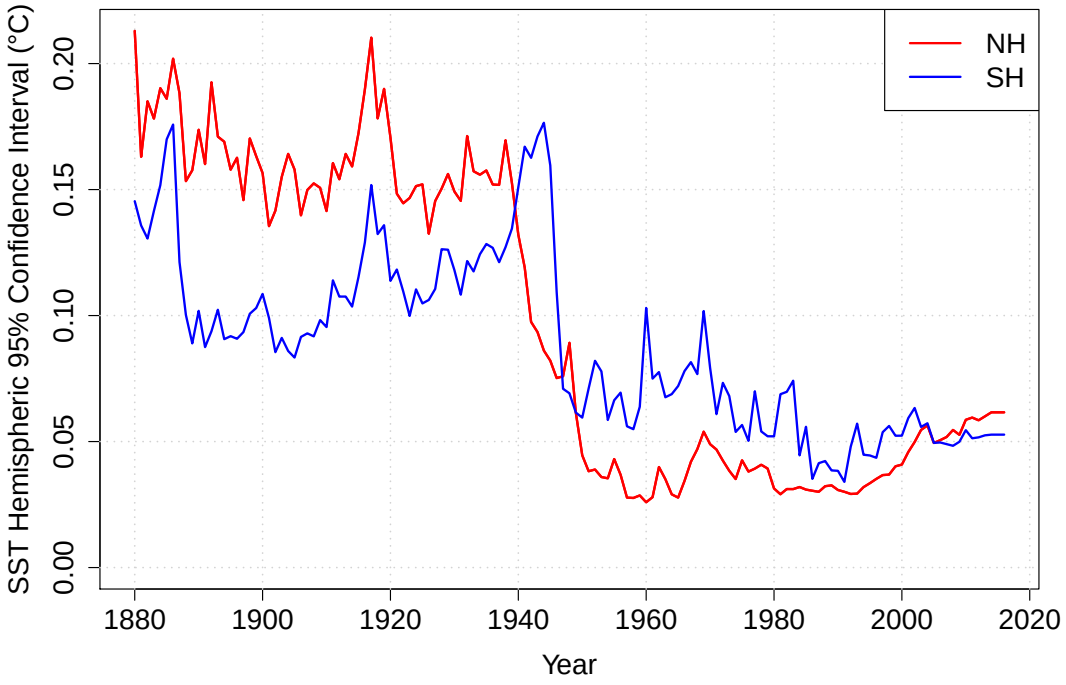

Figure 1.10: Annual sea surface temperature anomaly parametric uncertainty (2σ) per hemisphere calculated using the ERSSTv4 large ensemble with the GISTEMP averaging scheme.

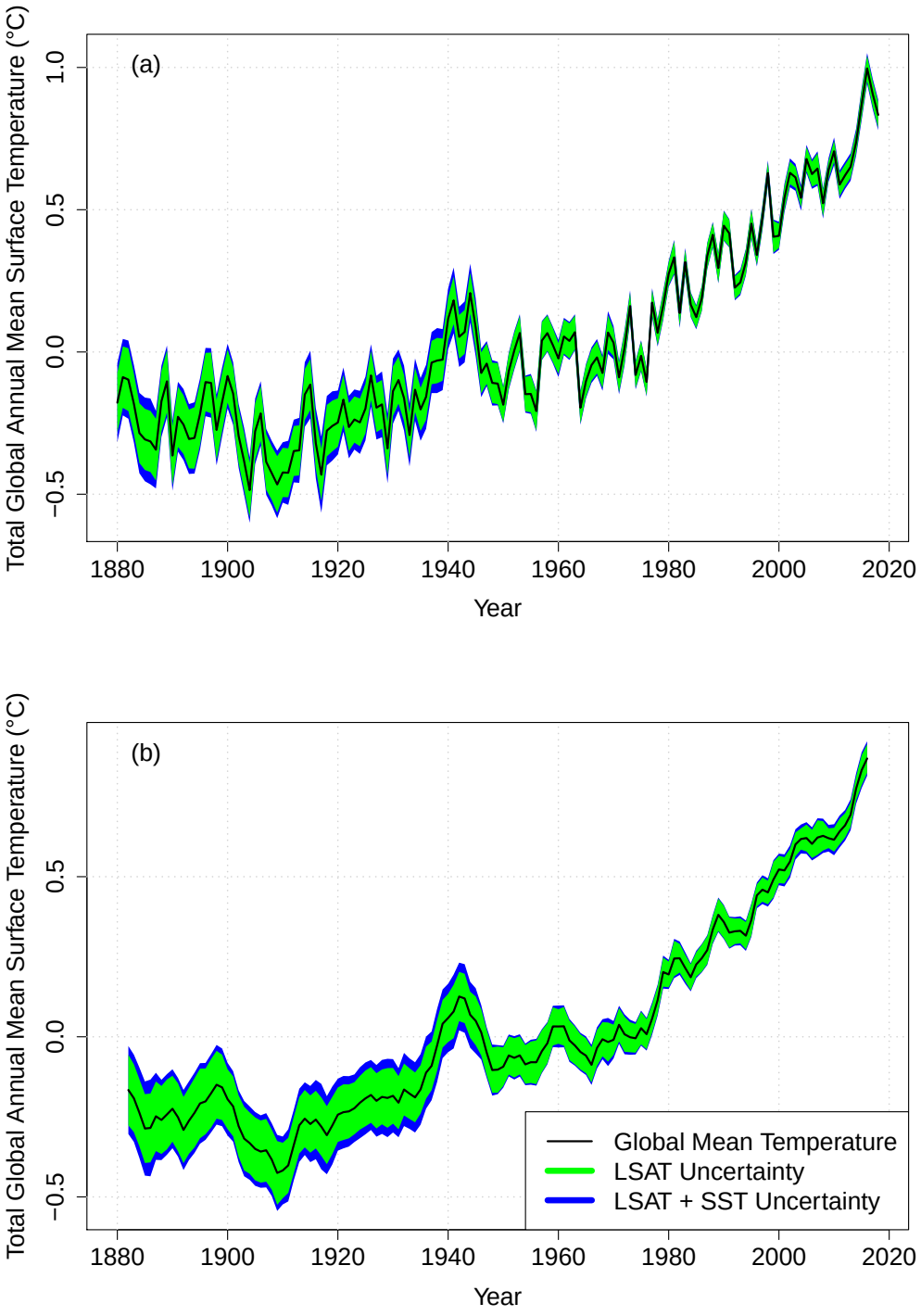

Figure 1.11: The production GISTEMP global mean temperature time series with the total (LSAT and SST) 95% confidence interval calculated in this study for (a) annual mean temperature and (b) annual mean temperature smoothed with LOWESS with 5-year bandwidth. Anomalies are calculated with respect to the 1951-1980 climatology. The annual uncertainty on the 5-year smoothed series is shown to illustrate that the trend has much larger magnitude than the uncertainty.

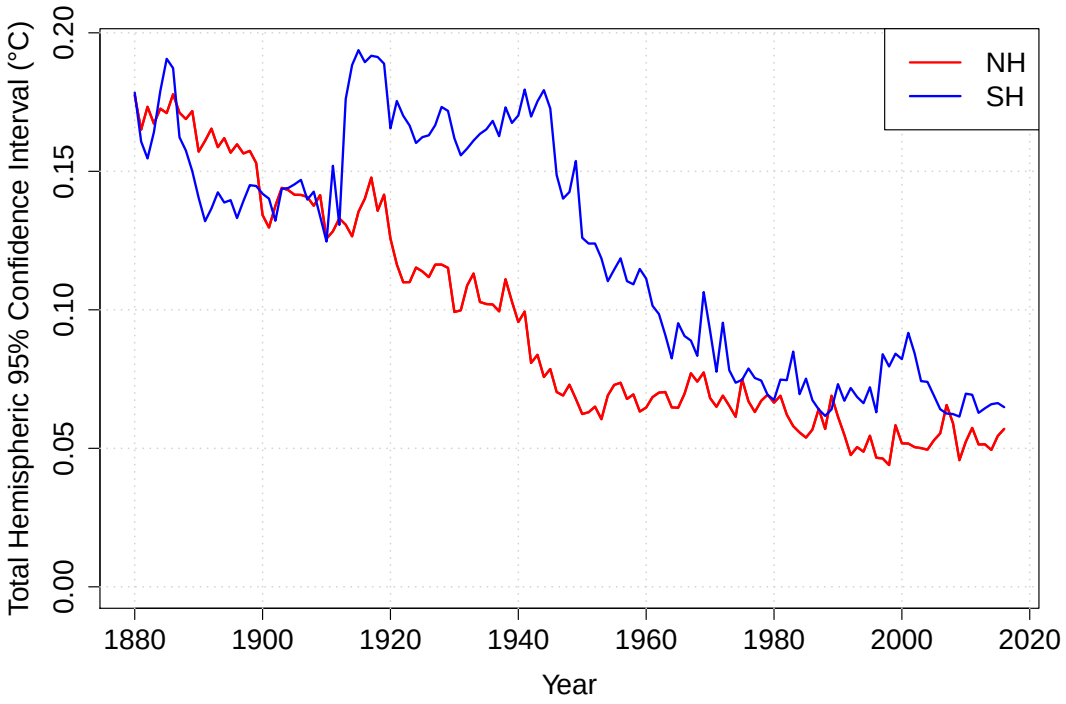

Figure 1.12: Annual mean temperature anomaly total uncertainty (2σ) per hemisphere.

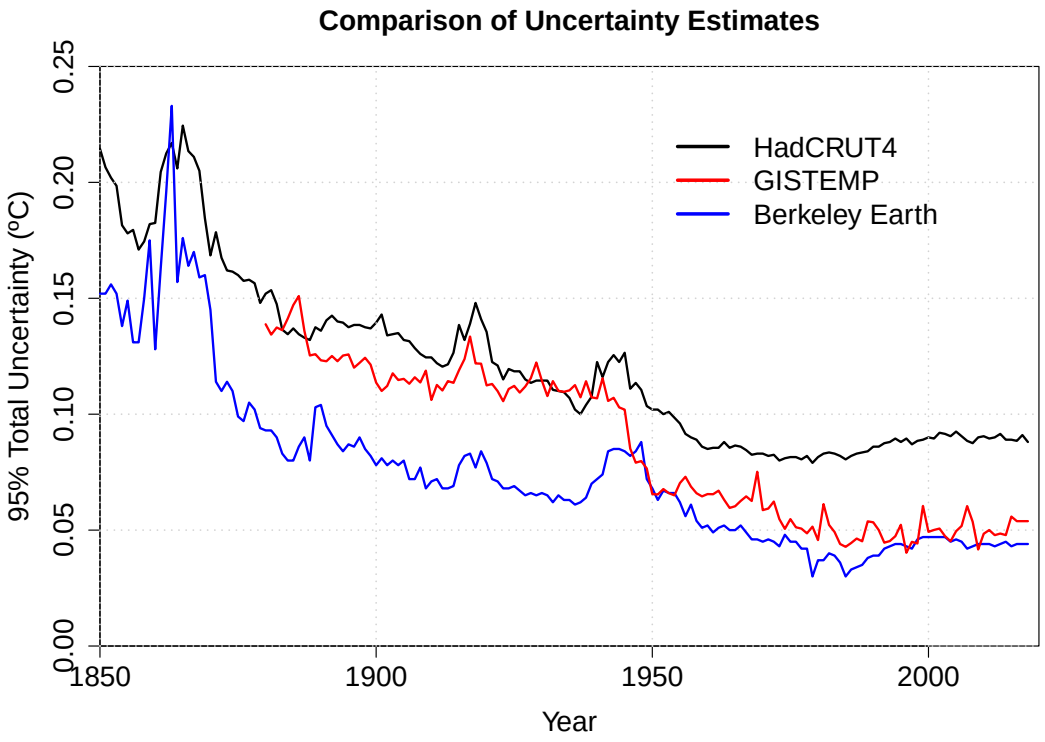

Figure 1.13: Comparison of total uncertainty (95% confidence interval) in three independent global analyses, HadCRUT4, GISTEMP (this paper), and Berkeley Earth.

1.6 Discussion

Since the first GISTEMP estimates in the 1980s, there have been large increases in the amount of data ingested, improvements in the homogenization of station data to remove non-climatic effects, and the incorporation of ocean data, but not much change to the global mean calculation methodology. These data changes have produced variations over time of the global annual mean record that, while not a controlled exploration, are indicative of the structural uncertainties in the product that arise indirectly through changes in data availability and processing. The new analysis presented here is far more complete, but it is appropriate that recent versions of GISTEMP fall within the uncertainties shown in Figure 1.11.

The improved assessment of uncertainty in the GISTEMP product is a function of three new developments: the Monte Carlo ensembles that have been done for the input data (ERSST and GHCN), the upgrading of the GISTEMP code base, and the evolving standards in uncertainty quantification in climate science. These threads have made the current study far more tractable than it would have been a decade ago.

The existence of the new uncertainty product now allows us to be more rigorous in assessing the strength of claims of records and trends in the data itself, but also to improve the propagation of that uncertainty into, for instance, detection and attribution exercises for constraining anthropogenic climate change.

One persistent question is whether it makes sense to extend the GISTEMP product prior to 1880, to perhaps as early as 1850 (for instance, to help estimate a 19th Century baseline climatology (Hawkins *et al.* 2017)). Figure 1.3.2 demonstrates that the sampling in the 1870s is not that much worse than the 1880s, but unfortunately, the homogenization analysis does not extend before 1880, and nor does the ERSST data. This is an issue that will continue to be explored.

1.6.1 Probability of a new warmest year record

The addition of the global annual mean uncertainty values calculated in this study to the widely distributed GISTEMP surface temperature product will enable users to include more informed probabilistic statements of uncertainty in their research. One such example is the probability of warmest year calculation which is often cited in scientific and popular literature.

Given the strong trend in global mean temperatures since the 1970s, NASA/GISS has frequently reported on new records for annual means over the instrumental period (11 times since 1988). This naturally leads to the question of how confident one can be in declaring that any particular record year in the GISTEMP index, was in fact the warmest year in the real world since 1880. Discussion of this uncertainty has been a focus of the NOAA and NASA annual briefings since 2014, which at the time was the warmest year in the record (NASA Public Affairs 2015). With the major El Niño event in 2015/2016, both subsequent years were notably warmer (NASA Public Affairs 2016; NASA Public Affairs 2017), but how certain are these statements?

A Monte Carlo estimate of the warmest year is made by determining which year has the highest temperature anomaly after either independent or autoregressive simulations of the uncertainties. The probability that a given year was the warmest year on record to date is then the number of simulations in which it is the warmest year divided by the total number of simulations. This method is used to reassess how well NASA's recent statements on the probability of warmest years match up to the updated uncertainty calculations presented in this study.

In January 2015, NASA reported that 2014 was likely the warmest year with 38% likelihood (NASA Public Affairs 2015) based on a simple assumption of linearly increasing uncertainty based on the (Hansen *et al.* 2010) estimates. This estimate was conservative and 2014 actually had a 79% chance of truly being the warmest year in the instrumental period. Assuming autocorrelated uncertainties, this reduces slightly to 75% since the next most probable warmest years were non-consecutive (2010 and 2005). The following year, NASA reported a likelihood that 2015 was the new record warmest year was 96%, which compares to a 99.99% probability calculated now (regardless of whether independent or autocorrelated uncertainties are used).

Assuming that uncertainties in the annual mean are independent from year to year, 2016 is most likely the warmest year in the last 139 (1880–2018) years with 86.2% certainty. The other years that could plausibly have been the warmest were 2017 (12.5 % probability), 2018 (1.2% probability), and 2015 (<0.1% probability). While the GISTEMP-estimated mean global temperature is larger in 2015 than in 2018, the uncertainty in the 2018 mean is larger, primarily due to an increase in the LSAT homogenization uncertainty. Therefore, 2015 will rank higher on the warmest years than 2018 on average, but the additional uncertainly in the 2018 mean gives it a greater chance of being the warmest year.

This probability is also calculated using autoregressive uncertainties. Unlike the uncertainty in temperature change, autoregressive uncertainties give more certainty to 2016 being the warmest year with a simulated 87.2% certainty. Since all of the candidate years have occurred in consecutive years, positive autocorrelation reduces expected difference in uncertainty.

While the AR(1) calculation is a reasonable choice for comparing anomalies over a short time period, such a calculation is not statistically sound for longer-term analyses using the uncertainties calculated in this study. Components of the uncertainty, particularly the homogenization uncertainty, persist over many decades reflecting large shifts in the record that propagate in time. These types of uncertainties are best represented in an uncertainty ensemble which has not yet been created for GISTEMP.

1.6.2 Comparison to other uncertainty estimates

Two of the other products shown in Fig. 1.1 have independently derived total uncertainties, specifically HadCRUT4 and BEST. Figure 1.13 shows the comparison of the three 95% confidence intervals. The overall magnitudes are similar, with close agreement with the HadCRUT4 uncertainty pre-1945 and with BEST post-1945. The character of the change around 1945 is driven primarily by the reduction in SST uncertainty in ERSST and reduction in the greater reduction in GISTEMP LSAT sampling uncertainty relative to HadCRUT4.

1.7 Conclusion

The presented uncertainty quantification of the global annual mean surface temperature anomaly in the GISTEMP product brings this analysis up to the enhanced standards of its peers. It is the hope that this study and its findings will aid the interpretation and utility of this widely used product. This study has focused on the global and hemispheric annual means, but the procedure can equally be used to improve the uncertainty analysis of regional and monthly data products and these will be pursued in further work.

Chapter 2: A NASA GISTEMPv4 Observational Uncertainty Ensemble: Regional and Monthly Uncertainty

In recent years, better characterization of global and hemispheric trends has become available (Lenssen *et al.* 2019; Morice *et al.* 2020; Huang *et al.* 2020; Rohde & Hausfather 2020), but the methodologies are not necessarily applicable to smaller regional areas, or monthly means, where station sparsity and other systematic issues contribute to greater uncertainty. This chapter describes the creation of a large ensemble of temperature reconstructions for the Goddard Institute for Space Studies (GISS) Surface Temperature product (GISTEMP) product (Hansen *et al.* 2010; Lenssen *et al.* 2019) that allows the characterization of regional and monthly uncertainty and applies this ensemble to characterize uncertainty in two problems of societal and scientific interest.

Chapter 1 of this dissertation presents an initial uncertainty model for GISTEMP (Lenssen *et al.* 2019). The study provides two critical components necessary for the characterization of uncertainty in regional and monthly temperature change; a framework of uncertainty for the GISTEMP procedure was formalized, and then the framework was applied to quantify the uncertainty on global and large-scale annual mean temperatures. Uncertainty was divided into independent, quantifiable components to represent the major sources of uncertainty in the GISTEMP product. In the Land Surface Air Temperature (LSAT) record, the primary sources of uncertainty are sampling uncertainty and station homogenization uncertainty. Sampling uncertainty is an umbrella term for uncertainties introduced into global and regional means due to incomplete spatial and temporal coverage. Station homogenization uncertainty accounts for possible errors arising from the adjustment of single station records to correct artificial break points due to changes in observing methods or station locations. Using this framework, operational GISTEMP now provides an

estimate of global mean uncertainty.

Extending the results of Chapter 1 to regional and monthly mean temperature is a significant undertaking. There are two primary difficulties: (1) moving from global and large-scale spatial means to small-scale spatial means and (2) quantifying the temporal dependence of the uncertainty to provide accurate estimates of the uncertainty in changes in the mean. The temporal structure of the uncertainty is the most urgent problem, and is particularly important to capture correctly for accurate uncertainty quantification in global and regional trends. The simple 95% confidence intervals for the global mean discussed in Chapter 1 do not provide any information about the temporal structure of uncertainty. It is well known that significant temporal autocorrelation in uncertainty exists, primarily driven over the land surface by the homogenization of the station record (Menne *et al.* 2018). The temporal structure of this homogenization uncertainty is highly persistent and not well represented by common statistical models for time series such as auto-regressive and other ARIMA models.

Creating ensembles of equally likely realizations of the global temperature record is the current best practice for quantifying and presenting uncertainty in gridded monthly historical temperature analyses. The Hadley Centre with HadCRUT4 (Morice *et al.* 2012) and HadCRUT5 (Morice *et al.* 2020) as well as NOAA's GlobalTemp Version 5 (Huang *et al.* 2020) have shifted their global temperature uncertainty products from simple confidence intervals to such uncertainty ensembles. These ensembles are able to represent the complex and persistent temporal structure of the uncertainties inherent in the global temperature record, enable more accurate estimates of uncertainty in global and regional temperature change, and make it straightforward to include observational uncertainty in subsequent analyses.

This chapter presents a GISTEMPv4 uncertainty ensemble from 1880-2020. Following the operational GISTEMP analysis, Land Surface Air Temperature (LSAT) is calculated from station records from NOAA's Global Historical Climatology Network (GHCN) monthly version 4 (GHCNm v4; Menne *et al.* 2018). Sea Surface Temperature (SST) data from NOAA's Extended Reconstructed Sea Surface Temperature version 5 (ERSSTv5; Huang *et al.* 2017) is merged with

the LSAT analysis to form the GISTEMP global land-ocean analysis (Hansen *et al.* 2010; Lenssen *et al.* 2019).

One of the primary motivations behind the GISTEMP uncertainty ensemble is to increase the use of observational uncertainty in studies relying on historical temperature data. The global historical temperature record, and GISTEMP in particular, is widely accessed, cited, and used in subsequent studies. From the 10-most cited papers that cite Lenssen *et al.* (2019), direct applications of GISTEMP include: the validation of historical runs of global general circulation models (Swart *et al.* 2019; Held *et al.* 2019; Danabasoglu *et al.* 2020; Notz & SIMIP Community 2020), retrospectively evaluating past climate model projections (Hausfather *et al.* 2020), verifying estimates of climate sensitivity (Tokarska *et al.* 2020), quantifying changes in mean climate and extremes over the historical period (Myhre *et al.* 2019), and estimating the cost of carbon emission in the global economy (Carleton *et al.* 2020). Despite the scientific and societal importance of the problems addressed in these studies, and their reliance on the historical global temperature record, none of them include observational uncertainty as part of their methodologies.

A potential reason for the near ubiquitous omission of observational uncertainty in analyses involving historical climate data is the lack of accessible, interpretable, and easily-implemented uncertainty products. Observational ensembles are a large step forward as posterior distributions of a key result in an analysis that relies on historical temperature can be constructed nearly trivially by running the analysis of interest on each uncertainty ensemble member. However, these ensembles are relatively new, only appearing in the last decade, with very few studies utilizing them. In this chapter, two applications of the uncertainty ensemble in relatively simple analyses are presented to illustrate the necessity of including observational uncertainty in many studies as well illustrate the ease of implementing observational uncertainty through running a single analysis over each ensemble member. The first example is the calculation of country-level monthly mean series as provided to the United Nation's Food and Agriculture Oragnization (FAO) for a map-room on global climate and agriculture conditions (FAO 2022). The second example is a study revisiting the common claim that the Arctic is warming twice as fast as the rest of the planet and updates this

statistic to an Arctic warming rate of around 4 times greater than the global mean (Jacobs *et al.* 2021).

The remainder of this chapter is organized as follows. Section 2.1 outlines the source data used for the analyses. Section 2.2 provides a brief background on the LSAT uncertainty model discussed in detail in Chapter 1. Section 2.3 presents the methods used to generate the GISTEMP uncertainty ensemble. Section 2.4 summarizes the statistical properties of the GISTEMP uncertainty ensemble. Section 2.5 gives a simple application of the ensemble by calculating the uncertainty in country-level mean temperature over 1960-2016. Section 2.6 provides a second application of the ensemble to the relative rates of Arctic and global warming in recent decades. Section 2.7 summarizes the results.

2.1 Input Data

2.1.1 LSAT Data: GHCNm Version 4

The GHCNm version 4 dataset is a quality-controlled collection of station-based land temperature records at the monthly temporal resolution (Menne *et al.* 2018). All station records included in the dataset are processed to correct for irregularities arising due to change of station location, measurement method, and surrounding land cover. In addition to a single authoritative station record, GHCNm v4 also contains a 100+ member uncertainty ensemble that spans the parametric uncertainty arising from choices in the homogenization procedure. This study uses the GHCNm v4 ensemble to capture the station and bias uncertainties as is discussed further in Section 2.2.

2.1.2 SST Data: ERSSTv5

The latest version of NOAA's gridded sea surface temperature analysis, ERSSTv5, is used to quantify the historical monthly SST anomalies globally (Huang *et al.* 2017). The product is distributed on a $2^\circ \times 2^\circ$ grid that is interpolated to the 8000 GISTEMP equal area boxes to be compatible with the operational GISTEMP python analysis. The uncertainty quantification in ERSSTv4/v5 breaks down ocean uncertainty into parametric uncertainty, or uncertainty arising

from choices the ERSST method, and reconstruction uncertainty, or uncertainty arising from estimating global SST from limited SST records. The ERSSTv5 uncertainty model contains small updates to the parameters from the ERSSTv4 uncertainty method, but is otherwise identical (Liu *et al.* 2015b; Huang *et al.* 2016b; Huang *et al.* 2017). ERSSTv5 provides a 1,000 member uncertainty ensemble of gridded SST fields as well as a 500 member operational uncertainty ensemble, enabling other operational products to take advantage of their uncertainty assessment.

2.1.3 ERA5 Reanalysis

This study uses the monthly ECMWF Reanalysis version 5 (ERA5) from 1951-2020 as an approximate, full-coverage historical LSAT record (Hersbach *et al.* 2020). The 2 m temperature field is averaged to the final $2^\circ \times 2^\circ$ GISTEMP uncertainty ensemble grid to facilitate direct comparison between ERA5 and GISTEMP. ERA5 is chosen as the reanalysis as it best replicates the observed global mean over its period (Lenssen *et al.* 2019; Hersbach *et al.* 2020). As shown in Chapter 1, global and large-scale uncertainty estimates derived from ERA5 agree with the JRA55 and MERRA2 reanalyses.

2.2 LSAT Uncertainty

There are three major, statistically independent, categories of uncertainty that arise in the LSAT record (Figure 2.1). A brief introduction to these uncertainties is provided, see Chapter 1, Morice *et al.* (2012), and Lenssen *et al.* (2019) for more details. The uncertainty ensemble model accounts for station and bias uncertainties through the GHCNm v4 ensemble as detailed in Section 2.3.1 and sampling uncertainties following the methodology outlined in Section 2.3.2.

Station uncertainty arises from errors in the temperature record of a single station. The first sources of station uncertainty are instrumental errors from limited thermometer precision. These are relatively small and uncorrelated in space and time, making them essentially a non-issue for monthly records (Morice *et al.* 2012). The other and more significant sources of station uncertainty are inhomogeneities, or non-climatic shifts in mean in station records. These can arise from local

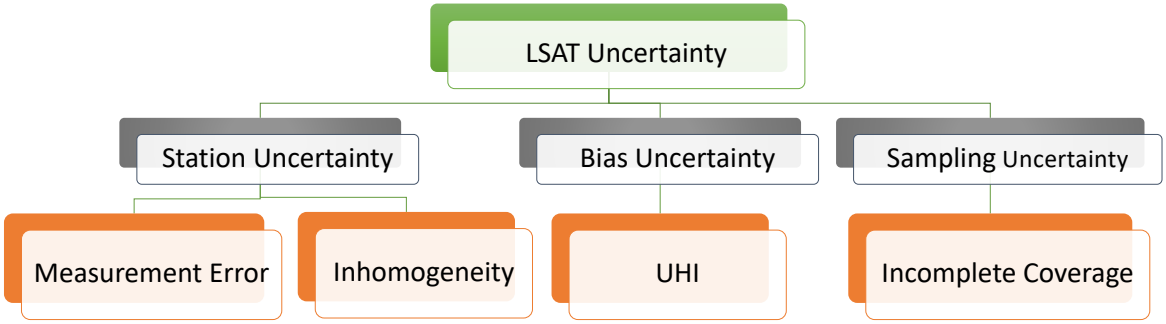

Figure 2.1: Decomposition the total LSAT uncertainty into the three major categories and the most common sources. The connections on the chart denote dependence, implying statistical independence between cells that are not connected.

microclimate shifts or changes in the station measurement method. As mentioned in Section 2.1.1, the GHCNm dataset detects and corrects for these homogeneities, but this is a difficult problem, and uncertainty due to statistical estimation of these corrections adds uncertainty to estimates of regional and global temperature while reducing any bias (e.g. Hausfather *et al.* 2013).

Bias uncertainty refers to anthropogenic changes in local climate that are not representative of changes in the regional or global climate system. Generally, this category of uncertainty refers to the enhanced warming observed in cities commonly referred to as the urban heat island (UHI) effect. Again, the GHCNm dataset accounts for these, but corrections add uncertainty to the surface temperature record. This issue has also been tackled in GISTEMP through the use of nightlights to characterize the more urban environments (Hansen *et al.* 2010).

Sampling uncertainty arises from estimating regional and global temperature due to limited spatial and temporal coverage. The coverage of the global observation network does not fully cover the land surface and has changed over time. GISTEMP uses the spatial correlation of temperature anomalies to increase the coverage. By interpolating the station-level anomalies, GISTEMP is able to make a more accurate estimate of the global temperature, but at a cost of introducing uncertainty into fine-scale regional means.

Due to the sources of these uncertainties, representing uncertainty at each location and month as independent or correlated Gaussian random variables is an incomplete method. In particular,

uncertainty arising due to errors in the homogenization process have long-term persistence that are not well-suited to ARIMA time series models. Construction of an LSAT uncertainty ensemble using an iterative process where each step accounts for one of the two major categories of uncertainty, is better able to better represent the spatiotemporal structure of the uncertainties, and can be understood in isolation.

2.3 Methods

2.3.1 GHCN-ERSST-GISTEMP Ensemble

The core of the GISTEMP uncertainty is the GHCN-ERSST-GISTEMP ensemble which is generated by running 100 potential station records from the GHCNm v4 uncertainty paired with 100 of the ERSSTv5 uncertainty ensemble members. These station record-ocean record pairs are then run through the operational python GISTEMP analysis code (Barnes & Jones 2011). This process is outlined visually in the code flowchart (Figure 2.2) through steps at the top of the chart leading up to the block labeled "Run GISTEMP with NOAA Ensemble Data." The GHCN ensemble is 100 possible station records in the same format as the version of GHCN used in production GISTEMP. Temperature fields and mean time series are calculated as described in Chapter 1 (Hansen *et al.* 2010; Lenssen *et al.* 2019). The 100 member GHCN-ERSST-GISTEMP ensemble accounts for all quantified SST uncertainty as well as homogenization and bias LSAT uncertainty. Thus, all that remains is to quantify the LSAT sampling uncertainty arising from limited station coverage as detailed in Section 2.3.2.

Managing the output of the ensemble members to ensure computations are working as intended and output is documented appropriately, is a critical part of the workflow. The steps in the flowchart prior to the "Run GISTEMP with GHCN/ERSST Ensemble" block describe the data and code management processes needed to organize the analysis on the NASA supercomputer DISCOVER. By porting the analysis to DISCOVER, the GHCN-ERSST-GISTEMP ensemble is able to be generated in under an hour as opposed to the days it would take to run on a typical laptop.

2.3.2 Sampling Uncertainty Ensemble

The 100 member GHCN-ERSST-GISTEMP ensemble detailed above in Section 2.3.1 accounts for the station and bias uncertainties. To incorporate the sampling uncertainty, 5 possible realizations of the sampling uncertainty are simulated and added to each of the 100 members of the GHCN/GISTEMP, resulting in a final uncertainty ensemble of 500 members. This step is performed in R (R Core Team 2020) and relies on the fields package (Nychka *et al.* 2017) for estimation of the spatial model. It is denoted by the blue "Estimation and Simulation of Sampling Uncertainty" on the analysis flowchart (Figure 2.2).

The covariance structure of the sampling uncertainty is quantified using an improved version of GISTEMP sampling uncertainty analysis described in Chapter 1 and Lenssen *et al.* (2019). The ERA5 reanalysis is used as an approximate historical climate with full global coverage. For each decade from the 1880-2020, a proxy station record is created by masking the full ERA5 record by the decadal station coverage. The GISTEMP interpolation step with 1,200km smoothing is applied to the masked ERA5 data resulting in estimates of regional temperature on a $2^\circ \times 2^\circ$ grid. The true temperature anomaly fields from ERA5 are differenced to calculate reconstruction error fields for each timestep in the ERA5 record. These reconstruction error fields are an estimate of the uncertainty in the LSAT field due to limited station coverage. The sampling uncertainty $\hat{\sigma}_{\ell,s}$ for decade ℓ and location s is then estimated through the standard deviation of the error series at that location.

Due to the interpolation in the GISTEMP method, the reconstruction error fields have spatial structure that must be accounted for. To ensure an invertible covariance matrix, a stationary Gaussian process with a Matérn covariance function is used to model the reconstruction error fields (Rasmussen & Williams 2005). The Matérn covariance function provides a flexible class of stationary spatial processes that relies on two key hyperparameters: the shape or smoothness parameter which controls and the range parameter which determines the length-scale of the spatial correlation (Rasmussen & Williams 2005). The Matérn hyperparameters are estimated fitting a simulated Matérn variogram to the variogram of the ERA5 reconstruction error fields (Cressie

2015). A variogram summarizes the spatial structure of a spatial field by determining the average variance of observations within a given radius. The more slowly variance increases with distance, the greater the spatial correlation within the data and visa versa (Cressie 2015).

With the Matérn hyperparameters estimated, covariance matrices $\hat{\Sigma}_\ell$ are calculated for each decade ℓ with variance equal to the sampling uncertainty estimate $\hat{\sigma}^2_{\ell,s}$ and a unit-variance Matérn covariance matrix M as

$$\hat{\Sigma}_\ell = \hat{\sigma}_\ell M \hat{\sigma}'_\ell \, . \tag{2.1}$$

This heteroskedastic, spatially correlated covariance structure does a good job of representing the spatial variability in the LSAT error field for all decades, and performs particularly well in the second half of the record (Figure 2.3).

Independent realizations of the sampling uncertainty for each decade ℓ are simulated as random draws from a mean-zero multivariate Gaussian distribution with covariance $\hat{\Sigma}_\ell$. Note that the use of "independent" here refers to the independence of each of the draws from this spatial process, as each simulated field contains the appropriate spatial covariance structure. To conservatively account for temporal persistence in the sampling uncertainty that is poorly represented by ARIMA models, a random persistence of 1-18 months is simulated for each sampling uncertainty draw. This is an extension of the 12 month persistence of uncertainty used in the HadCRUT5 method (Morice *et al.* 2020) and reduces artifacts in time series while still allowing temporal persistence of sampling errors.

2.4 Results and Discussion

The global and hemispheric annual mean series are calculated with the GISTEMP uncertainty ensemble by applying the GISTEMP averaging scheme to each of the 500 gridded ensemble members. The ensemble median matches very well with operational GISTEMP for each of these series (Figure 2.4). The 95% confidence intervals of mean series are constructed as the empirical 95% confidence interval from the 500 annual mean series. The 95% confidence interval of the ensemble

mean and hemispheric series covers the operational series at every time point, which along with the near-perfect agreement between the ensemble median and the operational series, validates the GISTEMP uncertainty ensemble's ability to accurately replicate the global mean calculation.

The ensemble estimate of uncertainty in the global mean changes somewhat from the calculation performed in Lenssen *et al.* (2019) with the ensemble showing less uncertainty in the first half of the record and more uncertainty in the second half (Figure 2.5). The global uncertainty as calculated with the ensemble changes very little from 1880-2020, dropping from values just over 0.1 °C to around 0.08 °C. This general pattern of global mean uncertainty is very similar to that of the Berkeley Earth product, which also shows a nearly flat uncertainty series in time (Figure 2.5; citerohde2020). The lack of decrease over time in the GISTEMP ensemble uncertainty is potentially due to the more complete inclusion of station homogenization uncertainty, which has been shown to dominate the global land and therefore total global uncertainty in the recent decades (Lenssen *et al.* 2019). It also could be due to the inclusion of sampling uncertainty at the monthly scale that is temporally persistent, leading to greater annual uncertainty than was estimated when assuming independence of annual uncertainty in the initial study. Regardless, the general agreement of the GISTEMP ensemble global mean uncertainty with previous studies again suggests that the GISTEMP ensemble is capturing the global mean and uncertainty pattern.

Looking at the GISTEMP latitudinal band mean estimates from the uncertainty ensemble, large uncertainty in the polar regions appears to be driving the global uncertainty (Figure 2.6). This uncertainty is driven by land and sea ice regions near the poles (Figure 2.7). Again, there is very good agreement between the operational and ensemble GISTEMP with the ensemble confidence interval always covering the operational series. In general, the land surface has greater uncertainty than the ocean at the monthly scale, except during the observationally sparse 1940s (Figure 2.7).

Additional insight on the LSAT uncertainty is gained in separating the total uncertainty fields into contributions due to the sampling uncertainty and the combined bias and station uncertainties as quantified in the GHCN LSAT ensemble (Figure 2.8). The data is plotted as confidence intervals for interpretability. However, since this confidence interval has the same units as the standard

deviation, the total uncertainty is approximately the root sum of squares of the two fields. Thus, the larger of the two uncertainty fields shown in Figure 2.8 will dominate the total uncertainty.

Sampling uncertainty is the dominant source of uncertainty for much of the land surface. The largest values occur in regions that are further than the interpolation range of 250 km from a station. These regions, such as Greenland and portions of the Sahara, have sampling uncertainty equal to the monthly temperature variability. The sampling uncertainties are also high for regions with sparse station coverage such as the Arctic, much of Africa, and the mountainous Andes and the Himalayan regions. Expanding the observation network in these regions would greatly decrease the total uncertainty.

Taking the log ratio of the two uncertainty fields allows for better qualitative comparison (Figure 2.8). The sampling uncertainty dominates over much of the land surface. In these regions improvements in the homogenization procedure will have little-to-no impact on the total uncertainty. However, in the purple regions occurring mostly in South America and Southern Africa, improvements to the homogenization through new methods or the incoporating of additional station records would decrease the total uncertainty.

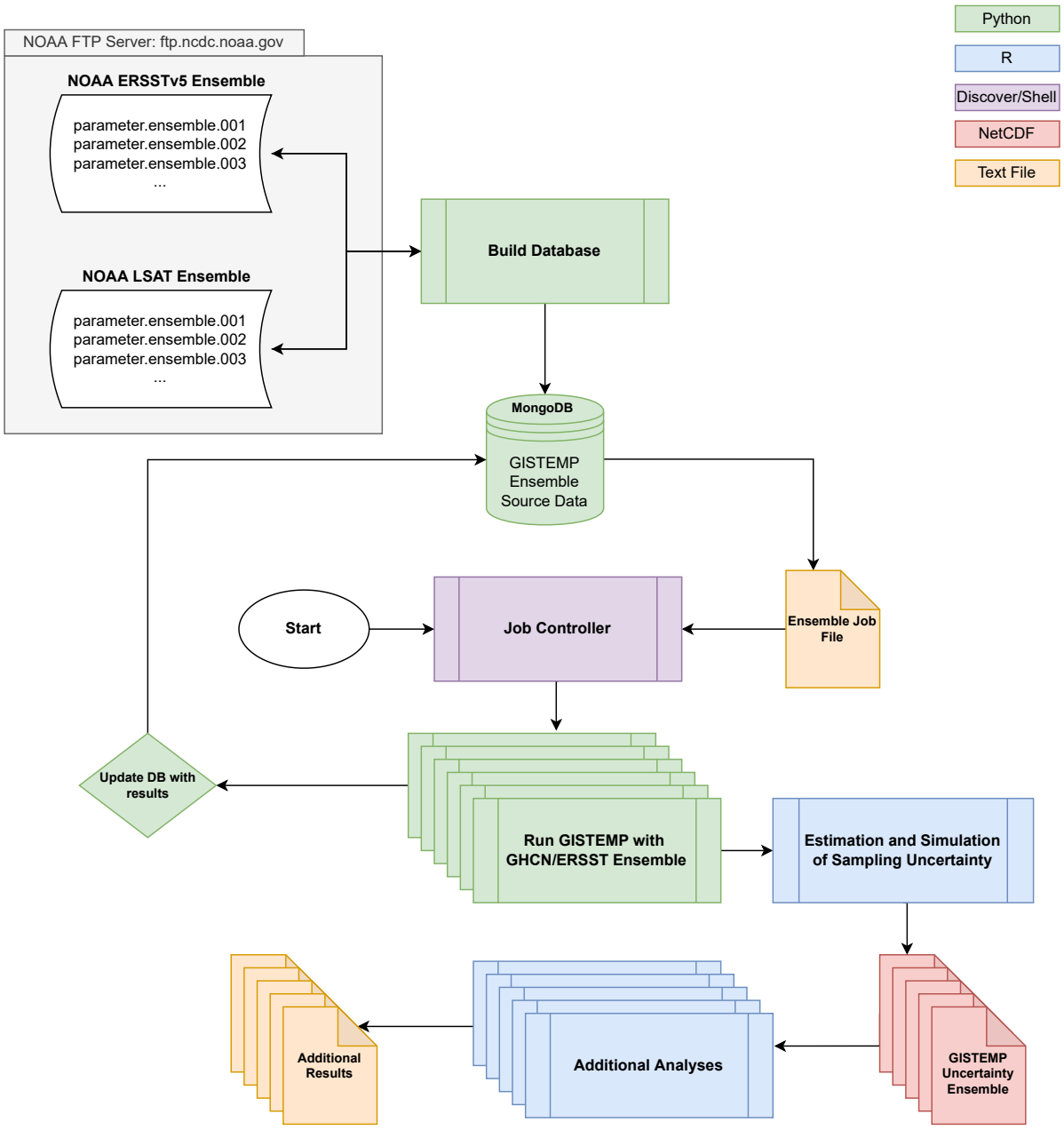

Figure 2.2: Organization of the analysis from the raw NOAA data in the upper-left corner to the final country-level mean estimates in the bottom-left corner. The legend in the upper-right denotes the primary language or datatype of each node.

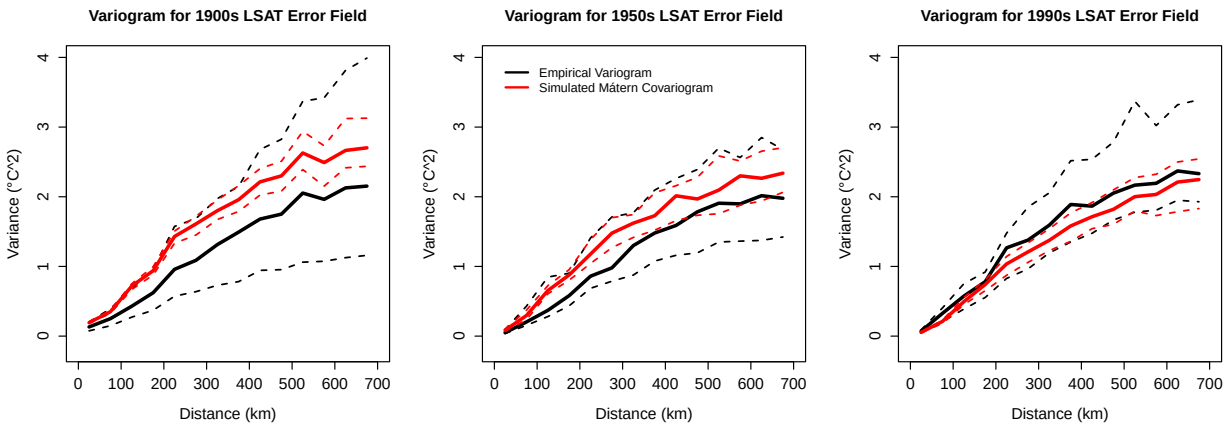

Figure 2.3: A summary of variograms from a random sample of 40 empirical ERA5 error fields and simulated variograms from 40 simulations from the heteroskedastic Matérn covariance structure with spatial locations corresponding to the LSAT coverage in that decade. The solid lines indicate the median of the 40 variograms and the dashed lines indicate the middle quartile.

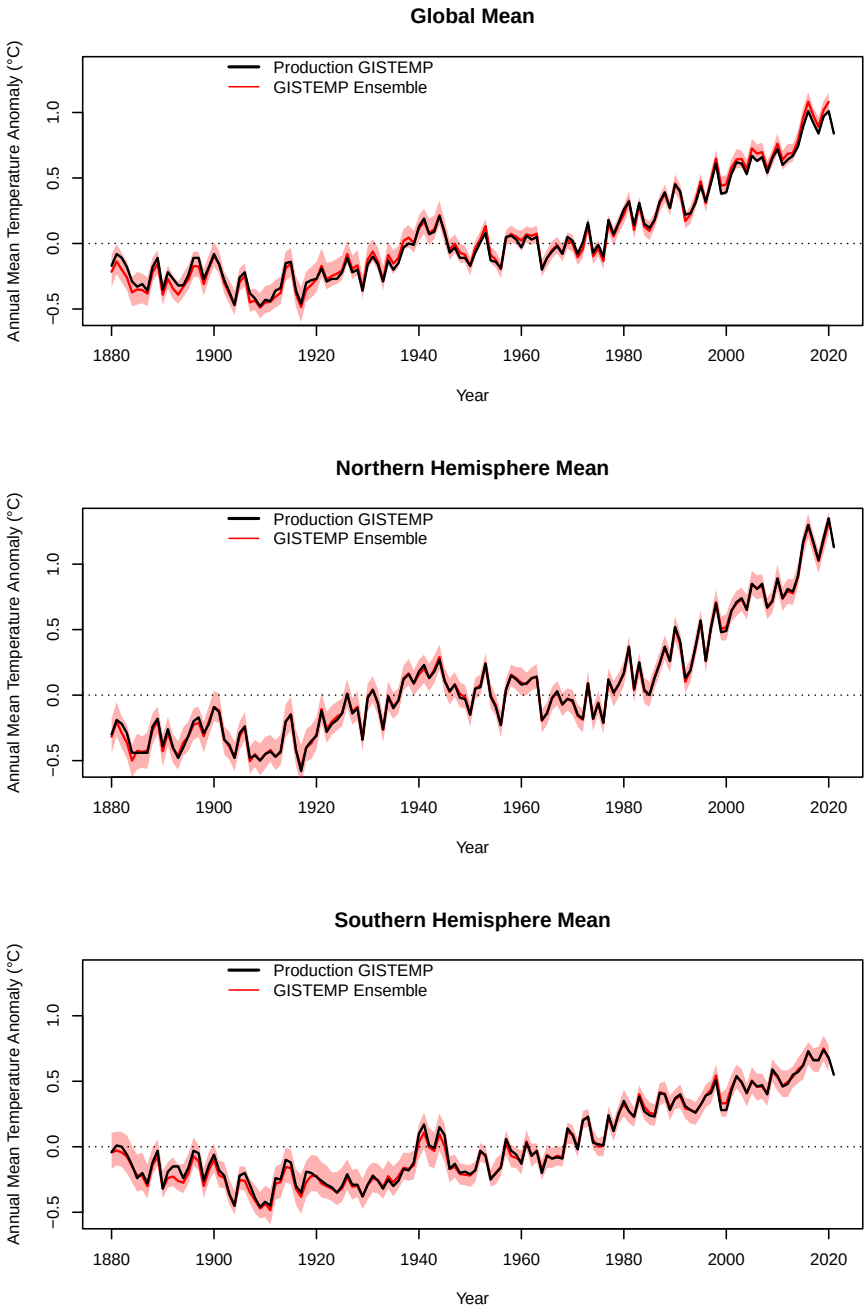

Figure 2.4: A comparison of the global and hemispheric annual mean series as calculated from operational gistemp and the GISTEMP ensemble. These uncertainty calculations can be used to update the graphs on the GISTEMP website.

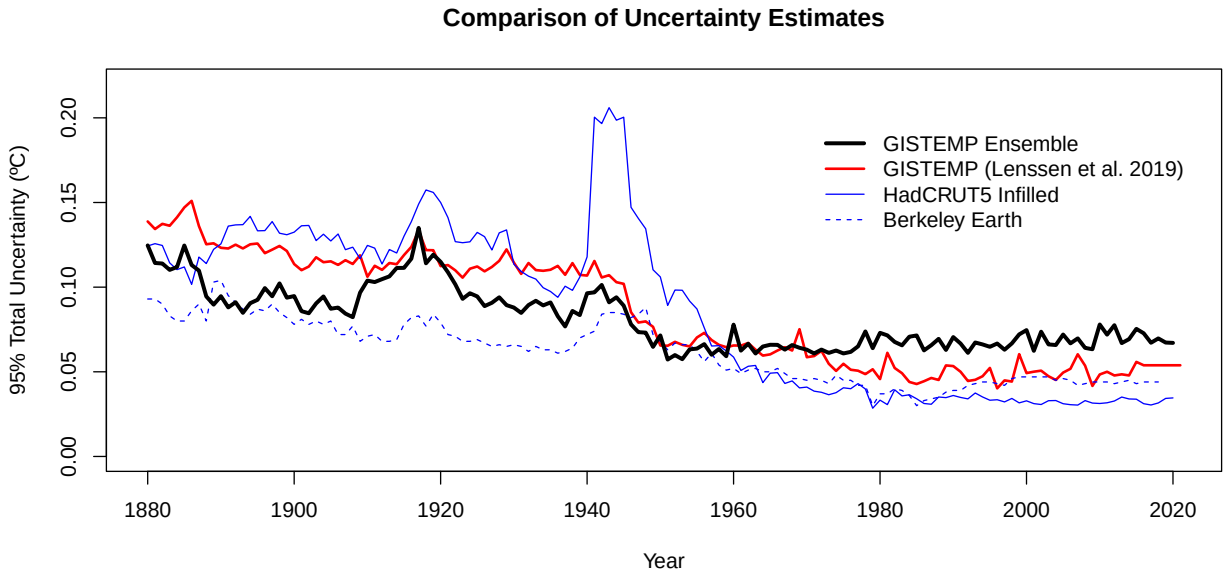

Figure 2.5: The global annual mean 95% confidence intervals for the new GISTEMP ensemble, the same calculation as performed in Lenssen *et al.* (2019), and the two products that publish operational confidence intervals.

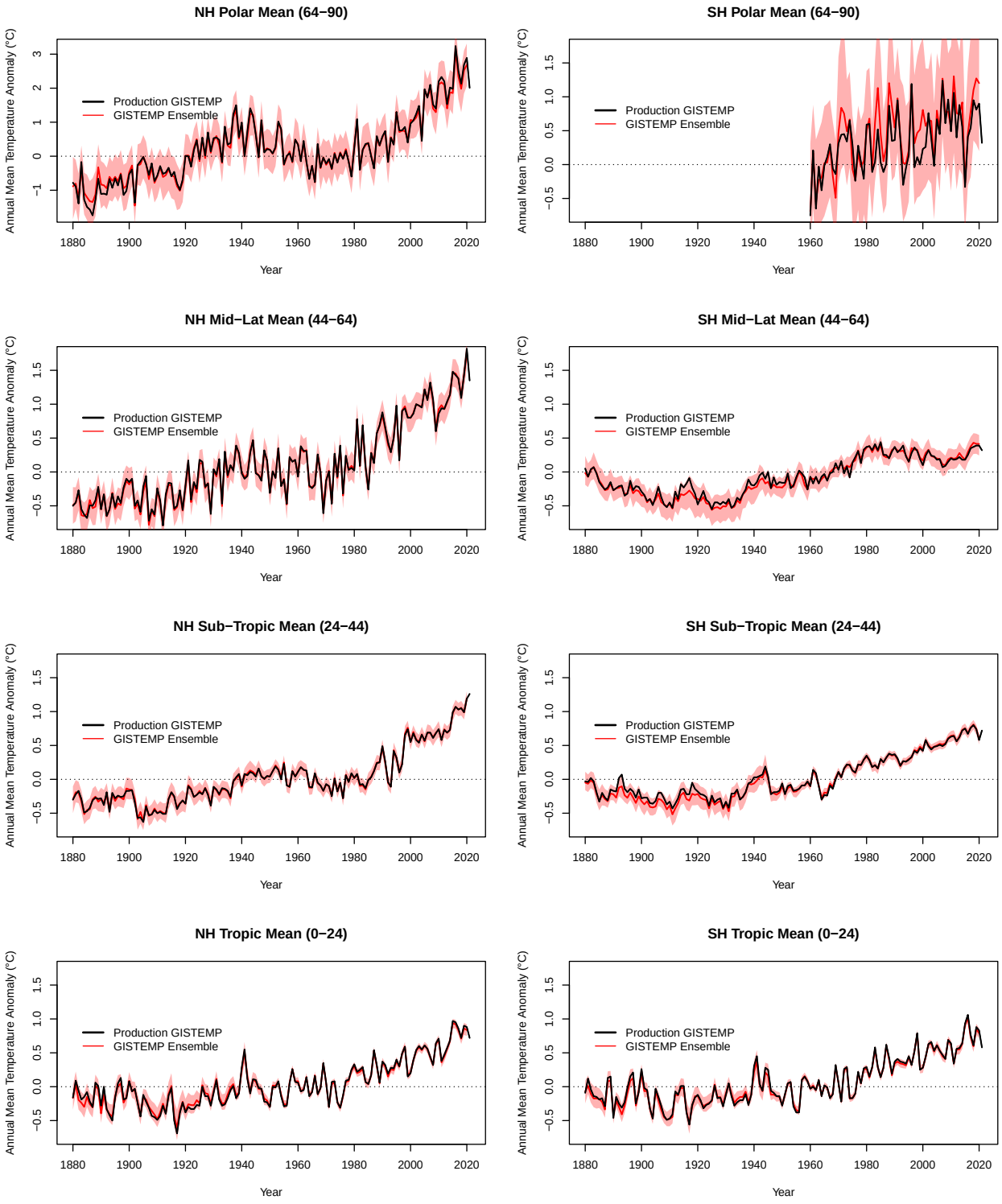

Figure 2.6: A comparison of the annual mean series from the 8 GISTEMP latitudinal bands as calculated from operational GISTEMP and the GISTEMP ensemble. These uncertainty calculations can be used to update the graphs on the GISTEMP website. Note the different y-scale on the top-left NH Polar plot.

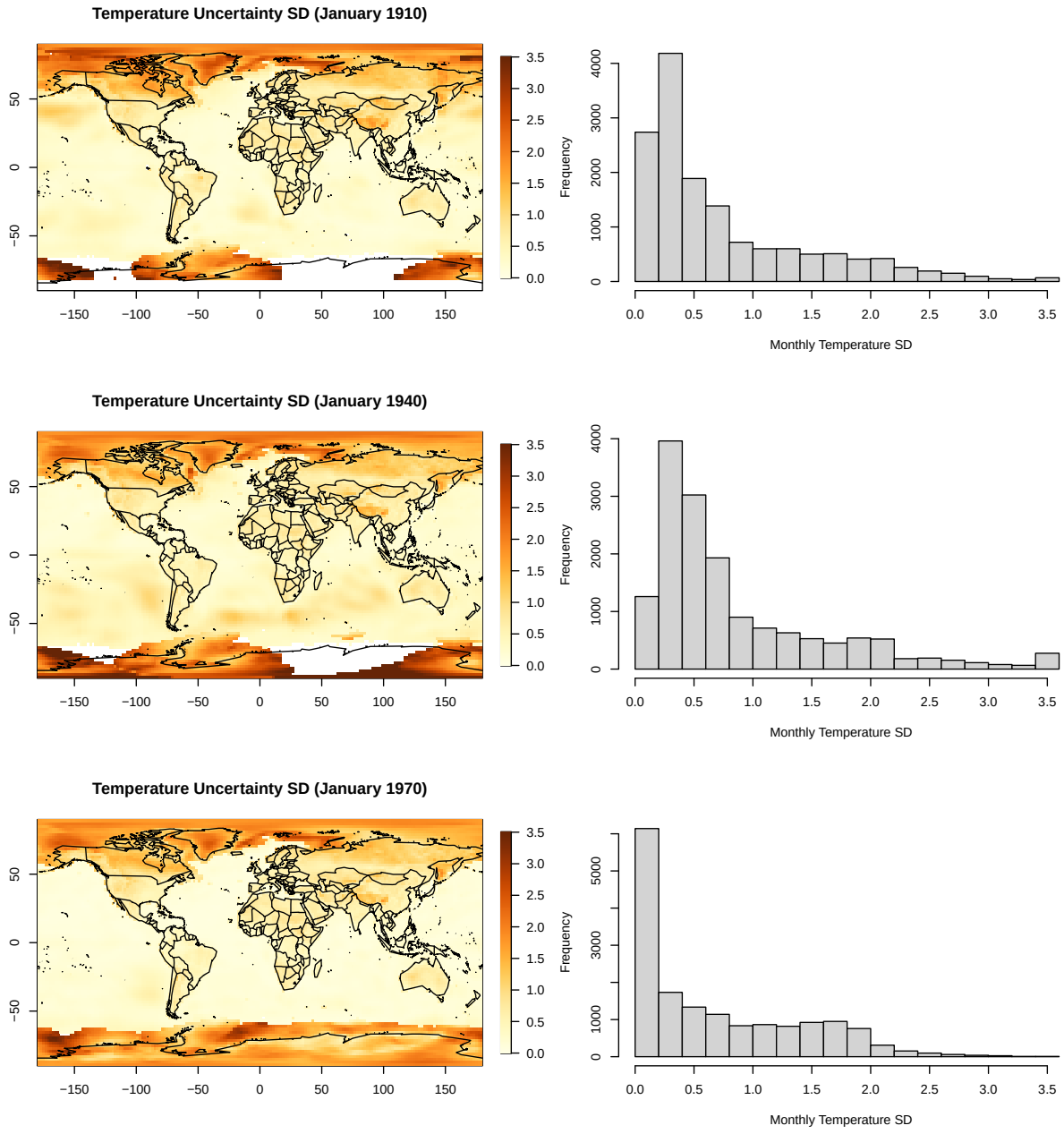

Figure 2.7: The standard deviation of the GISTEMP uncertainty ensemble for three monthly fields. The corresponding histogram to each field is shown to the right. The visualization has been capped at a standard deviation of 3.5 to avoid the very large antarctic uncertainty dominating the maps.

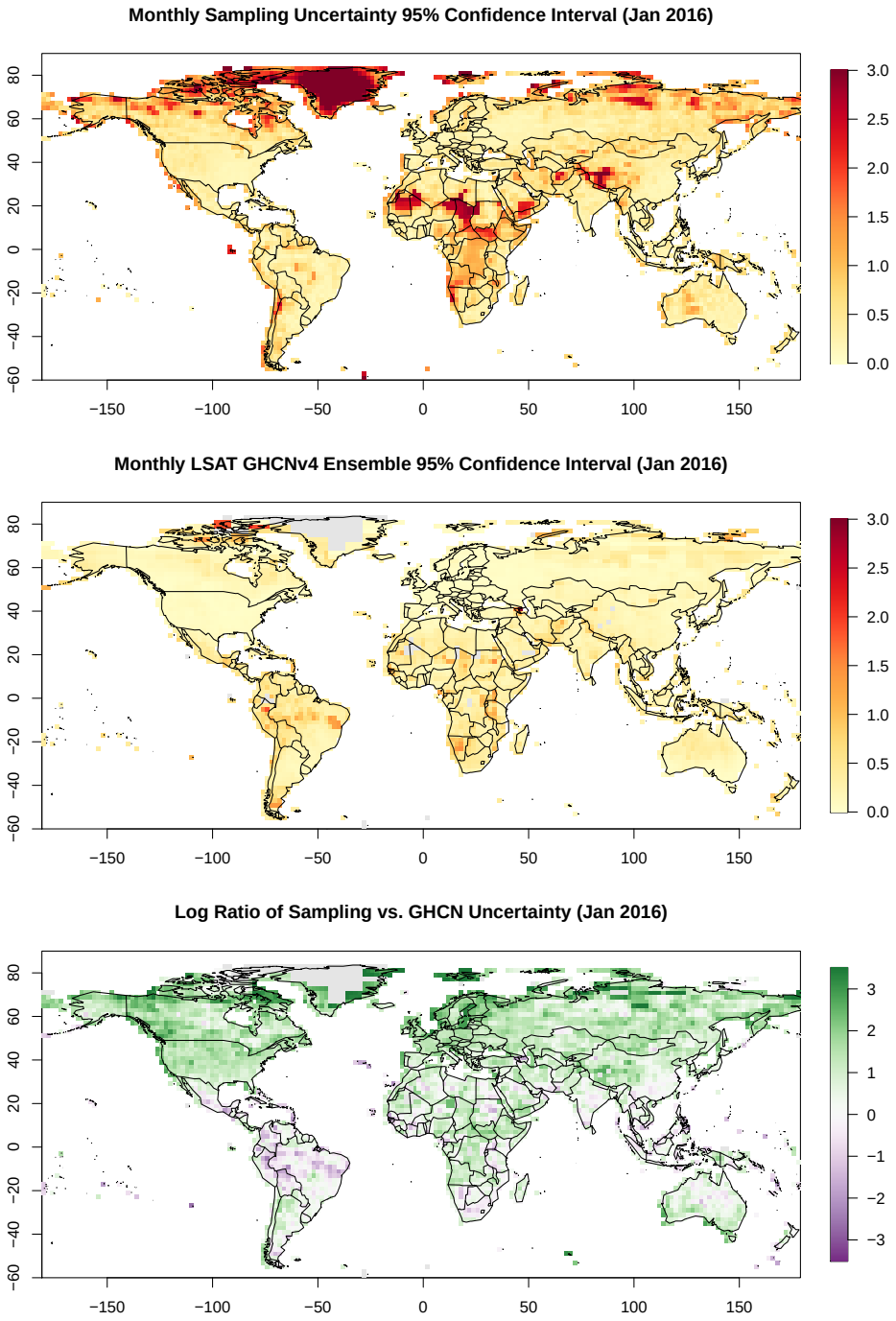

Figure 2.8: Global LSAT Uncertainty for the month of January 2016 decomposed into the contributions of (top) sampling uncertainty and (middle) bias and station uncertainty as quantified in the GHCN ensemble. (Bottom) the log ratio of sampling and GHCN uncertainty with green regions showing where sampling uncertainty dominates. Grey areas indicate regions where the GHCN uncertainty could not be estimated due to lack of coverage.

2.5 Application 1: Uncertainty in Country-Level Mean Series

Temperature products on latitude-longitude grids are often the most useful and easy to work with formats for climate and other science applications. However, applied climate science and the social and economic sciences often need climate information on maps defined by political boundaries. Here, the GISTEMP uncertainty ensemble is used to create a climate product with country-level information of the monthly and annual temperatures. Calculating country-level spatial means is a relatively easy exercise on a single gridded surface temperature record and serves as a good example of how an uncertainty ensemble can be utilized to account for observational uncertainty in subsequent studies. This country-level product is now provided operationally by NASA GISS to the UN Food and Agriculture Organization Corporate Statistical Database (FAOSTAT) for a map-room on global climate and agriculture conditions (FAO 2022).

2.5.1 Data and Methods

The method and code developed for the first deliverable of the FAOSTAT country-level temperature change project is used to calculate the country-level means to provide continuity with the existing product. The above method produces an ensemble of 500 equally likely reconstructions of the global temperature record. Each of these reconstructions is run through the country-level mean procedure, resulting in 500 potential temperature records for each country. Since this is an unwieldy amount of data to analyze and visualize, the uncertainty estimate for each country is provided as an empirical 95% confidence interval of this ensemble. While these confidence intervals are provided, all ensemble members are available and should be used in further analyses whenever possible to fully represent the spatiotemporal correlation structure of the uncertainty.

2.5.2 Results and Discussion

Calculation of the country-level mean using the GISTEMP LSAT ensemble from 1960–2016 finds similar results to the initial version of the country-level mean product produced for FAO-

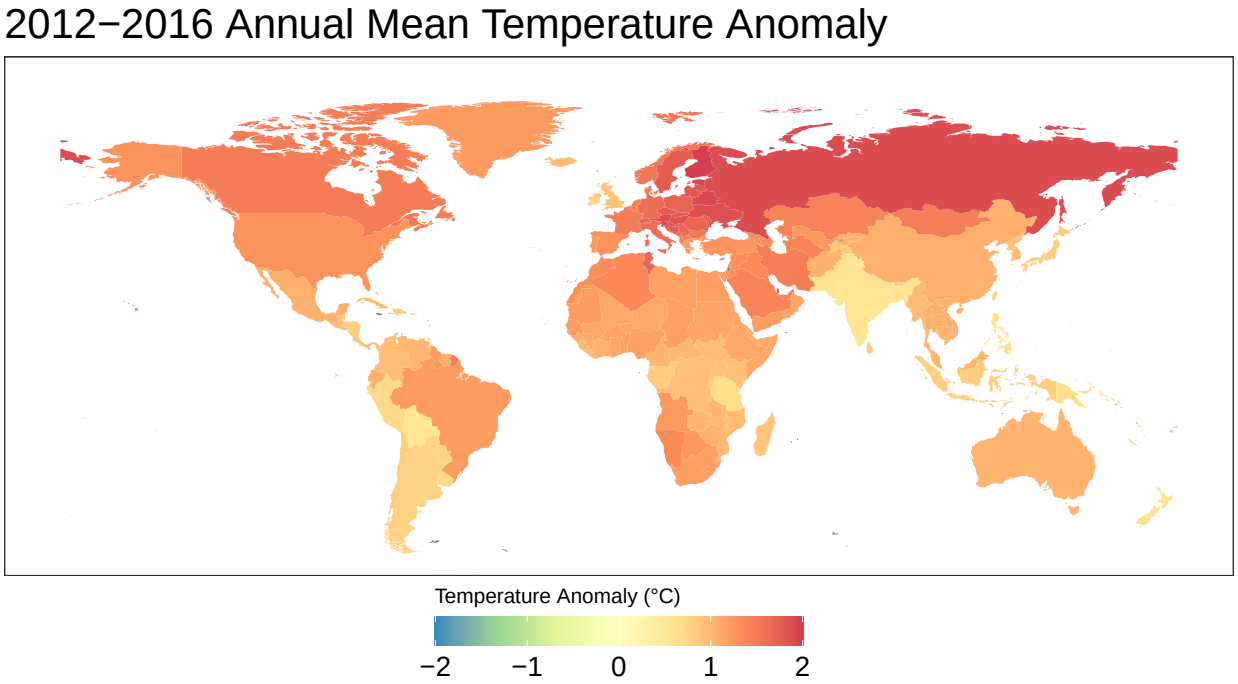

Figure 2.9: The country-level LSAT anomaly over 2012–2016, the most recent 5-year period in the GISTEMP-FAO ensemble.

STAT. Geographic patterns of warming (Figure 2.9) over 2012–2016 mirror the first generation of the country-level mean analysis as well as predicted regional temperature change following increased greenhouse gas emissions (Miller *et al.* 2014; Lenssen *et al.* 2019). In particular, the largest temperature anomalies occur in countries that have area in the Arctic, demonstrating the effects of Arctic amplification (Serreze & Barry 2011; Cohen *et al.* 2014).

Uncertainty in the country-level mean LSAT (Figure 2.10) is in agreement with studies on the density and reliability of the global observation network (Hansen *et al.* 2010; Menne *et al.* 2018). That is, there is low uncertainty, and high confidence, in country-level temperature in areas with dense station networks such as North America, Europe, and Australia. Greater uncertainty is found in regions with less dense and reliable networks such as parts of South America and Africa and the Middle East.

As expected, the largest uncertainties are found in polar Greenland and Antarctica, reflecting the very sparse temperature records in these regions. The choice of using GISTEMP with a 250 km

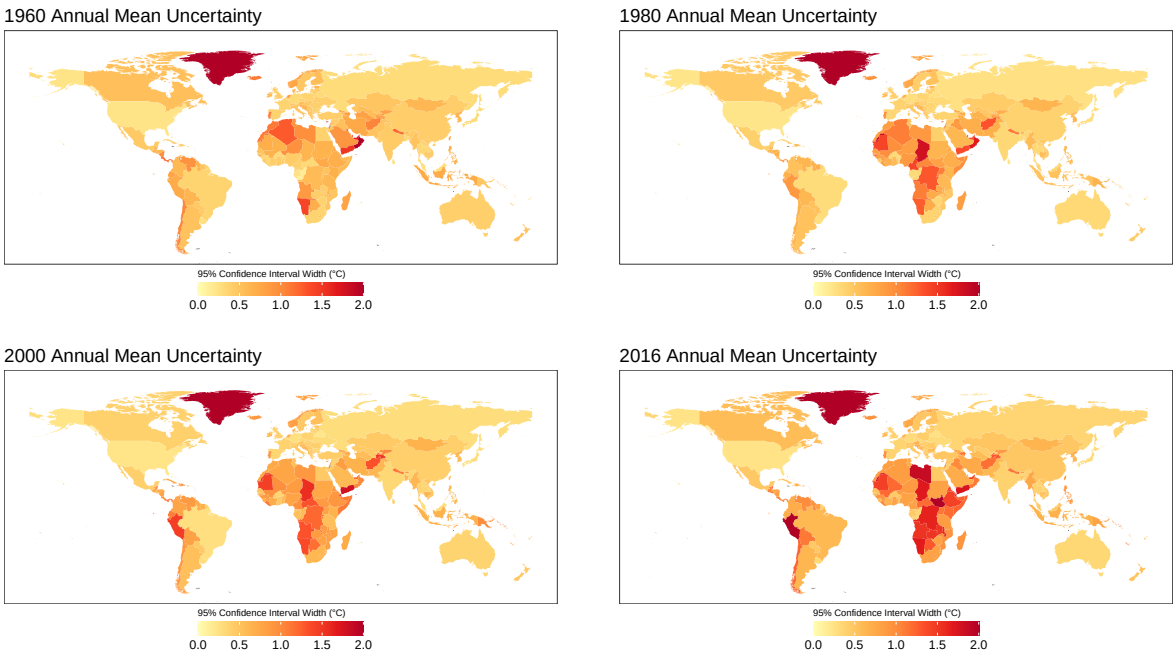

Figure 2.10: The total uncertainty in country-level annual LSAT for the years 1960, 1980, 2000, and 2016. The uncertainty is summarized by the empirical 95% confidence interval of the country-level annual means from the 500-member GISTEMP LSAT ensemble. The uncertainty for Greenland is greater than 2.0 °C with values varying between 2.2 °C and 2.6 °C.

interpolation, rather than the 1,200 km used for the calculation of the global mean, further increases uncertainty in the mean temperature of these two countries. However, the choice of using the 250 km interpolation is more than offset due to the much lower uncertainty in the vast majority of the country means.

Looking at the evolution of country-level uncertainty over the record (Figure 2.10), the spatial pattern remains consistent from 1960–2000. Uncertainty increases somewhat in 2016, the last year of the record. This is due to a decrease in global reporting of climate data in recent decades (Harris *et al.* 2020; Lenssen *et al.* 2020), which also increases uncertainty in the homogenization procedure (Menne *et al.* 2018). The uncertainty in 2000–2016 will likely decrease as more data is reported to international repositories as well as data past 2016 improving the performance of the homogenization method over this period.

When interpreting country-level confidence interval maps (Figure 2.10), it is important to consider the effect of country area. In general, larger countries will have lower uncertainty as averaging

over more area is equivalent to averaging over more data. The degree to which larger surface area decreases uncertainty is related to the spatial autocorrelation of the uncertainty. The result is that smaller countries that contain regions with high sampling uncertainty will not experience much of a reduction in uncertainty due to spatial averaging. Further reducing the effectiveness of spatial averaging, additional spatial structure is represented in the GHCN homogenization ensemble.

As an example, Italy and Ecuador have nearly identical land area at around 300,000 km^2, but have quite different annual mean LSAT uncertainty (Figure 2.11). Ecuador is among the more uncertain country-level means with a CI width growing to over 1 °C in around 2000 as a result of high station uncertainty. Italy has very low uncertainty, particularly for its size, with a CI width of around 0.45 °C for the entire record.

Comparing with Brazil and Australia at 8.5 million km^2 and 7.7 million km^2 respectively (Figure 2.11), two things stand out. First, the comparison between the smaller countries and these larger ones illustrates that uncertainty can only be reduced so far by spatial averaging over larger areas. Despite a high-quality station record and significantly greater land area, the uncertainty on the Australia annual mean temperature is still 0.3 °C, only slightly smaller than that of Italy. Second, the variation in the confidence interval width in Brazil from 0.25 °C to 0.6 °C shows that land area has limited ability to reduce uncertainty when there is underlying uncertainty in the station record. These periods of greater uncertainty in Brazil are due to station-level errors that are difficult to homogenize and introduce uncertainty into the estimates that cannot be corrected for by spatial averaging.

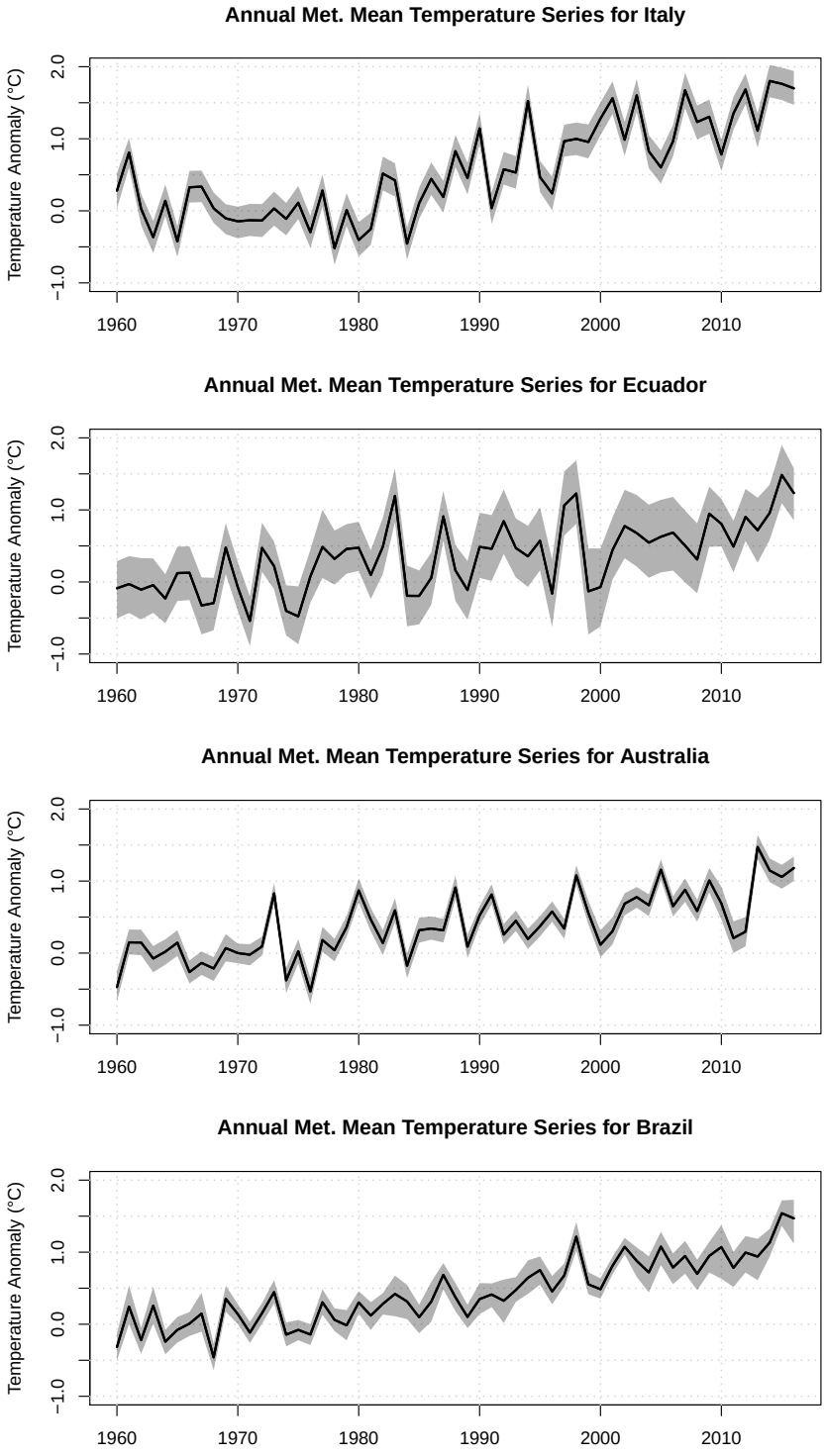

Figure 2.11: Annual meteorological (December-November) mean LSAT series for the approximately equal-area countries of Italy, Ecuador, Australia, and Brazil. The grey shading indicates the empirical 95% confidence interval and is not necessarily symmetric around the series.

2.6 Application 2: Relative Arctic Warming Rates

A commonly stated, but rarely cited statistic of climate change is that the Arctic is warming twice as fast the world (Cohen *et al.* 2014; Hope & Schaefer 2016; Dolgin 2017; Overland *et al.* 2019; Richter-Menge *et al.* 2020). However, back of the envelope calculations with operational GISTEMP suggests that the rate of relative Arctic warming, or Arctic Amplification (AA) ratio, is drastically underestimated, with the Arctic warming about 4 times faster than the global trend. The large discrepancy between the common knowledge statistic and this simple calculation prompts an investigation into the true current value of the AA statistic, as well as the effect of various methodological assumptions on the value of observed amplification.

In this study, 2 key assumptions for the calculation of the AA ratio as explored: the latitude range of the Arctic region and the statistical method used to quantify trends in historical temperature. The GISTEMP uncertainty ensemble is utilized to address these question while accounting for observational uncertainty. It is recommended to include observational uncertainty whenever possible, but it is essential to include observational uncertainty in studies involving Arctic temperature due to the large uncertainty in the region (Figure 2.6).

The first assumption addressed is the definition of the Arctic region. Traditionally, many studies have defined the Arctic as 60°N-90°N (Cohen *et al.* 2014). However, these is another definition of the Arctic region following the location of the Arctic Circle from 66.6°N-90°N. As the observed warming increases with latitude in the Arctic region, this difference in region is meaningful for calculating sensitive statistics of Arctic climate change such as the AA ratio. This study will consider the AA ratio for both 60°N and 66.6°N, investigating how this choice of region influences results.

A traditional method for assessing trends in climactic time series is linear regression over time periods of at least 30 years. This method is one of the two methods for summarizing historical global temperature change in the recent IPCC Working Group 1 report where they report observed linear trends over the past 40, 60, and 140 years (Table 2.4 in Gulev *et al.* 2021). Linear trends

have the advantage of being easy to implement and understand as well as providing an estimate of statistical uncertainty. However, the observed global temperature is not linear over the observed record and regional mean series such as the Arctic mean may be even less well-represented by linear trends (Figure 2.12). In light of this, the other method used by the IPCC is the difference in mean temperature between two time periods. Taking the different in means has the advantage of not assuming the functional form of the temperature series, but the disadvantage of requiring periods of sufficient length to average away internal variability and observational uncertainty before calculating the difference. As such, the IPCC adopts the difference in global temperature between 1850-1900 and three modern 20-year periods.

The disadvantages of both of these approaches are limiting when calculating the recent, or 1980-present, change in Arctic and global temperatures. The difference in means method is will dramatically underestimate the current warming as using even a 10 year period (from 2011-2020 for instance) will not capture the recent dramatic warming in the Arctic. The linear regression method is more appropriate, but the nonlinear nature of the mean Arctic temperature, particularly prior to 1990, means that a statistical model assuming linearity may misrepresent the rate of warming and thus the AA ratio.

In this study, the AA ratio as calculated by the ratio of linear trends is compared with an AA ratio as calculated by the difference in annual temperature after the series is smoothed using a Generalized Additive Model (GAM) (Wahba 1990; Wood 2006; Simpson 2018). By smoothing the annual mean series, the change in mean temperature can be robustly calculated as the difference between the first and last years of interest. Here, GAM's constructed with penalized regression splines are used as there is robust theory supporting the fitting of the critical smoothness parameter through cross validation approaches (Wahba 1990; Wood 2006). That is, the data is represented through a flexible class of polynomials where the "wigglyness" is determined by the data itself.

2.6.1 Data and Methods

Data Sources

In addition to using the 500 member GISTEMP uncertainty ensemble discussed in this chapter, the infilled and non-infilled HadCRUT5 uncertainty ensembles are used as points of comparison (Morice *et al.* 2020). Both of these ensembles account for the same sources of uncertainty as the GISTEMP ensemble. The non-infilled HadCRUT5 ensemble uses the method of HadCRUT4 and earlier versions of HadCRUT where the globe is divided into a $5° \times 5°$ grid and a grid-cell only has a value if it contains a station (Morice *et al.* 2012). The infilled version of HadCRUT5 uses Gaussian process regression to extrapolate station values, similarly to the method used in Berkeley Earth (Rohde *et al.* 2013a; Rohde & Hausfather 2020) and approximately equivalent to the interpolation method in GISTEMP (Lenssen *et al.* 2019).

For each of the three temperature products, global and Arctic means are calculated as the simple mean of grid-cells weighted by the cell area. Annual mean anomaly series are calculated for the global mean as well as the Arctic mean as defined by the mean of grid-cells between 60°N-90°N and 66.6°N-90°N. Grid-cells are appropriately weighted to account for the disagreement between the cutoff Arctic latitude and the native grid of the temperature products when necssary.

AA Ratio: Linear Regression

The first method used to quantify the AA uses linear regression to estimate the rate of change of global and Arctic warming. Given observed annual global/Arctic mean temperature series $y^{(g)}$ and $y^{(a)}$ and year x, the regression models are written as

$$y^{(g)} = \beta^{(g)} x + \varepsilon \tag{2.2}$$

$$y^{(a)} = \beta^{(a)} x + \varepsilon \, . \tag{2.3}$$

Then, the AA ratio is defined as the ratio of the slope parameters for the global temperature rate of change $\beta^{(g)}$ and $\beta^{(g)}$ or

$$\mathrm{AA}_{regression} = \frac{\beta^{(g)}}{\beta^{(g)}} \, . \tag{2.4}$$

Statistical uncertainty of the regression AA ratio is quantified through simulation. The slope are sampled following $\beta \sim \mathcal{N}(\hat{\beta}, \mathrm{SE}(\hat{\beta}))$ where SE stands for the standard error of prediction of the slope parameter. This process is repeated for each members of the observational uncertainty ensemble. Then, the final estimate is taken as the median and empirical 95% CI of the AA ratio statistic.

AA Ratio: Generalized Linear Models

In the GAM method, the annual temperature series are modeled with a single smooth function $f(x)$ where x is time from 1880-2020. Thus, the statistical models for the global and Arctic mean series are

$$y^{(g)} = f^{(g)}(x) + \varepsilon \tag{2.5}$$

$$y^{(a)} = f^{(a)}(x) + \varepsilon \, . \tag{2.6}$$

The form of f is determined by fitting a basis of polynomial functions and penalizing based on the square of the second derivate. This method is commonly referred to as a thin-plate spline or penalized regression of splines (Wood 2006). The smoothing parameter, or parameter controlling the strength of the penalty term, is determined using Generalized Cross Validation (GCV) (Wahba 1990; Wood 2006). The R package mgcv was used to fit all of the GAM functions for this study (Wood 2011). The statistical uncertainty of the GAM models is again represented through simulation where the entire GAM is simulated using the estimated covariance matrix.

The rate of change is then is estimated by taking the difference of the first and last year in the 30 year period from the GAM simulations and dividing by 30 to get a rate of change in deg/year while capturing the uncertainty in the statistical model. That is, for a 30 year period of GAM estimated

temperatures $\{\hat{y}_1^{(a)}, \ldots, \hat{y}_{30}^{(a)}\}$, the AA is written as

$$\mathrm{AA}_{GAM} = \frac{\hat{y}_{30}^{(a)} - \hat{y}_1^{(a)}}{\hat{y}_{30}^{(g)} - \hat{y}_1^{(g)}} \,. \tag{2.7}$$

This process is repeated for all members of the observational uncertainty ensemble. Then, the final estimate is taken as the median and empirical 95% CI of the AA ratio statistic.

2.6.2 Results and Discussion

It is helpful to first visualize the two statistical trend methods to understand why their estimates and uncertainty may deviate (Figure 2.12). The difference between the methods is most clear in the Arctic series plot. Looking at the period from 1987-2017, the linear regression method and estimate a rate of increase in Arctic mean temperature of 0.26 °C/decade where as the GAM method estimates a trend of 0.19 °C/decade better captures the non-linearity in the early portion of the time period. In addition, the uncertainty in the GAM is smaller at the early portion of the record as the GAM fit takes advantage of the full time series, rather than just relying on the 30-year period of interest. This property of the GAM fit uniformly results in lower uncertainties when estimating the global and Arctic trend compared to the linear regression method, leading to less uncertainty in the AA ratio (Figure 2.14).

Comparing the estimated AA ratio as a function of trend period motivates the use of longer periods when calculating trends (Figure 2.13). While the estimates remain relatively constant across trend period lengths from 10 to 30 years, the uncertainty in the AA ratio decreases substantially as the trend period lengthens. The increase in AA ratio uncertainty corresponding with a decrease in trend period is even more dramatic using linear regression method as expected (not shown).

The best estimates for AA ratio fall between 3 and 4, depending on the method and dataset used (Figure 2.14). As expected, the AA ratio is larger when defining the Arctic from 66.6°N. Also as expected, the estimates of AA are larger when using the linear regression method, but the uncertainty is also much greater than the GAM method. In general, the GISTEMP estimate falls

between the two HadCRUT5 estimates. This is somewhat surprising as it was initially expected that GISTEMP and the infilled HadCRUT5 would be in better agreement due to their similar interpolation of critical Arctic stations.

As shown in Figure 2.5 and discussed in Section 2.4, the uncertainty in the GISTEMP ensemble global mean is greater than HadCRUT5. This result extends to the AA ratio as the uncertainty in AA ratio as calculated with the GISTEMP uncertainty ensemble is the larger than that of both HadCRUT ensembles. This uncertainty difference in AA ratio is driven by the difference in observational uncertainty, particularly in the Arctic mean (Figure 2.15). Comparing the GISTEMP and infilled HadCRUT5 empirical 95% CIs, the GISTEMP global CI is around twice a large over the 1960-2020 period shown and the Arctic CI is between four and fives times as large suggesting that GISTEMP is overestimating the uncertainty and/or HadCRUT5 is underestimating it.

Returning to the motivation for this study for investigating the claim that the Arctic is warming twice as fast as the earth, it is now clear that this statistic is outdated and needs to be updated. Regardless of the method, data product, or definition of the Arctic, all but one of the AA 95% CIs are fully above 2 (Figure 2.14). Furthermore, when using the GAM method and defining the Arctic according to the Arctic circle at 66.6°N, the Arctic is warming 2.5-5 times faster than the planet. This common statistic should be updated in the discussion about global change as well as incorporated into studies on the impacts of Arctic warming. In addition, this studies serves as an important example for the inclusion of observational uncertainty through observational ensembles as well as using multiple products when possible. The robust results between the three products provides confidence in the answer, and the resulting confidence intervals enables reporting the AA ratio with proper uncertainty.

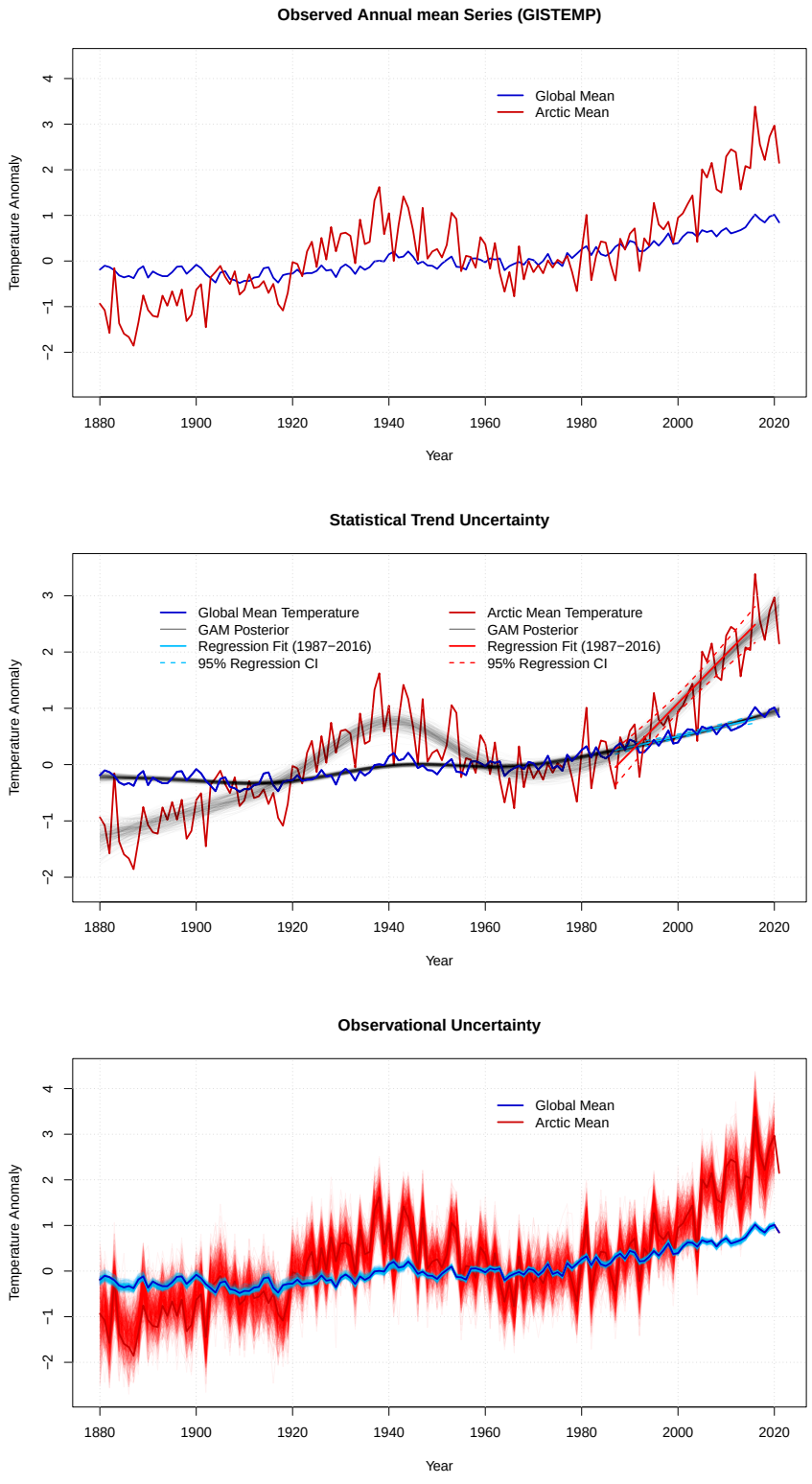

Figure 2.12: (Top) the GISTEMP operational annual mean global mean and Arctic (66.6°N-90°N) mean time series. (Middle) The linear regression and GAM fits to the global and Arctic mean series. (Bottom) the annual mean global mean and Arctic time series from each of the 500 uncertainty ensemble members.

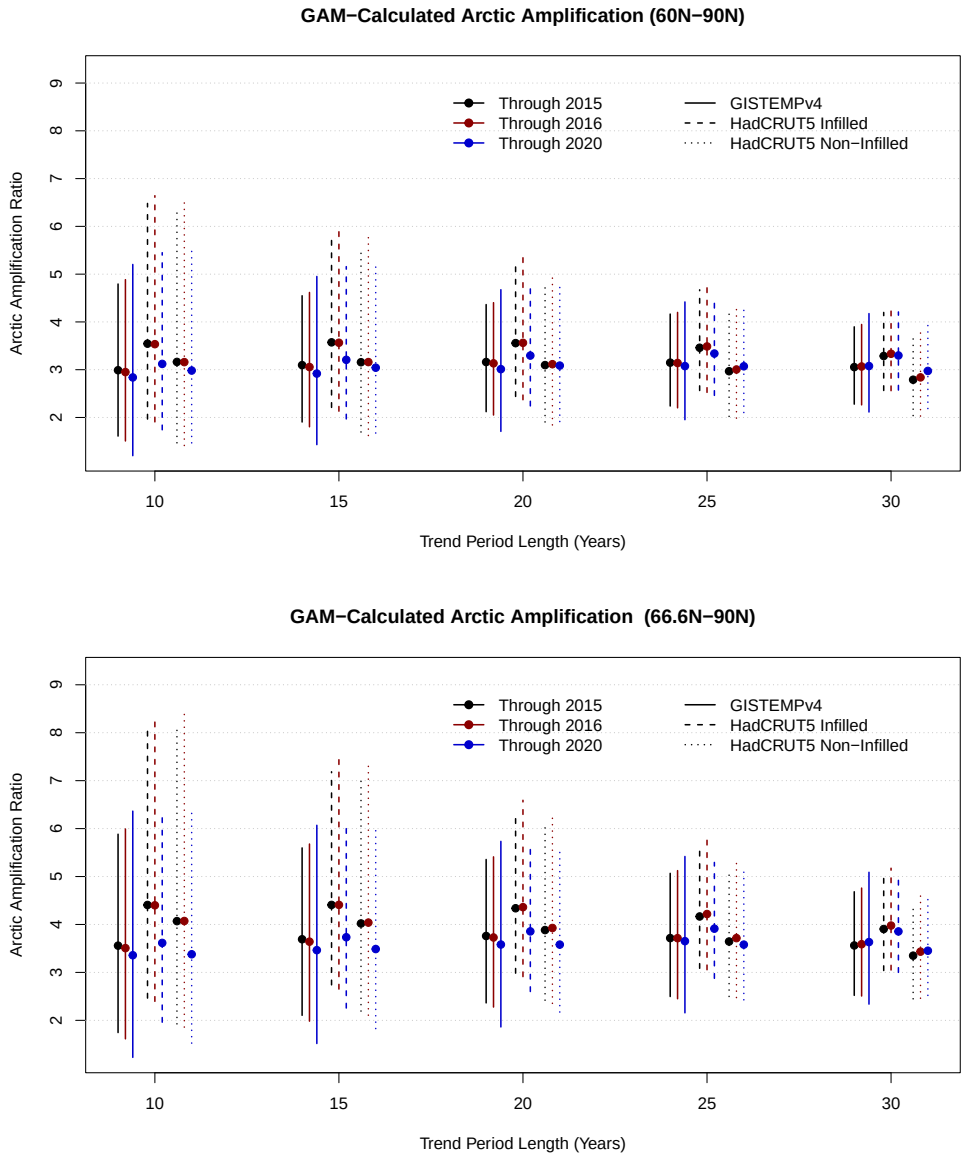

Figure 2.13: (Top) the GISTEMP operational annual mean global mean and Arctic (66.6°N-90°N) mean time series. (Bottom) the annual mean global mean and Arctic time series from each of the 500 uncertainty ensemble members.

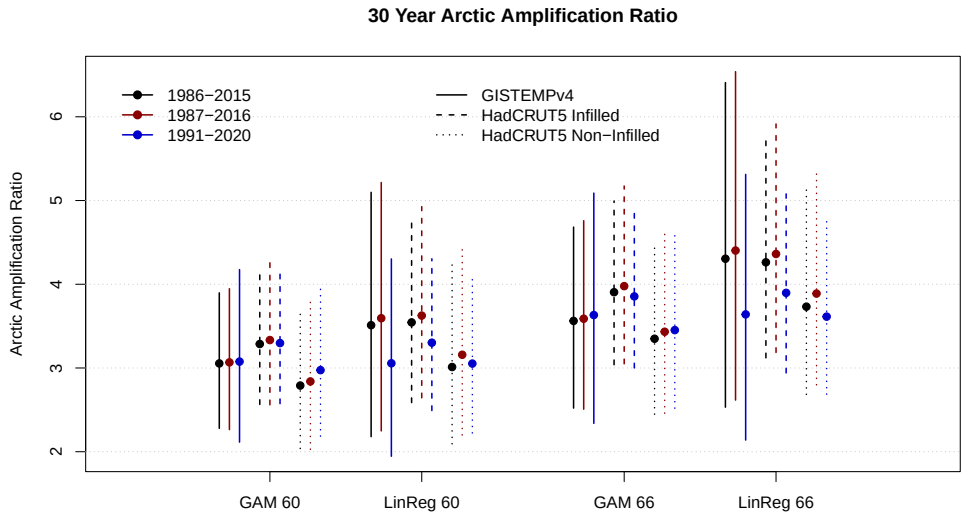

Figure 2.14: The AA ratio for each of the three products used for three 30 year periods: 1986-2015, 1987-2016, and 1991-2020. The methods for calculating the trend and defining the Arctic are shown by the groupings on the x-axis. The dot shows the ensemble median AA ratio and the whiskers show the empirical 95% confidence interval

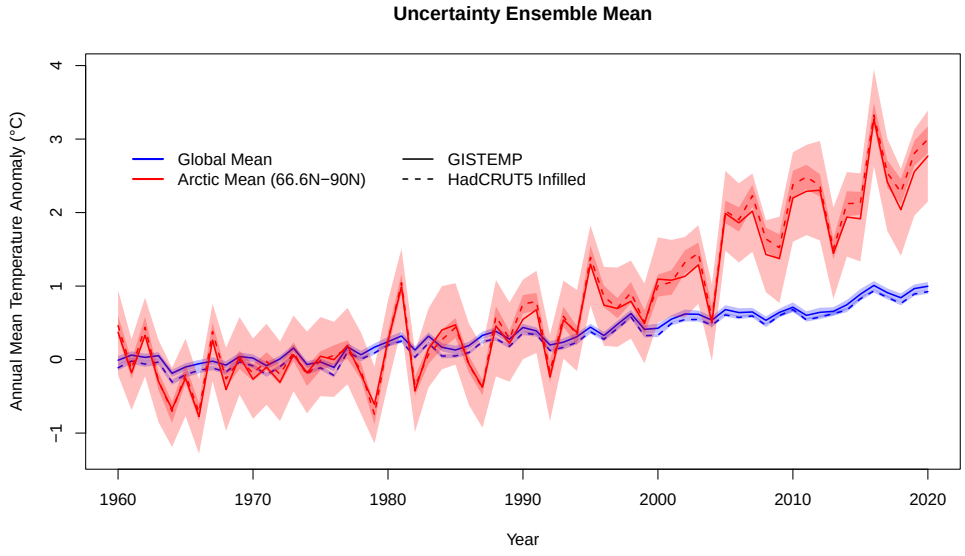

Figure 2.15: The global and Arctic (66.6°N-90°N) annual mean temperature anomaly series for GISTEMP and HadCRUT5 infilled. The plotted series are the ensemble mean series from the uncertainty ensemble for each product and the shading indicates the empirical 95% confidence interval.

2.7 Conclusions

In this chapter, an uncertainty ensemble for the GISTEMP temperature product has been presented presented and analyzed. Accounting for all sources of uncertainty at the monthly level increases enables inclusion of historical temperature uncertainty in future studies. The median estimates from the GISTEMP uncertainty ensemble agree very well with operational gistemp and the resulting global mean uncertainty agrees with the calculation of Lenssen *et al.* (2019). This work is a major step forward in the GISTEMP uncertainty model, enabling the inclusion of observational uncertainty in studies on historical global change. Two such studies are presented: the calculation of country-level means and an investigation into the relative rate of Arctic warming as defined by the AA ratio.

Uncertainty in the historical temperature record at the national level from 1960–2016 was quantified using the GISTEMP observational ensemble. This ensemble captures station, bias, and sampling uncertainties while properly accounting for their spatial and temporal correlation structures. The creation of this ensemble allows the calculation of country-level LSAT mean temperature and the corresponding uncertainty. The results were delivered to FAOSTAT as 95% empirical confidence intervals on the country-mean series and are available on the FAOSTAT map-room.

Revisiting the common claim that the Arctic is warming twice as fast as the globe using the GISTEMP as well as the HadCRUT5 uncertainty ensembles, shows that this statistic should be updated to, "The Arctic is warming 2.5-5 times faster than the globe." In this analysis, two statistical methods for quantifying trend were applied to the observational uncertainty ensembles continuing an important discussion of how best to quantify climactic trends.

It is the author's hope that the release of the GISTEMP uncertainty ensemble, alongside the already existing HadCRUT5 and NOAA GlobalTemp uncertainty ensembles, will prompt the community to incorporate observational uncertainty in studies whenever possible. The two examples presented here, and particularly the investigation into Arctic warming, show how large observational uncertainty can be, particularly when taking the means of smaller and less observed regions.

Uncertainty is particularly important for impacts as it can reveal worst-case scenarios that can be hidden in mean estimates. As this work continues, software for working with the uncertainty ensemble will be developed and made open source to ensure it is easy as possible to incorporate observational uncertainty in climate research.

Chapter 3: Seasonal Forecast Skill of ENSO Teleconnection Maps

This first-author work is published as Lenssen et al. *(2020) in Weather and Forecasting.*

The El Niño–Southern Oscillation (ENSO) is a major driver of precipitation variability worldwide (Ropelewski & Halpert 1987; Ropelewski & Halpert 1989; Mason & Goddard 2001). The robust precipitation teleconnections and associated societal impacts of ENSO make it the primary source of skill for seasonal forecasts (Livezey & Timofeyeva 2008; Barnston *et al.* 2010a). Immense work has gone into understanding the dynamics, variability, and predictability of ENSO (Yeh *et al.* 2018; Timmermann *et al.* 2018) to improve replication in Earth System Models (ESMs) critical to forecasts of climate variability and projections of climate change (Bellenger *et al.* 2014). Research on understanding, quantifying, and communicating seasonal forecasts and ENSO-driven precipitation variability reaches far beyond the physical sciences, with decision makers in fields such as agriculture (Rahman *et al.* 2016), water management (Crochemore *et al.* 2016), and public health (Borbor-Mendoza 2016) utilizing forecasts of seasonal precipitation.

The gold standard for seasonal forecasts is a dynamical forecast that has been post-processed to address systematic model biases, then tailored for specific users (Cash & Buizer 2005; Kumar *et al.* 2020). In addition, the forecast must be "translated" to users' applications effectively (Hansen *et al.* 2006). Effective tailoring and translation of seasonal forecasts is difficult and is a current focus of the World Meteorological Organization (Kumar *et al.* 2020). If the generation, translation, or communication of a forecasts fails, users do not have proper access to that forecast and must turn to an alternative forecast. One of the most available and interpretable alternative forecasts is a teleconnection map representing of ENSO impacts such as the "cartoon" map shown in Figure 3.1. Similar representations of ENSO impacts are also prominent on NOAA and WMO webpages.

The primary goal of this study is to quantify the skill of simple ENSO teleconnection forecasts

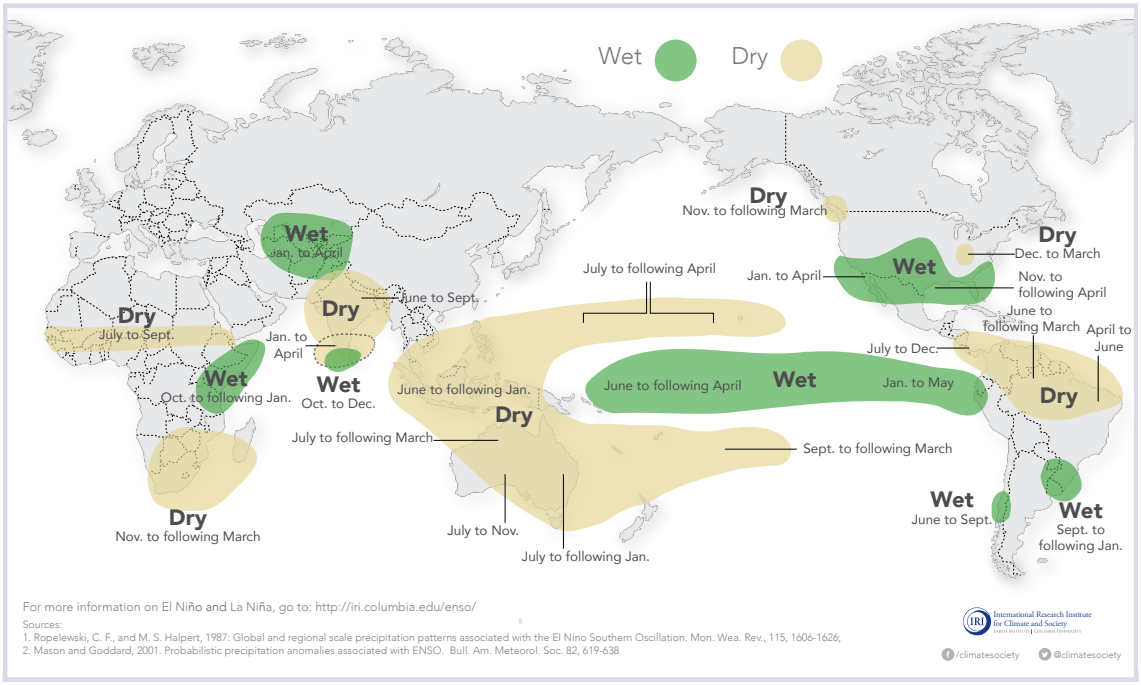

Figure 3.1: The "cartoon" El Niño teleconnections map issued by the IRI. Precipitation impacts are aggregated from Ropelewski & Halpert 1987 and Mason & Goddard 2001 and displayed in an easy to read format.

worldwide. The regions of the world with robust ENSO-precipitation teleconnections are detected with an update of Mason & Goddard (2001) (hereafter MG01). The skill of predicting regional precipitation from the deterministic and probabilistic teleconnection maps is quantified through the verification of these simple empirical seasonal forecasts. Generally, the goal of empirical forecasting is to maximize skill by including multiple climate indices and other predictors in a complex statistical model. The statistical forecast used in this study, hereafter referred to as an ENSO-based forecast (EBF), emulates the human use of ENSO teleconnection maps. The EBFs are made with simple statistical models that use the ENSO state to predict seasonal precipitation.

The simple ENSO-based forecasts also provide a benchmark for state-of-the-art forecasts. Seasonal forecast skill is quantified through comparison of some forecast attribute with a reference forecast. The reference is generally climatology: a forecast only containing information on the

long-term mean climate. However, the robust ENSO impacts on seasonal precipitation motivates using the simple ENSO-based forecasts as an alternative reference for seasonal forecasts. The use of more stringent alternative reference forecasts is not novel; the ENSO Climatology and Persistence (CLIPER) forecasting scheme of Knaff & Landsea (1997) uses simple statistical models as a physically-motivated reference for ENSO forecasting. Using an EBF as a reference for seasonal forecasting follows similar reasoning by defining skill as the value added over the empirical, historical impact of ENSO on precipitation.

In section 2, the historical precipitation, historical ENSO, and historical forecast data used in this study are outlined. In section 3, methodologies for the updated teleconnection maps of MG01 and the generation of EBFs are detailed. Section 4 presents results from the updated teleconnection maps and the EBF verification. Section 5 summarizes the results and provides some suggestions for future applications.

3.1 Data

3.1.1 Historical Precipitation

CRU TS 4.01

The Climatic Research Unit (CRU) TSv4.01 dataset is used to quantify historical ENSO precipitation impacts due to its long record with global coverage at high resolution, moderate interpolation, and its uncertainty quantification (Harris *et al.* 2014). Station data is quality controlled and interpolated onto a 0.5° grid using triangular linear interpolation. Full coverage of land grid cells is achieved by providing climatology values for grid-boxes with insufficient observational data. Data quality is reported through "number of contributing stations" fields and a grid-cell with two or fewer contributing stations is filled with the climatology value. The study is restricted to 1951-2016 due to data sparsity issues before 1950.

Properly accounting for grid-boxes that contain climatology-only values into their monthly time series is necessary for accurate ENSO impacts. Since the ENSO impacts are detected through

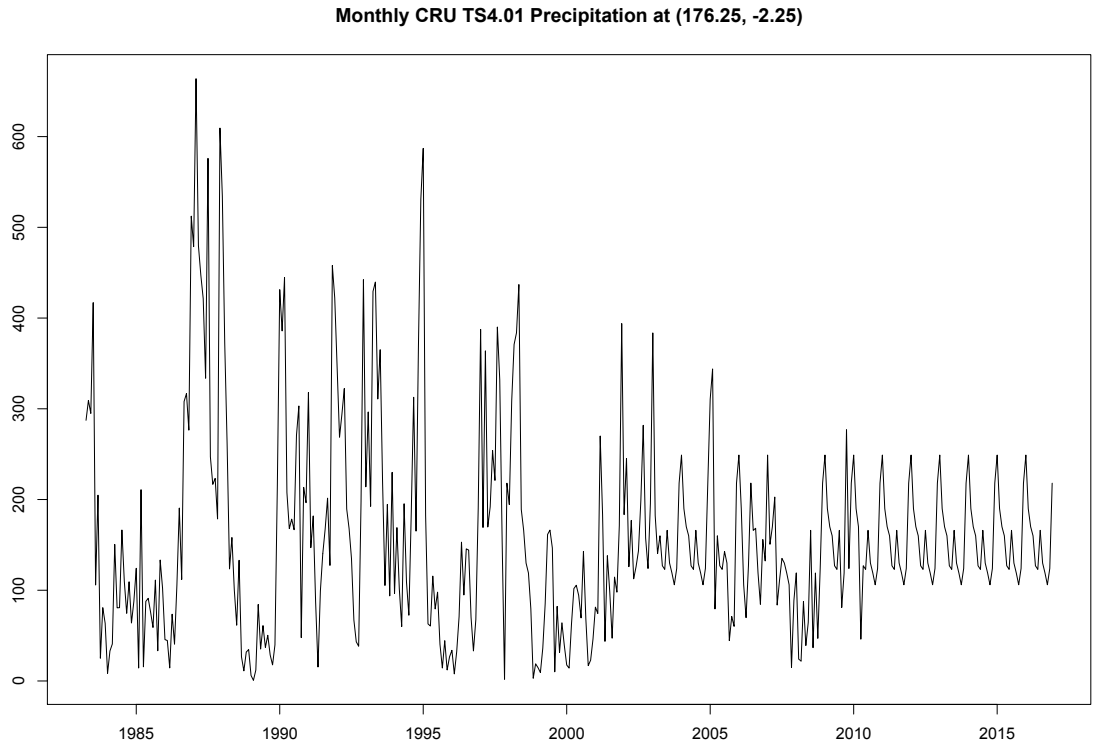

Figure 3.2: A single grid-cell time series in the maritime continent from CRU TS4.01. The change in variance is due to climatological precipitation being used when station records drop off in the last decades of the record.

anomalous wet or dry seasons, climatological values due to missing data will erroneously reduce the signal of ENSO on seasonal precipitation (Figure 3.2). Due to reduced reporting of the global observational record, the nearly stationary coverage CRU TS from 1951-1990 steadily decreases from 1990-present (Figure 3.3). The loss of coverage in recent decades is not uniform across the globe (Figure 3.3b); the losses are the greatest in the tropics where the precipitation signals from ENSO are most robust. The resulting probabilistic teleconnection maps should be viewed as a lower bound on the likelihood of the expected precipitation anomaly, particularly for the tropics region, as robust teleconnections may be hidden in the missing data. The recent reduced reporting along with other potential issues in data quality motivates future regional studies incorporating national data not reported to global sources.

The ENSO impacts analysis is performed on the native 0.5° grid as well as a 2.5° grid. The 2.5° resolution is chosen to match the resolution of the IRI seasonal forecasts. In addition, the

results from the 2.5° impacts analysis are used to confirm results from the native 0.5° resolution. A 2.5° grid-box has sufficient reporting station data if at least half of the contained 0.5° grid-boxes have sufficient reporting data, balancing reductions in signal due to climatological values while not over-aggressively masking.

CPC CMAP

The NOAA Climate Prediction Center (CPC) Merged Analysis of Precipitation analysis (CMAP) (Xie & Arkin 1997) is used to verify global seasonal forecasts over the 1997–2016 time frame. the CPC CMAP dataset incorporates remote sensed information on monthly precipitation allowing for greater global coverage over an era when station coverage is declining. It is also chosen to provide continuity with previous verification studies of the IRI seasonal forecasts such as Barnston *et al.* (2010a). Repeating the verification with the GPCP merged analysis (Adler *et al.* 2003) results in no major changes to the conclusions.

3.1.2 Seasonal Niño 3.4 SST Index

The state of ENSO at a seasonal timescale is represented by the NOAA CPC Oceanic Niño Index (ONI). The ONI is the seasonal mean of the monthly Niño 3.4 index. Again following the CPC, the monthly Niño 3.4 index is the monthly mean sea-surface temperature anomaly in the central equatorial Pacific (5N-5S, 170W-120W) calculated using the Extended Reconstructed Sea Surface Temperature version 5 (ERSSTv5) dataset (Huang *et al.* 2017).

3.1.3 Historical Forecasts

IRI Seasonal Forecasts

The IRI history of seasonal precipitation forecasts from 1997–2016 serves as an example state-of-the-art seasonal forecast. These seasonal forecasts were first issued October 1997 and were produced quarterly until August 2001, when the frequency of issuance increased to monthly. The forecasts are issued globally over land on a 2.5° grid with leads up to four months.

By using historical forecasts rather than hindcasts, this forecast data captures the evolution of seasonal forecast development over the time period. Prior to 2017, the core of the IRI forecast system was a primarily dynamical two-tiered model in which SST fields were predicted with a combination of dynamical and statistical models followed by the atmospheric response as simulated with dynamical atmospheric general circulation models (AGCMs) (Goddard *et al.* 2003; Barnston *et al.* 2010a). Final forecasts are determined after statistical post-processing of the multi-model AGCM ensembles and minor subjective modifications to reduce noise and known regional issues in the dynamical models and post-processing methods (Goddard *et al.* 2003; Barnston *et al.* 2010a).

Since 2017, the IRI's seasonal climate forecast system begins with raw output from NOAA's North American Multi-Model Ensemble Project (NMME) (Kirtman *et al.* 2014). After removal of mean, lead-time dependent biases, each model is corrected for systematic biases and calibrated using extended logistic regression (Wilks 2009). The calibrated forecasts from individual models are combined with equal weight into one multi-model forecast.

IRI ENSO Forecasts

The IRI began to issue realtime monthly ENSO forecasts in March 2002, incorporating the many dynamical and statistical ENSO forecasts run at climate centers around the world. These ENSO forecasts became a joint effort with NOAA's Climate Prediction Center (CPC) in late 2011. The forecasts are issued as probabilities of El Niño, Neutral, and La Niña conditions with lead times up to nine months. The forecasts are objective probabilities from a simple counting of the models in each ENSO category and are available from June 2003–2016.

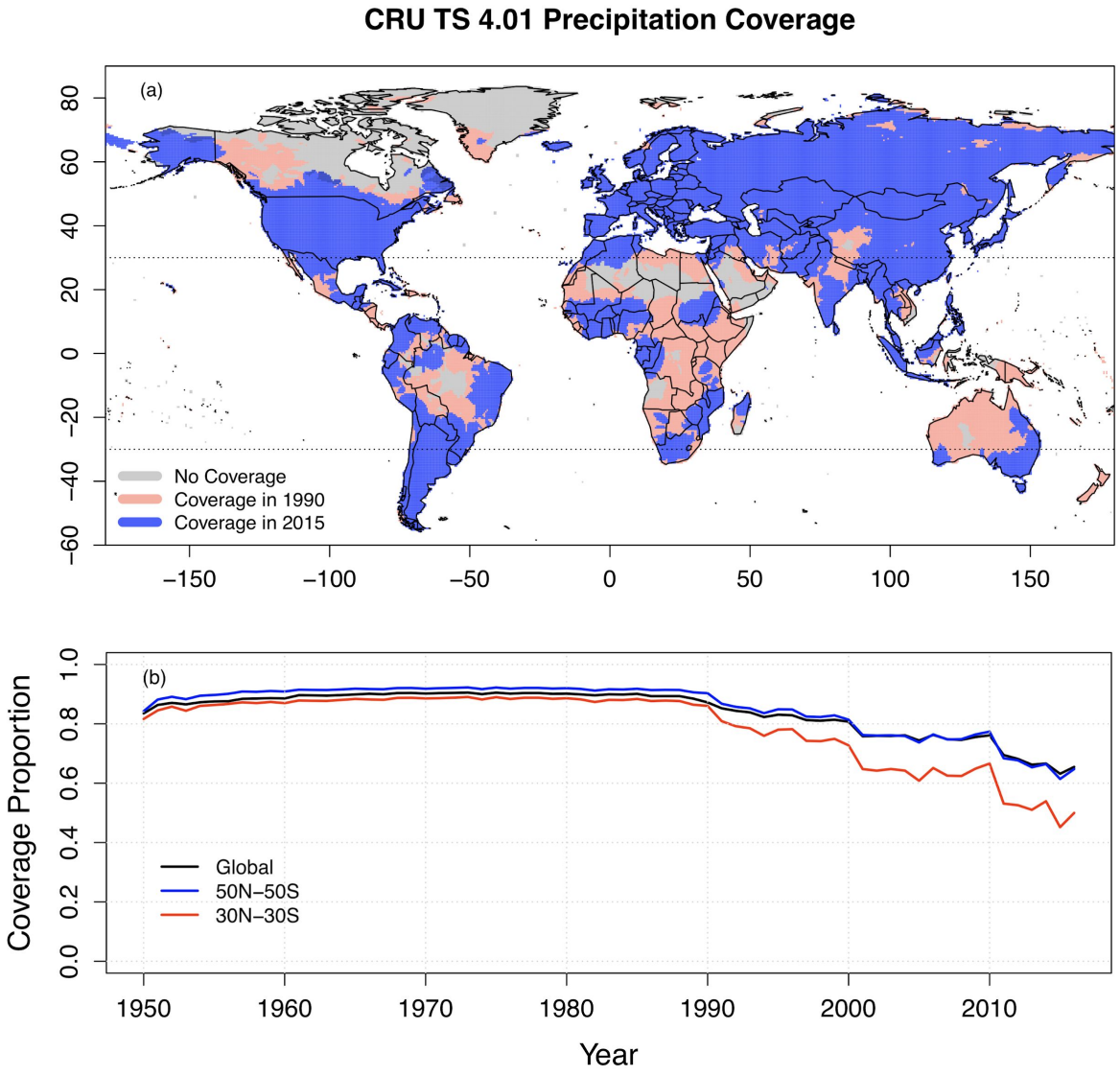

Figure 3.3: (a) The spatial distribution of coverage in the CRU TS 4.01 dataset visualized through the instantaneous coverage in 1990 when CRU has approximately maximum coverage and 2015 which is reflective of the present day coverage. (b) The time evolution of coverage in CRU TS 4.01 from 1950-2016.

3.2 Methods

3.2.1 Global ENSO Precipitation Impacts Methods

Teleconnection maps are constructed as in MG01; these maps indicate the local frequency of occurrence of categorized precipitation. Maps are computed for each season and both ENSO phases. Locations with a locally statistically significant response according to a hypergeometric test (modified to account for purely climatological data in the CRU dataset) are identified.

The method contains three steps; changes and updates to the MG01 method are italicized:

1. *Select all ENSO events since 1950 where the ONI exceeds an absolute anomaly of 0.5 degrees for at least 5 consecutive months and has a maximum absolute seasonal anomaly of at least 1 degree.*

2. Calculate the empirical conditional distribution of precipitation over each land surface grid point given the ENSO state according to the (below-normal, near-normal, above-normal) terciles *excluding years with climatology-only data*.

3. Determine statistical significance of above or below-normal signal according to a hypergeometric test with the sample sizes determined by the number of El Niño or La Niña events and total years *with sufficient data*.

Step 1 of the updated analysis differs from MG01 by defining ENSO events according to the ONI rather than taking the top 8 or 11 events. Providing an absolute definition of a moderate or stronger ENSO event, rather than a relative one as in the case of MG01, allows continuous updates as events that were included in previous versions will always be included. Selecting only ENSO events with maximum absolute anomalies of at least 1 degree excludes weak ENSO events which have reduced skill in seasonal precipitation forecasts due to corresponding weaker atmospheric responses (Goddard & Dilley 2005).

The changes to steps 2 and 3 are to exclude climatology-only values in the analysis by only using time points with sufficient station data. For each grid box independently, time points are

removed from the series where CRU TS does not have sufficient reporting station data. Masking for data sparsity results in intra-grid-box differences in the number of years on record and the number of El Niño and La Niña events. The differences in sample sizes and ENSO events are inherently taken into account in the p-values of the hypergeometric test. However, a small p-value does not necessarily imply a large effect, rather some combination of a large effect and a large sample size.

A p-value is determined using the hypergeometric test with the null hypothesis that each precipitation tercile is equally likely to occur regardless of the ENSO state. The corresponding alternative hypothesis is that ENSO does change the distribution of seasonal precipitation for a given location. As described in detail in MG01, this problem can be posed as a test for independence on a 3×2 contingency table (Agresti 1990). The corresponding null distribution is a hypergeometric distribution with tail probabilities calculated using Fisher's exact test (Fisher 1935; Agresti 1990; Lehmann & Romano 2005).

The counts of observed precipitation terciles for each ENSO phase cannot be easily compared for different locations because of the different sample sizes resulting from the masking of climatological values. Thus, hypergeometric tests are performed on each grid-box series to calculate a p-value. The resulting impact maps show historical empirical probabilities that have been masked for robust impacts, providing a measure of the effect of ENSO on precipitation while still accounting for statistical significance. Direct visualization of the p-values is avoided as a low p-value can be an indication of a large effect size and/or a large sample size (Sullivan & Feinn 2012).

Dry season areas are defined similarly to MG01. A location is determined "dry" for a given season if either (a) the climatology for that season is less than 15% of the annual total and greater than 50 mm or (b) the lower tercile for that season is less than 10 mm of rain. The absolute cutoff in criteria (b) is increased from 0 mm in MG01 to 10 mm to reduce artifacts in the dry mask arising from the interpolation in CRU TS.

3.2.2 ENSO-Based Forecast (EBF) Models

Three versions of empirical seasonal precipitation forecast models are developed with statistical models trained solely on the historical ENSO impacts (Table 3.1). Two known-ENSO EBF models, one deterministic and one probabilistic, assume that the state of ENSO is known. The probabilistic forecast-ENSO EBF model accounts for uncertainty in the seasonal forecast of precipitation due to limited predictability of the ENSO state.

The three EBF models represent increasing complexity in how a decision maker might factor known ENSO teleconnections into planning or preparedness during an ENSO event. As such, the skill of forecasts issued with these EBF models represent the skill of forecasts made with ENSO impact maps. Hindcasts made with the three EBF models are compared with the IRI seasonal forecast to quantify the value added by state-of-the-art seasonal forecasting and identify possible areas of improvement.

Broadly speaking, seasonal climate forecasts contain two sources of uncertainty arising from the underlying dynamical systems: uncertainty from predicting the global SST pattern and from uncertainty predicting the atmospheric response given SST patterns. Two known-ENSO EBF models assume perfect information of the ENSO state and effectively ignore the uncertainty in SST prediction. These serve as an upper bound in forecast skill of the ENSO teleconnection map or cartoon. The forecast-ENSO EBF model uses the historical ENSO forecasts to represent the uncertainty arising from prediction of the SST state. Note that this framework does not account for the uncertainty in seasonal forecasts arising due to model biases and post-processing.

Following a widely used format for seasonal forecasts, the EBF models issue probabilistic forecasts with the forecasted likelihood of each climatological tercile (above-normal, near-normal, and below-normal)=(AN, NN, BN) represented as probabilities that sum to one. An important special case is the climatological forecast of $(1/3, 1/3, 1/3)$.

Hindcasts are issued with each of the three EBF configurations over the period 1997-2016. This is the period of the IRI forecast, the example historical state-of-the-art forecast used for this study. The three EBF models described in detail below are based on precipitation anomaly maps

Table 3.1: Summary of the three ENSO-based Forecast (EBF) methods.

	Probabilistic?	Known ENSO State?
Deterministic Known-ENSO EBF	No	Yes
Probabilistic Known-ENSO EBF	Yes	Yes
Probabilistic Forecast-ENSO EBF	Yes	No

calculated from the out-of-sample time period of 1951–1996. When used in realtime, the EBF forecast models would be trained on the full historical record. While the precipitation impacts analysis was done on both the 0.5° and 2.5° grids, the forecasts in this study use the 2.5° global grid to allow direct comparison with the IRI historical forecasts.

Deterministic Known-ENSO EBF Model

The deterministic known-ENSO EBF model assumes perfect information of the ENSO state. Given an El Niño or La Niña, the deterministic known-ENSO EBF model issues a forecast with 100% above-normal or below-normal probability in regions with a robust impact. A climatology forecast is issued in grid-boxes without statistically significant impacts and globally for ENSO-neutral seasons. While a deterministic forecast is drastically overconfident for seasonal precipitation forecasting, it represents how decision makers may interpret and act upon ENSO teleconnection maps; during an ENSO event, the precipitation impacts are viewed as an expectation – as a certain forecast. This forecast is innately overconfident and expected to verify poorly, particularly when using probabilistic verification methods that measure reliability.

Probabilistic Known-ENSO EBF Model

Increasing in complexity, the probabilistic known-ENSO EBF model uses the same criteria for issuing a non-climatological forecast as the deterministic known-ENSO EBF, but instead issues the historical probability of observing each precipitation tercile. For a location and season with robust impacts, the forecast is the empirical probabilities of the three terciles from out-of-sample historical record. As with the deterministic known-ENSO EBF model, non-climatological forecasts are

only issued during active ENSO events. Terciles with zero empirical probability are given nominal probability ($\sim 2\%$) to avoid issuing zero probability forecasts. As before, a climatology forecast is issued globally during ENSO-neutral conditions and for any location that does not have a significant impact. The probabilistic known-ENSO EBF model provides a more realistic representation of the uncertainty in predicting atmospheric responses to ENSO and is expected to outperform the deterministic forecast on any probabilistic verification method that takes reliability into account.

Probabilistic Forecast-ENSO EBF Model

The probabilistic forecast-ENSO EBF model accounts for uncertainty in predicting the ENSO state in addition to the uncertainty in the atmospheric response. Historical IRI ENSO forecasts are used to quantify the limited predictability of ENSO in the hindcast study. The probabilistic forecast-ENSO EBF model issues forecasts as the weighted average of the El Niño, Neutral, and La Niña probabilistic known-ENSO EBF forecasts for a given season with weights set by the ENSO forecast. For example, given a probabilistic forecast of the ENSO state, the issued probability of AN precipitation is calculated as

$$\begin{aligned} P(AN) = {} & P(\text{El Niño}) \cdot P(AN|\text{El Niño}) + \\ & P(\text{Neutral}) \cdot P(AN|\text{Neutral}) + \\ & P(\text{La Niña}) \cdot P(AN|\text{La Niña}), \end{aligned} \tag{3.1}$$

where the first term in each line of the right hand side is the forecast probability of the ENSO state and the second term is the historical probability of AN precipitation under each ENSO state. The forecast probabilities for NN and BN precipitation are calculated similarly. The forecast-ENSO EBF model issue identical forecasts to the probabilistic Known-ENSO EBF model during ENSO events as the probability for El Niño or La Niña is then 100%.

Probabilistic forecast-ENSO EBF hindcasts are issued from mid 2004–2016 as the historical IRI ENSO forecast is first available in 2004. The hindcasts are issued at leads of 1–4 months,

Table 3.2: Descriptions of the forecast attributes referenced in the study.

Attribute	Question Answered	Score
Reliability	Does the outcome occur as frequently as forecasted on average?	Reliability Score
Resolution	Does the outcome differ given different forecasts?	Resolution Score
Discrimination	Do the forecasts distinguish higher categories from lower?	GROC
Skill	How does the forecast perform relative to a reference forecast?	RPSS

which permits investigation into the effect of lead time on skill of the different methods. The probabilistic forecast-ENSO EBF model is the most promising candidate for benchmarking state-of-the-art seasonal forecasting systems as it better represents the uncertainty in seasonal forecasts.

3.2.3 Forecast Verification Methods

The quality of the EBFs and IRI forecasts is quantified through the metrics of resolution, reliability, and discrimination (Table 3.2). See the Appendix for further details on the forecast verification scores used in the study.

The analysis is divided into three sections. First, the performance of the two known-ENSO EBFs and the IRI forecast are compared. All skill calculations in this portion of the study use climatology as the reference forecast, following the current practice in seasonal forecast verification (Jolliffe & Stephenson 2012; Mason 2018). The climatological forecast for tercile forecasts is a probability of $\frac{1}{3}$ issued to each of the precipitation terciles. The deterministic known-ENSO EBF represents realtime forecasts that could be, and often are, issued with the use of "cartoon" teleconnection maps such as Figure 3.1. The probabilistic known-ENSO EBF emulates a forecast issued using probabilistic teleconnection maps such as Figure 3.4. By verifying the known-ENSO EBFs, the skill of these very simple seasonal predictions is quantified.

Second, the skill of the IRI forecast is calculated relative to reference forecasts made with the probabilistic known-ENSO EBF model in addition to the traditional climatology reference. The skill of the IRI forecast relative to the EBF reference provides a measure of the value added by a calibrated MME forecast over ENSO teleconnection maps.

Third, skill as a function of lead time is calculated for the IRI forecast model and probabilistic forecast-ENSO EBF model. Understanding how skill falls off with lead time in the EBF model provides another important baseline metric for state-of-the-art seasonal forecast systems.

Both Brier- (Brier 1950) and Ignorance-based (Roulston & Smith 2002) scores were used to verify the ENSO-based and IRI forecasts. Since the results from the Brier- and Ignorance-based verifications qualitatively agree, only the Brier-based results are presented for three reasons. First, the Brier-based scores remain proper when averaged over both time and space and the spatial and temporal averages commute. Second, the resolution–reliability decomposition of the Brier score extends naturally to the tercile setting (Epstein 1969; Murphy 1971). The Ignorance-based score decomposition requires the non-local Ranked Ignorance Score, removing one of the supposed advantages of working with Ignorance based scores (Weijs *et al.* 2010; Tödter & Ahrens 2012). Finally, the Ignorance score of an incorrect deterministic forecast is infinite. While this feature can be argued to be appropriate, infinite values make comparison of the deterministic known-ENSO EBF and the other forecasts impossible.

3.3 Results

3.3.1 Global ENSO Teleconnection Maps

The global ENSO impacts analysis presented in section 3.23.2.1 is used to generate global maps of robust ENSO-related precipitation impacts for each season, ENSO state, and tercile category with empirical probabilities demonstrating the historical effect of ENSO in an intuitive way. Higher historical probability indicates more consistent, and therefore predictable precipitation anomalies given a non-neutral ENSO state. In addition, using probabilities to describe the effect of ENSO on precipitation motivates probabilistic precipitation forecasts according to historical impacts. As all maps are not able to be included here, they have been uploaded to the IRI data library[1] where they can be visualized online and downloaded in a variety of formats.

[1]Maps for all seasons and ENSO states can be found at `http://iridl.ldeo.columbia.edu/home/.lenssen/.ensoTeleconnections/`

As an example, a pair of maps showing the DJF precipitation anomalies during La Niña years are shown in Figure 3.4. Many established ENSO impacts are reproduced such as increased precipitation in Southern Africa Van Heerden *et al.* (1988), Northern South America (Ropelewski & Halpert 1987), Northwestern North America (Ropelewski & Halpert 1986), and Southeastern Australia (Ropelewski & Halpert 1987), and decreased precipitation is seen in Eastern Brazil (Grimm *et al.* 1998).

This study reproduces all of the teleconnections discussed in previous global teleconnection studies (Ropelewski & Halpert 1987; Mason & Goddard 2001). An additional 20 years of data with multiple strong ENSO events along with the improvements in methodology leads to greater statistical power and better estimates of the empirical probability of anomalous seasonal precipitation compared to MG01.

Several additional teleconnections are found during El Niño periods. During DJF (Figure online[2]), a much larger above-normal precipitation signal than MG01 is shown across the southern USA, Caribbean, and Mexico (Horel & Wallace 1981; Cayan *et al.* 1999), and a above-normal teleconnection is found in Lake Victoria region in equatorial eastern Africa. Below-normal precipitation is found in the El Niño DJF season in the southern tropical Andes region stretching from southern Peru through Bolivia into northern Chile and Argentina (Vuille *et al.* 2000; Garreaud & Aceituno 2001; Sulca *et al.* 2018) as well as the northern USA and Canada in agreement with the above-normal signal to the south (Horel & Wallace 1981; Cayan *et al.* 1999). For MAM, the above-normal anomalies in the southern USA, Caribbean, and Mexico (Horel & Wallace 1981; Cayan *et al.* 1999) and below-normal anomalies in the southern tropical Andes (Vuille *et al.* 2000; Garreaud & Aceituno 2001; Sulca *et al.* 2018) found in DJF are shown to persist into the spring. In addition, an above-normal anomaly is found in Peru in agreement with Sulca *et al.* (2018). For JJA, below-normal anomalies are found in southern Mexico into Central America (Magana *et al.* 2003). During SON, a greater spatial extent of below-normal precipitation is now detected throughout northern South America (Ropelewski & Halpert 1987; Grimm 2003).

[2]Maps for all seasons and ENSO states can be found at `http://iridl.ldeo.columbia.edu/home/.lenssen/.ensoTeleconnections/`

The updated study found fewer additional teleconnections beyond MG01 during La Niña periods in agreement with the fewer robust La Niña teleconnections already identified. Additional teleconnections discovered during DJF (Figure 3.4) are below-normal anomalies found across the southern USA, Caribbean, and Mexico (Cayan *et al.* 1999). For MAM, widespread below-normal anomalies stretch between eastern Afghanistan and Pakistan to the Caspian Sea (Barlow *et al.* 2002). For SON, no additional teleconnections from MG01 were found.

During JJA a coherent above-normal anomaly associated with El Niño is found centered in France. While a significant impact of ENSO on boreal winter and spring precipitation over Europe has been demonstrated (Brönnimann *et al.* 2007; Yeh *et al.* 2018), no work could be found documenting a possible summer teleconnection. Above-normal precipitation conditions occur frequently during La Niña JJA stretching from Nepal through Bangladesh. These signals motivate further analyses with regional and national data sources to verify the teleconnections in observations as well as dynamical studies of the possible underlying mechanisms.

The broad agreement of the teleconnections found with regional studies support the updated global teleconnection maps. These maps provide empirical estimates of regional teleconnections using consistent methodology across the entire land surface as opposed to piecing together disparate regional studies. The maps provide a starting place studies incorporating regional data sources as well as more sophisticated statistical and dynamical methodologies.

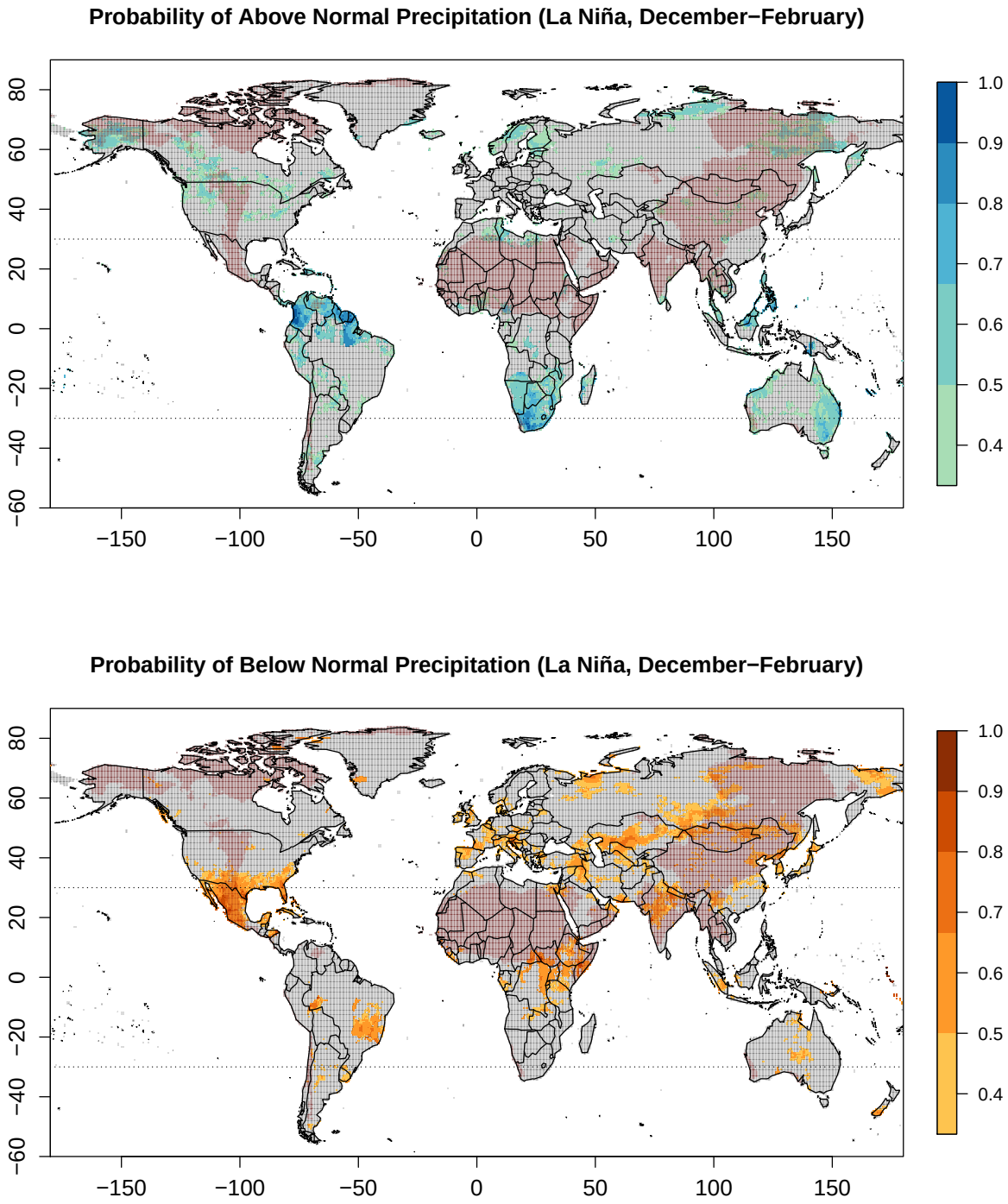

Figure 3.4: The empirical probability (from 1951-2016) of observing (a) above-normal and (b) below-normal seasonal anomalies in DJF during La Niña events. Areas considered "dry" are masked in light red and areas without a significant signal at the $\alpha = 0.10$ significance level are masked in gray. Maps for all 12 seasons and both ENSO states are available at `http://iridl.ldeo.columbia.edu/home/.lenssen/.ensoTeleconnections/`.

3.3.2 Assessment of Known-ENSO EBFs

The study's goal of quantifying the skill of simple ENSO teleconnection forecasts globally is addressed through verification of hindcasts from the EBF methods. Since the EBF methods utilize robust teleconnections, one would expect them to be skillful. As shown by the teleconnection maps, seasonal precipitation responds stochastically to ENSO. Comparing deterministic and probabilistic forecasts based on ENSO teleconnections illustrates the value of including uncertainty in the forecast method. Comparing the statistical EBFs with the IRI forecast, a forecast system primarily based on dynamical models, shows the added value of a state-of-the-art forecast and highlights geographic regions or ENSO events that need further investigation.

Comparison of the IRI forecast with the known-ENSO EBFs across various forecast attributes (Figures 3.5–3.11) indicates that the IRI forecast performs best, followed by the probabilistic known-ENSO EBF, and the deterministic known-ENSO EBF has the least skill. Additionally, the improvement in skill over time of a historical realtime forecast captures the evolution of the forecast system, both in the development of dynamical models and calibration/combination approaches. The general increase in IRI forecast skill over time illustrates these improvements, particularly during the larger ENSO events. The three most extreme events during the study period were the 1997–1998 and 2015–2016 El Niños and the 2010–2012 La Niña. Each metric presented shows the increased performance of the IRI forecast relative to the probabilistic known-ENSO EBF as time progresses. While the improvement is a qualitative result as it is made from a sample size of three for a highly variable system, it is encouraging to observe that the IRI forecast has captured each major ENSO event better than the previous one while the known EBF skill remains relatively constant.

Looking more specifically at the forecast attributes, the mean resolution, which measures the dependence of the outcome on the forecast (Murphy 1973), is calculated for the tropics (30N–30S) over time (Figure 3.5a). The corresponding spatial patterns of annual mean resolution for the IRI forecast and the known-ENSO EBFs (Figure 3.6) are similar with highest resolution in regions with strong ENSO teleconnections. The greatest resolution occurs in the maritime continent re-

gion where ENSO modulates the climate most directly. The IRI forecast generally has greater resolution than the EBFs in teleconnection regions, echoing the tropical mean time series. However, the known-ENSO EBFs exhibit comparable or greater resolution over northwest Colombia, the Namibia region in southwest Africa, and eastern Australia suggesting that some ENSO teleconnections may be inadequately represented in some or all of the dynamical models that contribute to the IRI MME forecast or that there may be low-quality observations. In addition, the EBF forecast exhibits some, albeit low, resolution throughout the extratropics with skill extending into subpolar regions in the northern hemisphere. These results are echoed in the spatial distribution of the discrimination (Figure 3.7) suggesting need for a more detailed investigation into the teleconnections as well as the cause of extratropical skill of the EBFs.

The discrimination (Figures 3.5b and 3.7) quantifies the dependence of the forecast on the outcome (Murphy 1991). The area under the Generalized Relative Operating Characteristics (GROC) curve (Mason & Weigel 2009) of the EBFs are very similar when they issue non-climatological forecasts since the GROC rewards forecasts that put the highest forecast probabilities on the category that is ultimately observed regardless of the probability issued. The IRI forecast and EBF have similar spatial patterns of skill with the IRI discrimination generally higher. The qualitative agreement between the resolution and discrimination in both the spatial and temporal averages suggests that the Brier-based scores capture the information content of the forecasts reasonably well.

The reliability time series (Figure 3.8) measures the ability of each forecast to represent the probability of outcomes Murphy (1973). The deterministic forecast has very poor and highly variable reliability due to the inherent uncertainty in predicting variations in seasonal precipitation. Reliability diagrams (Figure 3.9) are examined for mean and conditional forecast bias. Reflective of the bias-correction, the IRI exhibits very small conditional bias for each of the three terciles. The EBFs are over-confident for all terciles. That is, they systematically issue higher forecasts probability than the observed relative frequencies.

The combined reliability and resolution skill of the forecasts with respect to climatology is

quantified by the Ranked Probability Skill Scores (RPSS) (Figures 3.10 and 3.11). The deterministic known-ENSO EBF performs poorly in RPSS primarily due to its poor reliability. The ranking of forecast systems observed in resolution and discrimination scores holds with the IRI as the most skillful on average, and at most time points. The mean RPSS fields (Figures 3.11a,b) closely mirror the spatial pattern seen in the resolution (Figure 3.6) reflecting that the majority of the spatial variability in RPSS arises from the spatial variability in resolution.

The resolution-reliability decomposition and skill calculation was also performed with ignorance-based scores for the IRI and probabilistic EBFs with no substantial change to the results. The resolution, discrimination, and RPSS field calculations were performed seasonally in addition to the annual values presented and indicated that increased skill generally aligns with a region's rainy season, in agreement with Barnston *et al.* (2010a).

3.3.3 Climatology vs. ENSO Reference Forecasts

The added skill provided by IRI forecast over the probabilistic known-ENSO EBF forecast is calculated with the RPSS using the EBF as the reference forecast instead of the usual climatological forecast. The spatial pattern of this RPSS (Figure 3.11c) roughly mirrors that of the RPSS with a climatology reference forecast (Figure 3.11a) suggesting that dynamical models in the IRI forecast are adding additional skill in the majority of ENSO teleconnection regions. The widespread positive skill illustrates the value added by IRI forecast over the EBF for the majority of the world. However, a few teleconnection regions show negative skill indicating that the EBF forecast has higher skill over the study period. The most striking regions of negative IRI forecast skill relative to the EBF are the monsoon region of western India, southwestern Africa, and eastern Australia. These negative skill regions are also found in the rainy season RPSS in addition to the annual RPSS presented in Figure 3.11c further motivating closer investigation.

The global RPSS skill of the IRI forecast is slightly higher when the ENSO-based reference forecast is used as the baseline (Figure 3.12a), but the tropics skill decreases substantially (Figure 3.12b). The additional value of the IRI forecast over the EBF is summarized by the positive global

skill under the EBF reference during the majority of nearly every ENSO event in both the global and tropics means. Under the ENSO-based reference forecast, the total tropics skill decreases nearly 50% where the large decrease in reflects the greater presence of robust teleconnection in the tropics.

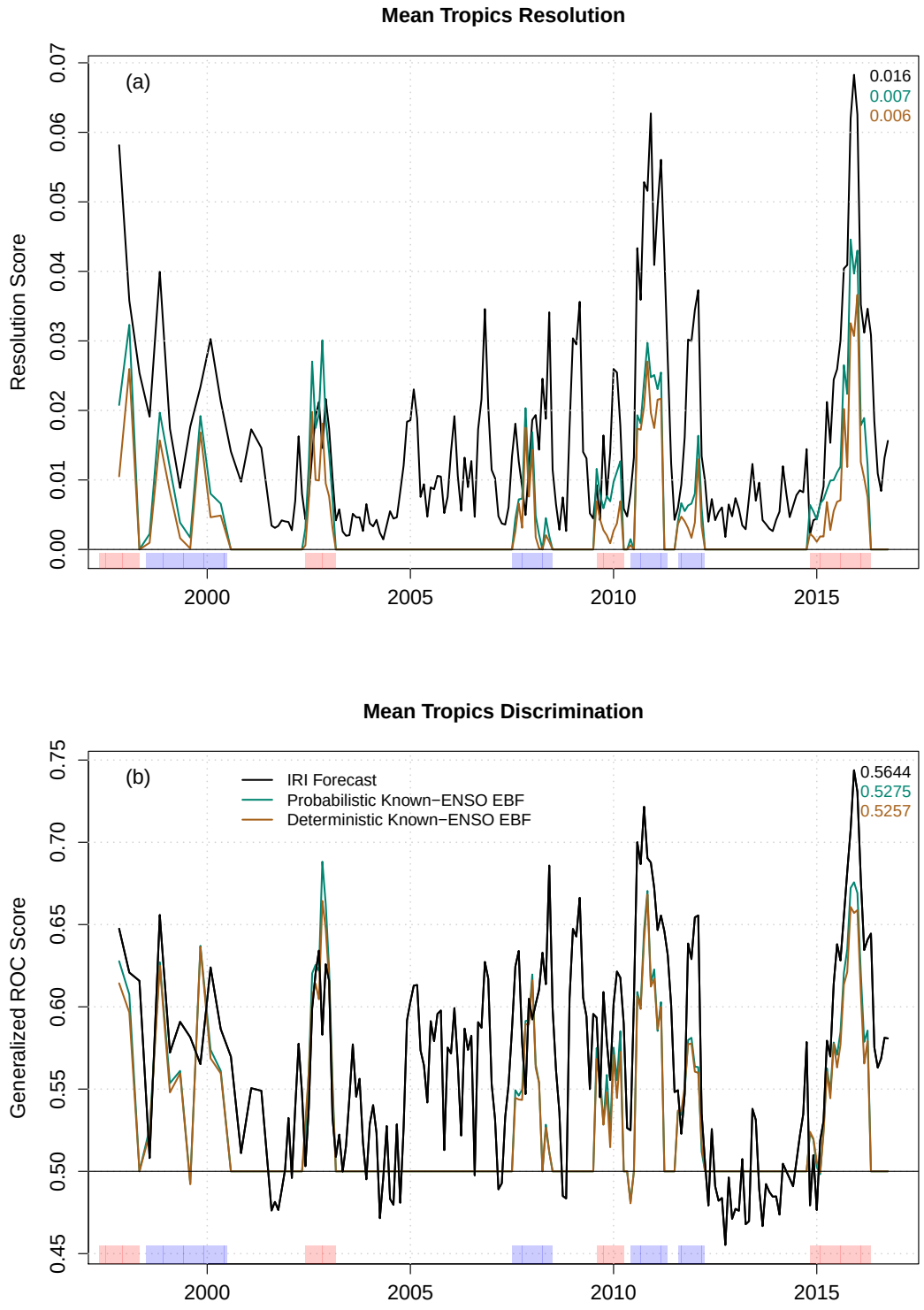

Figure 3.5: The (a) mean resolution and (b) mean discrimination over the tropics (30S–30N) of the three forecasts. The mean resolution and discrimination over the total record are denoted by the values with color corresponding to the forecasts.

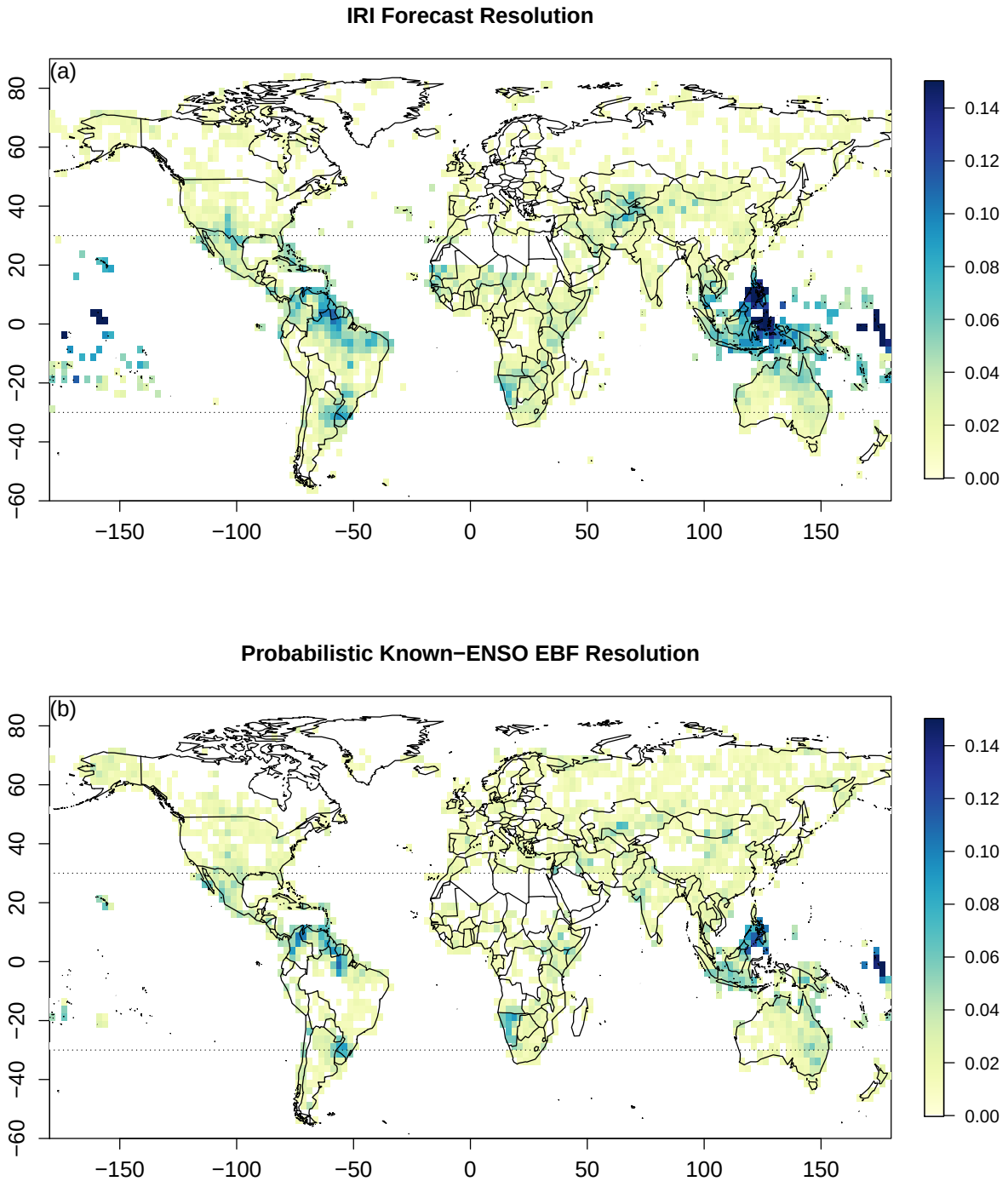

Figure 3.6: Spatial Distribution of the resolution score averaged over all 12 seasons for the (a) IRI forecast and (b) probabilistic known-ENSO EBF. Higher values are indicative of better forecasts as the outcome is more conditioned on the forecast probability.

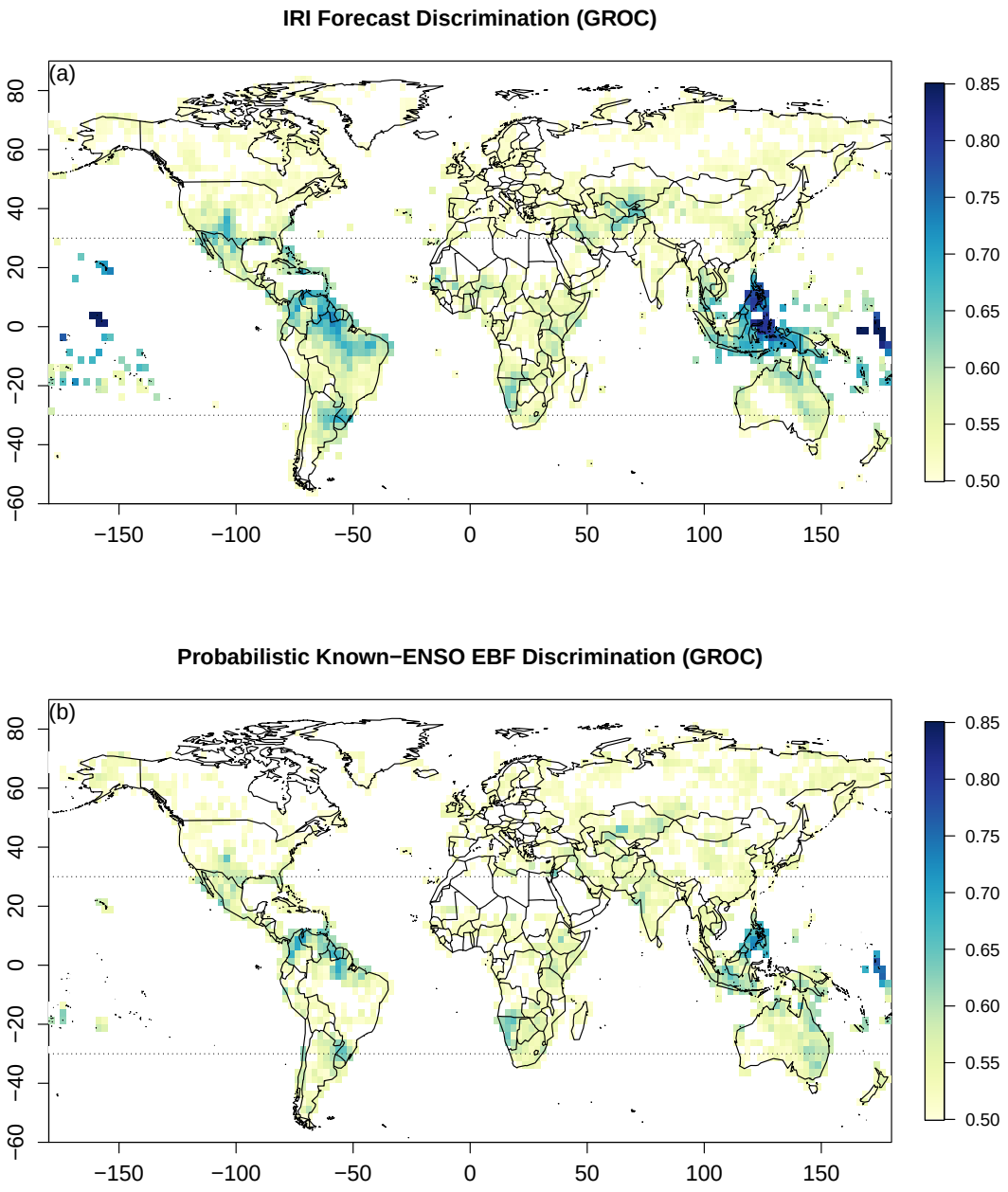

Figure 3.7: A comparison of the (a) IRI forecast and (b) probabilistic known-ENSO EBF discrimination as quantified by the GROC score. The EBFs issue maximum probability on the same category in nearly all cases resulting in similar discrimination.

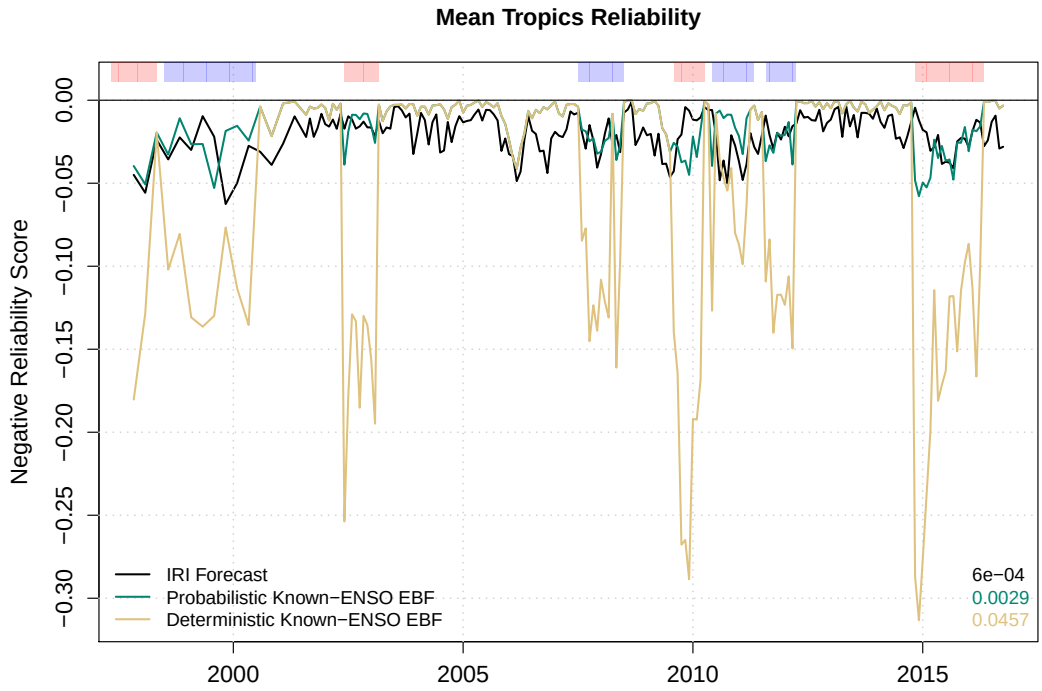

Figure 3.8: The mean negative reliability over the tropics (30S–30N) of the three forecasts. Negative reliability is plotted to remain consistent with the other verification plots where high values on the plot represent good forecast performance.

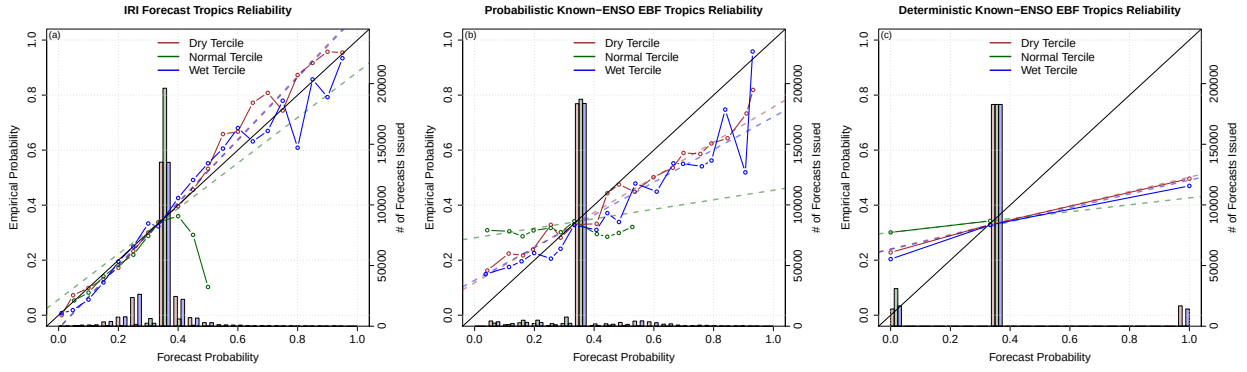

Figure 3.9: Reliability diagrams for the (a) IRI forecast, (b) Probabilistic known-ENSO EBF, and (c) Deterministic known-ENSO EBF with the forecast probability on the x-axis and the corresponding frequency of observed outcome on the y-axis. Histograms indicate the distribution and quantity of forecast probabilities issued. Dotted lines show a weighted linear fit of the reliability curve with weights determined by the number of forecasts issued in for a probability.

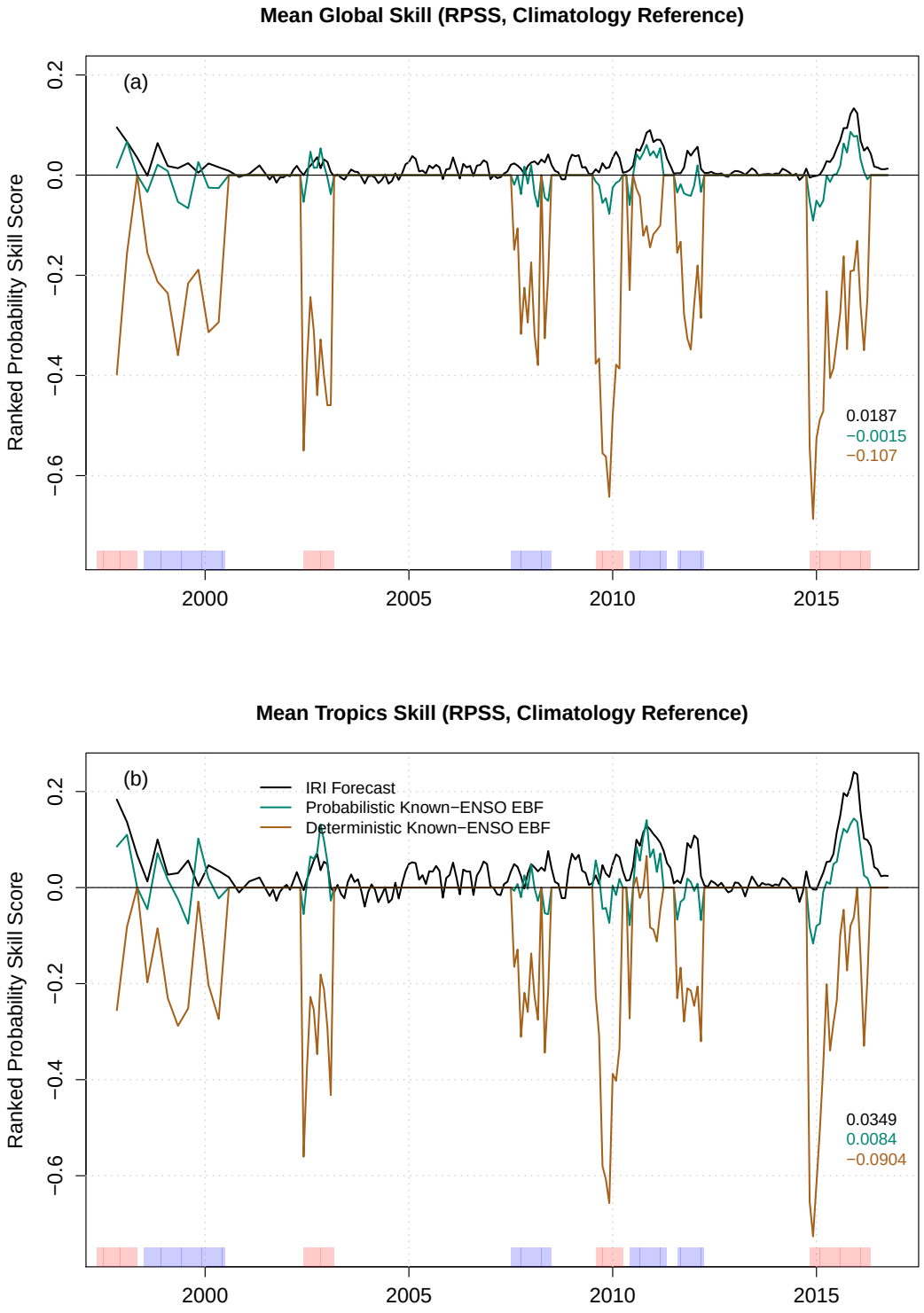

Figure 3.10: The (a) Global and (b) tropics mean RPSS for the IRI forecast and two known-ENSO EBFs. The results are generally consistent between the global and tropical series. Total RPSS scores over the record are given by the values in the bottom right corner.

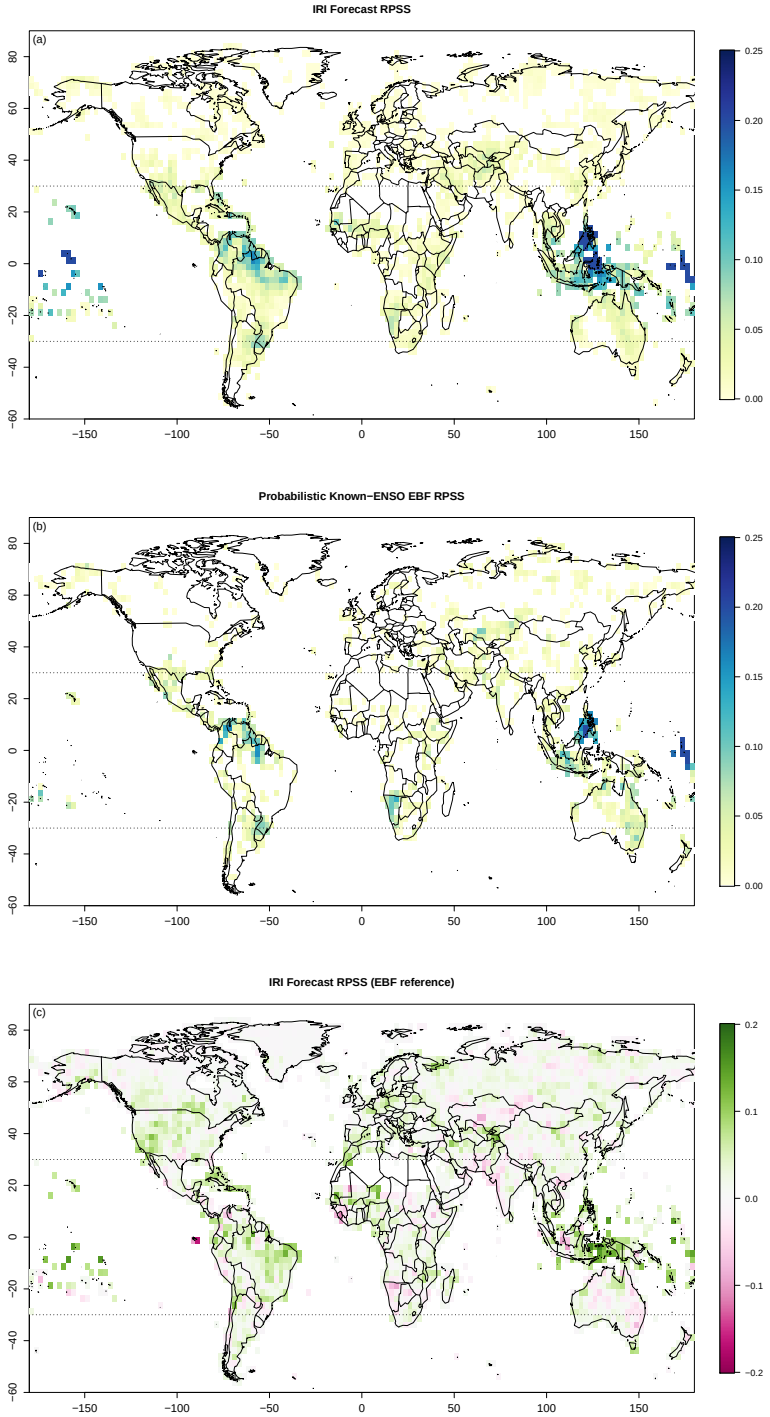

Figure 3.11: Spatial distribution of RPSS averaged over all 12 seasons for the (a) IRI forecast and (b) probabilistic known-ENSO EBF. (c) The RPSS of the IRI forecast using the probabilistic known-ENSO EBF as the reference is green where the IRI forecast has additional skill over the EBF and pink where it under-performs.

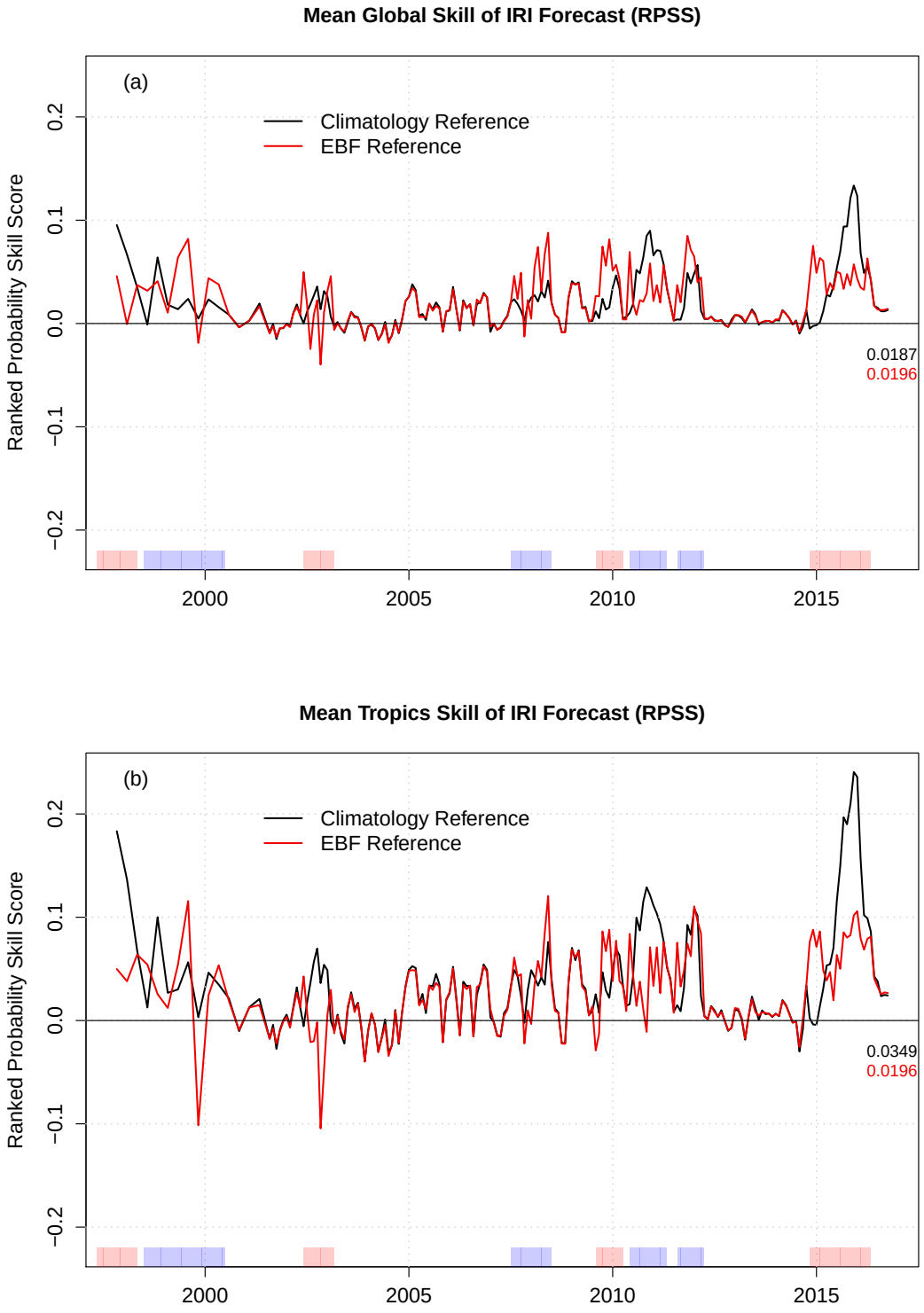

Figure 3.12: The (a) Global and (b) tropics mean RPSS for the three forecasts with reference forecasts of climatology and the probabilistic known-ENSO EBF. The EBF is equal to climatology in periods of neutral ENSO. Total RPSS scores over the record are given by the values in the bottom right corner.

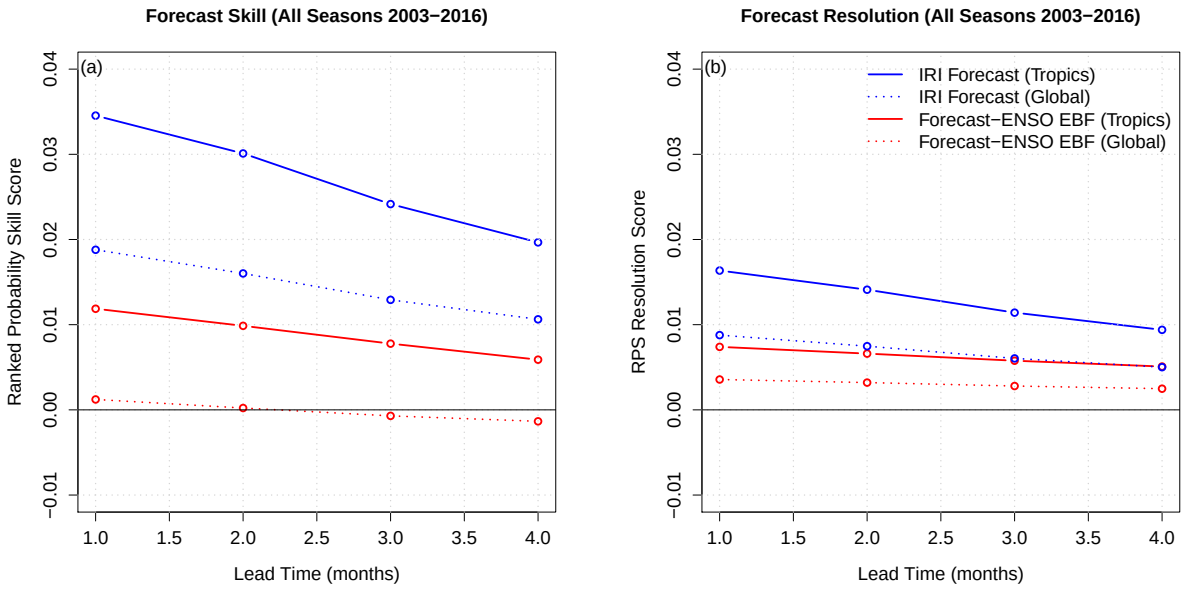

Figure 3.13: IRI forecast and probabilistic forecast-ENSO EBF forecast (a) RPSS and (b) Resolution as a function of lead time.

3.3.4 Lead Time Dependence

The RPSS and resolution skill as a function of lead time are calculated for the IRI and probabilistic forecast-ENSO EBFs forecasts (Figure 3.13). Both the IRI forecast and EBF have positive RPSS skill relative to climatology in the tropics at all lead times with the IRI forecast exhibiting uniformly greater skill. The higher resolution of the IRI forecast demonstrates that the additional skill of the IRI forecast is due to a greater information content, and not just improved reliability due to calibration. The IRI forecast has positive skill at the global level to at least 4-month lead times, but the EBF has nearly zero global skill over climatology. This is to say that the utility of the EBF in the extratropics is limited, but still may have longer lead skill in regions with strong teleconnections and regional skill such as the southwestern United States. As with the known-ENSO EBF verification, the verification was replicated using ignorance-based scores finding no substantive difference in results.

3.4 Discussion and Conclusions

This study updates Mason & Goddard (2001) with new data and better data quality control. The resulting probabilistic precipitation impact maps provide a dataset expanding the teleconnections found in MG01 to include many previously studied teleconnections. Investigation of the maps suggests a boreal summer El Niño above-normal precipitation teleconnection in France and a boreal summer La Niña below-normal teleconnection in the greater Bangladesh region, warranting further study.

There is some predictive skill in the simplest of the ENSO-based statistical forecast models as quantified by resolution and discrimination. Incorporating probabilistic information in the forecast increases resolution in addition to the expected increase in reliability. Additionally, including a realistic representation of the SST prediction uncertainty allows forecasts to be made at a variety of lead times and in borderline ENSO conditions where the deterministic and probabilistic forecasts only issue climatological information.

State-of-the-art seasonal forecasts using statistically-calibrated dynamical models provide more skill than the EBFs in nearly all regions and seasons. Including uncertainty arising from limited predictability of the state of the tropical Pacific at various lead times allows comparison of the EBF and IRI forecast as lead time changes. The probabilistic forecast-ENSO EBF exhibits skill in the tropics at leads out to at least four months, but the IRI forecast uniformly provides additional skill. Thus, users should generally avoid using simplified ENSO teleconnection maps as forecasts when possible.

The widespread implementation of state-of-the-art forecasts is an ongoing global project (Kumar *et al.* 2020). The trade off between simplicity and skill explored in the verification of the EBFs can inform the dissemination and suggested use of climate information. While the WMO continues to facilitate and encourage use of state-of-the-art forecasts, simplified EBFs may be more appropriate for users with limited resources as they are better forecasts than no information in many regions. However, the superior skill of the probabilistic EBF over the deterministic EBF shows

that communication of uncertainty in the impacts of ENSO is critical in the proper adoption of climate information for decision making (Suarez & Patt 2004; Hansen 2005; Roulston *et al.* 2006; Hirschberg *et al.* 2011). Potential communication methods could include the impact maps from this study or customized regional info-sheets that back up probabilities with historical data.

The verification study also reveals regions where forecast methods may be under-performing relative to the simple ENSO-based method. Predictions based on historical teleconnections exhibit greater skill in southwestern Africa, eastern Australia, and around the Bay of Bengal as shown by the probabilistic known-ENSO EBF. Also, greater EBF skill is observed in the parts of the northern hemisphere extratropics where more work is necessary to understand the nature of these results.

While the state-of-the-art forecast system generally outperforms the EBF system, especially in recent years, there is utility for these EBFs in the forecast development process. Moving forward, the IRI's realtime forecast will be verified with EBFs in addition to the standard climatology. The alternative verification provides new information about the additional skill provided by a state-of-the-art forecast in regions with and without known teleconnections.

Chapter 4: Advancing and Extending Seasonal Prediction with Model-Analogue Forecasts

This chapter is an expansion of first-author conference proceedings published as Lenssen et al. *(2022) in the 46th NOAA Climate Diagnostics and Predictability Workshop Digest.*

The work presented in Chapter 3 demonstrated the robust and widespread impact of the El Niño-Southern Oscillation (ENSO) on global precipitation as well as the substantial seasonal precipitation skill that can be achieved with teleconnection maps given a skillful forecast of the ENSO state. Forecasts of ENSO issued in realtime by forecasting centers, or *operational forecasts*, have steadily improved since the first ENSO forecasts in the late 1980s (Zebiak & Cane 1987). However, there still appears to be untapped skill in ENSO forecasts, particularly in the prediction of large ENSO events at leads longer than the 9-month forecast currently issued operationally by the International Research Institute for Climate and Society (IRI) and NOAA's Climate Prediction Center (Gonzalez & Goddard 2016; Dunstone *et al.* 2020).

Extending global ENSO prediction past these currently operational 9-month outlooks is of significant scientific and societal importance. The 2021 food crisis in the Horn of Africa highlights the need for, and potential of, multi-year ENSO predictions. Consecutive La Niña events often cause drought in the Horn of Africa (Hoell & Funk 2014; Lenssen *et al.* 2020) leading to reduced crop yields (Iizumi *et al.* 2014), which contributes to food crises (FEWSNet 2021). In this example, translation of the known predictability of consecutive La Niña events (DiNezio *et al.* 2017) into a skillful operational forecast would have provided critical early information to humanitarian organizations responding to the food crisis.

Key features of ENSO, particularly its duration, are theoretically predictable several years in advance (Gonzalez & Goddard 2016; DiNezio *et al.* 2017; Ham *et al.* 2019; Dunstone *et al.* 2020; Wu *et al.* 2021). Potential physical mechanisms leading to multi-year ENSO predictability include the subsurface heat content in the tropical Pacific Ocean (McPhaden 2003; Zhao *et al.* 2021), basin-scale Pacific Ocean dynamics (Vimont *et al.* 2003; Joh & Di Lorenzo 2019), and cross-basin interactions with the Indian (Mayer & Balmaseda 2021) and Atlantic Oceans (Ham *et al.* 2013). Despite our understanding of the dynamical processes leading to extended ENSO predictability, multi-year skill in traditionally initialized dynamical forecast systems has remained elusive.

Possible reasons why multi-year predictability does not always translate into predictive skill in initialized dynamical forecasts include model bias, errors due to initialization, and model drift. Nearly all coupled general circulation models (CGCMs) exhibit bias in the tropical Pacific mean state and variability, leading to unrealistic ENSO behavior (Li & Xie 2014; Planton *et al.* 2021). The link between ENSO bias and predictive skill is still poorly understood with some studies finding no relationship (Scaife *et al.* 2019), while others find a conclusive relationship (Ding *et al.* 2020). In addition, dynamical models have a preferred climate state that is different from the observed climate. When initialized to match the Earth's climate system, they will slowly return to their preferred climate over time in a process known as *model drift*. The effect of model drift on the mean and trend errors in forecasts can be reduced through proper bias correction (Kharin *et al.* 2012), but accounting for drifts in variability is still not well understood. Finally, *initialization shock*, or rapidly increasing forecast error due to the initial observed state being incompatible with the CGCM's dynamics, result in errors in the mean state and variability of ENSO (Mulholland *et al.* 2015; Hermanson *et al.* 2018), including a westward shift of the predicted ENSO anomaly that results in poor forecast skill in the western tropical Pacific (Newman & Sardeshmukh 2017). Understanding, reducing, and correcting initialization shock is necessary to improve our climate forecast systems. However, there has not been a comprehensive study on the effect of initialization on ENSO predictability in initialized dynamical prediction systems.

An additional complication in using initialized CGCMs for climate prediction research is the

immense amount of computing power and related time investment in model development and maintenance required to run prediction experiments. Large ensembles must be run at each initialization to capture the inherent uncertainty due to the chaotic nature of the coupled atmospheric-oceanic system. Recent work in the multi-model large ensemble project has shown that at least 10 ensemble members are needed to represent the possible internal variability with some regions needing over 40 members (Deser *et al.* 2020). In addition, forecasts over the historical record, or *hindcasts*, must be run at as many initializations as possible to span as many possible climate states as well as increase the sample size of the verification period. This is particularly important for ENSO and thus seasonal prediction as ENSO exhibits large multi-decadal variability in its activity and thus predictability (Wittenberg 2009; Wittenberg *et al.* 2014; Levine *et al.* 2017).

Recently, scientists have begun adapting model-analogue methods originally used in weather forecasting (Lorenz 1969a) to climate forecasting (Ding *et al.* 2018). Such model-analogue forecasts are a promising tool for investigating climate predictability as they do not suffer from initialization shock and require orders of magnitude less computing while taking advantage of the large number of publicly-available CGCMs simulations. Model-analogue forecasts are quite simple in principle; a forecast is made by matching the observed climate state to the closest states in a library of observed or CGCM-simulated climate states and using the evolution of the chosen library states as a forecast (Lorenz 1969a; Ding *et al.* 2018).

The first portion of this study quantifies the deterministic skill of model-analogue ENSO predictions. The hindcast skill of the model-analogue forecasts are found to be comparable to the state-of-the-art North American Multi-Model Ensemble (NMME) (Barnston *et al.* 2019). This similarity in skill, along with the findings of Ding *et al.* (2018) and Ding *et al.* (2019) suggest that model-analogue forecasts are appropriate for further investigations of model dynamics and predictability. The second portion of this study investigates the potential multi-year ENSO skill currently not captured in initialized prediction systems due to CGCM bias and initialization shock. The potential skill to be gained in initialized forecast systems will be assessed through a comparison of traditionally initialized forecasts with model-analogue forecasts, an empirical-dynamical

model based on CGCMs output.

4.1 Data

Note that all model and observational data are regridded to a $2^\circ \times 2^\circ$ grid prior to any analyses.

4.1.1 Observations

Historical observed monthly sea-surface temperature (SST) is taken from HadISST1.1 (Rayner *et al.* 2003). The observed Niño3.4 Index is calculated as the seasonal average of the SST anomalies over the standard domain of 5°N - 5°S and 170°W - 120°W. El Niño and La Niña events are defined as seasonal Niño3.4 anomalies that exceed the upper or lower quartile of the seasonal Niño3.4 index respectively, following Gonzalez & Goddard (2016). Historical observed sea-surface height anomalies (SSH) are taken from the ECMWF Ocean Reanalysis System 4 (ORAS4) (Balmaseda *et al.* 2013).

4.1.2 CGCM Output

Initialized predictions are from the CMIP6 Decadal Climate Prediction Project (DCPP) Component A hindcasts with annual initializations from 1960 - 2016 (Boer *et al.* 2016). This study was limited to the three models with complete data and corresponding pre-industrial control (piControl) runs of sufficient length on the Google Cloud CMIP6 archive: CanESM5 (Sospedra-Alfonso *et al.* 2021), CESM1-1-CAM5-CMIP5 (hereafter CESM1.1; Yeager *et al.* 2018), and MIROC6 (Kataoka *et al.* 2020). CESM1.1 and MIROC6 are both initialized in November and CanESM5 is initialized in January allowing for comparable verifications of multi-year ENSO skill. See Table 4.1.2 for a summary of the DCPP and piControl runs for the three models.

The Niño3.4 indices for the initialized hindcasts are bias corrected to account for both lead-dependent mean-state bias as well as trend bias due to model drift following Kharin *et al.* (2012). In addition, ENSO event thresholds in the initialized models are defined as the upper and lower quartile at each lead following Gonzalez & Goddard (2016). After bias correcting, probabilistic

Model	DCPP Members	DCPP Init. Month	Init. Method	piControl Years
CanESM5	20	January	Full-Field	2,000
CESM1.1	40	November	FOSI	1,800
MIROC6	10	November	Ocean Anomaly	800

Table 4.1: Details of the three CGCM DCPP and piControl experiments analyzed in this study. In the initialization method column FOSI stands for forced ocean sea ice initialization.

forecasts of ENSO state are made with the initialized hindcasts for JFM at lead years 0-5 for each CGCM.

4.2 Methods

4.2.1 Model-Analogue Forecasts

A model-analogue forecast is made by determining states in a library of model output that best match the observed climate state. The, the forecast ensemble is the evolution of each of these closest matching library states. Model-analogue forecasts follow the assumption for dynamical system theory that a pair of states that is initially similar will evolve along similar trajectories (Lorenz 1969a). Here, ENSO forecasts are made by minimizing the root-mean-square difference between observed Indo-Pacific SST and SSH to states in a library of CGCM piControl simulations following Ding *et al.* (2018).

Model-analogue hindcasts are made over the same 1960 - 2016 period using control runs from the same CGCM configuration as the traditionally initialized forecasts. While the model-analogue forecasts are able to issue forecasts with CGCM output while avoiding initialization shock, they have the disadvantage of larger error than traditionally initialized forecasts in the first month due to the analogues not perfectly matching the observed state. Probabilistic and deterministic forecasts of the Nino3.4 index are made with the closest 15 library members from as large of a library of piControl output as possible, following the findings of Ding *et al.* (2018).

4.2.2 Verification Metrics of Deterministic ENSO Skill

Two standard metrics are used to assess the deterministic skill of ensemble mean model-analogue Niño3.4 forecasts: the anomaly correlation and the mean squared error skill score (MSESS). Anomaly correlation is defined as the Pearson correlation coefficient of the ensemble mean Niño3.4 anomaly and the observed Niño3.4 anomaly. Statistical significance is determined using the t-distribution approximation (Lehmann & Romano 2005). The anomaly correlation measures forecast discrimination and is not affected by miscalibration of the forecast, making it a useful first-pass verification metric (Barnston *et al.* 2019).

The MSESS is a useful skill score that assess the accuracy by comparing the mean squared error (MSE) of a forecast with the MSE of a climatological forecast (Murphy 1988). The MSESS is defined as

$$\text{MSESS} = 1 - \frac{\text{MSE}_{forecast}}{\text{MSE}_{climatology}} \tag{4.1}$$

with positive values of the MSESS reflecting that the forecast has smaller MSE than climatology and therefore more accurate forecasts, negative values of the MSESS reflecting that the forecast has greater MSE than climatology and therefore less accurate forecasts. A MSESS of zero indicates that the forecast of interest and climatology have identical MSE and are equally accurate. Significant positive MSESS skill is assessed using a non-parametric sign test Lehmann & Romano 2005.

The MSESS is a particularly useful metric for assessing Niño3.4 predictions as it penalizes forecasts for incorrect amplitude. The MSESS can be decomposed into three components:

$$\text{MSESS} = cor^2 - \left(cor - \frac{SD_{forecast}}{SD_{obs}}\right)^2 - \left(\frac{mean_{forecast} - mean_{obs}}{SD_{obs}}\right)^2 \tag{4.2}$$

where the first term is the anomaly correlation, the second term is the bias in the forecast amplitude, and the third term is the bias in the forecast mean (Murphy 1988; Barnston *et al.* 2019). Simulations, and therefore forecasts, of ENSO often fail to produce sufficient in ENSO variability,

particularly in the winter (Figure 4.1). This lack of variability is captured in the MSESS and can be further quantified through the decomposition above.

4.2.3 Verification Metrics of ENSO Event Detection

Deterministic and probabilistic verification metrics of the Nino3.4 index are often too challenging of a standard for assessing ENSO predictability at leads longer than 6 month. Previous work suggests that assessing multi-year ENSO forecast skill in the prediction of individual ENSO events, rather than the exact evolution of the Niño3.4 index, reveals significant skill past the first year (Gonzalez & Goddard 2016). As such, the DCPP and model-analogue hindcasts are verified against the observed ENSO events as calculated through the upper and lower quartile of the Niño3.4 index. The forecasts of interest are probabilistic forecasts of ENSO events for which forecast skill is assessed by the area under the receiver operating characteristic (ROC) curve (Mason 1982; Hogan & Mason 2012). The ROC area is a measure of forecast discrimination, a skill measure relative to a climatological forecast, and can be generalized as a U-Statistic providing useful statistical properties (Mason & Graham 2002). In particular, statistical significance of skillful ROC scores is assessed using the assumption that the U-statistic is Gaussian, which holds for large sample sizes (Mason & Graham 2002).

4.3 CGCM ENSO Climatology

ENSO mean state and variability is also assessed in the uninitialized piControl runs and the initialized DCPP runs of the three CGCMs (Figure 4.1). The seasonal cycle of each of the piControl models roughly matched that of observations with temperature peaking in boreal summer and reaching minimum during boreal winter. However, the timing of the winter minimum is early in all three piControl simulations suggesting a mismatch in ENSO timing when compared to observations. In addition, each of the three models shows an ENSO climatological amplitude that is larger than observations, in agreement with other studies of CMIP5 and CMIP6 ENSO amplitude (Bellenger *et al.* 2014; Planton *et al.* 2021). The evolution of the seasonal cycle of the initialized

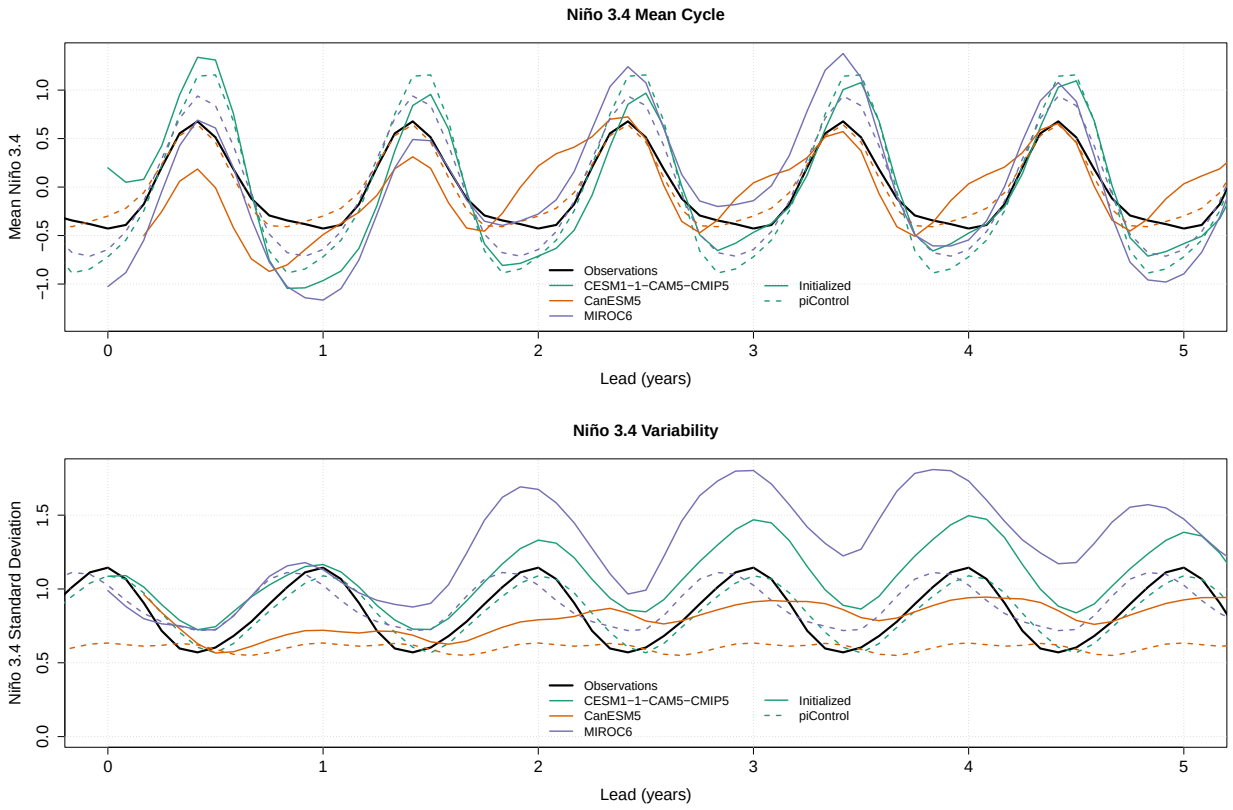

Figure 4.1: The climatology of the initialized and uninitialized ENSO simulations as compared with observations with the lead time zeroed at January of the first year. Shown is (top) the mean cycle of monthly Nino3.4 mean absolute temperature with the 12-month running mean removed and (bottom) variability of monthly Niño 3.4 absolute temperature. The observations (solid black line) are calculated over 1960-2016 and do not depend on lead time. The piControl climatologies (colored dashed lines) are calculated over the entire length of the piControl and also do not depend on lead time. The initialized model climatologies (colored solid lines) vary with lead time, reflecting the lead-dependent biases in mean and variability.

models shows the signature of errors arising from initialization (Figure 4.1). In particular, the initialized CanESM5 model shows a large and growing departure from observations in the critical winter months.

The variability of Niño3.4 temperatures is assessed to determine if the ENSO variability is properly represented in the models. CESM1.1 and MIROC6 both show quite realistic variability in the uninitialized piControl runs, but CanESM5 shows dramatically reduced ENSO variability in the key boreal winter months suggesting that CanESM5 struggles to simulate a diverse range of

ENSO amplitudes in agreement with (Planton *et al.* 2021). When the models are initialized, each of their ENSO variabilities grow as a function of lead time while preserving a similar seasonal cycle to that of the uninitialized versions. This behavior is potentially a signature of initialization shock-caused dynamical corrections in the ENSO system.

4.4 Deterministic Skill of Model-Analogue ENSO Forecasts

In the first portion of this study, the deterministic skill of model-analogue forecasts for the three CGCMs is analyzed. It is useful to perform a basic verification to confirm that the model-analogue forecasts from the particular CGCMs in this study approximately match the skill shown in previous studies of model-analogue forecasts (Ding *et al.* 2018; Ding *et al.* 2019). This deterministic verification also provides useful information about the seasonality of ENSO predictability as it is conducted with hindcasts initializations each month of the year, instead of just once a year as in the DCPP hindcasts. While we expect that the longest multi-year skill will arise from forecasts initialized around the peak of an ENSO event in boreal winter (DiNezio *et al.* 2017), it is useful to confirm this in the model-analogue setting.

The full-field SST anomaly correlation is assessed for each model at leads of 0 (Figure 4.2) and 6 months over the tropical Indo-Pacific (Figure 4.3). The anomaly correlation at lead-0 shows the initialization bias inherent in model-analogue forecasting. Critically, each model shows very high correlations in the equatorial central and eastern Pacific, suggesting that the analogues are being initialized with the correct ENSO state. Comparing the three models, there are few differences in the spatial pattern or magnitude of the lead-0 skill. Differences between the models emerge at lead-6 (Figure 4.3) with CESM1.1 exhibiting greater skill in the critical central Pacific. There is also substantial lead-6 skill in CESM1.1 and MIROC6 in the northwestern and southwestern Pacific, a region with skill associated with the atmospheric bridge (Alexander *et al.* 2002) and previously shown to have model analogue skill (Ding *et al.* 2018). However, CanESM5 does not exhibit much skill in the northwestern Pacific, making it likely that it will not be capable of skillful forecasts over land in the northern hemisphere as much of this predictability depends on ENSO teleconnections

through the atmospheric bridge mechanism (Alexander *et al.* 2002).

The Niño3.4 anomaly correlation skill for each model follows the general ranking of the models established in the full fields, with CESM1.1 having slightly greater and more persistent correlation (Left column of Figure 4.4). The typical spring predictability barrier pattern is evident, with skill dropping off more quickly as a function of lead time for forecasts made for the summer and early fall months. There is a local maximum of anomaly correlation skill in predicting January at a lead of 18 month for CESM1.1 and CanESM5 suggesting some predictability of a winter ENSO event from two summers prior. This pattern is also seen in the MSESS (Middle column of Figure 4.4), but this skill is not statistically significant. In general, the MSESS of each of the three models is comparable with a recent verification of the operational NMME (Barnston *et al.* 2019), adding support to the findings that the model-analogue approach is on par with operational seasonal forecast systems (Ding *et al.* 2018; Ding *et al.* 2019).

The MSESS is decomposed to assess the amplitude bias of the forecasts (Right column of Figure 4.4). From the analysis of the piControl ENSO variability (Figure 4.1), it is expected that model-analogue forecasts from CESM1.1 and CanESM5 will have reasonable ENSO variability, but MIROC6 will likely underestimate ENSO variability, particularly in the winter. The findings support this theory as the dark red diagonal line in the bottom portion of the MIROC6 amplitude bias (Figure 4.4) corresponds to model initializations in the boreal winter. More investigation is needed to understand how the lack of variability in MIROC6 boreal winter ENSO leads to large Niño3.4 amplitude bias for model-analogue forecasts initialized during this time.

The model-analogue results presented here are created with CGCM piControl libraries of different length, ranging from 851 years in MIROC6 to 2,000 years in CanESM5. The effect of library size on January Niño3.4 MSESS shows that skill generally increases with library size, but that the increases in skill are very small after a library size of 250 years (Figure 4.5). The one exception to the observed monotonic increase in skill with library size is the skill maximum in CanESM5 lead-9 skill at a library size of 250 years (Figure 4.5). However, these findings suggest that model-analogue forecasts made from piControl runs of at least 500 years are reasonable to

compare. This analysis notably does not quantify the effect of library size on the prediction of large or rare multi-year ENSO events where it is likely that large library sizes will be critical to span the possible ENSO variability spectrum (Wittenberg 2009; Timmermann *et al.* 2018).

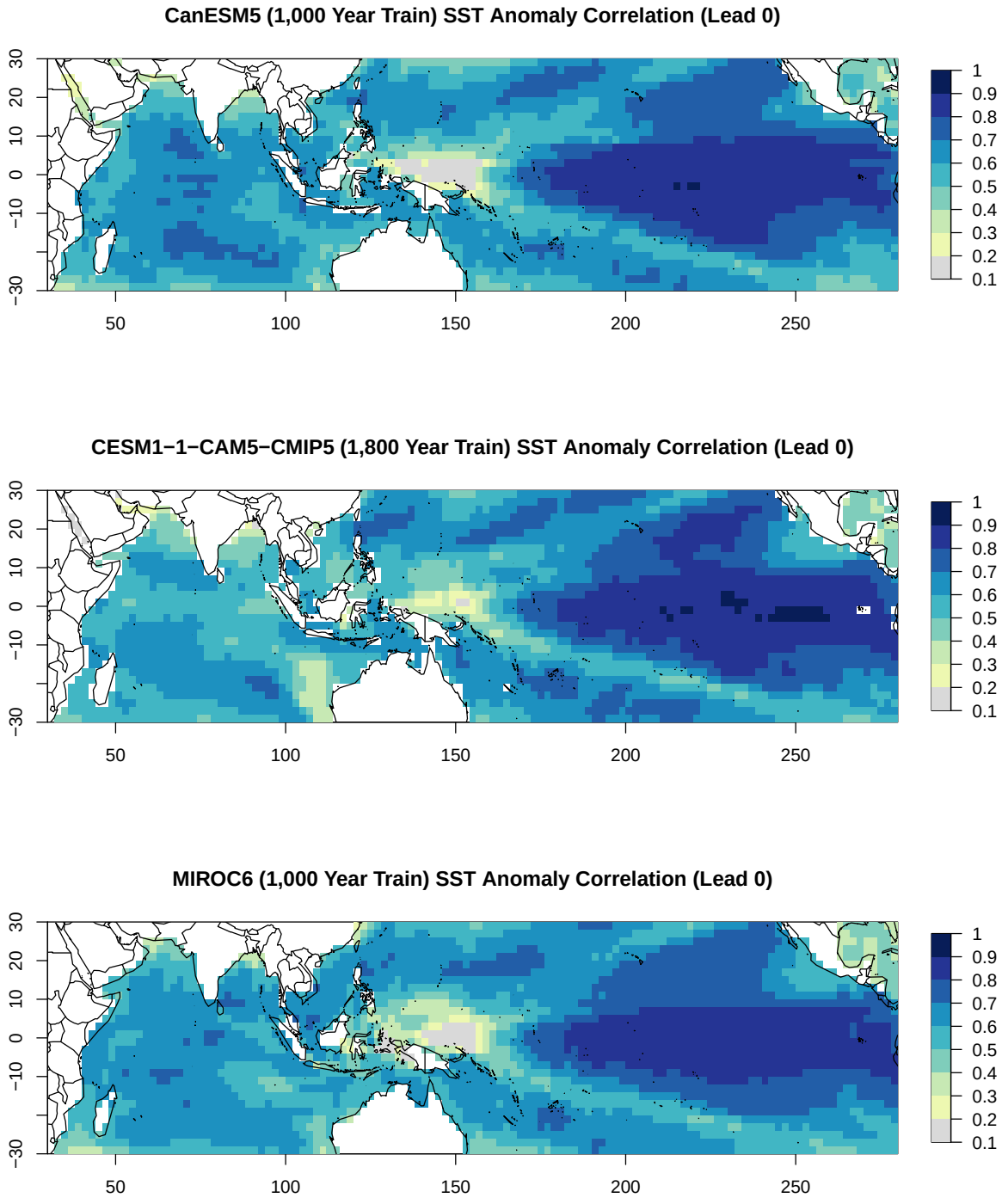

Figure 4.2: SST anomaly correlation of 0 month lead model-analogue hindcasts for each of the three models in the study.

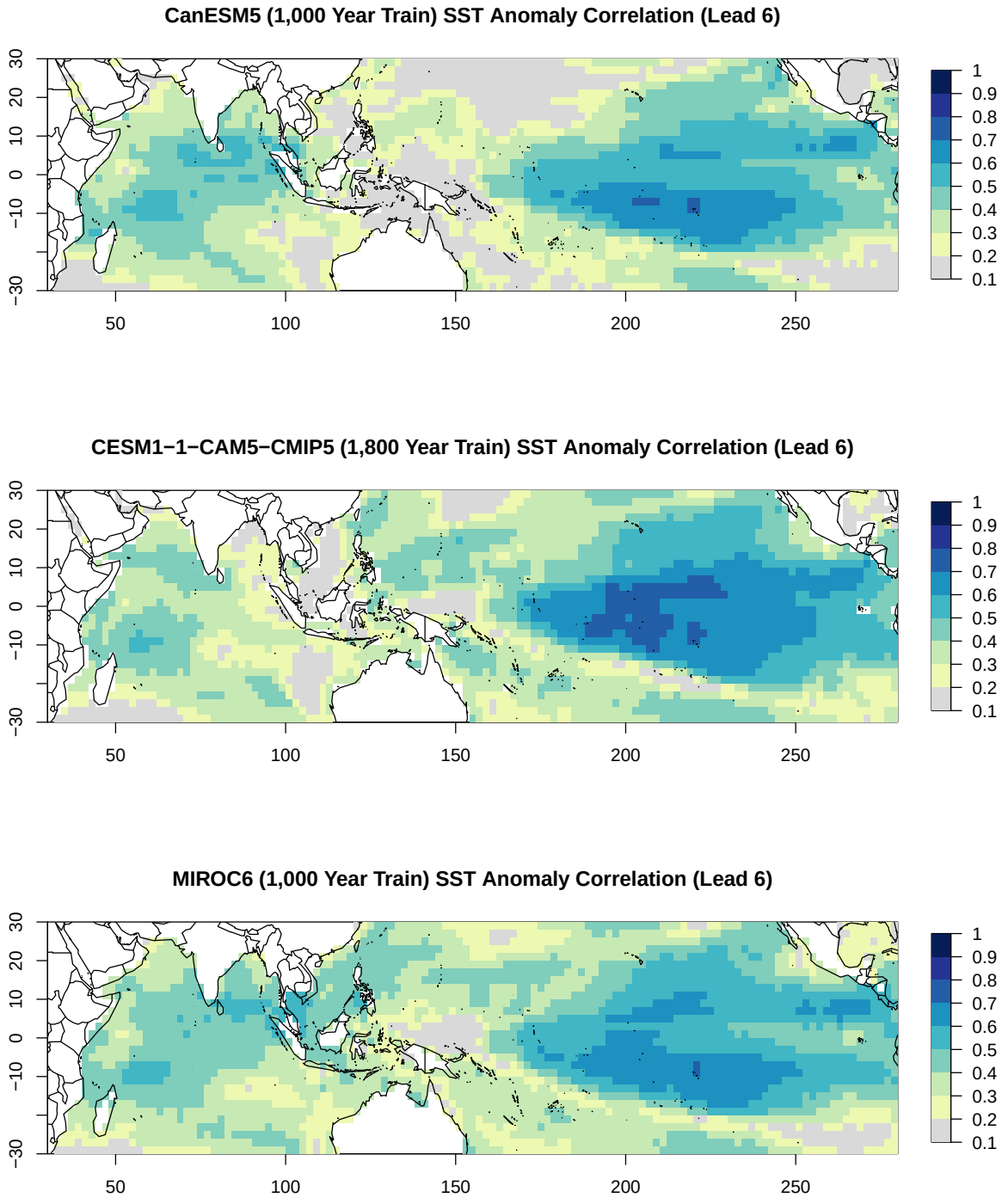

Figure 4.3: SST anomaly correlation of 6 month lead model-analogue hindcasts for each of the three models in the study.

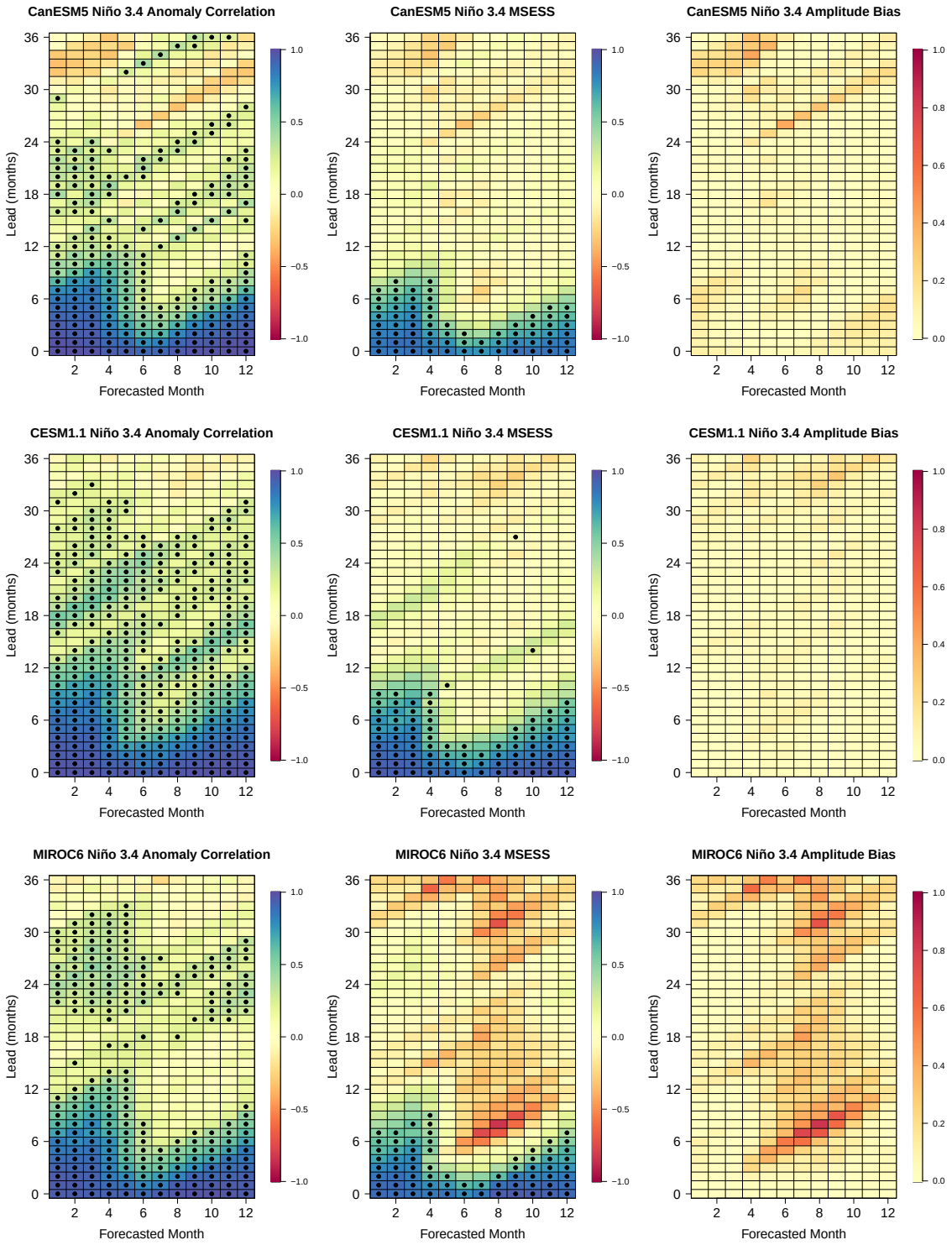

Figure 4.4: Deterministic verification of Niño 3.4 hindcasts for (top row) CanESM5, (middle row) NCAR CESM1-1-CAM5-CMIP5 and (bottom row) MIROC6. (Left column) The anomaly correlation (AC) with statistically significant AC shown with a dot. (Middle column) MSESS with statistically significant positive skill shown with a dot. (Right column) The amplitude bias component of the MSESS decomposition.

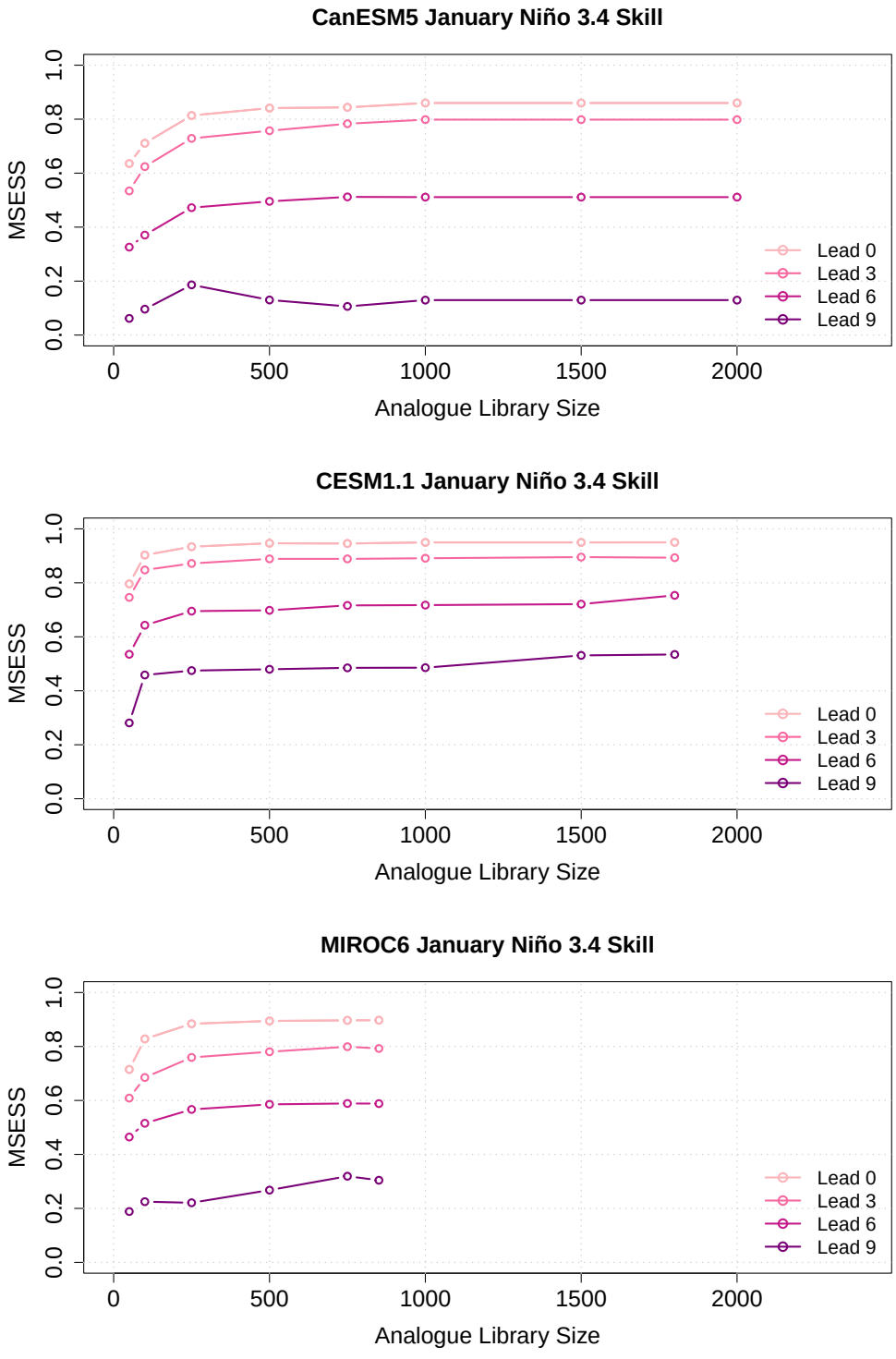

Figure 4.5: Deterministic ENSO prediction skill as measured by January Niño3.4 MSESS as a function of analogue library size and lead time for the three models in the study.

4.5 Initialization Shock and Long-Lead ENSO Predictability

In the second portion of this study, the skill of predicting ENSO events is compared between traditionally initialized CGCM prediction systems and model-analogue predictions. It is well established that initialization leads to model drift as the dynamical model returns to its preferred climatology (Smith *et al.* 2007; Balmaseda & Anderson 2009; Kharin *et al.* 2012; Sanchez-Gomez *et al.* 2016). In addition, there is growing evidence that such initialization shock also leads to spurious dynamics, particularly arising due to unrealistic biases in the air-sea heat fluxes, leading to not only bias in mean climate, but incorrect representation of key climate dynamics (Mulholland *et al.* 2015; Hermanson *et al.* 2018). Since ENSO dynamics rely on proper representation of the coupled air-sea system in the tropical Pacific, it is likely that such initialization shocks in initialized dynamical prediction systems are degrading the predictive skill. In addition, it is likely that initialization shocks are particularly affecting ENSO skill at longer leads as the errors arising due to initialization shock can grow over time (Mulholland *et al.* 2015) and that the error growth in ENSO are dependent on the initial state (Hermanson *et al.* 2018).

In the previous section, the strength of model-analogue predictions for conducting climate forecast research with limited computational resources was highlighted. Here, the other strength of model-analogue forecasts is highlighted; that model-analogue forecasts utilize dynamical simulations of the climate system, but do not suffer from initialization shock. This motivation was put forward in Ding *et al.* (2018), the first paper to explore seasonal prediction with model analogues. However, model-analogue forecasts suffer a tradeoff for having no initialization shock as they have substantial lead-0 bias due to the lack of perfect matches of the observed state in the analogue library. Here, this tradeoff in model-analogue forecasts of no initialization shock vs. greater lead-0 bias is explored in the context of multiyear ENSO forecasts.

The initialized models all show significant skill in predicting La Niña events at 1-year leads, extending the findings of DiNezio *et al.* (2017) that multi-year La Niña events were predictable in CESM1.1 to multiple CGCMs (Figure 4.6). However, the uninitialized model analogues do

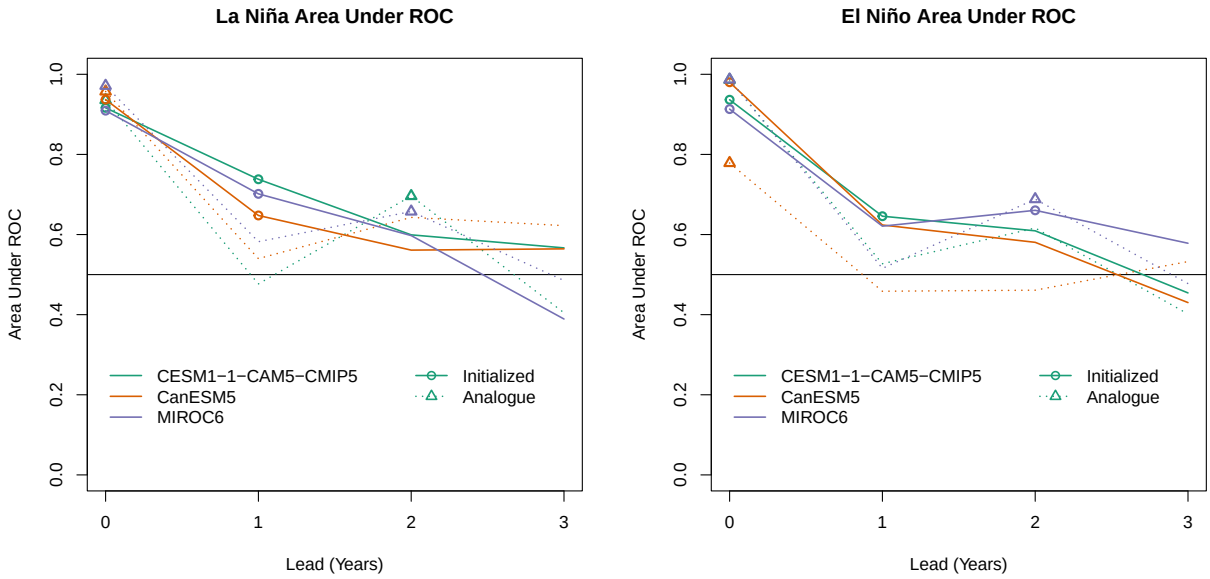

Figure 4.6: The ROC skill for ENSO event detection at 0-3 year leads. Climatological skill is 0.5 and marked by the solid black line. Statistically significant positive skill is marked with a circle for initialized forecasts or triangle for model-analogue forecasts.

not show this same skill at 1 year. Curiously, these model analogue forecasts are more skillful at 2 years than at 1 year, a finding which warrants further investigation. This year 2 skill is robust across all forecast probabilities providing further evidence that this year 2 skill is robust (Figure 4.7 - 4.9). There is less skill for El Niño events at 1 year, with only the initialized CESM1.1 showing significant discrimination when compared to climatology (Figure 4.6). Again, the increase in skill from lead-1 to lead-2 is observed in many of the initialized and model-analogue predictions, though the individual increases are not statistically significant. Notably, both initialized and model-analogue MIROC6 show significant skill in lead-2 El Niño despite not showing skill in lead-1. No conclusive link between the CGCM ENSO climatological biases and predictive skill was found, but such investigations will be revisited as more models are added into the analysis.

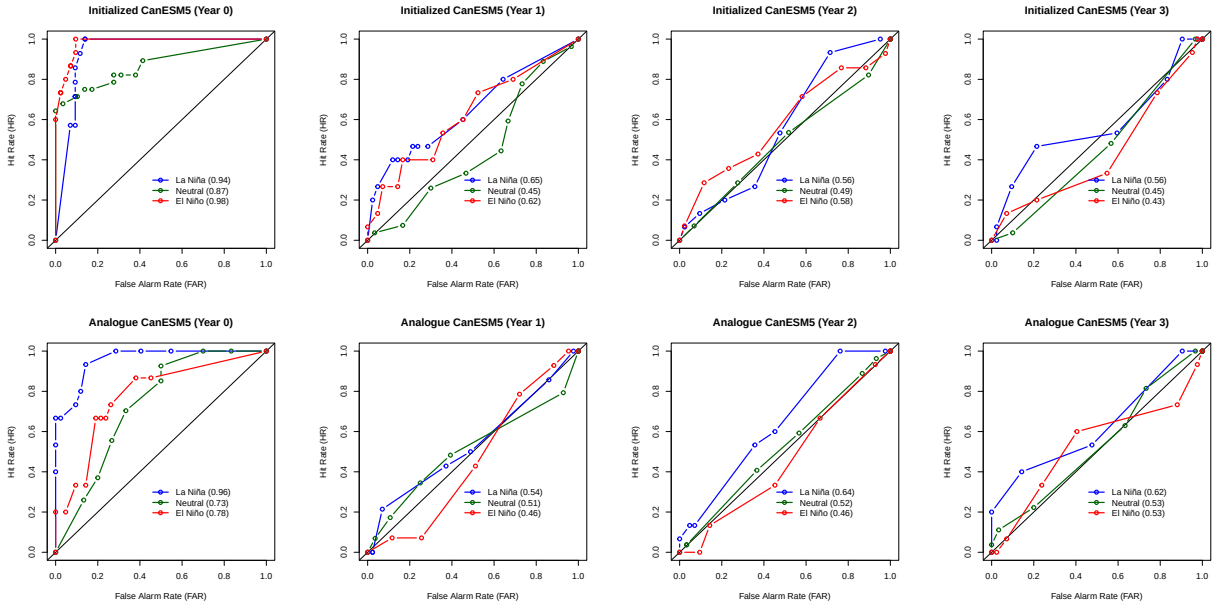

Figure 4.7: The ROC Diagrams for ENSO event detection at 0-3 year leads for (top row) CanESM5 initialized hindcasts and (bottom row) CanESM5 hindcasts using model-analogues.

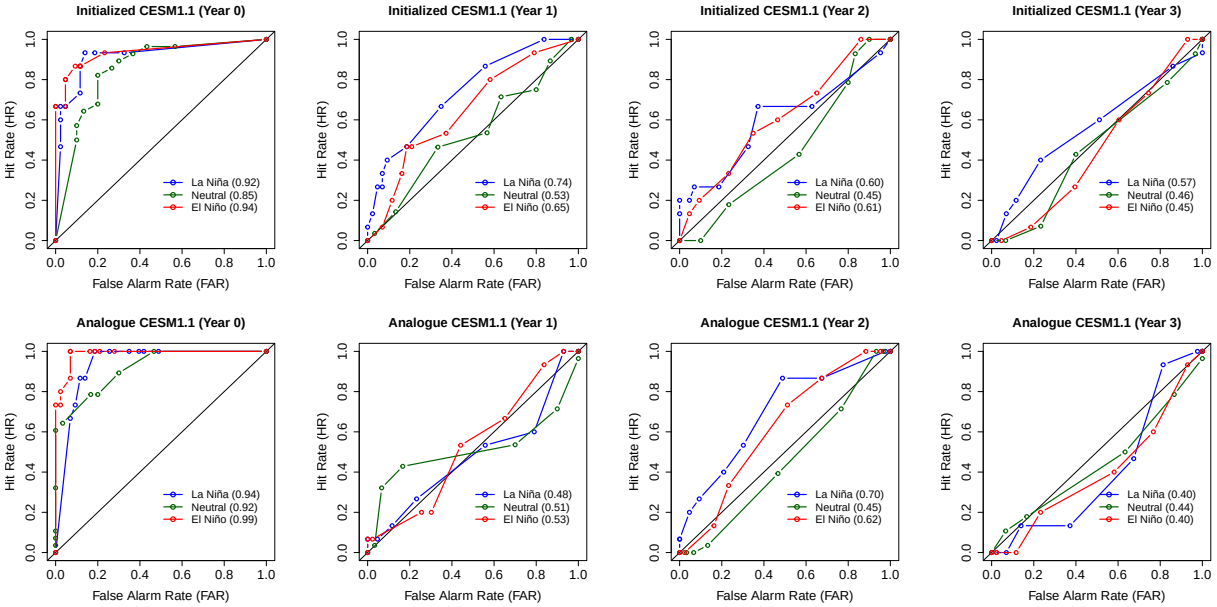

Figure 4.8: The ROC Diagrams for ENSO event detection at 0-3 year leads for (top row) CESM1.1 initialized hindcasts and (bottom row) CESM1.1 hindcasts using model-analogues.

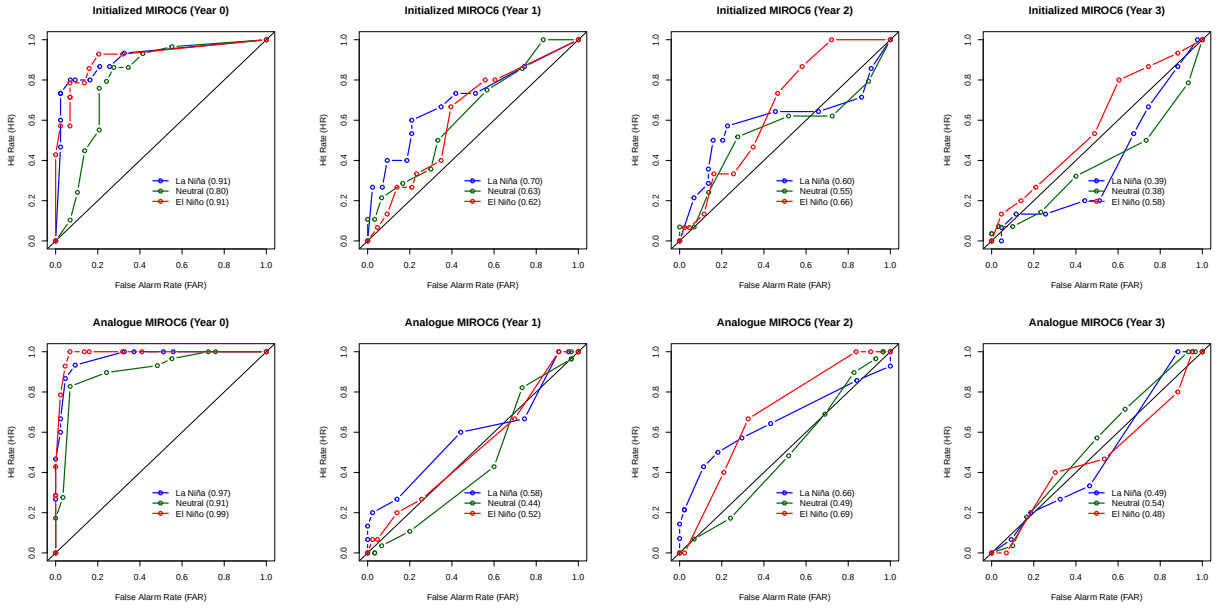

Figure 4.9: The ROC Diagrams for ENSO event detection at 0-3 year leads for (top row) MIROC6 initialized hindcasts and (bottom row) MIROC6 hindcasts using model-analogues.

4.6 Conclusions

This work is a promising initial investigation of multi-year prediction of ENSO events using CMIP6-class initialized models as well as model-analogue predictions using existing control run simulations. Here, evidence for the effect of initialization on CGCM ENSO dynamics is presented. However, the hypothesis that these growing errors due to initialization shock would make model-analogue predictions superior at longer leads was not conclusively shown as the initialized forecasts show more discrimination in forecasting ENSO events at most leads. However, the significant two year lead predictive skill shown by model-analogue forecasts for both La Niña and El Niño events suggests that model-analogue forecasts may be a useful tool for extending ENSO prediction past the first year.

Looking forward, this work will continue by expanding the number of models analyzed up to the 11 CGCMs that plan to submit runs to the DCPP, allowing for a more rigorous investigation of the link between model ENSO dynamics and predictive skill. In addition, the model-analogue methods will be compared with the predictions from linear inverse models (LIMs), which have comparable skill to initialized predictions Newman & Sardeshmukh 2017. Finally, an investigation of successful long-lead forecasts will be conducted to determine if there are certain characteristics of ENSO events and/or CGCM systems that lend themselves to greater predictive skill.

Chapter 5: Decomposition and Attribution of Observed Climate Variability

The prediction of low-frequency evolution of climate at the scales from a year to a decade, or interannual-to-decadal prediction, is an active and growing area of research (Cassou *et al.* 2018; Kushnir *et al.* 2019; Merryfield *et al.* 2020). Sitting at the transition between an initial condition and boundary value problem, decadal prediction draws upon the potential predictability of decadal variations in ocean dynamics as well as natural and anthropogenic changes to the atmospheric composition. To this point in time, significant global skill has only been shown in massive multi-model ensembles of state-of-the-art decadal prediction systems (Smith *et al.* 2019) due to the relativity small decadal signal present in most of the world as well as the complexity of interactions between various components of the Earth system.

The traditional view of climate prediction involves three steps (Goddard 2012). First, a predictand of interest is identified for some region and time scale. Then, the predictand must be represented as a manifestation of dynamical climate processes. Finally, it must be shown that the underlying dynamical processes have some predictable nature that can be represented in a model. If these criteria are met, a prediction has the potential to be skillful and ultimately useful. These steps are relatively clear in seasonal climate prediction. Regional variability in temperature and precipitation (predictands) are linked to variations in the El Niño-Southern Oscillation (ENSO) (Ropelewski & Halpert 1987; Mason & Goddard 2001) whose evolution can be predicted at lead times of months (Zebiak & Cane 1987; Zebiak 1989) to years (Gonzalez & Goddard 2016).

The search for skillful decadal prediction has proceeded with a similar philosophy, but has run into difficulties. There is a strong observational and model evidence for a decadal mode in the Atlantic Ocean known as the Atlantic multidecadal oscillation (AMO) (Kushnir 1994; Delworth & Mann 2000; Zhang *et al.* 2019). However, the analogous Pacific decadal oscillation (PDO)

(Mantua *et al.* 1997; Mantua & Hare 2002) is not as physically well-defined and now thought to be a combination of many dynamical processes (Newman *et al.* 2016). For the well-defined AMO, limited robust impacts have been found over land (Knight *et al.* 2006; Ting *et al.* 2011; Ting *et al.* 2014). Furthermore, dynamical simulations struggle to accurately represent the AMO teleconnections globally (Han *et al.* 2016; Liu & Di Lorenzo 2018).

While seasonal climate prediction gets the majority of skill from the initial conditions of the climate system, the longer timescales in decadal climate prediction necessitate accounting for changing boundary conditions due to changes in radiative forcings. This time-evolution of radiative forcings make decadal prediction a more difficult problem, but also provide another path towards predictive skill. For example, the majority of decadal predictability is currently linked to the forcing in regions outside of the North Atlantic (Corti *et al.* 2015; Yeager *et al.* 2018).

This study focuses on the second step of a climate prediction system as outlined above: attributing observed temperature variability to climate processes. That is, what are the physical drivers of interannual-to-decadal variability of temperature? The study is an extension of the empirical decadal prediction method of (Suckling *et al.* 2017) which uses linear regression to statistically model annual temperature series at global and regional scales. The attribution of interannual temperature variability to interannual to multi-decadal drivers in the climate system partitions the forced and internal variability and reveals potential teleconnections for further study. Once tested on observational data, these methods can be extended to climate model output to evaluate the simulation of impacts affected by decadal variability.

While the early stages of this study follow in the footsteps of empirical decadal prediction, it is important to note that the ultimate goal is not to develop a statistical prediction system as has been traditional (Hawkins *et al.* 2011). Rather, the goal is to link ocean variability and the temporal evolution of radiative forcings to observed global and regional temperature variability. Methodologically, this necessitates the development of statistical models for attributing the temperature variability to potential drivers of variability.

5.1 Data

The analysis is conducted at annual temporal resolution over 1900–2000. The start of the analysis period is limited by the quality of land and sea surface temperature data and is unlikely to be extended. The end of the analysis period is limited by the climate model derived radiative forcing data and will extended with data from the upcoming release of the Climate Model Intercomparison Project 6 (CMIP6) forcing experiments.

5.1.1 Historical Temperature data

The primary surface temperature anomaly data used is the Berkeley Earth analysis (Rohde *et al.* 2013a) which has global coverage of monthly surface air temperature and sea surface temperature anomalies on a $1^\circ \times 1^\circ$ grid. Anomalies are calculated relative to a 1951–1980 climatology. The Berkeley Earth method provides nearly full coverage during the early 20^{th} century through interpolation via Kriging. Annual temperature anomalies are calculated as the simple mean of monthly temperature anomalies.

The sensitivity of the results to the raw data homogenization method is tested by replicating the analysis with version 4 of the NASA Goddard Institute for Space Studies (GISS) Surface Temperature Analysis (GISTEMPv4) analysis (Hansen *et al.* 2010; Lenssen *et al.* 2019). GISTEMPv4 uses a comparable methodology to the Berkeley Earth analysis, but the homogenization from the fourth-generation Global Historical Climatology Network (GHCNv4) monthly data set (Menne *et al.* 2018) whereas Berkeley Earth uses their own, independent method for homogenization (Rohde *et al.* 2013b).

5.1.2 Forcing Data

The impact of changing atmospheric properties is quantified through radiative forcings: the energetic imbalance of the planet prior to equilibrium. Annual instantaneous radiative forcings were calculated from 1850–2000 as part of the GISS Model E2 CMIP5 historical runs (Miller

et al. 2014). An iRF is the radiative imbalance at the tropopause before any adjustment to the troposphere or stratosphere and is useful for characterizing the interannual radiative influence of forcing agents with time-evolving concentrations.

Effective radiative forcing calculations (ERF) are another form of radiative forcings that allow for the full adjustment of the troposphere, accounting for rapid feedbacks such as changes in water vapor and cloud properties (Hansen *et al.* 2005). These are much more computationally intensive to compute, but provide a better measure of the eventual equilibrium surface temperature. However, the current methods for calculating ERFs causes errors in the land surface temperature on shorter time scales (Shindell *et al.* 2013; Miller *et al.* 2014). As we are focused on the interannual to decadal variations in temperature, the iRF serves better.

Anthropogenic Forcings

Radiative forcings with primarily anthropogenic origins are split into three groups: well-mixed greenhouse gases (WMGHG), tropospheric aerosols, and ozone. The annual global iRF for the WMGHG combines the radiative forcing due to changes in atmospheric CO_2, CH_4, N_2O, and chlorofluorocarbons (CFCs) into a single radiative forcing. The tropospheric aerosol radiative forcing includes the aerosol direct effect: the forcing due to the absorption or reflection of radiation by the particles themselves, as well and the aerosol indirect effect: the change in cloud albedo and longevity due to increase nucleation sites. The ozone radiative forcing is dominated by increased tropospheric ozone, but also accounts for the decreasing stratospheric ozone due to CFCs.

The global radiative forcings due to the WMGHG, ozone, and tropospheric aerosols are highly collinear as they are all driven by increasing emissions due to industrialization. Collinearity, or correlation between predictors, is problematic for regression models and leads to increased uncertainty. As an initial workaround, these three forcings are combined into a single annual anthropogenic forcing (AF) series for the decomposition of global mean temperature (Figure 5.1). For the regional analyses, the WMGHG and tropospheric aerosols are no longer necessarily correlated. In particular, sulfate aerosol emissions and thus forcing peaked in the 1980s for North America and

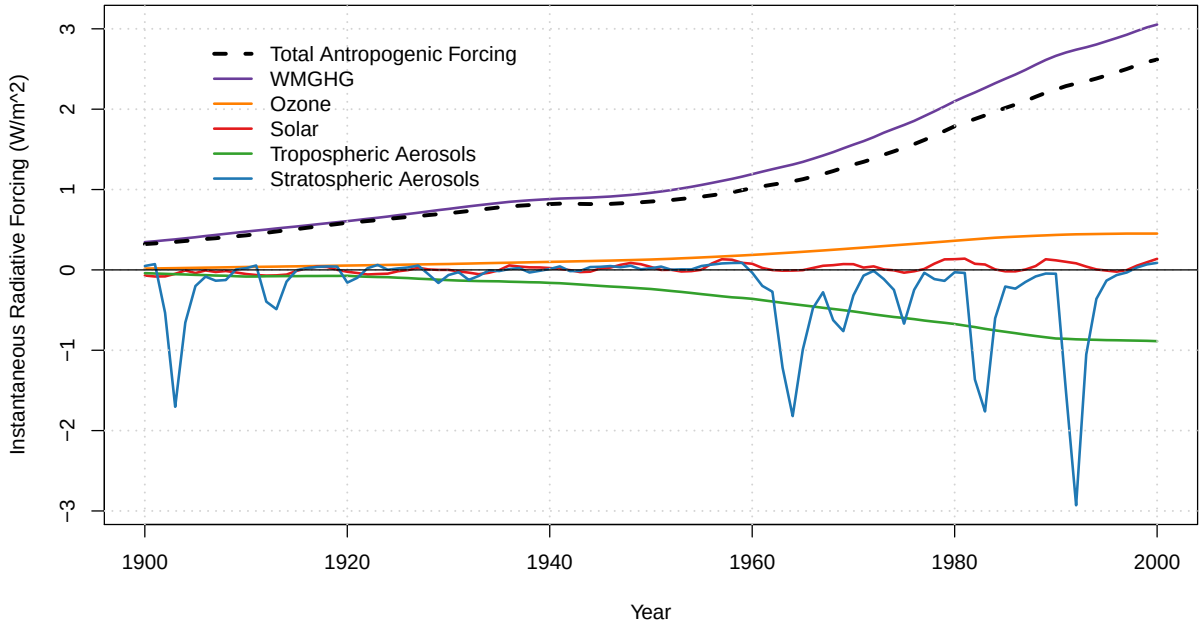

Figure 5.1: The annual iRF series over the time period of the study as calculated with the CMIP5 version of GISS Model E2 (Miller *et al.* 2014). The total anthropogenic forcing (dashed black line) is the sum of the well-mixed greenhouse gases (WMGHG), Ozone, and Tropospheric Aerosol forcings and reflects the total energy imbalance due to the three collinear anthropogenic emissions.

Europe, while the rest of the world has experienced increasing aerosol emissions (Bauer *et al.* 2020). Thus, the regional temperature decomposition will incoporate the regional tropospheric aerosol forcing as a predictor where the aerosol forcing in the annual iRF at the tropopause as estimated by GISS ModelE2.1 (Bauer *et al.* 2020; Miller *et al.* 2021). Following the global analysis, annual WMGHG and ozone forcings are combined into one predictor to reduce collinearity.

Other Radiative Forcings

Global annual radiative forcings due to stratospheric aerosols and variations in the solar output are also obtained from the (Miller *et al.* 2014) GISS Model E2 model output and shown in Figure 5.1. The solar radiative forcing reflects changes in solar irradiance which oscillates on approximately 11 year cycles (Wang *et al.* 2005). The stratospheric aerosol radiative forcing is primarily

due to large tropical volcanic eruptions. The largest negative peaks in the series are the eruptions of Santa Maria (1902), Agung (1963), El Chichon (1982), and Pinatubo (1991).

5.1.3 Indices of Climate Variability

While the eventual goal is to develop a model that selects relevant indices from sea surface temperature and pressure fields, a natural starting point for an analysis is to understand the predictive power of established climate indices. Following the model of Suckling *et al.* (2017), three annual SST-derived indices are used: one representing ENSO, the AMO, and the PDO.

ENSO Index: Niño 3.4

ENSO is the dominant interannual mode of variability in the climate system. While ENSO is primarily an equatorial Pacific phenomenon, it drives shifts in convection that affect climate globally through the propagation of Rossby and Kelvin waves. Incorporating ENSO as a predictor in any climate prediction system is of interest as it is a predictable phenomenon at lead times of up to one year.

A standardized monthly Niño 3.4 sea surface temperature (SST) index is created from the NOAA Physical Sciences Division (PSD) Niño 3.4 SST index derived from the Met Office Hadley Centre's sea ice and SST data set (HadISST1) (Rayner *et al.* 2003). The methodology for calculation of an annual index from the monthly index is addressed in Section 5.2.

AMO Index: North Atlantic Variability Index (NAVI)

Recent work by Haustein *et al.* (2019) suggests that the traditional AMO index of Delworth & Mann (2000) and subsequent improvements such as Trenberth & Shea (2006) and van Oldenborgh *et al.* (2009) fail to remove the entire anthropogenic warming signal and provide an overestimation of the correlation of Atlantic multidecadal variability with global mean temperature.

In lieu of these findings, the analysis uses the proposed NAVI of Haustein *et al.* (2019) which is the annual average SST at 40–60N and 15–50W minus annual average Northern Hemisphere

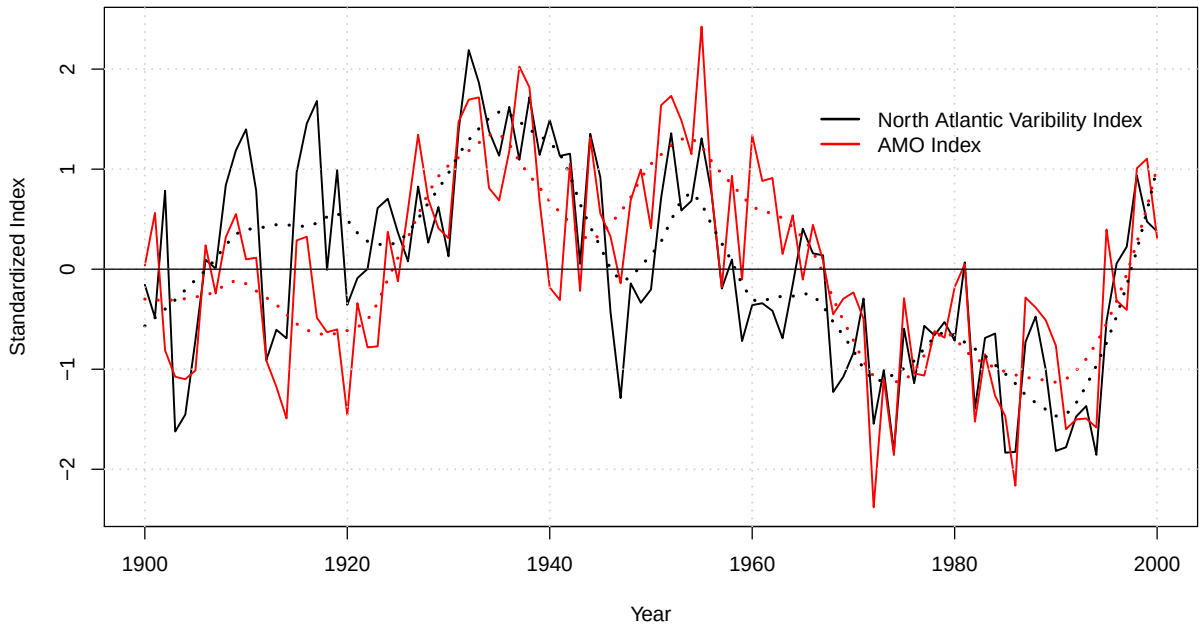

Figure 5.2: The NAVI of Haustein *et al.* (2019) and the AMO index of Trenberth & Shea (2006) over the time period of the study. The solid lines are the annual value of the index and the dotted lines show a loess smooth with a window equivalent to a 10-year moving average.

ocean temperature. The NAVI is similar to the more traditional AMO index from Trenberth & Shea (2006) which is defined as the average of Atlantic SST north of the equator where the global mean SST is subtracted from each grid-box prior to averaging. A comparison of these indices is shown in Figure 5.2. While the two indices roughly agree, there is a difference in sign in the early record prior to 1925 and a difference in magnitude of the oscillation from 1940–1970.

PDO Index

The PDO is the primary mode of variability in monthly North Pacific SST (Mantua *et al.* 1997; Mantua & Hare 2002). It can be viewed as the decadal-scale expression of ENSO throughout the entire Pacific basin (Newman *et al.* 2016). The index used to quantify the PDO is the standardized values of the leading principal component of monthly SST anomalies in the North Pacific Ocean, poleward of 20N (Mantua *et al.* 1997).

5.2 Methods

The first portion of the study is a replication of the global and regional attribution of annual surface temperature anomalies performed in Suckling *et al.* (2017). The second analysis is a refinement of the statistical methods to select optimal predictors for each location. The variable selection method provides a better representation of the relevant radiative forcings and modes of climate variability at the regional scale as modes of variability relevant to regional temperature variability, but not global temperature variability, can now be included. Results for the regional temperature attribution are presented over Central and South America to allow a deeper discussion, but can be trivially extended to the global land surface.

In both methods, the goal is to select a subset of predictors for optimal prediction of the annual mean temperature anomaly. There are six potential predictors as detailed in Section 5.1: three radiative forcing predictors (AF, stratospheric aerosol forcing, and solar forcing) and three indices of climate variability (ENSO, AMO, and PDO).

5.2.1 Replication of Suckling *et al.* (2017)

The empirical model of Suckling *et al.* (2017) is developed in two phases. First, a subset of predictors is determined by their ability to model the evolution of the global mean temperature. Then, this empirical model is fit at each grid-point independently to determine the contribution of the predictors to regional mean temperature.

Model of Global Mean Temperature

A global annual mean temperature anomaly series T is modeled by the multiple linear regression

$$T(t) = \beta_0 + \sum_{i=1}^{N} \beta_i x_i(t - \ell_i) + \epsilon \tag{5.1}$$

where the β_i are the regression parameters of interest, x_i are the predictor series, and the ϵ_i are the error terms assumed to be independent, identically distributed Gaussian errors. The ℓ_i allow a

lagged response in temperature to changes for each predictor.

The model is fit in two stages: first, the lag parameters are estimated through univariate linear regressions. That is, for the i^{th} parameter, a regression of the form

$$T(t) = \beta_0 + \beta_i x_i(t - \ell_i) + \epsilon_i \tag{5.2}$$

is fit for an array of possible ℓ values where $\ell > 0$ to restrict to physically relevant and potentially skillful lags. The ℓ value that maximizes the variance explained by the model is chosen. The two predictors that exhibit an non-zero lag for the global mean temperature are ENSO which was found to have a fourth month lag, that is, using the September-August annual mean, in agreement with the value found in Suckling *et al.* (2017) and the anthropogenic radiative forcing which was found to have a 12 year lag, slightly longer than the 10 year lag found in Suckling *et al.* (2017).

The second stage of the model fit is variable selection. Since there are only six predictors, all possible permutations are fit and the model with the highest Akaike information criterion (AIC) is selected (Akaike 1973). The optimal model contains the lagged anthropogenic forcing, the stratospheric aerosol forcing, the solar forcing, and the lagged ENSO index, in agreement with Suckling *et al.* (2017).

Initial Model of Regional Mean Temperature

The first model for regional historical temperature uses the four predictors selected for optimal prediction of the global mean temperature: the anthropogenic forcing, the stratospheric aerosol forcing, the solar forcing, and ENSO. The model presented Equation (5.1) is used where T now represents the mean temperature of a $0.5^\circ \times 0.5^\circ$ grid-box. Each grid-box uses the lag parameters estimated for the global mean as discussed in Section 5.2.1. While this study applies the regional model to the Central and South America region. However, there is no mathematical or computational reason to prevent the method from being extended to the entire global land surface.

The regional regressions are fit with standardized predictors allowing direct comparison of

the resulting coefficients. By scaling all predictor series to the same variance, comparisons of resulting estimates of the β_i directly show the relative importance of each parameter to predicting the evolution of regional temperature. This is worth the small loss of physicality as the goal of the study is to understand the relative effect of the predictors on temperature.

As with the global mean temperature model, this initial statistical model for regional temperature prediction and attribution follows the methodology of Suckling *et al.* (2017). The method does not incorporate predictors with limited spatial skill in a regional model if they do not provide predictive power in the global model. This is a problematic limitation as it has been shown that all of the SST-variability predictors have temperature teleconnections with limited spatial extent. These teleconnection are well studied for ENSO (Ropelewski & Halpert 1986; Ropelewski & Halpert 1987; Larkin & Harrison 2005), and has also been shown for the AMO (Ting *et al.* 2011; Danabasoglu *et al.* 2019) as well as the IPO (Dong & Dai 2015), primarily through its modulation of ENSO (Dong *et al.* 2018).

5.2.2 Selection of Optimal Predictors for Regional Temperature

To better capture these teleconnections with limited spatial extent the predictor set for the regression model in Equation (5.1) are selected for each grid-point series independently. This is opposed to the regional temperature model presented in Section 5.2.1 where the predictors were all chosen based on their ability to predict the global mean temperature. By allowing the best predictors to be selected at each location, information from the AMO and PDO can be utilized despite the indices not providing skill for the global mean. As with the initial regional model, the predictors are all standardized to simplify interpretation of the results.

The best set of predictors is selected using stepwise regression where the preferred model at each step is chosen by maximizing AIC. The AIC is used as it penalizes for overfitting, or the selection of too many parameters, and has a number of asymptotically optimal statistical properties. As this work proceeds the results found with stepwise regression should be confirmed with a regularized regression that constrains the predictor set as part of the model fit rather than as a *post-hoc*

adjustment. One such approach is the least absolute shrinkage and selection operator (LASSO) model.

In addition to changing the predictor subset, a method for regionally-varying lag coefficients is introduced. To avoid model overfit, lags were only allowed on the AF and ENSO predictors; the predictors that had non-zero lag in the global mean case.

5.3 Results and Discussion

The predicted global mean temperature anomaly matches the low and high frequency variability in the true global mean well (Figure 5.3a). The overall fit is good with a correlation of 0.93 and nearly all of the observed values falling inside the predictive 95% confidence interval. However, this model fails to reach the correlation of 0.95 found in Suckling *et al.* (2017) and falls quite below the correlation of over 0.99 archived by the impulse model method of Haustein *et al.* (2019), albeit with series that had been low-pass filtered.

The remaining panels of Figure 5.3 show the warming or cooling contribution of each of the selected predictors over time. As expected, the response to the anthropogenic forcing (AF) is greatest (Figure 5.3b), accounting for over a degree of warming over the century. The stratospheric aerosol response is also relevant at the global scale with large cooling spikes aligning with major volcanic eruptions (Figure 5.3c). The anomaly due to changes in solar irradiance is more minor (Figure 5.3d), but the response to ENSO (Figure 5.3e) is large and highlights it as a source of interannual variability in the global mean temperature. In general, the results presented are in close agreement with those found in (Suckling *et al.* 2017)

An estimate of the transient climate response (TCR) is made by fitting the global model with separate predictors for WMGHG and tropospheric aerosols and assuming the forcing resulting from a doubling of CO_2 is 3.7 Wm^{-2} (Myhre *et al.* 2017). The modeled response to WMGHG is found to be 0.37 ± 0.09 resulting in a TCR of 1°C–1.7°C. This value is lower than the IPCC accepted range of 1.5°C to 2.8°C, but may be more reasonable according to studies suggesting this range is high (Otto *et al.* 2013; Shindell 2014).

The second portion of the Suckling *et al.* (2017) replication is the confirmation of the regional results. The AF and stratospheric aerosol forcing estimates (Figures 5.4a and 5.4d) exhibit physically consistent values at all locations; a positive forcing corresponds to an increase in temperature. In addition, they both have significant coefficients nearly everywhere respectively showing their skill in explaining the 20^{th} century temperature evolution regionally.

The solar forcing (Figure 5.4c) does not follow a physically consistent pattern everywhere with a large, significant negative patch in southern Brazil. Such a result most likely suggests an ill-posed statistical model, rather than an increased solar irradiance leading to a widespread cooling as the negative coefficient aligns with the area of decreased declining model fit as shown by the adjusted-R^2 field (Figure 5.5a). The ENSO response agrees with the general understanding of temperature teleconnections (Figure 5.4b), showing general warming under positive (El Niño) conditions with significant cooling in the southern USA and northern Mexico.

The sparse, but generally spatially coherent patterns of significance in the solar forcing (Figure 5.4c) and stratospheric aerosol forcing (Figure 5.4d) coefficients provide motivation and justification for the variable selection method for modeling regional temperature. If the globally-optimal predictors are not useful, they should not be included in the statistical model. The variable selection method provides estimates of the important predictors to temperature variability, ideally leading to more robust attribution of temperature variability.

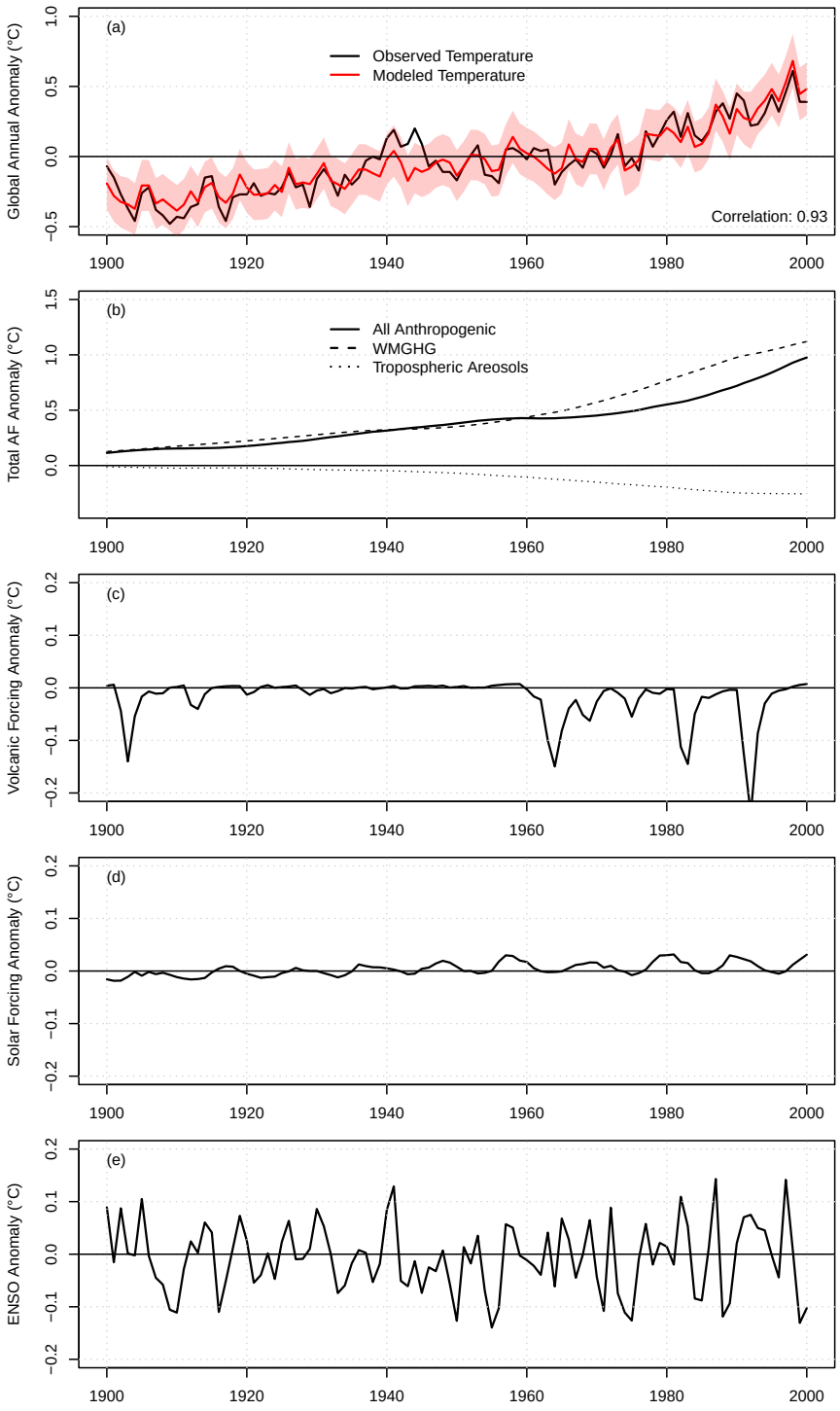

Figure 5.3: (a) The global mean temperature according to the Berkeley Earth analysis and the predicted global mean temperature with the attribution model. (b)–(e) The effect series of the four predictors. Note that the scales of (b) and (c)–(e) are different to account for the large response to the AF Forcing.

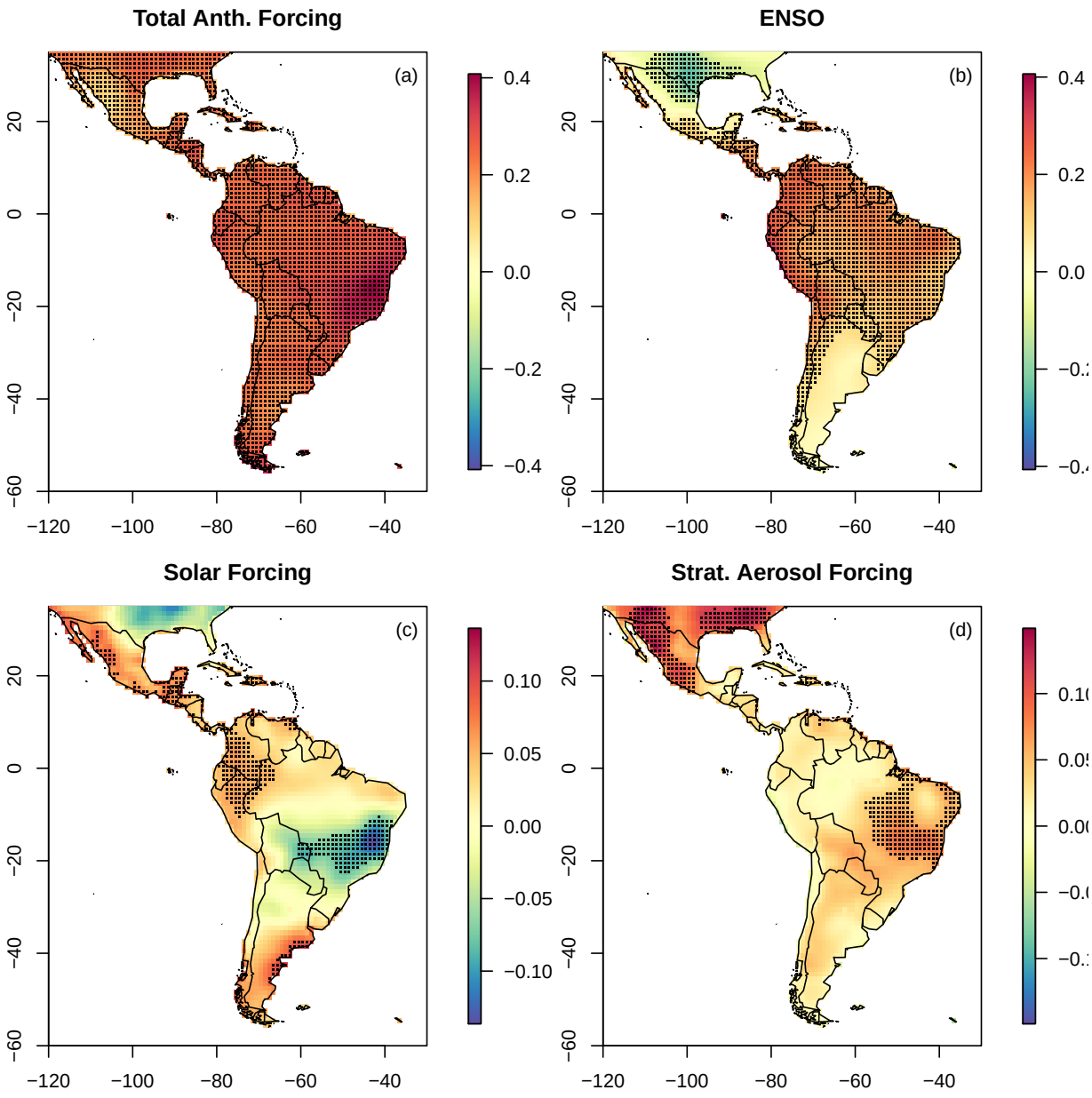

Figure 5.4: The standardized regression coefficients for predicting regional temperature from the Suckling *et al.* (2017) regional model for (a) the total anthropogenic forcing (b) ENSO (c) the solar forcing and (d) the stratospheric aerosol forcing. Stippling indicates the coefficient is significant at the 0.05 level. Note that the range for (a) and (b) differ from the range on (c) and (d).

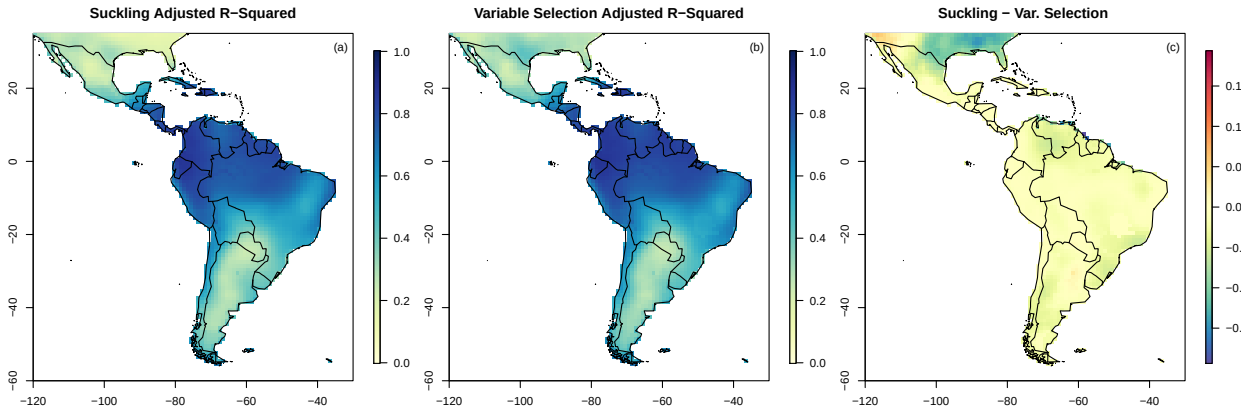

Figure 5.5: The adjusted-R^2 statistic for the (a) Suckling *et al.* (2017) regional model and (b) variable selection model. (c) The difference of (a)-(b) with cool colors indicating an increase in performance with the variable selection model.

In agreement with the results of the initial regional model, each predictor in the variable selection model shows coherent spatial clusters (Figure 5.6). The relevant predictors are more clear in the model selection case as predictors are only included when the variance they explain is worth the increased complexity of the model. While these predictors may provide more physical intuition into the drivers of temperature variability, the model with selected predictors does not provide a substantial improvement in terms of the variance explained (Figure 5.5) other than the southern US where the addition of the AMO predictor explains around 10% more variance. In general, both regional models show higher skill in the tropics and lower skill in the extratropics in agreement with seasonal forecasts (Barnston *et al.* 2010b; Scaife *et al.* 2019).

As with the Suckling *et al.* (2017) regional model, the AF and ENSO forcings are associated with temperature nearly everywhere (Figures 5.6a,d). Despite the stratospheric aerosols having a large effect on the global mean, there appears to be only some regions where cooling in the regional mean is linked to large volcanic eruptions. More work is needed into understanding if this temperature response to stratospheric aerosols is linked to some physical process or an artifact of an insufficient statistical model (Figure 5.6b). The solar irradiance forcing is again non-physical in the same band across Brazil suggesting that is some correlated process needed to properly model temperature in that region (Figure 5.6c).

Of particular interest are the coherent clusters in the AMO and PDO which suggest possible teleconnections. The AMO coefficient (Figure 5.6e) shows that a positive phase of the AMO leads to a warming in the USA/Mexico region, a weak warming in northern Brazil, and some cooling in the Uruguay region. The USA/Mexico warming agrees with the results found in (Knight *et al.* 2006; Ting *et al.* 2011). While there is a large body of working linking the AMO to precipitation in northeastern Brazil through modulation of the Atlantic dipole (Hastenrath & Heller 1977), a relationship between temperature and the AMO is less clear (Knight *et al.* 2006). No direct support could be found linking a positive AMO with a cooling in the Uruguay region other than weak evidence in the pre-industrial control runs analyzed in (Ting *et al.* 2011). Additionally, a mechanism for the AMO controlling the ENSO teleconnection over Uruguay is presented in (Kayano &

Capistrano 2014) where a positive AMO leads to decreased wet anomalies during El Niño years.

There are two potential PDO-linked teleconnections that suggest a positive PDO leads to warming in southern Brazil and cooling around the Gulf of Mexico. However, the global analysis of Dong & Dai (2015) does not agree, finding incoherent patterns in both regions. However, the Dong & Dai (2015) study uses a statistical model that only controls for ENSO and the warming trend. They find PDO-driven variability in regions that the variable selection model associates with the AMO such as southeastern Brazil and the southern US. Regardless of whether these variations are driven by the Atlantic or Pacific, the comparison of these results provides a cautionary tale for avoiding confounding when utilizing regression-based approaches to associate variability with specific climate modes.

Visual inspection of the coefficients from both regression models (Figures 5.4 and 5.6) show numerous artifacting patterns, likely due to data quality and availability issues. There are a few 'water drop' patterns in Brazil as well as an artificially sharp line across northern Chile and Argentina. The water drop appears to be from a single bad station and may be able to be corrected. The source of the sharp line is less clear, but is likely due to data sparsity in the early record.

Exploration of residual series showed that a potential low-frequency signal was missed by the linear regression model. A quadratic curve was fit to the residuals to determine if there were regions where the linear statistical model may be doing a poor job of capturing the functional form of temperature (Figure 5.7). A large area from Mexico through eastern equatorial South America shows a negative quadratic behavior in the residuals. Looking a little deeper into series in Mexico and Colombia (Figure 5.8 with locations shown in Figure 5.7c) shows that the current statistical model is unable to capture the large temperature modulation in the 1940s. A similar poor performance would be expected when the model is extended to include the warming slowdown of the 2000s. Interestingly, this quadratic pattern in the residuals does not directly correspond with the model's ability to explain variance as measure by adjusted-R^2 (Figure 5.5b).

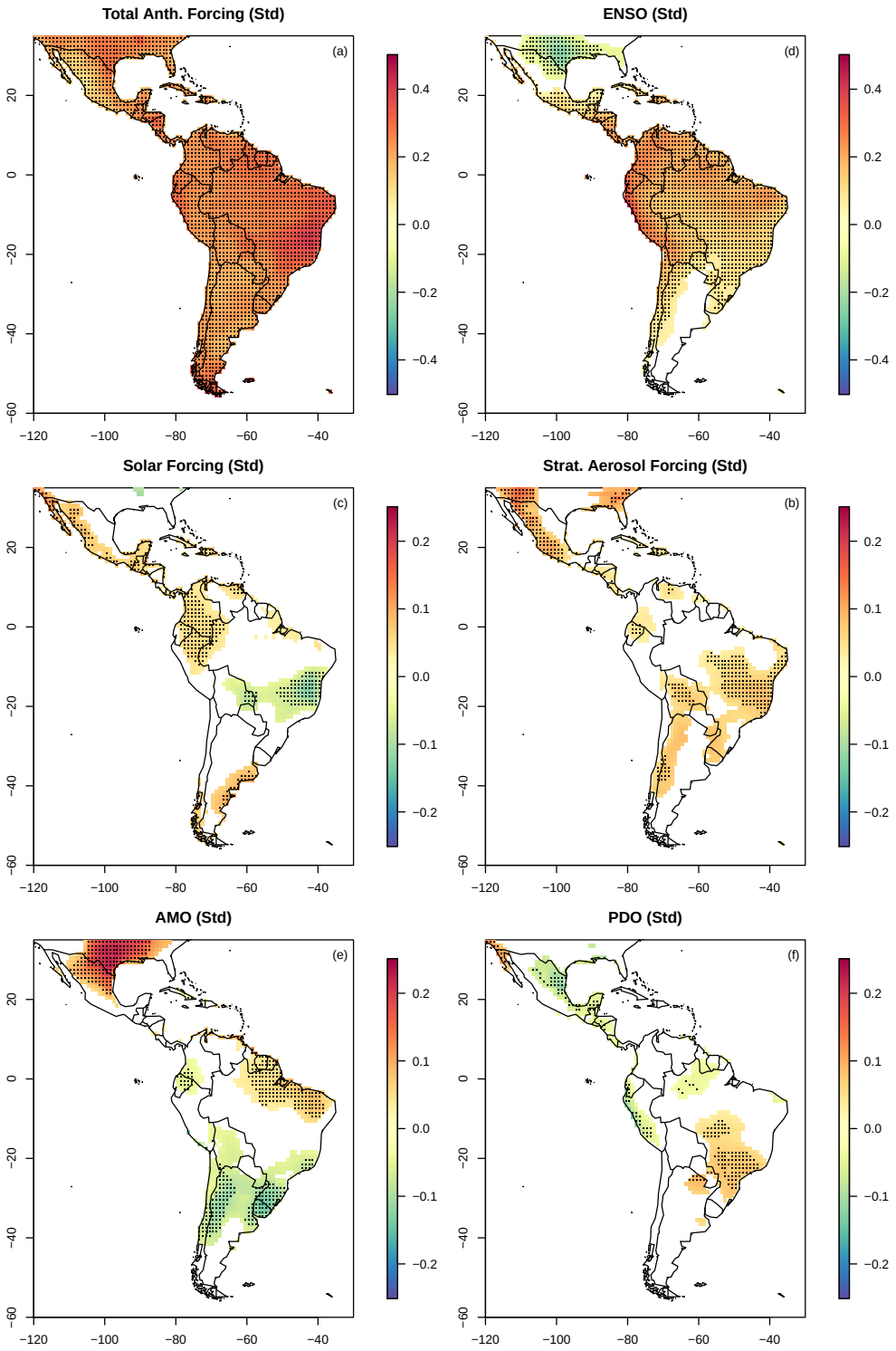

Figure 5.6: The standardized regression coefficients for predicting regional temperature from the variable selection regional model. White regions over land indicate a variable did not explain sufficient variance at that location. Stippling indicates the coefficient is significant at the 0.05 level. Note that the range for (a) and (b) differ from the range on (c)-(f).

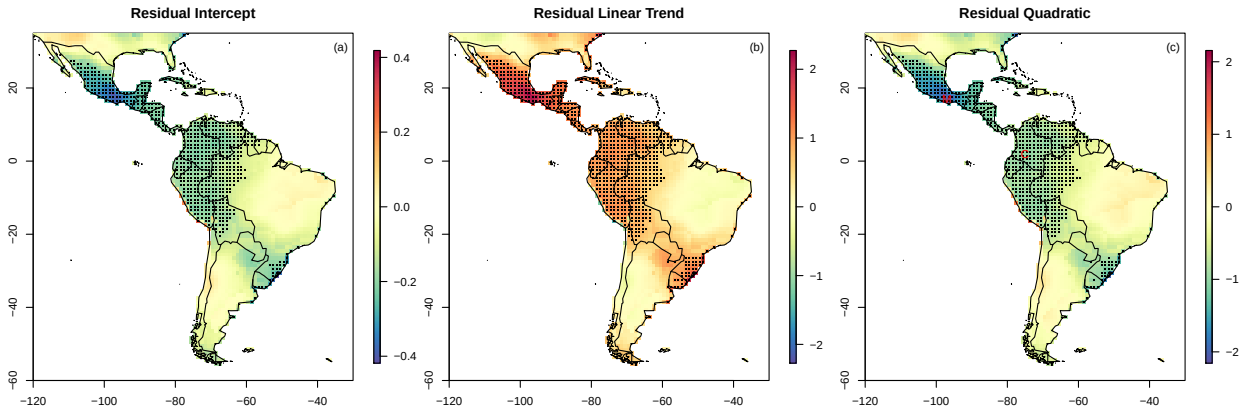

Figure 5.7: The coefficients of a quadratic fit on the residuals from the variable selection regional model. Stippling indicates the coefficient is significant at the 0.05 level and suggest that the residuals are not random, mean zero Gaussian noise. The 'C' and 'M' in (c) indicate the locations of the Colombia and Mexico series shown in Figure 5.8.

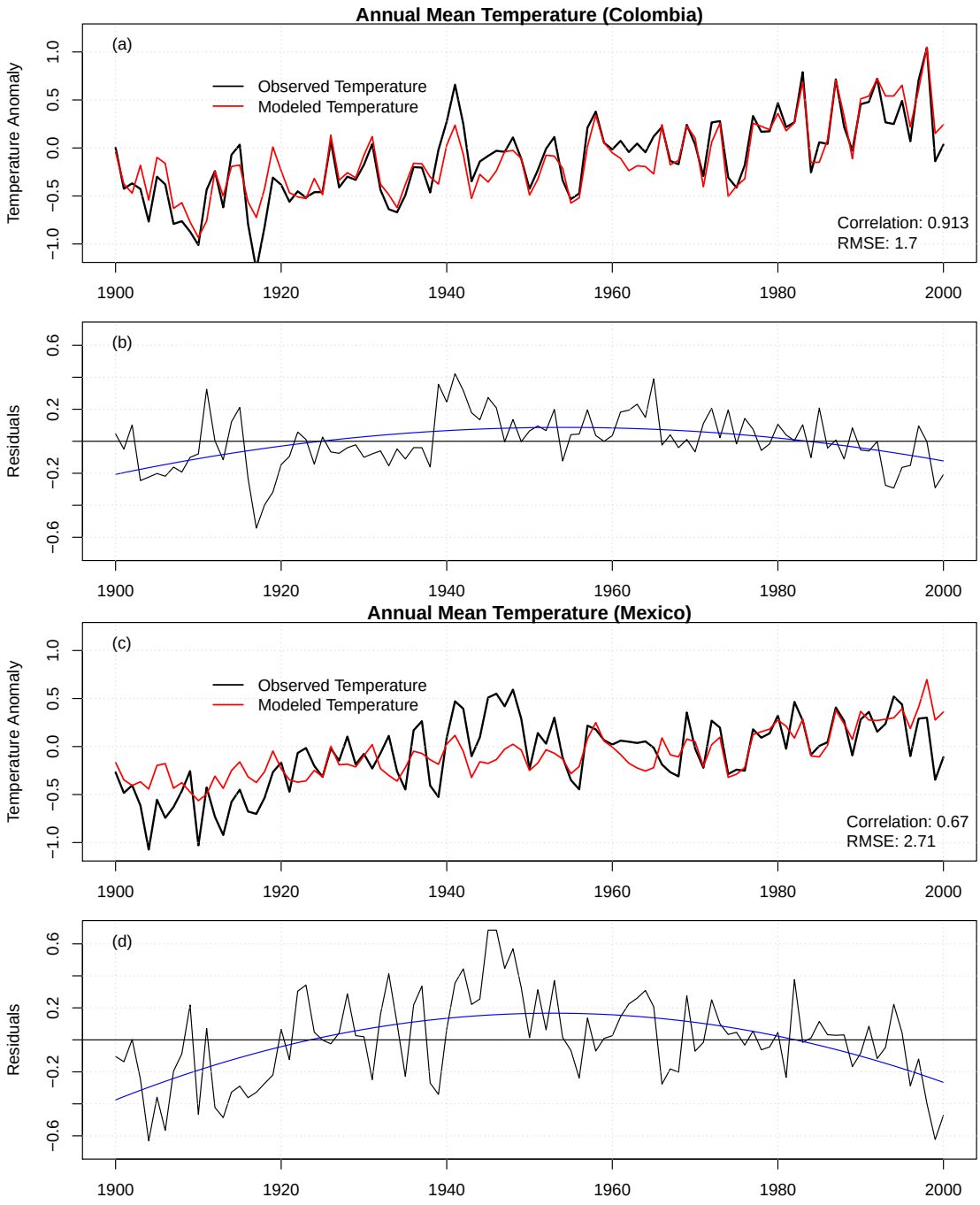

Figure 5.8: The observed and modeled temperature for a grid-box in (a) Colombia and (c) Mexico. The respective residuals with the quadratic fit shown in Figure 5.7 are shown in Figures (b) and (d).

5.4 Conclusions and Future Work

The findings presented show promise for the attribution of interannual to decadal variability in temperature. A replication of Suckling *et al.* (2017) provides a baseline for subsequent development as well as a deep understanding of the current state of the field. The spatial heterogeneity of predictors used in variable selection model suggest that a "one size fits all" model for attributing regional temperature variability is incomplete. However, the lack of improvement in variance explained motivates more creative thinking for the fundamental model used to predict the evolution of temperature.

An exciting next step is the application of these attribution methods to climate model output. Doing so will achieve a similar goal as the observational study, but for the climate of the model as opposed to the climate of the true earth. Through attributing low-frequency variability to dynamical models and comparing to the observational results, a deeper understanding of the limitations of models for decadal prediction can be gained. Additionally, larger datasets from large ensembles and decadal prediction hindcasts can be leveraged to increase the sample size and therefore the power and/or complexity of statistical models for attribution and prediction.

Tandem to the proposed work with climate model output will be an investigation of the potential AMO and PDO teleconnections discovered here. The attribution model presented can suggest potential teleconnections, but dynamical support needs to be found through climate model experiments. Recent Atlantic pacemaker studies may be useful in addition to the experiments mentioned above.

As the statistical model development progresses, a priority is a more sophisticated model for incorporation of anthropogenic forcings. Combining these forcings is useful in the regression model, but obscures physical relationships of interest. The GISS experiment also provides a spatial iRF from anthropogenic aerosols that is currently not utilized, and could be useful for prediction of regional temperature.

The choices of smoothing and filtering are difficult questions for the detection and quantifi-

cation of decadal climate variability. Some amount of smoothing is inherent in any climate data: observations are aggregated over space and time to create climatological averages. This problem becomes more complex as slowly evolving variability is sought. Some work has gone into investigating the effects of temporal filtering and smoothing such as Narapusetty *et al.* (2009) and Cane *et al.* (2017), but there is little into the role of spatial aggregation and smoothing on quantifying the climate.

Conclusion

This five chapters of this dissertation provided insights into three fundamental questions in modern climate science: what is the natural variability of the climate; how predictable is this variability; and how does climate change affect variability and predictability? Two broad categories of data are available to investigate the Earth's climate: a spatially and temporally limited and error-filled observational record of past climate and weather, and imperfect simulations using dynamical climate models. In addressing these questions, each chapter encountered at least one of the two core challenges in making physical or statistical inferences about the climate system.

Chapters 1 and 2 presented a complete uncertainty analysis of the NASA GISTEMP observational global surface temperature product. The described work enhances the GISTEMP product to be in-line with other cutting edge surface temperature products, to provide critical validation of key statistics of global warming. The global surface temperature record is an essential tool for understanding climate variability and change as the longest, global direct observational data and uncertainty quantification is necessary for defining the limits of these data. The uncertainty ensemble implemented and analyzed in Chapter 2 is an exciting tool for scientists who utilize historical temperature data in their analyses. The example analyses in Chapters 1 and 2 show how including observational uncertainty is necessary to understand the entire picture when asking common questions in applied climate science such as, "What is the warmest year on record?" and "How much more quickly is the Arctic warming than the Earth?"

The uncertainty ensemble of Chapter 2 opens up many potential future studies to investigate the impact of observational uncertainty on established and new questions that rely on observed

temperature data. Among future work to be taken following this dissertation, it will be useful to revisit the variability attribution in Chapter 5, incorporating observational uncertainty. Showing that a radiative effect or teleconnection is robust to observational uncertainty can direct future modeling studies to identify the thermodynamic or dynamical causal link between the predictor and resulting interannual temperature variability.

Chapters 3 and 4 provided new information about the predictability and global impact of the El Niño-Southern Oscillation (ENSO), the largest source of interannual climate variability. Chapter 3 showed the global impact of ENSO on seasonal precipitation and provides maps at varying information levels detailing these teleconnections for decision makers. The predictive skill of these maps is quantified and shows that historical ENSO impacts provide more useful information than the absence of a forecast does, but also that state-of-the-art calibrated dynamical forecast systems remain generally superior. The widespread ENSO teleconnections motivated Chapter 4 where multi-year prediction of ENSO was investigated. Initialization shock, or non-physical dynamical processes arising from incompatible initial conditions, is a major issue in modern forecasting systems. The model-analogue method explored in Chapter 4 avoids initialization shock and shows promising performance in predicting La Niña events two years in advance.

Chapter 4 presented some promising results and has opened up many interesting questions. As the study is expanded to include more CGCMs, it will be interesting to see how the model ENSO simulation dynamics relate to predictability in both initialized and model-analogue prediction. In addition, it is worth revisiting the traditional ENSO forecast paradigms of either predicting ENSO events or predicting the time-evolution of some SST index. Determining when a long lead forecast is confident is a holy grail of forecasting and can be explored through an extension of the hindcast study presented. Finally, the dynamical sources of long-lead ENSO prediction will be investigated to determine if there are certain characteristics of ENSO events and/or CGCM systems that lend themselves to greater predictive skill at leads greater than 12 months.

Chapter 5 used a regression-based method to decompose historical annual temperature into sources associated with radiative forcings and sources associated with known modes of natural

variability. It confirms findings that greenhouse gasses are necessary to explain observed warming at regional and global scales. In addition, teleconnections of potentially predictable modes of variability such as ENSO and Atlantic multidecadal variability are revealed. As discussed above, a natural next step in this study is to replicate using the uncertainty ensemble presented in Chapter 2 to determine the robustness of the results to the now well described observational uncertainty.

As a single work, this dissertation identified three monumental questions in climate science and provides examples of how physics, statistics, and computer science are critical for answering them. The three fundamental questions posed will occupy the field for the next decades or centuries and will rely on advances in our understanding of the physical system, novel statistical and machine learning methods for making sense of exabytes of data, and computer software and hardware for running and analyzing the necessary dynamical and statistical models. As humanity faces rapidly more devastating impacts of climate change throughout the world, a better understanding of how consequential it will be on human and natural systems is critical. In tandem, the potential for improved climate and weather prediction provides optimism. With better knowledge of upcoming weather and climate extremes at all timescales, humans can prepare and protect those who are most at risk.

References

1. Adler, R. F. *et al.* The Version-2 Global Precipitation Climatology Project (GPCP) Monthly Precipitation Analysis (1979–Present). *Journal of Hydrometeorology* **4,** 1147–1167 (2003).

2. Agresti, A. *Categorical data analysis* 1st, 558 (Wiley, New York [u.a.], 1990).

3. Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. *Proceedings of the 2nd International Symposium on Information Theory,* 267–281 (1973).

4. Alexander, M. A. *et al.* The Atmospheric Bridge: The Influence of ENSO Teleconnections on Air–Sea Interaction over the Global Oceans. *Journal of Climate* **15,** 2205 –2231 (2002).

5. Balmaseda, M. & Anderson, D. Impact of initialization strategies and observations on seasonal forecast skill. *Geophysical Research Letters* **36** (2009).

6. Balmaseda, M. A., Mogensen, K. & Weaver, A. T. Evaluation of the ECMWF ocean reanalysis system ORAS4. *Quarterly journal of the royal meteorological society* **139,** 1132–1161 (2013).

7. Barlow, M., Cullen, H. & Lyon, B. Drought in Central and Southwest Asia: La Niña, the Warm Pool, and Indian Ocean Precipitation. *Journal of Climate* **15,** 697–700 (2002).

8. Barnes, N. & Jones, D. Clear Climate Code: Rewriting Legacy Science Software for Clarity. *IEEE Software* **28,** 36–42 (2011).

9. Barnston, A. G. *et al.* Verification of the First 11 Years of IRI's Seasonal Climate Forecasts. *Journal of Applied Meteorology and Climatology* **49,** 493–520 (2010).

10. Barnston, A. G. *et al.* Verification of the First 11 Years of IRI's Seasonal Climate Forecasts. *Journal of Applied Meteorology and Climatology* **49,** 493–520 (2010).

11. Barnston, A. G., Tippett, M. K., Ranganathan, M. & L'Heureux, M. L. Deterministic skill of ENSO predictions from the North American Multimodel Ensemble. *Climate Dynamics* **53,** 7215–7234 (2019).

12. Bauer, S. E. *et al.* Historical (1850–2014) Aerosol Evolution and Role on Climate Forcing Using the GISS ModelE2.1 Contribution to CMIP6. *Journal of Advances in Modeling Earth Systems* **12.** e2019MS001978 2019MS001978, e2019MS001978 (2020).

13. Bellenger, H., Guilyardi, E., Leloup, J., Lengaigne, M. & Vialard, J. ENSO representation in climate models: from CMIP3 to CMIP5. *Climate Dynamics* **42,** 1999–2018 (2014).

14. Bjerknes, J. Atmospheric teleconnections from the equatorial Pacific. *Monthly weather review* **97,** 163–172 (1969).

15. Boer, G. J. *et al.* The decadal climate prediction project (DCPP) contribution to CMIP6. *Geoscientific Model Development* **9,** 3751–3777 (2016).

16. Borbor-Mendoza, M. *et al.* in *WHO/WMO Climate Services for Health* (eds Shumake-Guillemot, J & Fernandes-Montoya, L) 108–109 (WHO/WMO, Geneva, 2016).

17. Brier, G. W. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review* **78,** 1–3 (1950).

18. Brohan, P., Kennedy, J. J., Harris, I., Tett, S. F. B. & Jones, P. D. Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *Journal of Geophysical Research* **111,** D12106 (2006).

19. Brönnimann, S., Xoplaki, E., Casty, C., Pauling, A. & Luterbacher, J. ENSO influence on Europe during the last centuries. *Climate Dynamics* **28,** 181–197 (2007).

20. Budyko, M. I. The effect of solar radiation variations on the climate of the Earth. *Tellus* **21,** 611–619 (1969).

21. Callendar, G. S. The artificial production of carbon dioxide and its influence on temperature. *Quarterly Journal of the Royal Meteorological Society* **64,** 223–240 (1938).

22. Callendar, G. S. Temperature fluctuations and trends over the earth. *Quarterly Journal of the Royal Meteorological Society* **87,** 1–12 (1961).

23. Cane, M. A., Zebiak, S. E. & Dolan, S. C. Experimental forecasts of El Niño. *Nature* **321,** 827–832 (1986).

24. Cane, M. A., Clement, A. C., Murphy, L. N. & Bellomo, K. Low-Pass Filtering, Heat Flux, and Atlantic Multidecadal Variability. *Journal of Climate* **30,** 7529–7553 (2017).

25. Carleton, T. A. *et al. Valuing the global mortality consequences of climate change accounting for adaptation costs and benefits* tech. rep. (National Bureau of Economic Research, 2020).

26. Cash, D. W. & Buizer, J. Knowledge–Action Systems for Seasonal to Interannual Climate Forecasting:Summary of a Workshop. *National Academy Press, Washington DC* (2005).

27. Cassou, C. *et al.* Decadal Climate Variability and Predictability: Challenges and Opportunities. *Bulletin of the American Meteorological Society* **99,** 479–490 (2018).

28. Cayan, D. R., Redmond, K. T. & Riddle, L. G. ENSO and Hydrologic Extremes in the Western United States. *Journal of Climate* **12,** 2881–2893 (1999).

29. Changnon, S. A. *El Niño 1997-1998: the climate event of the century* (Oxford University Press, 2000).

30. Cohen, J. *et al.* Recent Arctic amplification and extreme mid-latitude weather. *Nature Geoscience* **7,** 627 EP – (2014).

31. Copernicus Climate Change Service (C3S). *ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate.* Copernicus Climate Change Service Climate Data Store (CDS). 2017.

32. Corti, S. *et al.* Impact of Initial Conditions versus External Forcing in Decadal Climate Predictions: A Sensitivity Experiment. *Journal of Climate* **28,** 4454–4470 (2015).

33. Cowtan, K. & Way, R. G. Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Quarterly Journal of the Royal Meteorological Society* **140,** 1935–1944 (2014).

34. Cowtan, K. *et al.* Robust comparison of climate models with observations using blended land air and ocean sea surface temperatures. *Geophysical Research Letters* **42,** 6526–6534 (2015).

35. Cressie, N. *Statistics for spatial data* (John Wiley & Sons, 2015).

36. Crochemore, L., Ramos, M.-H. & Pappenberger, F. Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. *Hydrology and Earth System Sciences* **20,** 3601–3618 (2016).

37. Danabasoglu, G., Landrum, L., Yeager, S. G. & Gent, P. R. Robust and Nonrobust Aspects of Atlantic Meridional Overturning Circulation Variability and Mechanisms in the Community Earth System Model. *Journal of Climate* **32,** 7349–7368 (2019).

38. Danabasoglu, G. *et al.* The Community Earth System Model version 2 (CESM2). *Journal of Advances in Modeling Earth Systems* **12,** e2019MS001916 (2020).

39. Delworth, T. L. & Mann, M. E. Observed and simulated multidecadal variability in the Northern Hemisphere. *Climate Dynamics* **16,** 661–676 (2000).

40. Deser, C. *et al.* Insights from Earth system model initial-condition large ensembles and future prospects. *Nature Climate Change* **10,** 277–286 (2020).

41. DiNezio, P. N., Deser, C., Okumura, Y. & Karspeck, A. Predictability of 2-year La Niña events in a coupled general circulation model. *Climate dynamics* **49,** 4237–4261 (2017).

42. Ding, H., Newman, M., Alexander, M. A. & Wittenberg, A. T. Skillful climate forecasts of the tropical Indo-Pacific Ocean using model-analogs. *Journal of Climate* **31,** 5437–5459 (2018).

43. Ding, H., Newman, M., Alexander, M. A. & Wittenberg, A. T. Diagnosing secular variations in retrospective ENSO seasonal forecast skill using CMIP5 model-analogs. *Geophysical Research Letters* **46,** 1721–1730 (2019).

44. Ding, H., Newman, M., Alexander, M. A. & Wittenberg, A. T. Relating CMIP5 model biases to seasonal forecast skill in the tropical Pacific. *Geophysical Research Letters* **47,** e2019GL086765 (2020).

45. Dolgin, E. Climate change: As the ice melts. *Nature* **543,** S54–S55 (2017).

46. Dong, B. & Dai, A. The influence of the Interdecadal Pacific Oscillation on Temperature and Precipitation over the Globe. *Climate Dynamics* **45,** 2667–2681 (2015).

47. Dong, B., Dai, A., Vuille, M. & Timm, O. E. Asymmetric Modulation of ENSO Teleconnections by the Interdecadal Pacific Oscillation. *Journal of Climate* **31,** 7337–7361 (2018).

48. Dunstone, N., Smith, D., Yeager, S., Danabasoglu, G., *et al.* Skilful interannual climate prediction from two large initialised model ensembles. *Environmental Research Letters* **15,** 094083 (2020).

49. Epstein, E. S. A Scoring System for Probability Forecasts of Ranked Categories. *Journal of Applied Meteorology* **8,** 985–987 (1969).

50. FAO. Temperature change statistics 1961-2021: Global, regional, and country trends. *FAOSTAT Analytical Brief Series* (2022).

51. FEWSNet. Worsening drought threatens Horn of Africa as conflict-driven emergency persists in northern Ethiopia. `https://fews.net/east-africa/alert/october-27-2021` (2021).

52. Fischer, E. M. & Knutti, R. Observed heavy precipitation increase confirms theory and early models. *Nature Climate Change* **6,** 986–991 (2016).

53. Fisher, R. A. The Logic of Inductive Inference. *Journal of the Royal Statistical Society* **98,** 39–82 (1935).

54. Freeman, E. *et al.* ICOADS Release 3.0: a major update to the historical marine climate record. *International Journal of Climatology* **37,** 2211–2232 (2016).

55. Garreaud, R. & Aceituno, P. Interannual Rainfall Variability over the South American Altiplano. *Journal of Climate* **14,** 2779–2789 (2001).

56. Gelaro, R. *et al.* The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *Journal of Climate* **30,** 5419–5454 (2017).

57. Ghil, M. Natural climate variability. *Encyclopedia of global environmental change* **1,** 544–549 (2002).

58. Gill, A. E. Some simple solutions for heat-induced tropical circulation. *Quarterly Journal of the Royal Meteorological Society* **106,** 447–462 (1980).

59. Goddard, L., Barnston, A. G. & Mason, S. J. Evaluation of the IRI's "Net Assessment" Seasonal Climate Forecasts: 1997–2001. *Bulletin of the American Meteorological Society* **84,** 1761–1782 (2003).

60. Goddard, L. *et al.* A verification framework for interannual-to-decadal predictions experiments. *Climate Dynamics* **40,** 245–272 (2013).

61. Goddard, L. Climate Change Modeling Methodology: Selected Entries from the Encyclopedia of Sustainability Science and Technology. *Climate Predictions, Seasonal-to-Decadal* (ed Rasch, P. J.) 261–301 (2012).

62. Goddard, L. & Dilley, M. El Niño: Catastrophe or Opportunity. *Journal of Climate* **18,** 651–665 (2005).

63. Goddard, L. & Philander, S. G. The energetics of El Niño and La Niña. *Journal of climate* **13,** 1496–1516 (2000).

64. Goddard, L. *et al.* Current approaches to seasonal to interannual climate predictions. *International Journal of Climatology: A Journal of the Royal Meteorological Society* **21,** 1111–1152 (2001).

65. Gonzalez, P. L. M. & Goddard, L. Long-lead ENSO predictability from CMIP5 decadal hindcasts. *Climate Dynamics* **46,** 3127–3147 (2016).

66. Grimm, A. M. The El Niño Impact on the Summer Monsoon in Brazil: Regional Processes versus Remote Influences. *Journal of Climate* **16,** 263–280 (2003).

67. Grimm, A. M., Ferraz, S. E. T. & Gomes, J. Precipitation Anomalies in Southern Brazil Associated with El Niño and La Niña Events. *Journal of Climate* **11,** 2863–2880 (1998).

68. Gulev, S. *et al.* in *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (eds MassonDelmotte V., P., Zhai, A., Pirani, S., *et al.*) (Cambridge University Press, 2021).

69. Ham, Y.-G., Kug, J.-S. & Park, J.-Y. Two distinct roles of Atlantic SSTs in ENSO variability: North tropical Atlantic SST and Atlantic Niño. *Geophysical Research Letters* **40,** 4012–4017 (2013).

70. Ham, Y.-G., Kim, J.-H. & Luo, J.-J. Deep learning for multi-year ENSO forecasts. *Nature* **573,** 568–572 (2019).

71. Han, Z. *et al.* Simulation by CMIP5 models of the Atlantic multidecadal oscillation and its climate impacts. *Advances in Atmospheric Sciences* **33,** 1329–1342 (2016).

72. Hansen, J. *et al.* Climate Impact of Increasing Atmospheric Carbon Dioxide. *Science* **213,** 957–966 (1981).

73. Hansen, J., Ruedy, R., Glascoe, J. & Sato, M. GISS analysis of surface temperature change. *Journal of Geophysical Research: Atmospheres* **104,** 30997–31022 (1999).

74. Hansen, J. *et al.* A closer look at United States and global surface temperature change. *Journal of Geophysical Research: Atmospheres* **106,** 23947–23963 (2001).

75. Hansen, J. *et al.* Efficacy of climate forcings. *Journal of Geophysical Research: Atmospheres* **110** (2005).

76. Hansen, J. *et al.* Climate simulations for 1880–2003 with GISS modelE. *Climate Dynamics* **29,** 661–696 (2007).

77. Hansen, J., Ruedy, R., Sato, M. & Lo, K. Global Surface Temperature Change. *Reviews of Geophysics* **48.** RG4004 (2010).

78. Hansen, J. & Lebedeff, S. Global trends of measured surface air temperature. *Journal of Geophysical Research: Atmospheres* **92,** 13345–13372 (1987).

79. Hansen, J. W. Integrating seasonal climate prediction and agricultural models for insights into agricultural practice. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360,** 2037–2047 (2005).

80. Hansen, J. W., Challinor, A., Ines, A., Wheeler, T. & Moron, V. Translating climate forecasts into agricultural terms: advances and challenges. *Climate Research* **33,** 27–41 (2006).

81. Harris, I., Jones, P., Osborn, T. & Lister, D. Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 Dataset. *International Journal of Climatology* **34,** 623–642 (2014).

82. Harris, I., Osborn, T. J., Jones, P. & Lister, D. Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Scientific Data* **7,** 109 (2020).

83. Hastenrath, S. & Heller, L. Dynamics of climatic hazards in northeast Brazil. *Quarterly Journal of the Royal Meteorological Society* **103,** 77–92 (1977).

84. Hausfather, Z. *et al.* Quantifying the effect of urbanization on U.S. Historical Climatology Network temperature records. *Journal of Geophysical Research: Atmospheres* **118,** 481–494 (2013).

85. Hausfather, Z., Drake, H. F., Abbott, T. & Schmidt, G. A. Evaluating the performance of past climate model projections. *Geophysical Research Letters* **47,** e2019GL085378 (2020).

86. Haustein, K. *et al.* A Limited Role for Unforced Internal Variability in Twentieth-Century Warming. *Journal of Climate* **32,** 4893–4917 (2019).

87. Hawkins, E. & Jones, P. D. On increasing global temperatures: 75 years after Callendar. *Quarterly Journal of the Royal Meteorological Society* **139,** 1961–1963 (2013).

88. Hawkins, E., Robson, J., Sutton, R., Smith, D. & Keenlyside, N. Evaluating the potential for statistical decadal predictions of sea surface temperatures with a perfect model approach. *Climate Dynamics* **37,** 2495–2509 (2011).

89. Hawkins, E. *et al.* Estimating Changes in Global Temperature since the Preindustrial Period. *Bulletin of the American Meteorological Society* **98,** 1841–1856 (2017).

90. Hayes, S., Mangum, L., Picaut, J., Sumi, A & Takeuchi, K. TOGA-TAO: A moored array for real-time measurements in the tropical Pacific Ocean. *Bulletin of the American Meteorological Society* **72,** 339–347 (1991).

91. Hegerl, G. C. *et al.* Climate Change Detection and Attribution: Beyond Mean Temperature Signals. *Journal of Climate* **19,** 5058 –5077 (2006).

92. Held, I. *et al.* Structure and performance of GFDL's CM4. 0 climate model. *Journal of Advances in Modeling Earth Systems* **11,** 3691–3727 (2019).

93. Held, I. M. & Soden, B. J. Robust Responses of the Hydrological Cycle to Global Warming. *Journal of Climate* **19,** 5686 –5699 (2006).

94. Hermanson, L. *et al.* Different types of drifts in two seasonal forecast systems and their dependence on ENSO. *Climate Dynamics* **51,** 1411–1426 (2018).

95. Hersbach, H. *et al.* The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society* **146,** 1999–2049 (2020).

96. Hirschberg, P. A. *et al.* A Weather and Climate Enterprise Strategic Implementation Plan for Generating and Communicating Forecast Uncertainty Information. *Bulletin of the American Meteorological Society* **92,** 1651–1666 (2011).

97. Hoell, A. & Funk, C. Indo-Pacific sea surface temperature influences on failed consecutive rainy seasons over eastern Africa. *Climate dynamics* **43,** 1645–1660 (2014).

98. Hogan, R. J. & Mason, I. B. in *Forecast verification: a practitioner's guide in atmospheric science* (eds Jolliffe, I. T. & Stephenson, D. B.) (John Wiley & Sons, 2012).

99. Hope, C. & Schaefer, K. Economic impacts of carbon dioxide and methane released from thawing permafrost. *Nature Climate Change* **6,** 56–59 (2016).

100. Horel, J. D. & Wallace, J. M. Planetary-Scale Atmospheric Phenomena Associated with the Southern Oscillation. *Monthly Weather Review* **109,** 813–829 (1981).

101. Huang, B. *et al.* Extended reconstructed sea surface temperature version 4 (ERSST. v4). Part I: Upgrades and intercomparisons. *Journal of climate* **28,** 911–930 (2015).

102. Huang, B. *et al.* Extended Reconstructed Sea Surface Temperature Version 4 (ERSSTv4). Part I: Upgrades and Intercomparisons. *Journal of Climate* **28,** 911–930 (2015).

103. Huang, B. *et al.* Further Exploring and Quantifying Uncertainties for Extended Reconstructed Sea Surface Temperature (ERSST) Version 4 (v4). *Journal of Climate* **29,** 3119–3142 (2016).

104. Huang, B. *et al.* Further exploring and quantifying uncertainties for extended reconstructed sea surface temperature (ERSST) version 4 (v4). *Journal of Climate* **29,** 3119–3142 (2016).

105. Huang, B. *et al.* Extended Reconstructed Sea Surface Temperature, Version 5 (ERSSTv5): Upgrades, Validations, and Intercomparisons. *Journal of Climate* **30,** 8179 –8205 (2017).

106. Huang, B. *et al.* Uncertainty Estimates for Sea Surface Temperature and Land Surface Air Temperature in NOAAGlobalTemp Version 5. *Journal of Climate* **33,** 1351 –1379 (2020).

107. Iizumi, T., Luo, J.-J., Challinor, A. J., *et al.* Impacts of El Niño Southern Oscillation on the global yields of major crops. *Nature communications* **5,** 1–7 (2014).

108. Ishihara, K. Calculation of global surface temperature anomalies with COBE-SST. *Weather Service Bulletin* **73,** S19–S25 (2006).

109. Jacobs, P., Lenssen, N. J., Schmidt, G. A. & Rohde, R. A. *The Arctic Is Now Warming Four Times As Fast As the Rest of the Globe* in *AGU Fall Meeting 2021* (2021).

110. Joh, Y. & Di Lorenzo, E. Interactions between Kuroshio Extension and Central Tropical Pacific lead to preferred decadal-timescale oscillations in Pacific climate. *Scientific Reports* **9,** 13558 (2019).

111. *Forecast Verification: A Practitioner's Guide in Atmospheric Science* 2nd (eds Jolliffe, I. T. & Stephenson, D. B.) (Wiley, 2012).

112. Jones, P. *et al.* Northern Hemisphere surface air temperature variations: 1851–1984. *Journal of Applied Meteorology and Climatology* **25,** 161–179 (1986).

113. Jones, P., Raper, S. & Wigley, T. Southern Hemisphere surface air temperature variations: 1851–1984. *Journal of Applied Meteorology and Climatology* **25,** 1213–1230 (1986).

114. Jones, P. D., Wigley, T. M. & Wright, P. B. Global temperature variations between 1861 and 1984. *Nature* **322,** 430–434 (1986).

115. Jones, P. The reliability of global and hemispheric surface temperature records. *Advances in Atmospheric Sciences* **33,** 269–282 (2016).

116. Karl, T. R. *et al.* Possible artifacts of data biases in the recent global surface warming hiatus. *Science* **348,** 1469–1472 (2015).

117. Kataoka, T. *et al.* Seasonal to Decadal Predictions With MIROC6: Description and Basic Evaluation. *Journal of Advances in Modeling Earth Systems* **12** (2020).

118. Kayano, M. T. & Capistrano, V. B. How the Atlantic multidecadal oscillation (AMO) modifies the ENSO influence on the South American rainfall. *International Journal of Climatology* **34,** 162–178 (2014).

119. Keenlyside, N., Latif, M., Jungclaus, J, Kornblueh, L & Roeckner, E. Advancing decadal-scale climate prediction in the North Atlantic sector. *Nature* **453,** 84–88 (2008).

120. Kennedy, J. J., Rayner, N. A., Smith, R. O., Parker, D. E. & Saunby, M. Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 1. Measurement and sampling uncertainties. *Journal of Geophysical Research* **116** (2011).

121. Kennedy, J. J., Rayner, N. A., Smith, R. O., Parker, D. E. & Saunby, M. Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization. *Journal of Geophysical Research* **116** (2011).

122. Kennedy, J. J. A review of uncertainty in in situ measurements and data sets of sea surface temperature. *Reviews of Geophysics* **52,** 1–32 (2014).

123. Kharin, V., Boer, G., Merryfield, W., Scinocca, J. & Lee, W.-S. Statistical adjustment of decadal predictions in a changing climate. *Geophysical Research Letters* **39** (2012).

124. Kirtman, B. P. *et al.* The North American Multimodel Ensemble: Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal Prediction. *Bulletin of the American Meteorological Society* **95,** 585–601 (2014).

125. Knaff, J. A. & Landsea, C. W. An El Niño–Southern Oscillation Climatology and Persistence (CLIPER) Forecasting Scheme. *Weather and Forecasting* **12,** 633–652 (1997).

126. Knight, J. R., Folland, C. K. & Scaife, A. A. Climate impacts of the Atlantic Multidecadal Oscillation. *Geophysical Research Letters* **33** (2006).

127. Kobayashi, S. *et al.* The JRA-55 Reanalysis: General Specifications and Basic Characteristics. *J. Meteorol. Soc. Japan. Ser. II* **93,** 5–48 (2015).

128. Kumar, A. *et al.* in (Guidance prepared under the auspices of the World Meteorological Organization Commission for Climatology (CCI) and Commission for Basic Systems (CBS), 2020).

129. Kushnir, Y. Interdecadal Variations in North Atlantic Sea Surface Temperature and Associated Atmospheric Conditions. *Journal of Climate* **7,** 141–157 (1994).

130. Kushnir, Y. *et al.* Towards operational predictions of the near-term climate. *Nature Climate Change* **9,** 94–101 (2019).

131. Larkin, N. K. & Harrison, D. E. Global seasonal temperature and precipitation anomalies during El Niño autumn and winter. *Geophysical Research Letters* **32** (2005).

132. Lehmann, E. L. & Romano, J. P. *Testing statistical hypotheses* Third (Springer, New York, 2005).

133. Leith, C. The standard error of time-average estimates of climatic means. *Journal of Applied Meteorology (1962-1982),* 1066–1069 (1973).

134. Lenssen, N. J. L. *et al.* Improvements in the GISTEMP Uncertainty Model. *Journal of Geophysical Research: Atmospheres* **124,** 6307–6326 (2019).

135. Lenssen, N. J. L., Goddard, L. & Mason, S. Seasonal Forecast Skill of ENSO Teleconnection Maps. *Weather and Forecasting* **35,** 2387–2406 (2020).

136. Lenssen, N. J., Goddard, L., Mason, S. & Kushnir, Y. Initialized and uninitialized ENSO predictability in year 2+. *Climate Prediction S&T Digest, 46th NOAA Climate Diagnostics and Prediction Workshop,* 41–45 (2022).

137. Levine, A. F. Z., McPhaden, M. J. & Frierson, D. M. W. The impact of the AMO on multidecadal ENSO variability. *Geophysical Research Letters* **44,** 3877–3886 (2017).

138. Li, G. & Xie, S.-P. Tropical biases in CMIP5 multimodel ensemble: The excessive equatorial Pacific cold tongue and double ITCZ problems. *Journal of Climate* **27,** 1765–1780 (2014).

139. Liu, W. *et al.* Extended Reconstructed Sea Surface Temperature Version 4 (ERSST.v4): Part II. Parametric and Structural Uncertainty Estimations. *Journal of Climate* **28,** 931–951 (2015).

140. Liu, W. *et al.* Extended Reconstructed Sea Surface Temperature Version 4 (ERSST.v4): Part II. Parametric and Structural Uncertainty Estimations. *Journal of Climate* **28,** 931 –951 (2015).

141. Liu, Z. & Di Lorenzo, E. Mechanisms and Predictability of Pacific Decadal Variability. *Current Climate Change Reports* **4,** 128–144 (2018).

142. Livezey, R. E. & Timofeyeva, M. M. The First Decade of Long-Lead U.S. Seasonal Forecasts. *Bulletin of the American Meteorological Society* **89,** 843–854 (2008).

143. Lorenz, E. N. Atmospheric predictability as revealed by naturally occurring analogues. *Journal of Atmospheric Sciences* **26,** 636–646 (1969).

144. Lorenz, E. N. The predictability of a flow which possesses many scales of motion. *Tellus* **21,** 289–307 (1969).

145. Madden, R. A. Estimates of the Natural Variability of Time-Averaged Sea-Level Pressure. *Monthly Weather Review* **104,** 942–952 (1976).

146. Magana, V., Vázquez, J., Párez, J. & Pérez, J. Impact of El Niño on precipitation in Mexico. *Geofisica Internacional* **42,** 313–330 (2003).

147. Manabe, S. & Bryan, K. Climate calculations with a combined ocean-atmosphere model. *J. atmos. Sci* **26,** 786–789 (1969).

148. Manabe, S. & Wetherald, R. T. Thermal Equilibrium of the Atmosphere with a Given Distribution of Relative Humidity. *Journal of Atmospheric Sciences* **24,** 241–259 (1967).

149. Mantua, N. J. & Hare, S. R. The Pacific Decadal Oscillation. *Journal of Oceanography* **58,** 35–44 (2002).

150. Mantua, N. J., Hare, S. R., Zhang, Y., Wallace, J. M. & Francis, R. C. A Pacific Interdecadal Climate Oscillation with Impacts on Salmon Production. *Bulletin of the American Meteorological Society* **78,** 1069–1080 (1997).

151. Mason, I. A model for assessment of weather forecasts. *Aust. Meteor. Mag* **30,** 291–303 (1982).

152. Mason, S. J. Guidance on Verification of Operational Seasonal Climate Forecasts. *WMO Commission for Climatology* (2018).

153. Mason, S. J. & Goddard, L. Probabilistic Precipitation Anomalies Associated with ENSO. *Bulletin of the American Meteorological Society* **82,** 619–638 (2001).

154. Mason, S. J. & Graham, N. E. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography* **128,** 2145–2166 (2002).

155. Mason, S. J. & Weigel, A. P. A Generic Forecast Verification Framework for Administrative Purposes. *Monthly Weather Review* **137,** 331–349 (2009).

156. Masson-Delmotte, V. *et al.* Climate change 2021: the physical science basis. *Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change,* 2 (2021).

157. Mayer, M. & Balmaseda, M. A. Indian Ocean impact on ENSO evolution 2014–2016 in a set of seasonal forecasting experiments. *Climate Dynamics* **56,** 2631–2649 (2021).

158. McCarty, W. *et al.* MERRA-2 Input Observations: Summary and Assessment. *NASA Technical Report Series on Global Modeling and Data Assimilation* **46** (2016).

159. McPhaden, M. J. Tropical Pacific Ocean heat content variations and ENSO persistence barriers. *Geophysical Research Letters* **30** (2003).

160. McPhaden, M. J., Santoso, A. & Cai, W. *El Niño Southern Oscillation in a changing climate* (John Wiley & Sons, 2020).

161. Meehl, G. A. *et al.* Decadal prediction: can it be skillful? *Bulletin of the American Meteorological Society* **90,** 1467–1486 (2009).

162. Meehl, G. A. *et al.* Initialized Earth System prediction from subseasonal to decadal timescales. *Nature Reviews Earth & Environment* **2,** 340–357 (2021).

163. Menne, M. J. & Williams, C. N. Homogenization of Temperature Series via Pairwise Comparisons. *Journal of Climate* **22,** 1700–1717 (2009).

164. Menne, M. J., Williams, C. N. & Vose, R. S. The U.S. Historical Climatology Network Monthly Temperature Data, Version 2. *Bulletin of the American Meteorological Society* **90,** 993–1008 (2009).

165. Menne, M. J., Williams, C. N. & Palecki, M. A. On the reliability of the U.S. surface temperature record. *Journal of Geophysical Research* **115** (2010).

166. Menne, M. J., Williams, C. N., Gleason, B. E., Rennie, J. J. & Lawrimore, J. H. The Global Historical Climatology Network Monthly Temperature Dataset, Version 4. *Journal of Climate* (2018).

167. Merryfield, W. J. *et al.* Current and emerging developments in subseasonal to decadal prediction. *Bulletin of the American Meteorological Society* (2020).

168. Miller, R. L. *et al.* CMIP5 historical simulations (1850–2012) with GISS ModelE2. *Journal of Advances in Modeling Earth Systems* **6,** 441–478 (2014).

169. Miller, R. L. *et al.* CMIP6 Historical Simulations (1850–2014) With GISS-E2.1. *Journal of Advances in Modeling Earth Systems* **13** (2021).

170. Mitchell Jr., J. M. Recent Secular Changes of Global Temperature. *Annals of the New York Academy of Sciences* **95,** 235–250 (1961).

171. Morice, C. P. *et al.* An Updated Assessment of Near-Surface Temperature Change From 1850: The HadCRUT5 Data Set. *Journal of Geophysical Research: Atmospheres* **126** (2021).

172. Morice, C. P., Kennedy, J. J., Rayner, N. A. & Jones, P. D. Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *Journal of Geophysical Research: Atmospheres* **117** (2012).

173. Morice, C. P. *et al.* An updated assessment of near-surface temperature change from 1850: the HadCRUT5 dataset. *Journal of Geophysical Research* (2020).

174. Mulholland, D. P., Laloyaux, P., Haines, K. & Balmaseda, M. A. Origin and impact of initialization shocks in coupled atmosphere–ocean forecasts. *Monthly Weather Review* **143,** 4631–4644 (2015).

175. Murphy, A. H. A Note on the Ranked Probability Score. *Journal of Applied Meteorology* **10,** 155–156 (1971).

176. Murphy, A. H. A New Vector Partition of the Probability Score. *Journal of Applied Meteorology* **12,** 595–600 (1973).

177. Murphy, A. H. Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient. *Monthly Weather Review* **116,** 2417 –2424 (1988).

178. Murphy, A. H. Forecast verification: Its Complexity and Dimensionality. *Monthly Weather Review* **119,** 1590–1601 (1991).

179. Myhre, G., Myhre, C. L., Forster, P. M. & Shine, K. P. Halfway to doubling of CO2 radiative forcing. *Nature Geoscience* **10,** 710–711 (2017).

180. Myhre, G. *et al.* Frequency of extreme precipitation increases extensively with event rareness under global warming. *Scientific reports* **9,** 1–10 (2019).

181. Narapusetty, B., DelSole, T. & Tippett, M. K. Optimal Estimation of the Climatological Mean. *Journal of Climate* **22,** 4845–4859 (2009).

182. NASA Public Affairs. *NASA, NOAA Find 2014 Warmest Year in Modern Record* `https://www.giss.nasa.gov/research/news/20150116/`. last-accessed July 9, 2018. 2015.

183. NASA Public Affairs. *NASA, NOAA Analyses Reveal Record-Shattering Global Warm Temperatures in 2015* `https://www.giss.nasa.gov/research/news/20160120/`. last-accessed July 9, 2018. 2016.

184. NASA Public Affairs. *NASA, NOAA Data Show 2016 Warmest Year on Record Globally* `https://www.giss.nasa.gov/research/news/20170118/`. last-accessed July 9, 2018. 2017.

185. National Research Council. *Carbon Dioxide and Climate: A Scientific Assessment* (The National Academies Press, Washington, DC, 1979).

186. Newman, M. & Sardeshmukh, P. D. Are we near the predictability limit of tropical Indo-Pacific sea surface temperatures? *Geophysical Research Letters* **44,** 8520–8529 (2017).

187. Newman, M. *et al.* The Pacific Decadal Oscillation, Revisited. *Journal of Climate* **29,** 4399–4427 (2016).

188. Notz, D. & SIMIP Community. Arctic sea ice in CMIP6. *Geophysical Research Letters* **47,** e2019GL086749 (2020).

189. Nychka, D., Furrer, R., Paige, J. & Sain, S. *fields: Tools for spatial data* R package version 11.5. Boulder, CO, USA: University Corporation for Atmospheric Research, 2017.

190. Otto, A. *et al.* Energy budget constraints on climate response. *Nature Geoscience* **6,** 415–416 (2013).

191. Overland, J. *et al.* The urgency of Arctic change. *Polar Science* **21,** 6–13 (2019).

192. Parker, D. E. Effects of changing exposure of thermometers at land stations. *International Journal of Climatology* **14,** 1–31 (1994).

193. Phillips, N. A. The general circulation of the atmosphere: A numerical experiment. *Quarterly Journal of the Royal Meteorological Society* **82,** 123–164 (1956).

194. Planton, Y. Y., Guilyardi, E., Wittenberg, A. T., *et al.* Evaluating climate models with the CLIVAR 2020 ENSO metrics package. *Bulletin of the American Meteorological Society* **102,** E193–E217 (2021).

195. Pörtner, H. O. *et al.* Climate change 2022: impacts, adaptation and vulnerability. *Contribution of working group II to the sixth assessment report of the Intergovernmental Panel on Climate Change* (2022).

196. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2020).

197. Rahman, T., Buizer, J. & Guido, Z. a. The Economic Impact of Seasonal Drought Forecast Information Service in Jamaica, 2014-15. *Paper prepared for USAID* (2016).

198. Rao, Y., Liang, S. & Yu, Y. Land Surface Air Temperature Data Are Considerably Different Among BEST-LAND, CRU-TEM4v, NASA-GISS, and NOAA-NCEI. *J. Geophys. Res. Atmos.* **123,** 5881–5900 (2018).

199. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)* (The MIT Press, 2005).

200. Raymo, M. E. & Ruddiman, W. F. Tectonic forcing of late Cenozoic climate. *Nature* **359,** 117–122 (1992).

201. Rayner, N. A. *et al.* Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research: Atmospheres* **108** (2003).

202. Rennie, J. J. *et al.* The international surface temperature initiative global land surface databank: monthly temperature data release description and methods. *Geoscience Data Journal* **1,** 75–102 (2014).

203. Richter-Menge, J., Druckenmiller, M. L. & Thoman, R. L. Arctic Report Card 2020: 15 Years of Arctic Observation: A Retrospective. *United States. National Oceanic and Atmospheric Administration. Office of Oceanic and Atmospheric Research and University of Alaska Fairbanks. Institute of Northern Engineering and National Snow and Ice Data Center (U. S.) and Alaska Center for Climate Assessment and Policy (U. S.) and International Arctic Research Center. Arctic Report Card* (2020).

204. Rohde, R *et al.* A New Estimate of the Average Earth Surface Land Temperature Spanning 1753 to 2011. *Geoinformatics & Geostatistics: An Overview* (1 2013).

205. Rohde, R *et al.* Berkeley Earth Temperature Averaging Process. *Geoinformatics & Geostatistics: An Overview* **1,** 20–100 (2 2013).

206. Rohde, R. A. & Hausfather, Z. The Berkeley Earth Land/Ocean Temperature Record. *Earth System Science Data* **12,** 3469–3479 (2020).

207. Ropelewski, C. F. & Halpert, M. S. North American Precipitation and Temperature Patterns Associated with the El Niño/Southern Oscillation (ENSO). *Monthly Weather Review* **114,** 2352–2362 (1986).

208. Ropelewski, C. F. & Halpert, M. S. Global and Regional Scale Precipitation Patterns Associated with the El Niño/Southern Oscillation. *Monthly Weather Review* **115,** 1606–1626 (1987).

209. Ropelewski, C. F. & Halpert, M. S. Precipitation Patterns Associated with the High Index Phase of the Southern Oscillation. *Journal of Climate* **2,** 268–284 (1989).

210. Roulston, M. S. & Smith, L. A. Evaluating Probabilistic Forecasts Using Information Theory. *Monthly Weather Review* **130,** 1653–1660 (2002).

211. Roulston, M. S., Bolton, G. E., Kleit, A. N. & Sears-Collins, A. L. A Laboratory Study of the Benefits of Including Uncertainty Information in Weather Forecasts. *Weather and Forecasting* **21,** 116–122 (2006).

212. Sanchez-Gomez, E., Cassou, C., Ruprich-Robert, Y., Fernandez, E. & Terray, L. Drift dynamics in a coupled model initialized for decadal forecasts. *Climate Dynamics* **46,** 1819–1840 (2016).

213. Scaife, A. A. & Smith, D. A signal-to-noise paradox in climate science. *npj Climate and Atmospheric Science* **1,** 1–8 (2018).

214. Scaife, A. A. *et al.* Tropical rainfall predictions from multiple seasonal forecast systems. *International Journal of Climatology* **39,** 974–988 (2019).

215. Schwarzwald, K. & Lenssen, N. J. L. The Importance of Internal Climate Variability in Climate Impact Projections (*in review*).

216. Sellers, W. D. A Global Climatic Model Based on the Energy Balance of the Earth-Atmosphere System. *Journal of Applied Meteorology and Climatology* **8,** 392–400 (1969).

217. Serreze, M. C. & Barry, R. G. Processes and impacts of Arctic amplification: A research synthesis. *Global and Planetary Change* **77,** 85–96 (2011).

218. Sherwood, S. C. *et al.* An Assessment of Earth's Climate Sensitivity Using Multiple Lines of Evidence. *Reviews of Geophysics* **58** (2020).

219. Shindell, D. T. *et al.* Radiative forcing in the ACCMIP historical and future climate simulations. *Atmos. Chem. Phys.* **13,** 2939–2974 (2013).

220. Shindell, D. T. Inhomogeneous forcing and transient climate sensitivity. *Nature Climate Change* **4,** 274–277 (2014).

221. Simmons, A. J., Willett, K. M., Jones, P. D., Thorne, P. W. & Dee, D. P. Low-frequency variations in surface atmospheric humidity, temperature, and precipitation: Inferences from reanalyses and monthly gridded observational data sets. *Journal of Geophysical Research* **115** (2010).

222. Simmons, A. J. *et al.* A reassessment of temperature variations and trends from global reanalyses and monthly surface climatological datasets. *Quarterly Journal of the Royal Meteorological Society* **143,** 101–119 (2016).

223. Simpson, G. L. Modelling Palaeoecological Time Series Using Generalised Additive Models. *Frontiers in Ecology and Evolution* **6** (2018).

224. Smith, D. M. *et al.* Robust skill of decadal climate predictions. *npj Climate and Atmospheric Science* **2,** 13 (2019).

225. Smith, D. M. *et al.* Improved Surface Temperature Prediction for the Coming Decade from a Global Climate Model. *Science* **317,** 796–799 (2007).

226. Smith, D. M. *et al.* North Atlantic climate far more predictable than models imply. *Nature* **583,** 796–800 (2020).

227. Smith, T. M. & Reynolds, R. W. A Global Merged Land–Air–Sea Surface Temperature Reconstruction Based on Historical Observations (1880–1997). *Journal of Climate* **18,** 2021–2036 (2005).

228. Sospedra-Alfonso, R. *et al.* Decadal climate predictions with the Canadian Earth System Model version 5 (CanESM5). *Geoscientific Model Development* **14,** 6863–6891 (2021).

229. Suarez, P. & Patt, A. G. Cognition, caution, and credibility: the risks of climate forecast application. *Risk, Decision and Policy* **9,** 75–89 (2004).

230. Suckling, E. B., van Oldenborgh, G. J., Eden, J. M. & Hawkins, E. An empirical model for probabilistic decadal prediction: global attribution and regional hindcasts. *Climate Dynamics* **48,** 3115–3138 (2017).

231. Sulca, J., Takahashi, K., Espinoza, J.-C., Vuille, M. & Lavado-Casimiro, W. Impacts of different ENSO flavors and tropical Pacific convection variability (ITCZ, SPCZ) on austral summer rainfall in South America, with a focus on Peru. *International Journal of Climatology* **38,** 420–435 (2018).

232. Sullivan, G. M. & Feinn, R. Using Effect Size-or Why the P Value Is Not Enough. *Journal of Graduate Medical Education* **4.** JGME-D-12-00156[PII], 279–282 (2012).

233. Susskind, J, Schmidt, G. A., Lee, J. N. & Iredell, L. Recent global warming as confirmed by AIRS. *Environmental Research Letters* **14** (2019).

234. Swart, N. C. *et al.* The Canadian earth system model version 5 (CanESM5. 0.3). *Geoscientific Model Development* **12,** 4823–4873 (2019).

235. Thorne, P. W. *et al.* Towards a global land surface climate fiducial reference measurements network. *International Journal of Climatology* **38,** 2760–2774 (2018).

236. Timmermann, A. *et al.* El Niño–Southern Oscillation complexity. *Nature* **559,** 535–545 (2018).

237. Ting, M., Kushnir, Y., Seager, R. & Li, C. Robust features of Atlantic multi-decadal variability and its climate impacts. *Geophysical Research Letters* **38** (2011).

238. Ting, M., Kushnir, Y. & Li, C. North Atlantic Multidecadal SST Oscillation: External forcing versus internal variability. *Journal of Marine Systems* **133,** 27–38 (2014).

239. Tödter, J. & Ahrens, B. Generalization of the Ignorance Score: Continuous Ranked Version and Its Decomposition. *Monthly Weather Review* **140,** 2005–2017 (2012).

240. Tokarska, K. B. *et al.* Past warming trend constrains future warming in CMIP6 models. *Science advances* **6** (2020).

241. Trenberth, K. E. & Shea, D. J. Atlantic hurricanes and natural variability in 2005. *Geophysical Research Letters* **33** (2006).

242. Trenberth, K. E., Fasullo, J. T. & Shepherd, T. G. Attribution of climate extreme events. *Nature Climate Change* **5,** 725–730 (2015).

243. Van Oldenborgh, G. J., te Raa, L. A., Dijkstra, H. A. & Philip, S. Y. Frequency- or amplitude-dependent effects of the Atlantic meridional overturning on the tropical Pacific Ocean. *Ocean Science* **5,** 293–301 (2009).

244. Van Heerden, J., Terblanche, D. E. & Schulze, G. C. The southern oscillation and South African summer rainfall. *Journal of Climatology* **8,** 577–597 (1988).

245. Vimont, D. J., Wallace, J. M. & Battisti, D. S. The seasonal footprinting mechanism in the Pacific: Implications for ENSO. *Journal of Climate* **16,** 2668–2675 (2003).

246. Vitart, F. & Robertson, A. W. The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. *npj Climate and Atmospheric Science* **1,** 3 (2018).

247. Vose, R. S. *et al.* NOAA's Merged Land–Ocean Surface Temperature Analysis. *Bulletin of the American Meteorological Society* **93,** 1677–1685 (2012).

248. Vuille, M., Bradley, R. S. & Keimig, F. Interannual climate variability in the Central Andes and its relation to tropical Pacific and Atlantic forcing. *Journal of Geophysical Research: Atmospheres* **105,** 12447–12460 (2000).

249. Wahba, G. *Spline models for observational data* (SIAM, 1990).

250. Walker, G. T. Correlations in seasonal variations of weather. I. A further study of world weather. *Mem. Indian Meteorol. Dep.* **24,** 275–332 (1924).

251. Wang, Y.-M., Lean, J. L. & N. R. Sheeley, J. Modeling the Sun's Magnetic Field and Irradiance since 1713. *The Astrophysical Journal* **625,** 522–538 (2005).

252. Weijs, S. V., van Nooijen, R. & van de Giesen, N. Kullback–Leibler Divergence as a Forecast Skill Score with Classic Reliability–Resolution–Uncertainty Decomposition. *Monthly Weather Review* **138,** 3387–3399 (2010).

253. Wilks, D. S. Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorological Applications* **16,** 361–368 (2009).

254. Williams, A. P., Cook, B. I. & Smerdon, J. E. Rapid intensification of the emerging southwestern North American megadrought in 2020–2021. *Nature Climate Change* **12,** 232–234 (2022).

255. Williams, C. N., Menne, M. J. & Thorne, P. W. Benchmarking the performance of pairwise homogenization of surface temperatures in the United States. *Journal of Geophysical Research: Atmospheres* **117** (2012).

256. Wittenberg, A. T. Are historical records sufficient to constrain ENSO simulations? *Geophysical Research Letters* **36** (2009).

257. Wittenberg, A. T., Rosati, A., Delworth, T. L., Vecchi, G. A. & Zeng, F. ENSO modulation: Is it decadally predictable? *Journal of Climate* **27,** 2667–2681 (2014).

258. Wood, S. N. *Generalized additive models: an introduction with R* (chapman and hall/CRC, 2006).

259. Wood, S. N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73,** 3–36 (2011).

260. Wu, X., Okumura, Y. M., Deser, C. & DiNezio, P. N. Two-Year Dynamical Predictions of ENSO Event Duration during 1954–2015. *Journal of Climate* **34,** 4069–4087 (2021).

261. Wyrtki, K. El Niño—the dynamic response of the equatorial Pacific Oceanto atmospheric forcing. *Journal of Physical Oceanography* **5,** 572–584 (1975).

262. Xie, P. & Arkin, P. A. Global Precipitation: A 17-Year Monthly Analysis Based on Gauge Observations, Satellite Estimates, and Numerical Model Outputs. *Bulletin of the American Meteorological Society* **78,** 2539–2558 (1997).

263. Yeager, S. G. *et al.* Predicting Near-Term Changes in the Earth System: A Large Ensemble of Initialized Decadal Prediction Simulations Using the Community Earth System Model. *Bulletin of the American Meteorological Society* **99,** 1867–1886 (2018).

264. Yeh, S.-W. *et al.* ENSO Atmospheric Teleconnections and Their Response to Greenhouse Gas Forcing. *Reviews of Geophysics* **56,** 185–206 (2018).

265. Zebiak, S. E. On the 30–60 day oscillation and the prediction of El Nino. *Journal of Climate,* 1381–1387 (1989).

266. Zebiak, S. E. & Cane, M. A. A Model El Niño-Southern Oscillation. *Monthly Weather Review* **115,** 2262 –2278 (1987).

267. Zhang, R. *et al.* A Review of the Role of the Atlantic Meridional Overturning Circulation in Atlantic Multidecadal Variability and Associated Climate Impacts. *Reviews of Geophysics* **57,** 316–375 (2019).

268. Zhao, S., Jin, F.-F. & Stuecker, M. F. Understanding Lead Times of Warm Water Volumes to ENSO Sea Surface Temperature Anomalies. *Geophysical Research Letters* **48** (2021).