Process-based machine learning for observational constraints on temperature in past and future climates

Sophie Wilkinson

Registration No. 100328862

Supervisors: Prof. Peer Nowack & Prof. Manoj Joshi

A Thesis submitted for the degree of Doctor of Philosophy

School of Environmental Sciences

University of East Anglia

October 2023

Abstract

This thesis develops a novel method using machine learning (ML) to combine existing climate model output with observations in order to impose new constraints on regional near-surface temperature anomalies from the Coupled Model Inter-comparison Project Phase 6 (CMIP6) in Northern Hemisphere summertime. The ML model is trained with reanalysis data to predict near-surface temperature anomalies from a set of process-based dynamic and thermodynamic predictor variables. Test predictions over the historical period capture day-to-day variation in temperature anomalies well, including the magnitude of more extreme events (for example the Europe 2003 heatwave) in most land regions. The ML method is applied for bias correction of historical climate model output and to constrain uncertainty in future warming projections. This future warming constraint points to a potential over-sensitivity of several CMIP6 models with the constraint tending to result in a small downward correction of the projected future temperature range. The ML bias correction technique performs competitively with a traditional variance scaling approach and, using interpretable AI methods, can be decomposed into contributions from each predictor variable to reveal the presence of potential compensating biases in climate models. The ML method is also applied in a climate model evaluation context by additionally training CMIP6 ML emulators using historical climate model data. Learnt coefficients are then compared between climate models and reanalysis to identify process simulation differences. Finally, the reanalysis-based model is applied as a tool for investigating historical heatwave drivers which are compared with existing literature. The ML method can also be used for climate model evaluation in this context, evaluating the representation of extreme event processes in climate models and testing the ability of climate models to reproduce the magnitude of historical heatwave events when provided with predictor variables from reanalysis.

Acknowledgements

I would like to thank my supervisors, Peer and Manoj, for their advice and support over the last three years, and for giving me the opportunity to take on this project in the first place. I'm extremely grateful for all of your feedback and have learnt so much from you both.

I would also like to thank my family and friends for their support, in particular Katie and Tomas.

Contents

List of Figures

List of Tables

1 Introduction

Whilst increasing global mean temperatures have been unequivocally linked with anthropogenic influence (IPCC, 2021), climate change signals are also identifiable on a regional scale (Doblas-Reyes et al., 2021). Changes in regional climate paint a much more complex picture than global mean warming, with some regions warming at a faster rate than others and being subject to different degrees of variability (Xie et al., 2015). There is still a large spread in regional climate sensitivity across the latest generation of global climate models, a range that has seen little improvement compared with previous climate model ensembles (Seneviratne and Hauser, 2020). Factors controlling regional mean state climate have knock-on effects for climate extremes - a trend of increasing frequency, intensity and duration of warm temperature anomalies has been observed over several decades (Perkins-Kirkpatrick and Lewis, 2020). While such heatwave events have already resulted in devastating impacts in many regions globally (*e.g.* Robine et al., 2008; Russo et al., 2015; Xu et al., 2016; Yang et al., 2019), this trend is only projected to continue under future warming. Increasing our understanding of how temperature anomalies are represented by existing climate models and constraining uncertainty in projections of these anomalies under future climate change are key issues for informing policy and limiting future impacts. Machine learning (ML) techniques offer new opportunities for investigating these questions and bridging the gap between observations and climate models.

In this thesis, a novel method is proposed to merge historical reanalysis and climate model data for improved simulations of summertime temperature anomalies in the Northern Hemisphere. Training a process-based ML model with reanalysis data to predict daily temperature anomalies, this method learns relationships between drivers of temperature anomalies and their extremes, which can be applied to, and contrasted with, existing climate model data. With applications in bias correction, climate model evaluation, constraining uncertainty in future projections and interpreting extreme event drivers, the method is used to address key questions in climate science.

This introduction begins by considering the factors which control mean state temperature across the globe and how extreme heat events can be defined as deviations from this mean state. Secondly, the drivers of mid-latitude heat anomalies are considered including dynamical anomalies and land-surface feedbacks. Global climate models are introduced with consideration of their component parts and sources of uncertainties. The representation of heatwave drivers in current global climate models is investigated, as well as the predicted future influence of climate change on the global distribution of temperature and specifically its impact on extreme heat events. Next, consideration is given to the availability of global observations of temperature and its drivers as well as the benefits and limitations of reanalysis

data sets. Finally, ML is introduced as a complementary tool to existing climate modelling methods followed by an outline of the structure for the rest of this thesis.

1.1 Temperature and its Extremes

Since pre-industrial times, an increase of 0.8-1.2^oC in global mean temperature has already been observed with larger increases in some regions (Allen et al., 2018). The Intergovernmental Panel on Climate Change (IPCC) states that evidence for human influence on global climate is now 'unequivocal' (IPCC, 2021). Increasing global and regional temperatures not only have direct impacts on human health but are also associated with knock-on effects on water demand, air quality, wildfire risk and more (Patz et al., 2005; Kovats and Kristie, 2006; McMichael et al., 2006; Ahima, 2020).

While changes in temperature over the last few decades exhibit a clear global warming signal, patterns naturally exist in the distribution of mean state temperature. Figure 1.1 shows the average summertime (JJA) temperature from the ECMWF Reanalysis (ERA) 5 data set (Hersbach et al., 2020) for 1979-2022. Here, a clear equator to pole temperature gradient is visible with the hottest temperatures on average recorded over land in the Tropics and much cooler temperatures at the poles. Areas of high altitude such as the Andes, the Himalayas and the Tibetan plateau also stand out as relatively cooler.

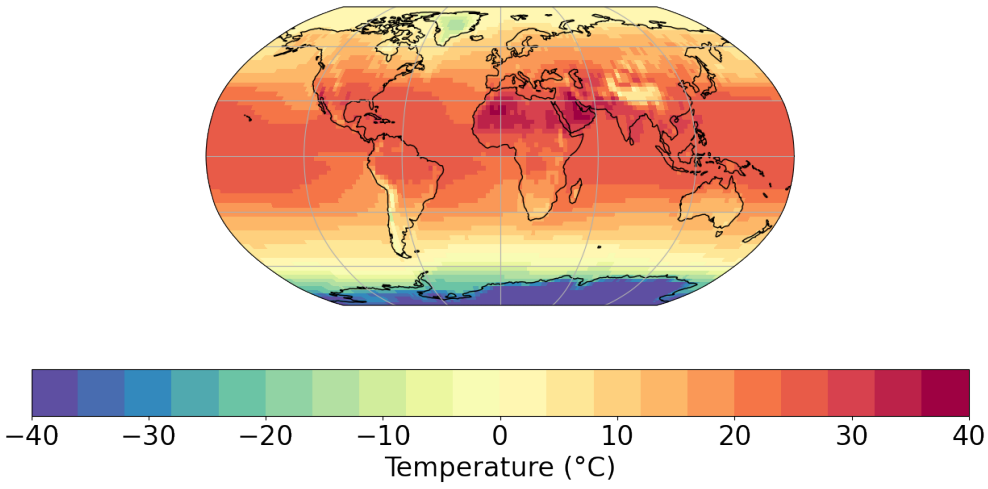

Figure 1.1: Average daily-mean temperature for JJA 1979-2021 from ERA5.

Local, mean state temperature conditions are controlled by a variety of factors including geographic position, altitude, albedo feedbacks and differential heating of land *vs.* water

(Baede et al., 2001). A key factor in determining mean state temperature is latitude because of the distribution of insolation - incoming solar radiation - over the Earth's surface. The equator receives the most insolation and also has the smallest intra-annual variation. Insolation decreases towards the poles and also shows greater seasonal variance at higher latitudes. The result of this is an equator to pole gradient in mean temperature conditions as seen in Figure 1.1.

The factors controlling temperature on a regional basis are of course more complex than simply the degree of direct radiative heating. The heat gradient between equator and poles is moderated by the currents of the global ocean along with atmospheric circulation patterns and properties of the land such as albedo, vegetation cover and soil moisture interactions (Baede et al., 2001; McGuffie and Henderson-Sellers, 2001). Land areas change temperature more rapidly and to greater extremes than the ocean as a result of specific heat capacity and the influence of ocean currents which affect local temperatures via heat transport and mixing. Additionally, water can also evaporate at the surface of the ocean resulting in cooling via latent heat without the same restriction of moisture availability as is present over land regions.

These variations in baseline, background temperature as well as factors like preparedness and resource availability contribute to the distribution and impact of extreme temperature anomalies. There is no universal, quantitative definition of what constitutes a heatwave, mostly attributable to the dependence of the definition on the intended application. The IPCC defines a heatwave as "a period of abnormally hot weather" which may last anywhere from "two days to months" (IPCC, 2021) and an extreme event as one that "is rare at a particular place and time of year" (IPCC, 2021). When considering human impacts, factors like exposure, vulnerability and adaptation may also need to be taken into account (Robinson, 2001). For the purposes of the analysis performed here, a generalisable and quantitative definition is required.

Various percentile and fixed value thresholds quantifying intensity, duration and/or frequency have been used in the literature to examine extreme events (*e.g.* Goodess, 2013; Perkins, 2015). Commonly used quantitative definitions often refer to percentile thresholds calculated from the historical distribution of temperatures at a given location and time of year, *e.g.* events exceeding the 95^{th} or 99^{th} percentile (Fischer and Schär, 2010; Stefanon et al., 2012; Schoetter et al., 2015). In some instances it may be more appropriate to consider fixed rather than relative thresholds, for example when studying human health impacts (Robinson, 2001; Ullah et al., 2022).

Recent years have seen multiple daily temperature records being broken across the globe, for example temperatures of $49.6^{o}C$ during the Pacific Northwest heatwave of 2021 (White et al., 2023). The intensity of such record breaking extremes can be quantified by how

far they exceed the previous record either as an absolute value or a number of standard deviations (Fischer et al., 2021). Alternative statistical measures consider extreme event return periods: a measure of the expected time between events of such magnitude occurring. Or, conversely, the magnitude of an N-year return period event. While the indices and measures described already allow intensity and frequency to be quantified, they typically refer only to a single time point and location measurement but can be extended to quantify the duration of extreme event conditions. For example, duration may be quantified by the number of days/weeks/months meeting a particular threshold (Meehl and Tebaldi, 2004; Fischer and Schär, 2010; Schoetter et al., 2015).

The next subsection outlines the major drivers which contribute to mid-latitude heatwaves. These are the processes that will need to be captured by the process-based ML model in order to construct effective constraints on existing climate model output.

1.2 Heatwave Drivers

Considering a loose definition of an 'extreme' temperature as one lying in the tails of the temperature distribution for a particular location and time of year, heatwave drivers are the mechanisms which drive the climate system away from mean state conditions and towards those extremes. Any given heatwave event may arise as a result of a single primary driving mechanism or a complex interplay of contributions from a range of processes conditioned on the background state of the system. Frequently identified driving forces are: anomalies in the atmospheric circulation (including remote teleconnection effects and blocking); land-surface feedbacks such as soil moisture deficits; and sea surface temperature anomalies (Palmer, 2013; Miralles et al., 2014; Wehrli et al., 2019; Liu et al., 2020; Barriopedro et al., 2023). Understanding these physical processes is essential for accurate representation of them in climate models and facilitating more reliable simulations of how extreme temperatures will evolve in the future under climate change.

Heatwaves result from a combination of these driving factors changing the regional energy budget, primarily diabatic heating (radiation from the sun, release of latent heat, transfer of sensible heat), advection (horizontal transport of warm air) and adiabatic heating of descending air (Domeisen et al., 2023). Northern Hemisphere midlatitude heat extremes are frequently associated with anticyclones or 'blocking' (Stephenson, 2008; Träger-Chatterjee et al., 2013; Miralles et al., 2014; Woollings et al., 2018; Chan et al., 2022), characterised by large scale disturbance of mid to high latitude atmospheric circulation by persistent near-surface high pressure conditions. These high pressure, slow moving systems divert faster moving systems and prolong atmospheric conditions under which advection, clear-sky radia-

tive heating and subsidence with adiabatic heating contribute to hot weather. See Figure 1.2 for a summary of these processes which contribute to the accumulation of heat and temperature extremes.

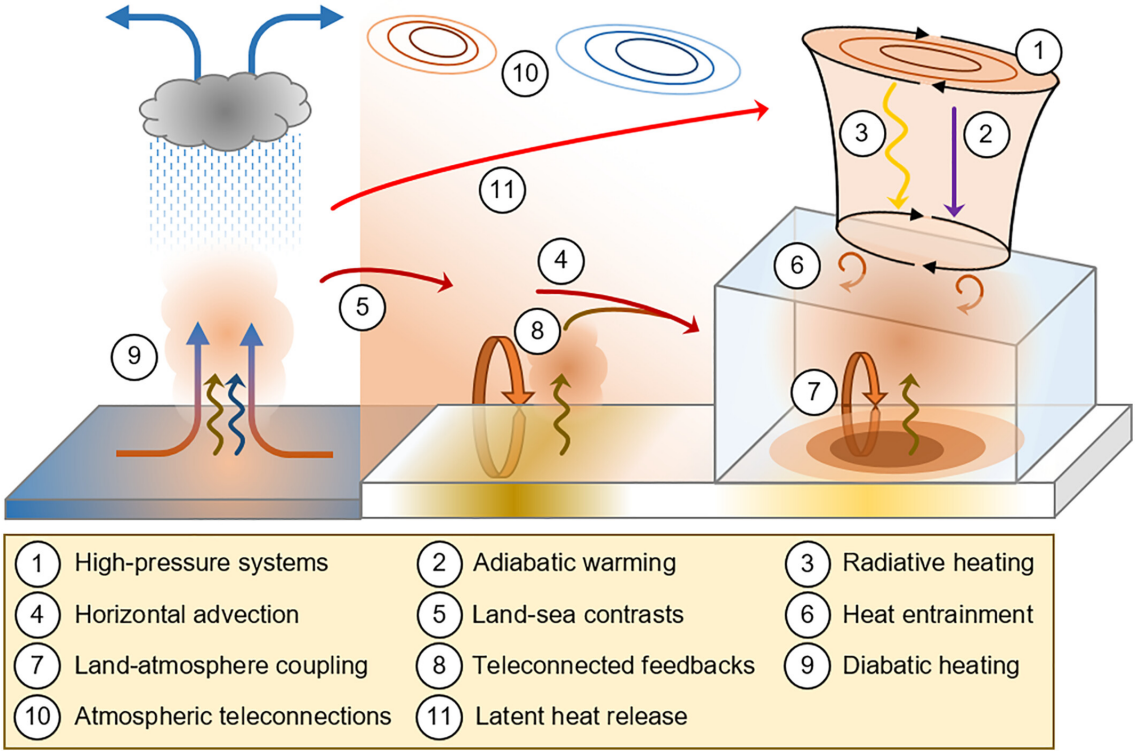

Figure 1.2: Diagram summarising driving factors which contribute to heatwaves. *Source: Barriopedro et al., 2023.*

There is a strong link between heat and drought resulting from the long duration anticyclones that tend to drive such conditions, for example in Northern Eurasia (Schubert et al., 2014). Diversion of storm systems by the persistent anticyclone reduces precipitation in the region while simultaneously increasing temperatures via downward transport of warm air. This high pressure system of descending air is also conducive to clear sky conditions leading to greater surface heating via solar radiation. An associated soil moisture feedback whereby reduced precipitation leads to drier soils and therefore a reduced capacity for evaporative cooling is another contributing factor to the association. For example, the 2010 Russian heatwave which was characterised by sustained and extremely high temperatures (Wehrli et al., 2019) resulted from a persistent period of atmospheric blocking (Otto et al., 2012) which was exacerbated by a soil moisture deficit (Hauser et al., 2016; Schubert et al., 2014) attributed to a lack of precipitation in the first half of the year and early spring snow melt (Barriopedro et al., 2011).

Aside from dynamical, high pressure anomalies, land-surface feedbacks also contribute to heatwave conditions. The amount of moisture available in the soil affects the ratio of latent and sensible heat fluxes (Alexander, 2010; Perkins, 2015). When there is moisture available, latent heat is the dominant flux and evaporation increases near-surface relative humidity promoting cloud formation resulting in a net cooling effect. Conversely, under drought conditions, sensible heat fluxes dominate, contributing to a deeper and warmer atmospheric boundary layer and less cloud, setting up a positive feedback loop which prolongs and intensifies already hot and dry conditions (Alexander, 2010; Miralles et al., 2014). Thus, pre-conditions such as seasonal precipitation deficits leading to dry soils can result in more intense heatwaves (Quesada et al., 2012; Miralles et al., 2019).

There are also associations between regional temperature extremes and global scale modes of variability. For instance, the El Niño Southern Oscillation (ENSO) which is the most significant mode of inter-annual variability (Bjerknes, 1969), is associated with a global pattern of teleconnections which influence weather across the globe (Bjerknes, 1969; Trenberth et al., 1998; Diaz et al., 2001). Teleconnections refer to statistically significant correlations between climate anomalies in remote regions. They may be caused by Rossby wave trains, displacement of mid-latitude jet or storm tracks or circulation fluctuations; they can be initiated by modes of climate variability which can result in time lagged effects (IPCC, 2021). In the case of ENSO, potential teleconnections have for example been identified with heat extremes in North America (Loikith and Broccoli, 2014), Eastern Europe (Behera et al., 2013), China (Luo and Lau, 2019), Western Australia (Feng et al., 2013) and drought in Southern Spain (Muñoz-Díaz and Rodrigo, 2005).

The next subsection provides some background on global climate models followed by a review of how these models capture the heatwave driving processes outlined here.

1.3 Global Climate Models

The IPCC defines the climate system as composed of "five major components: the atmosphere, the hydrosphere, the cryosphere, the lithosphere and the biosphere and the interactions between them" (IPCC, 2021). Traditional climate modelling methods have represented an increasingly comprehensive understanding of this system; from simple, one-dimensional energy-balance models which could be calculated by hand to the latest Global Climate Models (GCMs) which are run on supercomputers (McGuffie and Henderson-Sellers, 2001). GCMs produce regularly gridded output for a suite of climatic variables. Since it is not possible to perform targeted climate-scale experiments in reality, climate models constitute a primary tool for study of Earth's climate system and climate change (Le Treut et al., 2007).

Atmospheric models were the earliest to be developed and contain a dynamical core which models the large scale motions of the atmosphere according to the equations of fluid dynamics (Gettelman and Rood, 2016). This includes accounting for fundamental laws conserving mass and momentum as well as the effects of gravity described by Newton's laws of classical mechanics. The dynamical core must also encapsulate the laws of thermodynamics from conservation of energy to the Stefan-Boltzmann Law which quantifies the thermal radiation emitted by a mass. Physical properties of the atmosphere (and oceans) can be modelled by the Navier-Stokes equations, a set of differential equations which describe the dynamics of fluid parcels on a rotating sphere. The equations are solved numerically to quantify the exchange of conserved quantities between adjacent model grid cells in both the horizontal and vertical dimensions so that the value of various climatic variables can be calculated at each time step, see Figure 1.3. Alternatively, in a spectral model, dynamical fields are

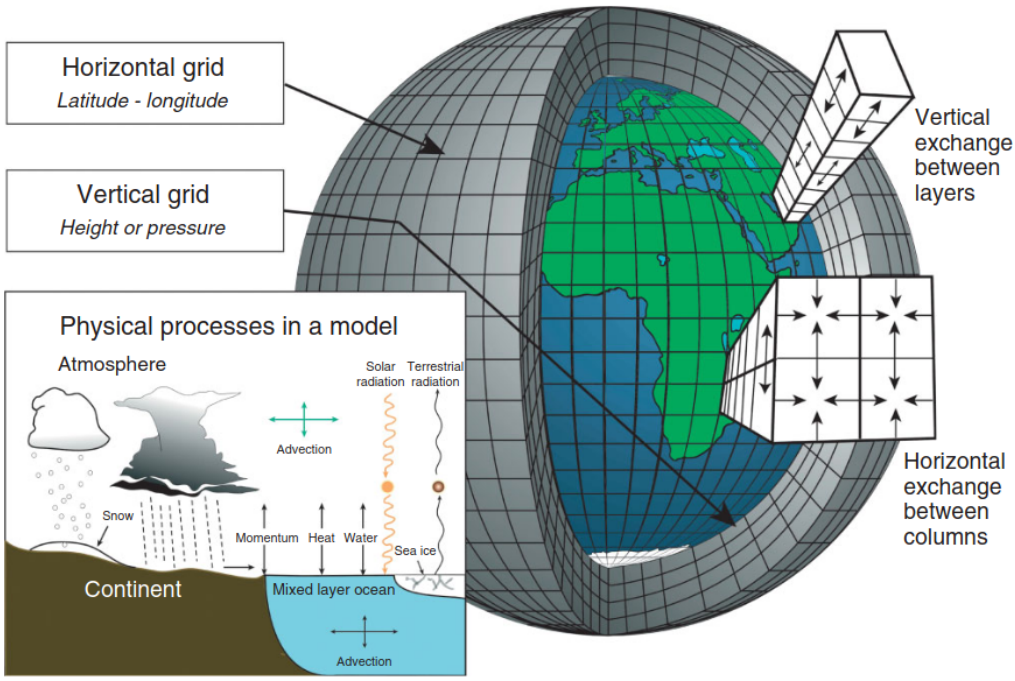

Figure 1.3: Schematic of global climate model displaying a horizontal longitude-latitude grid and vertical levels alongside the key processes that are represented (inset). *Source: Edwards, 2011.*

represented as waves using spherical harmonics. Alongside this dynamical core, the model physics account for all other significant processes from radiative transfer to cloud formation and friction (Edwards, 2011). In reality, many of these processes occur on a much smaller scale than can be resolved by a typical GCM, it is therefore necessary to represent some processes indirectly using parameterisations. Parameterisations represent the large scale

effects of sub-grid scale processes without having to simulate the processes explicitly. A key step in model development is tuning: adjusting parameter values to stabilise model output and bring particular model features closer in line with observed quantities (Mauritsen et al., 2012; Hourdin et al., 2017).

With increasing computing power and efficiency, climate model resolution and the length of climate model runs also increased (McGuffie and Henderson-Sellers, 2001). New generations of supercomputers also enabled the next step towards comprehensive simulation of the Earth's climate system, coupling general circulation atmospheric models to ocean models (Le Treut et al., 2007) in Atmosphere-Ocean General Circulation Models (AOGCMs). Development has continued in this direction with the latest Earth System Models (ESMs) coupling AOGCMs to vegetation, hydrology and cryosphere models (Flato, 2011), see Figure 1.4. Throughout the development of climate models, collaboration between institutes and knowl-

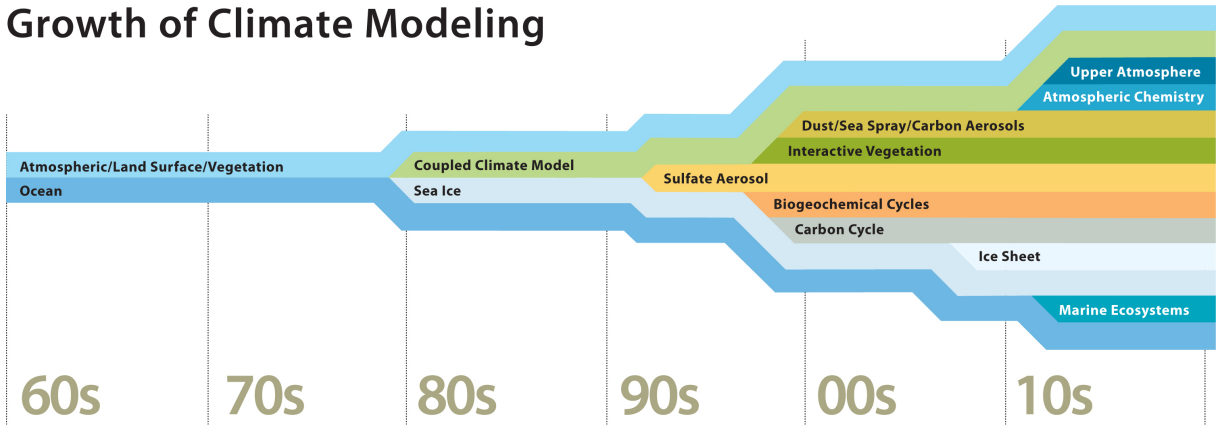

Figure 1.4: Timeline of climate model development. *Source: UCAR.*

edge sharing has contributed to a certain degree of overlap in the underlying components powering GCMs, for example different atmospheric models being coupled to the same ocean model. Combined with a mix-and-match approach to coupled modelling, this has resulted in a complex entanglement of model performance and interdependence which does present some challenges when it comes to model evaluation and uncertainty quantification (Pirtle et al., 2010; Masson and Knutti, 2011; Pennell and Reichler, 2011; Knutti et al., 2013; Sanderson et al., 2015*a*; Annan and Hargreaves, 2017).

By providing models with different forcings; realisations of Earth's climate under different scenarios can be generated resulting in vast archives of data for analysis (Eyring et al., 2016). The Coupled Model Inter-comparison Project (CMIP) set up under the auspices of the World Climate Research Programme represents a global effort to understand Earth's past and future climate. Participating modelling centres run prescribed scenario experi-

ments to allow for inter-comparison of model output with the aim of better understanding climate change. The sequential generations of CMIP models reflect ongoing progress in the field of climate modelling from improved representations of physical processes to better understanding of climate phenomena and increasing computing capabilities. The most recent generation, CMIP6 (Eyring et al., 2016), represents the latest in state-of-the-art climate modelling. Within CMIP6, experiments include simulation of historical climate using observations-based forcings as well as projections of future climate under a range of Shared Socio-economic Pathways (SSPs) (O'Neill et al., 2015).

Although great progress has been made and the latest generation of GCMs perform impressively in many areas, there are still challenges to overcome. Certain phenomena or dynamics are governed by sub-grid scale processes which are not feasible to resolve on a global scale with current computational constraints (Maraun et al., 2017), for example; cloud feedbacks, changes in land cover and biological feedbacks (Pitman et al., 2012). So, even though essential for representing important small scale processes, parameterisations are responsible for some of the largest uncertainties and discrepancies between climate models (Hargreaves, 2010). In general, parameterisations vary between modelling centres and contribute to inter-model uncertainty as well as potential biases in model output (Räisänen, 2007). Additionally, certain processes associated with extreme events in particular, for example atmospheric blocking, are known to be relatively poorly simulated (Lupo, 2021) when model output is compared with observations. Climate models naturally do not have the ability to perfectly reproduce reality; the climate system is inherently chaotic and non-linear. Current computing power imposes limits on time and spatial resolution; and processes still remain which are not yet fully understood (Le Treut et al., 2007). All of these factors can contribute to individual biases with knock-on effects leading to variation between models (Fan et al., 2020).

When seeking to summarise output between models across time, individual models provide different climate distribution statistics - a result of different modelling processes and biases. When projected out to the end of the century, model differences can result in uncertainties of for example several degrees for even relatively simple quantities such as daily-mean surface temperature change (Sherwood et al., 2020). How should output from different models be combined? Does every model deserve equal weight? Are models equally independent? Bias correction and uncertainty constraint are broad terms for methods designed to counter some of the challenges outlined above via post-processing of climate model output. A variety of these existing methods will be explored in detail in later research chapters. The next subsection specifically addresses the representation of known heatwave drivers by existing climate models in the historical and present-day climate.

1.4 Representation of Heatwaves in Climate Models

This subsection analyses the performance of existing climate models in the context of extreme event modelling, specifically, considering the representation of extreme temperatures or heatwaves in Northern Hemisphere summertime as these events will be the focus of the process-based ML model.

Despite significant effort to improve the quality and number of processes represented by GCMs (Flato et al., 2013), less progress has been made in the simulation of extremes compared with mean state behaviour (Di Luca et al., 2020; Wehner et al., 2020). Figure 1.5 shows a comparison of CMIP5 and CMIP6 performance in simulating mean heatwave attributes (Hirsch et al., 2021). The following paragraphs highlight the ability of current

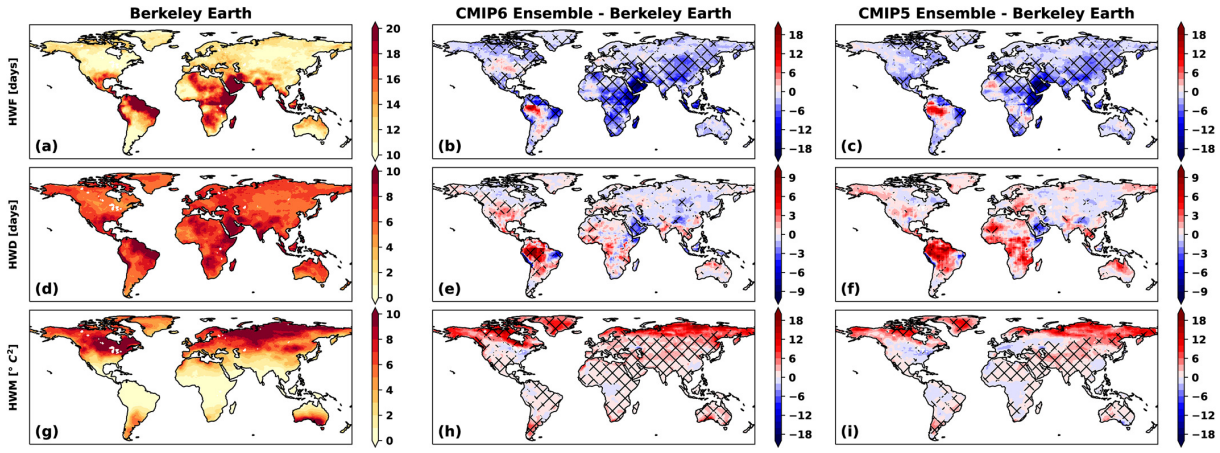

Figure 1.5: Global maps of mean heatwave attributes; mean frequency of heatwave days (HWF, a-c), length of the longest heatwave (HWD, d-f) and average heatwave magnitude (HWM, g-i). *Source: Hirsch et al., 2021.*

GCMs to capture the processes which drive extreme temperature anomalies.

The previous subsection highlighted atmospheric circulation anomalies such as blocking as key driving forces for extreme heat in the mid-latitudes (Section 1.2). Blocking occurs when high pressure systems disrupt the prevailing westerly flow, prolonging circumstances which allow temperature and moisture anomalies to accumulate resulting in hot and dry conditions. In general, blocking frequency is underestimated across CMIP3/5/6 in comparison with observations (Masato et al., 2013; Woollings et al., 2018; Davini and D'Andrea, 2020). CMIP5 models underestimate the frequency and duration of blocking events (Davini and D'Andrea, 2016). The newer generation of CMIP6 models show some improvements but are still known to underestimate blocking frequency and duration (Davini and D'Andrea, 2020; Schiemann et al., 2020). Factors which have been identified as important for improving

the representation of blocking in models include model resolution, sea surface temperature biases, physical parameterisations (*e.g.* drag), and representation of orography (Matsueda, 2009; Jung et al., 2012; Berckmans et al., 2013; Pithan et al., 2016; Woollings et al., 2018; Schiemann et al., 2020).

Land-atmosphere interactions also play a key role in moderating heatwave conditions with the potential for dry soils to exacerbate extreme heat. Comparison of CMIP6 soil moisture fields with reanalyses indicates that global patterns of soil moisture are well represented (Yuan et al., 2021; Wang et al., 2022). However there is a wet bias in the simulation of higher latitude Northern Hemisphere soil moisture where cold-season processes contribute significantly rather than soil moisture being dominated by the relationship between precipitation and evaporation (Qiao et al., 2021; Sang et al., 2021). In general, the multi-model mean shows better agreement with observations/reanalyses whilst there is still significant variation between models (Yuan et al., 2021; Qiao et al., 2021; Wang et al., 2022).

GCMs are not a perfect representation of Earth's climate system; model biases, poorly simulated and missing processes contribute to an imperfect representation of reality. The case of modelling extreme events is particularly challenging given their scarcity in the observational record. This is where post-processing calculations to bias correct model output and constrain the range of future projections are useful. In this thesis, an ML based approach is taken to address these questions; combining existing climate model data with observations to construct novel, observational constraints on CMIP6 temperature output.

1.5 Heatwaves and Climate Change

The previous subsections gave consideration to factors controlling temperature and its extremes in the present day climate, however, climate change signals attributable to human factors are already detectable and affecting the probability of such extreme events occurring (Diffenbaugh et al., 2017; Fischer et al., 2021). The expected continuation of global warming trends into the future has a direct effect on many heatwave drivers as well as mean state temperature conditions. As a result, heatwave events are expected to become more frequent, more intense and longer in duration (Seneviratne et al., 2021).

Increased levels of greenhouse gases in the atmosphere contribute to an enhanced greenhouse effect resulting in warming across the globe, although these warming effects are not evenly distributed, see Figure 1.6. One of the clearest patterns in the global warming signal is the land-sea warming contrast; future climate model simulations indicate that the land warms at a faster rate than the oceans on average (Sutton et al., 2007). There are several factors which contribute to this pattern, firstly the thermal inertia of the global ocean. The

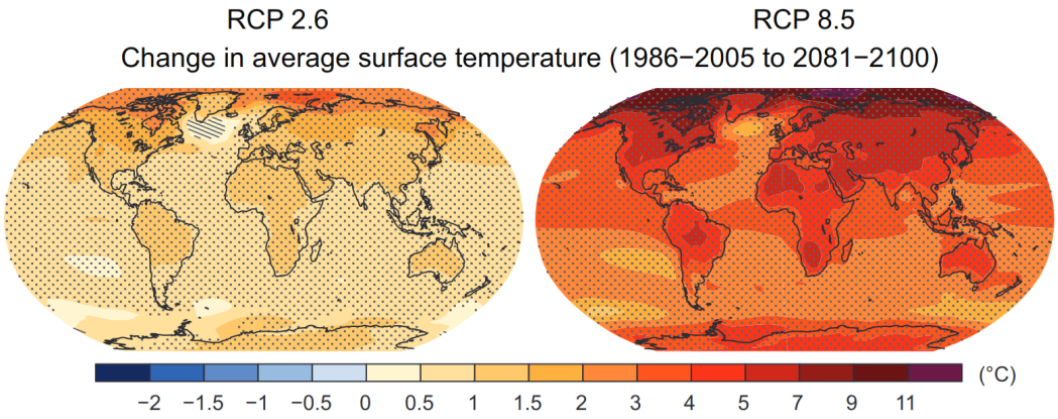

Figure 1.6: Predicted future change in surface temperature under low emission future scenario *RCP 2.6* and high emission future scenario *RCP 8.5*. *Source: IPCC, 2013.*

specific heat capacity of water is three times that of the land, meaning a larger amount of energy is required to increase the same mass of water by one degree as compared with the land. However, this effect alone cannot explain the disparity (e.g. Dong et al., 2009). The land sea warming contrast is associated with differing lapse rates (rate of vertical temperature change) between the land and ocean (Joshi et al., 2008). Over land, this rate is calculated as the dry, adiabatic lapse rate. However, when the atmosphere is saturated - which occurs more frequently over the ocean where relative humidity is higher on average - lapse rate is calculated as the saturated adiabatic lapse rate, which is less than the dry adiabatic lapse rate. As such, the average lapse rate over the ocean is frequently less than the average lapse rate over land. As a function of saturation specific humidity, the saturated adiabatic lapse rate is directly affected by future warming in accordance with the Clausius-Clapeyron equation (Held and Soden, 2006) and subsequently so is the average ocean lapse rate. As specific humidity increases, the saturated adiabatic lapse rate, and hence average lapse rate over the ocean, will decrease. Assuming uniform warming sufficiently high in the atmosphere (Joshi et al., 2008), these differing lapse rates point to a pattern of greater warming over the land surface than the ocean. There are also other contributing factors, for example a decrease in relative humidity over land as moisture availability becomes limited reduces cloud cover and allows for greater radiative heating of the surface (Manabe et al., 1992). The land-sea contrast in models is also sensitive to changes in large scale cloud representation and inclusion or exclusion of stomatal closure (Joshi et al., 2008).

Another pattern of future warming can be observed over the Arctic. Arctic temperatures have warmed more than twice as fast as the global mean over the last several decades (Rantanen et al., 2022), a phenomenon referred to as Arctic amplification (Manabe and Wetherald,

1980). A primary factor contributing to this is the surface albedo feedback (Hall, 2004; Serreze et al., 2009; Screen et al., 2018; Taylor et al., 2013). The surface albedo feedback arises from the difference in albedo between snow/ice and the land/ocean. As the climate warms, snow and ice melt exposes land and ocean surfaces which have a lower albedo and therefore reflect less radiation (Hall, 2004). The greater absorption of solar radiation by these surfaces leads to further heating, and subsequent snow/ice melt. Other proposed mechanisms contributing to the polar amplification warming pattern include ocean heat transport (Beer et al., 2020), cloud feedbacks (Taylor et al., 2013), lapse rate feedbacks (Stuecker et al., 2018) and Planck feedbacks (Pithan and Mauritsen, 2014).

While different parts of the world may be warming at different rates, generally warmer conditions result in a shifting of the entire temperature distribution, with an increase in both frequency and intensity of warm days and a corresponding decrease in the number and intensity of cold days as shown in Figure 1.7. The width of the distribution could also be affected by increased variability impacted by factors such as soil moisture and changing weather patterns (Perkins, 2015). Recently observed heatwave events have produced unprecedented temper-

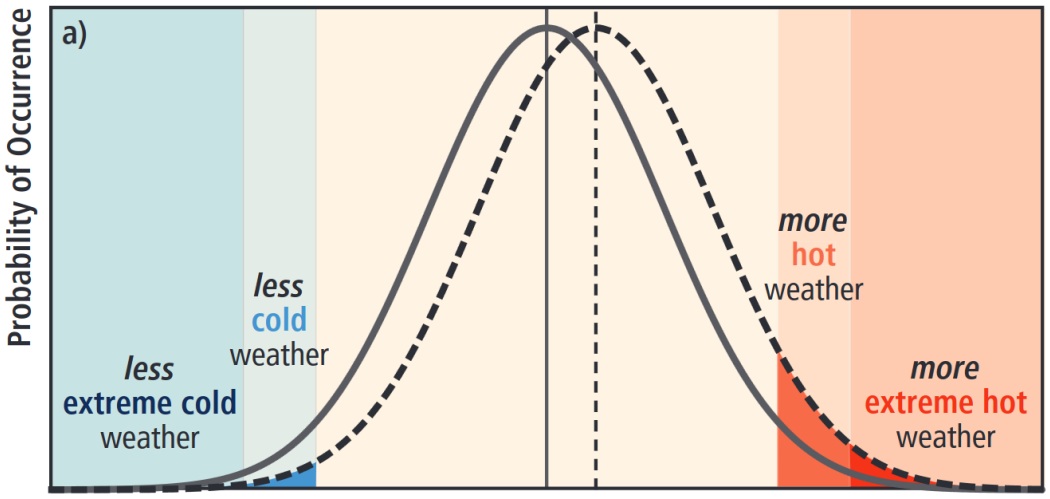

Figure 1.7: Illustration of the effect of a mean temperature shift on the frequency and intensity of hot days. *Source: IPCC, 2012.*

atures and broken heat records by substantial margins, for example the Pacific Northwest heatwave of 2021 (Philip et al., 2021; White et al., 2023). This trend for increasing intensity of extremes is projected to continue under future warming with the probability of longer duration, record breaking heatwaves becoming more than 3 times the present day value by the end of the century (Fischer et al., 2021). There is evidence that the trends for heatwave duration and frequency are not only increasing but accelerating (Perkins-Kirkpatrick and

Lewis, 2020). To understand the observed and projected changes in heatwave intensity, the effects of global warming on the driving mechanisms behind such events are considered.

Future warming will impact land-atmosphere interactions such as soil moisture evaporation, vegetation interactions and albedo feedbacks. Increasing global mean temperatures result in a corresponding increase in atmospheric water vapour, according to the Clausius-Clapeyron relationship: each degree of warming results in an approximately 7% increase in atmospheric water vapour (Held and Soden, 2006). Particularly in the mid-latitudes, the soil drying trend associated with continued warming is expected to exacerbate the trend for increasing heatwave severity (Qiao et al., 2021; Lal et al., 2023), where lack of moisture in the soil reduces capacity for evaporative cooling and increases sensible heat fluxes. At higher latitudes, the reduction in seasonal snow cover that comes with a warmer climate forms part of a positive feedback with decreasing albedo and increasing surface temperature, amplifying the most extreme events, particularly in areas where snow was typically most abundant (Diro and Sushama, 2020). Effects of warming on plant transpiration and high carbon dioxide concentrations on stomatal resistance also contribute to an amplification of temperatures over land (Cao et al., 2010); a potential contributor to the land-sea warming contrast. These effects could also moderate the ability of vegetation to mitigate heatwave conditions with stomatal closure resulting in a significant decrease in cooling potential in the mid-latitudes (Zhao et al., 2023).

There are also dynamical effects to consider. As mentioned previously, atmospheric anomalies such as blocking contribute significantly to mid-latitude heatwave events (Pfahl et al., 2012; Sousa et al., 2018). Relatively poor representation of blocking in existing climate models poses a challenge for simulating the effect of climate change on blocking and the subsequent impacts on heatwave events (Woollings et al., 2018; Lupo, 2021). Projections from the older generation of CMIP3 models indicated a general reduction in blocking frequency under future climate change (Barnes et al., 2012) however this is not consistent with later CMIP5 models which predicted a more complex picture of regionally dependent changes in blocking intensity (Masato et al., 2013). Even in the latest generation of climate models, CMIP6, future blocking trends are not significant although models generally predict a decrease in blocking activity with the exception of the Ural mountains (Davini and D'Andrea, 2020). It is difficult to place a high degree of confidence in blocking projections from existing climate models, as such the influence of blocking on heatwaves under future climate change remains somewhat of an open question.

These observed and projected increases in heatwave frequency, duration, extent and intensity combined with the remaining uncertainties in future climate projections motivate the application of the ML method presented here to constrain uncertainties in future regional temperature projections and understand heatwave drivers in the present day climate. The

following subsection introduces the concept of ML in general, and its potential applications in climate science.

1.6 Machine Learning

While there is no single, agreed upon definition of machine learning (ML), here the term is used to refer to a subset of artificial intelligence where models learn patterns in data during training without context-specific programming. Throughout the training process, model parameters are tuned such that the model adapts to the use case and is able to make useful and generalisable predictions. Here, a supervised learning approach is taken: the model is exposed to correct sets of labelled input and output variables during training, and parameter tuning is directed by minimisation of a cost function which quantifies the difference between the current model prediction and the true value (Hastie et al., 2009). By contrast, in an unsupervised learning approach, the model is provided with unlabelled data in order to identify patterns or trends.

ML is increasingly being used in the field of climate science and in particular as a complement to existing climate modelling methods. Key advantages of an ML approach in the context of climate modelling include the ability of an ML model to learn directly from observations or reanalysis data, the ability to identify non-linear relationships and represent patterns in complex data. The remainder of this subsection summarises a selection of ML applications for climate modelling.

A key area where progress can still be made in GCMs is in the parameterisation of sub-grid scale processes, for example clouds, which contribute to model uncertainty. Physics-informed ML methods which can incorporate some prior knowledge of system physics can be applied in this area to emulate steps in these complex processes with the goal of replacing some or part of the existing parameterisation scheme (Schneider et al., 2017; Chantry et al., 2021). There is even progress being made towards purely data driven forecasting models in the weather modelling domain (Rasp et al., 2020; Schultz et al., 2021). A key advantage of ML in many of these use cases is computational efficiency which can make it a much more viable option than traditional approaches.

In terms of post-processing of GCM data, ML methods may also be applied in downscaling (Kashinath et al., 2021) - translating a lower resolution GCM output field to a higher resolution more useful for studying local phenomena. As discussed already, bias correction and uncertainty constraint are essential post-processing tasks when dealing with data from GCMs and ML also has applications here. Once again, the ability of ML models to learn directly from observations is a key advantage for these use cases. A more detailed review of

ML for these applications will be presented in later results chapters.

Here, a new approach is taken, applying ML algorithms to merge existing climate model output with observations to produce new temperature simulations. A key component of this method is to learn process-based relationships between temperature anomalies and their drivers which are derived from observations. The next subsection addresses the availability of such observational data sets and considers the use of reanalysis data as a proxy for observations.

1.7 Observations and Reanalysis

Detailed observational records of temperature date back to the mid-19th century when regular thermometer-based records began (Le Treut et al., 2007). These records are unevenly distributed and unsuitable for studying regionally impactful events at a global scale. The beginning of the satellite era in the 1970s represented a step closer to true, global observation often now conveniently compiled into reanalysis data sets.

Reanalysis can construct regularly-gridded, global data sets based on observational measurements both from satellites and in-situ data. Available observations are combined with short-range weather forecasts through data assimilation (Hersbach et al., 2020) to produce a comprehensive snapshot of the Earth's climate system at a particular point in time. While it may not be possible for reanalyses to precisely reconstruct true observations globally, they represent the most complete record of historical weather and climate that it is currently possible to produce. As such, reanalysis data are used as a proxy for observations in training the observations-based ML models presented here.

Prominently available global reanalysis data sets include ECMWF Re-Analysis (ERA) 5 (Hersbach et al., 2020), the Japanese 55-year Reanalysis (JRA-55) (Kobayashi et al., 2015), the Modern-Era Retrospective analysis for Research and Applications (MERRA) 2 (Bosilovich et al., 2015) and the Climate Forecast System Reanalysis (CFSR) (Saha et al., 2010). As reanalyses rather than true global observations, these data sets are subject to biases and errors which must be taken into consideration, particularly when they are being implemented in a climate model evaluation context. In this thesis, reanalysis data is used as a proxy for observations in order to train the process-based ML model to predict daily-mean temperature anomalies. The variables required to represent the necessary physical processes to model these temperature anomalies were not available at sufficient spatial/temporal resolution or spatial extent via direct observational measurements.

The next question to consider is which reanalysis data set to use. Reanalyses can be evaluated by comparison against true observational measurements. Although no single re-

analysis comes out on top across all measures of mean bias, variability and historical trends, ERA5 performs well in comparison with other reanalyses (*e.g.* Graham et al., 2019; Ramon et al., 2019; Arshad et al., 2021) and is frequently used in the literature in place of observations (*e.g.* Kharin et al., 2013; Sillmann et al., 2013; Qiao et al., 2021; Schumacher et al., 2022; Röthlisberger and Papritz, 2023). Additionally, MERRA2 is only available from 1980 onwards and does not directly assimilate 2m air temperature. As a relatively less used reanalysis, CFSR has been evaluated less than the other listed data sets and it is also uncoupled meaning ocean-atmosphere interactions are not directly modelled. Here, ERA5 is used as the basis for training the process-based ML model. The possibility of extending this work to compare results across reanalyses is considered in Chapter 6. The next subsection outlines the structure for the rest of the thesis.

1.8 Thesis Outline

The methods in this thesis are based on the application of a process-based ML model (Ridge-ERA5) trained on reanalysis data to learn observations-based relationships between inputs and temperature anomalies. These relationships are then combined with existing output from GCMs to construct novel observational constraints on temperature anomalies in present and future climates.

In Chapter 2 three ML methods - Ridge regression, Random Forest regression and Gaussian Process regression are introduced. The suitability of these methods for the intended application of constraining climate model uncertainty is assessed. The reanalysis and climate model data used to obtain the results throughout this thesis are introduced, including necessary pre-processing steps. A range of potential predictor variables are tested to find an optimum variable setup for the process-based ML model to predict daily-mean temperature anomalies.

In Chapter 3, performance of Ridge-ERA5 on held-out test reanalysis data is assessed as well as testing the limits of the models' predictive skill in terms of rarer and more extreme events. Ridge-ERA5 is then applied to historical data from CMIP6 models for bias correction of summertime temperature anomalies. Coefficients learnt by Ridge-ERA5 from reanalysis data are combined with inputs from historical CMIP6 data to produce 'observationally constrained' historical temperature distributions across the Northern Hemisphere. The Ridge-ERA5 method is also applied in a climate model evaluation context, comparing coefficients learnt by Ridge from ERA5 with those that Ridge can learn from historical climate model data. Patterns of coefficients across the Northern Hemisphere are plotted, producing unique fingerprints for models which can be compared with Ridge-ERA5. Total distance between

Ridge-ERA5 and the CMIP6 Ridge emulators is calculated as a performance metric against which end of century future warming is plotted. These results motivate the application of Ridge-ERA5 for future uncertainty constraint in the following chapter.

In Chapter 4, Ridge-ERA5 is applied to constrain uncertainty in future projections of regional temperature from existing climate models. Prior to this, to investigate the question of climate invariance (do the relationships learnt by Ridge from historical data hold in the future under climate change?) a set of Ridge-CMIP emulators are trained on historical CMIP6 data. End of century temperature predictions from these Ridge-CMIP emulators are compared with raw temperature anomaly projections from the corresponding CMIP6 model. To construct a novel observational constraint on future temperature anomalies, inputs from the high emissions scenario *SSP585* in the CMIP6 archive are combined with the observations-based coefficients from Ridge-ERA5. The final observational constraint consistently excludes models which project the highest degree of regional warming across Northern hemisphere regions, potentially suggesting that the degree of warming projected by these particular models is incompatible with observational evidence.

Chapter 5 focuses on extreme events. SHapley Additive exPlanation (SHAP) values are used to analyse contributions from individual input variables to Ridge-ERA5 predictions during historical heatwaves. The inferences that can be formed from this analysis are compared with existing literature. SHAP contributions are compared between the Ridge-ERA5 model and Ridge-CMIP emulators, showing how the contributions of each predictor variable vary for temperature anomalies in the extreme tails of the temperature distribution in different regions. Additionally SHAP value contributions to temperature anomaly predictions by Ridge-CMIP emulators are compared across past and future climates. This analysis is intended to aid in the interpretability of the Ridge-ERA5 predictions and to confirm that predictions are consistent with understood physical mechanisms.

Finally, Chapter 6 contains general discussion and conclusions with suggestions for extensions and future work.

2 Methods

The methods presented in this thesis are based on the idea of applying observations-based ML models to existing global climate model (GCM) near-surface temperature output during Northern Hemisphere summer. This methods chapter begins with a general introduction to supervised learning and the training process followed by an overview of the three ML methods tested for these applications: Random Forest regression; Gaussian Process regression; and Ridge regression. A series of preliminary tests are carried out using reanalysis data to determine which of these three methods is most appropriate for the intended applications. A range of climate predictor variables are tested to produce a final variable setup used to train the ML model.

2.1 Introduction to Machine Learning

Machine learning (ML) refers to a wide range of methods with foundations primarily in statistics and computer science; what connects them all is the ability of the model to learn from the data alone. Predictive skill comes from the progressive adjustment of trainable parameters directed by the aim of incrementally reducing error in the algorithm's predictions. As such, ML is now finding widespread applications in data-driven modelling across many scientific disciplines, including climate science (*e.g.* Herman and Schumacher, 2018; Karpatne et al., 2019; Barnes et al., 2020; Pan et al., 2021). This section introduces the fundamentals of supervised ML techniques including: training (Section 2.1.1), the bias-variance trade off (Section 2.1.2), cross-validation (Section 2.1.3), normalisation (Section 2.1.4) and performance measures (Section 2.1.5).

2.1.1 Supervised Learning

Supervised learning is an approach to training an ML model which exposes the model to labelled example or 'training' cases so that it can learn relationships between the predictor variables and the target variable(s). Subsequently, these relationships, or predictive functions, can then be used to make accurate predictions given new predictor values as inputs.

In supervised learning, data is separated into a training set (which is further sub-divided for cross-validation to tune hyperparameters in ML models, see Section 2.1.3) and a test set. The ML model chosen will be defined by a function which takes values of the predictor variables and any model parameters as inputs in order to produce a prediction for the target output variable. For a training set comprised of N samples each with values for P predictor variables, the following procedure is carried out for each sample, $i = 1, 2, ..., N$, in the training

set:

1. The predictor variables, $\mathbf{x}_i$, are passed into the model using the output function to get a prediction, $\hat{y}_i$, for the target variable, y_i.

2. The prediction, $\hat{y}_i$, and the actual value, y_i, are passed into the cost function which quantifies the error in the prediction.

3. The error is distributed between the trainable parameters of the model according to their influence on the output and is used to make an adjustment to each trainable parameter with the aim of improving prediction accuracy.

The repetition of this process across all samples in the training set should result in sufficient adjustment of the trainable parameters to enable the ML model to make reasonable predictions. However, as discussed in detail for the three regression algorithms outlined later in the chapter, highly flexible ML functions can eventually suffer from an over-adjustment to the training data, a phenomenon known as *overfitting*. To find the best predictive function, a typical training process therefore usually requires additional constraints on the learning process in order to find the optimal complexity function. An example of such a constraint is L2-regularisation discussed in the context of Ridge regression in Section 2.2.1. The general issue of finding the optimum complexity function that can represent the variance in the data but does not overfit, is commonly referred to as the bias-variance trade off (Bishop, 2006) which is discussed in more detail in the following subsection.

2.1.2 The Bias-Variance Trade Off

The bias-variance trade off refers to the idea of finding a balance between a model which does not overfit to the training data set - so remains generalisable to unseen test data - but is not underfit to the extent that the predictions are ineffective for the particular use case. These ideas are demonstrated in Figure 2.1 which plots error from model bias and variance against model complexity as well as total error - the combination of variance and bias plus irreducible error.

The left-hand side of Figure 2.1 indicates an underfitting regime where error is dominated by model bias. Essentially, the model has not been tuned sufficiently for the use case resulting in poor predictions for training and test data. Conversely, the right-hand side of Figure 2.1 illustrates an overfitting regime. Here, the model has been tuned too closely to the training data set and, whilst training scores will be high, the model does not generalise well to unseen test cases. Optimal model complexity is found at the boundary between these two regimes

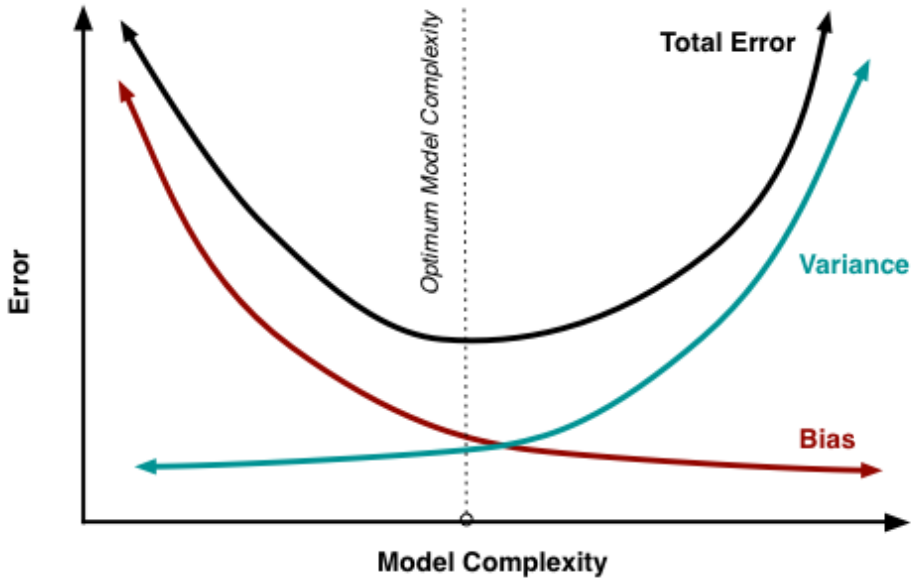

Figure 2.1: Illustration of the bias-variance trade off. Tuning a model of optimum complexity requires a balance between error resulting from model bias and error resulting from overfitting to training data. *Source: Fortmann-Roe, 2012.*

where error from bias and variance can be minimised. A key step to achieving this is the tuning of model hyperparameters which happens via cross-validation.

2.1.3 Cross-Validation

An additional level of complexity is added to the training process with the inclusion of cross-validation. Cross-validation is a process which allows the optimum value of the tunable parameters - hyperparameters - in the ML model to be determined by testing out different hyperparameter values. Tuning of hyperparameters is an essential step to help guide the complexity of the model and find the best generalising predictive function. This is closely linked to the idea of the bias-variance trade off - the wrong choice of hyperparameter for a particular ML model and data set can result in over or under fitting. Once the optimum value of hyperparameters has been determined in cross-validation, they do not change throughout the training process. Values are set when the model is initiated and do not change after this.

A standard method to find the best set of hyperparameters for a given ML function is k-fold cross-validation (Hastie et al., 2009). In this method, the training data is separated into k subsets or folds, see Figure 2.2. For each value under consideration for each hyperparameter, the model is trained k times, each time omitting a different one of the k subsets to use for

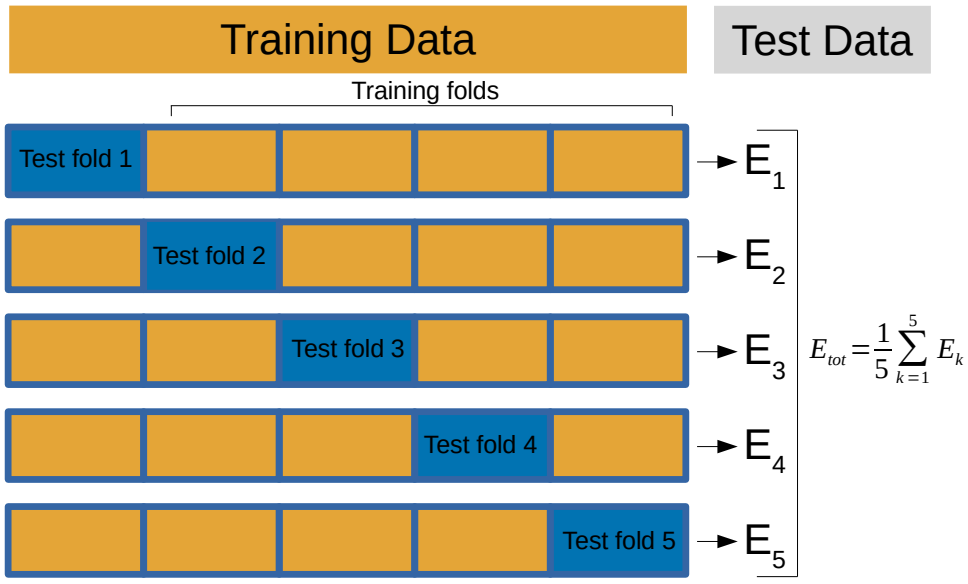

Figure 2.2: Illustration of a 5-fold cross-validation. Training data is divided into 5 subsets or folds, each of which is sequentially omitted for independent cross-validation. ML models are trained on the remaining 4 folds using different values of model hyperparameters which are tested on the unseen fold. The cross-validation scores, E_k, are averaged over the 5 folds to get final scores for each hyperparameter value.

testing. This allows k models to be trained and cross-validated independently on the same data. Each time, the model performance is quantified based on its performance on the omitted fold which was not seen during training. The cross-validation scores (E_k in Figure 2.2) can then be averaged over the k models trained for each possible value of the hyperparameter, and the optimum hyperparameter values are selected based on these cross-validation scores.

Typically, the final model is then trained on the training data set in its entirety using the optimum hyperparameter setup identified during cross-validation. A test score is then obtained using the independent test set that has not been used during training or cross-validation up to this point.

2.1.4 Normalisation

Normalisation is typically applied to predictor variables prior to inputting them to the ML model. The aim of normalisation is to transform predictor variables to a comparable scale (Ali and Faraj, 2014), which is, for example, important in Ridge regression where predictor variables with larger value ranges would otherwise be penalised less by the L2-Regularisation (see Section 2.2.1). Normalisation can improve performance with some ML algorithms because of the assumptions they make about the input data and also aids in the interpretability

of trained coefficients for linear methods.

A standard normalisation transforms a variable to zero mean and unit standard deviation:

$$x'_j = \frac{x_j - \mu_j}{\sigma_j}. \tag{1}$$

Here, a specific value of the j^{th} feature variable, x_j is transformed by subtracting the mean value (μ_j) of x_j over the training data set and dividing by the standard deviation (σ_j) of the training values for x_j.

An alternative to standard normalisation is a min-max normalisation:

$$x'_j = \frac{x_j - min_j}{max_j - min_j}. \tag{2}$$

Here, x_j is normalised relative to the minimum and maximum values of the variable in the training data set, min_j and max_j respectively. Compared with standard normalisation, a min-max procedure is much more sensitive to outliers since the bounds of the data (min and max) determine the range of the rescaling. A min-max normalisation may be preferable to a standard normalisation in instances where a defined final range of values is required as a min-max normalisation produces values in the range zero to one by definition.

2.1.5 Performance Measures

An ML model has two purposes; to accurately predict the target variable but also to provide structural information about the relationship between predictors and output (Breiman et al., 1984). The trainable parameters learnt by a model can provide new information about potentially causal relationships between predictor and target variables. The success of this aspect of the model depends, inter alia, on how interpretable the model is.

In terms of quantifying prediction accuracy there are a variety of measures of performance, for example the Mean Absolute Error (MAE):

$$MAE(y_i, \hat{y}_i) = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i|. \tag{3}$$

MAE quantifies the average magnitude of difference between predictions and their true value.

Another measure is the R^2 score, a method of comparing the variance explained by the model with a mean guess model (Pedregosa et al., 2011):

$$R^2(y_i, \hat{y}_i) = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{4}$$

where $\bar{y}$ indicates the mean value of the target variable across training samples. A perfect model would have a maximum, optimum R^2 score of 1 whilst a model equivalent to a mean

guess would have an R^2 score of 0. A negative R^2 score indicates that the model performs worse than simply guessing the mean value of y for all test cases.

The Pearson correlation coefficient measures the strength of relationship between two continuous variables:

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x}) - \sum_{i=1}^{N}(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}, \tag{5}$$

where $\bar{x}$ and $\bar{y}$ indicate the mean of values x_i and y_i respectively. A positive or negative Pearson correlation coefficient corresponds to a positive or negative correlation between variables while the magnitude of the coefficient (up to a maximum magnitude of 1) indicates the strength of that correlation. A high correlation coefficient score indicates that both samples (x_i and y_i) increase and decrease at the same time but without accounting for the amplitude of these increases. Alternatively, the R^2 score also accounts for the magnitude of the increase or decrease giving a fuller picture of the proportion of variation that is explained by the model's predictions.

2.2 Machine Learning Methods

Three ML algorithms were tested as the basis for the method applied in this thesis: Ridge regression; Random Forest Regression; and Gaussian Process Regression. Having introduced the fundamentals of training ML models in the previous subsection, the functions associated with these models are now explained in further detail.

2.2.1 Ridge Regression

Ridge regression (Hoerl and Kennard, 1970) is a simple, interpretable method that deals relatively well with many highly correlated input variables (*e.g.* Dormann et al., 2013). As with any regression approach, the aim is to quantify the relationship between the independent and dependent variables. This relationship is assumed to take the form:

$$y = f(\mathbf{x}) + \epsilon. \tag{6}$$

Where the dependent variable, y, is modelled as a function of $\mathbf{x} = (x_1, x_2, ..., x_p)$, a p-dimensional vector of predictor variables with residual error, ϵ.

In Ridge regression, the relationship between the j^{th} predictor variable, x_j, and the target variable, y, is characterised by a single coefficient, β_j, as in a simple multiple linear regression:

$$\hat{y} = \beta_0 + \sum_{j=1}^{P} x_j \beta_j. \tag{7}$$

The difference between ordinary least squares linear regression and Ridge is that the Ridge regression cost function contains an L2-regularisation term which additionally penalises the total coefficient magnitudes:

$$E = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{P}\beta_j^2, \tag{8}$$

where N is the number of samples in the training data set, P is the number of predictor variables and λ is the regularisation parameter which is a hyperparameter that is tuned to adjust the influence of the additional term. This helps to prevent overfitting (see Section 2.1.2 on the bias-variance tradeoff) by evenly distributing predictive power amongst correlated input variables. The first term in Equation 8 quantifies the mean squared error across the training samples and the second term measures total squared coefficient magnitude. The influence of the second term can be tuned by adjusting the value of hyperparameter λ. The optimum value of λ is determined through a cross-validation procedure (Stone, 1974) after which the final version of the model can be trained with the entire training data set.

2.2.2 Gaussian Process Regression

Gaussian processes are a probabilistic supervised learning method which may be applied in a regression or classification context (Williams and Rasmussen, 1995). Gaussian processes are versatile methods (through selection of different kernels) and provide probabilistic predictions allowing for computation of empirical confidence intervals. As with the Ridge regression method outlined above, the aim with Gaussian Process Regression (GPR) is to calculate a function to describe the relationship between dependent variable y and predictor variables, $\mathbf{x} = (x_1, x_2, ..., x_p)$. In theory there are an infinite number of such functions, GPR aims to define a distribution over these functions, the mean function of which is then used for regression predictions.

GPR is a Bayesian approach in the sense that the kernel or prior distribution of possible functions is selected before any calculations with training data. When the method is applied to training data, the functions which do not fit with the training data points can be discarded which leads to the posterior - the distribution of functions that actually agree with the observed data points. Put simply, the posterior is the prior updated with information from observed data points. So, to apply this method, a prior over the function space needs to be specified, and Gaussian processes can be used for this task.

A Gaussian process (GP) is a collection of random variables such that any subset of variables will have a joint Gaussian distribution. An n-dimensional variable $\mathbf{x}$ following a

multivariate Gaussian distribution has a joint probability density given by:

$$p(\mathbf{x}|\mu, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^{n/2}|\mathbf{\Sigma}|}} \exp(-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)). \tag{9}$$

In the case of GPR, each random variable is a function, $f(\mathbf{x})$. Mathematically speaking, a GP is fully quantified by its mean, μ, and covariance, Σ, functions:

$$\mu(\mathbf{x}) = E[f(\mathbf{x})], \tag{10}$$

$$\Sigma(\mathbf{x}, \mathbf{x}') = E[(f(\mathbf{x}) - \mathbf{x})(f(\mathbf{x}') - \mathbf{x}')]. \tag{11}$$

This means that the GP is defined by the mean and covariance of the possible functions in the function space:

$$f(\mathbf{x}) \sim GP(\mu(\mathbf{x}), \Sigma(\mathbf{x}, \mathbf{x}')). \tag{12}$$

Prior knowledge about the expected function space can be incorporated through the choice of the mean and covariance functions. Often the mean of the distributions is assumed to be zero or a constant value such as the mean of the training data. The covariance function or kernel is defined such that the function value of two points which are similar in kernel space will also be similar to each other.

Specification of the covariance function naturally implies a distribution over possible functions. This can be demonstrated through the example of a Radial Basis Function (RBF), defined by:

$$\Sigma(\mathbf{x}, \mathbf{x}') = \exp(-\frac{1}{2\sigma^2}||\mathbf{x} - \mathbf{x}'||^2). \tag{13}$$

Assuming that the mean of the GP is zero, samples can be drawn from the distribution of functions evaluated at any number of input points. Figure 2.3 shows five such functions sampled at random from the GP prior with RBF kernel. The next step is to condition this prior on a set of observations.

As stated previously, GPR takes a Bayesian approach where a prior assumption is updated with training information to form the posterior according to Bayes' Theorem:

$$\text{posterior} = \frac{\text{likelihood x prior}}{\text{marginal likelihood}}. \tag{14}$$

In the context of GPR, the posterior is used to make predictions, $\mathbf{y_b} = f(\mathbf{x_b})$, based on the Gaussian process prior updated with training data points, $(\mathbf{x_a}, \mathbf{y_a})$. From the properties of Gaussian distributions, $\mathbf{y_a}$ and $\mathbf{y_b}$ are jointly Gaussian since they both come from the same multivariate Gaussian distribution. For a finite number of samples it thus follows that:

$$\begin{pmatrix} \mathbf{y_a} \\ \mathbf{y_b} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\mathbf{a} \\ \mu_\mathbf{b} \end{pmatrix}, \begin{pmatrix} Cov(x_a, x_a) & Cov(x_a, x_b) \\ Cov(x_b, x_a) & Cov(x_b, x_b) \end{pmatrix}\right). \tag{15}$$

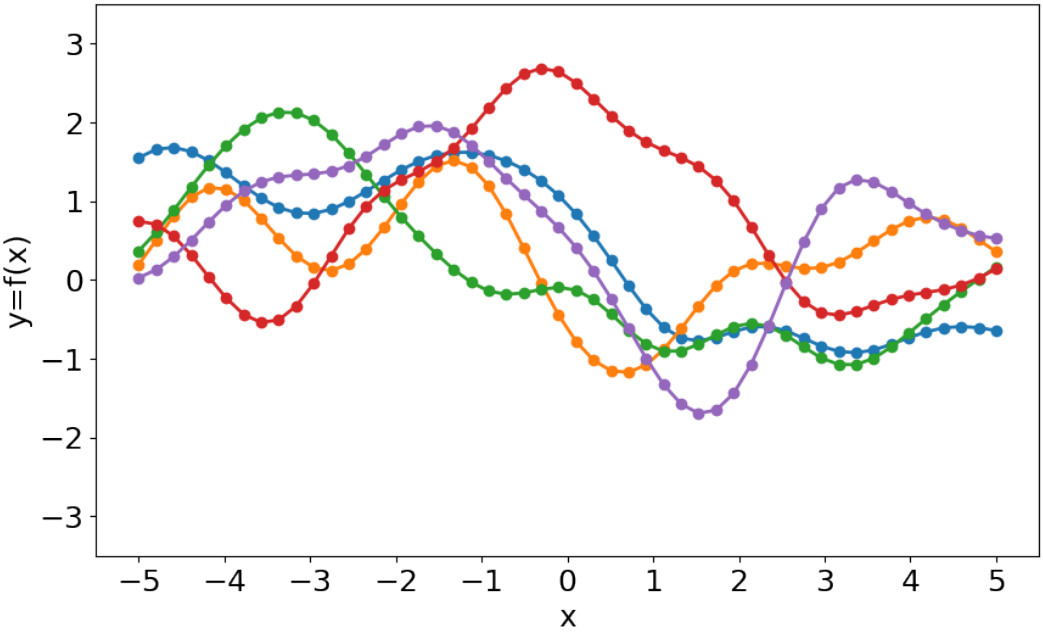

Figure 2.3: Five functions drawn at random from a Gaussian Process prior with zero mean and Radial Basis Function kernel evaluated at 50 equidistant points covering the x-range -5 to 5.

If there are n_a training points and n_b test points then $\mathbf{Cov}(\mathbf{x_a}, \mathbf{x_b})$ represents an $n_a \times n_b$ matrix of covariance pairs between all training and test samples. Similarly for $\mathbf{Cov}(\mathbf{x_a}, \mathbf{x_a})$, $\mathbf{Cov}(\mathbf{x_b}, \mathbf{x_a})$ and $\mathbf{Cov}(\mathbf{x_b}, \mathbf{x_b})$. To obtain the posterior, the prior distribution of functions is restricted to only those functions which are compatible with the observed training data points. This corresponds to conditioning the joint Gaussian prior distribution on the observations:

$$p(\mathbf{y_b}|\mathbf{y_a}, \mathbf{x_b}, \mathbf{x_a}) = N(\mu_{\mathbf{b|a}}, \mathbf{\Sigma_{b|a}}), \tag{16}$$

where;

$$\mu_{\mathbf{b|a}} = \mu_{\mathbf{b}} + \mathbf{\Sigma_{ba}}\mathbf{\Sigma_{aa}^{-1}}(\mathbf{y_a} - \mu_{\mathbf{a}}), \tag{17}$$

which can be further simplified if the prior distribution is assumed to have zero mean, and;

$$\mathbf{\Sigma_{b|a}} = \mathbf{\Sigma_{bb}} - \mathbf{\Sigma_{ba}}\mathbf{\Sigma_{aa}^{-1}}\mathbf{\Sigma_{ab}}. \tag{18}$$

Applying this to the example from Figure 2.3 for four different sets of training observations illustrates how the same prior distribution assumptions will lead to different posterior distributions based on the training data, see Figure 2.4.

Choice of kernel affects the patterns that the functions included in the prior will be able to capture. This means that different choices of kernel would be required to, for example, capture harmonics *vs.* trends. Figure 2.5 gives some examples of kernels and combinations of kernels which can produce different function shapes.

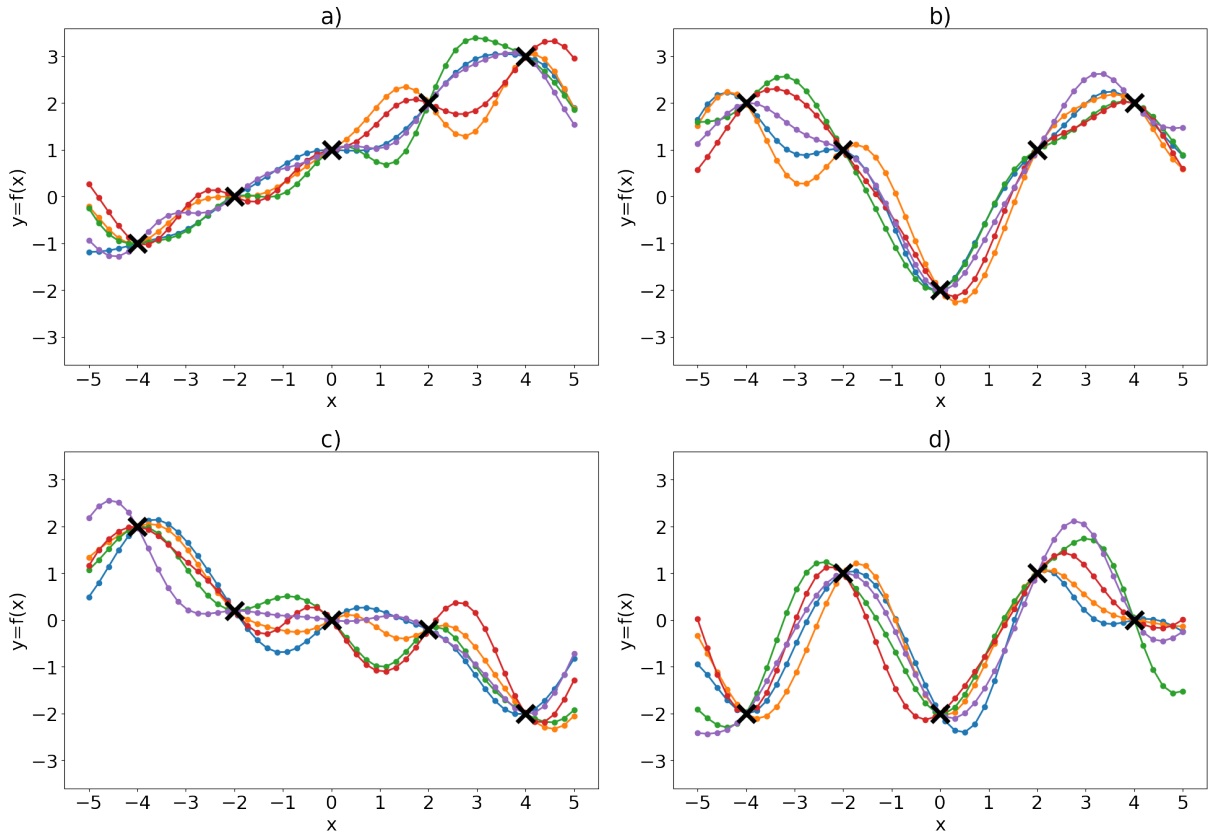

Figure 2.4: Five functions (coloured lines) drawn at random from a Gaussian Process posterior obtained by conditioning the same prior on for four different sets of training observations (black crosses).

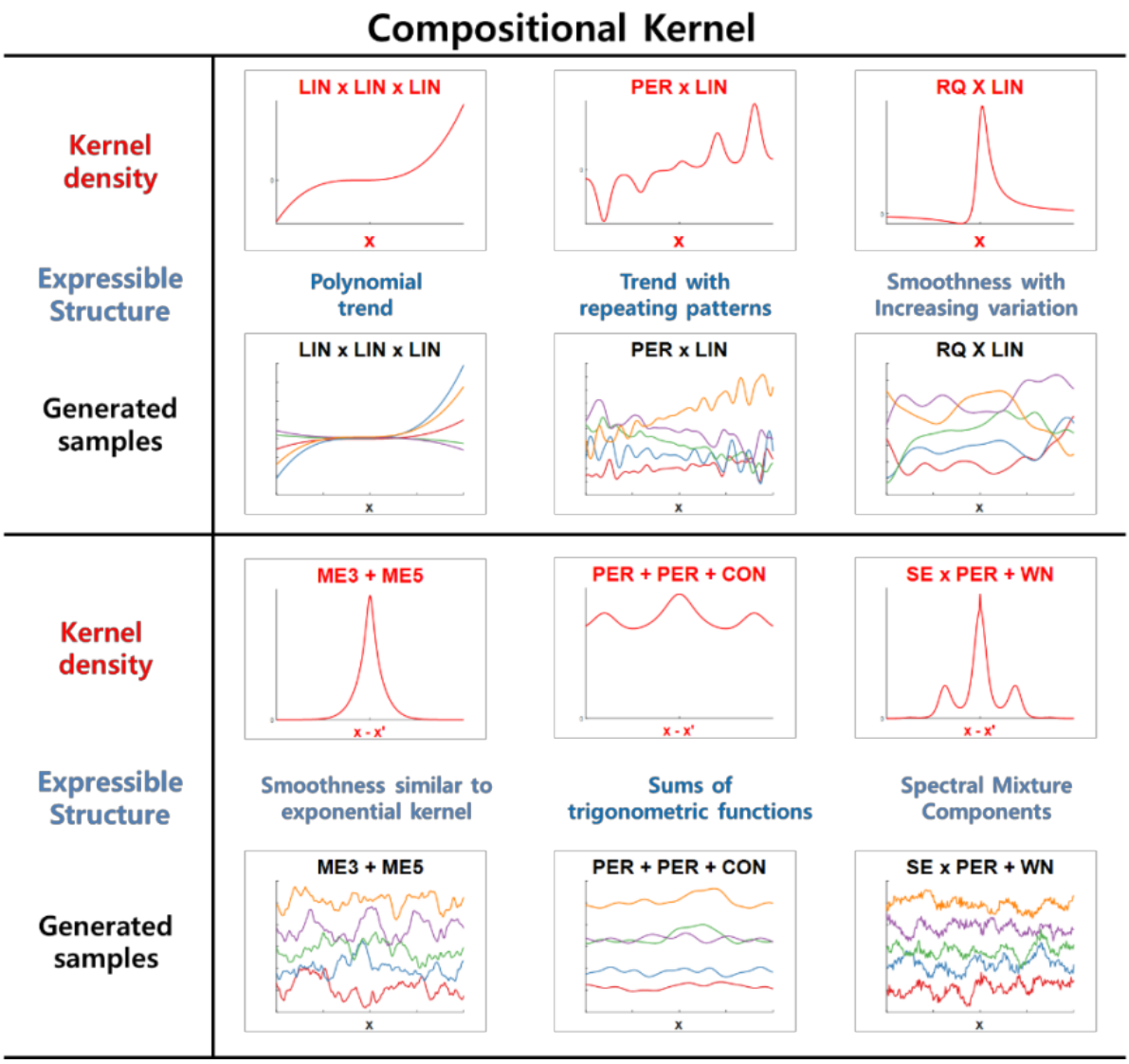

Figure 2.5: Demonstration of functions that can be modelled using GPR by choosing different combinations of kernels. (WN=white noise; CON=constant; LIN=linear; PER=periodic; SE=squared-exponential; RQ=rational quadratic; ME3/ME5=Matérn). *Source: Jin, 2020.*

2.2.3 Random Forest Regression

Similarly to Gaussian processes, Random Forest models can be used in both a regression and a classification context. Random forests are ensemble learning methods and are commonly referred to as a highly interpretable ML approach.

A Random Forest (RF) is a collection of decision trees over which predictions are aggregated to give the final RF prediction (Ho, 1995; Breiman, 2001). A decision tree can be drawn like a decision chart or flow chart - a series of conditional branching points leading down to the *leaves* of the tree. At each branching point the 'best split' is determined by minimisation of the cost function. This process continues iteratively, partitioning the training data set further and further until stopping criteria, for example maximum branch depth, is reached. A simple procedure to generate a decision tree is to use the sum of squared errors to determine the optimum branching points. At each node, the resulting sum of squared errors for all possible binary splits of each predictor variable is calculated. The split that would result in the greatest reduction of total tree error is then selected. This process continues for subsequent branching points until either all samples at the node have the same y value or the stopping criteria are satisfied.

To demonstrate this, an illustrative decision tree is created to predict temperature from soil moisture, sea level pressure, relative humidity and cloud cover anomalies, shown in Figure 2.6. The soil moisture variable has been selected for the first branching point partitioning

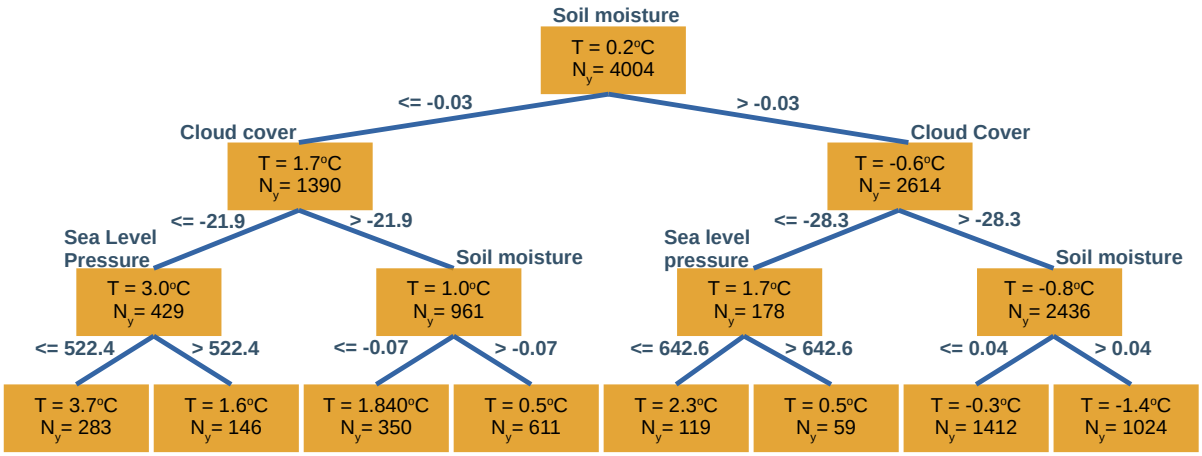

Figure 2.6: Illustrative decision tree with maximum depth of 3 for prediction of temperature from soil moisture, sea level pressure, relative humidity and cloud cover. N_y indicates the number of training samples at each node and T indicates the average temperature anomaly of those samples.

the training data set into a group of 1390 samples with soil moisture anomalies less than or

equal to -0.03 and a second group of 2614 samples with a soil moisture anomaly more positive than -0.03. This partitioning continues to a maximum depth of three resulting in eight terminal nodes or 'leaves'. T in Figure 2.6 indicates the average temperature anomaly of samples collected at a particular point in the tree.

Subsequent predictions are made by inputting the predictor variable values for each sample into the tree. The value of these variables will determine which terminal node or leaf of the tree the sample is filtered into. To make a prediction for this sample in a regression use case, the target variable values for training cases in the same terminal node as the test sample are averaged. Taking the example tree in Figure 2.6, for a test sample with soil moisture less than -0.03, cloud cover less than -21.9, and sea level pressure less than 522.4; the test sample will be filtered into the left most terminal node. Training samples which ended up in this node had a mean temperature anomaly of 3.7^oC, so this value is taken as the prediction for subsequent test cases which meet the same criteria.

There are drawbacks to using a single decision tree for a regression problem. Decision trees have a tendency towards overfitting - they will perform very well on training data that they have learnt from but do not generalise well to novel test data (Ho, 1995). Also, a single decision tree provides discrete, non-continuous predictions, whereas a smoother prediction range can be achieved by averaging the predictions of multiple trees. This is where the idea of ensemble learning comes in, a collection of trees are trained on different subsets of predictor variables and training samples. Sub-selection of different predictor sets for different trees, also called 'feature bagging', aids in de-correlation of individual trees (Hastie et al., 2009). Bootstrapping helps to prevent overfitting by training each tree on a different randomly sampled (with replacement) subset of the available training data (Hastie et al., 2009).

Other model parameters that can be tuned via cross-validation include the maximum tree depth (the number of branching points before the terminal node), the number of trees in the forest, the maximum number of features involved at each branching point and the minimum number of samples at each terminal node.

In the case of RF regression, the mean prediction across all decision trees constitutes the final prediction:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} f_b(\mathbf{x}), \tag{19}$$

where f_b represents the b^{th} decision tree from a total of B trees. Uncertainty in this prediction can then be quantified by the standard deviation in predictions across all trees:

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(\mathbf{x}) - \hat{y})^2}{B-1}}. \tag{20}$$

2.3 Method Development

The key results of this thesis stem from the application of a process-based ML model which learns from reanalysis data and, separately, climate simulations conducted by a range of climate models. Key desirable features of the final model are interpretability, physical consistency and the ability to extrapolate the learned relationships under significantly changing climatic conditions. In order to infer heatwave drivers, and to have confidence that relationships learnt by the model during training are consistent with physical processes, it was important that the contribution of predictor variables to final predictions could be quantified and directly compared. Secondly, the ML model should be able to make good predictions for the right reasons - the modelled relationships between predictors and target temperature anomalies should be consistent with our understanding of the physical processes of the climate system. Finally, in order to apply the observations-based model for constraint of future climate projections, the model must have the ability to extrapolate beyond the range of temperatures it has been exposed to during training, a requirement also referred to as climate invariance (e.g. Beucler et al., 2021; Nowack et al., 2023). In the following subsections, the reanalysis and climate model data sets are introduced, ML methods are compared and a range of predictor variables for modelling Northern Hemisphere, summertime, near-surface temperature anomalies are tested and selected. These tests will provide the baseline model used in the three application results chapters on: climate model evaluation and observational constraints on historical climate model simulations (Chapter 3); observational constraints on future regional changes in temperature (Chapter 4); and case studies of extreme heatwave events and their interpretation (Chapter 5).

2.3.1 Reanalysis Data

The following subsection outlines the reanalysis data sources used including where the data was accessed and how it was processed. The ECMWF Reanalysis 5th Generation (ERA5) data set (Hersbach et al., 2020) is used as a proxy for observations in order to train the process-based ML model and apply observations-based constraints on climate model simulations. The ERA5 data was accessed from the Corpernicus Climate Data Store (Hersbach et al., 2023*a*,*b*) at hourly time resolution and $0.25^o \times 0.25^o$ spatial resolution. The data was re-gridded to a $3^\circ \times 3^\circ$ longitude-latitude grid and daily-mean values were calculated using Climate Data Operators (CDO) (Schulzweida, 2022) commands *remapcon* and *dailymean*.

For the purpose of this analysis, the historical period is defined as 1979-2022 to match the period of the reanalysis data and therefore maximise the volume of training data within the satellite era. In order to focus on warm temperature anomalies and heatwave events,

only summertime data was considered, here defined as the months of June, July and August (JJA) in addition focusing the analysis spatially on the Northern Hemisphere. Land grid cells were selected via a land area fraction threshold derived from ERA5. Figure 2.7 shows a map of land area fraction across the Northern Hemisphere. Here land grid cells are defined as grid cells with a land area fraction exceeding 0.7.

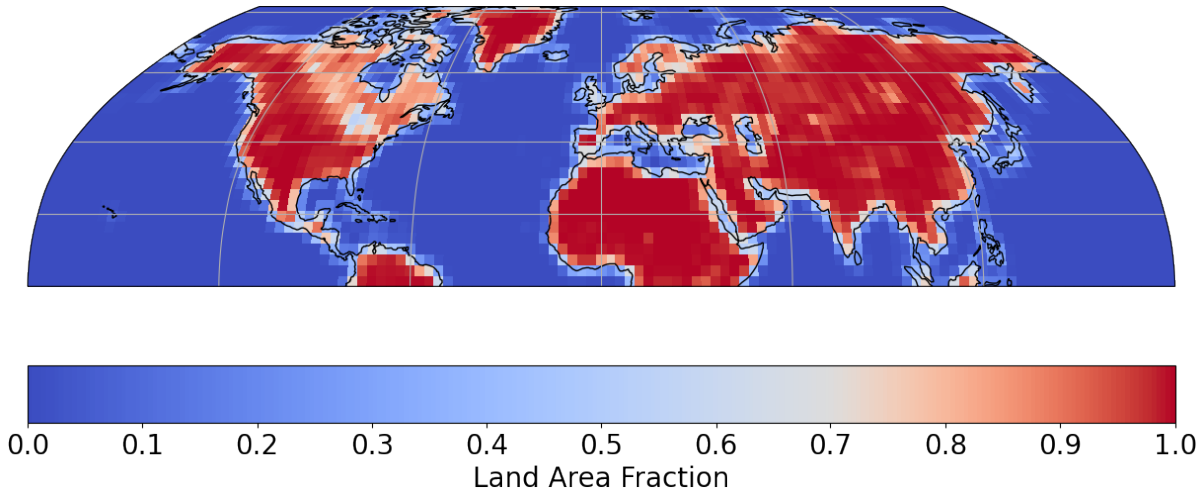

Figure 2.7: Land area fraction from ERA5 across the Northern Hemisphere, grid cells with a land area fraction < 0.7 are excluded from analysis (blue) while grid cells with a land area fraction > 0.7 are included (red).

A smoothed mean seasonal cycle was subtracted from all data across the historical and future periods. This removes the possibility of having trivial predictive skill by only predicting the seasonal cycle historically. The seasonal cycle was calculated via a centred, running-mean of 31 days over the time period 1979-2022 which was then averaged on each day of the calendar year to produce a smoothed seasonal cycle. Figure 2.8 shows this seasonal cycle for 2m temperature at an example location over Western Europe. These smoothed seasonal cycles are calculated for each variable at every grid cell location so that all variables are measured as anomalies to the seasonal cycle.

Most variables are available directly in the ERA5 data set, however near-surface relative humidity variables have to be calculated from other variables. Relative humidity, R, is calculated as the ratio of the partial pressure of water vapour, e, to the saturation partial pressure, e_{sat}, (ECMWF, 2016):

$$R = \frac{e}{e_{sat}(T)}, \tag{21}$$

which is multiplied by 100 to obtain the value as a percentage. The partial pressures are

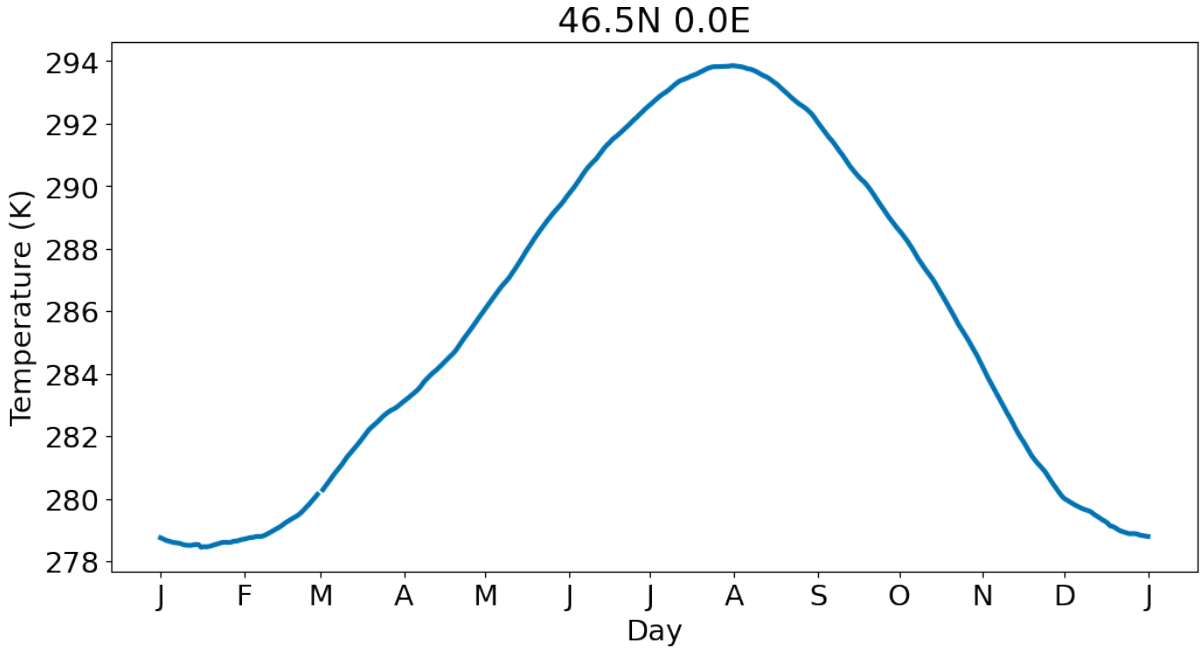

Figure 2.8: Smoothed seasonal cycles for 2m temperature at 46.5N 0E.

derived from dew-point temperature, T_d, and near-surface temperature, T, according to:

$$e = a_1 \exp\left(a_3 \frac{T_d - T_0}{T_d - a_4}\right), \tag{22}$$

and saturation water vapour pressure according to Teten's formula;

$$e_{sat} = a_1 \exp\left(a_3 \frac{T - T_0}{T - a_4}\right), \tag{23}$$

where $a_1 = 611.21$, $a_3 = 17.502$, $a_4 = 32.19$ and $T_0 = 273.15$.

2.3.2 Climate Model Data

The process-based ML model is designed as a novel way to combine information from Earth observations with climate model simulations. The model will be applied to CMIP6 data in order to impose observations-based constraints on climate model output, but can also be applied in the context of model evaluation to compare coefficients learnt from reanalysis with those that can be learnt from historical climate model data. Climate model data from the Coupled Model Intercomparison Project (CMIP) Phase 6 (Eyring et al., 2016) was accessed from the Centre for Environmental Data Analysis (CEDA) via JASMIN and the Earth System Grid Federation (ESGF). In the first instance, all CMIP6 models for which the complete set of predictor variables and 2m temperature were available for both the historical forcing

scenario and future warming scenario following Shared Socio-Economic Pathway (SSP) 585 (see Section 4.1.1 for more details) were selected. CMIP6 data was accessed as daily-mean values at native resolution of each model, see Table 1 for these resolutions, then re-gridded to $3^\circ \times 3^\circ$ resolution using the same conservative remapping procedure as for ERA5. This resolution was selected to be coarser than the coarsest resolution climate model data available which was $2.8^\circ \times 2.8^\circ$.

2.3.3 Problem Definition

Having introduced the data sets that will be used throughout the analysis in this thesis, this subsection defines the aims of the ML method to be trained. The idea is to learn a generalisable function, f_{era5}, that predicts daily-mean, near-surface temperature anomalies, y_{era5}, based on dynamic and thermodynamic predictor anomalies, x_{era5}. The choice of these predictors is motivated by them being proxies for processes known to influence temperature. For example: dynamical/large-scale weather patterns; soil dryness; humidity; and cloudiness (refer to Section 1.2 for further details). The process-based ML model which learns from re-analysis data (as a proxy for observations) should represent these true, physical relationships as authentically as possible. The relationships that the ML model learns will be quantifiable by the model parameters, θ_{era5}, which are learnt during training. Predictions ($\hat{y}$) of historical ERA5 temperature anomalies by the process-based, ML model will take the form:

$$\hat{y} = f_{era5}(\theta_{era5}, x_{era5}). \tag{24}$$

This ML model can then be combined with predictor anomalies from historical climate model data for bias correction (Chapter 3):

$$\hat{y} = f_{era5}(\theta_{era5}, x_{cmip6,hist}), \tag{25}$$

and combined with predictor anomalies from future projections of existing climate models under defined future warming scenarios to constrain uncertainty in future projections of temperature under climate change (Chapter 4):

$$\hat{y} = f_{era5}(\theta_{era5}, x_{cmip6,fut}). \tag{26}$$

The desire to include patterns of predictor anomalies rather than single, local values and the number of variables involved make it necessary to approach this as a high-dimensional regression problem. The following subsection tests the suitability of the three ML algorithms previously introduced (Ridge, RF and GPR) for this application.

Model Name	Resolution lat(^{o}) $\times$ lon(^{o})
ACCESS-CM2	1.3x1.9
ACCESS-ESM1-5	1.3x1.9
CanESM5	2.8x2.8
CMCC-CM2-SR5	0.9x1.3
CMCC-ESM2	0.9x1.3
GFDL-CM4	1x1
HadGEM3-GC31-LL	1.3x1.9
HadGEM3-GC31-MM	0.6x0.8
INM-CM4-8	1.5x2
MIROC6	1.4x1.4
MIROC-ES2L	2.8x2.8
MPI-ESM1-2-HR	0.9x0.9
MPI-ESM1-2-LR	1.5x1.5
MRI-ESM2	1.1x1.1
NorESM2-LM	1.9x2.5
NorESM2-MM	0.9x1.3
TaiESM1	0.9x1.3
UKESM1-0-LL	1.3x1.9

Table 1: List of CMIP6 models and native resolutions.

2.3.4 Method Selection

The task for the ML model is to predict the daily-mean temperature anomaly at the target grid cell from a set of predictor variables that provide information about the local state of the climate system. Preliminary testing was carried out to determine which of the three statistical learning methods described earlier in this chapter was most suitable for this application. As well as good predictive skill, key features required of the selected method were extrapolation and interpretability. Interpretability is key to ensuring that the final model makes good predictions for the right reasons and that the relationships identified between predictors and the target temperature are consistent with our understanding of the climate system. The ability of the method to extrapolate is required for the application of constraining future climate change. Naturally, as temperatures increase under high future emissions scenarios, temperatures will eventually begin to consistently exceed those which the model was exposed to during training on historical data. Additionally, the ability to extrapolate is also likely to aid in the prediction of more extreme events in the historical and present day climate.

Predictive Skill

In order to investigate the ability of Ridge, RF and GPR to meet these required criteria, preliminary testing was carried out at a set of 50 randomly selected $3° \times 3°$ land grid cells (with land area fraction greater than 0.7), see Figure 2.9. As an initial test, models were

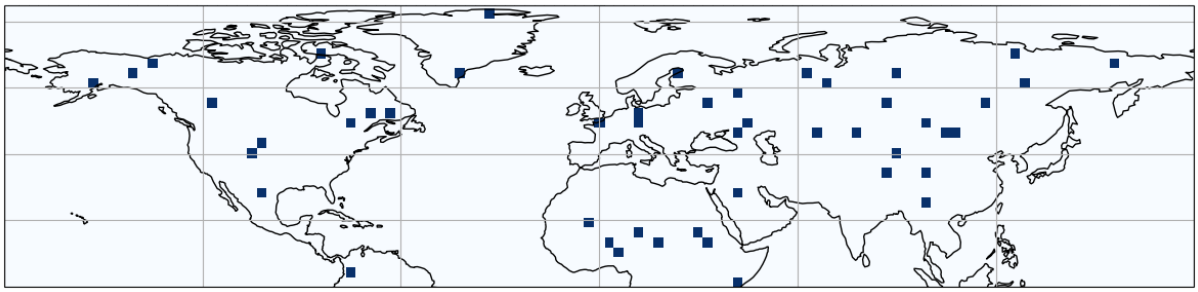

Figure 2.9: Location of 50 randomly selected test grid cells on 3×3 longitude-latitude grid with land area fraction > 0.7.

trained to predict the daily 2m temperature anomaly relative to a smoothed mean seasonal cycle. A set of initial variables to use as local (same grid cell as target variable) predictors for this testing were identified based on a correlation analysis with the target variable, 2m temperature anomaly, across the 50 test grid cells. Figure 2.10 shows the mean and 10^{th} to 90^{th} percentile spread of correlation between 11 climatic variables and temperature. The five variables with the greatest degree of correlation were selected: near-surface relative

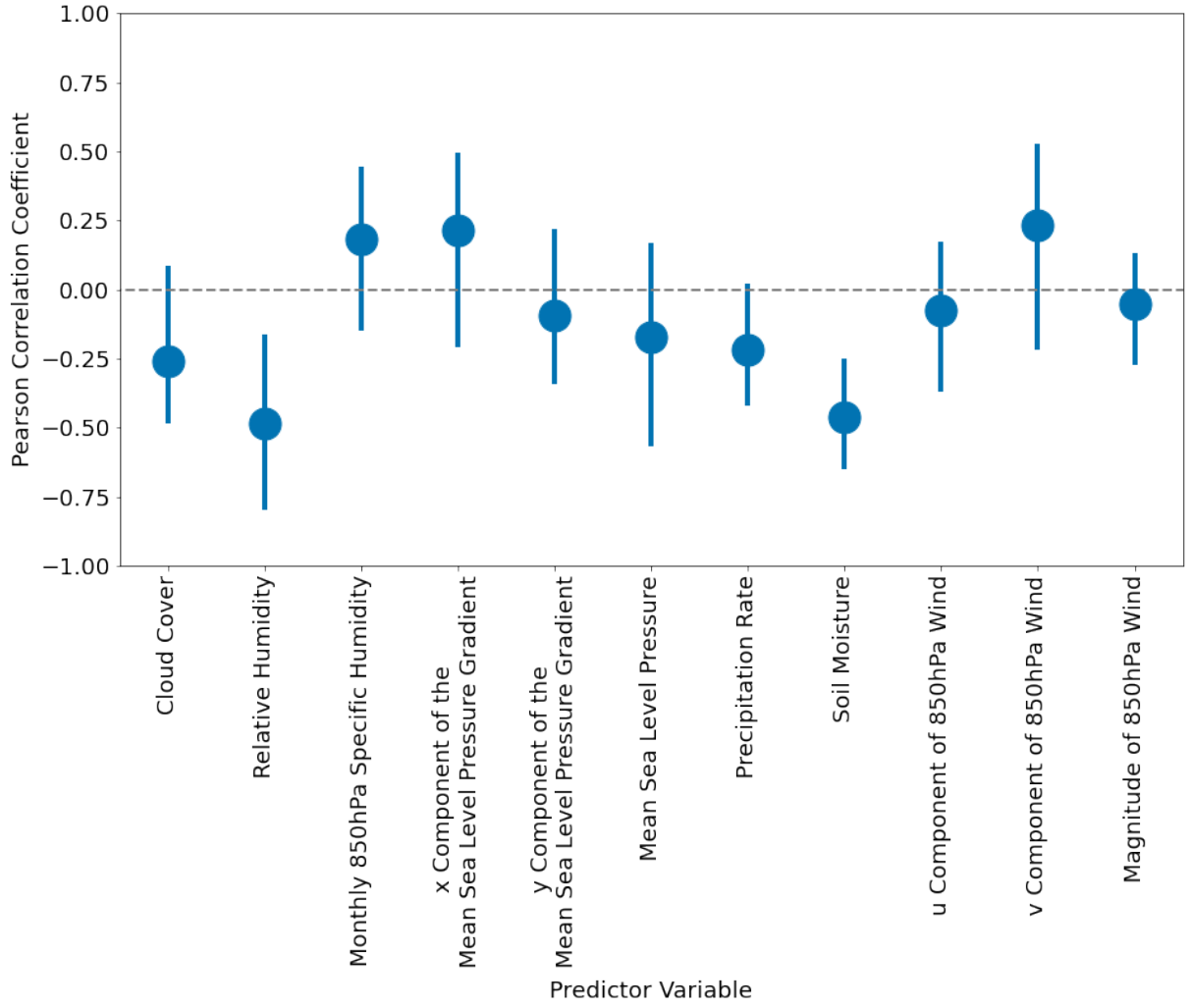

Figure 2.10: Mean (circle), 10^{th} and 90^{th} percentile (line) of Pearson correlation coefficients between predictor variables and temperature across 50 randomly selected test locations.

humidity; soil moisture; monthly specific humidity at 850hPa; the x-component of the mean sea level pressure gradient; and the v-component of the 850hPa wind. Ridge, RF and GPR models were trained on ERA5 data to predict the daily temperature anomaly using these five variables (also measured as anomalies to the seasonal cycle). Four years, 2000-2003, were excluded to use as unseen test data, leaving 40 years of data between 1979-1999 and 2004-2022 inclusive for training and cross-validation. For an overview of performance between the three methods, R^2 scores and mean absolute errors were calculated using the test data (2000-2003), as shown in Figure 2.11. Based on these metrics there is little to separate the

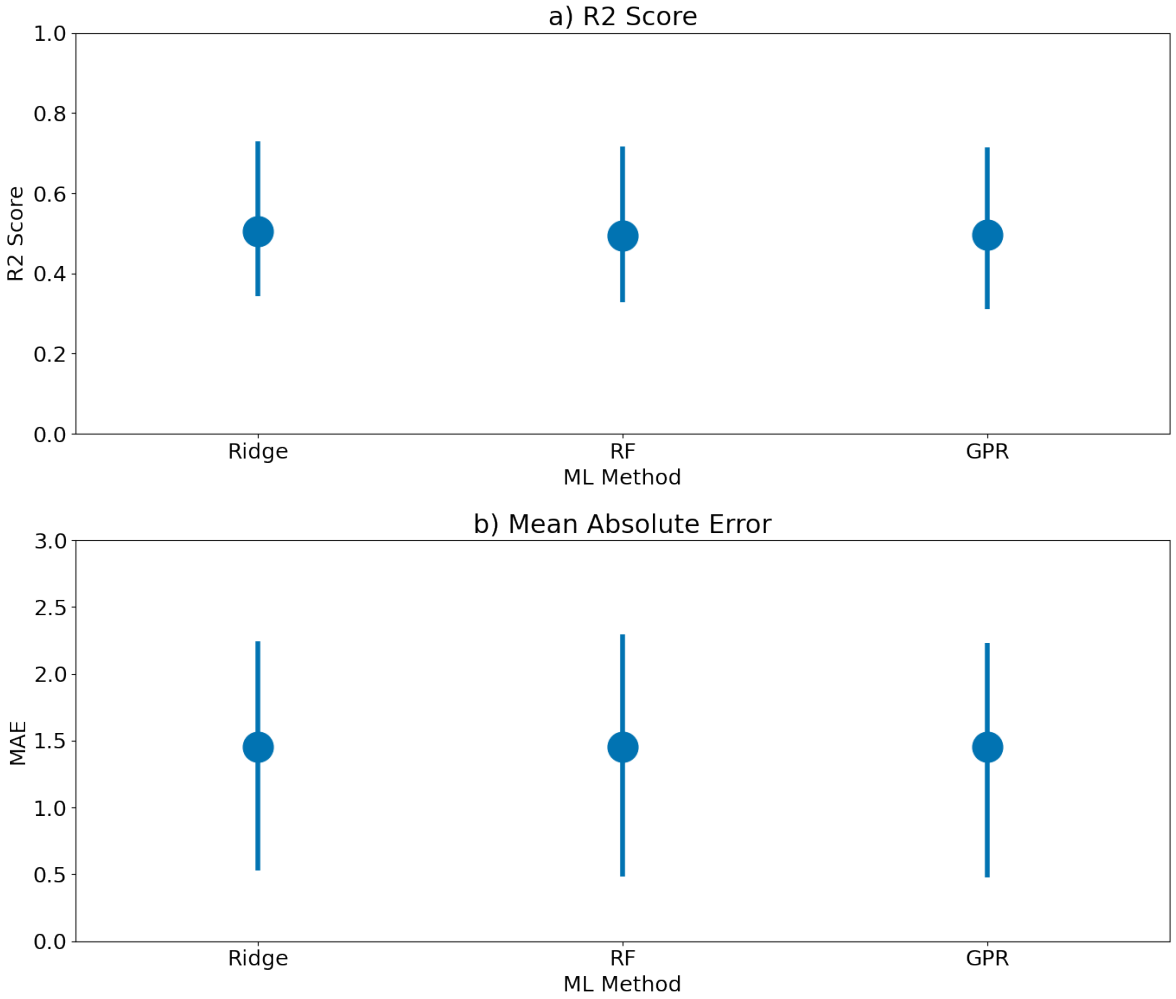

Figure 2.11: Mean (circle), 10^{th} and 90^{th} percentile (line) of R^2 scores (a) and mean absolute error (b) for Ridge, RF and GPR predictions of ERA5 temperature anomalies using relative humidity, soil moisture, monthly 850hPa specific humidity, x-component of the mean sea level pressure gradient and v-component of the 850hPa wind as predictors.

three models with this variable setup; this is also clear when time series of predictions at individual locations are inspected. Figure 2.12 shows prediction time series from each ML

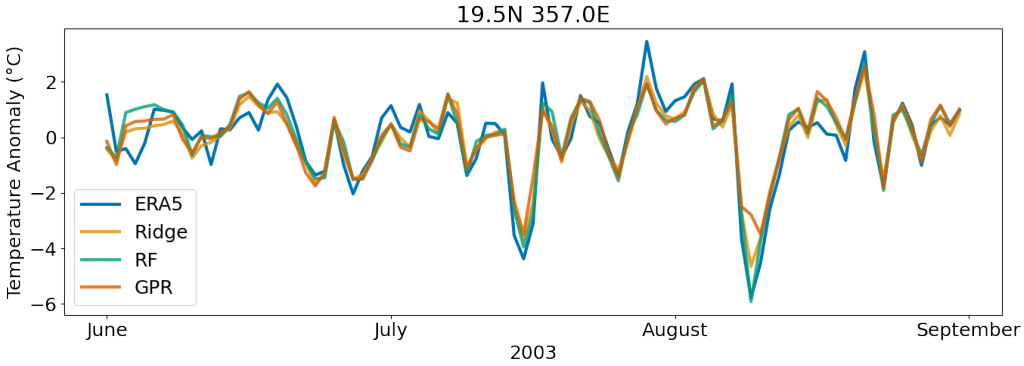

Figure 2.12: Comparison of Ridge, RF and GPR temperature predictions for ERA5 temperature anomalies during summer 2003 at 19.5N 357E.

model at a grid box over West Africa where all three models show generally good agreement with ERA5 and make similar predictions to each other. Next, consideration was given to the interpretability of the three ML methods and their ability to extrapolate.

Interpretability

Interpretability will be important not only for considering the drivers of particular extreme events but also to increase confidence that the model makes good predictions for the right reasons, and that predictive skill is derived from physically consistent relationships. There are two aspects to interpretability, one being the overall interpretability of the model - which variables does the model derive the most predictive power from? The second being the interpretation of individual predictions - in this case temperature anomalies on specific days. One method to enable this specific interpretation is SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) which quantify the contribution of each predictor variable to the overall prediction for a given temperature anomaly (see Chapter 2.4.2 for more detail). In the case of Ridge regression, this is much simplified by the linear relationships between each of the predictor variables, x_j, and the target, y, being quantified by a single learnt coefficient, $\theta_{era5,j}$.

As variables are normalised prior to regression, coefficient magnitude can also indicate which variables dominate predictions of the model in general. For decision trees, plotting out the branching points can be a powerful tool for interpretability with small data sets, however for a random forest which may contain hundreds of trees this is no longer a helpful approach. Instead, there are metrics which measure feature importance across the forest taking into

account factors such as node impurity (Scornet, 2021), a measure of how well a node splits the tree, and permutation importance which measures the decrease in performance on test data when the value of a particular feature is randomly assigned (Gregorutti et al., 2017). There are issues with impurity-based metrics including the fact that impurity-based statistics are calculated on training data so give no information about generalisation to test data, and impurity metrics show a heavy bias towards high-cardinality features (Shih and Tsai, 2004). These same metrics can also be used to rank the importance of individual predictor variables for specific predictions. As ensemble methods, RFs, like GPR models, also provide intrinsic measures of uncertainty from spread across the ensemble. GPR itself does not have any 'built-in' metrics to measure contributions of individual predictor variables to overall predictive skill or specific predictions. External analysis methods such as SHAP value analysis (Lundberg and Lee, 2017) or local linear approximations (Yoshikawa and Iwata, 2021) can instead be applied to infer contributions of individual predictor variables.

The initial testing above was carried out with local predictor variables only. To take advantage of spatial information about each input variable, predictor variables to the final model will be provided in square $n \times n$ grid cell domains centred on the target grid cell for prediction. This results in n^2 inputs to the model for each predictor variable, see Figure 2.13. Naturally, values of the same climatic variable at adjacent grid cells are likely to be highly

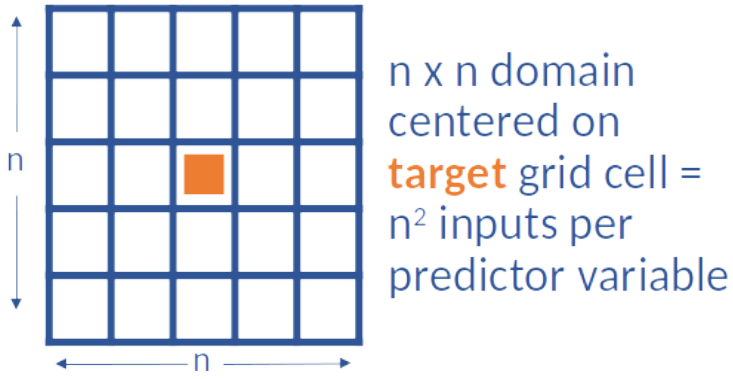

Figure 2.13: Structure of predictor variable domains.

correlated with each other and so it is important that the chosen ML algorithm is able to cope with this. To test this ability, Ridge, RF and GPR models were trained with the same set of predictor variables used previously but with increasing domain sizes. Again models were trained on ERA5 data from 1979-2022 excluding 2000-2003 to use as independent testing data. Figure 2.14 shows how test performance varies for each of the three ML methods as the domain size of input variables is increased according to two metrics; R^2 score and mean absolute error. Whilst Ridge and RF models gain skill with the increased spatial

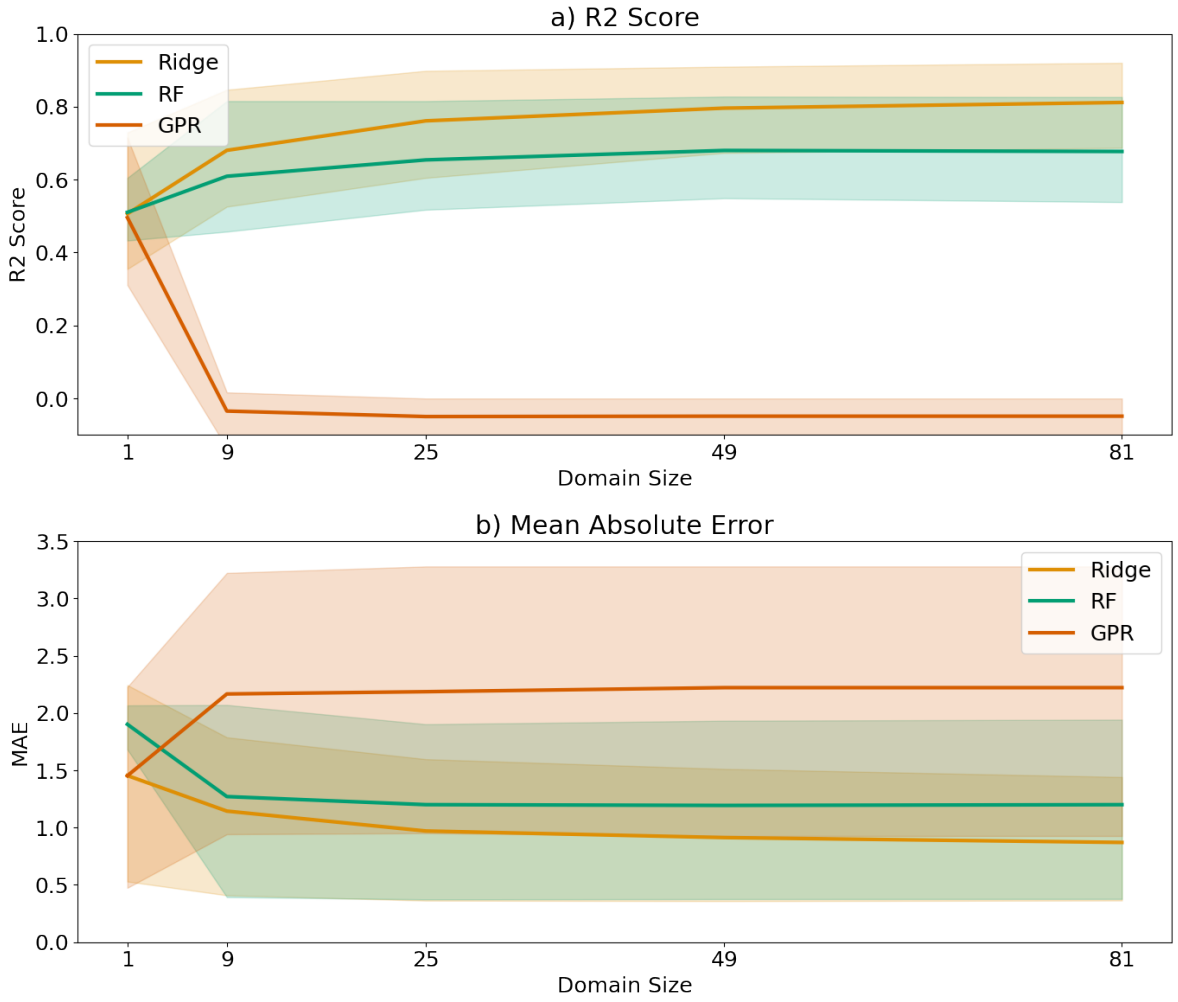

Figure 2.14: Comparison of mean (lines) and 10^{th}-90^{th} percentile range (shading) for Ridge (orange), RF (green) and GPR (red) test performance when trained with different domain sizes for predictor variables using R^2 scores (a) and mean absolute error (b).

information, with Ridge exhibiting the best performance scores, GPR skill decreases as a result of overfitting to co-linear input variables.

Extrapolation

To test the ability of the ML methods to extrapolate, models were trained on data from a CMIP6 model, UKESM1-0-LL, covering the reanalysis period 1979-2022. Data from 1979-2014 were taken from the *historical* forcing scenario and data for 2015-2022 were taken from future emissions scenario *SSP585*. Ridge, RF and GPR models were trained at each of the 50 test locations using the same procedure as with the ERA5 data. These trained models were then tested with inputs from the *SSP585* future emissions scenario covering 2023-2100. Figure 2.15 plots how the mean absolute error in predictions from each ML method evolves

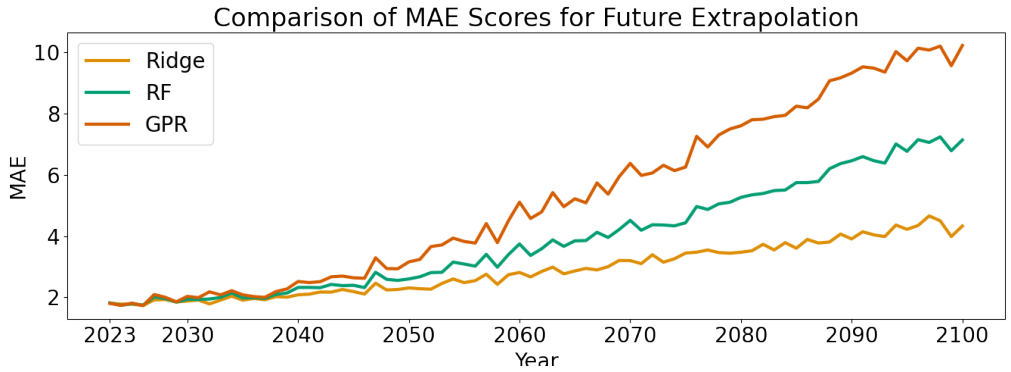

Figure 2.15: Comparison of future temperature prediction performance for Ridge, RF and GPR trained on historical UKESM1-0-LL data using mean absolute error scores calculated every year and averaged over 50 test locations.

over time with increasing warming over the course of the 21st century under future warming scenario *SSP585*. While there is a degree of increasing error across all three models, GPR and RF clearly struggle much more than Ridge to extrapolate predictions to temperatures projected from around 2040 on wards in a high emissions future. For GPR, this is likely attributable to the overfitting seen in Figure 2.14

Model Selection

Based on the analysis presented above, Ridge proved much better able to cope with co-linearity amongst predictor variables in a high-dimensional predictor space. This is important in an interconnected system, especially if adjacent grid cell measurements of the same variables are to be included as predictors in order to provide spatial information to the ML model. While RF showed a smaller improvement in performance with additional information

from adjacent grid cells, GPR suffered from significant problems with over-fitting under these conditions. As a linear-based model, Ridge is naturally much simpler to interpret than either RFs or GPR although methods like SHAP value analysis do counteract this problem and, as ensemble learning methods, RF and GPR have the advantage of providing an intrinsic quantification of uncertainty in their final predictions. Ultimately, the performance in the extrapolation test excludes RF and GPR from being selected; the decay in performance as temperatures continue to exceed the training range is unacceptable for constraining uncertainty in future warming projections (see Chapter 4). While the overfitting problems with GPR could potentially be resolved by further fine tuning, other advantages of Ridge such as computational cheapness, ease of interpretation, simplicity of training, and relative ability to cope with co-linear input variables leads to Ridge being selected as the method to take forward.

2.3.5 Variable Selection

Having selected Ridge regression as the most promising ML method for the intended applications, a systematic testing of predictor variable combinations and domain sizes was carried out. A range of variables were identified for initial testing via a literature review of mechanisms contributing to recent heatwave events; from variables affecting radiative heat transfer to dynamical terms and soil moisture feedbacks. The variables selected for the final model should provide information both about the background thermodynamic state as well as dynamical fluctuations.

Variable Definitions and Interpretations

Predictor variables are selected with the physical processes they represent in mind. With the aim of trying to constrain existing climate model output based on the relationships learnt from observations, the model is not intended to simply recalculate the energy budget. The combination of variables selected should not so definitively define temperature as to make model skill trivial. The variables identified for testing are listed below:

- soil moisture content in the top 7cm
- near-surface relative humidity
- running monthly mean of near-surface relative humidity
- fractional total cloud cover
- mean total precipitation rate

- mean sea level pressure
- x-component of the mean sea level pressure gradient
- y-component of the mean sea level pressure gradient
- magnitude of the 850hPa wind
- u-component of the 850hPa wind
- v-component of the 850hPa wind
- running monthly mean of 850hPa specific humidity.

For the variables defined at 850hPa, a common mask is applied (to climate model and reanalysis data) to remove grid cells which would be obscured by topography in reality. This mask is plotted in Figure 2.16 and is applied to both CMIP6 and ERA5 data for all 850hPa variables.

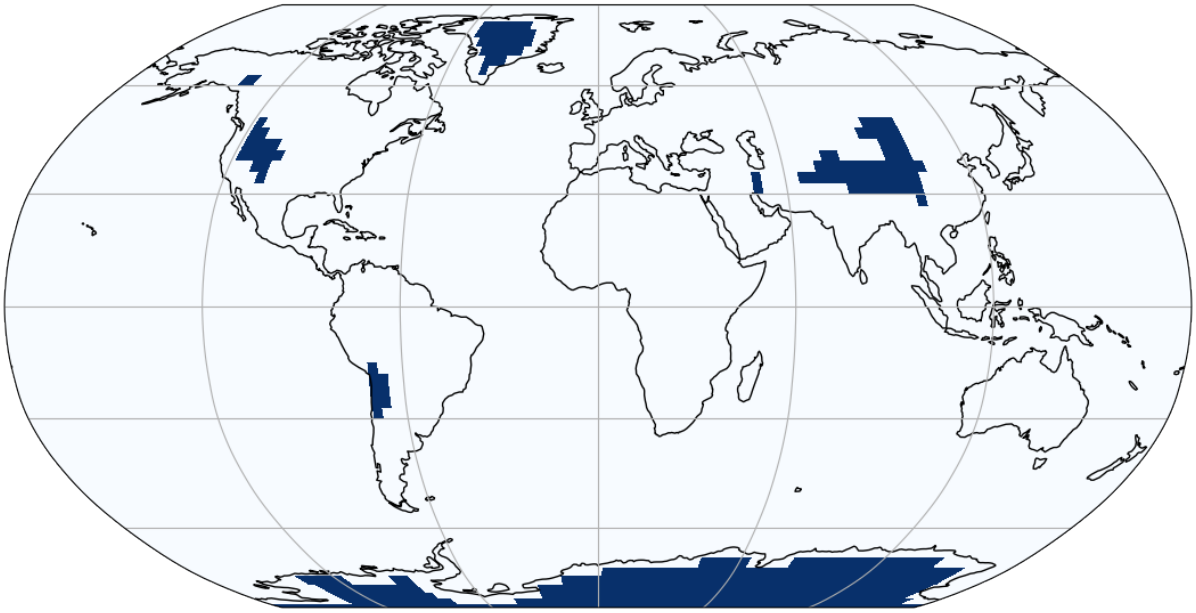

Figure 2.16: Grid cells excluded by 850hPa mask (blue) to avoid large interference of topography and associated features such as ice and snow cover.

To give an overview of the relationship between the potential predictor variables and temperature, maps of Pearson correlation coefficients (see Section 2.1.5) are plotted for each combination of predictor variable with temperature for JJA across the reanalysis period (1979-2022), see Figure 2.17. The remainder of this subsection explains the processes that

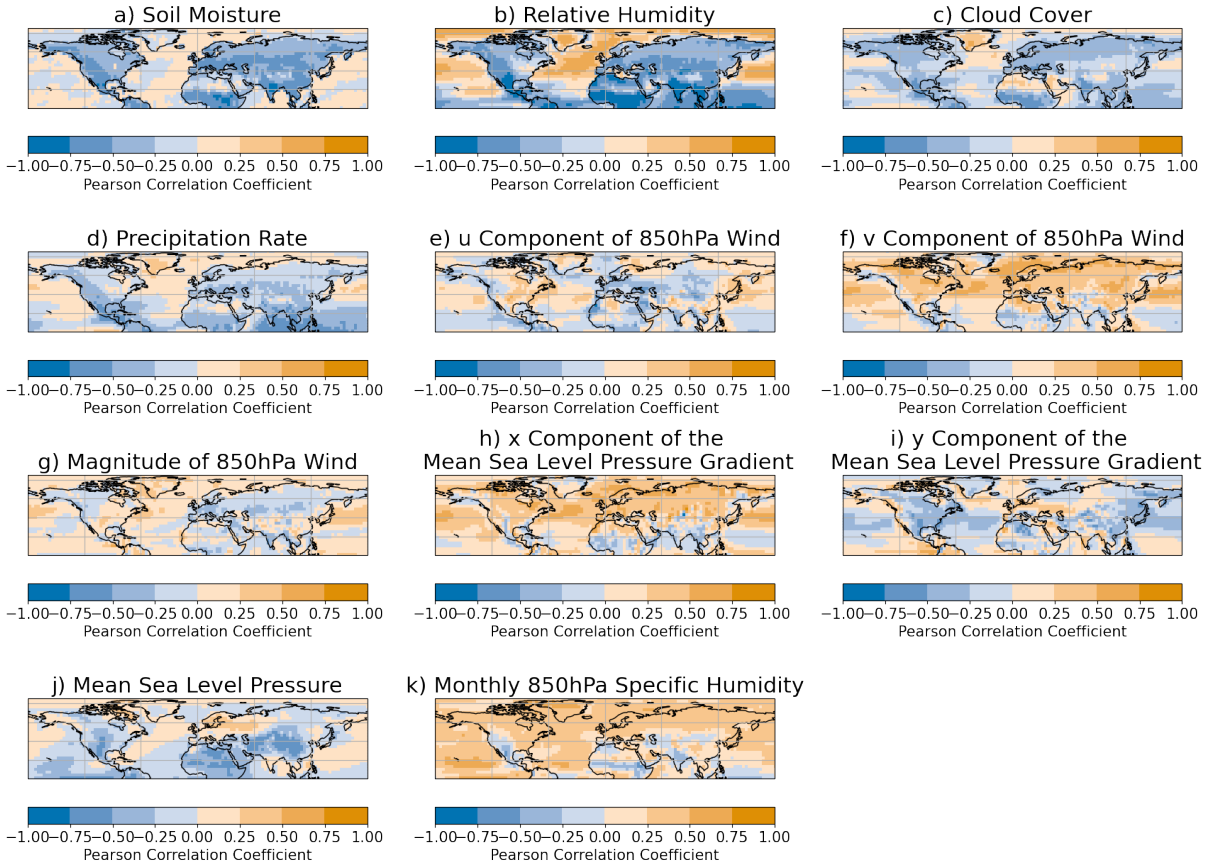

Figure 2.17: Heatmap of Pearson correlation coefficients with local temperature anomalies for all variable combinations.

each variable is intended to represent in the process-based ML model.

Daily near-surface relative humidity is closely linked to temperature - the capacity of the air to hold water is related to temperature via the Clausius-Clapeyron relationship. Other factors such as the atmospheric temperature structure and the type of forcing also affect this relationship (*e.g.* Hodnebrog et al., 2019). Near-surface humidity may also give an indication of the degree of evaporation occurring from the surface (Mahfouf, 1991).

To capture dynamical anomalies such as blocking, mean sea level pressure and the magnitude of the 850hPa wind are included as potential predictor variables. Mid-latitude continental heatwaves frequently occur under blocking conditions (*e.g.* Pfahl et al., 2012; Schaller et al., 2018; Domeisen et al., 2023), anomalous, anti-cyclones which halt or even reverse the prevailing wind conditions diverting other systems and allowing temperature anomalies to build up (Woollings et al., 2018). For directional information relating to heat transport, the x and y components of the mean sea level pressure gradient as well as the u and v components of the 850hPa wind are included.

Heatwave intensity and duration is often sensitive to pre-conditions (*e.g.* Träger-Chatterjee et al., 2013; Seo et al., 2020), as such monthly means of near-surface relative humidity and 850hPa specific humidity are included to provide the model with longer term information about the state of the system. These variables are also intended to aid in detection of the global warming signal which will be helpful for constraining future temperature projections (see Chapter 4). Several mid-latitude heatwaves in the last decade have been associated with exacerbating soil moisture deficits which reduce the capacity for evaporative cooling at the surface, for example Europe 2003 (Black et al., 2004; Ferranti and Viterbo, 2006; Fink et al., 2004; Fischer et al., 2007; Zaitchik et al., 2006) and Russia 2010 (Hauser et al., 2016; Schubert et al., 2014). Daily soil moisture volume present in the top layer of the soil is included as an indicator of the level of drought and capacity for cooling. Total cloud cover is included to give an indication of the degree of direct surface radiative heating. In Europe 2003, for example, high temperatures were found to be intensified by clear sky conditions (Garcia-Herrera et al., 2010). Finally, daily-mean precipitation rate gives an indication of moisture availability for evaporative cooling.

Domain Sizes

To identify optimum domain sizes for each variable, Ridge regression models were trained over a range of domain sizes for each predictor variable at a set of 50 randomly selected land grid cell test locations (as in the previous subsection). The test R^2 scores for these models shown in Figure 2.18 indicate that performance is not considerably improved by increasing domain sizes beyond 5×5 or 7×7 grid cells. Additionally, the same test set up as shown in Figure 2.18 was also completed combining all variables in one Ridge regression model for each

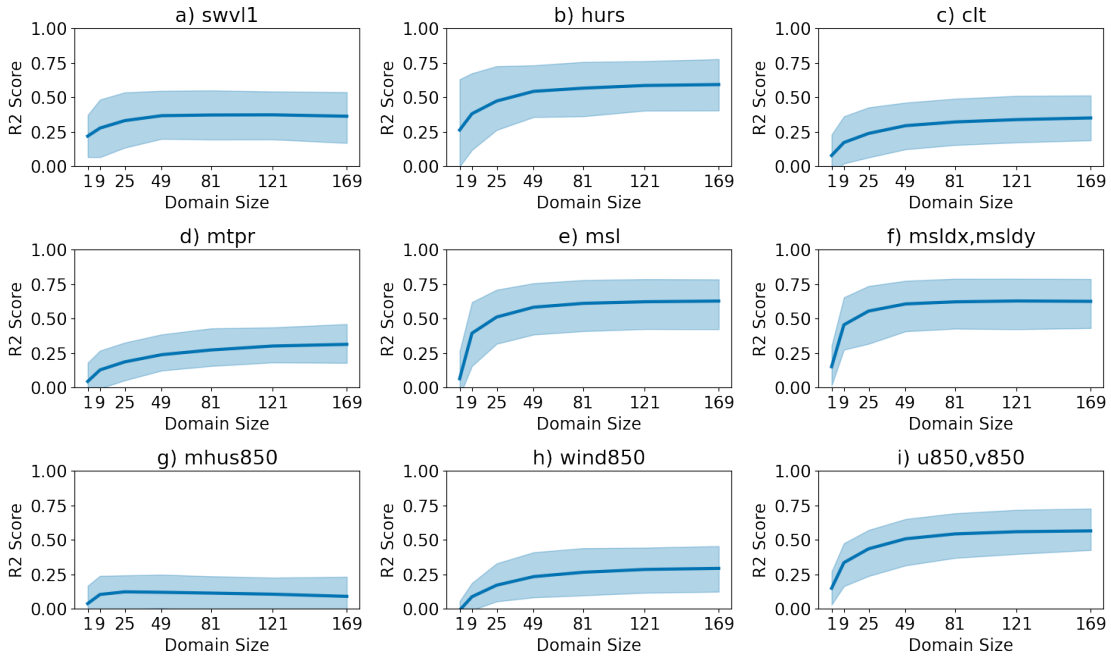

Figure 2.18: Domain size of input variable *vs.* mean test R^2 score (blue line) with 10% to 90% percentile range (blue shading) averaged over 50 randomly selected Northern Hemisphere land grid cells for Ridge regressions with a) volumetric soil moisture content in the top 0-7cm, b) near-surface relative humidity, c) total cloud cover fraction, d) mean total precipitation rate, e) mean sea level pressure, f) x and y components of the mean sea level pressure gradient, g) monthly 850hPa specific humidity, h) magnitude of 850hPa wind or i) u and v components of the 850hPa wind as predictor variables.

test domain size, shown in Figure 2.19. The effect of increasing domain size on performance is consistent with the individual variable Ridge models. With more variables now included in the regression, beyond domain sizes of 5×5 there is very little gain in performance from the increasing volume of coefficients. Based on these results, for most variables, a domain size of

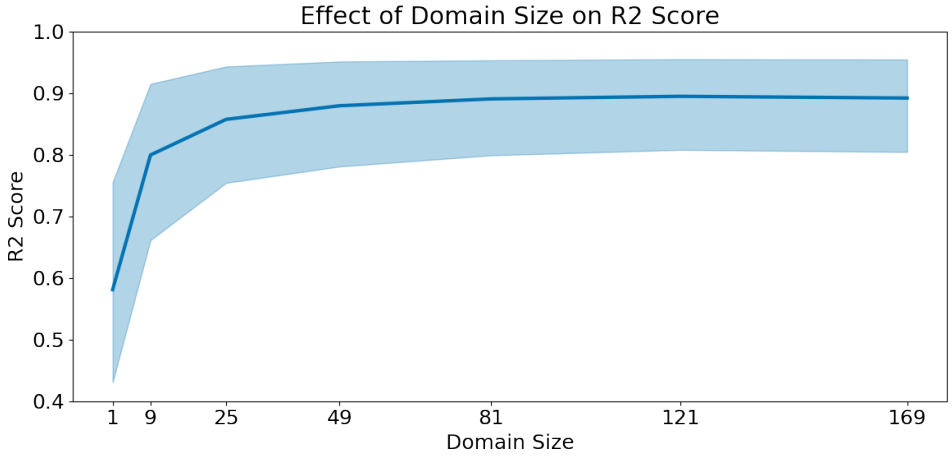

Figure 2.19: As in Figure 2.18 for Ridge regression models with near-surface relative humidity, volumetric soil moisture content in the top 0-7cm, mean sea level pressure, total cloud cover fraction, mean total precipitation rate, mean sea level pressure, x & y components of the mean sea level pressure gradient, magnitude of the 850hPa wind, u & v components of the 850hPa wind and monthly 850hPa specific humidity as predictor variables.

5×5 was selected. With a balance between model performance and physical consistency in mind, soil moisture domains were further restricted. Including larger domains of soil moisture improves local predictions of daily temperature anomalies but decreases interpretability from physical mechanisms. Additionally the greatest gain in performance occurs at smaller domain sizes for soil moisture so for this variable, inputs were taken only from the target grid cell.

Overfitting

As well as considering which variables added skill to the model, the intended applications of the model and the physical consistency of the relationships that Ridge was learning between the predictor variables and target temperature anomaly also have to be considered. Inclusion of dynamical variables (those derived from the mean sea level pressure and 850hPa wind) showed a tendency towards overfitting in some locations (Sippel et al., 2019). The optimum value of the regularisation parameter determined during cross-validation was usually the smallest value available. This diminishes the influence of the additional regularisation term

in the Ridge cost function and results in models which tend towards a standard multiple linear regression. Subsequently, Ridge models trained with these variable combinations had noisier distributions of coefficients associated with them. This is demonstrated at 45N 0E where a Ridge model was trained with 5×5 domains of near-surface relative humidity, cloud cover fraction, precipitation rate, monthly 850hPa specific humidity, mean sea level pressure and x & y components of the mean sea level pressure gradient along with local soil moisture. The distributions of these coefficients learnt by Ridge from ERA5 (θ_{era5}) as well as those learnt from the CMIP6 models listed in Table 1 (θ_{cmip6}) are plotted in Figure 2.20. By

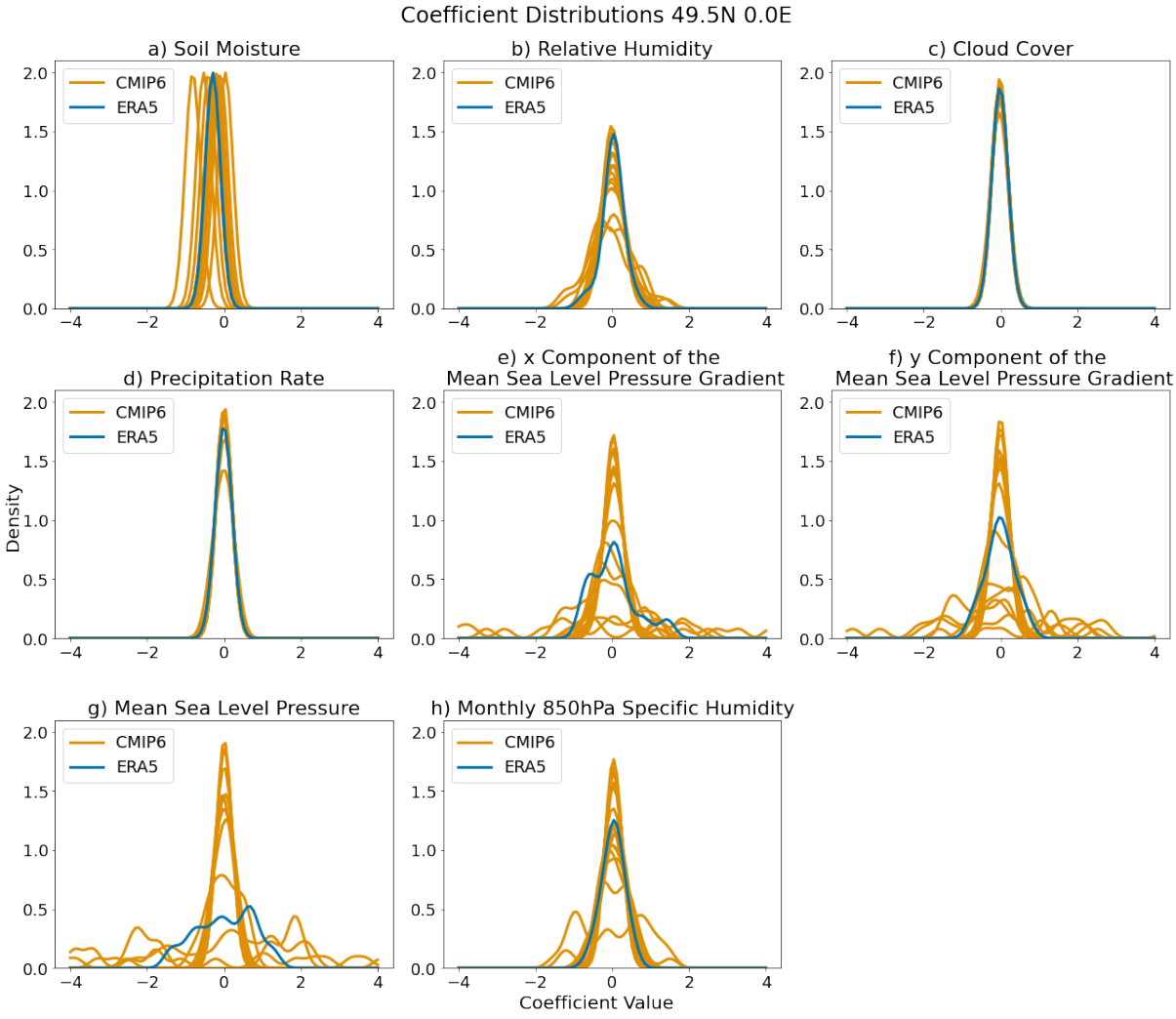

Figure 2.20: Distribution of Ridge coefficients (θ) learnt from ERA5 (blue) and CMIP6 (orange) data at 45N 0E for a) soil moisture, b) relative humidity, c) cloud cover, d) precipitation rate, e) x component of the mean sea level pressure gradient, f) y component of the mean sea level pressure gradient, g) mean sea level pressure, h) monthly 850hPa specific humidity.

contrast, when the mean sea level pressure variables are replaced with components of the 850hPa wind, the coefficient distribution learnt from ERA5 looks smoother, with some noise still apparent in the coefficients learnt from historical CMIP6 data, shown in Figure 2.21.

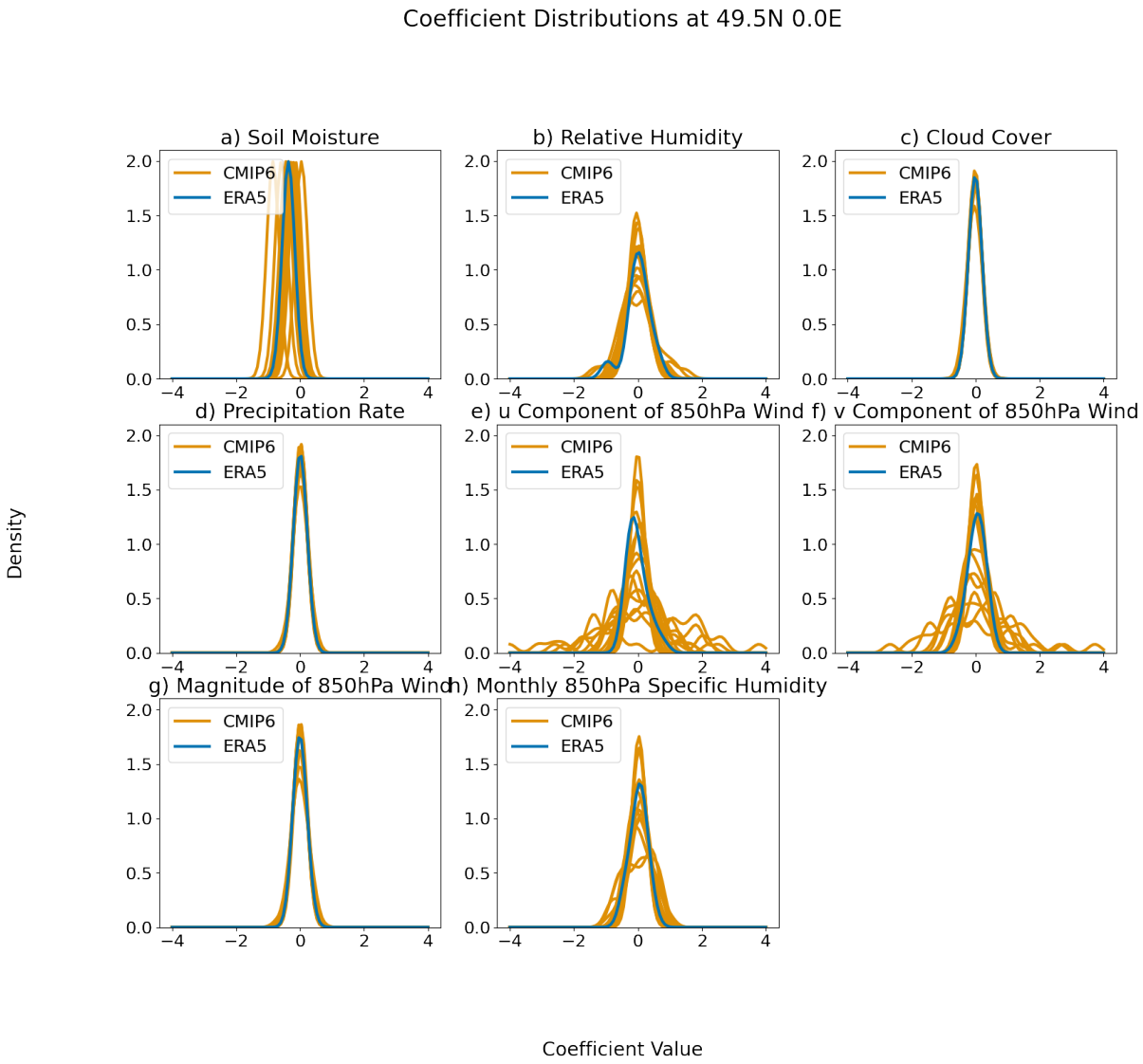

Figure 2.21: As in Figure 2.20 but with predictor variables of a) soil moisture, b) relative humidity, c) cloud cover, d) precipitation rate, e) 850hPa zonal wind, f) 850hPa meridional wind, g) magnitude of 850hPa wind, h) monthly 850hPa specific humidity.

This noise in the coefficient distributions resulting from overfitting is important to consider, not only in relation to making predictions for unseen ERA5 data using our observations-based Ridge-ERA5 model, but also for applying Ridge-ERA5 to climate model data in order to impose observations-based constraints. Both the sea level pressure and wind variables

provide information about the surrounding dynamical state. Inferring information about wind from sea level pressure does require the assumption of geostrophy but this assumption is closest to being true during the season of interest, summer, when winds are generally at their weakest. Figure 2.22 plots the optimum value of the regularisation parameter determined over a 5-fold cross validation during training to determine where (and how frequently) overfitting is likely to be occurring for each variable setup across the Northern Hemisphere. For Ridge regressions trained on ERA5 data, it seems that the tendency to overfit at land

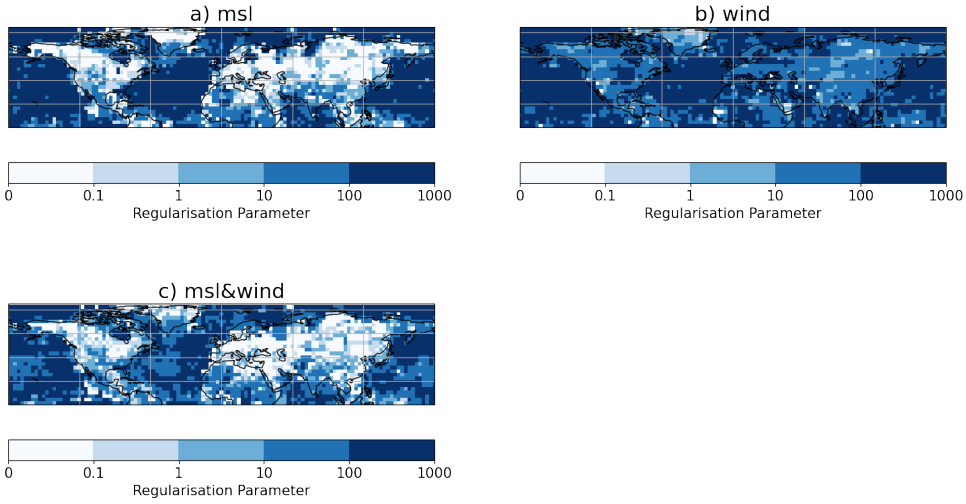

Figure 2.22: Optimum value of regularisation parameter determined via 5-fold cross-validations for variable setups including mean sea level pressure variables (a), variables derived from the 850hPa wind (b) and both wind and pressure variables (c).

grid cells is introduced by the inclusion of mean sea level pressure variables as predictors. The wind variable setup shows a significantly lower frequency of overfitting.

Final Variable Setup

The result of the testing above is a final variable combination of soil moisture, near-surface relative humidity, fractional cloud cover, mean total precipitation rate, meridional and zonal 850hPa wind, magnitude of 850hPa wind and monthly running means of both near-surface relative humidity and 850hPa specific humidity. This setup with domain sizes is summarised in Table 2 and will be referred to as Ridge-ERA5.

Variable Name	**Domain Size**
Soil Moisture	1x1
Near-Surface Relative Humidity	5x5
Monthly Near-Surface Relative Humidity	5x5
Cloud Cover Fraction	5x5
Precipitation Rate	5x5
u Component of the 850hPa Wind	5x5
v Component of the 850hPa Wind	5x5
Magnitude of the 850hPa Wind	5x5
Monthly 850hPa Specific Humidity	5x5

Table 2: Final predictor variable setup for Ridge-ERA5.

2.4 Analytical Techniques

The following subsections outline analytical techniques that will be applied in the later results chapters.

2.4.1 KDEs

Kernel Density Estimators (KDEs) are used to represent temperature distributions for comparison between reanalysis data, model data and ML predictions. For a set of N data points $X_1, X_2, ..., X_N$, a KDE estimates the true probability density curve according to the following formula:

$$KDE(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{X_i - x}{h}\right), \tag{27}$$

where h is the bandwidth with $h > 0$, and K is the kernel function such that $K(y) \geq 0$ and $\int_0^\infty K(y)dy = 1$ (Rosenblatt, 1956). The meaning of this formula is demonstrated using a simple rectangular kernel function which is given by a uniform distribution over the interval $(-\frac{1}{2}, \frac{1}{2})$ (Węglarczyk, 2018), see Figure 2.23. With a rectangular kernel, $K\left(\frac{X_i - x}{h}\right)$ is non-

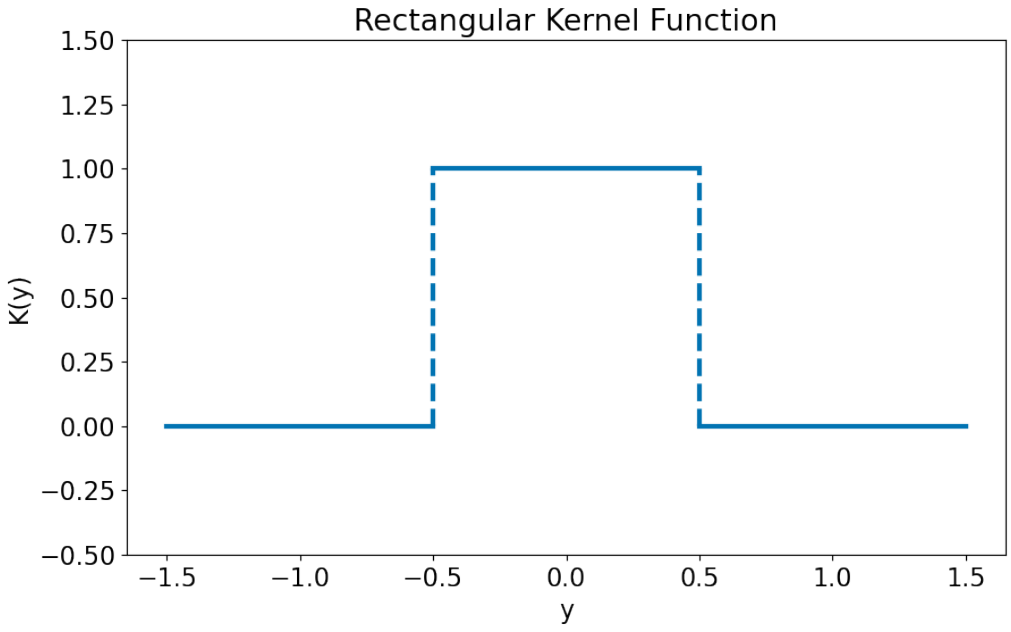

Figure 2.23: Rectangular kernel function given by a uniform distribution over the interval $(-\frac{1}{2}, \frac{1}{2})$.

zero for a data point X_i if and only if, X_i falls within $\frac{h}{2}$ of x. As the value of $K(y)$ can only be one or zero in this case, $\sum_{i=1}^{N} K\left(\frac{X_i - x}{h}\right)$ is simply equal to the number of observations

within $\frac{h}{2}$ of x. So, the value of the KDE at a specific point, x, is constructed based on the average number of neighbouring observations which fall within a 'bin width' of x. The overall estimate, $KDE(x)$, is then equivalent to dividing the average number of data points in the bin by the bin width, h. This is directly comparable to the construction of a traditional histogram where frequency density is calculated according to the number of observations in the bin divided by the bin width. The difference here for the KDE is that the bins are not discrete; bin locations are dependent on the values of x and may overlap.

The rectangular kernel has discrete jumps at $\pm\frac{1}{2}$ which results in a non-continuous KDE. Often the rectangular kernel is replaced by a smooth kernel function which results in a continuous KDE since sums of continuous functions are always continuous themselves. An example of a smooth kernel function is a Gaussian kernel (Węglarczyk, 2018) given by:

$$K_{Gaussian}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \tag{28}$$

which is defined such that the weight given to each observed data point, X_n, decreases smoothly moving away from x rather than simply dropping to zero outside of the bandwidth interval, see Figure 2.24. In general, for symmetrical kernel functions, the choice of kernel

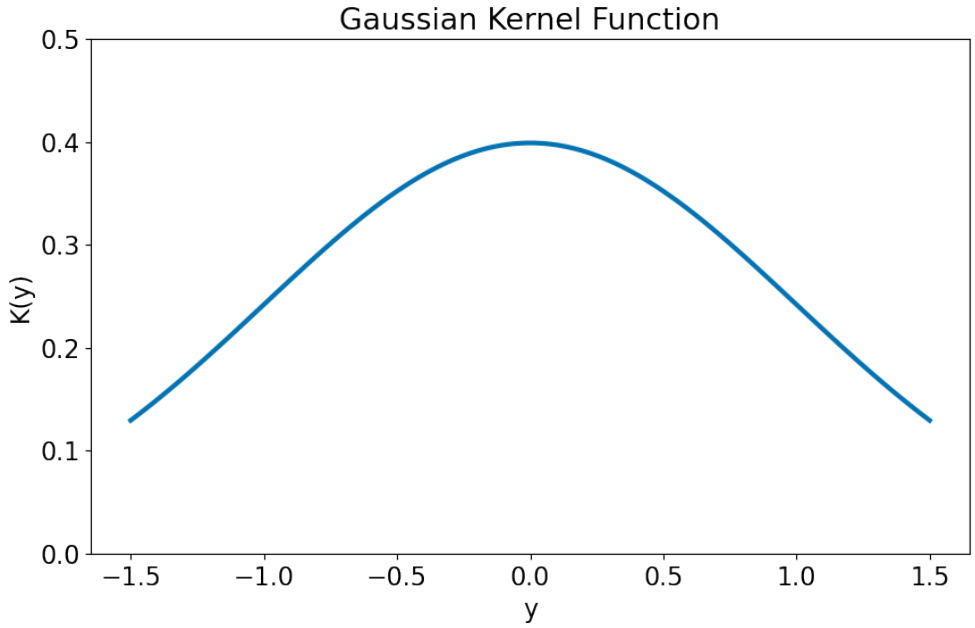

Figure 2.24: Gaussian kernel function evaluated over the interval $(-1.5, 1.5)$.

shape has relatively little impact on the final estimator, it is the choice of bandwidth which is much more critical (Marron, 1988).

Choice of bandwidth affects the smoothness of the fitted function and is associated with a bias-variance trade off - similarly to training an ML algorithm. A bandwidth that is too small results in under-smoothing meaning that variance is large whereas a bandwidth that is too large results in over-smoothing and large bias (Chen, 2017). For small values of bandwidth this occurs because, over a smaller interval, fewer observed data points, X_n, are averaged over and there is less overlap in bin width for adjacent points, x. As a result, the final KDE is over-sensitive to variability in the observed data points - analogous to overfitting. For large choices of bandwidth, observations which are 'too far' from x are included and features of interest in the pattern of observed points are smoothed away - a similar effect to underfitting an ML model. There are many selection techniques by which an optimal value of bandwidth can be chosen (*e.g.* Marron, 1988; Jones et al., 1992) depending on whether or not the *true* underlying curve is available. A commonly used rule of thumb with a Gaussian kernel and derived from the asymptotic mean integrated square error (Silverman, 1986) is calculated based on sample size and variance:

$$\hat{h} = 1.06\sigma N^{-\frac{1}{5}}. \tag{29}$$

2.4.2 SHapley Additive exPlanation (SHAP) Values

An important question in the application of ML models in any context, but especially in the field of climate science, is their interpretability. Is model skill derived from relationships that are compatible with observable physical mechanisms or do they result from more spurious statistical anomalies? In order to analyse the contributions of specific predictor variables to the predicted temperature anomaly on a given day, SHapley Additive exPlanation (SHAP) values (Lundberg and Lee, 2017) are calculated.

The idea of SHAP values is originally derived from Shapley values (Shapley, 1953) a concept in game theory used to quantify the contribution of each "player" to the "game". In the context of interpretable ML, the "game" is the ML model and the "players" are the input variables which contribute to the model predictions. Shapley values are calculated by looking at the outcome of the game for each possible combination of players or in this case the prediction of the ML model given each possible combination of input variables considered. This allows the marginal contribution to be calculated which refers to the effect of adding a given variable of interest to the model.

The schematic in Figure 2.25 shows an example of how this could work using a toy Ridge model trained with local surface sensible heat flux, near-surface specific humidity and mean sea level pressure at a grid cell over the UK. Each node represents a possible combination of input variables for a model to be trained with alongside the resulting prediction from that

combination of inputs. The contribution of a given variable, x_i to each model can be read from the difference in prediction between the model that does contain x_i as a predictor and the model that does not. There are multiple such values for each variable in the diagram; a weighted average is taken to calculate the variable's overall marginal contribution. Thus,

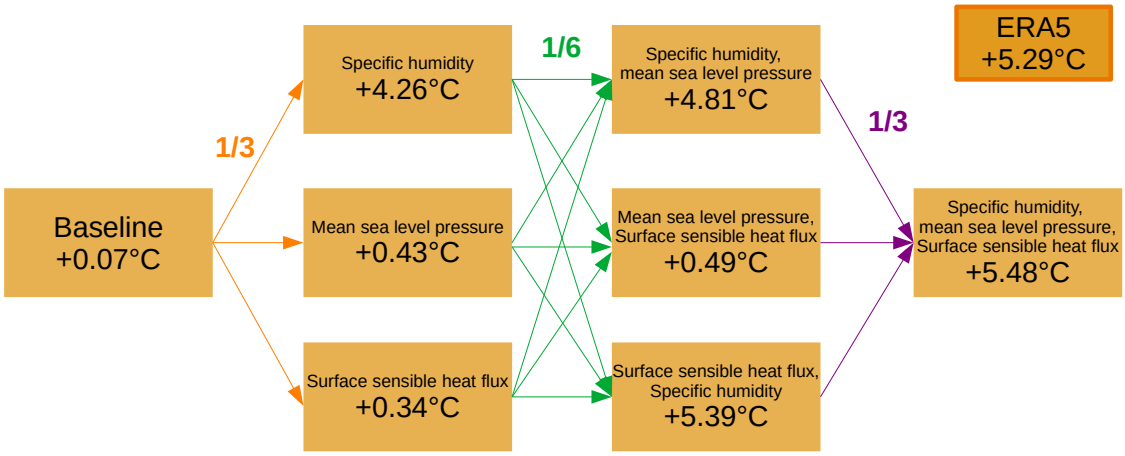

Figure 2.25: Schematic diagram of SHAP value calculations for a toy model with local surface sensible heat flux, near-surface specific humidity and mean sea level pressure as predictor variables. Baseline represents the fit of an intercept, the next column represents ML models fit using each predictor variable separately, the second column represents predictions by ML considering each possible combination of two variables, the rightmost column represents the model using all three predictors, with the actual value according to ERA5 in the top right.

SHAP values quantify the impact a certain value of a given input variable has on the final prediction compared to a baseline value. This allows individual predictions to be explained based on the contribution of each feature of the model. SHAP values are additive meaning that the sum of SHAP values for all variables adds up to the difference between the prediction of the model and the baseline value:

$$\hat{y} = \phi_0 + \sum_{j=1}^{P} \phi_j. \tag{30}$$

In Equation 30, the prediction of the ML model, y', is expressed as the sum of the baseline value, ϕ_0, and the contributions, ϕ_j, from the P predictor variables.

In the context of ML, it is often not computationally feasible to actually train models with all combinations of input variables; the number of models required to be trained scales as 2^P, where P is the number of predictor variables. Instead, as per the *shap* Python library (Lundberg and Lee, 2017), the final model with all possible features is trained and then the

marginal contributions are calculated by replacing the absent inputs with a random sample from the data. In the case of linear models like Ridge, the calculation is further simplified by the form of the model function. This is easily derived by comparison of the linear regression function (Equation 39) with the SHAP value definition above (Equation 30). The baseline SHAP value, ϕ_0, is equal to the intercept, β_0, and the SHAP value of a given variable, ϕ_j, is simply the product of the corresponding coefficient, β_j, and the observed value for that case, x_j:

$$\phi_j = \beta_j x_j. \tag{31}$$

3 Process-based Machine Learning for Bias Correction of Historical CMIP6 Temperature Distributions and Climate Model Evaluation

In this chapter, performance of the process-based ML model, Ridge-ERA5, is evaluated on held-out ERA5 test data - data not seen during training. This is carried out using a range of performance metrics, considering both spatial patterns of high and low performance by looking at maps of, for example, R^2 scores but also at individual time series at particular locations. Limitations of the method in terms of its prediction performance for events in the extreme tails of the temperature distribution as well as for events which exceed the range of values seen at a given location during the respective training period are also investigated. Two applications of Ridge-ERA5 for historical climate are demonstrated. Firstly, as a means of performing an observations-based 'bias correction' to existing climate model output by feeding input variables from the CMIP6 historical forcing scenario into Ridge-ERA5 to produce observationally moderated versions of CMIP6 simulations of near-surface temperature anomalies. Secondly, the method is applied in a model evaluation context. Coefficients learnt by a set of CMIP6 emulator Ridge-CMIP models, trained with the same variable setup as Ridge-ERA5, are compared with those learnt by Ridge-ERA5. This allows similarities between climate models, as well as closeness to observations-based relationships, to be analysed.

3.1 Introduction

Continued progress in the field of climate modelling has seen an increasingly complex picture of Earth's climate represented in global climate models. Whilst these models show a reasonable performance in representing key elements of the climate system and historical climate change, there are still important and well-known limitations in our understanding and modelling of global climate (*e.g.* Toreti and Naveau, 2015; Eyring et al., 2019; Davini and D'Andrea, 2020). At the same time, ML is increasingly being applied in climate science to address some of these challenges, and also to complement existing modelling tools (see Section 1.6). This introduction begins with consideration of various model evaluation strategies and existing methods for bias correction followed by a literature review of ML applications for bias correcting climate model output.

3.1.1 Climate Model Evaluation

For as long as weather and climate models have been in use it has been necessary to evaluate their performance, not only to measure the reliability of their output but also as part of the model development process. Whilst weather models aim to produce predictions of precisely what will be observed and time series can therefore be directly compared with observations, free-running climate model outputs represent only one possible realisation of the climate system and on much longer, decadal timescales. This prompts the question of how climate model performance can be evaluated against the observed climate and how performance can be compared between models given that natural variability is expected - models may differ from what is observed but still produce reasonable realisations of climate. There is also the possibility of compensating biases where simulation of a particular variable may show good agreement with observed quantities but only as a result of multiple errors cancelling each other out (Gleckler et al., 2008). To summarise, climate model evaluation seeks to ensure that the model simulates a realistic result for reasons that are physically consistent. This subsection outlines how model output can be compared with observations, from climatological averages to representations of different modes of variability as well as specifically evaluating climate model performance in simulating extreme events.

Historical scenario forcings provide climate models with information about anthropogenic and natural forcings, for example, historical atmospheric carbon dioxide concentrations (O'Neill et al., 2015) but, due to the natural variability of the climate system, no climate model is expected to reproduce the historical record exactly. The most widely used strategy to evaluate climate model output whilst accounting for this fact is to compare historical climate model output with observed climatological averages (Edwards, 2011). Climatological averages may include monthly mean statistics for variables of interest which can be compared via: root mean square errors (RMSE); correlation; and ratio of variances (*e.g.* Taylor, 2001; Gleckler et al., 2008; Pincus et al., 2008). Errors in simulation of these climatological mean values may also be combined into a single performance score designed to summarise model performance across a range of metrics (*e.g.* Murphy et al., 2004; Schmittner et al., 2005; Reichler and Kim, 2008; Waugh and Eyring, 2008).

Comparison of models with observations provides information about both the quality of the climate simulations produced and the accuracy of process simulations with more recent evaluation metrics taking into account not only averages, but variability, trends and emergent constraints as well (Eyring et al., 2019). The ability of climate models to simulate well documented modes of climate variability, *e.g.* El Niño Southern Oscillation (ENSO), are key metrics given that these dynamical patterns affect weather and climate on a global scale (Phillips et al., 2014). Climate model simulations of known modes of variability can be

quantified using defined indices (*e.g.* Stoner et al., 2009) or Empirical Orthogonal Functions / Principle Component Analysis (Casado et al., 2009). Once calculated, the location and intensity of the modes, as well as variability over time, can be compared across models (Stoner et al., 2009; Casado and Pastor, 2012). Seasonal cycles can be evaluated by comparing the mean seasonal cycle across a time period to identify model biases and over- (or under-) estimated variation (Chou et al., 2014) as well as evaluating seasonal means for variables of interest (Vidale et al., 2003). Emergent trends are another important consideration, particularly if climate model data is to be relied upon for future projections or attribution. Climate model performance in simulating already observed warming can be evaluated against time series of *e.g.* global mean temperature (Flato et al., 2013) by comparing underlying trends over many years, although time-dependent forcings do have to be considered for these comparisons to be realistic (Allen et al., 2000). Alternatively, regression-based methods test if trend coefficients differ significantly between models and observations (Santer et al., 2008; Mckitrick et al., 2010).

Although many climate model evaluation metrics focus on comparison of mean state values, emphasis is increasingly being placed on the ability to model extreme events at the tails of the distribution (Fischer and Knutti, 2015; Fischer et al., 2021). These low frequency, higher-order statistics are naturally more challenging to accurately represent in global climate models than mean state behaviour (Flato et al., 2013). Previously mentioned methods to quantify skill in modelling particular climate processes are certainly relevant here, where there is frequently a desire to understand the underlying drivers of a particular event. Comparisons between models and observations may focus on the reproduciblity of extreme event statistics such as daily minima/maxima or number of days exceeding a fixed threshold (Klein Tank et al., 2009; Fan et al., 2020). The threshold of N-year return period events can be compared between models and observations or conversely the return period of an event of defined magnitude can be calculated (Toreti et al., 2013; Angélil et al., 2016). Parameters of fitted extreme value distributions can also be compared; either modelling those values which exceed a defined threshold (*e.g.* Chan et al., 2014) or a so-called 'block maximum' approach where extreme value are obtained by selecting the maximum value recorded in each time block (Kharin and Zwiers, 2000; Klein Tank et al., 2009). Newly developed statistical approaches seek to compare the tails of distributions without fitting a parameterised distribution instead applying adapted statistical tests to events exceeding an extreme value threshold (e.g. Toreti and Naveau, 2015). The null hypothesis that the two sets of samples (climate model and observations) are realisations of the same underlying distribution is rejected if the test statistic (which quantifies the difference between the samples) exceeds the critical value at the chosen significance level.

These comparisons of models with observations can be performed between spatially av-

eraged quantities or at local and regional spatial scales. Global patterns of performance also provide a means for model comparison. This idea can be adapted to average performance statistics over different regimes, for example circulation *vs.* thermodynamics, to identify processes which require improvement (Flato et al., 2013). Correlation between the performance metrics achieved by a particular model for different climate variables can also be a valuable model evaluation and comparison tool (Gleckler et al., 2008).

With GCMs providing output on a global scale, corresponding observational data sets are required to produce the ground truth climatology against which to evaluate models. Observational data sets also play a role in model development and tuning processes prior to final model evaluation (Mauritsen et al., 2012); these steps in the model development process are important to consider to place model evaluation scores in full context. Whilst global observations play an essential part in the climate model evaluation process, the role of measurement uncertainties and biases also needs to be considered (Collins et al., 2013; Kotlarski et al., 2019), as well as the timescales of particular variable definitions. For example, comparing daily-mean values from climate models with once daily measurements from satellites (Eyring et al., 2019). One approach to dealing with these instrumental uncertainties is to use instrumental simulators which mimic the output a satellite would provide if it were 'observing' the climate model (Flato et al., 2013). As an alternative to direct observations, reanalysis data sets, such as the ERA5 data set used in this thesis, combine satellite observations with other measurements and short-range weather forecasts through data assimilation (Hersbach et al., 2020) to produce a comprehensive picture of the global climate system at a particular point in time.

While the methods outlined above allow for comparison between individual model output and observations, there is also value in comparing climate models directly with each other. An early example of this strategy is the Atmospheric Model Inter-comparison Project (AMIP): participating modelling centres ran simulations with a specified set of boundary conditions and parameter values (Phillips, 1996). By focusing on the atmospheric component of the model and forcing the model with observed, *e.g.* sea surface temperatures (SSTs), biases in SST patterns can be controlled for. This prevents a cyclical effect where SST biases lead to biases in atmospheric variables, which subsequently turn into feedbacks on SSTs such that the original source of the bias - ocean or atmospheric components - is harder to identify. AMIP has since evolved into the generations of the Coupled Model Inter-comparison Project (CMIP) which simulates not only historical climate but also includes pre-industrial control runs, climate sensitivity experiments and projections of future climate under defined future emissions scenarios (Eyring et al., 2016). Variance across the model ensemble can also provide information about uncertainty in future model projections, this will be discussed in more detail in Chapter 4. Large, multi-model ensembles constitute powerful tools for climate

model evaluation, although simply identifying the presence of a bias does not necessarily lead directly to the root cause or a modelling solution. This is where it becomes necessary to apply bias correction methods as a post-processing step.

3.1.2 Bias Correction

While climate models certainly have the predictive skill to provide useful outputs, the presence of climate model biases must also be recognised. The identification of model biases in existing global climate model output motivates the application of bias correction - a suite of methods to bring climate model output in line with historical observations. Traditionally the term bias correction refers to a statistical post-processing calculation to bring the distribution of, for example, temperature into better agreement with observations. Here the term is used more broadly to describe any process which improves agreement of the output temperature distribution of a global climate model with a reference reanalysis distribution (in this case ERA5). Traditional statistical approaches are contrasted with the new methodology proposed here to "bias correct" the physical relationships which underpin the output temperature distribution by applying the relationships learnt by Ridge-ERA5. In general bias correction methods are assumed to be stationary in time so the same correction process would be applied to both historical and future scenario simulations (Teutschbein and Seibert, 2012). With bias corrected climate model output frequently being used as a basis for real-world decision-making in terms of climate change adaptation and mitigation planning (Hawkins and Sutton, 2009; Eyring et al., 2016), it is essential that the methods used are justifiable (Maraun, 2016). This subsection outlines a range of methods that have been applied in this field including consideration of the advantages and potential drawbacks of each method.

Traditional bias correction approaches are based on some form of scaling where climate model output is re-scaled such that it better agrees with observed values. The simplest example of this would be linear scaling (e.g. Shrestha et al., 2017), where a scaling between the mean value of the model data and the observed mean is applied to all model output. To apply an additive linear scaling to temperature (Londhe et al., 2023) at daily time resolution from a model simulation, $T_{hist,d}$ with corresponding observations, $T_{obs,d}$, and long-term, discrete monthly means $\mu(T_{hist,m})$ and $\mu(T_{obs,m})$ respectively, the corrected daily model output, $T^*_{hist,d}$, on the d^{th} day of the m^{th} month is given by :

$$T^*_{hist,d} = T_{hist,d} + (\mu(T_{obs,m}) - \mu(T_{hist,m}))\,. \tag{32}$$

This bias correction method is demonstrated with artificial observational and model data in Figure 3.1. Applying linear scaling to the model data shifts the distribution so that the

mean of the model data agrees with the observed data. With this approach, by definition,

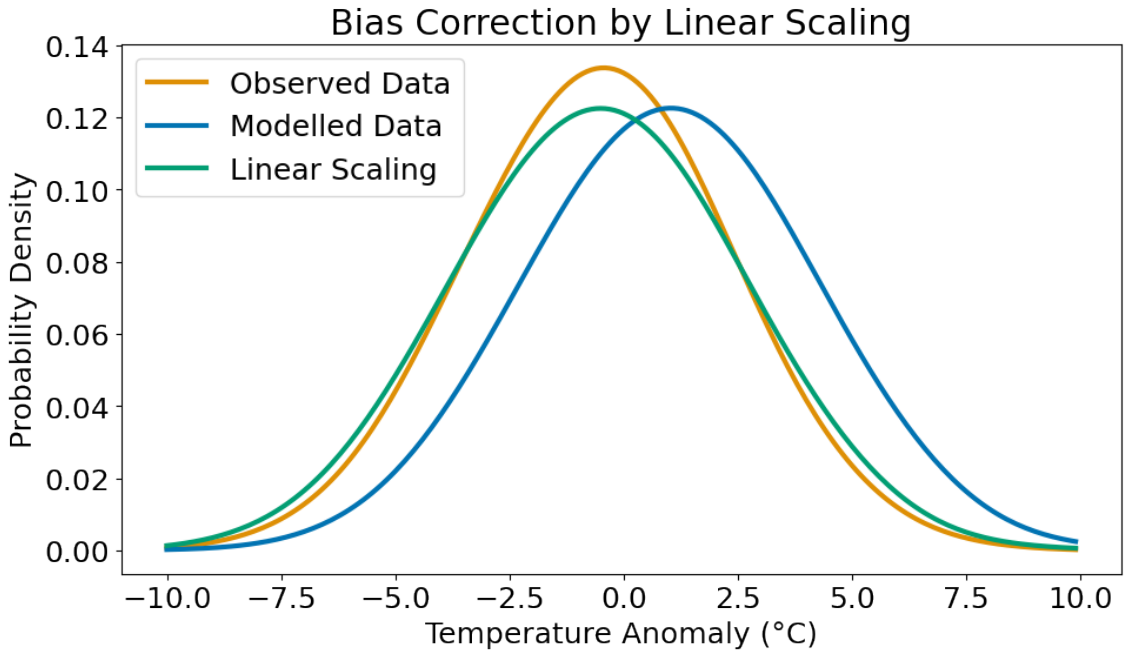

Figure 3.1: Demonstration of a linear scaling applied to artificial data. Model data (blue) is corrected relative to observed data (orange) to produce the bias corrected distribution (green).

the corrected monthly mean values from the model will perfectly agree with those that were observed (Teutschbein and Seibert, 2012). A linear scaling only has the ability to correct a mean bias; to correct discrepancies in variance requires alternative techniques.

Power transformation can be applied to variables like precipitation to adjust the standard deviation of the time series however this is not appropriate for temperature which is known to be approximately normally distributed as the power transformation would result in a non-normal output (Terink et al., 2010). Instead, temperature values can be mean shifted and scaled relative to the observed variance, $\sigma(T_{obs,m})$, the so-called *variance scaling* method (Terink et al., 2010; Chen et al., 2011), resulting in a similar correction:

$$T^*_{hist,d} = \mu(T_{obs,m}) + (T_{hist,d} - \mu(T_{hist,m})) \frac{\sigma(T_{obs,m})}{\sigma(T_{hist,m})} \tag{33}$$

Variance scaling is demonstrated in Figure 3.2 using artificial data. In contrast with the linear scaling example in Figure 3.1, the variance scaling factor adjusts the width of the distribution as well as shifting the mean.

The family of distribution mapping, or quantile mapping, techniques seeks to map the distribution of modelled values on to the distribution of observed values. There is more than

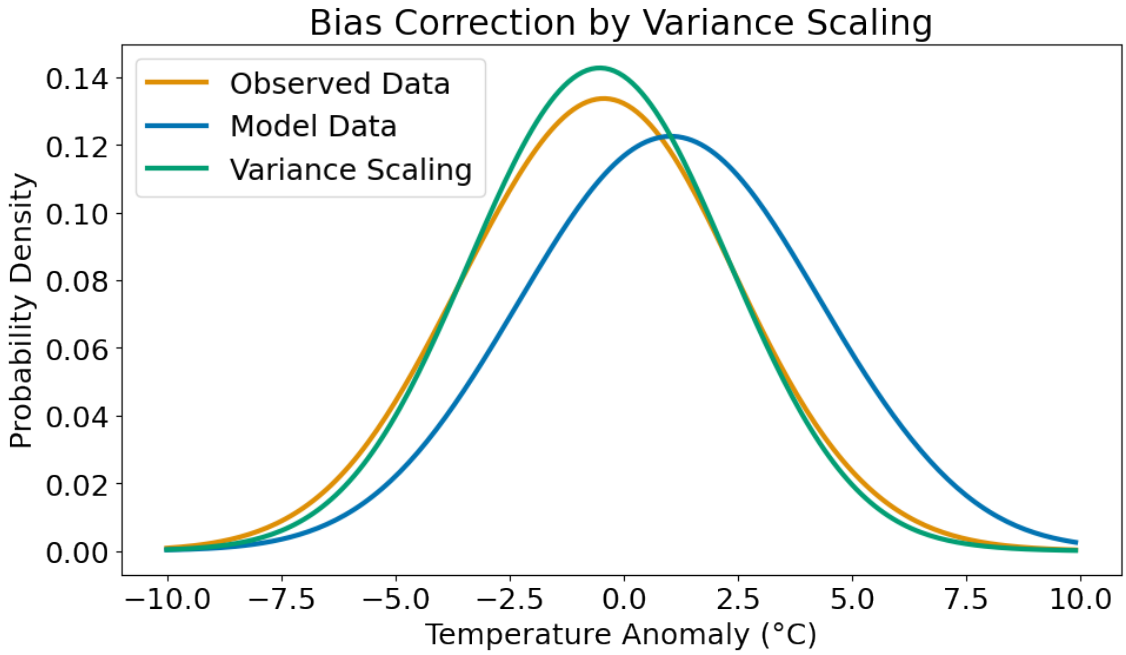

Figure 3.2: Demonstration of a variance scaling applied to artificial data. Model data (blue) is corrected relative to observed data (orange) to produce the bias corrected distribution (green).

one way to approximate a solution to this transformation (e.g. Gudmundsson et al., 2012). For variables which can be assumed to follow a normal distribution, a normal distribution mapping can be applied (Qian and Chang, 2021) by estimating model and observed values for the distribution parameters to calculate the modelled and observed cumulative density functions (CDFs). To transform a modelled value, X_{hist}, its quantile in the historically modelled distribution is first calculated using the modelled CDF, F_{hist}. The observed inverse CDF, F_{obs}^{-1}, is then applied to this quantile to obtain the bias corrected value, X_{hist}^{*}:

$$X_{hist}^{*} = F_{obs}^{-1}(F_{hist}(X_{hist})). \tag{34}$$

The quantile mapping approach is demonstrated on artificial data in Figure 3.3. This method works very well on this data as the data was generated by sampling from a normal distribution meaning the assumption that the data is normally distributed clearly holds true, however this is not the case for many climatic data sets *e.g.* rainfall. A common, non-parametric approach is to use empirical quantiles; CDFs are estimated at regular quantile intervals for both the model output and observations with linear interpolation being used for values which fall within these intervals (Gudmundsson et al., 2012). Advantages of this distribution mapping approach include the ability to scale both mean and variance whilst maintaining extreme values (Londhe et al., 2023), however assumptions may need to be made about the expected distribution of the data.

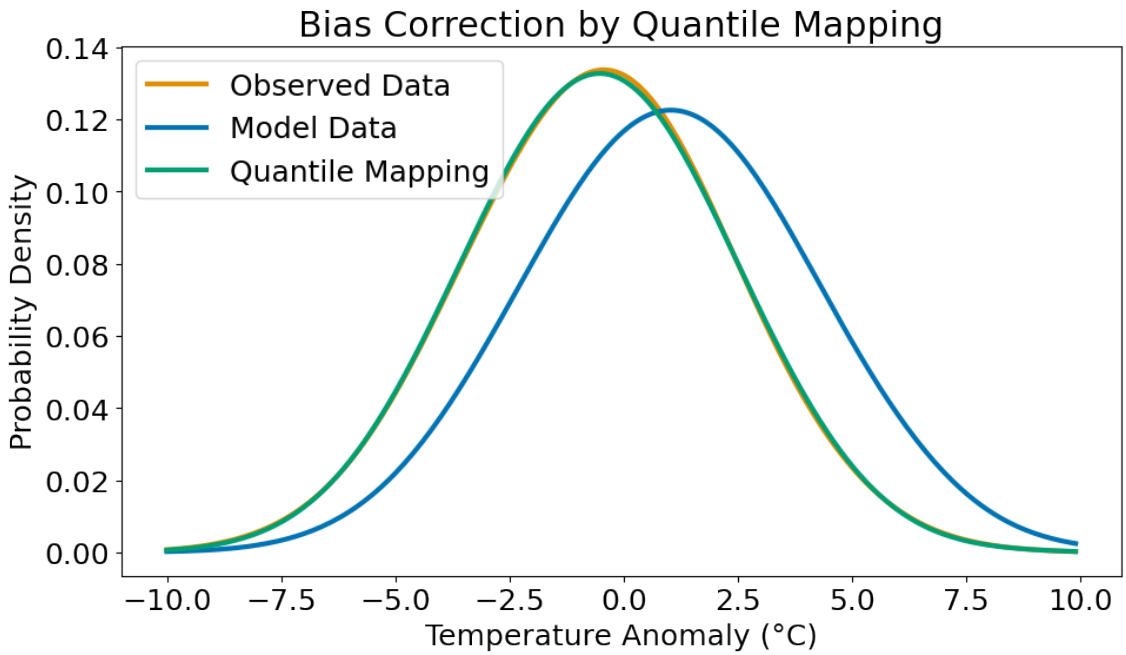

Figure 3.3: Demonstration of a quantile mapping assuming normally distributed data applied to artificial data. Model data (blue) is corrected relative to observed data (orange) to produce the bias corrected distribution (green).

An additional layer of complexity is considered in multivariate bias correction methods where multiple variables are corrected simultaneously while taking into account variable inter-dependencies (*e.g.* Cannon, 2016), such approaches may also consider spatial dependencies of locations on surrounding grid cells (Wang et al., 2022). These approaches aim to reduce the possibility of producing corrected climate model outputs which are physically inconsistent *i.e.* outputs which are incompatible with known physical laws (Agbazo and Grenier, 2020). Despite several theoretical advantages of multivariate bias correction, results of inter-comparison with uni-variate techniques are not clear-cut (Li et al., 2014; Chen et al., 2018; François et al., 2020) particularly under climate change conditions (Van De Velde et al., 2022). This may be attributable to assumptions about the inter-variable dependency relationships being the same in models as they are in observations which may not be true in general (Meyer et al., 2019).

3.1.3 Machine Learning for Bias Correction

With the increasing prevalence of ML in the environmental sciences, and particularly in the field of climate science (see Section 1.6), it is not surprising that ML models have been applied in the context of bias correction. The ability of ML to identify patterns between variables clearly appears promising for attempting transformations of climate model output for better agreement with historical observations, particularly given recent progress in image to image translation (*e.g.* Pang et al., 2022) which provides an avenue for capturing the relationships between model outputs and observations. This section highlights some specific examples of these methods.

An obvious application of ML models for bias correction is to learn relationships which transform model output to better agree with observed climate. For correcting forecast models this is a relatively simple process since the predicted values aim to perfectly match the observed time series. An ML model can be applied to learn any systematic differences between the forecast and reality and correct those differences for subsequent predictions, see Figure 3.4. Models such as Artificial Neural Networks (ANNs) and Random Forests (RFs) have been applied in this field showing competitive and even superior results to traditional statistical alternatives (Watson, 2019; Cho et al., 2020). ML models have several other advantages over traditional methods; they deal better with multiple co-linear input variables, are easily automated and can capture non-linearities.

An ML approach to bias correcting climate model output requires a slightly different approach given that, as mentioned previously, climate models are not intended to replicate the observed historical time series on a daily basis and therefore no ground truth exists for unbiased climate model output. To get around this problem, relationships can be learnt be-

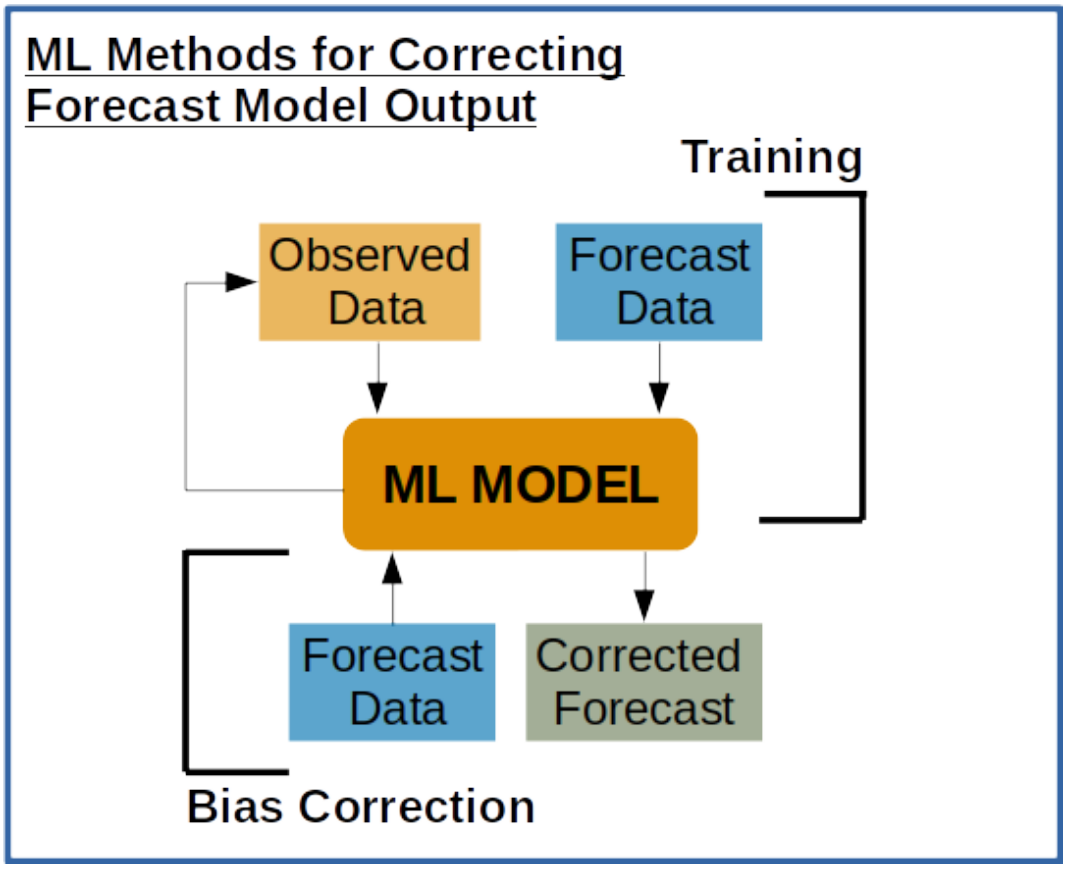

Figure 3.4: Steps to apply machine learning for bias correction of forecast model data relative to observations.

tween other climate variables and the target variable for bias correction from observational or reanalysis data and before applying the trained ML model to climate model output (Tan et al., 2021). Alternative methods (Pan et al., 2021; Fulton et al., 2023), some combining bias correction with down-scaling (e.g. Ballard and Erinjippurath, 2022), apply Generative Adversarial Networks (GANs) for unpaired (necessary because of the absence of ground truth data sets) image to image translation (Park et al., 2020). The GAN works by concurrently training a generator: in this case attempting to produce bias corrected climate model fields, and a discriminator: learning to distinguish between those outputs and fields from observations, see Figure 3.5. As both components of the network improve so too does the degree of bias correction; producing fields which are more and more difficult to distinguish from reality.

3.2 Performance Evaluation of Ridge-ERA5

A detailed description of the training process for the process-based Ridge-ERA5 model including ML method choice and variable selection can be found in Chapter 2, here only a brief overview is given. Ridge-ERA5 is trained at each $3^\circ \times 3^\circ$ Northern Hemisphere grid location to predict the 2m temperature anomaly to a smoothed mean seasonal cycle during the summer months of June, July, August (JJA).

In order to construct a prediction time series covering 1979-2022 without predicting on data already seen during training, a leave-one-year-out approach is taken. For each year in the time period (1979-2022), that year is held out whilst all other years are used for training and cross-validation. The resulting model is then used for prediction on the held-out test year. Repeating this process for each year in the series results in a complete test time series of summertime temperature anomalies. Spatial variance of Ridge-ERA5 performance is analysed by plotting spatial maps of quantitative performance measures. Predictive skill of Ridge-ERA5 is measured with R^2 scores (see Section 2.1.5) which quantify the proportion of variance explained by the model across the time series compared with the true ERA5 variance. Mean absolute error plots are also used to show how performance varies by location. At individual locations, time series are plotted to compare Ridge-ERA5 predictions with ground truth ERA5 data.

To give an overview of Ridge-ERA5 performance on held out ERA5 test data, R^2 scores and mean absolute errors of test predictions across the historical time period (1979-2022) obtained via the leave-one-year-out approach are plotted in Figure 3.6. Most land areas exhibit relatively high R^2 scores (> 0.7) with expected similar patterns visible in the mean absolute error scores. Performance suffers the most over the ocean and in Northern Siberia, this is most likely related to predictor variables being chosen for mid-latitude summer rather

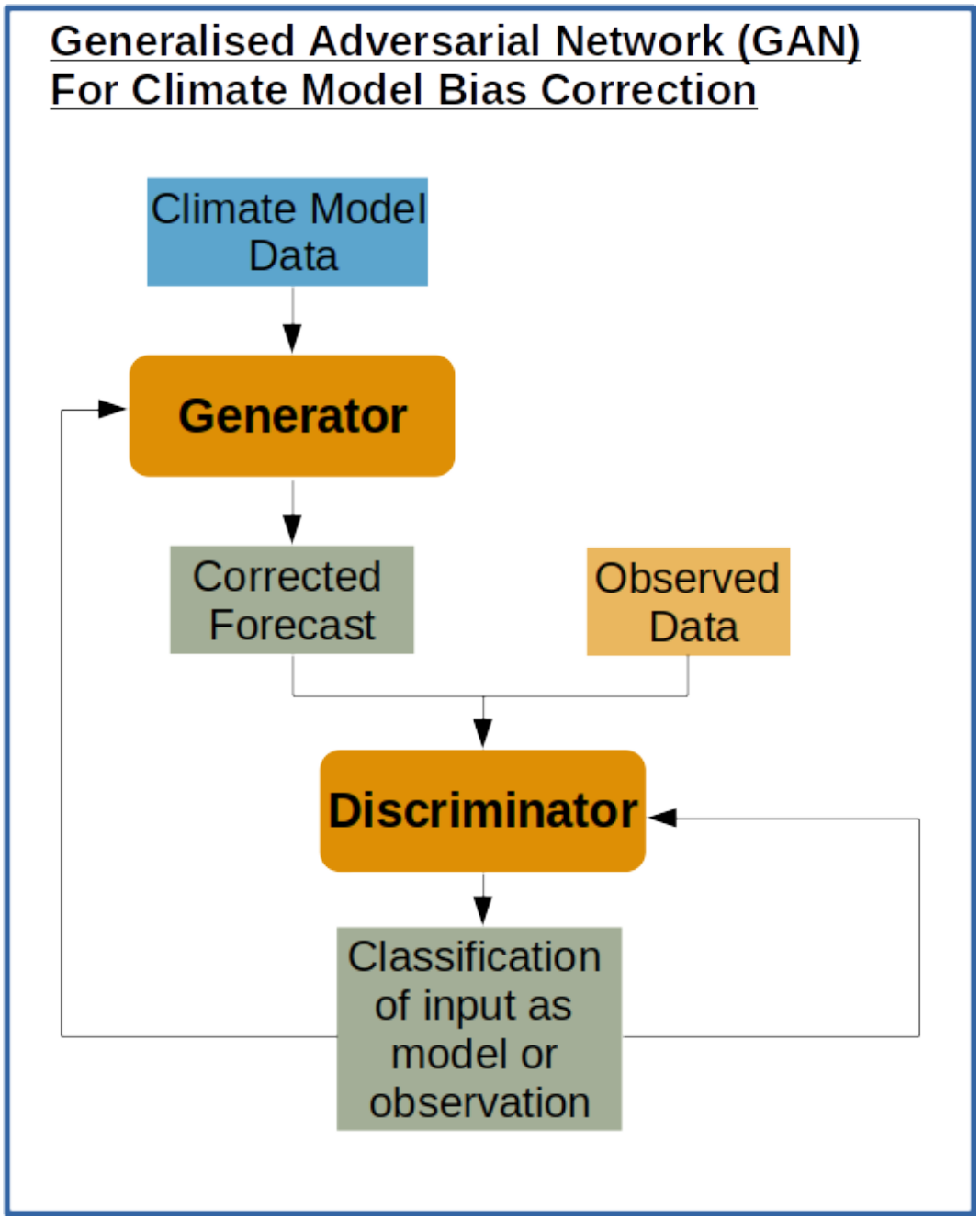

Figure 3.5: Schematic diagram of a Generalised Adversarial Network (GAN) for image to image translation.

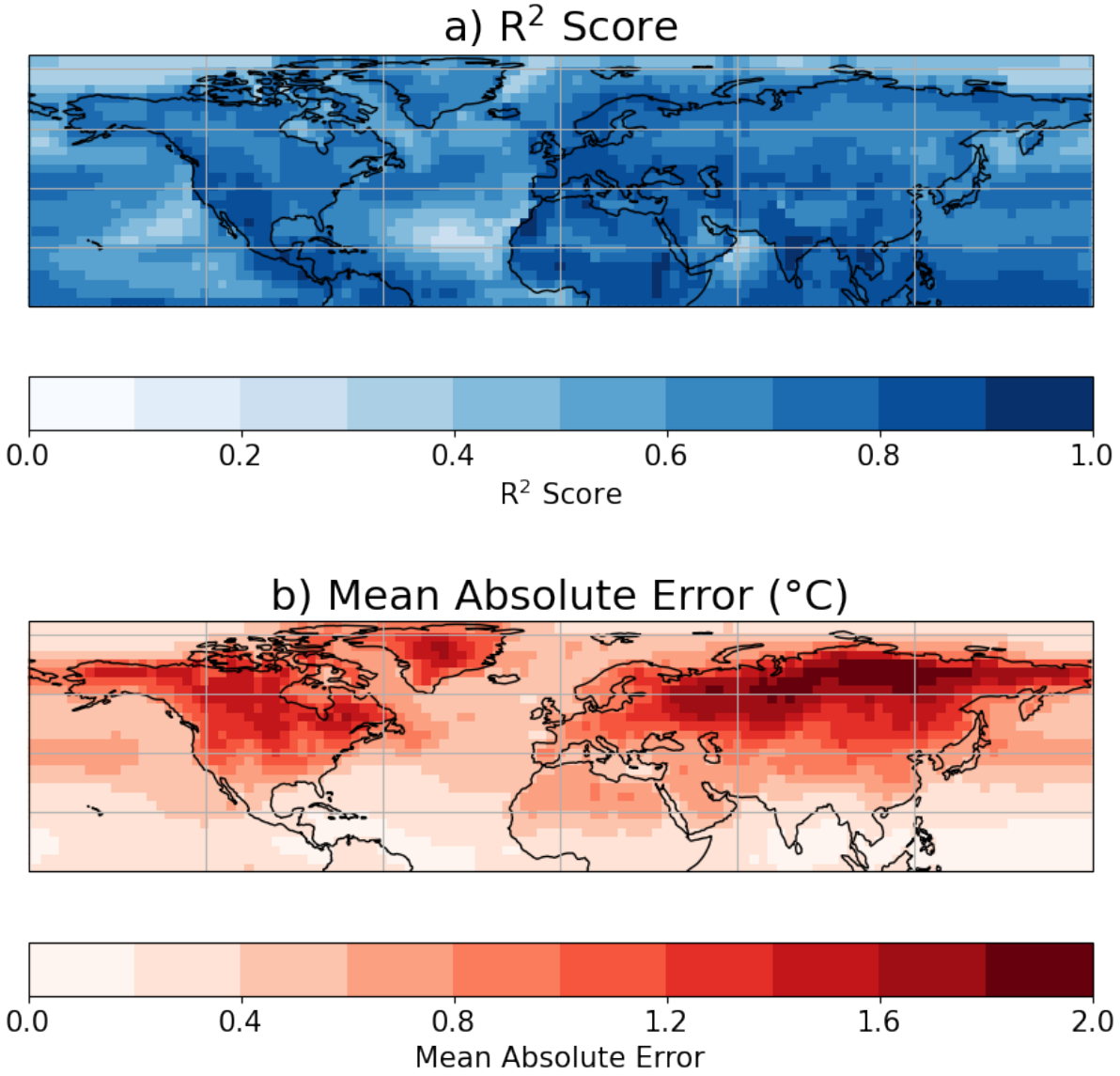

Figure 3.6: R^2 scores (a) and mean absolute errors (b) evaluating Ridge-ERA5 predictions against ERA5 data for summertime (JJA) temperature anomalies 1979-2022.

than winter. For example, there are no snow variables or measures of surface albedo changes included in the model which would likely be necessary to capture key processes such as moisture availability for evaporation and the surface-albedo feedback which undoubtedly play a role in heat extremes in these regions (Sato and Nakamura, 2019; Zhang et al., 2020; Marquardt Collow et al., 2022).

Some examples which demonstrate this distribution of skill are plotted in Figure 3.7 which shows time series of predictions from Ridge-ERA5 for one year of the historical time period at three locations. Where performance scores are high, the Ridge-ERA5 prediction time series

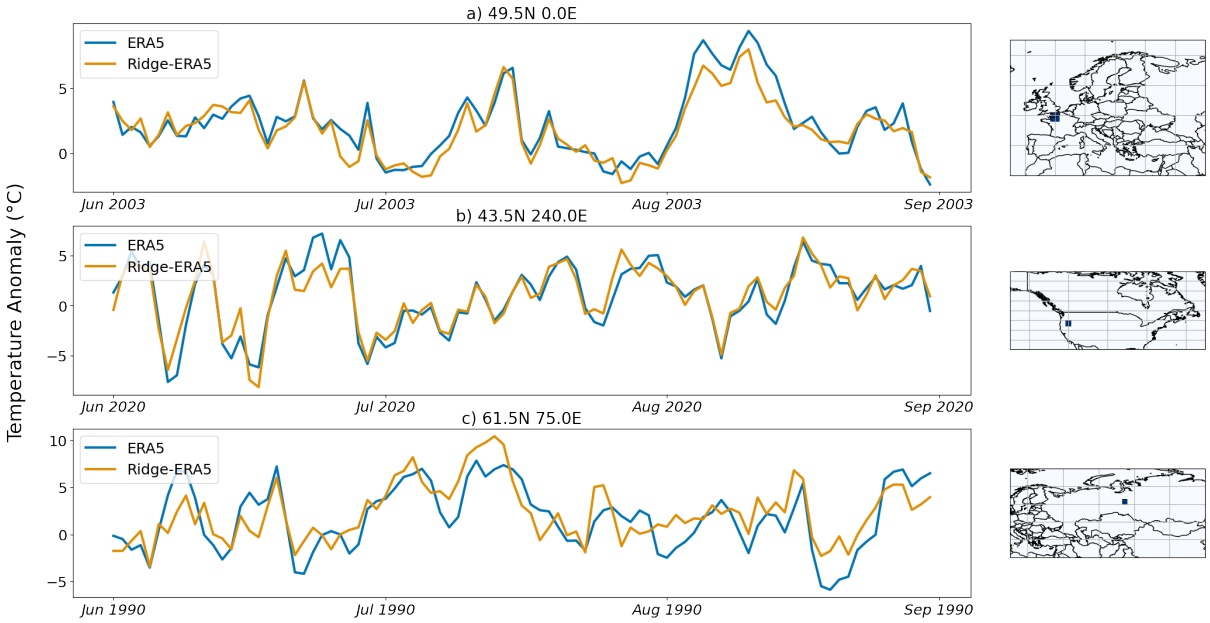

Figure 3.7: Time series of Ridge-ERA5 predictions (orange) *vs.* actual ERA5 data (blue) at 49.5N 0E in 2003 (a), 43.5N 240E in 2020 (b) and 61.5N 75E in 1990 (c).

shows excellent day-to-day agreement with the true ERA5 time series (Figure 3.7a, b). The magnitude of most anomalies is accurately captured as well as the patterns of daily variation; even during exceptionally warm years such as the 2003 European heatwave which is investigated in more detail in Chapter 5. In areas where Ridge-ERA5 struggles, such as in Siberia, there are instances both of correctly modelled patterns but incorrect temperature anomaly magnitudes but also temporal incoherence and false patterns of variation (Figure 3.7c).

Analysis now focuses on the limitations of Ridge-ERA5 and the boundaries at which performance starts to break down. Setting one year aside for testing leaves a 44-year training time series with 92 days of data from JJA of each year resulting in 4048 training data points. So, for 95^{th} percentile events, there are 202 data points, for 99^{th} percentile events there are 40 training data points and for 99.9^{th} percentile events there are only 4 data points. These

limitations are important to quantify as they will inform the definition of extreme event that can reasonably be applied to this data as well as giving context for predictions of historical heatwaves which are analysed in more detail in Chapter 5 using a 95^{th} percentile threshold. A first indication of the distribution of performance across the temperature distribution is given in Figure 3.8 which plots how mean absolute error varies across the percentiles of Northern Hemisphere summertime temperature anomalies over land in the ERA5 data set. Performance naturally is better towards the centre of the distribution where most training

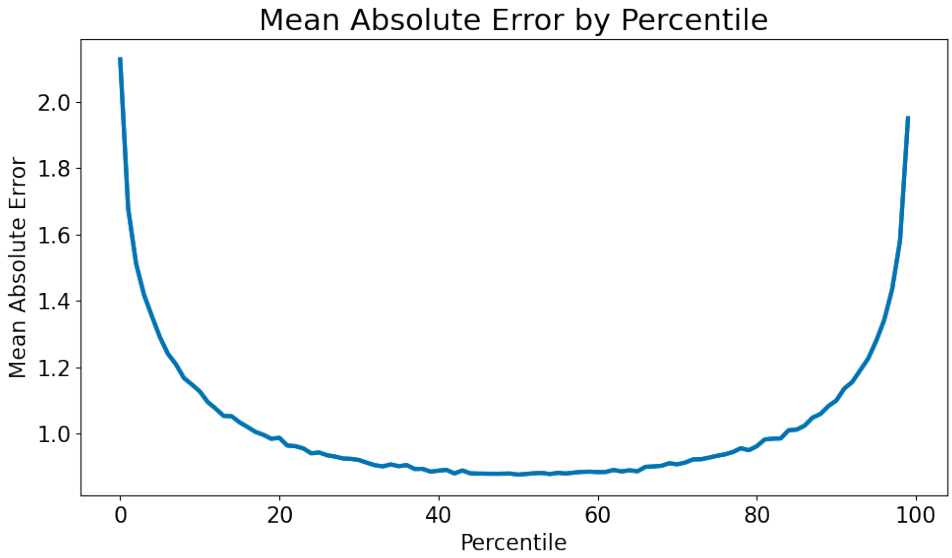

Figure 3.8: Mean absolute error (MAE) for temperature anomalies in each percentile of the ERA5 compiled on a grid cell by grid cell basis then averaged over Northern Hemisphere land locations.

cases lie and Ridge is able to learn the most robust relationships between predictors and the target temperature. Average mean absolute error starts to rise above one degree around the 80^{th} percentile, however it is important to bear in mind that with temperatures increasing at higher percentiles, larger magnitude of errors are not necessarily larger percentage errors.

To understand how this performance across the temperature distribution varies spatially, predictions *vs.* actual anomalies are plotted on a regional basis. AR6 regions (Iturbide et al., 2020) divide the Earth into 46 land and 15 ocean regions shown in Figure 3.9. They aim to represent consistent climatic regimes that can be used for reference and comparison in studies.

Plots of Ridge-ERA5 predictions *vs.* true ERA5 temperature anomalies by AR6 region are shown in Figures 3.10 and 3.11. They give an indication of how performance is distributed across the range of temperatures measured at a particular location and the point at which Ridge-ERA5 performance start to diminish. Ideally Ridge-ERA5 predictions should cluster

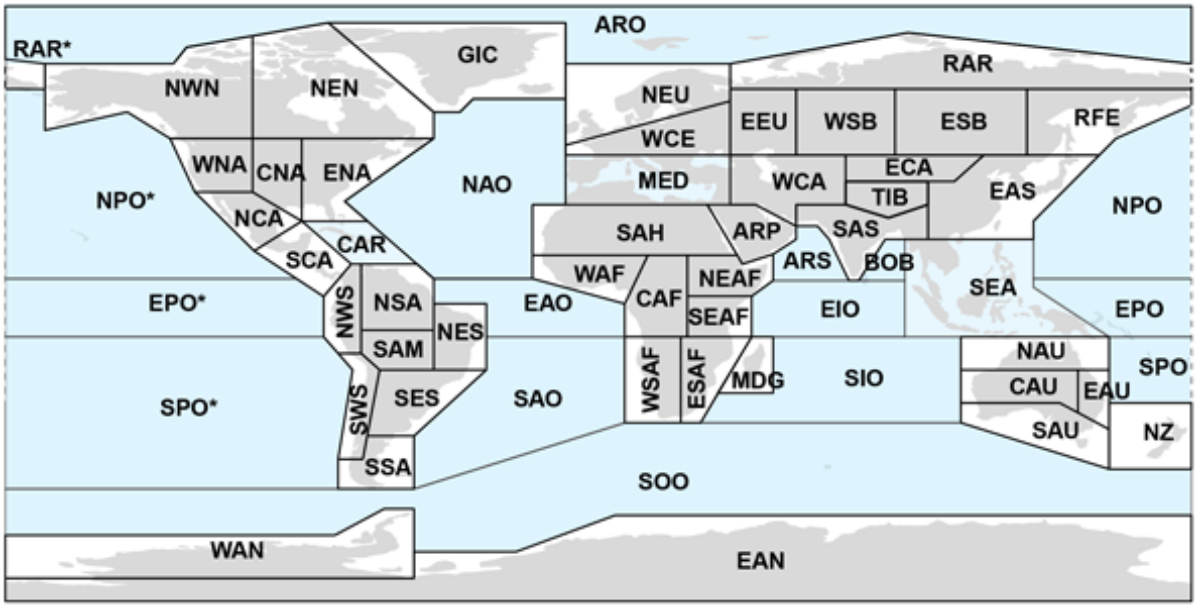

Figure 3.9: Global map of AR6 regions, *Source: Iturbide et al., 2020*

closely to the one-to-one line indicating small errors from true ERA5 temperature anomalies across the range of temperatures observed in a particular region. Across most regions, overall agreement with the one-to-one line is good even for temperature anomalies close to 10^oC compared with the historical seasonal average. Only a small minority of regions show a systematic under-prediction of warm anomalies and this is primarily confined to the more extreme days at a particular location, for example in Western North America (Figure 3.10a) and East Asia (Figure 3.11j). The spread of points around the one-to-one line gives an idea of the error range in predictions for a particular region. The performance over Northern (Figure 3.11c) and Western Africa (Figure 3.11a) as well as Central America (Figure 3.10d, e) and South Asia (Figure 3.11l) is particularly good in this aspect with a tighter fit of points to the one-to-one line.

As the included variables were selected with drivers for mid-latitude, land regions in mind, subsequent analysis will only include those grid cells with a land area fraction greater than 0.7. Poor performance over Siberia and Northern Russia, is likely explainable by lack of variables providing information about snow cover, snow melt, snow depth or surface albedo. These regions, along with regions of similar latitude, (North East/West North America and Greenland/Iceland) are excluded from future summary statistics and figures.

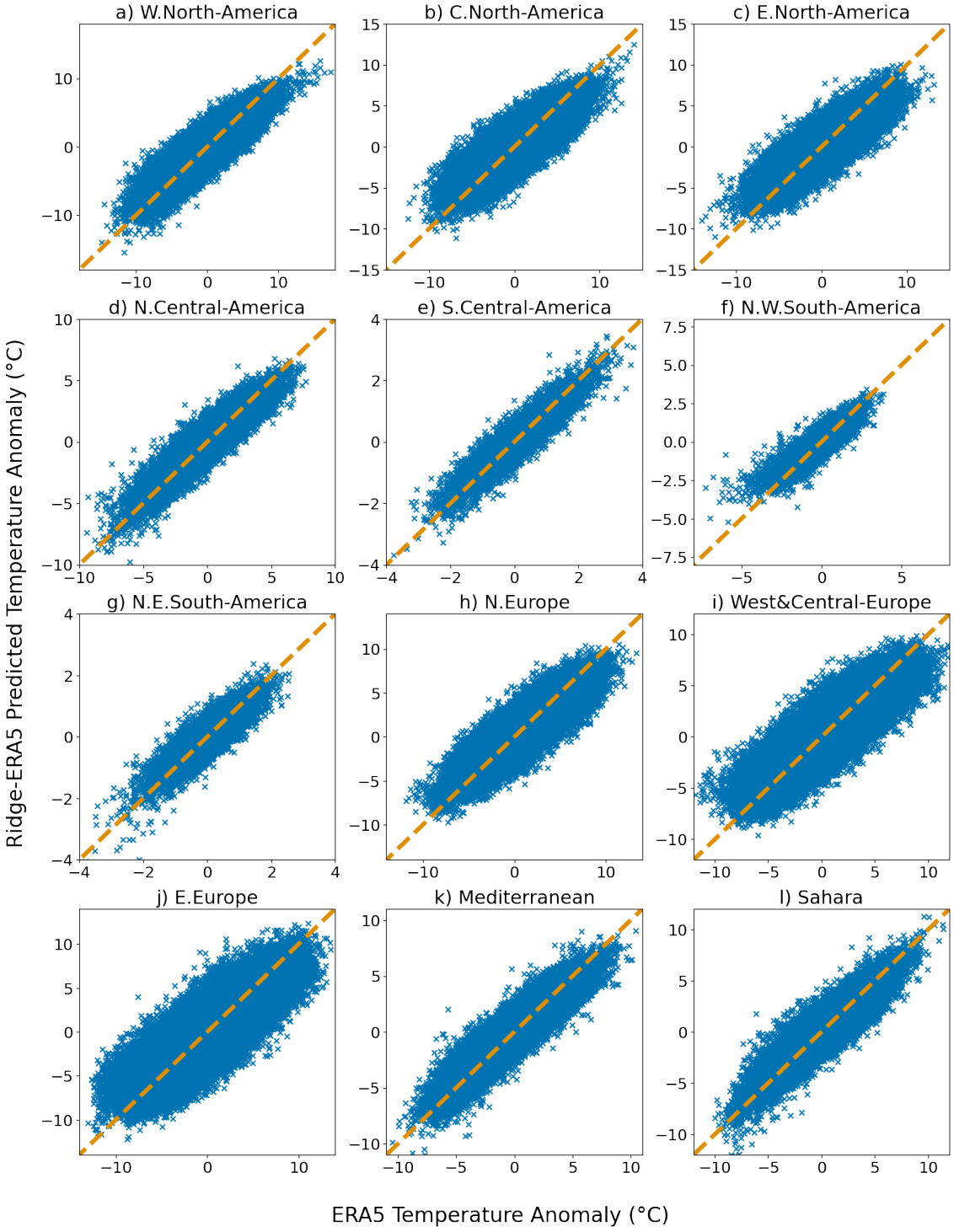

Figure 3.10: ERA5 temperature anomalies (x-axis) *vs.* Ridge-ERA5 test predictions (y-axis) for JJA 1979-2022 for all grid cells in each Northern Hemisphere AR6 region.

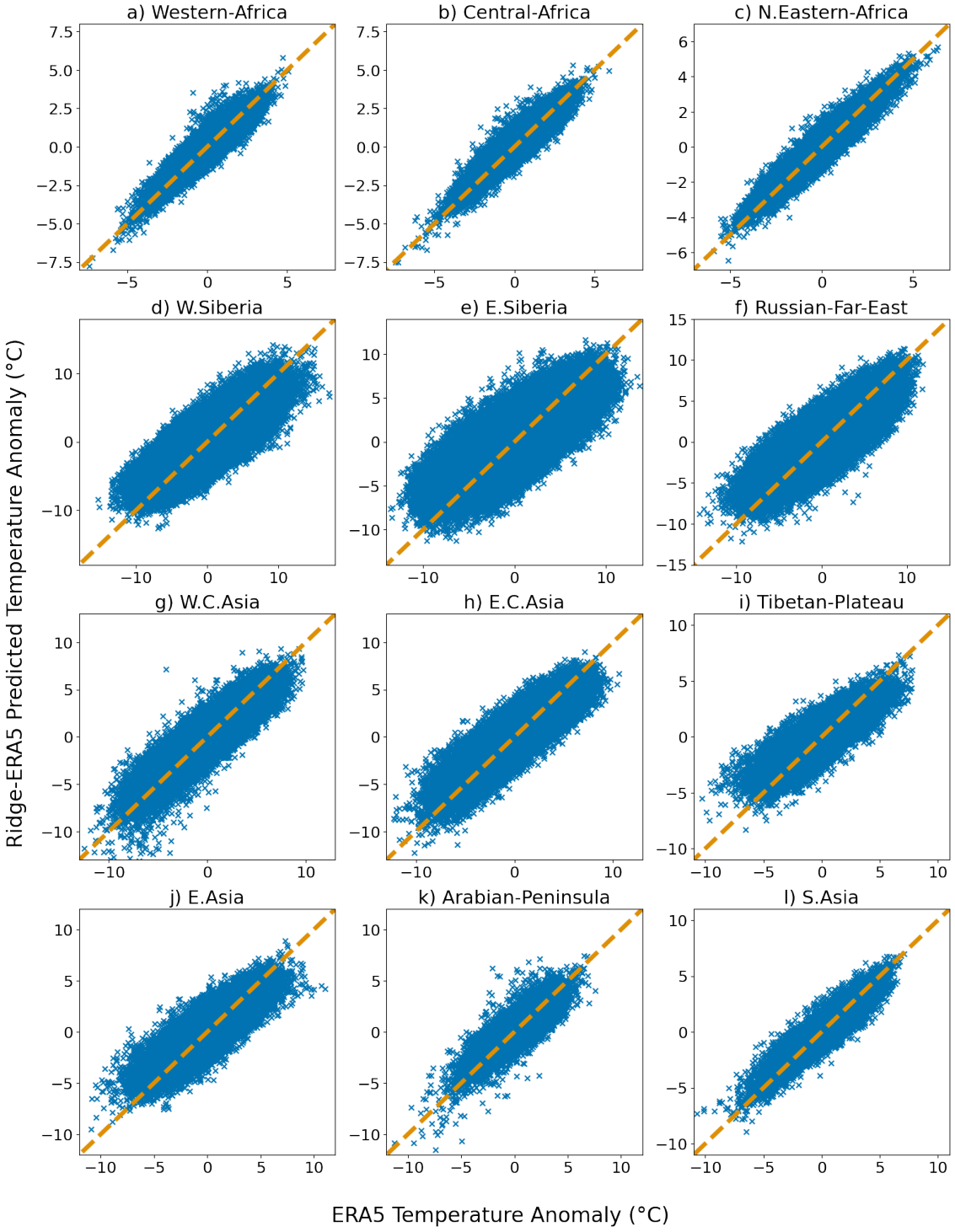

Figure 3.11: ERA5 temperature anomalies (x-axis) *vs.* Ridge-ERA5 test predictions (y-axis) for JJA 1979-2022 for all grid cells in each Northern Hemisphere AR6 region.

3.3 Interpreting Ridge-ERA5 Coefficients

In the following sub-sections, coefficients learnt by Ridge-ERA5 will be used alongside existing climate model data both for bias correction of historical temperature anomalies and as a model evaluation tool with implications for future warming projections. These methods are laid out in Figure 3.12 which summarises how Ridge-ERA5 is used to complement existing climate model data. Ridge-ERA5 is trained (Figure 3.12a) using historical reanalysis data to predict daily-mean temperature anomalies to an average seasonal cycle during Northern Hemisphere summer (JJA). For bias correction (Figure 3.12b), Ridge-ERA5 coefficients are combined with predictor variables from CMIP6 models to produce new realisations of historical temperature anomaly (y_{bc}) distributions:

$$y_{bc} = f_{era5}(\theta_{era5}, x_{cmip6}). \tag{35}$$

This allows an observational, process-based bias correction to be applied. For climate model evaluation (Figure 3.12c), Ridge-CMIP emulators are trained on historical CMIP6 data to reproduce the pattern of temperature anomalies simulated in each CMIP6 model. Coefficients learnt by the Ridge-CMIP emulators, which represent the processes simulated by the CMIP6 models themselves, can then be compared with Ridge-ERA5 coefficients to perform climate model evaluation.

Before applying Ridge-ERA5 to CMIP6 data, the regression coefficients learnt by Ridge-ERA5 are interpreted to identify the role of individual predictor variables and the physical processes they may represent. Mean coefficient values averaged across the domain size of each predictor variable for Ridge-ERA5 models trained at each Northern Hemisphere grid cell are plotted in Figure 3.13. All predictor variables are normalised prior to training so coefficient magnitudes can be directly compared across predictor variables to infer relative importance.

As expected, soil moisture (Figure 3.13a) shows a broadly negative relationship with temperature; drier soils mean less evaporation and thus warmer temperatures. The contribution from soil moisture is stronger at mid to high latitudes; regions with greater moisture availability (Jiang et al., 2022) potentially because these regions are less frequently in a non-linear soil moisture regime where maximum soil dryness is reached and subsequent soil moisture anomalies remain constant. In these higher latitude regions, soil moisture is one of the larger contributors consistent with existing literature that frequently cites soil moisture deficits as an exacerbating factor in mid-latitude heatwaves (see Section 1.2).

The relationship between cloud cover and temperature is broadly negative representing increased radiative heating resulting from clear sky conditions (Figure 3.13d). As the cloud cover variable measures total vertical cloud cover the effects of high and low cloud are not sep-

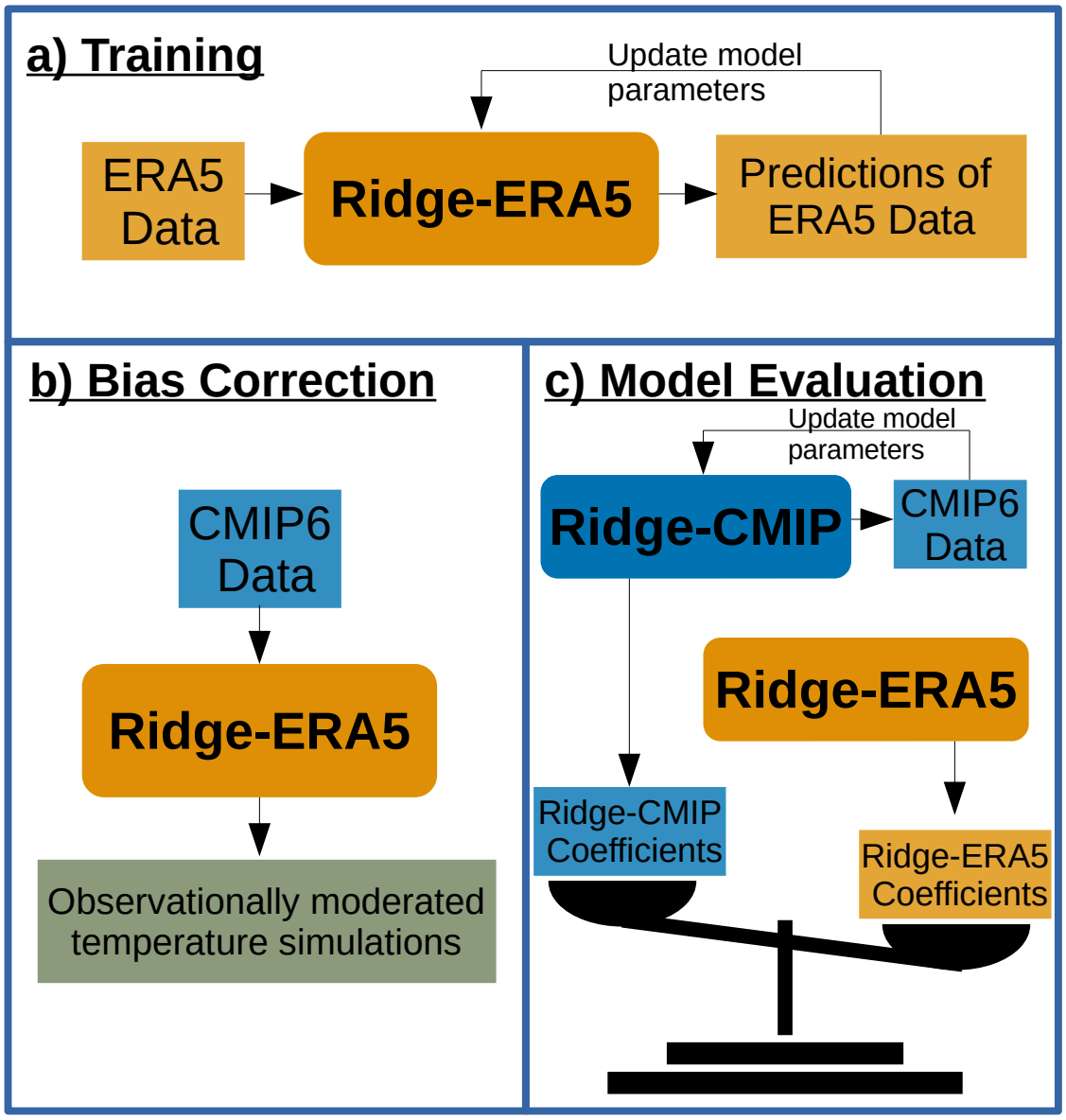

Figure 3.12: The process of training the Ridge-ERA5 model on reanalysis data (a), combining CMIP6 predictor variables with coefficients learnt from reanalysis to produce bias corrected historical temperature simulations (b) and training Ridge-CMIP emulators of existing CMIP6 models so that coefficients representing the processes simulated by those CMIP6 models can be compared with Ridge-ERA5 coefficients for climate model evaluation (c).

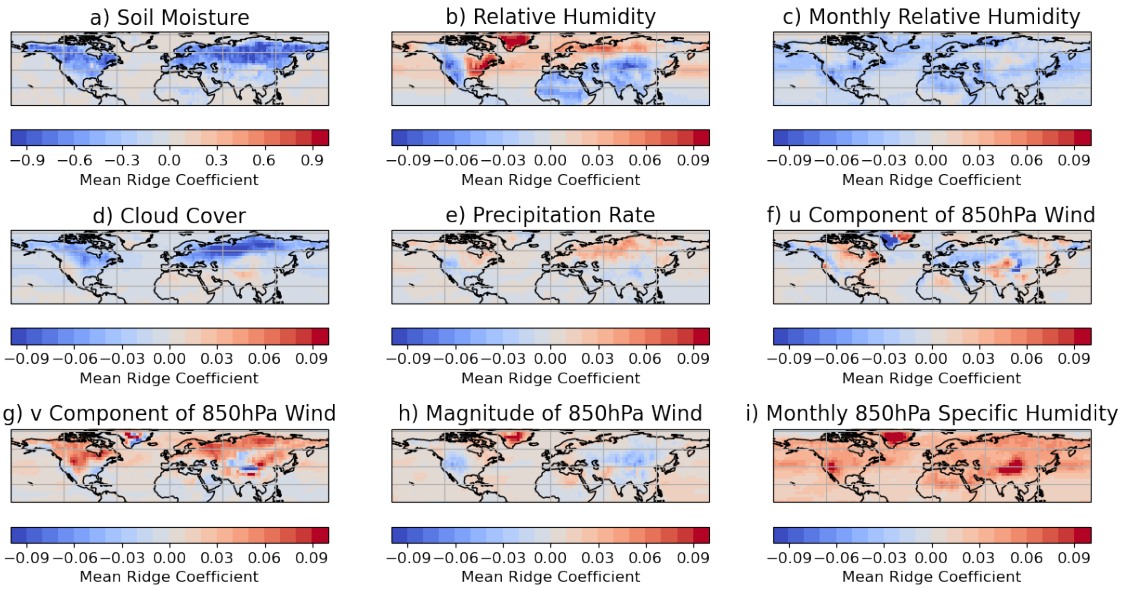

Figure 3.13: Maps of mean regression coefficients across domain size for Ridge-ERA5 models trained at each 3° Northern Hemisphere grid cell to predict daily 2m temperature anomaly during JJA.

arated. The variable provides Ridge with different information depending on the distribution of cloud cover with lower clouds towards the poles while higher clouds are most prominent in the Tropics (Shikwambana, 2022). In areas where this relationship is positive, higher altitude clouds are likely associated with heat trapping. Similarly to soil moisture, cloud cover coefficients make a relatively large contribution to temperature anomalies at higher latitudes where the relationship is negative.

The long-term mean variables of near-surface relative humidity (Figure 3.13c) and 850hPa specific humidity (Figure 3.13i) exhibit relationships with temperature that are consistent across the hemisphere, this is to be expected as the purpose of including these variables was to provide Ridge-ERA5 with information about the background warming state. They are also relatively weaker contributors likely because they are unable to provide the model with information about variations on a daily timescale. The positive coefficients representing the relationship between monthly 850hPa specific humidity and temperature as well as the negative coefficients representing the relationship between monthly near-surface relative humidity and temperature can both be related back to the Clausius-Clapeyron relationship (*e.g.* Held and Soden, 2006). The saturation vapour pressure increases by approximately 7% for every 1K increase in temperature meaning that long term temperature rise results in a well documented future trend of increasing specific humidity (Willett et al., 2007). Monthly relative humidity shows a much weaker correlation, but coefficients indicate a negative relationships consistent with a small reduction in relative humidity under future warming as

moisture availability becomes a limiting factor (Held and Soden, 2006; Hardwick Jones et al., 2010; Byrne, 2021).

On a daily basis, distinct relative humidity regimes emerge (Figure 3.13b). For a given value of specific humidity or water transport from the ocean, relative humidity decreases as temperature increases since the capacity for moisture to be held in the air is greater (Hardwick Jones et al., 2010). In the Tropics, where temperature gradients are generally lower, weather systems tend to transport moisture rather than warmer air, when this air is dried out, temperature increases but relative humidity goes down (Byrne, 2021) resulting in a negative relationship between temperature and relative humidity (Figure 3.13b). At higher latitudes, weather systems are also advecting warmer air poleward and raising the temperature so the transport of moisture and warm air is connected and relative humidity has a positive relationship with temperature (Russo et al., 2017). This is particularly clear in Eastern North America where very warm, moist air is transported from the Gulf of Mexico.

The wind variables were included with the intention of providing the model with information about the systems dynamical state. Looking at the coefficient maps for the u and v components of the 850hPa wind (Figure 3.13f, g), patterns of negative coefficients consistent with coastal cooling from an onshore breeze (Zhou et al., 2019) are visible. Across much of the inland extra-tropics, the v component of the 850hPa wind is associated generally with warming which makes sense for pole-ward transport of warm air (Schneider, 2006). There is also some indication of patterns of coefficients surrounding areas of higher altitude.

Some of the weaker contributions come from total wind magnitude (Figure 3.13h) which has a largely negative relationship with temperature over many land regions (Figure 3.13 h) indicating that Ridge has potentially picked up on the relationship between blocking conditions and high temperature anomalies (Nakamura and Huang, 2018) and generally higher winds are associated with cooling. The only areas where this relationship is significantly positive is over Greenland, which is much cooler than the surrounding ocean and wind anomalies are therefore likely to be associated with warm air transport (Jiménez-Esteve and Domeisen, 2022).

The precipitation rate coefficients are also relatively less important contributors. Coefficients between precipitation and temperature (Figure 3.13e) are negative closer to the equator indicating that moisture is generally associated with cooling and hot anomalies occur under drought-like conditions. However, there are some regions at higher latitudes where this relationship is reversed, here increased precipitation rates seem to be tied to transport of warm, moist air rather than cooling (Trenberth and Shea, 2005).

The individual contributions of each predictor variable will be analysed in more detail in Chapter 5 where SHAP value analysis is applied to quantify contributions to historical heatwave events.

3.4 Bias Correction of Historical CMIP6 Simulations Using Ridge-ERA5

In this section, the process-based ML model, Ridge-ERA5, which learns from reanalysis data to predict daily temperature anomalies, is applied as a potential tool for bias correction of climate model output. By combining predictor variables from models in the CMIP6 archive with coefficients learnt from reanalysis data; new realisations of historical temperature time series are produced which are moderated by relationships with temperature that are derived from observations. A key advantage of this approach is the interpretability of the high-dimensional Ridge-ERA5 model; every prediction can be broken down into a contribution from each predictor variable and corresponding coefficient.

In order to plot the distributions of temperature for comparison, Kernel Density Estimators or KDEs (Rosenblatt, 1956; Parzen, 1962) are fitted to each time series. KDEs are superior to plotting histograms as they produce a smooth estimate of the probability density function and are less restrictive than fitting parametric distributions which may make prior assumptions about distribution shape and symmetry (see Section 2.4.1).

To perform bias correction with Ridge-ERA5, predictor variables from CMIP6 models are combined with Ridge-ERA5 coefficients learnt from reanalysis data to produce new temperature simulations. In order to combine CMIP6 inputs with ERA5 coefficients, CMIP6 data needs to be normalised relative to variance in the ERA5 data set:

$$x'_{CMIP} = \frac{x_{CMIP} - \mu_{CMIP}}{\sigma_{ERA5}}. \tag{36}$$

For each predictor variable, from each CMIP6 model, the standard normalisation procedure is carried out using the particular CMIP6 model mean and ERA5 variance according to Equation 36. The resulting historical temperature simulations are compared with the original CMIP6 temperature time series as well as with ERA5 data. Ridge-ERA5 predictions of Ridge-ERA5 temperatures are calculated as:

$$\hat{y}_{era5} = f_{era5}(x_{era5})$$

these prediction can be compared with actual ERA5 temperature anomalies as follows:

$$y_{era5} = \hat{y}_{era5} + \epsilon_{era5} = f_{era5}(x_{era5}) + \epsilon_{era5}.$$

This can be seen in Figures 3.10 and 3.11 where y_{era5} is plotted against $\hat{y}_{era5}$ and ϵ_{era5} is represented by the scatter of points about the one-to-one line. In the case of the bias correction application, the bias corrected prediction can be represented as:

$$\hat{y}_{bc} = f_{era5}(x_{cmip}),$$

which is equal to the actual ERA5 temperature plus an error term:

$$y_{era5} = \hat{y}_{bc} + e_{bc}.$$

The aim of bias correction is to reduce the error in existing climate model outputs, represented below by e_{cmip}:

$$y_{era5} = y_{cmip} + e_{cmip}.$$

So if e_{bc} is smaller than e_{cmip} then the bias correction is successful by this measure. As climate models are not designed to perfectly reproduce the historical time series of observations it is not possible to directly compare time series predictions as in Section 3.2 for performance evaluation, however the principle of reducing error relative to existing climate model performance still stands.

As time series predictions cannot be directly compared in this context, the bias correction method is instead evaluated in terms of characteristics of the distribution of temperatures produced. Following the steps outlined above, the initial performance of the CMIP6 model output is evaluated against the ERA5 temperature distribution as follows:

$$\int Y_{era5} - Y_{cmip} = E_{cmip},$$

where Y_{era5}, Y_{cmip} represent continuous variables over the temperature distribution. Similarly for evaluating the Ridge-ERA5 bias correction outputs:

$$\int Y_{era5} - Y_{bc} = E_{bc}.$$

The performance metric p is the calculated as the difference between these two errors:

$$p = E_{bc} - E_{cmip}.$$

This performance measure is calculated on a regional basis averaging over land grid cells in each Northern Hemisphere AR6 region and is shown in Figures 3.14 and 3.15.

To summarise these results, application of the Ridge-ERA5 bias correction method produces historical temperature distributions which match as well or better with the ERA5 distribution - compared with taking raw CMIP6 temperature anomalies - in many Northern Hemisphere regions. In general, application of Ridge-ERA5 results in distributions which are narrower, where in many regions the raw CMIP6 distributions over-estimate the density of extreme events. For example, in the Mediterranean (Figure 3.14h), West Central Asia (Figure 3.14j) and the Sahara (Figure 3.14i), combining the Ridge-ERA5 coefficients with predictor variables from CMIP6 models results in a historical distribution of temperatures showing much closer agreement with ERA5 than the CMIP6 temperature outputs themselves.

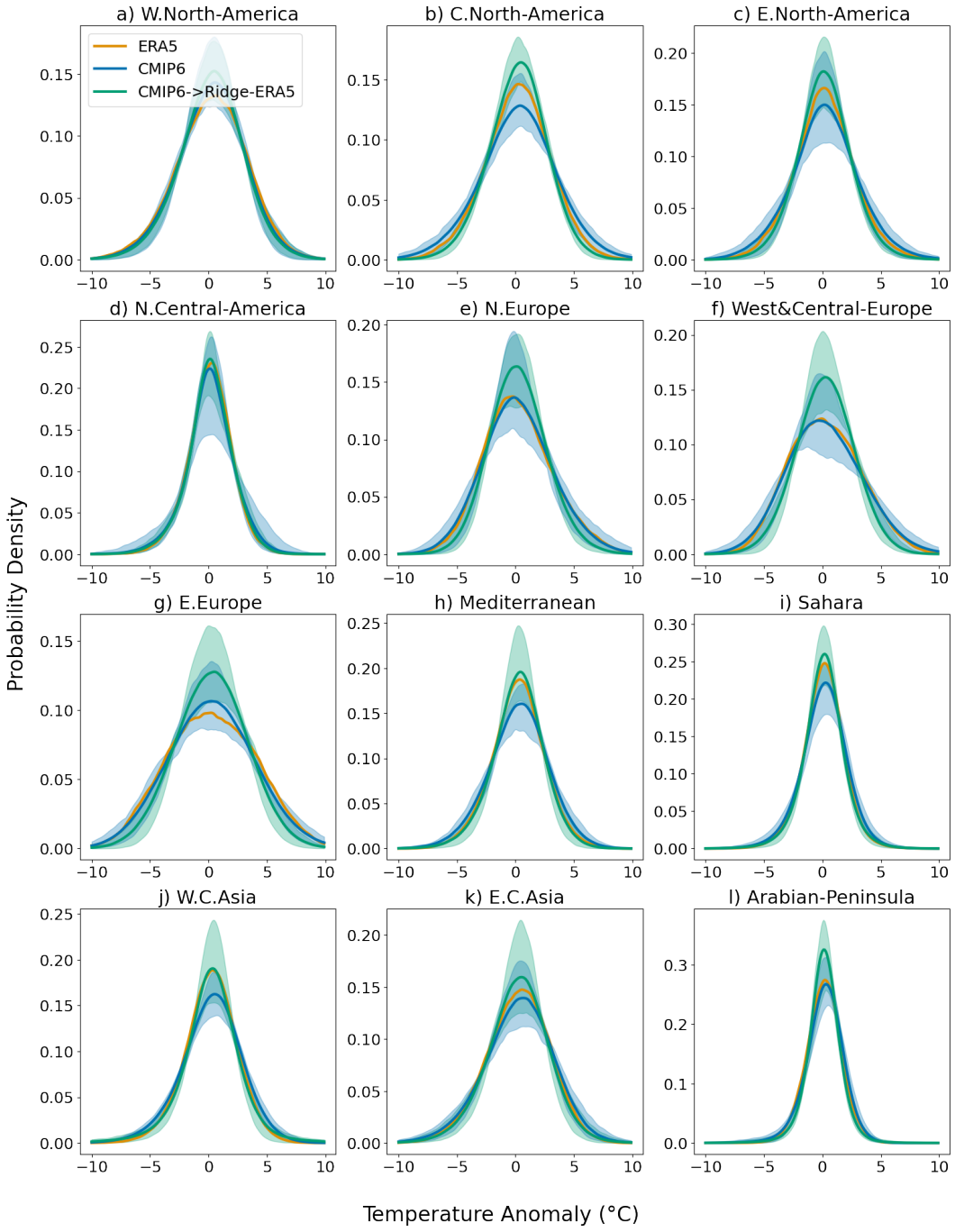

Figure 3.14: Kernel Density Estimators fitted to ERA5 data (orange line), CMIP6 mean (blue line) and minimum-maximum range (blue shading) and Ridge-ERA5 corrected temperatures from CMIP6 model input mean (green line) and minimum-maximum range (green shading) average over land grid cells in Northern Hemisphere AR6 regions.

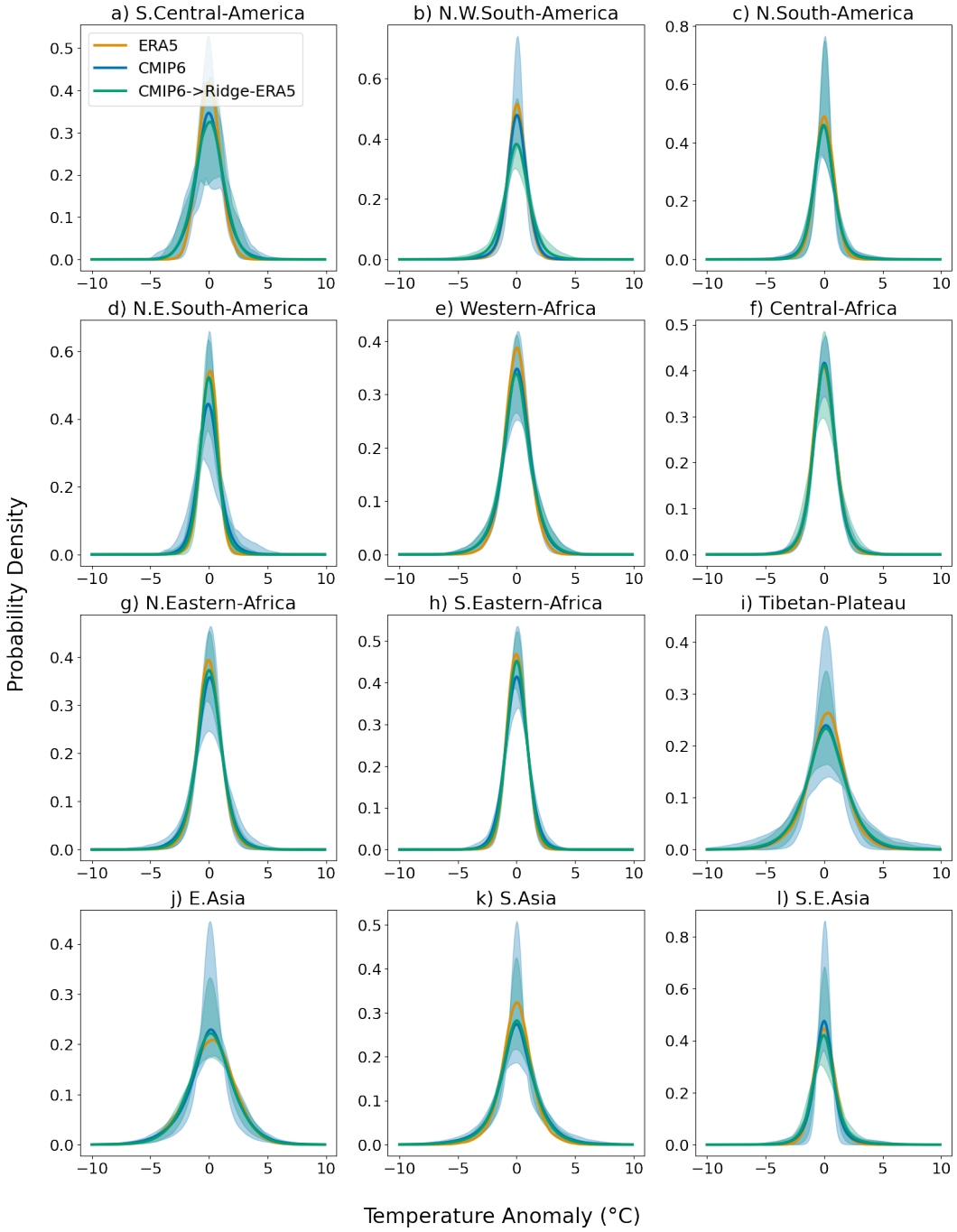

Figure 3.15: As in Figure 3.14 for remaining Northern Hemisphere AR6 regions.

The regions where distance from the ERA5 distribution is increased by this method correspond to regions of generally poorer performance of Ridge-ERA5 even when predicting ERA5 itself. For example, in the Northern, Central and Eastern European regions (Figure 3.14e, f and g respectively). These regions showed the greatest spread around the one-to-one line when historical Ridge-ERA5 test predictions were plotted against true ERA5 anomalies (Figure 3.10) and application of Ridge-ERA5 for bias correction in these regions results in an over-narrowing of the temperature anomaly distribution compared with taking raw CMIP6 anomalies. This pattern can also be seen in the high-latitude regions that are excluded based on the lack of suitable predictor variables in the model (see Section 3.2 for why these regions are excluded), regions where Ridge-ERA5 performance is generally lower.

This information is summarised in a different format in Figure 3.16 which indicates which individual models are corrected (or not) relative to ERA5 by application of the Ridge-ERA5 coefficients. This is calculated by comparing the area between the simulated temperature distribution curve (Ridge-ERA5 coefficients combined with CMIP6 predictors) and the ERA5 distribution with the corresponding area between the distribution of raw CMIP6 model anomalies and ERA5. Comparing the distributions using the area between the curves makes clear that the temperature distribution of most models in most regions is brought closer to the reanalysis distribution by applying Ridge-ERA5 (blue squares in Figure 3.16). Eastern, West & Central and Northern Europe stand out as regions where the method does not work as intended. Since the temperature anomalies are a simple function of coefficients and predictor anomalies, the source of this result must be either in the coefficients learnt from ERA5 in these regions or differences in the distribution of predictor variables between ERA5 and CMIP6. This is investigated in greater detail in the following subsection.

To place the difference metric scores in Figure 3.16 in context a variance scaling bias correction is also applied to historical CMIP6 temperature anomalies for comparison. These results are plotted in Figure 3.17. Across most regions, the Ridge-ERA5 bias correction performs competitively or even outperforms a traditional variance scaling approach with the previously highlighted European regions being the only outliers in this respect.

Compared with other methods that seek to map quantiles or apply a transformation to the mean and variance of climate model data for closer agreement with observations; the Ridge-ERA5 approach does not directly model the relationship between observed and modelled distributions. The correction is purely derived from the process-based relationships represented by Ridge-ERA5, which may differ from the relationships between temperature anomalies and drivers represented in climate models. These differences are analysed in the following subsection, comparing learnt coefficients to locate the sources of these differences in performance and understand why the method results in a correction in most regions but poorer agreement with ERA5 in others.

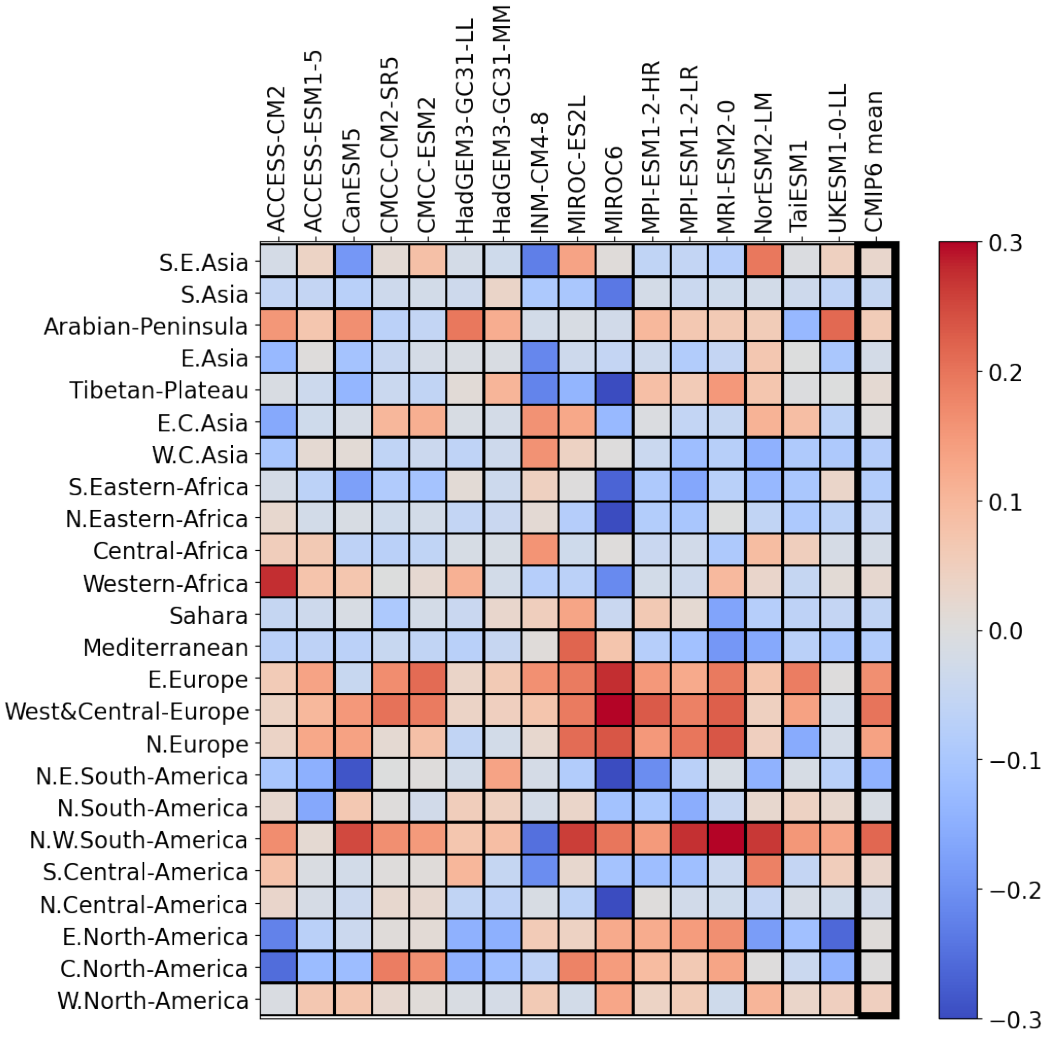

Figure 3.16: Difference between the distance from the ERA5 historical temperature distribution curve comparing Ridge-ERA5 temperature simulations with CMIP6. Negative/positive values (blue/red) indicate model and region combinations where applying Ridge-ERA5 brings the climate model distribution closer/further from the ERA5 curve.

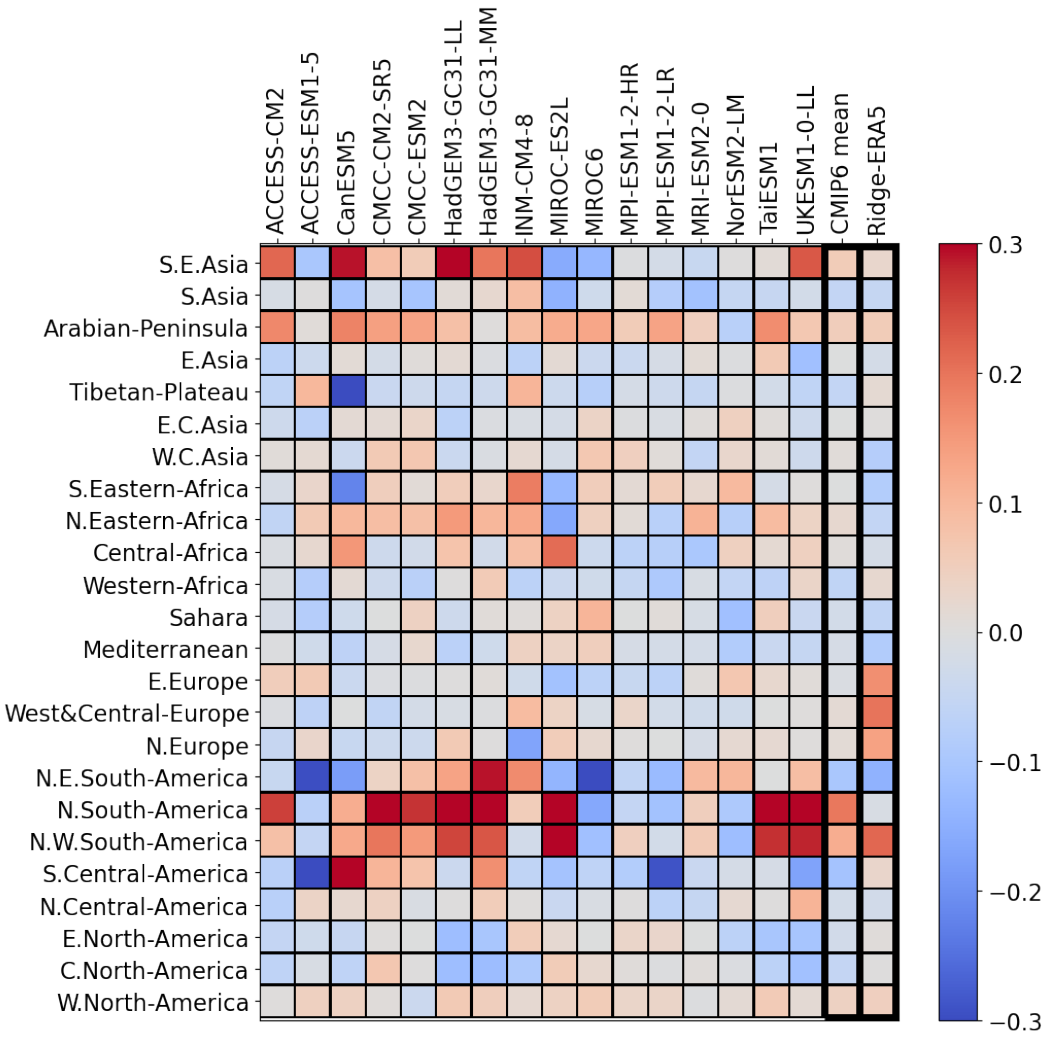

Figure 3.17: Difference between the distance from the ERA5 historical temperature distribution curve comparing variance scaled temperature anomalies with raw CMIP6 values. The final column shows the mean correction across the CMIP6 ensemble for the Ridge-ERA5 bias correction methods. Negative/positive values (blue/red) indicate model and region combinations where applying the correction brings the climate model distribution closer/further from the ERA5 curve.

3.5 Ridge-ERA5 for Climate Model Evaluation

Having investigated the predictive abilities of the Ridge-ERA5 model and the application of the model for bias correction of historical CMIP6 temperature anomalies, the method is now applied in a model evaluation context. By comparing coefficients learnt from ERA5 data with those learnt by Ridge from individual CMIP6 models, systematic differences between modelled relationships and observed patterns can be identified as well as performing model inter-comparisons. The results in this subsection highlight key differences in the simulation of dynamical variables and their relationship with temperature in models *vs.* reanalysis helping to explain the regional dependence of the bias correction performance.

In order to apply Ridge-ERA5 in a model evaluation context, a set of Ridge-CMIP models are trained as CMIP6 emulators using historical climate model data. These Ridge-CMIP emulators learn to predict daily temperature anomalies from the same set of predictor variables used for Ridge-ERA5, the only difference is that they learn from climate model data rather than reanalysis so that their learnt coefficients will represent the processes simulated by the corresponding CMIP6 model. To match the volume of training data with Ridge-ERA5, data covering 1979-2022 needs to be used. For the period of 1979-2014 data from the *historical* forcing scenario in CMIP6 is taken and for 2015-2022, data from future emissions scenario *SSP585* is used. Each Ridge-CMIP emulator is trained with the same predictor variable setup as for Ridge-ERA5 to predict the daily 2m temperature anomaly. And, also as for ERA5, all variables are measured as anomalies to a smoothed, mean seasonal cycle.

Each Ridge model, both Ridge-ERA5 and the Ridge-CMIP emulators, learns a set of coefficients which can be used to represent the relationship between each predictor variable and the target variable - the daily temperature anomaly. To use the Ridge-ERA5 method for climate model evaluation, the coefficients learnt by Ridge from ERA5 - which represent relationships between variables derived from observations - are compared with those learnt from CMIP6 data. The idea being that Ridge-CMIP emulators with coefficient maps that are more similar to those of Ridge-ERA5 may correspond to climate models which have more realistic representations of particular processes.

First, performance of the Ridge-CMIP emulators on historical CMIP6 data is checked to confirm that the Ridge-CMIP models offer a reasonable representation of the climate model they have learnt from. Figure 3.18 shows R^2 scores across the Northern Hemisphere obtained by using Ridge-CMIP emulators to predict on held-out test data from the corresponding CMIP6 model. These scores indicate that Ridge-CMIP models can simulate the variance in their corresponding CMIP6 model very well with the same patterns of lower performance over the oceans and at higher latitudes - regions which are excluded from this analysis.

Secondly, a metric of difference between coefficients learnt from ERA5 (Ridge-ERA5 coef-

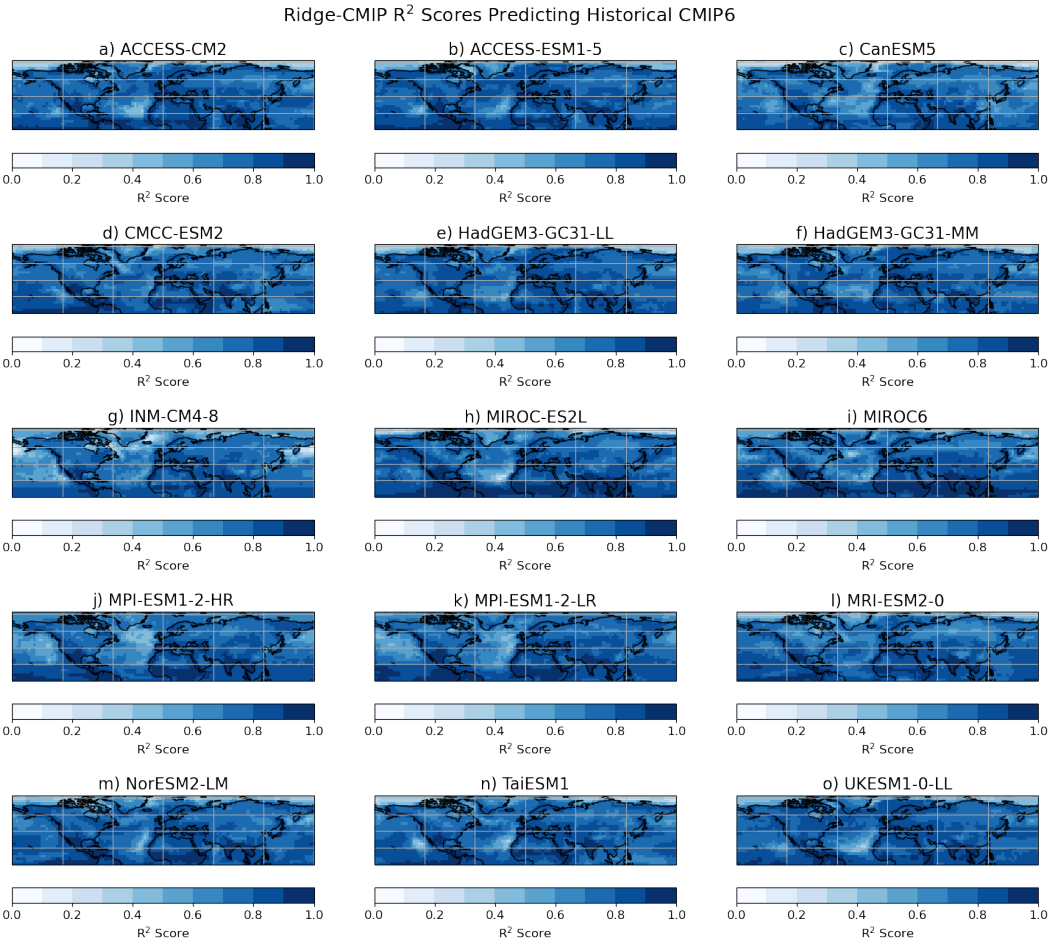

Figure 3.18: R^2 score for Ridge-CMIP models tested on historical CMIP6 data.

ficients) and those learnt from each CMIP6 model (Ridge-CMIP coefficients) for each climate predictor variable averaged over the domain size is plotted in Figure 3.19. A larger score here indicates greater difference between coefficients learnt from a particular CMIP6 model and those learnt from ERA5 and therefore potentially a model that simulates temperature drivers in a way that is less realistic. These difference maps produce unique 'fingerprints' for each

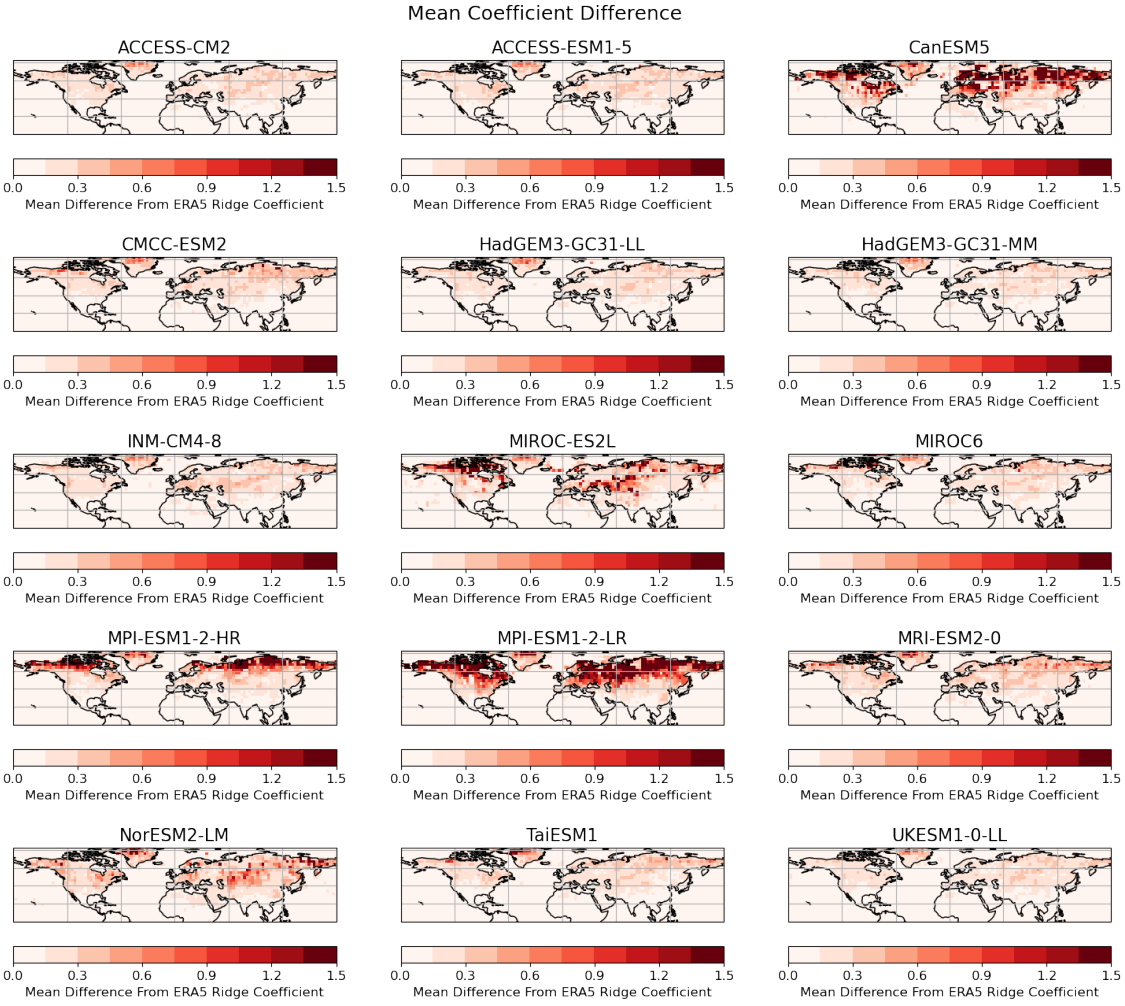

Figure 3.19: Mean coefficient difference between Ridge coefficients learnt from CMIP6 models and those learnt from ERA5.

climate model which give some indication of similarity (or lack thereof) with ERA5. Initial comparison with Figure 3.16 reveals some potential correlations between models and regions with larger magnitude coefficient differences and those that are made worse by the Ridge-ERA5 'bias correction' (European regions and higher latitudes). These maps are available on an variable by variable basis in Appendix A.1.

In order to study these patterns in more detail, coefficient difference is plotted on a

regional and model basis. Total magnitude of difference across all coefficients is plotted in Figure 3.20 for each CMIP model and AR6 region. Higher values here indicate that the coefficients learnt to represent relationships between temperature drivers and anomalies are generally more different between Ridge-ERA5 and the particular Ridge-CMIP model. Comparison of Figure 3.20 with Figure 3.16 now indicates some clear correlations between

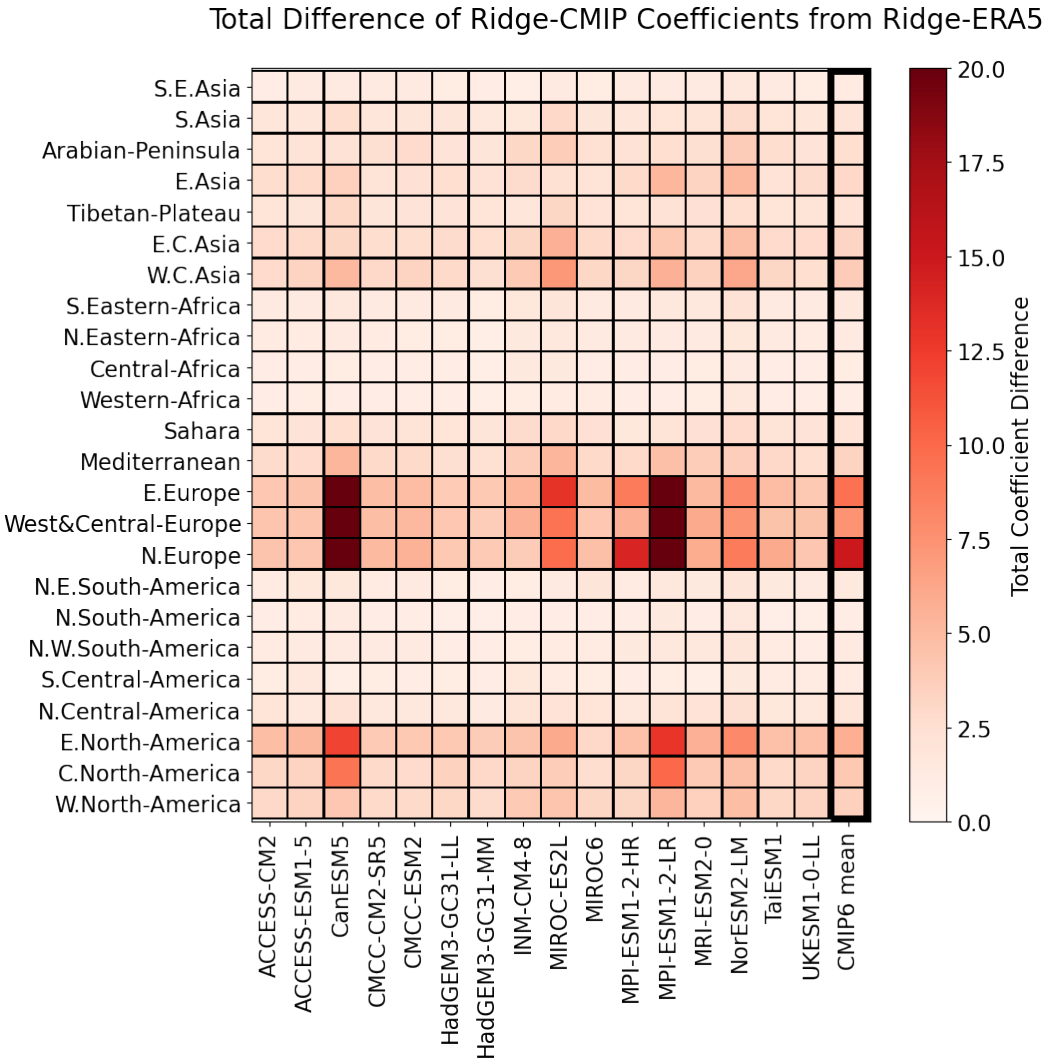

Figure 3.20: Total absolute differences between Ridge-CMIP emulators and Ridge-ERA5 coefficients averaged over Northern Hemisphere AR6 regions for each CMIP6 model.

coefficient difference and performance of the Ridge-ERA5 bias correction method. In regions where the bias correction fails (Northern, West & Central and Eastern Europe) there are larger magnitude differences between CMIP and ERA5 coefficients, particularly for certain models.

Inspection of the coefficients for individual models in the European regions where the bias correction fails reveals that the magnitude of (positive and negative) wind coefficients is larger for many CMIP6 models than for ERA5. The effects that this has on the overall contribution of wind variables to the final Ridge-CMIP temperature anomaly predictions is not as large as the total coefficient difference may imply as the magnitude of coefficients is over-estimated for some grid cells but under-estimated for others. So, while total difference magnitude is high, these effects are internally balanced out to some degree such that the overall effect is a slightly hotter contribution of wind anomalies to the temperature anomalies in Ridge-CMIP predictions compared with Ridge-ERA5 predictions. This over-prediction of temperature anomalies from the wind predictors in Ridge-CMIP models is compensated for by weaker soil moisture coefficients compared with Ridge-ERA5. This suggests that, in the climate models themselves, the strength of soil moisture feedbacks may be underestimated and dynamical associations with temperature are somewhat over-sensitive whilst patterns of dynamical associations with temperature are highly variable. This compensating effect is visible when the Ridge-ERA5 bias corrected temperature simulations are decomposed into SHAP value contributions (see Section 2.4.2) from each predictor variable, see Figure 3.21. Contributions from soil moisture variables (Figure 3.21a) to positive temperature anomalies are increased when Ridge-CMIP coefficients are swapped out for Ridge-ERA5 coefficients whilst contributions from u (Figure 3.21b) and v (Figure 3.21c) components of the 850hPa wind are decreased relative to Ridge-CMIP simulations.

The compensating effect between soil moisture and wind variables outlined above does not explain the narrowing of the temperature distribution that results from applying the Ridge-ERA5 bias correction method in the European AR6 regions since soil moisture contributions are scaled up at the same time as wind contributions are scaled down. The contribution of each predictor variable to the temperature anomaly, Δy, can be broken down into the contribution from the predictor variable, x, and the contribution from the coefficient, θ:

$$\Delta y = \theta x.$$

If, for a particular CMIP model, the coefficient $\theta_{cmip,p}$ for predictor variable x_p is very large (small) then an anomaly in x_p on a particular day will result in a relatively large (small) contribution to the overall temperature anomaly. If instead that coefficient is replaced with the corresponding coefficient from ERA5, $\theta_{era,p}$, which is much smaller (larger), then the same predictor variable x_p will contribute a relatively smaller (larger) temperature anomaly and the distribution of bias corrected temperature will be narrower (wider).

There are also large differences in the coefficients representing the relationship between daily near-surface relative humidity, with coefficients differing not only in magnitude but also in sign (positive or negative). When these coefficients are replaced with smaller, Ridge-ERA5

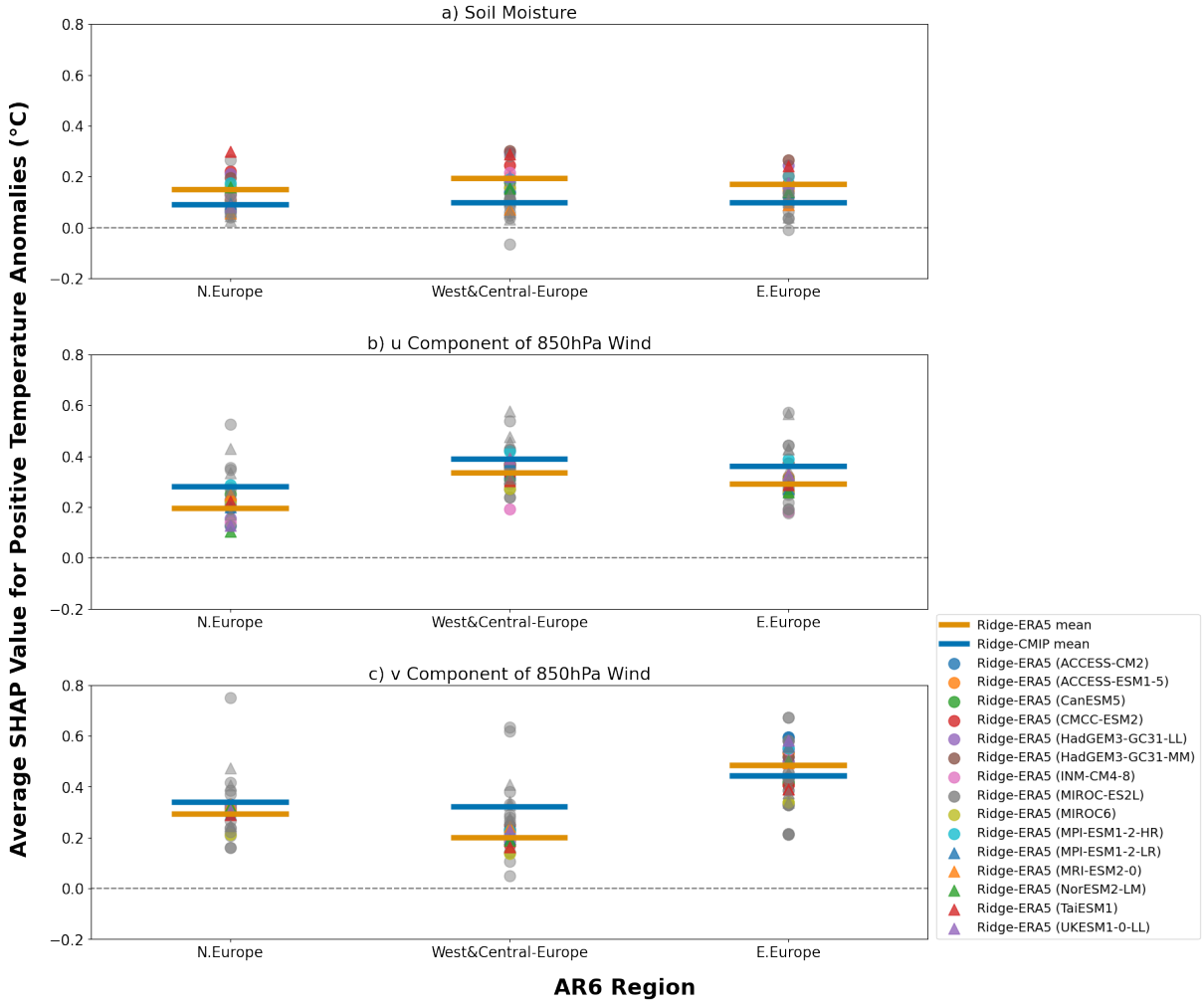

Figure 3.21: Mean SHAP value contributions to positive temperature anomaly predictions from Ridge-ERA5 (orange line) and Ridge-CMIP (blue line) across the CMIP6 models when provided with historical CMIP6 predictor variables. Individual SHAP value contributions for each CMIP6 model are also plotted for Ridge-CMIP (grey points) and bias corrected Ridge-ERA5 outputs (coloured points).

coefficients (during the bias correction procedure); smaller temperature anomalies are produced, see Figure 3.22, and there is a narrowing of the temperature anomaly distribution. For Ridge-ERA5, coefficients representing the relationship between near-surface relative hu-

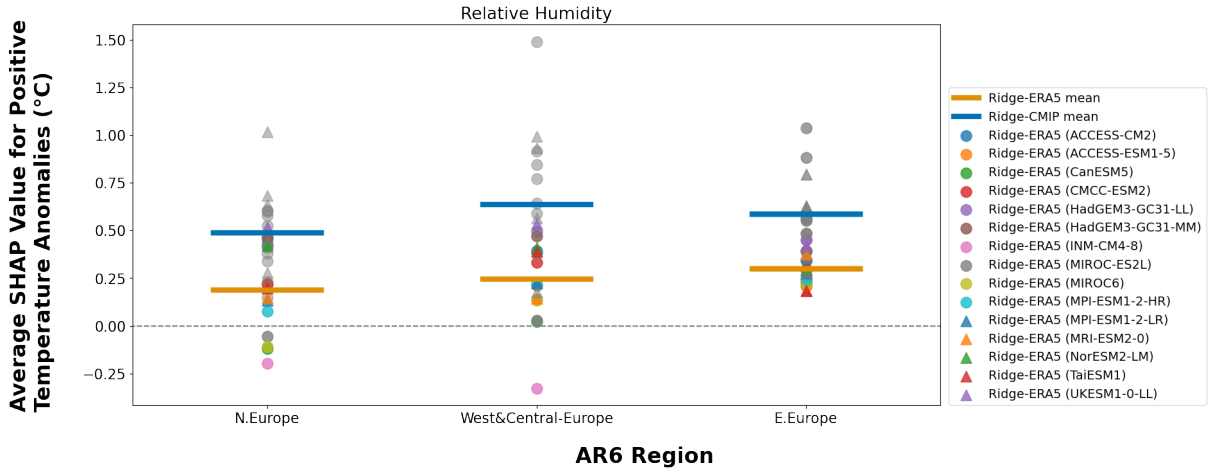

Figure 3.22: As in Figure 3.21 for near-surface relative humidity.

midity and temperature over Europe are generally positive and small in magnitude. For the Ridge-CMIP emulator models, these same coefficients tend to be larger in magnitude and are also mixed between positive and negative values. As such, negative relative humidity anomalies can still result in positive temperature anomalies. This can be seen in Figure 3.22 for Northern Europe where all contributions from Ridge-CMIP models (grey points) are positive but application of the Ridge-ERA5 bias correction results in negative SHAP values for some models. This implies a gap between the simulation of near-surface relative humidity in many of the models and ERA5 which could potentially have implications for soil moisture feedbacks and cloud formation processes in climate models (Liu et al., 2023). Altogether, for the bias correction method this implies that, for models which are 'different enough' from ERA5 in their representation of particular processes and/or variables which Ridge-ERA5 relies on for predictive skill, the bias correction method does not function as intended.

Returning to the application of Ridge-ERA5 in a model evaluation framework, these coefficient differences are symptomatic of underlying differences in modelled processes between CMIP6 and ERA5. Comparing these differences provides a tool for model ranking as well as highlighting simulation of individual processes and variables for closer inspection. In some regions, larger differences are also seen in the 850hPa specific humidity coefficients, particularly for CanESM5. Since monthly 850hPa specific humidity provides information about the background warming state, this may relate to the fact that CanESM5 is the highest sensitivity CMIP6 model. These differences in rate of background warming relative to ERA5

in the historical period have knock-on implications for future projections of such models. To investigate this further, difference metrics are plotted on a variable by variable basis against the end of century warming projected by each model under future scenario *SSP585*, see Figure 3.23. For the representation of some variables, *e.g.* precipitation rate (Figure 3.23e)

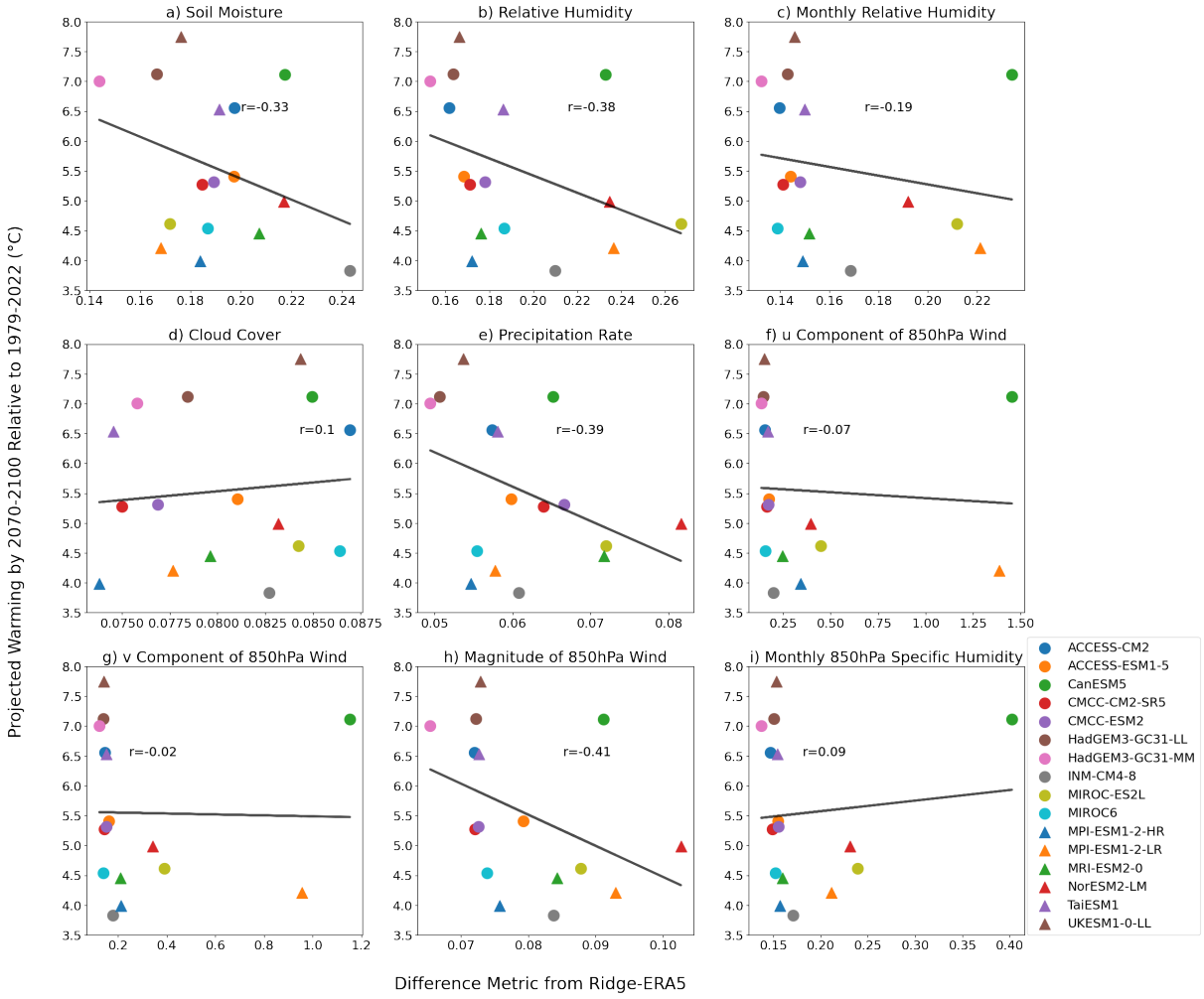

Figure 3.23: Difference metric between Ridge-CMIP and Ridge-ERA5 regression coefficients *vs.* projected end of century warming with linear fit to points (black line) and Pearson correlation coefficients (r).

and 850hPa wind magnitude (Figure 3.23h), an emergent pattern between historical model performance, judged against Ridge-ERA5, and the degree of future warming is potentially identifiable. For other variables, *e.g.* cloud cover (Figure 3.23d), no such link is visible. The possibility of applying Ridge-ERA5 to constrain uncertainty in future warming projections is explored in more detail in Chapter 4.

3.6 Conclusions

In this chapter, performance of the process-based ML model, Ridge-ERA5, is demonstrated on held-out test data from the ERA5 data set. Ridge-ERA5 is trained to predict the daily-mean temperature anomaly to a smoothed mean seasonal cycle using information from domains of 9 climate predictor variables; soil moisture, near-surface relative humidity, monthly mean of near-surface relative humidity, cloud cover, precipitation rate, u & v components of the 850hPa wind, magnitude of the 850hPa wind, and monthly-mean of 850hPa specific humidity. The monthly mean variables are intended to give information about the background state on top of which anomalies are superimposed. Wind variables give information about the local and regional dynamical state including potential indications of heat transport or the presence of blocking conditions. Other variables give information about potential exacerbating factors contributing to heat anomalies such as clear sky conditions or dry soils.

Summary statistics including R^2 scores of > 0.7 at most Northern Hemisphere land grid locations indicate that variance in the data set is well captured with notable decreases in performance over the oceans and some high latitude regions. Corresponding plots of mean absolute error tell a similar story of improved performance over land *vs.* ocean regions and also reveal a gradient of improved performance moving from higher latitudes towards the Tropics. Time series at individual locations show good day-to-day agreement between Ridge-ERA5 predictions and unseen summertime temperature anomalies, including the magnitude of warm anomalies during an extreme event such as the 2003 European heatwave. As expected, plotting mean absolute test prediction error against percentile indicates decreasing performance towards the tails of the temperature anomaly distribution although hot anomalies are better captured than the cold tail of the distribution. Plotting predictions *vs.* actual ERA5 temperature anomalies on a region by region basis gives a clearer picture of how this performance across the temperature distribution is laid out geographically. Patterns in these plots show excellent agreement with the one-to-one line although spread of points is greater in some regions than others. Performance appears particularly good in regions covering Northern and Western Africa as well as Central America whilst spread of points is larger over Eastern Europe and Siberia where key snow and albedo feedbacks are not represented by Ridge-ERA5.

A first potential application of Ridge-ERA5, bias correction, is demonstrated by applying Ridge-ERA5 coefficients to predictor variables from models in the CMIP6 archive. Kernel density estimators (KDEs) are fitted to the resulting temperature time series across the historical period (1979-2022) and compared with distributions of temperature anomalies from ERA5 and the CMIP6 models themselves. These distributions are averaged on a regional basis considering land grid cells only. Results indicate that this method can offer improved

agreement with ERA5 distributions in many regions in comparison with taking raw CMIP6 temperature anomalies. The Ridge-ERA5 bias correction method also performs competitively, and in some regions outperforms, a traditional variance scaling approach. These improvements are impressive considering that, unlike other bias correction methods, the ML model applied is not trained with the aim of bringing CMIP6 temperature anomaly distributions into closer agreement with those recorded in the reanalysis. The bias correction is a by-product of the process-based approach; applying coefficients representing observations-based relationships between drivers and temperature anomalies. Additionally, decomposing the bias corrected temperature anomaly predictions into contributions from each predictor variable reveals potential compensating errors between soil moisture feedbacks and dynamical patterns represented by the 850hPa wind variables in CMIP6 models.

The motivation behind this methodology is the idea that the Ridge-ERA5 model may be capturing certain influences on temperature more faithfully than the (much more complex) global climate model. Of course, by looking at performance scores of the Ridge-ERA5 model it is clear that this assumption is more likely to be true in some regions than others and therefore caution should be taken in applying Ridge-ERA5 as a blanket bias correction method. This is especially important when considering that the Ridge-ERA5 model is not trained with bias correction in mind whereas traditional bias correction approaches are formulated as a function which is designed to translate between raw climate model output and climate model output which agrees with a reference observational distribution. Here, the improvements that are seen in some regions comparing the "bias corrected" distribution with the raw CMIP6 outputs arise purely from the differences between relationships learnt by Ridge-ERA5 to represent temperature drivers from reanalysis data and the representation of such driving forces in the climate models themselves.

In regions where the method works well, for example the Mediterranean and West Central Asia, application of the Ridge-ERA5 coefficients results in improved agreement with ERA5 across the temperature distribution indicating that the method is resulting in a 'significant' improvement. In other regions, for example Eastern North America, while the new distribution agrees better with ERA5 than the raw CMIP6 output at some temperatures it also looks worse in others indicating that additional testing may be required to identify if this is a significant improvement or not. Further analysis over different time and spatial scales would give an indication of the robustness of the bias correction skill of Ridge-ERA5 as well as work to identify why the method may result in an improved agreement in some temperature ranges but worsening of the distribution in others. Additionally, alternative metrics of bias correction performance could also be considered for example comparing specific statistics of the output temperature distributions such as mean, variance, minimum, maximum or median. These scores could be considered on an average basis summarised over regions and

time periods or as time series and maps averaged over smaller spatial and temporal scales. The advantage of using these statistical measures is that there would be no need to fit a distribution to the temperature anomalies, scores could be calculated from the raw temperature outputs themselves.

The Ridge-ERA5 bias correction method is capable of altering both the mean and the variance of the distribution, and considers information from relevant dependent variables. Additionally, the ML method itself (Ridge regression) is much simpler to train and interpret than other approaches used previously in this field (*e.g.* GANs). However, there are regions where application of the Ridge-ERA5 coefficients results in increased distance from the historical reanalysis distribution. In parts of Europe, Northern North America and Siberia, application of the bias correction results in an over narrowing of the distribution meaning that the frequency of higher magnitude temperature anomalies is under-predicted. Differences in the relationship between CMIP6 near-surface relative humidity and temperature relative to the same quantities in ERA5 are identified as the root cause of this issue in Europe. These results suggest that a similarity threshold may need to be met in order for the Ridge-ERA5 method to confidently be applied in a bias correction context.

The second application for Ridge-ERA5 shown in this chapter is climate model evaluation. Having previously identified the potential to 'bias correct' existing climate model temperature anomalies by applying coefficients learnt from ERA5 to relevant driver variables, questions arise about where these differences between reanalysis and climate model data originate. By training Ridge models as emulators of individual CMIP6 models, corresponding coefficients representing the relationships between temperature drivers and anomalies can be compared on a location by location basis. Plotting differences between the coefficients learnt across the input domains of each climate predictor variable reveals distinctive model 'fingerprints' which give information about the level of deviation of the model from reanalysis. Plotting the coefficient difference metrics against projected future warming on a model by model basis reveals the possibility of a constraint on future warming based on historical climate model performance judged against Ridge-ERA5 and motivates the exploration of Ridge-ERA5 for constraining future uncertainty in Chapter 4.

4 Process-Based Machine Learning for Observational Constraints on Future Regional Warming Projections

The method used in the previous chapter to understand biases and deviations from observations in climate model output is not only useful for analysing historical temperature anomalies, but can also be applied to account for uncertainties in future projections of climate change. This chapter demonstrates application of the process-based ML model, Ridge-ERA5, for constraining uncertainty in projections of average future temperature change by the end of the century (2070-2100) compared with a historical reference period (1979-2022). The constraint is applied to data from the CMIP6 archive on a regional basis with a focus on Northern Hemisphere mid-latitudes. First, some background on the sources of uncertainties in climate change projections is provided (Section 4.1.1) followed by a description of traditional approaches to constraining those uncertainties covering: model weighting approaches (Section 4.1.2); emergent constraints (Section 4.1.5); detection and attribution based approaches (Section 4.1.3); and Bayesian statistical methods (Section 4.1.4). The question of climate invariance - do the relationships learnt from the present climate still hold under future climate change - is then considered. By using Ridge-CMIP emulator models trained on historical CMIP6 data, it is possible to test that the relationships learnt by the Ridge method from historical data can in principle be used to make predictions under future climate change (future emission scenario *SSP585*). The steps of the Ridge-ERA5 future constraint method are then derived (Nowack et al., 2023) before being applied on a regional basis to constrain average end-of-century warming. The final constraint has implications for climate model sensitivity with those models which predict the greatest degree of warming being excluded by the constraint in many Northern Hemisphere regions.

4.1 Introduction

Despite improvements in resolution and modelling of physical processes between subsequent generations of CMIP models, constraining uncertainty in future projections of climate change remains a key challenge. While models may agree on the direction of many climate change signals there are still discrepancies in the magnitude of the projected response. Uncertainty range changes between CMIP5 and CMIP6 are mixed with some variables showing little difference or even wider ranges (Li et al., 2021; Zhang and Chen, 2021; Cos et al., 2022), potentially related to more processes being added to models which then contribute additional uncertainties. For instance, the equilibrium climate sensitivity of several CMIP6 models exceeds the range of the previous generation of CMIP5 models (Tokarska et al., 2020). Lack

of certainty in end of century warming projections constitutes a major barrier to consensus agreement on climate change adaptation and mitigation policy. Various sources of uncertainty are explored in more detail in the next subsection followed by a literature review of existing methods for uncertainty constraint.

4.1.1 Sources of Uncertainty

Uncertainty in future projections of climate can be divided amongst three primary sources: *scenario uncertainty*; *model uncertainty*; and *internal variability* (Hawkins and Sutton, 2009). Projection of future temperatures requires estimates of forcings like anthropogenic greenhouse gas emissions and aerosols resulting in scenario uncertainty. Future values of these forcings will be determined by a complex combination of social and economic factors *e.g.* land use changes, population changes, migration, energy use, technological progress *etc.*, so assumptions have to be made in order to model future climate. In the CMIP6 archive, a range of possible future emissions scenarios are represented by Shared Socioeconomic Pathways (SSPs). The SSPs form part of ScenarioMIP (Eyring et al., 2016) designed to facilitate research of plausible future scenarios including quantification of uncertainties. The outcomes of the different SSPs in terms of differing 'levels' of climate change are designed to reflect different combinations of socio-economic challenges in relation to adaptation and mitigation (O'Neill et al., 2015). There are five SSPs;

- SSP1 - a gradual but definite shift towards sustainability, low challenge to both adaptation and mitigation.
- SSP2 - develops trends consistent with historical patterns, moderate challenges to adaptation and mitigation with variance between and within countries.
- SSP3 - increasing focus on domestic/regional issues with little international collaboration, high challenges to both adaptation and mitigation.
- SSP4 - increasing stratification and inequality both between and within countries, challenges for adaptation remain high, particularly for less developed populations.
- SSP5 - rapid innovation and technological progress powered by fossil fuels resulting in low challenges to adaptation but a high challenge to mitigation as a result of reliance on fossil fuels.

These scenarios and their varying approaches to climate change adaptation and mitigation result in an uncertainty range for projected future warming. To illustrate this, annually

and globally averaged temperature anomalies are plotted under each SSP for CMIP6 model UKESM1-0-LL in Figure 4.1.

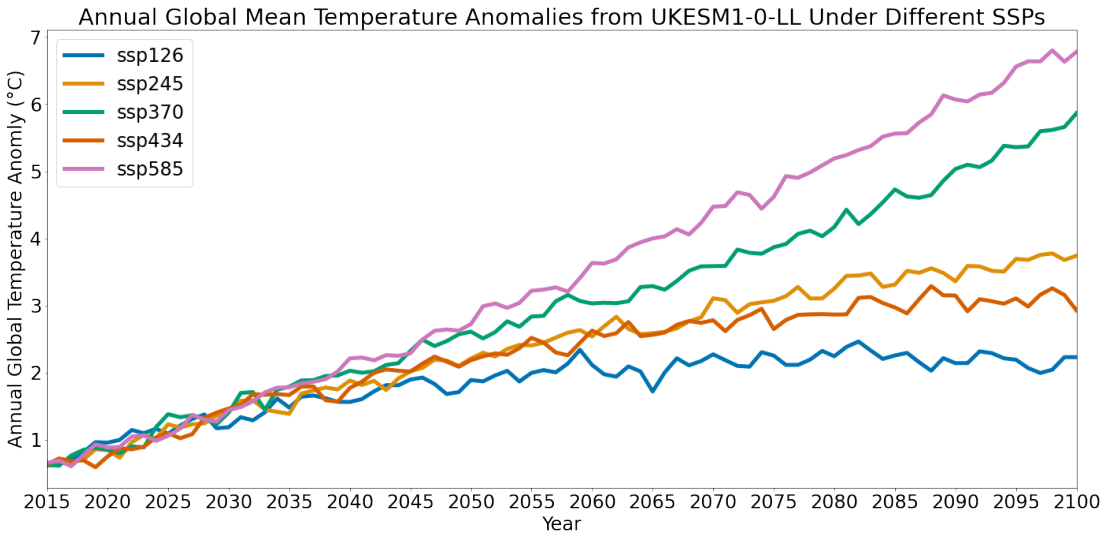

Figure 4.1: Annually and globally averaged temperature anomalies projected by UKESM1-0-LL under *SSP1*, *SSP2*, *SSP3*, *SSP4* and *SSP5* (colors as labelled) from 2015 to 2100.

Model uncertainty refers to the fact that different climate models will produce different projections of, for example, temperature even when provided with the same forcing scenarios. These variations arise from: difference in model structure; formulations of parameterisations (*e.g.* clouds, convection, aerosols); the processes that are included (*e.g* the carbon cycle, atmospheric chemistry, biogeochemistry); as well as model resolution (Hawkins and Sutton, 2009; Foley, 2010; Knutti and Sedláček, 2012). There are also other sources of uncertainty, for example model inadequacy (Sansom et al., 2020): gaps exist between the simulation of processes in climate models and our observations of those processes in the current climate or simply processes which are missing altogether from climate models. Parameter uncertainty can be analysed through Perturbed Physics Ensembles (PPEs) where the same simulations are repeated with the same model but differing values of parameters (*e.g.* Vellinga and Wu, 2008; Jackson and Vellinga, 2013). Inter-model uncertainty is illustrated in Figure 4.2, where annually averaged temperature anomalies from *SSP585* are plotted for five models from the CMIP6 archive.

A third source of uncertainty is internal variability or fluctuations in climate variables as a result of processes that are internal to the climate system *e.g.* El Niño Southern Oscillation (Lucas-Picher et al., 2008; Flato et al., 2013; Phillips et al., 2014). This natural variability can be thought of as a sampling uncertainty - the Earth's climate system is chaotic in na-

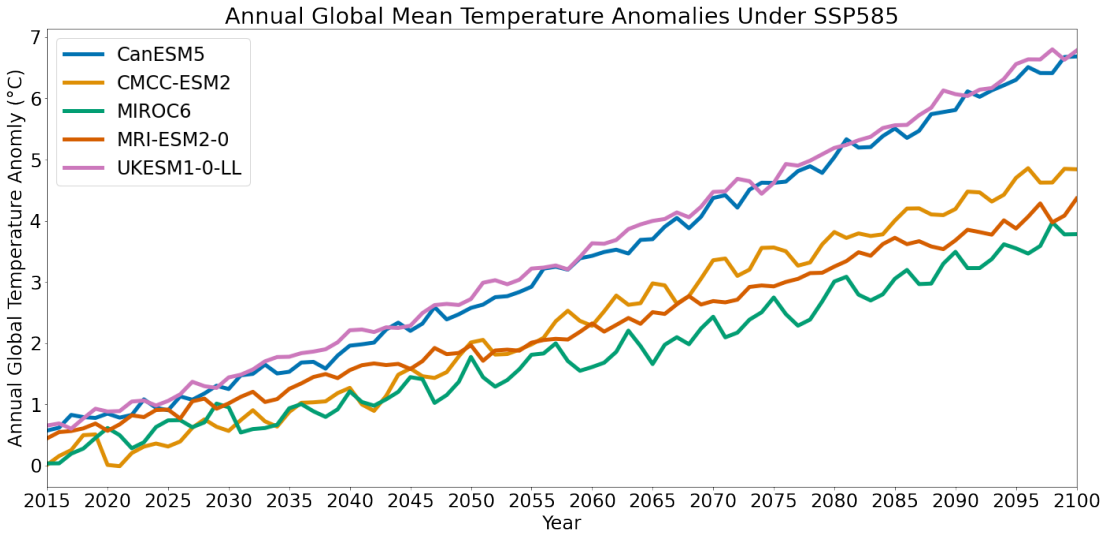

Figure 4.2: Annually and globally averaged temperature anomalies projected by CanESM5 (blue), CMCC-ESM2 (orange), MIROC6 (green), MRI-ESM2-0 (red) and UKESM1-0-LL (purple) under *SSP585* relative to a 1979-2022 reference period

ture and only one realisation of possible conditions is observed given the system's current state. The same climate model can produce different trajectories from slightly different initial dynamical conditions. Because of this, there will be variability between different ensemble members of the same model which is referred to as internal variability uncertainty. Internal variability can be sampled by running initial condition large ensembles, where each ensemble member receives slightly perturbed initial conditions but the same forcing scenario (Deser et al., 2012; Hawkins et al., 2016; Bengtsson and Hodges, 2019). Internal variability requires most consideration over the short-term whereas, on longer timescales, where global warming becomes more prominent, scenario and model uncertainty will start to dominate (Räisänen, 2001). Scenario uncertainty is mitigated by defining specific future conditions under which warming will be simulated, *e.g.* the SSPs in CMIP6, this leaves model uncertainty as the dominating factor to be quantified and constrained.

In the following subsections, previous approaches to constraining uncertainties in global and regional climate change projections are outlined. Specifically, these include a set of observational constraints methods such as model skill weighting (*e.g.* REA, ClimWIP), Bayesian statistical approaches, methods motivated by detection and attribution techniques and finally emergent constraint frameworks.

4.1.2 Model Weighting Approaches

A simple step to address model uncertainty in future projections is to combine projections across multiple models. In its most basic form, models are equally weighted with standard deviations or ranges representing uncertainty spread around the mean - the so called one-model-one-vote approach (see for example the IPCC Assessment Reports). Although equally weighted, multi-model means on average improve projection reliability (Weigel et al., 2010), there are cases where performance can clearly be further optimised by different weighting schemes (Knutti et al., 2017). Alternatives to a 'model democracy' approach (Knutti, 2010) require consideration of historical model performance measures and model interdependence to identify models which may be more (or less) fit for purpose when modelling a particular variable or region (Brunner et al., 2019) and to prevent over-weighting individual modelling groups or model components.

Ensembles of climate models like CMIP5 or CMIP6 are not necessarily compiled with model independence in mind. These so-called, 'ensembles of opportunity' (Tebaldi and Knutti, 2007) are unlikely to reflect a comprehensive range of possible model uncertainty with a complex entanglement of model inter-dependencies and parameter choices for a given ensemble member as well as including multiple versions of the same or very similar models which differ only in resolutions or single components (Lorenz et al., 2018). This has motivated efforts to explore alternative approaches to model weighting where models deemed to have a lower projection confidence are down-weighted and contribute less to the overall prediction. Projection confidence can be quantified by a variety of metrics from poor agreement with observations for historical data to representation of specific physical mechanisms and independence from other models.

Using the Reliability Ensemble Averaging (REA) method (Giorgi and Mearns, 2002) models are weighted by a reliability factor which quantifies both performance in present day climate and convergence of simulated future changes. Model performance is measured as the bias compared to the observed mean in the present day period, and convergence is calculated on an iterative basis as the distance of an individual model from the REA average change. These calculations are performed on a variable by variable basis, *i.e.* to quantify uncertainty in temperature, only bias and convergence in temperature are considered.

Alternatively, Climate Model Weighting by Independence and Performance (ClimWIP) aggregates information from several variables relevant to the target prediction and incorporates information about both model performance and model dependence (Knutti et al., 2017 building on Sanderson et al., 2015*a*,*b*). Distances of a model from observations and other models are quantified by, for example, root mean square differences (Knutti et al., 2017; Lorenz et al., 2018; Brunner et al., 2019). The diagnostics which contribute to the distance

quantification may cover different variables, regions, seasons and measure distance between means or variances (Brunner et al., 2019). Diagnostic variables are usually identified with physical links to the target variable in mind (Knutti et al., 2017) while only those which provide additional independent information to those already selected are kept (Brunner et al., 2019). This can be achieved by, for example, measuring correlation between diagnostic and target variables as well as correlation between the diagnostic variables themselves (Lorenz et al., 2018) in order to choose variables that are as independent from each other as possible. The method can also be extended to consider multiple ensemble members from the same model by adjusting the weighting calculations taking into account that each ensemble member represents an equally likely outcome of the climate system as simulated by a particular model (Merrifield et al., 2020; Brunner et al., 2020).

Although on the surface model weighting approaches appear quite intuitive, any model weighting approach will require making critical yet subjective decisions (Knutti et al., 2017) about the performance or independence metrics to rely on, the relevant quantities to measure these metrics with and the translation of these reliability measures into a weighting scheme.

4.1.3 Detection and Attribution Based Methods

Fingerprinting techniques from detection and attribution applications can be used to quantify uncertainty in future climate change (Brunner et al., 2020). This detection-attribution approach has commonly been applied to constrain global mean surface temperature where very clear linear relationships between 20^{th} and 21^{st} century warming have been identified for particular models and warming scenarios (*e.g.* Stott et al., 2000). This subsection introduces fingerprinting techniques for detection and attribution before outlining the steps of this method for uncertainty constraint (Allen and Tett, 1999; Stott et al., 2000; Stott and Kettleborough, 2002; Kettleborough et al., 2007).

Detection refers to the process of identifying a statistically significant change in the climate which is unlikely to have resulted simply from internal variability (IPCC, 2021). Patterns of change in response to an external forcing also referred to as 'fingerprints' (*e.g.* Hasselmann, 1997; Allen and Stott, 2003) can be quantified using climate model simulations. Ease of detection is affected by the signal to noise ratio - if the signal is large in comparison to natural variability it is more easily detected. As such, careful consideration of the relevant time and spatial scales is required to filter out internal variability as far as possible. Detection alone does not imply the change occurred as a result of an assumed cause, only that the signal is unlikely to have occurred as a result of natural variability. Attribution on the other hand refers to the process of identifying the most likely cause of a detected change within a defined confidence interval (IPCC, 2021), dividing observed climate changes into components

explainable by external forcings *e.g.* changes in greenhouse gas emissions and changes in solar radiation (Hegerl and Zwiers, 2011).

Simple detection and attribution approaches which use both climate model data and observations involve comparison of climate models driven by natural forcings, anthropogenic forcings and both natural and anthropogenic forcings with the observational temperature record to identify which forcings are required to reproduce historical warming (Hegerl and Zwiers, 2011). Fingerprinting techniques take these methods a step further and consider both spatial and temporal patterns of climate change as well as separating the comparison of different forced responses. Early applications of fingerprinting analysed how the correlation of spatial patterns - or fingerprints - generated from climate models under different forcings with observations evolved during the 20^{th} century (Santer et al., 1996). Underlying fingerprint-based detection-attribution methods is the assumption that an observed climate change signal, Y, can be represented as a linear combinations of forcings, X_i, (e.g. changes in greenhouse gas emissions, aerosol, ozone *etc.*) and internal variability, ϵ:

$$Y = \sum_{i=1}^{N} w_i X_i + \epsilon, \tag{37}$$

where w_i represents the scaling factor associated with the i^{th} forcing (Allen and Stott, 2003). An optimal least squares solution for the scaling factors, w_i, can be calculated using a maximum likelihood estimator (Hasselmann, 1997). If the uncertainty range for the scaling factor, w_i, associated with the i^{th} forcing, X_i, does not include 0 then the forcing is likely present in the observations *i.e.* a signal is detected. If, additionally, the uncertainty range for the value of W_j does not include 1, then the model response to forcing X_j likely requires re-scaling in order to be compatible with observations: for a scaling factor less (more) than one it is implied that the response is over- (under-) estimated so the models' pattern of response would need to be scaled down (up) for consistency with observations (Knutson, 2017). Later applications of this method allow the fingerprint to vary both spatially and temporally (Hegerl and Zwiers, 2011).

The detection and attribution methods described above can be applied to constrain uncertainties. In the context of constraining uncertainty, to constrain a variable Y, the simulated spatial and temporal pattern of the response of Y to a particular external forcing is compared with observations using optimal fingerprinting (Hasselmann, 1997; Allen and Tett, 1999). The basic assumption of this approach is that a climate model which over- (or under-) estimates the response of Y in the present-day climate will continue to over- (or under-) estimate that response in the future to a similar extent (Stott et al., 2000; Stott and Kettleborough, 2002). Uncertainty in the simulated historical response of Y is projected into the future using a transfer function which quantifies the relationship between the historical and

future responses (Kettleborough et al., 2007). Finally, measures of internal variability and the error associated with assuming a linear transfer function are accounted for to construct the overall uncertainty range (Kettleborough et al., 2007).

The primary drawback to the detection and attribution approach to uncertainty constraint is that the method can only be applied to variables for which the forced signal has already emerged and can be detected in observations (Brunner et al., 2020). This limits applicability where observations are sparse or the target region is not large enough for a signal to emerge from the noise of internal variability. Also, where the model pattern of the forced response deviates significantly from that which has been observed, it may not be clear if the response is forced or a result of internal variability. Additionally, the method makes the assumption that the pattern of the response will remain the same in the future, but it will actually change, simply because of changes in climate state (*e.g.* Jones et al., 2013; Ribes and Terray, 2013; Ribes et al., 2017)

4.1.4 Bayesian Statistical Approaches

Bayesian approaches to uncertainty constraint follow the basic steps of Bayesian statistics: a *prior* is constructed, then updated with information to form a *posterior* from which final predictions are sampled. In the context of an observational constraint, a prior constructed from historical climate model simulations is conditioned on available historical observations to remove future climate trajectories that are not compatible with historically observed climate change.

One such Bayesian approach referred to as UKCP09 (Murphy et al., 2009) and updated by UKCP18 (Murphy et al., 2018) makes use of a perturbed physics ensemble (PPE) of a single model where each ensemble member has a different set of values for the model parameters. The prior is constructed as a probability density function over this ensemble for the variable of interest (Murphy et al., 2018). The prior can be conditioned on observations by weighting each model variant by their likelihood relative to the observed climate (Sexton et al., 2012). There are some limitations to the UKCP approach, for instance the decisions about which parameters to be perturbed and the range of perturbation are naturally somewhat subjective, as is the underlying climate model choice.

A second, more recent statistical method (Ribes et al., 2021; Qasmi and Ribes, 2022; Ribes et al., 2022) referred to as HistC (from Brunner et al., 2020), constrains 21^{st} century temperature projections from CMIP6 using observations in a Bayesian framework. Historical forced responses of each CMIP6 model are estimated using a Generalised Additive Model (GAM) where the response to natural climate change is calculated from an energy balance model and the anthropogenic forcing is assumed to be smooth in time (Ribes et al., 2020).

The model-based prior is constructed as a normal distribution over an ensemble of these forced responses with mean and variance calculated over the ensemble. A historical, observational constraint is applied to the prior to exclude trajectories which are incompatible with observations once internal variability has been accounted for and form the posterior. This means that where a signal is not detectable in the observations, the posterior remains the same as the prior and no constraint is applied (Brunner et al., 2020).

4.1.5 Emergent Constraint Framework

An alternative approach to model weighting is the emergent constraint framework. Derived from physically explainable relationships between model simulations of a variable X in the current climate and projections of a different variable Y in the future climate, observational constraints on historical variable X are translated into a constraint on future projections of Y (Hall and Qu, 2006; Hall et al., 2019). This is illustrated in Figure 4.3 using artificially generated data. The term 'emergent' refers to the fact that the relationship between X and Y only *emerges* from the aggregation of information from multiple climate models across an ensemble. Although the spread in values of a historical variable X and future projections of second variable Y may be large across the models, what is important for the emergent constraint is that the relationship between them is distinct. Or mathematically speaking:

$$Y = f(X) + \epsilon, \tag{38}$$

where ϵ is small. Development of emergent constraints only becomes possible with the existence of organised large multi-model ensemble projects (Hall et al., 2019), for example CMIP3/5/6, which provide historical and future experiments that are uniform across models.

The earliest reported emergent constraint was for the snow-albedo feedback (Hall and Qu, 2006) where the correlation between feedback strength in the seasonal cycle and feedback strength under climate change enabled the latter to be constrained using historical reanalysis data. Since then, the method has been applied across the climate system from the hydrological cycle (*e.g.* O'Gorman, 2012; Li et al., 2017) to the carbon cycle (*e.g.* Cox et al., 2013; Wenzel et al., 2014) and equilibrium climate sensitivity (*e.g.* Brient and Schneider, 2016; Cox et al., 2018).

Emergent constraints represent a powerful tool for calculating observations-based constraints on future projections of key climate variables, however there are limitations to such an approach. The uncertainty that an emergent constraint can quantify is limited by the uncertainty range that is simulated by the climate model ensemble (Hall et al., 2019). For instance, systematic biases in a particular variable or process across the ensemble cannot

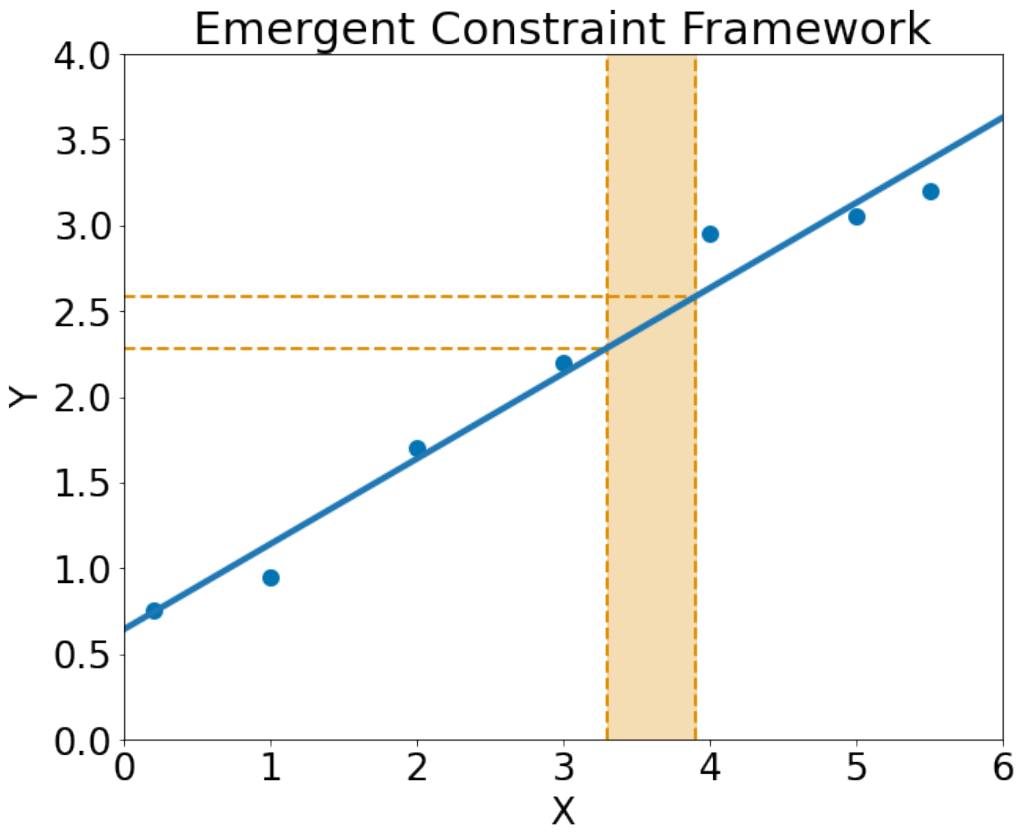

Figure 4.3: Illustration of the emergent constraint framework; an emergent relationship between a historical variable (X) and future projection of another variable (Y) is identified in climate models (blue circles). The range of present day observations of variable X (orange shading) can be translated into a constraint on future projections of Y (orange dashed lines) using the emergent relationship (blue line).

be corrected with this method. There may also be concerns about the explainability of the emergent relationship (Caldwell et al., 2018; Ferguglia et al., 2023) with the possibility that some such relationships emerge purely by chance (Caldwell et al., 2014) as a result of the relatively small number of truly independent climate models compared with the number of available output variables. Emergent relationships may provide over-confident constraints on future climate change signals given the usual assumption that the emergent relationship itself is exchangeable between the model and real worlds, and across ensembles of models (Sanderson et al., 2021). Attempts have been made to account for this by quantifying the likelihood that each model differs from reality (Williamson and Sansom, 2019).

4.2 Deriving the Ridge-ERA5 Future Constraint

As introduced in Chapter 2, the basis of this method for future constraint is the process-based ML model trained on reanalysis data, Ridge-ERA5. At each Northern Hemisphere grid cell, Ridge-ERA5 is trained on historical data from the ERA5 reanalysis data set covering the time period 1979-2022 for June, July, August (JJA) to predict the daily-mean 2m temperature anomaly relative to a smoothed, mean seasonal cycle (see Chapter 2 for further details). The basis of the observational constraint is to apply the coefficients learnt by Ridge-ERA5 from reanalysis to existing climate model output in order to constrain uncertainty in the relationships between temperature and driving factors (Nowack et al., 2023). To produce the observational constraint, it is necessary to combine Ridge-ERA5 coefficients with predictor variables from a future scenario in the CMIP6 archive, here analysis is focused on the most extreme future warming scenario, *SSP585* (see Section 4.1.1), which represents the most challenging extrapolation case for Ridge-ERA5.

Figure 4.4 outlines the processes involved in constructing the constraint which will be described in more detail throughout this and the following subsections. First, two types of Ridge models are defined, Ridge-CMIP emulators (Figure 4.4a) which learn from historical climate model data and Ridge-ERA5 (Figure 4.4b) which is trained on historical reanalysis data. Ridge-ERA5 is intended to represent as closely as possible the 'true' relationships observed in reanalysis while the Ridge-CMIP emulators recreate the historical (1979-2022) climate of individual climate models. Ridge-CMIP emulators can be used to make future predictions (Figure 4.4b) of temperature anomalies by providing them with inputs from future warming scenario *SSP585*. This allows climate invariance of the Ridge-CMIP models to be tested since the projections of their future temperature anomalies already exist for comparison. Of course, future values of predictor variables are not possible for ERA5, so in order to use Ridge-ERA5 for future predictions, inputs also need to be taken from CMIP6 data.

CMIP6 variables can be combined with Ridge-ERA5 coefficients by de-biasing and normalising relative to the historical ERA5 variance. The resulting future temperature predictions are composed of CMIP6 predictor variables conditioned on observations-based relationships represented by the Ridge-ERA5 coefficients (Figure 4.4d). These predictions form the basis of the observational constraint.

4.2.1 Climate Invariance

Before applying Ridge-ERA5 to future inputs from CMIP6, the 'climate invariance' of the relationships that Ridge is able to learn from historical data is tested. Do the relationships that Ridge learns from historical data still hold predictive power in the future under significant climate change? In order to do this, a set of 15 Ridge-CMIP model emulators are trained on historical data from the corresponding CMIP6 model, using years 1979-2022 to match the volume of training data used for Ridge-ERA5. In CMIP6, the historical forcing scenario runs from 1850 to 2014 so the remaining years (2015-2022) are filled in with data from *SSP585* (*e.g.* Sippel et al., 2020). Inputs from future scenario *SSP585* from 2023 to 2100 taken from the matching CMIP6 model are then provided to each Ridge-CMIP emulator. By comparing the temperature outputs from Ridge-CMIP with the actual future temperature projections from CMIP6 the climate invariance of the relationships that it is possible to learn can be assessed.

In Chapter 3 performance of the Ridge-CMIP emulators was demonstrated on test data (not seen during training) during the historical period. Here, to quantify the performance of the Ridge-CMIP emulator models under extrapolation conditions (future warming relative to the mean historical seasonal cycle), the change in performance over the course of the 21st century is calculated. The Ridge-CMIP models are applied to predict daily temperature anomalies under future emissions scenario *SSP585* from 2023 onwards in JJA, and the change in mean absolute error, in comparison with predictions during the historical period 1979-2022 is plotted in Figure 4.5. While some models show some decay in performance by the end of the century, most models exhibit a change in error of less than 0.5^oC by 2100 even for out of sample temperature anomalies of up 10^oC which far exceed the historically observed range.

The performance of Ridge-ERA5 in relation to some of these large, out of sample temperature anomalies is demonstrated in Figure 4.6 where Ridge-CMIP predictions for UKESM1-0-LL are compared with actual UKESM1-0-LL projections in 2095. Even for a Ridge-CMIP model which exhibits some of the greatest performance decay under future warming, performance of the Ridge model is still good in locations where historical Ridge performance was also good. For example, in Figure 4.6a and b, Ridge-CMIP shows clear predictive skill even for temperature anomalies far exceeding 10^oC. Locations where Ridge struggles are typically

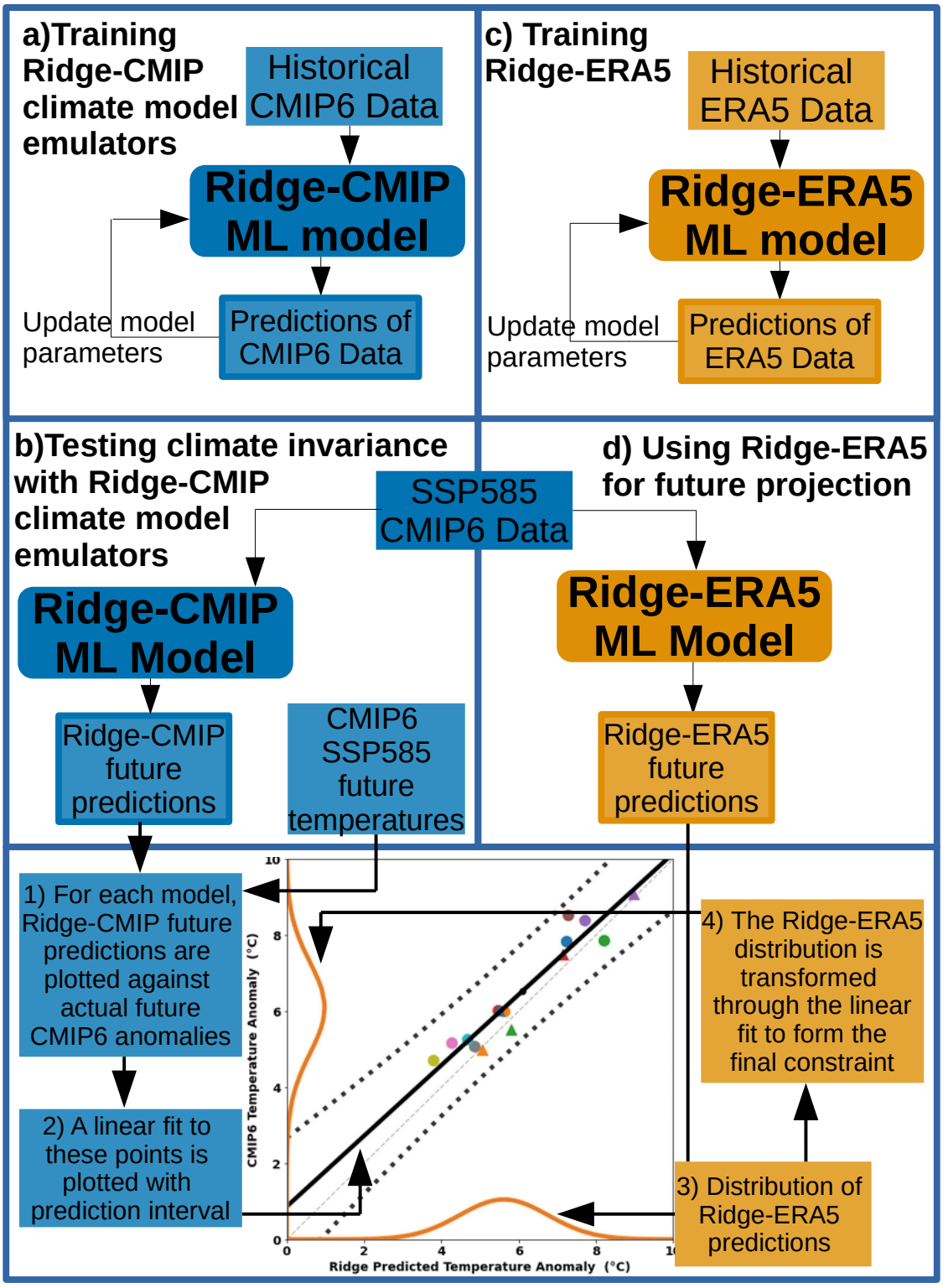

Figure 4.4: Illustration of the Ridge-ERA5 observational constraint.

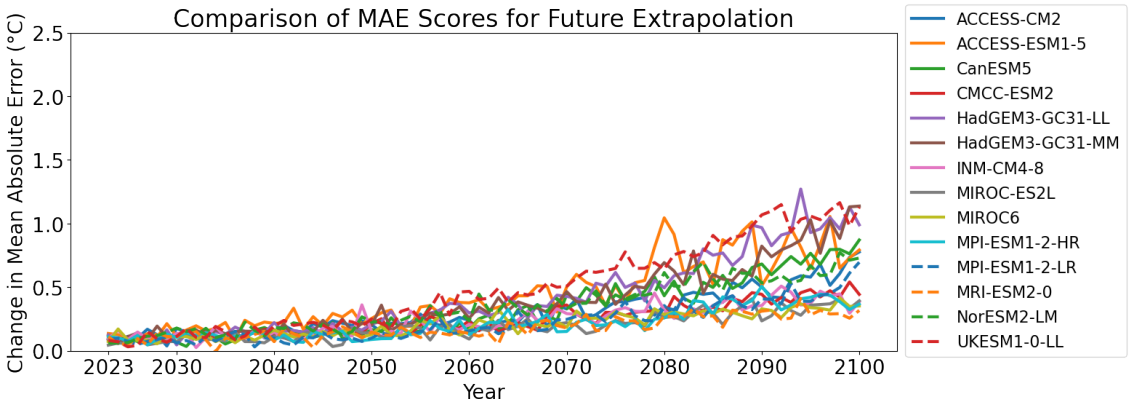

Figure 4.5: Change in Mean Absolute Error (MAE) for predictions under future emission scenario *SSP585* compared with the historical period (1979-2022) calculated each year over JJA for each model and averaged across the test locations highlighted in Figure 2.9.

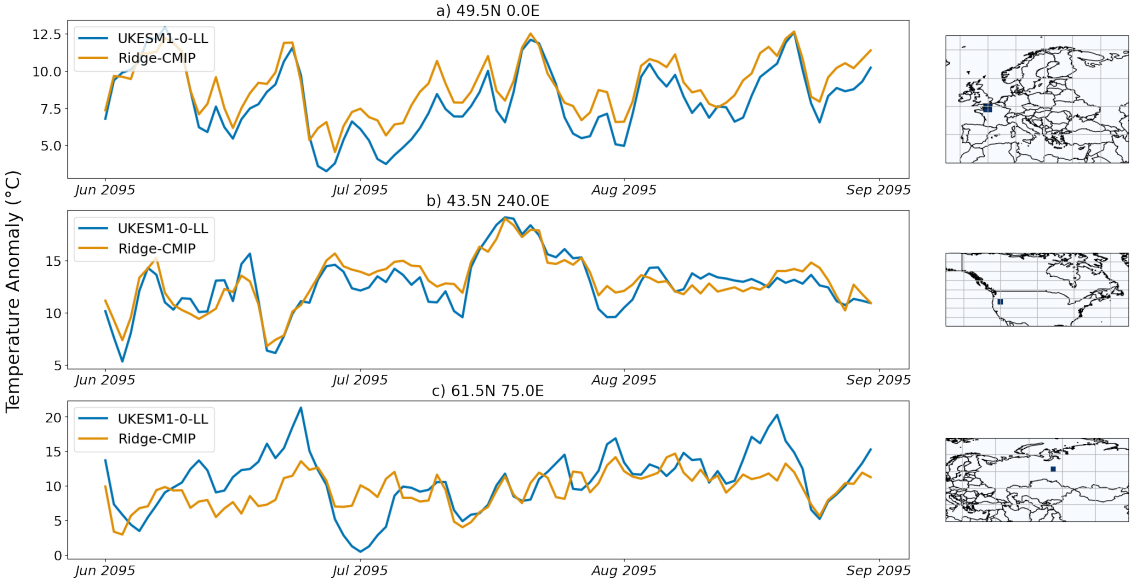

Figure 4.6: Ridge-CMIP temperature predictions (orange) compared with actual UKESM1-0-LL future projections (blue) at three test locations; a) 45N 0E, b) 43.5N 240E, c) 61.5N 70E during JJA 2095.

regions where historical performance skill was also poorer - *e.g.* Siberia in Figure 4.6c - see Section 3.2 for discussion of why this may be the case, and why higher latitude regions are excluded from summary statistics.

The extrapolation performance of the Ridge-CMIP models is then analysed on a regional basis. To quantify end of century warming, the mean temperature change compared to a historical period of 1979-2022 averaged over 2070-2100 is calculated. Ridge-CMIP predictions of this quantity are plotted against the future projections from *SSP585* for each model, as shown in Figures 4.7 and 4.8. The points show, for each CMIP6 model, the predicted future temperature change obtained by combining Ridge-CMIP coefficients with input variables from *SSP585*, plotted against the actual projected future temperature change under *SSP585*. A linear fit to these points, with prediction intervals, allows for comparison of this relationship with the one-to-one line. The orange distributions included in these plots will be explained as the Ridge-ERA5 observational constraint is derived in the following subsection.

The linear fit to the points (solid black lines in Figures 4.7 and 4.8) quantifies the relationship between the Ridge-CMIP predictions and the true CMIP6 temperature projections:

$$y_{CMIP} = a + by_{Ridge}. \tag{39}$$

A closer match to the one-to-one line indicates greater climate invariance in the relationships learnt, and therefore more accurate simulation by Ridge-CMIP of the future warming projected by the climate models. A prediction interval range (Wilks, 2006), $y \pm PI$, is calculated according to:

$$PI = \sqrt{S^2 \left(1 + \frac{1}{N} + \frac{(y - \bar{x})^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right)}. \tag{40}$$

Where S^2 is given by the mean squared error (see Section 2.1.5) between Ridge-CMIP predictions and CMIP6 projections and $\bar{x}$ is the mean of the N Ridge-CMIP predictions, $\hat{x}$. A prediction interval is wider than a confidence interval as it accounts for both the sampling uncertainty from the choice of coefficients and the irreducible error, ϵ, which is not explainable by the Ridge model. The prediction interval calculation is based on the simple mean squared error but is additionally moderated by the number of data points and the distance of data points from the mean. The prediction interval is shown by the dashed black lines in Figures 4.7 and 4.8.

Inclusion of the longer term mean variables of near-surface relative humidity and 850hPa specific humidity are both necessary to capture the magnitude of end-of-century warming. Likely due to the role these variables play in providing the model with information about the background warming state.

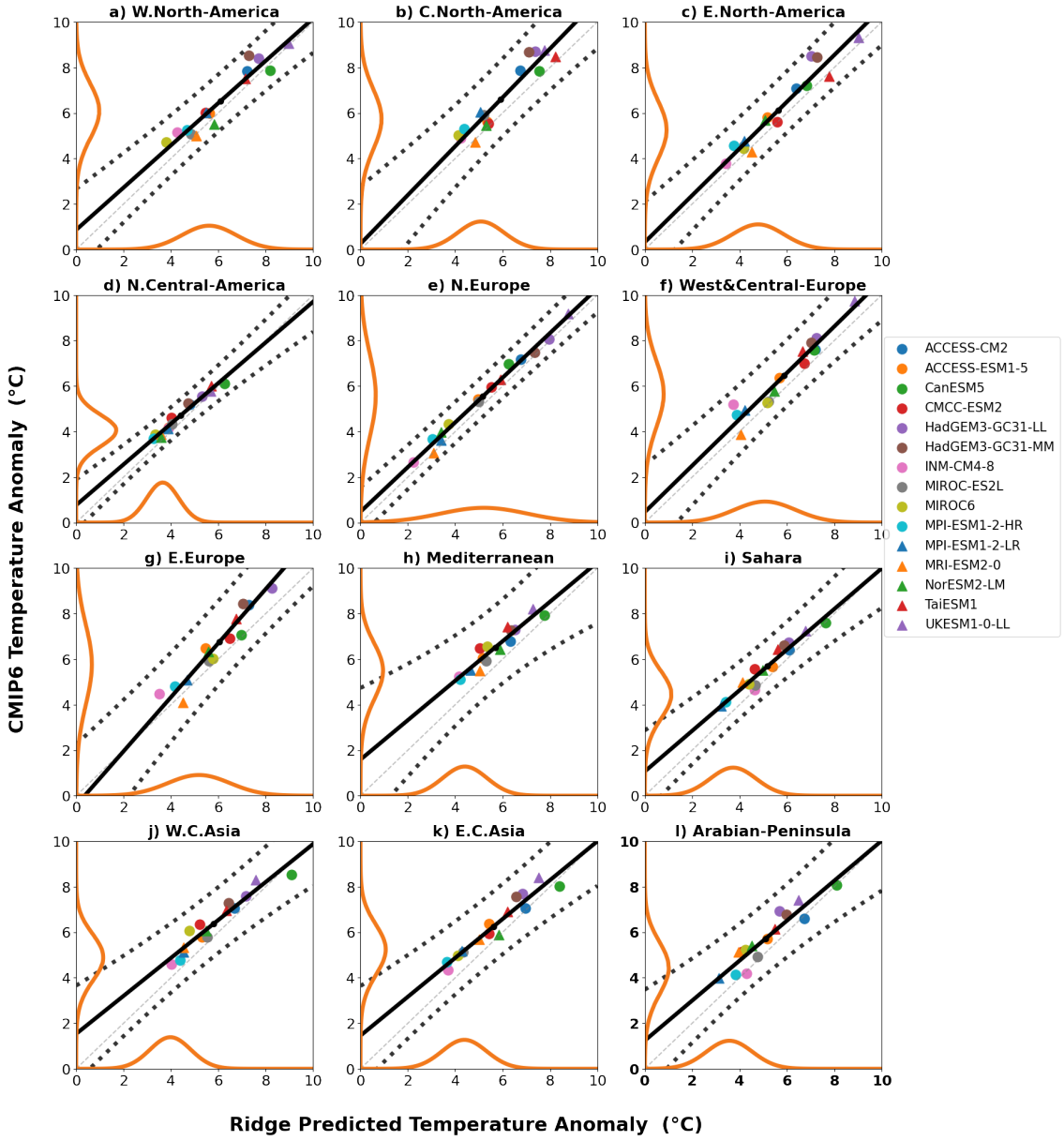

Figure 4.7: Projected mean daily temperature change (by 2070-2100) under *SSP585* during JJA from CMIP6 (y-axis) plotted against predicted change by Ridge-CMIP (x-axis) with least squares fit to points (black line) & prediction interval (black dashed line) for different AR6 regions. Probability distributions (orange lines) for Ridge-ERA5 future predictions given *SSP585* inputs (x-axis) convolved with Ridge prediction error for final constraint (y-axis).

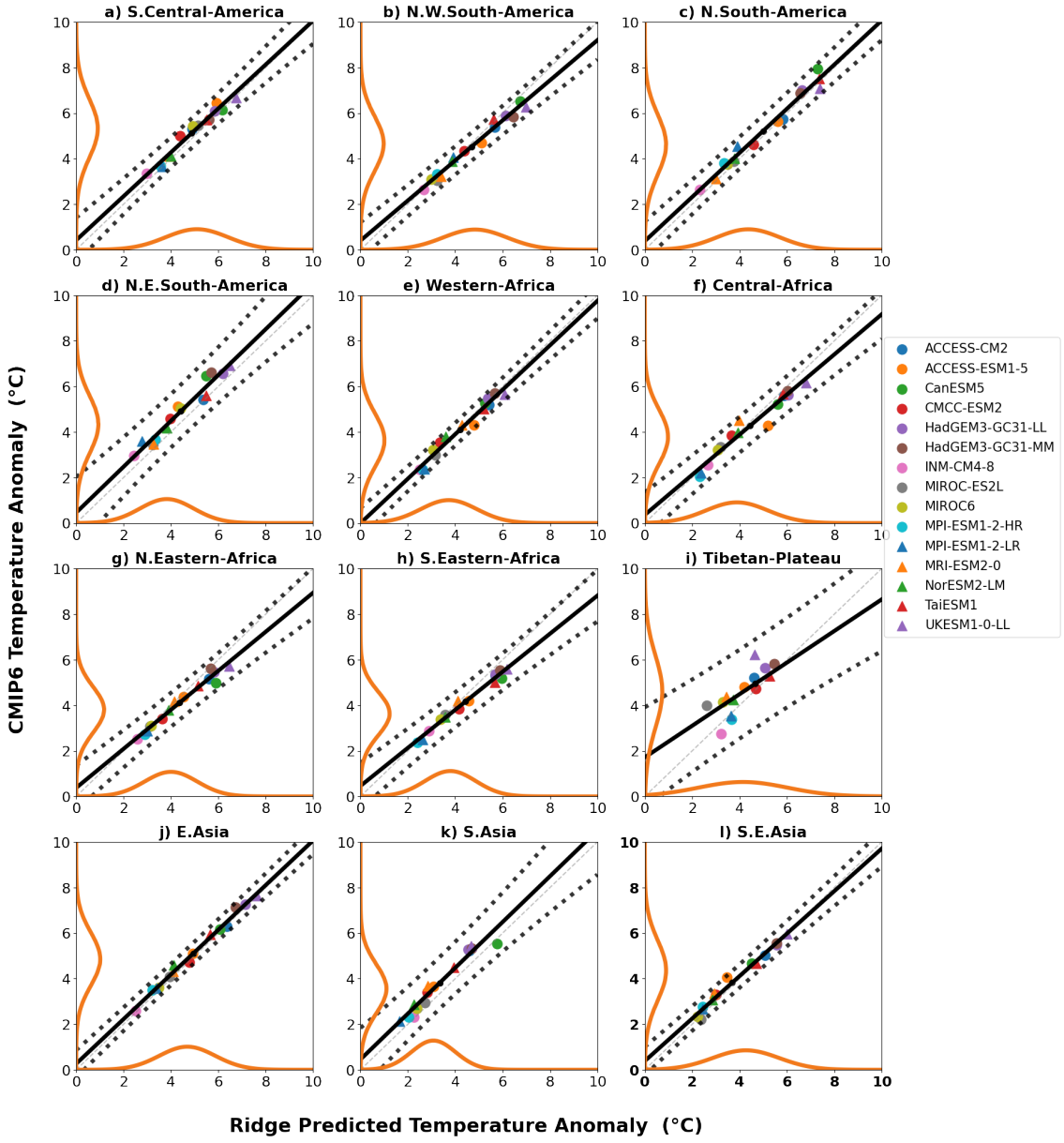

Figure 4.8: As in Figure 4.7 for remaining Northern Hemisphere AR6 land regions.

4.2.2 Using Ridge-ERA5 for Future Prediction

Next, Ridge-ERA5 is applied for future prediction to construct the orange distributions shown in Figures 4.7 and 4.8 which form the basis of the observational constraint. Projected temperature by the end of the century (2070-2100) compared with the historical period (1979-2022) is constrained on a regional basis for JJA, averaging over land grid cells (land area fraction > 0.7) in each AR6 region.

By normalising the predictor variables from *SSP585* in the CMIP6 archive (x) relative to the variance of the ERA5 data:

$$x' = \frac{x - \mu_{CMIP6,hist}}{\sigma_{ERA5,hist}}, \tag{41}$$

Ridge-ERA5 can be applied to the future predictor variables from CMIP6 in order to make predictions of daily-mean temperature anomalies out to the end of the century:

$$\hat{y}_{RidgeERA5} = \theta_{0,ERA5} + \sum_{j=1}^{P} \theta_{j,ERA5} x'_{j,CMIP}. \tag{42}$$

Where, $\theta_{j,RidgeERA5}$ is the coefficient for the j^{th} predictor variable learnt from ERA5 data and $x'_{j,CMIP}$ is the j^{th} predictor anomaly from the CMIP6 model normalised relative to ERA5 (according to Equation 41). Uncertainty in the future predictions, Δy, can be decomposed into uncertainty in the inputs, Δx, and uncertainty in the relationship between $\mathbf{x}$ and y, $\Delta\theta$. Compared with simply taking raw CMIP6 temperature projections, while Δx for the Ridge-ERA5 future temperature predictions ($\hat{y}_{RidgeERA5}$) still comes from the CMIP6 inputs, uncertainty in the relationships between $\mathbf{x}$ and y is constrained by using the observations-based coefficients, $\theta_{j,RidgeERA5}$.

To represent the distribution of Ridge-ERA5 future predictions, 10^7 equidistant points, $\mathbf{z}$, are sampled from a normal distribution with mean and standard deviation calculated across the $\hat{y}_{RidgeERA5}$ predictions for each model. This converts the 15 discrete points representing the Ridge-ERA5 future prediction for each CMIP6 model into a more continuous distribution. In order to account for any biases in the 'climate invariant' relationship, this distribution of temperatures for each region is substituted into the linear fit to points between the CMIP6 future projections and the Ridge-CMIP future projections (Equation 39):

$$\mathbf{z}' = a + b\mathbf{z}. \tag{43}$$

In order to fit a distribution to the translated temperatures, $\mathbf{z}'$, 10^7 equidistant points are sampled from a normal distribution fitted to $\mathbf{z}'$ with standard deviation calculated as the prediction interval. A Gaussian kernel density estimator (see Section 3) with bandwidth

0.01 is then fitted to these points to obtain the final distribution of constrained values. The observational constraint itself is calculated as a mean, 66% and 90% confidence range of this distribution. This process is repeated for each Northern Hemisphere AR6 region.

To obtain the final observational constraint, the mean and 5^{th} and 95^{th} percentiles of the y-axis distributions plotted in Figures 4.7 and 4.8 are calculated. These final constraints are plotted in comparison with the range of individual CMIP6 model projections for each AR6 region in Figure 4.9. Plotting the results in this way clearly reveals a pattern in the Ridge-ERA5 constraint that is consistent across the majority of the Northern Hemisphere AR6 regions. The constraint tends to exclude the warmest few models (frequently including CanESM5, HadGEM3-GC31-LL/MM, and UKESM1-0-LL) suggesting that these models are incompatible with the observational relationships learnt by Ridge-ERA5 and are potentially too sensitive. These results are consistent with existing literature which indicates that the increase in sensitivity of models between CMIP5 and CMIP6, resulting primarily from the representation of cloud feedbacks (Meehl et al., 2020; Zelinka et al., 2020), may not be consistent with observational evidence (Jiménez-de-la Cuesta and Mauritsen, 2019; Nijsse et al., 2020; Tokarska et al., 2020; Zhu et al., 2020). However, regional factors such as land surface feedbacks also play a role (Seneviratne and Hauser, 2020). The source of this constraint in terms of the contributing variables is analysed in more detail on a region by region basis in the next subsection.

4.3 Interpreting the Ridge-ERA5 Future Constraint

In the previous subsection, the Ridge-ERA5 observational constraint was applied to constrain average end of century summertime temperature change on a regional basis across the Northern Hemisphere. The resulting constraint generally results in a downward correction of the range of CMIP6 temperature projections with the models which predict the highest degree of warming being excluded by the constraint in many regions. In this subsection, SHAP value analysis is applied to the future Ridge predictions to identify which climatic predictor variables are the source of this downward correction in different regions.

To identify the source of the constraint in each region, Shapley Additive ExPlanation (SHAP) values (see Section 2.4.2) are calculated for Ridge-ERA5 predictions of future temperature anomalies as well as for Ridge-CMIP predictions of future temperature anomalies across the Northern Hemisphere. SHAP values quantify the contribution (in ^oC) of each predictor variable to the predicted temperature anomaly on a particular day. For predictor variables which cover a 5x5 domain, contributions from each input variable can simply be summed to give the total contribution of that climatic variable to the regression prediction.

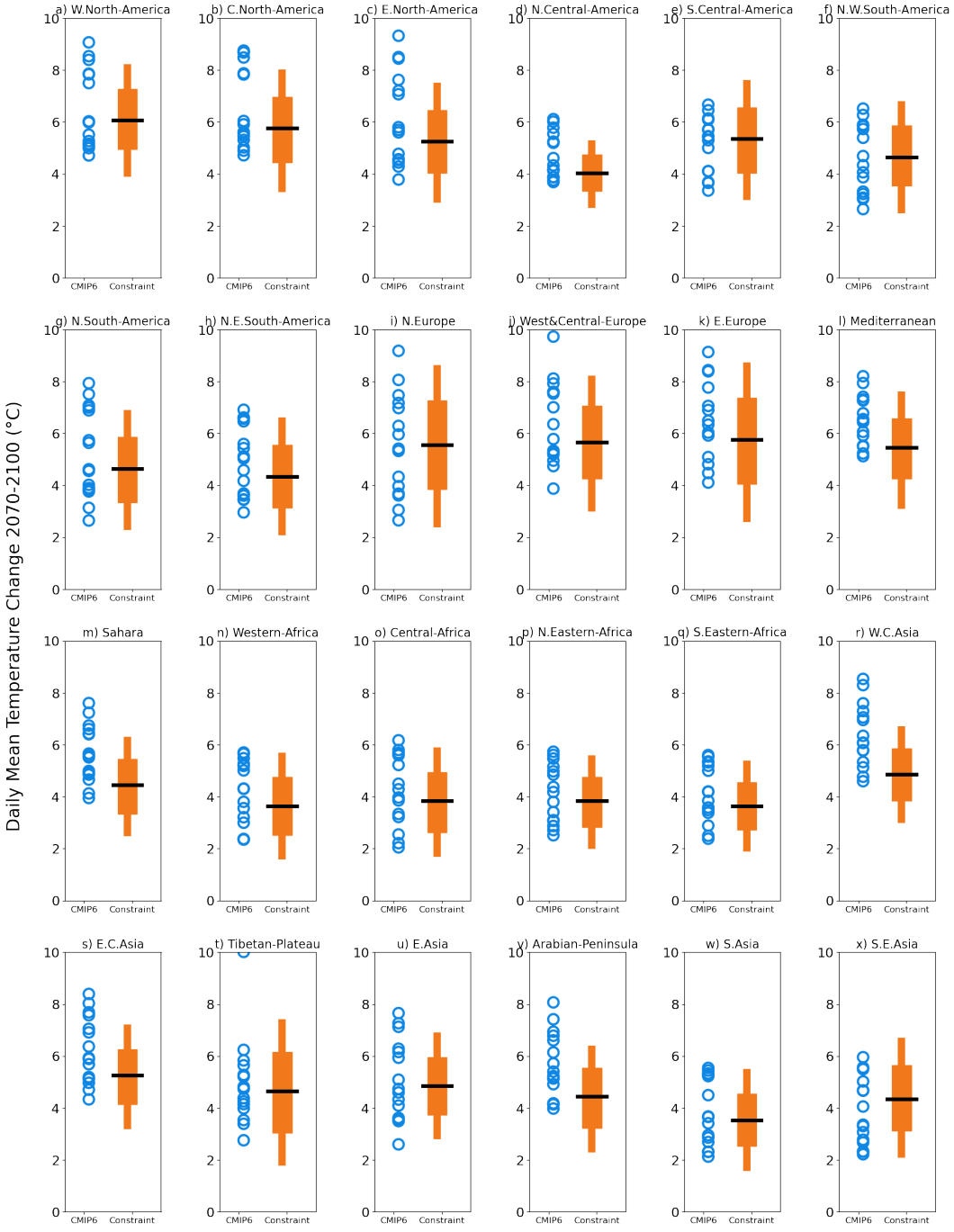

Figure 4.9: Mean temperature change during JJA projected by each CMIP6 model (blue circles) alongside observationally constrained mean (black line), 60% (wide orange bar) and 90% (thin orange bar) intervals for each Northern Hemisphere AR6 region.

So, for each predicted future temperature anomaly, the difference between the Ridge-ERA5 prediction and the Ridge-CMIP prediction can be decomposed into contributions from each climate predictor variable to identify the source of the observational constraint. These SHAP value differences are plotted for various AR6 regions throughout this subsection in order to explain the performance of the Ridge-ERA5 future constraint.

Regions where the constraint is widest are Northern (Figure 4.9i), West & Central (Figure 4.9j) and Eastern (Figure 4.9k) Europe and potentially the Tibetan Plateau (Figure 4.9t). The European regions also correspond to areas where the historical bias correction failed (Figure 3.14 e, f, g) and performance of Ridge-ERA5 on ERA5 test data was worse (Figure 3.10 h, i, j). The wider uncertainty range here makes sense given that Ridge-ERA5 performs worse here in general. The fact that poorer Ridge-ERA5 performance results in a wider uncertainty range is itself a positive feature of the method rather than obtaining an over-confident but narrower constraint. SHAP value differences between contributions of each predictor variable to Ridge-ERA5 and Ridge-CMIP future temperature anomaly predictions are plotted in Figure 4.10. The largest contributions to the difference between Ridge-ERA5 and Ridge-CMIP predictions come from soil moisture, daily and monthly near-surface relative humidity and the 850hPa monthly specific humidity with the relative humidity variables having the greatest uncertainty in the difference value. The soil moisture contribution seems to suggest that the soil moisture feedback is under-estimated by Ridge-CMIP models.

To understand the size of the constraint on the Tibetan Plateau region, SHAP value differences are plotted for each predictor variable. Figure 4.11 shows the average difference between Ridge-CMIP and Ridge-ERA5 SHAP values for end of century (2070-2100) temperature predictions over the Tibetan Plateau. For most predictors the contributions to Ridge-CMIP and Ridge-ERA5 predictions are very similar - the SHAP value differences are clustered around zero. However, for the 850hPa monthly specific humidity, there is a large spread in the SHAP value differences. This spread in 850hPa SHAP value differences may be related to the altitude of the region. As described in Chapter 2, a common mask is applied to all 850hPa variables across ERA5 and CMIP6 in order to mask values which would be obscured my the topography. In high altitude regions, such as the Tibetan Plateau, this reduces the number of values in the 5x5 domain of, for example 850hPa specific humidity, compared with the full 25 inputs that would be used in lower altitude regions (see Figure 2.16). This is particularly relevant when Ridge-ERA5 is applied for future prediction as the monthly mean variables of near-surface relative humidity and 850hPa specific humidity are essential for capturing the future warming trend and make a larger contribution to predictions than in the historical climate. Differences could also arise as a result of how the topography is represented in different CMIP6 models versus ERA5, particularly given that the native resolution of most CMIP6 models is much lower than the native resolution of ERA5.

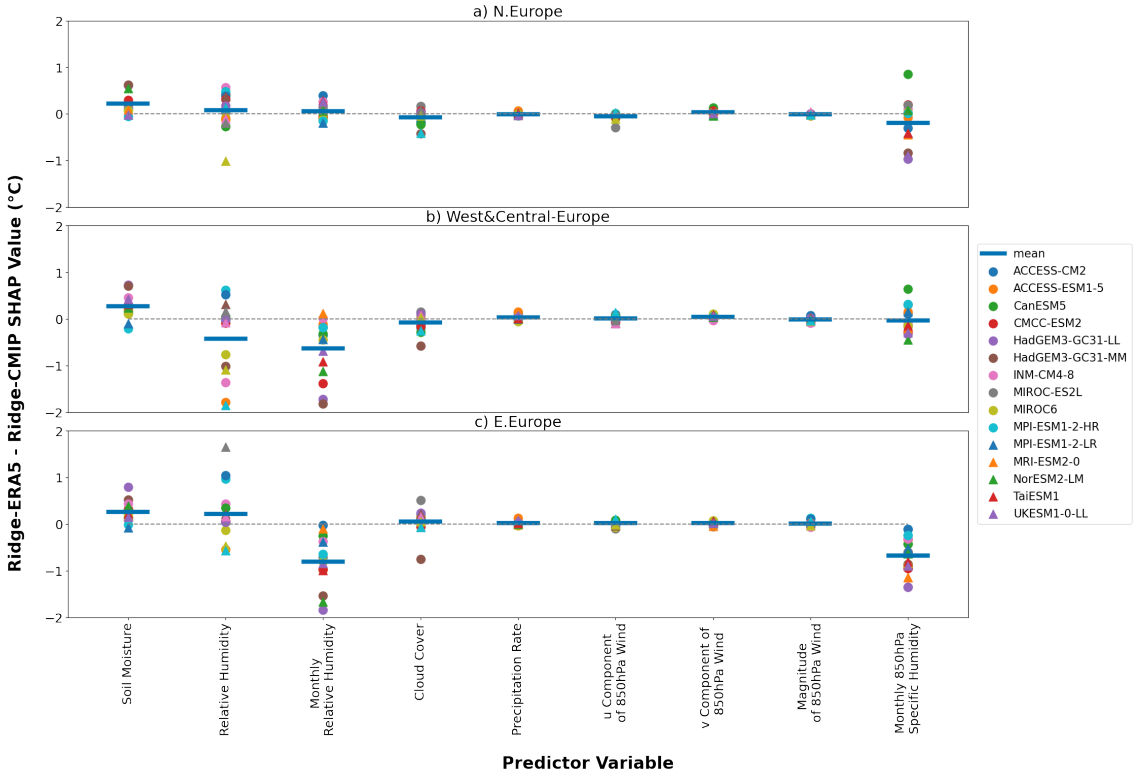

Figure 4.10: Average difference in SHAP values (Ridge-ERA5 minus Ridge-CMIP) for end-of-century temperature predictions for each Ridge predictor variable over Northern Europe (a), West & Central Europe (b) and Eastern Europe (c). A positive (negative) value of N^oC for a particular predictor variable indicates that the variable contributes an average N^oC of warming (cooling) more in Ridge-ERA5 than the same predictor variable in Ridge-CMIP.

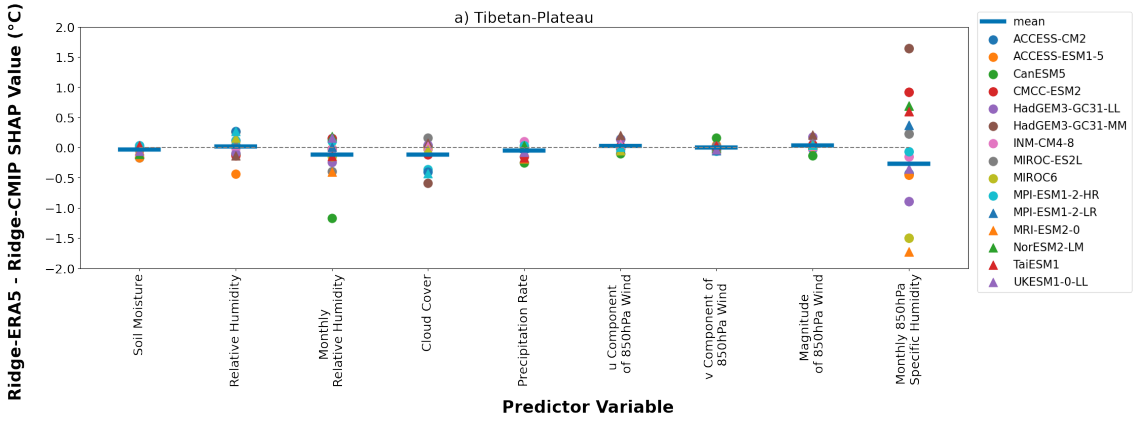

Figure 4.11: As in Figure 4.10 for the Tibetan Plateau AR6 region.

In the following regions, the Ridge-ERA5 future constraint results in an up-shift of the range of expected temperatures: Southern Central America (Figure 4.9e) and South-East Asia (Figure 4.9x). The average SHAP value differences between Ridge-ERA5 and Ridge-CMIP future predictions are plotted for these regions in Figure 4.12. Comparing the SHAP

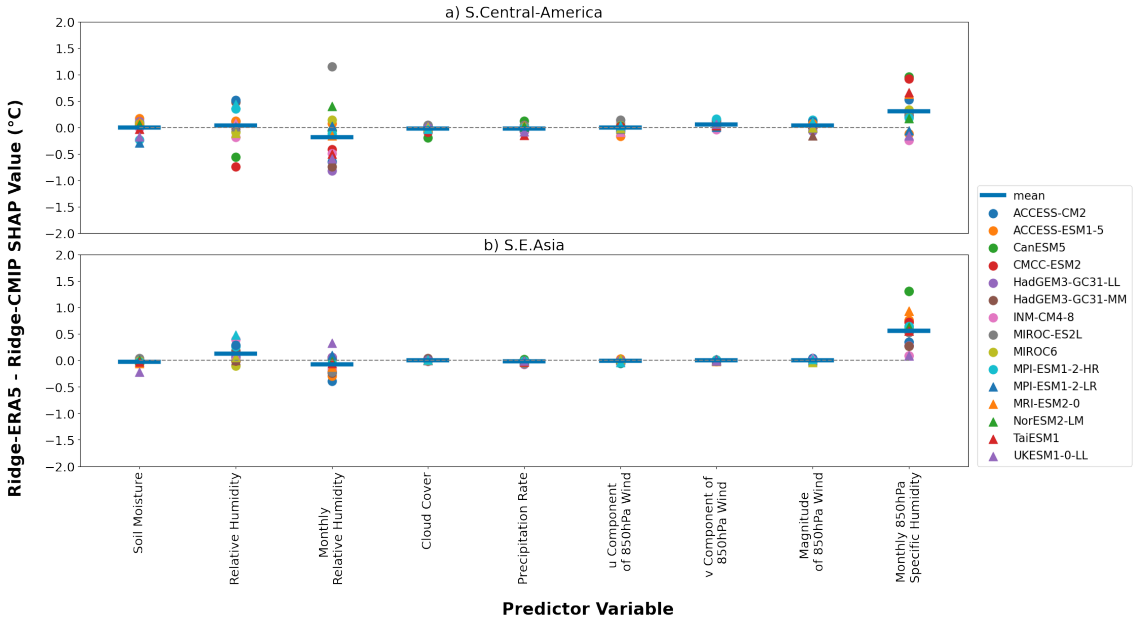

Figure 4.12: As in Figure 4.10 for Southern Central America (a) and South East Asia (b).

value differences between each predictor variable, it is clear that the up-shift in the temperature range comes primarily from the contributions of the monthly 850hPa specific humidity. The Ridge-ERA5 850hPa specific humidity coefficients appear to be more sensitive than those learnt by the Ridge-CMIP models, *i.e.* a given increase in the background specific humidity state is consistent with a larger temperature anomaly according to Ridge-ERA5.

The greatest downward shifting of the expected range of temperature anomalies can be found in Northern Central-America (Figure 4.9d), the Sahara (Figure 4.9m), West Central Asia (Figure 4.9r), East Central Asia (Figure 4.9s) and the Arabian Peninsula (Figure 4.9v), all relatively dry regions. As with South Central America and South East Asia which exhibit an up-shift in the range of expected temperature anomalies, the primary factor controlling the downshift of the distribution is the 850hPa specific humidity. Although in this case, it appears that the Ridge-ERA5 coefficients are *less* sensitive than those learnt by Ridge-CMIP resulting in a smaller degree of projected warming. There are also negative SHAP value difference contributions from the relative humidity variables in most regions in Figure 4.13 and a small average negative contribution from soil moisture over the Arabian Peninsula.

In other Northern Hemisphere AR6 regions the constraint results in a moderate downward

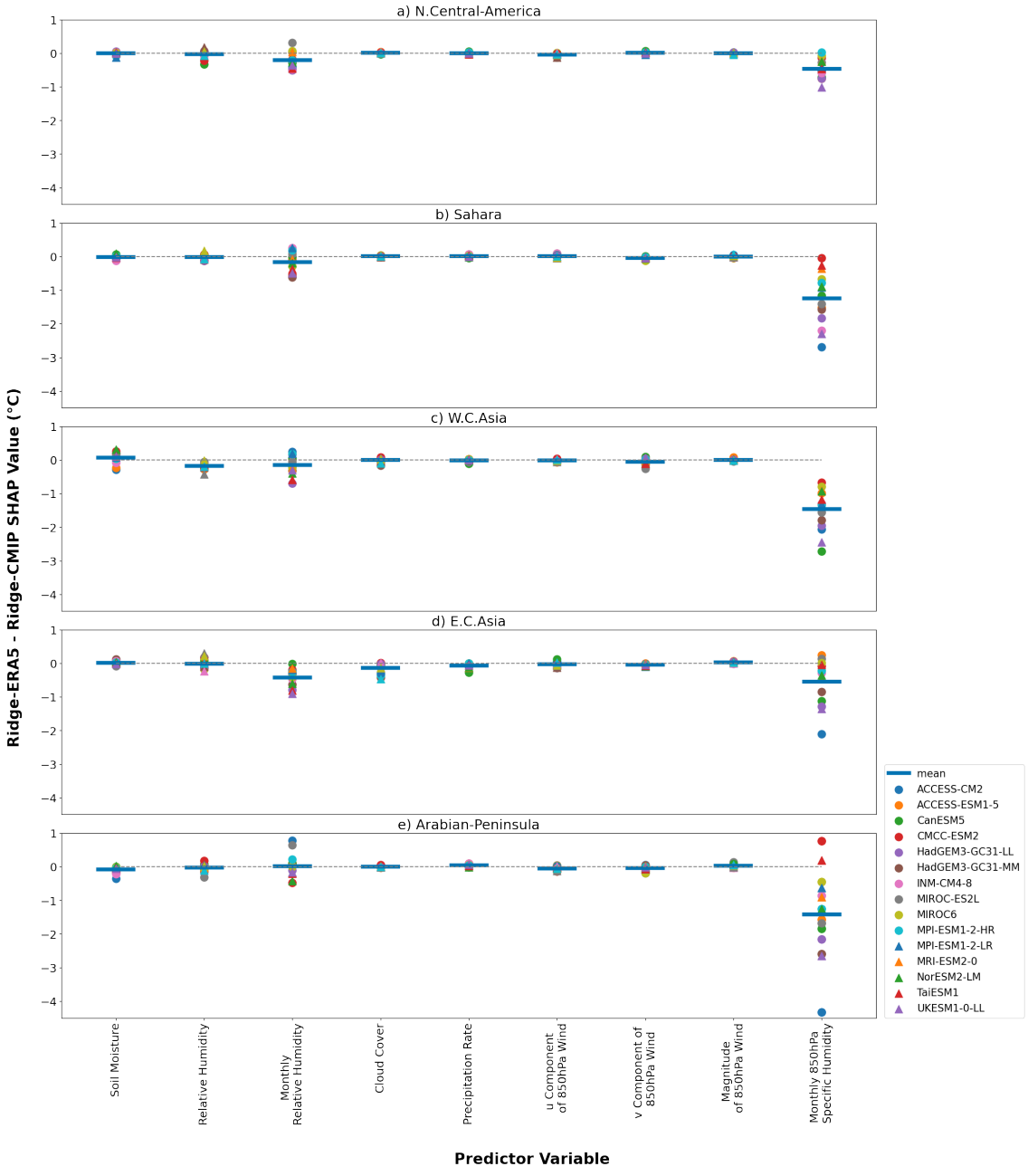

Figure 4.13: As in Figure 4.10 for Northern Central America (a), the Sahara (b), West Central Asia (c), East Central Asia (d) and the Arabian Peninsula (e).

correction, excluding the top few models which predict the most warming and extending the lower bound of the uncertainty range to include temperatures up to $\approx 1^oC$ lower than the original CMIP6 range. As for the regions analysed already, the dominant contributors to this constraint appear to be the monthly near-surface relative humidity and 850hPa specific humidity variables. However soil moisture and cloud cover also make small contributions to the constraint in North America, the Mediterranean and South Asia.

4.4 Conclusion

Coefficients learnt by the process-based ML model, Ridge-ERA5, are applied to predictor variables from future emission scenario *SSP585* from the CMIP6 archive during JJA to produce new future temperature prediction time series. These future Ridge-ERA5 predictions form the basis of a novel, observational constraint on future regional warming over land in the Northern Hemisphere. Climate invariance of the coefficients Ridge is able to learn to represent relationships between temperature drivers and anomalies is demonstrated by comparing end of century climate projections from CMIP6 models with temperature anomalies predicted by corresponding Ridge-CMIP emulators (which learn from historical climate model data). These Ridge-CMIP ML models show excellent performance, accurately predicting temperature anomalies which exceed those they have been exposed to during training under the most extreme future warming scenario representing a rigorous extrapolation challenge.

The key finding from these results is the exclusion across the Northern Hemisphere of those models which project the greatest degree of daily warming on a regional basis by the end of the century. There has been much discussion surrounding the increased equilibrium climate sensitivity between the previous generation of climate models in the CMIP5 archive and those in CMIP6 (e.g. Meehl et al., 2020). The indication of this observational constraint is that some models may be 'too warm' now, and in fact incompatible with observed relationships between temperature and other driving factors that have been modelled here. This contributes to a growing body of evidence that the sensitivity of some models in the CMIP6 archive may not be compatible with the observed historical climate (Jiménez-de-la Cuesta and Mauritsen, 2019; Nijsse et al., 2020; Tokarska et al., 2020; Zhu et al., 2020).

Comparison of SHAP value differences which quantify the average difference in contribution between predictor variables to Ridge-ERA5 and Ridge-CMIP future predictions appears to indicate that the monthly mean variables of near-surface relative humidity and 850hPa specific humidity are the greatest contributors to the constraint. And also, that soil moisture feedbacks may be too weak in some regions.

Compared to traditional climate model weighting approaches to constraining uncertainty,

a key advantage of this method is the possibility for the range of the constraint to exceed the range of the existing temperature projections. This is important because climate model ensembles, like CMIP6, are frequently not designed with the idea of sampling the feasible uncertainty range in mind. The upper and lower limits of these ensemble projections therefore cannot be expected to provide reliable boundaries for the possible range of future temperature anomalies. By constraining the relationships between variables with Ridge-ERA5 coefficients, models which have relationships between temperature and predictor variables that vary greatly from those observed in ERA5 may produce future temperature predictions which exceed the limits of the CMIP6 temperature distribution when inputs from that scenario are combined with coefficients learnt from ERA5. This is also relevant to the idea of inter-dependencies between models. Models with similar predictor variable projections should produce similar future projections of temperature when fed into Ridge-ERA5.

The constraint is based upon the relationships learnt by Ridge from observational data so represents a physical, process-based constraint, without the problems of other statistical approaches, such as the emergent constraint framework, where links between historical changes and future projections may result from spurious links in the data rather than representing connections which actually have a physical basis in the climate system. With the Ridge-ERA5 method, the same process-based function can be used for past and future rather than relying on indirect, though physically-motivated, correlations. The constraint can also be applied as a bias correction to historical climate model data, see Chapter 3, for consistent transformation of both historical and future climate model output.

5 Applying Process-Based Machine Learning to Analyse Extreme Events

The aim of this chapter is to apply the process-based ML model trained on reanalysis data, Ridge-ERA5, to historical heatwave events. A Shapley Additive ExPlanation (SHAP) value analysis is used to quantify the contributions from individual predictor variables over the course of four case study Northern Hemisphere summer heatwaves. The inferences this provides as to the physical mechanisms underlying these heatwaves are compared with existing literature (Section 5.1). Predictor anomalies during these heatwave events are then combined with Ridge coefficients from historical CMIP6 data to assess how these events are reproduced through the emulated behaviour of individual climate models (Section 5.2). Using a 95^{th} percentile threshold to define extreme events, SHAP values for extreme temperature are calculated across the Northern Hemisphere for each Ridge-CMIP emulator and Ridge-ERA5 to compare the representations of extremes which can be learnt from observations *vs.* climate model data (Section 5.3.1). Finally, the Ridge-CMIP analysis is extended to consider the representation of heatwaves under future climate change scenario *SSP585*, comparing the SHAP values for 95^{th} percentile events in historical and future climates (Section 5.3.2).

5.1 Interpreting Historical Heatwave Drivers Using Ridge-ERA5

This analysis of extreme events using the process-based Ridge-ERA5 model begins by focusing on four case study Northern Hemisphere summer heatwaves: Europe 2003; Europe 2018; the Pacific Northwest 2021; and China 2022. Figure 5.1 indicates the grid cell locations included in each region. Ridge-ERA5 is used to make test predictions for daily temperature anomalies during each of these heatwaves which are then assessed against the true temperature anomalies recorded in the ERA5 data set.

The Ridge-ERA5 temperature anomaly predictions are then broken down into contributions from each climatic predictor variable using SHAP value analysis (see Section 2.4.2). To calculate SHAP values for, for example the Europe 2003 heatwave, Ridge-ERA5 models are trained at each grid cell in the vicinity of the event using ERA5 data from 1979-2022 but excluding 2003. The prediction time series is then produced using predictor variables from JJA 2003 which were not seen previously during training. SHAP values are calculated by multiplying the learnt coefficients and corresponding input values at each location and at each time step to produce time series of SHAP values over the course of the heatwave. For example, multiplying the soil moisture anomaly by the soil moisture coefficient gives the contribution of that variable in ^oC to the temperature anomaly prediction on a given day. For

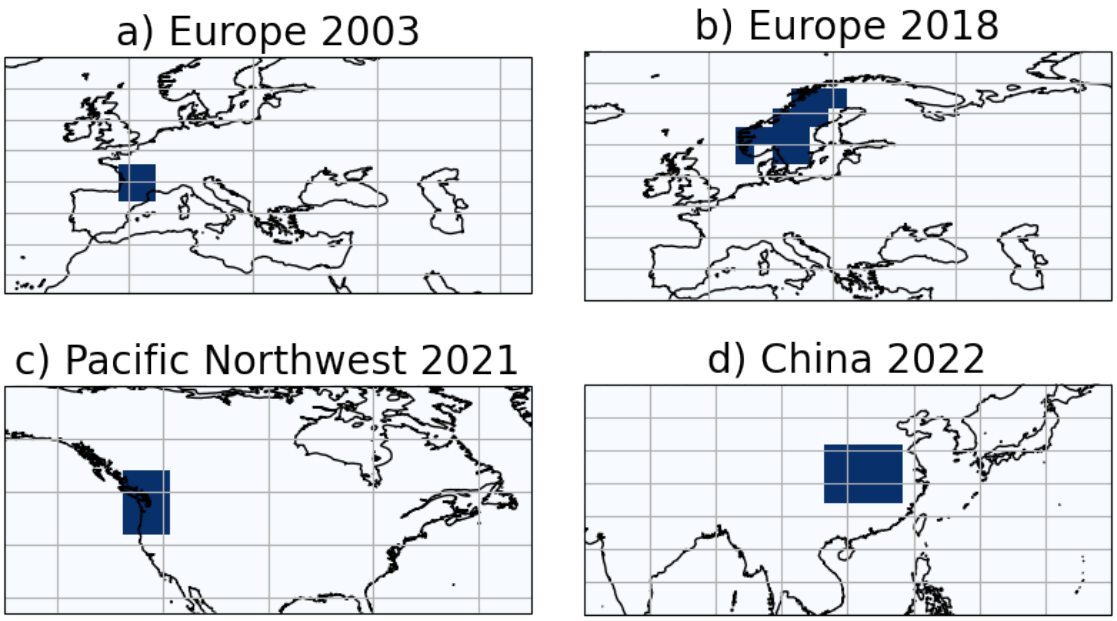

Figure 5.1: Regions defined for each historical heatwave on a $3° \times 3°$ longitude-latitude grid, a) Europe 2003, b) Russia 2010, c) Pacific Northwest 2021.

predictor variables which cover a 5×5 domain, contributions can simply be summed to give the net contribution of that climatic variable to each daily temperature anomaly prediction. The inferences that this analysis can provide as to the drivers behind each heatwave event are then compared with existing literature.

5.1.1 Europe 2003

The first heatwave event considered is Europe 2003, one of the hottest European summers on record (Luterbacher et al., 2004; Schär and Jendritzky, 2004), which resulted in an estimated 70,000 excess deaths (Robine et al., 2008) across Europe. Ridge-ERA5 predictions are plotted alongside ERA5 temperature anomalies for JJA 2003 and SHAP value contributions for each predictor variable in Figure 5.2. All anomalies are averaged over the region affected (see Figure 5.1a). Predictions of Ridge-ERA5 follow the pattern of temperature anomalies throughout the summer of 2003 well although they fail to consistently predict the extent of the temperature anomalies during the heatwave itself, from early to mid August.

In the literature, three major contributing factors to the anomalously hot summer of 2003 have been identified: persistent atmospheric blocking; soil moisture deficits; and remote sea surface temperature anomalies (Black et al., 2004; Cassou et al., 2005; Domeisen et al., 2023; Ferranti and Viterbo, 2006; Fink et al., 2004; Fischer et al., 2007; Stefanon et al., 2012;

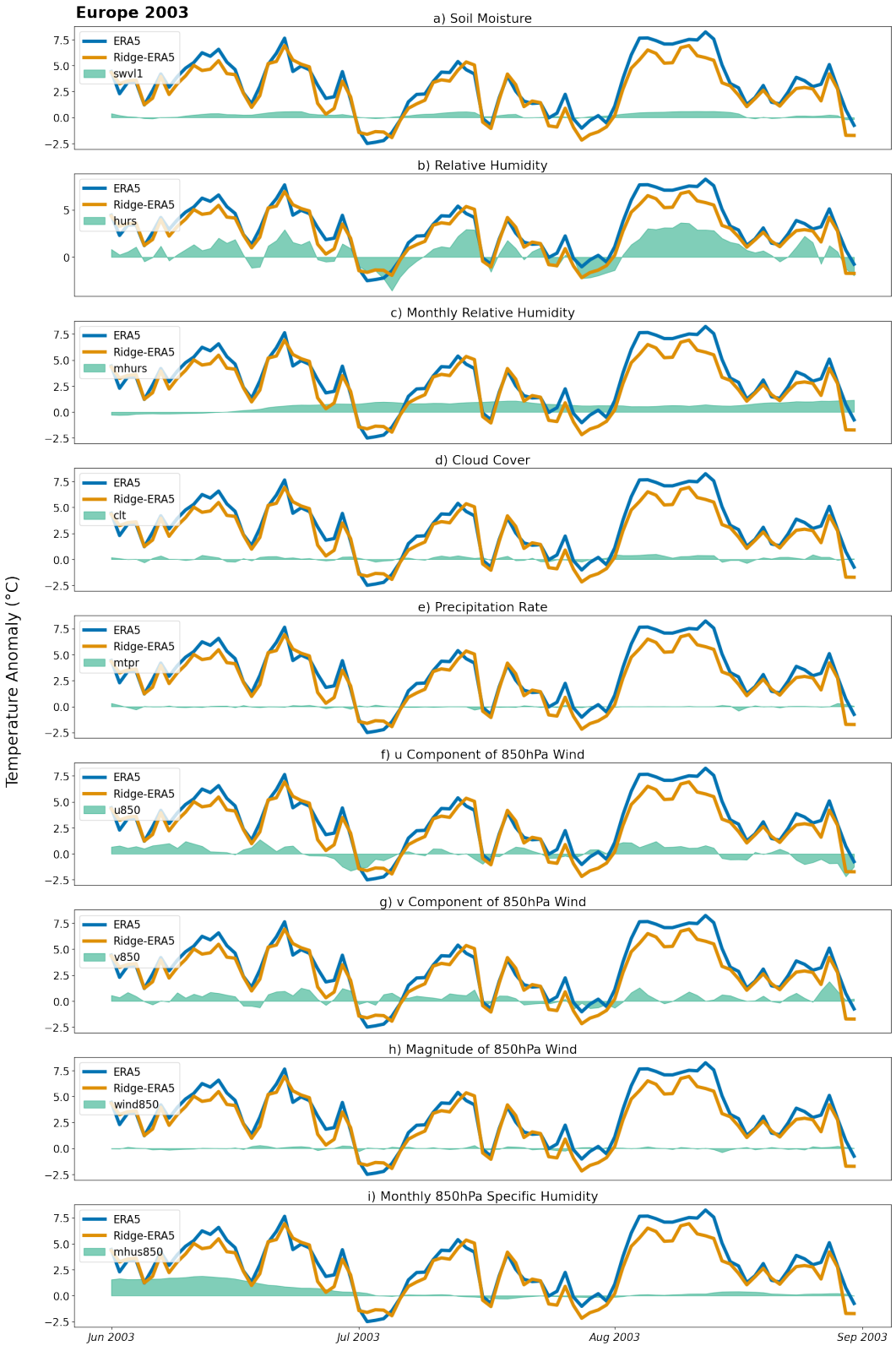

Figure 5.2: SHAP value contributions (green shading) to Ridge-ERA5 temperature anomaly predictions (orange lines) during the Europe 2003 heatwave (blue lines).

Zaitchik et al., 2006). Other potential influences include teleconnection patterns as well as anomalously clear skies and subsequent downward radiative fluxes (Garcia-Herrera et al., 2010).

Based on Figure 5.2, the predictor variables which contribute most to the Ridge-ERA5 predictions from early to mid August are relative humidity anomalies (Figure 5.2b) which build up to a peak and then diminish, dynamical anomalies in the u and v components of the 850hPa wind (Figure 5.2f, g) and soil moisture anomalies (Figure 5.2a). The soil moisture SHAP values are consistent with previously identified soil moisture deficits which would reduce the capacity for evaporative cooling. Also consistent with longer term soil moisture deficits, monthly near-surface relative humidity deficits were contributing to positive temperature anomalies the onset of which preceded the heatwave itself by several weeks.

The initiation of the heatwave coincides with positive SHAP value contributions from the u component of the 850hPa wind variables (Figure 5.2f) which may indicate anomalous synoptic conditions resulting from a slowing down or reversal of typical westerly flow (such as blocking) given that the associated coefficients have a negative relationship with temperature. These contributions remain positive throughout the duration of the heatwave, turning negative as temperatures start to cool again, perhaps indicating a return to more typical synoptic conditions. The v component of the 850hPa wind also contributes warm anomalies throughout the heatwave (Figure 5.2g), potentially indicative of warm air transport given the latitude of the test region. Positive contributions from the daily near-surface relative humidity (Figure 5.2b) begin alongside the contributions from the v component of the 850hPa wind, which would be consistent with poleward transport of warm, moist air. There are also small, positive contributions from cloud cover predictors (Figure 5.2d) which are consistent with clear skies and enhanced radiative heating which is associated with blocking conditions.

In summary, analysis of SHAP value contributions shows a broad agreement with previously identified drivers of the Europe 2003 heatwave. Particularly for soil moisture and cloud cover contributions which are linked clearly to physical processes: the capacity for evaporative cooling and the degree of surface radiative heating. While clear positive contributions are identified for the dynamical variables, disentangling the precise source of these anomalies is more challenging, although contributions from the u component of the 850hPa wind do seem to be consistent with atmospheric blocking.

5.1.2 Europe 2018

The second case study heatwave is Europe 2018, when daily temperatures higher than 33^oC were recorded in Scandinavia throughout July (NOAA, 2018). Ridge-ERA5 predicted temperature anomalies during the event are plotted over Northern Europe (refer to Figure 5.1b)

with SHAP value contributions from each predictor variable in Figure 5.3. Overall, Ridge-ERA5 predictions show excellent agreement with the true time series, capturing the magnitude of the warmest temperature anomalies.

Similarly to Europe 2003, the heatwave in 2018 was initially driven by an anomalous high pressure system, the result of a positive North Atlantic Oscillation (NAO) and a Rossby Wave-7 pattern (Drouard et al., 2019; Kornhuber et al., 2019; Liu et al., 2020). Atmospheric conditions enhanced the temperature anomalies (Yiou et al., 2020), with a persistent period of high pressure over Finland during July and a warmer than average lower to mid troposphere (Sinclair et al., 2019). In contrast with Europe 2003, soil moisture anomalies in Finland during July 2018 were relatively small (Liu et al., 2020) despite drought conditions stretching back as far as March across central and Northern Europe (Toreti et al., 2019). The magnitude of the high temperatures may instead be attributable to increased advection of warm air and solar heating (Liu et al., 2020) given that sensible heat fluxes were higher than average (Sinclair et al., 2019).

The SHAP value contribution from the u component of the 850hPa wind is very similar to the 2003 heatwave: positive contributions from this variable (Figure 5.2f) align with the onset of the heatwave while negative contributions occur as the heatwave ends. This is consistent with the anomalous high pressure system which initiated the heatwave. Contributions from cloud cover predictor variables (Figure 5.3d) begin a couple of days after this and result in larger positive anomalies compared with Europe 2003 (Figure 5.2d), suggesting that clear sky conditions enabled greater solar heating in 2018.

The gradual increase in monthly 850hPa specific humidity (Figure 5.3i) may be consistent with transport of warm, moist air resulting from warmer than usual sea surface temperatures (Byrne and O'Gorman, 2018) which is in agreement with existing literature that identifies warm air advection as a key factor. Contributions from the v component of the 850hPa wind (Figure 5.3g) are also positive indicating warm air transport from lower latitudes. Positive soil moisture contributions (Figure 5.3a) preceding the start of the heatwave indicate that drought pre-conditions may have contributed to the magnitude of the heatwave.

As with the Europe 2003 heatwave, interpretation of SHAP value contributions shows general agreement with existing literature. Although, as before, certain variables (*e.g.* 850hPa wind), are more complex to interpret than others.

5.1.3 Pacific Northwest 2021

The Pacific Northwest heatwave of 2021 was larger in magnitude but shorter in duration than either the Europe 2003 or Europe 2018 heatwaves. Estimated at the time to be a 1-in-1000 year event (Philip et al., 2021), record breaking temperatures of 49.6^oC were recorded in

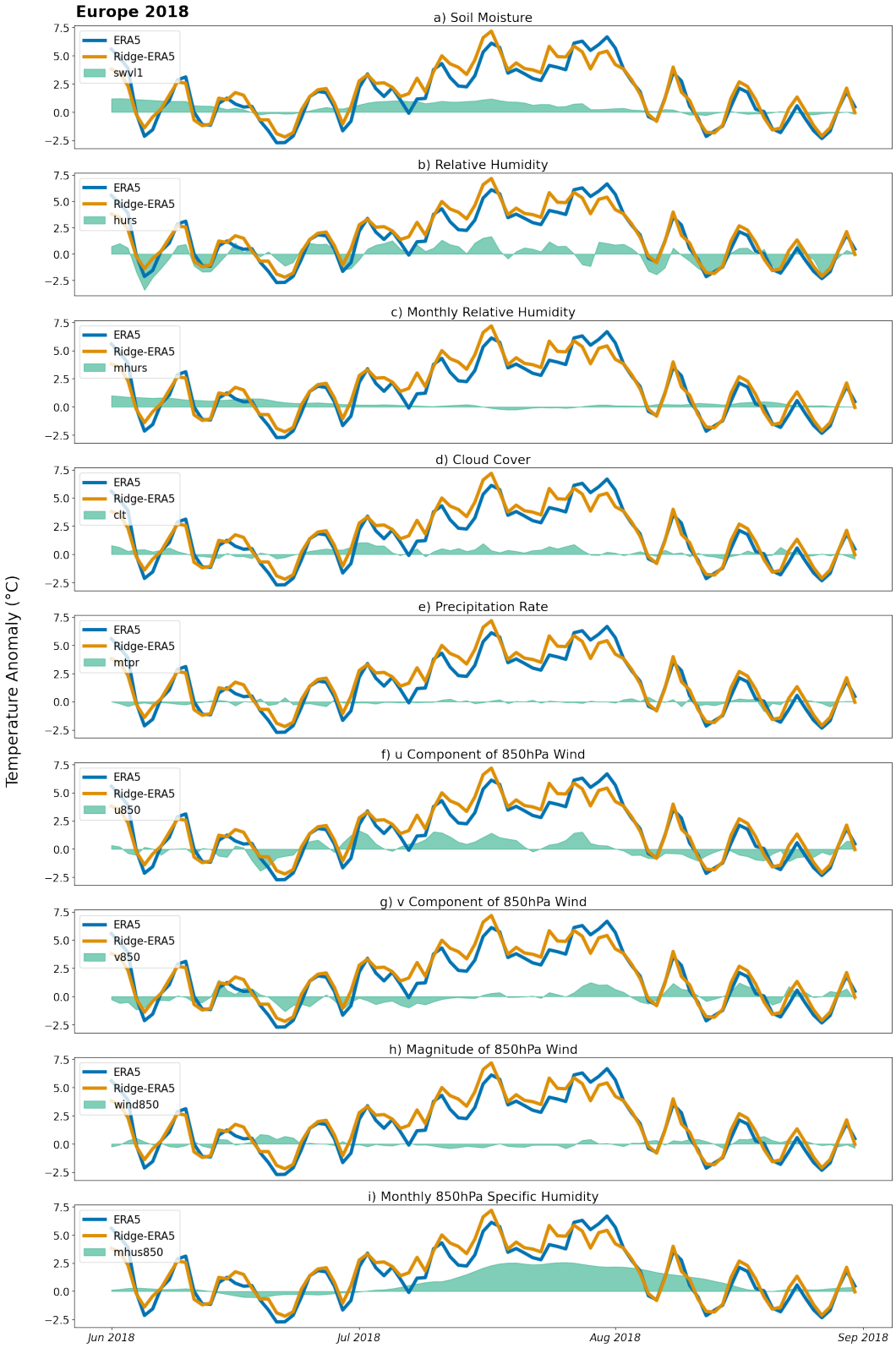

Figure 5.3: As in Figure 5.2 for the Europe 2018 heatwave.

Canada (White et al., 2023), far outside the range of historical observations. Ridge-ERA5 temperature anomaly predictions for the summer of 2021 over the Pacific Northwest (refer to Figure 5.1c) are plotted with SHAP value contributions from each predictor variable in Figure 5.4.

Although closely matching daily temperature anomalies for most of the summer, Ridge-ERA5 predictions fail to capture the magnitude of the late June heatwave, although a clear peak in predictions is present. This likely speaks to the unprecedented nature of this event. The reanalysis period 1979-2022 provides 44 years of data and focusing on the three summer months (JJA) corresponds to 92 days per year for a total of 4048 data points. For anomalies exceeding the 95^{th} percentile results there are approximately 200 such events for training whilst a 99^{th} threshold would correspond to only $\tilde{4}0$. Bearing in mind that all such temperature anomalies at a given location are not driven by the same combination of mechanisms and initial conditions, the true value of similar training cases for a given heatwave event may be considerably smaller than these values, particularly for events which are unprecedented in the observational record.

The Pacific Northwest heatwave has been associated with highly anomalous atmospheric blocking (Neal et al., 2022; Overland, 2021; White et al., 2023) with an already warm transported air mass being further locally heated by subsidence and sensible heat fluxes (Neal et al., 2022; Schumacher et al., 2022). The event was likely amplified by land-atmosphere feedbacks (Bartusek et al., 2022; Conrick and Mass, 2023; Schumacher et al., 2022). These dynamical effects are potentially identifiable in the daily near-surface relative humidity (Figure 5.4b) and the 850hPa monthly specific humidity (Figure 5.4). For the monthly 850hPa specific humidity, a gradual build up of heat is identifiable possibly representing transport of the warm air mass while the daily near-surface relative humidity contributions follow the timing and shape of the event itself when this warm air mass is further heated. A positive anomaly in the v component of the 850hPa wind (Figure 5.4g) several days prior to the hottest peak may also be consistent with warm air transport. Cloud cover contributions (Figure 5.4d) consistent with clear sky radiative heating are present throughout the event along with small contributions from soil moisture effects (Figure 5.4a).

5.1.4 China 2022

Finally, the fourth heatwave considered is from China in 2022 where daily maximum temperatures exceeded the historical (1981-2010) average by more than 10^oC (Hua et al., 2023). As before, Ridge-ERA5 temperature anomaly predictions are plotted with SHAP value contributions from each predictor variable in Figure 5.5. Excellent agreement is found between Ridge-ERA5 predictions and actual ERA5 temperature anomalies throughout the summer,

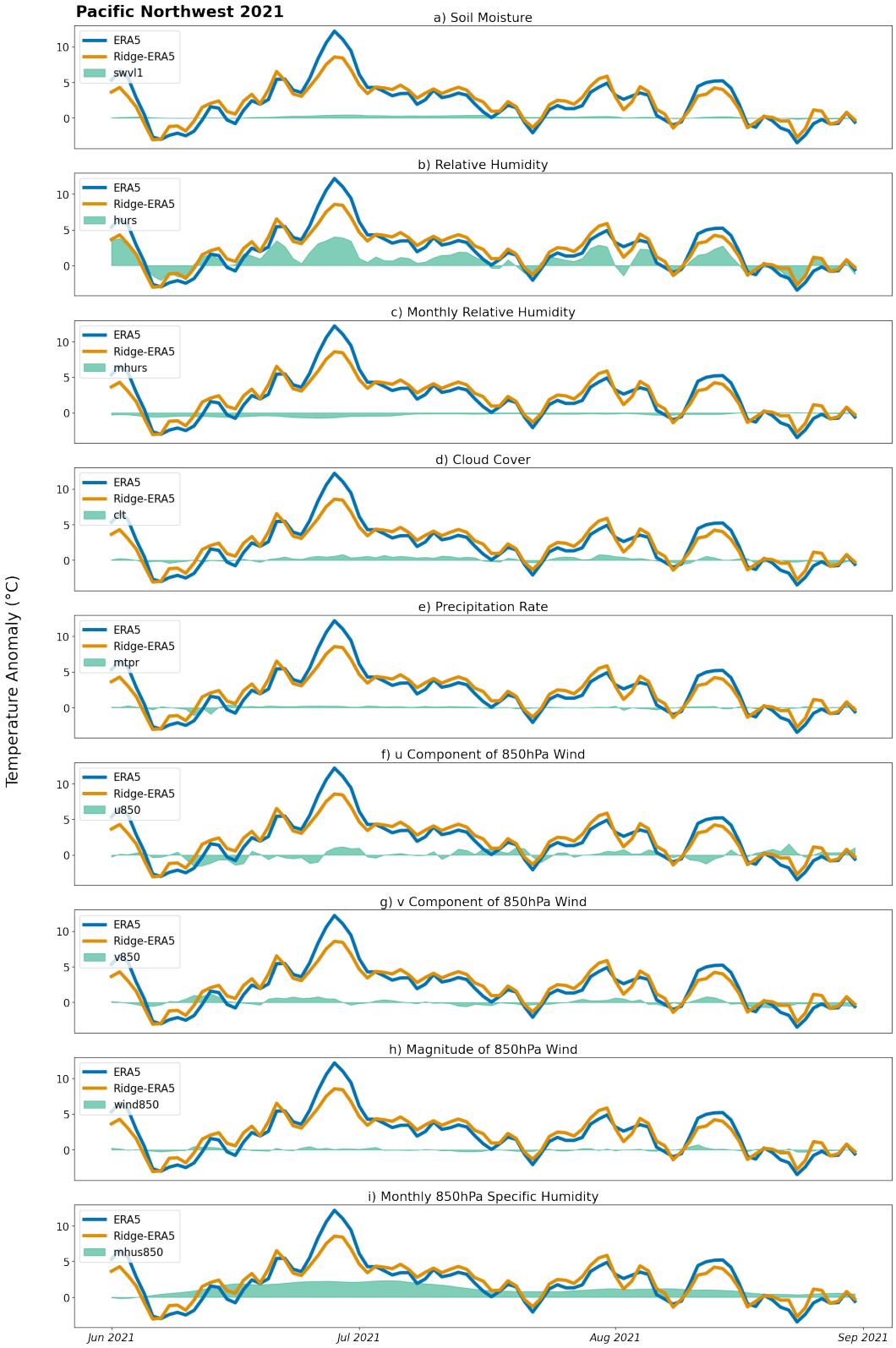

Figure 5.4: As in Figure 5.2 for the Pacific Northwest 2021 heatwave.

including for the hottest temperature anomalies in mid to late August.

The high temperatures observed in August have been attributed to: a period of quasi-stationary, anomalous high pressure over the region (Chen and Li, 2023) alongside increased surface radiative heating; adiabatic heating from subsidence; and warm air advection (He et al., 2023; Hua et al., 2023). As with previous case study heatwaves where warm air advection played a key role, there is a gradual build up in contributions from the 850hPa monthly specific humidity (Figure 5.5i) leading up to the height of the heatwave. Contributions from meridional flow (Jiang et al., 2023) are consistent with the temperature anomaly contributions from the v component of the 850hPa wind (Figure 5.5g) which remain positive throughout much of the summer possibly representing warm air advection from lower latitudes. Positive contributions from the u component of the 850hPa wind (Figure 5.5f) are consistent with anomalous atmospheric flow.

Land-atmosphere feedbacks also contributed to the heatwave with a soil drying trend throughout the summer resulting in unusually dry soils by August (Jiang et al., 2023). Soil moisture anomalies also contribute to the Ridge-ERA5 predictions (Figure 5.5a), particularly to the higher temperatures experienced throughout August. The large contributions from daily near-surface relative humidity (Figure 5.5b) may also be related to these land surface feedbacks.

5.2 Simulating Historical Heatwaves Using CMIP6 Climatologies

Next, as a model evaluation step, the ability of 15 Ridge-CMIP emulators trained on historical climate model data to reproduce the historical heatwaves presented in the previous subsection is tested. The Ridge-CMIP emulators are trained using historical climate model data to predict daily temperature anomalies to the seasonal cycle using the same set of predictor variables as Ridge-ERA5 (as described in Chapter 4). Predictor variable anomalies during the four heatwave events (Europe 2003, Europe 2018, the Pacific Northwest 2021 and China 2022) from ERA5 are de-biased and normalised relative to the variance of each CMIP6 model. These inputs can then be combined with Ridge-CMIP coefficients learnt from each CMIP6 model to reproduce the heatwave conditioned on the model climatology. Time series of these Ridge-CMIP simulations are plotted alongside Ridge-ERA5 predictions and the actual ERA5 temperature anomalies in Figure 5.6.

For the Europe 2003 heatwave (Figure 5.6a), Ridge-CMIP simulations show generally good agreement with the pattern of day-to-day variations of the actual ERA5 anomalies and the Ridge-ERA5 predictions. During the heatwave event itself from early to mid August, the Ridge-CMIP predictions follow a very similar pattern to the Ridge-ERA5 predictions with

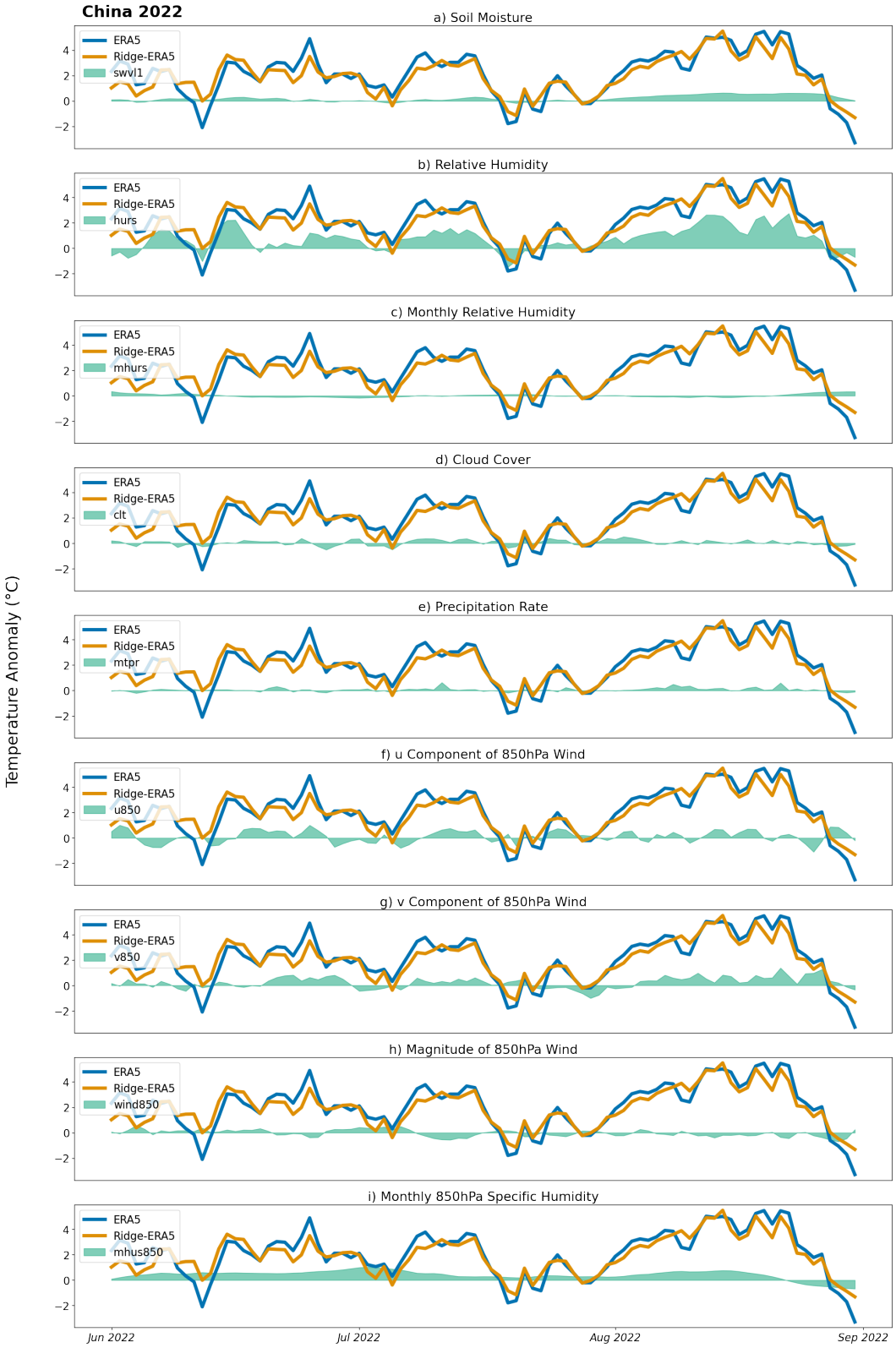

Figure 5.5: As in Figure 5.2 for the China 2022 heatwave.

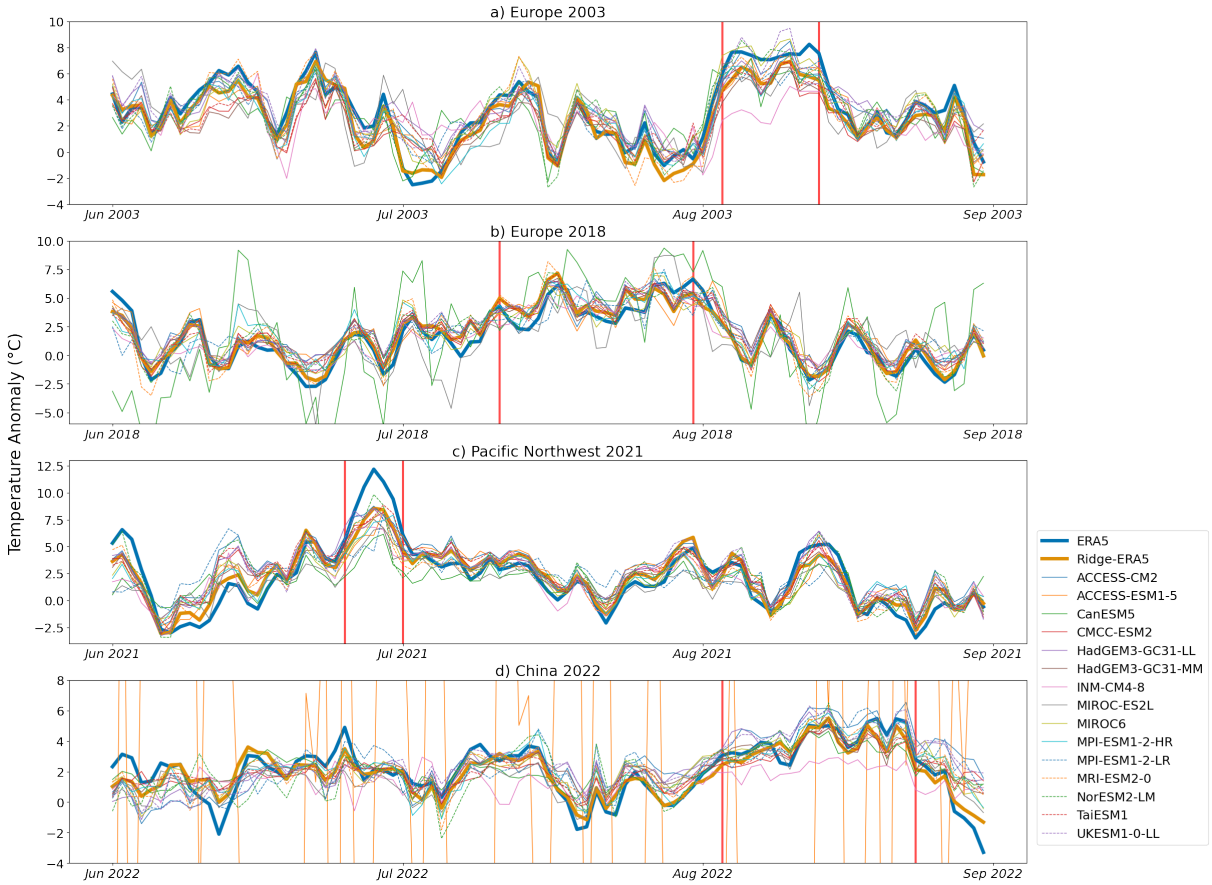

Figure 5.6: Ridge-ERA5 predictions (orange) plotted alongside actual ERA5 temperature anomalies (blue) with simulations from Ridge-CMIP emulators provided with ERA5 predictor anomalies for a) Europe 2003, b) Europe 2018, c) Pacific Northwest 2021 and d) China 2022. Vertical red lines define the duration of each heatwave event for later analysis.

some variation in the magnitude of the peaks predicted. The fact that these patterns match so well, but fail to exactly capture the true temperature anomaly time series is potentially indicative of missing processes which the Ridge models are collectively unable to capture or non-linearities in the processes that are represented. The predictions of the Ridge-INM-CM4-8 model stand out as outliers during the event itself, under-predicting the magnitude of the hot extremes by several degrees and exhibiting a unique pattern of day-to-day anomalies compared with other Ridge models.

To identify the cause of this difference, SHAP values for each Ridge-CMIP prediction are plotted with SHAP values from Ridge-ERA5 predictions for each predictor variable during the time periods of each heatwave (indicated by the vertical red lines in Figure 5.6) in Figure 5.7a. Breaking the predictions down in this way reveals that Ridge-INM-CM4-8 (pink circles in Figure 5.7a) is an outlier in terms of the contributions from the u component of the 850hPa wind and soil moisture. While other Ridge-CMIP emulators and Ridge-ERA5 produce a positive temperature anomaly based on this wind variable during the heatwave, Ridge-INM-CM4-8 actually produces a cooling effect, and similarly for soil moisture. Though other models do stand out as outliers for under-prediction in some predictor variables, *e.g.* Ridge-CanESM5 and Ridge-MIROC-ES2L, these effects are typically compensated for by higher predictions compared with the average in other variables, *e.g.* the v component of the 850hPa wind for Ridge-CanESM5. As with the analysis in Chapter 3, comparing the prediction of the Ridge-CMIP emulators with Ridge-ERA5 on a variable by variable basis can reveal compensating errors. The final temperature anomaly predictions of Ridge-CanESM5 appear to match quite closely with Ridge-ERA5 even though individual variable contributions look very different.

Ridge-CMIP simulations of the Europe 2018 heatwave (Figure 5.6b) perform similarly well to Ridge-ERA5 predictions, they largely capture the extent and pattern of the temperature anomalies. The most notable outlier is the Ridge-CanESM5 predictions which produce peaks in temperature an order of magnitude larger than any other Ridge-CMIP emulator or the Ridge-ERA5 predictions. Breaking the Ridge-CMIP simulations down into SHAP value contributions from each predictor variable and comparing them with ERA5 (see Figure 5.7b) the source of this over-sensitivity seems to be the u and v components of the 850hPa wind for which Ridge-CanESM5 has the largest contributions and also monthly near-surface relative humidity.

For the Pacific Northwest in 2021, the Ridge-CMIP simulations generally follow a similar pattern to the Ridge-ERA5 predictions (Figure 5.6c) - they match the true temperature anomaly time series well for most of the summer but fail to capture the magnitude of the heatwave days at the end of June. Decomposing the Ridge outputs into contributions from each predictor variable (Figure 5.7c), Ridge-CMIP and Ridge-ERA5 SHAP values are comparable

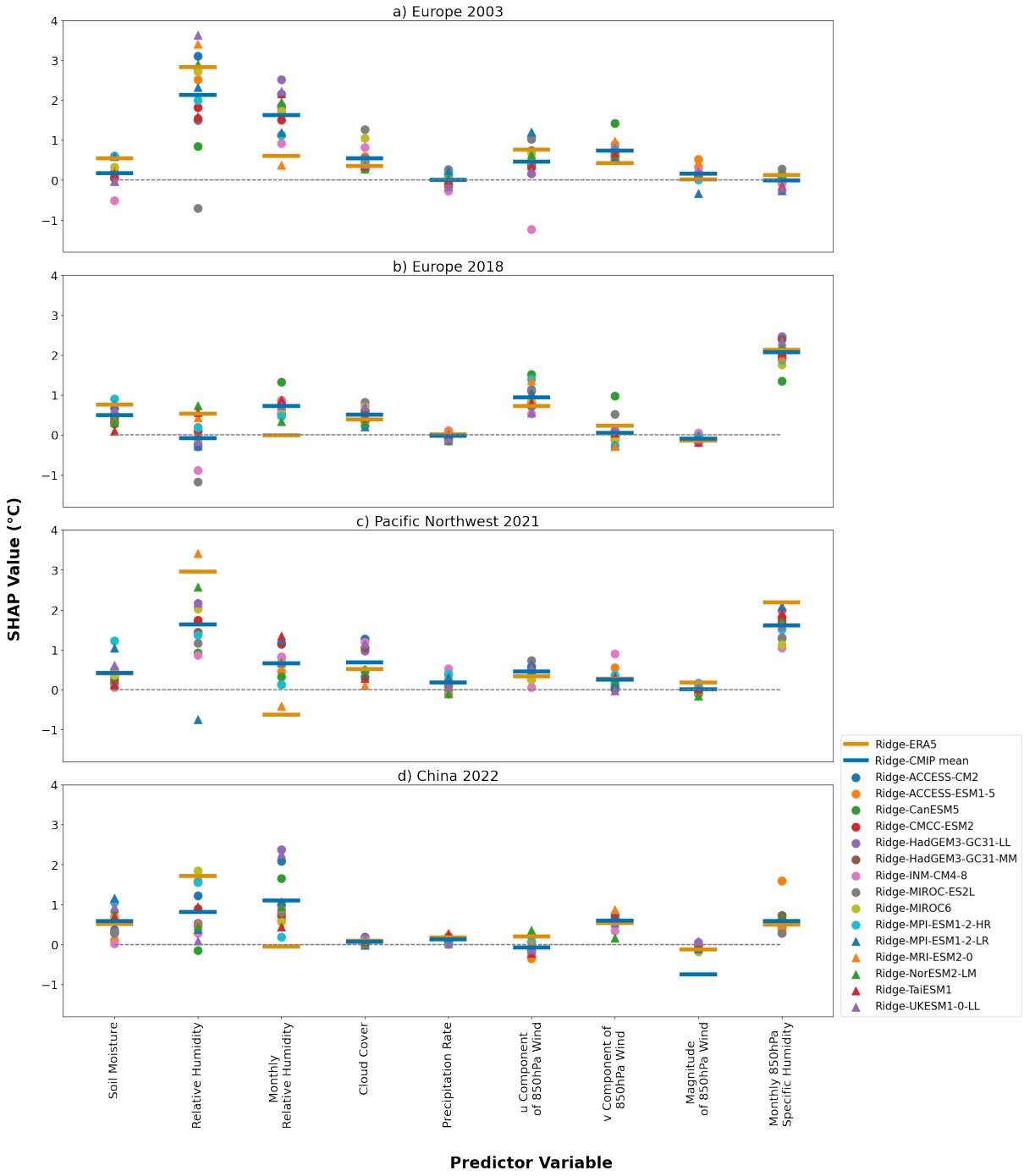

Figure 5.7: SHAP value contributions from each predictor variable to temperature anomaly predictions from Ridge-ERA5 (orange line), individual Ridge-CMIP (points) and Ridge-CMIP mean (blue line) during the Europe 2003 (a), Europe 2018 (b), Pacific Northwest 2021 (c) and China 2022 (d) heatwaves. Note that for 850hPa wind magnitude during the China 2022 heatwave, the contribution for Ridge-ACCESS-ESM1-5 is at $-10°C$ which skews the Ridge-CMIP mean (blue) downwards.

for most variables the notable exceptions being the daily and monthly near-surface relative humidity and the 850hPa monthly specific humidity. The relative humidity contributions for Ridge-ERA5 seem to cancel each other out to some extent with negative contributions from monthly relative humidity and a daily relative humidity contribution higher than most Ridge-CMIP models. The monthly 850hPa specific humidity contribution does make a difference for Ridge-ERA5, making the final Ridge-ERA5 temperature anomaly predictions amongst the highest predicted by any of the Ridge models - and therefore closer to the true peak - and certainly hotter than the Ridge-CMIP multi-model mean.

The Ridge-CMIP simulations of the China 2022 heatwave (Figure 5.6d) produce similar results to the simulations of the Europe 2018 heatwave (Figure 5.6b). In general, Ridge-CMIP simulations agree fairly closely with Ridge-ERA5 predictions of the event which themselves capture the true temperature anomaly time series reasonably well. The Ridge-INM-CM4-8 simulations are also outliers here, again resulting in an under-prediction of the warmest days throughout August and in mid-July. The primary outlier in this case though is the Ridge-ACCESS-ESM1-5 simulations, this clear sensitivity issue appears to come from the contribution of the magnitude of the 850hPa wind. In Figure 5.7d, the Ridge-CMIP mean (blue line) is skewed lower for the 850hPa wind magnitude by the Ridge-ACCESS-ESM1-5 contribution of ≈ -10. The lower predictions of Ridge-INM-CM4-8 appear to result from a weaker soil moisture response in comparison with other models. As with the Ridge-ERA5 predictions of the Pacific Northwest heatwave, there also seems to be a compensatory effect between Ridge-ERA5 contributions from daily and monthly near-surface relative humidity.

Overall, outlier predictions from Ridge-CMIP emulators models can most frequently be attributed to the dynamical 850hPa wind variables and to soil moisture differences. For both European heatwaves, soil moisture responses in the CMIP6 models appear to be too weak in comparison with Ridge-ERA5. This apparent weakness in the soil moisture response may also be related to the under prediction of daily near-surface relative humidity contributions compared with Ridge-ERA5 contributions, although this difference is common to all four heatwaves. At the same time, contributions from the monthly near-surface relative humidity tend to be much stronger in the climate models. This case study analysis is extended in the next subsection where SHAP value analysis is applied to compare common characteristics of 95^{th} percentile events in models and reanalysis. Additionally, comparisons are also made between extreme events in historical and future GCM climates.

5.3 Representations of Heatwaves in Climate Models *vs.* Reanalysis

Following on from the specific case studies in the previous subsections, next average SHAP values across all extreme events are plotted - here defined as daily temperature anomalies exceeding the 95^{th} percentile at a particular grid cell during JJA. This allows a consistent definition to be applied across the Northern Hemisphere and provides a reasonable chance for the ML model, Ridge-ERA5, to make reasonable predictions for 'extremes' given the volume of training data available. Percentiles are calculated over the complete reanalysis time period 1979-2022. These extreme SHAP values are compared between Ridge-ERA5 and Ridge-CMIP emulators as well as between historical Ridge-CMIP simulations and simulations under future warming scenario *SSP585*.

5.3.1 Heatwaves in the Present Day Climate

This subsection begins with an analysis of average SHAP value contributions to extreme temperature anomalies (days exceeding the 95^{th} percentile) according to Ridge-ERA5 predictions, followed by a comparison of these values between Ridge-ERA5 and Ridge-CMIP models. Figure 5.8 shows maps of Ridge-ERA5 SHAP value contributions to extreme events from ERA5 during JJA 1979-2022.

The largest magnitude contributions in general come from the daily near-surface relative humidity (Figure 5.8b) which broadly has a warming effect associated either with advection of warm moist air in *e.g.* Eastern North America, or relative humidity deficits over drier regions (Russo et al., 2017). There are however some regional exceptions to this rule, where daily near-surface relative humidity anomalies result in cooling. For example, over Northern Eurasia there is a band of cooling associated with relative humidity anomalies during extreme heat events. In this region, relative humidity coefficients are positive meaning that negative relative humidity anomalies are resulting in a prediction of cooling by Ridge-ERA5. These negative relative humidity anomalies are likely associated with soil moisture deficits and clear sky radiative heating which produces hot and dry conditions. With these events being extreme and therefore infrequent, it is potentially the case that this behaviour is not captured by Ridge-ERA5. The coefficient values are dominated by mean state behaviour where increased relative humidity is associated with warm, moist air advection from the Tropics and so the modelled relationship is positive while during extreme events the opposite is in fact true. This may also help to explain why performance of the Ridge-ERA5 model was generally poorer in this region.

Monthly relative humidity (Figure 5.8c) on the other hand provides much smaller con-

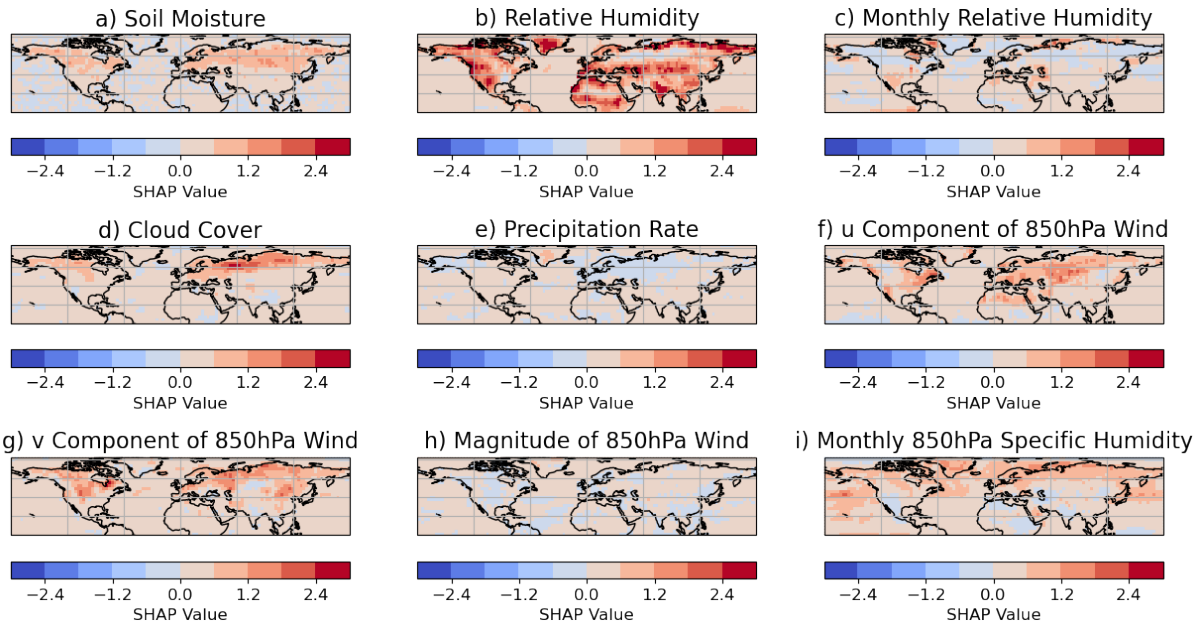

Figure 5.8: Maps of average SHAP contributions from each predictor variable to Ridge-ERA5 daily temperature anomaly predictions for events which exceed the local 95^{th} percentile.

tributions to extreme temperature anomalies with some notable warming over drier regions such as the Arabian Peninsula and Mexico. This may be related to land-surface feedbacks and soil moisture availability. Monthly 850hPa specific humidity (Figure 5.8i) contributions are also quite small, and are associated with a small warming contribution in most regions. As with the coefficient magnitudes themselves, the monthly mean variables provide less direct information about day-to-day fluctuations and are therefore likely to offer smaller contributions.

Soil moisture interactions (Figure 5.8a) contribute positive anomalies of $\approx 1^oC$ on average to mid to high latitude heat extremes attributable to land-atmosphere feedbacks but makes less significant contributions elsewhere. Similarly cloud cover anomalies (Figure 5.8d) contribute to warming at higher latitudes, potentially linked with blocking effects and clear sky radiative heating, but contribute in smaller magnitudes in other regions even producing a small cooling effect in South East Asia. Precipitation rates (Figure 5.8e) are one of the smallest contributors with a small cooling effect visible at higher latitudes, potentially as a result of moisture supply for evaporative cooling and little contribution elsewhere.

The v component of the 850hPa wind (Figure 5.8g) contributes positively to extreme temperature anomalies in most regions, particularly at higher latitudes. This is likely linked to transport of warmer air from lower latitude regions. The u component of the 850hPa wind (Figure 5.8f) is also broadly positive, particularly at mid to high latitudes, likely attributable

to atmospheric anomalies associated with blocking conditions. Total 850hPa wind magnitude (Figure 5.8h) is a relatively small contributor and primarily seems to result in cooling. Some of these cooling patterns along coastlines could be associated with an onshore breeze.

Next, to compare the representation of extreme values learnt from reanalysis with those that it is possible to learn from climate models, SHAP contributions from each variable for 95^{th} percentile events in Ridge-CMIP and Ridge-ERA5 are plotted for comparison, see Figures 5.9 and 5.10.

As seen in Figure 5.8, precipitation rate (Figure 5.9e) tends to contribute the least and this is common across the Ridge-CMIP emulators and Ridge-ERA5. Soil moisture anomalies (Figure 5.9a) contribute most at mid-high latitudes across Europe, North America and East Asia. In Europe and North America, the soil moisture contributions from Ridge-ERA5 are larger than the average SHAP value contribution from most Ridge-CMIP emulators suggesting that soil moisture feedbacks with temperature may be under-sensitive in the Ridge-CMIP emulators. Cloud cover contributions (Figures 5.9d) show a similar pattern with the greatest contributions to extreme heat events in Europe, North America and East Asia. Again, for European regions, this effect is generally under-simulated by the relationships learnt from CMIP6 data.

Contributions of daily near-surface relative humidity (Figures 5.9b) and monthly near-surface relative humidity (Figures 5.9c) vary on a regional basis but the average Ridge-ERA5 SHAP value contribution generally falls close to the Ridge-CMIP mean, although for some regions, *e.g.* West & Central Europe, the range of contributions across Ridge-CMIP emulators is quite large. The only notable exception to this is Central North America where the Ridge-CMIP mean contribution for daily relative humidity exceeds the Ridge-ERA5 contribution by $\approx 1.5^oC$ and the range of all Ridge-CMIP emulators excludes the Ridge-ERA5 contribution entirely.

The influence of dynamical anomalies as measured by the u (Figures 5.10a) and v (Figures 5.10b) components of the 850hPa wind is generally strongest in Europe, Eastern North America and East Asia where atmospheric anomalies have frequently been linked with heatwave events (Section 1.2). For the 850hPa wind magnitude (Figure 5.10c), contributions are small across all regions and contributions between Ridge-ERA5 and Ridge-CMIP models are closely matched. Finally, for the 850hPa monthly specific humidity (Figure 5.10d) contributions from Ridge-CMIP emulators are generally higher than from Ridge-ERA5. This is particularly clear in South America, the Sahara, the Arabian Peninsula, Central and North Eastern Africa, mostly relatively dry regions. Given that the monthly 850hPa specific humidity gives an indication of background warming, this seems to indicate a general over-sensitivity of the CMIP6 models, particularly for CanESM5 which is a notable outlier in the Tibetan Plateau region.

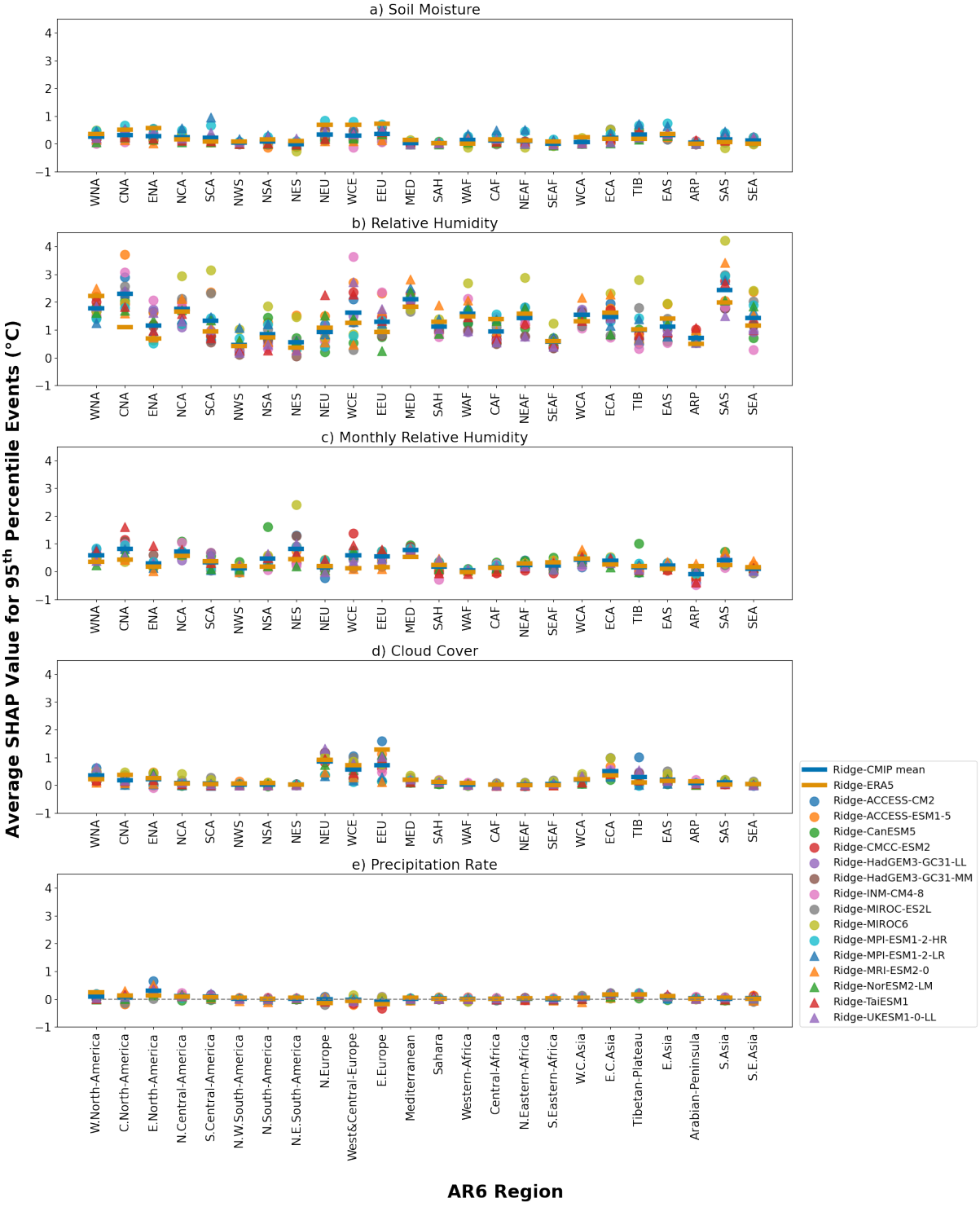

Figure 5.9: Average SHAP values contributing to 95^{th} percentile events in Ridge-ERA5 (orange line), individual Ridge-CMIP models (points) and the Ridge-CMIP mean (blue line) for each predictor variable: soil moisture (a); daily relative humidity (b); monthly relative humidity (c); cloud cover (d); and precipitation (e).

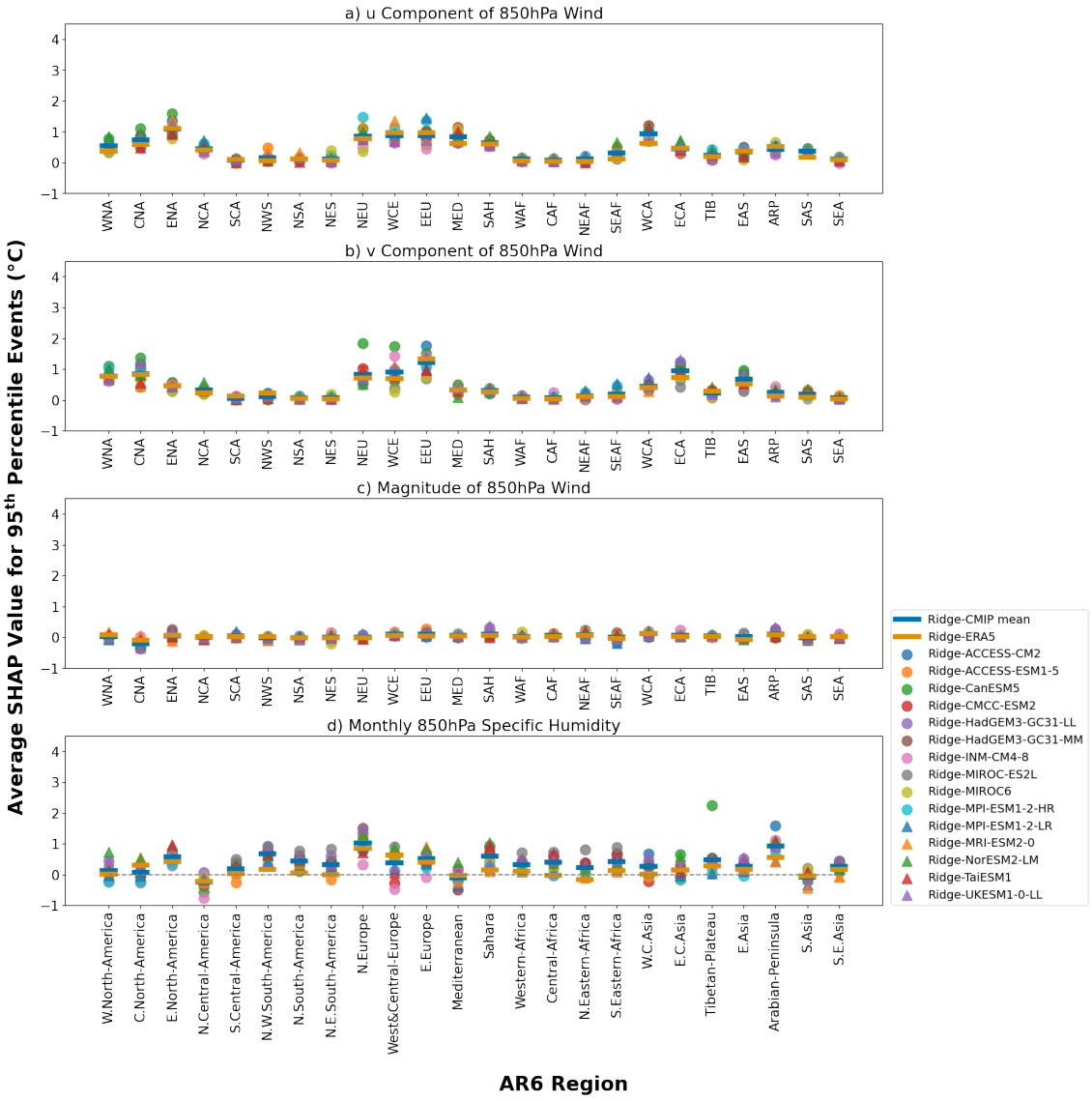

Figure 5.10: As in Figure 5.9 for: u component of the 850hPa wind (a); v component of the 850hPa wind (b); magnitude of the 850hPa wind (c); and monthly 850hPa specific humidity (d).

5.3.2 Heatwaves in Future Climates

The same analysis applied in the previous subsection to compare the representation of extreme events in reanalysis and climate model data in the present day climate is used here to analyse how heatwave events are represented in future climates in GCMs. Figures 5.11 and 5.12 show the SHAP value contributions to Ridge-CMIP predictions of 95^{th} percentile events in the historical climate (1979-2022) compared with 95^{th} percentile events in the future climate (2070-2100) under future warming scenario *SSP585*. The grey points and grey lines show the average contributions of each Ridge-CMIP emulator and the Ridge-CMIP mean to 95^{th} percentile daily-mean temperature anomalies in the historical climate (1979-2022). The coloured points and blue lines show the average SHAP value contributions to Ridge-CMIP predictions of daily-mean temperature anomalies between 2070 and 2100 which exceed the 95^{th} percentile of projected temperatures in that time range (2070-2100).

Contributions from soil moisture (Figure 5.11a) look the same in many regions but show an increase in Europe and North America. Under this intense future warming larger soil moisture anomalies result in large soil moisture SHAP value contributions to heatwave predictions. Cloud cover SHAP values (Figure 5.11d) show a similar pattern in terms of increasing contributions in Europe and North America. Monthly relative humidity contributions (Figure 5.11c) exhibit an increase in most regions, an increase which is particularly large over Europe. This may be linked in part with the increased soil moisture contributions also found in Europe pointing to generally drier conditions. Daily relative humidity (Figure 5.11b) generally looks quite similar with the future Ridge-CMIP mean (blue line) falling within the range of historical contributions (grey points), especially considering that this variable shows the greatest spread historically. The regional variations in daily relative humidity contributions are consistent with the fact that this variable represents different processes in different regions as discussed in the previous subsection.

The largest contribution to the warmer temperature anomalies by far comes from the monthly 850hPa specific humidity (Figure 5.12d). This is indicative of the background warming state by the end of the century. Ridge-CanESM5 stands out as an outlier here predicting the largest contributions from this variable likely because CanESM5 is the highest sensitivity CMIP6 model. The dynamical variables (Figure 5.12a,b,c) generally show very little change between 1979-2022 and 2070-2100.

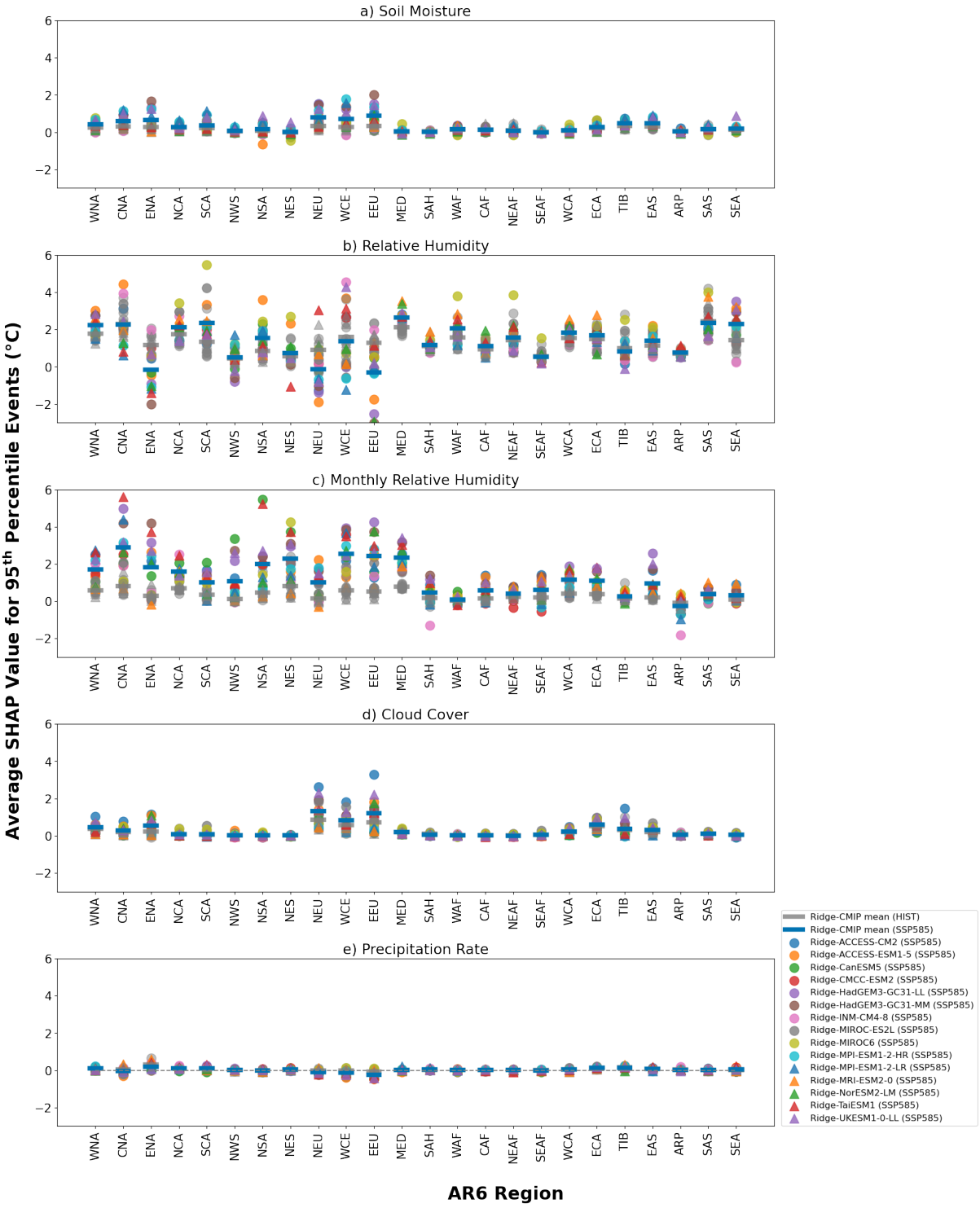

Figure 5.11: Average SHAP values contributing to *SSP585* 95^{th} percentile events for individual Ridge-CMIP models (points) and the Ridge-CMIP mean (blue line) for each predictor variable: soil moisture (a); daily relative humidity (b); monthly relative humidity (c); cloud cover (d); and precipitation (e).

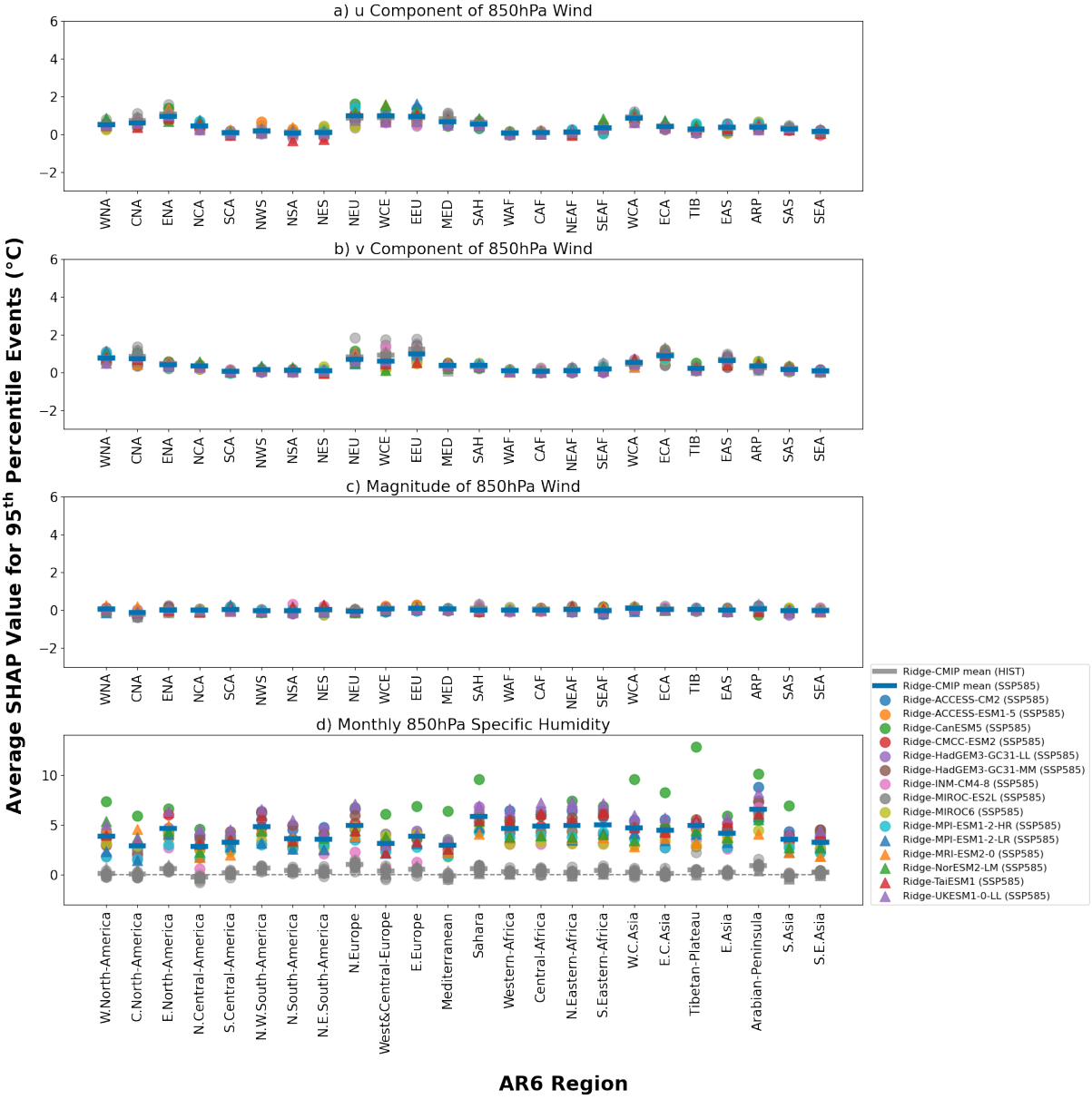

Figure 5.12: As in Figure 5.11 for: u component of the 850hPa wind (a); v component of the 850hPa wind (b); magnitude of the 850hPa wind (c); and monthly 850hPa specific humidity (d).

5.4 Conclusions

In this chapter, Ridge-ERA5 predictions of four historical heatwave events are evaluated. While Ridge-ERA5 captures the patterns of the temperature anomalies observed in these cases well, the magnitude of the anomalies is under-predicted in some instances, for example the Pacific Northwest heatwave of 2021. This is potentially explainable by the rareness of the event; given that the temperatures recorded at that location were unprecedented in the historical record, Ridge-ERA5 was not exposed to an event of such magnitude in that region during training. This potentially has implications for the ability of Ridge-ERA5 to extrapolate and make predictions under future climate change, although it is important to place the magnitude of these temperature anomalies in context. The height of the Pacific Northwest heatwave saw temperature anomalies of $> 12^oC$, while CMIP6 models project an average future warming range of $4-9^oC$ in the region. However, as shown in Chapter 4, Ridge predictions were able to accurately capture daily temperature anomalies of $10-12^{\circ}C$ across different regions under end-of-century, high emission scenario warming conditions. Another potential explanation to reconcile this fact is intrinsic non-linearities associated with the driving forces of this particular heatwave which prevented Ridge-ERA5 from being able to capture the full magnitude of the temperature anomalies. Ridge-ERA5's performance on the Pacific Northwest heatwave certainly highlights potential areas for improvements in the ML model which are discussed in more detail in Chapter 6.

The Ridge-ERA5 predictions for historical heatwaves were then decomposed into SHAP value contributions from each predictor variable. Generally speaking, the contributions from different variables could be aligned with existing literature to explain heatwave drivers. However, it should be noted that, due to the coupled nature of the climate system, the relationships modelled by Ridge-ERA5 are not necessarily independent, causal relationships. So although the contribution of each predictor variable can be calculated precisely, interpretation of this value in terms of physical mechanisms can still be challenging. This is similar to the case of existing GCMs, although the SHAP value calculations presented here are much cheaper computationally, as is the training process for Ridge-ERA5, and Ridge-ERA5 can learn directly from observations-based data for comparison.

Another advantage of Ridge-ERA5 in terms of computational costs is that coefficients can easily be combined with climate model output. To reproduce the four case study heatwaves in a climate model world, Ridge-ERA5 inputs are combined with Ridge-CMIP emulator coefficients (Ridge-CMIP models are trained on historical climate model data). These Ridge-CMIP emulator predictions of historical heatwaves generally show close agreement with the Ridge-ERA5 predictions. The fact that predictions look so similar between the Ridge-CMIP emulators and Ridge-ERA5 indicates that, in the instances where the temperature anomalies

are under-predicted, there may be some systematically excluded or poorly simulated processes *e.g.* non-linearities that Ridge is unable to reproduce. Some notable outliers from this analysis include Ridge-INM-CM4-8, which under-predicts the magnitude of both the Europe 2003 heatwave and the China 2022 heatwave. Analysis of SHAP value contributions reveals the source of this under-prediction to be soil moisture and wind sensitivities which are not compensated for by other variables in the same way as other Ridge-CMIP models. There were also some instability issues with combining ERA5 inputs with Ridge-CMIP coefficients in some cases, CanESM5 for Europe 2018 and ACCESS-ESM1-5 for China 2022: this is traced back to incompatibilities in the 850hPa wind variables, these differences would need to be investigated in more detail to identify if, for example similar patterns are modelled but these patterns are displaced in some models, the magnitude of these feedbacks is different in the listed models or a combination of factors, see Chapter 6 for further discussion of extensions and future work.

Continuing the comparison of Ridge-CMIP emulators with Ridge-ERA5 for representing extreme events, analysis was extended to compare the representation of all temperature anomalies exceeding a local 95^{th} percentile threshold. These predictions were decomposed into SHAP value contributions for comparison. Humidity variables, soil moisture, cloud cover and u and v components of wind were found to be the largest contributors to these extreme heat anomalies generally. In most regions there was good agreement between the magnitude of contributions from Ridge-ERA5 and the Ridge-CMIP emulators. Some exceptions include weaker contributions from soil moisture and cloud cover from Ridge-CMIP emulators in Europe and North America. Contributions from monthly 850hPa specific humidity seem symptomatic of generally higher climate sensitivity in CMIP6 models, and the range of daily relative humidity anomalies contributed to much of the spread between models.

Comparing contributions of Ridge-CMIP predictor variables to historical extreme temperature anomalies with SHAP values for predictions of end of century (2070-2100) extremes, a clear increase in the background warming state is visible in the monthly 850hPa specific humidity. This is particularly clear for higher sensitivity models, for example CanESM5, as would be expected. Since the soil moisture response was identified as being relatively weaker in the Ridge-CMIP models compared with Ridge-ERA5 over Europe and North America historically, increased contributions from soil moisture over these regions in the future indicates larger soil moisture anomalies in the future. However, questions are raised about the time of emergence of the soil moisture signals in these regions if the modelled response is in fact weaker than what is historically observed. Cloud cover anomalies are also observed to increase contributions to extreme heat in both Europe and North America, which needs to be recognised in terms of the importance of improving understanding and representation of cloud processes in climate models.

6 Conclusion

This chapter summarises the results presented in this thesis, placing the conclusions in context and providing suggestions for extensions and future work.

6.1 Summary

In this thesis, a novel method has been developed to merge existing climate model output with observations via the use of ML techniques. A Ridge regression model, Ridge-ERA5, is trained on reanalysis data to predict daily summertime (JJA) temperature anomalies in the Northern Hemisphere on a grid cell by grid cell basis. To make these predictions, the model is provided with information about the local soil moisture anomaly, and regional domains of daily near-surface relative humidity, cloud cover, precipitation rate, u and v components of the 850hPa wind, magnitude of the 850hPa wind and monthly values of near-surface relative humidity and 850hPa specific humidity. Testing of the model on held out data from ERA5 produces R^2 scores of > 0.7 across most Northern Hemisphere land regions and plotting predicted *vs.* actual anomalies on a region by region basis shows that the model performs best in Western, Central & North Eastern Africa, Central America, Northern South America and South Asia. Regions of poor performance at higher latitudes are likely due to the absence of variables representing snow or albedo effects. Advantages of applying Ridge regression for this purpose are its interpretability as a simple linear model, relative capacity to cope with co-linear predictors and ability to extrapolate beyond the range of temperatures seen during training. The following subsections reflect on the application of Ridge-ERA5 for: bias correction; climate model evaluation; constraining future uncertainty; and understanding extreme events.

6.1.1 Ridge-ERA5 for Bias Correction

In Chapter 3, Ridge-ERA5 is applied for bias correction of existing climate model output. Coefficients from Ridge-ERA5 are combined with predictor variables from CMIP6 models to produce constrained simulations of historical temperature anomalies which are moderated by the process-based relationships learnt by Ridge-ERA5 from reanalysis.

The bias correction results in improved temperature anomaly distributions across many Northern Hemisphere AR6 regions, typically narrowing the distribution of CMIP6 temperature anomalies which tend to over-estimate variability. The Ridge-ERA5 constraint performs competitively with a traditional variance scaling approach, even out-performing variance scaling in several regions. This is an interesting and impressive result considering that the bias

correction which arises from applying the Ridge-ERA5 coefficients is purely derived from the process-based approach to modelling temperature anomalies. Ridge-ERA5 does not learn a direct relationship between climate model output and observations as with other bias correction approaches.

In regions where the bias correction fails (Northern, West & Central and Eastern Europe), the constraint results in an over-narrowing of the temperature anomaly distribution such that variability is under-represented compared with the ERA5 distribution. The source of this issue was identified as differing representations of near-surface relative humidity and its relationship with temperature between the CMIP6 models and ERA5. While further investigation may be required to identify the underlying cause behind these differences, a similarity check between Ridge-CMIP and Ridge-ERA5 coefficients is suggested before application of Ridge-ERA5 for bias correction. A key advantage of this method of bias constraint is that the constraint itself can be decomposed into contributions from each predictor variable in order to identify compensating biases.

6.1.2 Ridge-ERA5 for Climate Model Evaluation

In order to apply Ridge-ERA5 for climate model evaluation, a set of Ridge-CMIP emulators were utilised. The Ridge-CMIP ML models learnt to emulate the climate and processes of respective CMIP6 models so that their coefficients could be compared with those learnt by Ridge-ERA5 as a means of model evaluation.

Plotting maps of these coefficient differences revealed unique fingerprints of model performance. Patterns in these maps indicated correlations with regions of greater coefficient difference and regions where the Ridge-ERA5 bias correction failed. Comparing the bias corrected temperature anomalies (Ridge-ERA5 coefficients and CMIP6 predictor anomalies) with historical Ridge-CMIP predictions (Ridge-CMIP coefficients and CMIP6 predictor anomalies) allowed compensating errors to be identified. Employing SHAP value analysis to investigate the individual contributions of predictor variables revealed that applying Ridge-ERA5 coefficients to CMIP6 soil moisture anomalies resulted in an upward correction at the same time as a downward correction of 850hPa wind contributions. This suggests that a weak soil moisture feedback in CMIP6 models (Qiao et al., 2021; Talib et al., 2023) may be compensated for by stronger dynamical effects.

A summarising metric of total coefficient difference was plotted against projected end of century warming under future scenario *SSP585* on a variable by variable basis. Aside from a potential correlation between precipitation coefficient difference and warmer end of century projections, no clear patterns emerge. This apparent lack of relationship between historical model skill (by this metric) and future warming motivate the application of Ridge-ERA5 to

construct an observations-based constraint.

6.1.3 Ridge-ERA5 for Constraining Future Uncertainty

In Chapter 4, Ridge-ERA5 was applied to construct a novel, observational constraint on future regional warming projections during Northern Hemisphere summer. Initially, the Ridge-CMIP emulator models were applied to make end of century (2070-2100) temperature predictions which could be compared with the actual CMIP6 projections. This setup was designed to test the climate invariance of the coefficients learnt by Ridge from historical data and confirm that they still had predictive power in the future under climate change. This test was applied using data from the *SSP585* future warming scenario which represents the most extreme extrapolation case and therefore the most challenging test for Ridge. In general, the Ridge-CMIP models perform well capturing temperature anomalies which exceed the local, historical average by more than 10^oC.

Next, Ridge-ERA5 coefficients were combined with future values of predictor variables from scenario *SSP585* in CMIP6 in order to make future predictions conditioned on the observations-based relationships learnt by Ridge-ERA5. These predictions formed the basis of a novel constraint on future regional warming.

The constraint itself tends to exclude those CMIP6 models which project the greatest degree of warming, this has implications for the climate sensitivity of these models (Jiménez-de-la Cuesta and Mauritsen, 2019; Nijsse et al., 2020; Tokarska et al., 2020; Zhu et al., 2020). The greatest downshifting of the expected range of future temperatures tends to be associated with drier regions (*e.g.* the Sahara, Northern Central America, and the Arabian Peninsula) suggesting that CMIP6 models temperature responses may be too sensitive in these regions. The greatest up-shift in the expected temperature range is found in South Central America and South East Asia where sensitivity of climate models seems to be weaker compared with the relationships learnt by Ridge-ERA5.

A SHAP value analysis to decompose the Ridge-ERA5 constrained temperature anomalies into contributions from each predictor variable indicates that the main sources of the constraint are daily near-surface relative humidity and monthly 850hPa specific humidity. As an indicator of the background warming state, the monthly 850hPa specific humidity is likely to be closely tied to model climate sensitivity, and multiple processes to be constrained playing into this relationship.

6.1.4 Understanding Extremes with Ridge-ERA5

In Chapter 5, Ridge-ERA5 is applied specifically to extreme events to analyse the driving factors contributing to historical heatwaves and compare the representation of heat extremes in climate models and reanalysis. In terms of reproducing historical heatwaves, Ridge-ERA5 performs relatively well capturing patterns of warming and the magnitude of both the Europe 2018 and the China 2022 heatwaves. However, the magnitude of rarer events is sometimes under-simulated, for example the unprecedented Pacific Northwest heatwave of 2021.

In terms of interpreting drivers from the SHAP values, broad agreement can be drawn between SHAP values for each predictor variable and drivers previously identified in existing literature. However, care must be taken not to over-interpret these SHAP value contributions. Given the coupled nature of the climate system, these coefficients do not represent direct causal relationships but rather correlations which can be linked back to physical mechanisms. The difficulty is disentangling these relationships is common to traditional climate modelling methods as well, GCMs are also limited to some degree in their capacity for interpretation. The advantages of Ridge-ERA5 here are the computational cheapness, the ability to simply decompose the temperature predictions into contributions from predictor variables and the ability to learn directly from observations which traditional climate models cannot.

Comparison of average contributions from each predictor variable to 95^{th} percentile temperature anomalies between Ridge-CMIP and Ridge-ERA5 models showed generally good agreement. However, contributions from soil moisture and cloud cover were weaker in Ridge-CMIP models over most of Europe and North America. Increased contributions in both of these variables were also found when comparing Ridge-CMIP predictions of future heatwaves (2070-2100) with historical heatwaves (1979-2022). This is important to highlight as the future soil drying trend has a direct impact on future heat extremes and simulation of cloud processes is still a source of uncertainty in GCMs.

6.2 Extensions and Future Work

The following subsections explore how the results presented in this thesis for different applications of the Ridge-ERA5 method could be extended.

6.2.1 Changes to the Ridge-ERA5 ML Model

This subsection considers changes to the ML method itself in terms of both alternative predictor variables and alternative ML algorithms as well as further analysis of the limitations of the existing Ridge-ERA5 approach. Ridge was initially selected as the ML method of choice

due to its relative simplicity, ability to manage co-linear predictor variables and to extrapolate beyond the range of temperatures seen during training. Ridge is also computationally cheap and simple to train with only one hyperparameter for tuning making application of the method across every Northern Hemisphere grid cell relatively simple. However, there are limitations to Ridge regression to consider as well as the effects of choices made during the training process.

Further steps could be taken to test the limitations of the Ridge-ERA5 ML approach and the knock-on effects of training and validation choices on the final results. For example further testing of different train-test-validation splits, with specific regard to the time series / sequential nature of the data being used. Taking the test period from different points in the time series or spread across the time series for example, or leaving larger time gaps between a historical training set and more recent test set (or vice versa) to further test the robustness of the relationships being learnt by Ridge-ERA5. These kinds of limitations could also be further tested using the Ridge-CMIP model emulators where much longer time series of data are available as well as alternative warming scenarios, for example quadrupling CO_2. Naturally any of these changes which effect the final trained Ridge model are likely to have implications for the final results in terms of bias correction, future uncertainty constraint etc. The extent of these changes would largely be dependent on the robustness of the Ridge-ERA5 coefficients, i.e. how much the coefficients vary between the different choices of training data described above.

A key tuning factor in Ridge regression is the regularisation parameter which controls the contribution of the final term in the Ridge cost function and acts to distribute predictive power amongst correlated input variables by penalising the squared sum of coefficient magnitudes. In other applications, steps are sometimes taken to artificially reduce the value of the regularisation parameter below the "natural" limit determined during cross-validation. Such decisions may be motivated by a desire to smooth the map of coefficients learnt (e.g. Sippel et al., 2020). In this instance with Ridge-ERA5, the optimum regularisation parameter value as determined by the 5-fold cross validation is selected. Choices to select a larger regularisation parameter value may reduce total difference between Ridge-ERA5 and Ridge-CMIP coefficients as no single coefficient would remain as large in magnitude. Depending on how far the value of regularisation parameter is limited, there is also likely to be an impact on Ridge-ERA5 predictive skill with a greater move away from the optimum regularisation parameter value resulting in a more notable decrease in predictive performance.

Naturally as a linear model, Ridge may fail to capture certain processes accurately, for example non-linear soil moisture regimes. This motivates further consideration of alternative ML approaches, particularly given that interpretable AI methods such as SHAP value analysis enable easier interpretation of more complex algorithms. Experiments with Gaussian Process

Regression (GPR) methods could be taken further to consider more complex combinations of kernels and different kernel setups for different geographical regions in order to improve the performance seen during initial testing. The kernel used in initial testing here was the radial basis function also known as the squared exponential kernel. An alternative to this would be a rational quadratic kernel which is simply equivalent to adding multiple radial basis functions with varying lengthscale parameters together. Kernels can also be combined to produce more complex functions by addition or multiplication, this can allow different characteristics of the training data to be captured. For example, a linear kernel which is simply a straight line function and a periodic kernel which is a regularly repeating pattern may be combined by multiplication to produce a function which is periodic with increasing amplitude. Without consideration of time constraints, more complex ML methods could also be experimented with, for instance Convolutional Neural Networks (CNNs), which may be able to model the non-linear relationships that Ridge struggled to capture. Naturally this increase in model complexity comes with additional computational costs and a prolonged training and fine tuning process.

Another avenue for exploration would be to train the ML model with alternative reanalysis data sets, for example JRA55 (Kobayashi et al., 2015). Training the same model at a given location with multiple reanalysis data sets would also allow for a quantitative measure of the degree of uncertainty in the learnt model parameters by comparing coefficients across reanalysis data sets. This would also have implications for the future uncertainty constraint which are discussed in Section 6.2.3. Another potential option would be to use actual observations for model training. Although this was not possible with the variables required on a hemispheric scale for the results in this thesis, on a smaller scale it may be viable to train the ML model purely with direct observational data rather than reanalysis. This would require consideration of the volume of available training data in terms of the viability of training an ML model and the ability of the model to cope with more extreme events. Additionally, the climate models for comparison would need to be at much higher resolution introducing potential biases and errors associated with down-scaling techniques. Further to this there is also a wealth of climate model data available that would allow for further exploration of the methods presented here. For example, additional scenarios in the CMIP6 archive which provide hundreds of years of data including historical forcing scenarios, pre-industrial control runs and a range of future warming experiments. Other modelling output which produce large volumes of data such as ensemble forecasts would also be an interesting application, particularly the possibility to use SHAP value analysis to identify the variables driving the range of trajectories in the forecast.

Other options for extension, either with Ridge or an alternative ML method, could include experimenting with additional - or replacing existing - predictor variables. Of course, the

caveat that the model is not intended to simply recalculate the entire radiative energy budget still stands as the prediction problem then becomes trivial. However, there are still other predictors which could be tested. Perhaps the most notable omission from the current variable setup is that of time-lagged variables. Although initial testing with a selection of time-lagged variables was not able to provide the Ridge model with any additional skill, alternative time-lagged variables could be an interesting avenue for exploration. Particularly considering the influence of pre-conditions on heatwave events, for example anomalous dry soils preceding the Russia 2010 heatwave (Barriopedro et al., 2011; Schubert et al., 2014; Hauser et al., 2016). Additionally considering predictor variables over larger domain sizes or higher order indices may also help to capture other important heatwave drivers such as teleconnection influences and large scale modes of variability. As well as offering the ability to attribute them more directly than is possible with the currently included dynamical variables.

Extending the method to look at other world regions and seasons would also necessitate further consideration of alternative predictor variables. For instance, the results in this thesis have focused on land regions only; the method could in principle be applied to ocean or coastal regions provided an appropriate predictor variable setup was identified. The model could also be extended for consideration of Southern Hemisphere land heat anomalies.

6.2.2 Bias Correction and Climate Model Evaluation

In terms of applying the Ridge-ERA5 method for bias constraint and climate model evaluation, the results presented in this thesis have focused on one ensemble of global climate models: CMIP6. However the same approach could also be taken to bias correct output from, for example, HighResMip (Haarsma et al., 2016). Similar to CMIP6, participating models run a set of defined experiments allowing for comparison between models, the difference being that HighResMip models are required to have a horizontal resolution of at least 50km in the atmosphere and 0.25^o in the ocean. Comparison of climate model evaluation results between HighResMip and CMIP6 could provide intuitions as to which process biases can be improved with increased resolution.

A side effect of the Ridge-ERA5 bias correction process is the ability to diagnose compensating errors in predictor variables and their relationships with, in this case, temperature. The method could be applied as a tool for identifying compensating errors in other relatively uncertain processes in climate modelling, for example cloud process simulations (Ceppi and Nowack, 2021). This approach would work by constructing a predictive model from reanalysis or observation to predict a quantity of interest based on a series of process-based predictor variables. Comparison of the learnt coefficients between predictors and the target variable in reanalysis and climate models allows for quantitative measures of process realism.

Comparing climate model anomalies with reanalysis coefficients provides a means for bias correction, bearing in mind the caveats for relationships which show a high degree of dissimilarity between climate models and reanalysis. Finally, comparing SHAP value contributions to bias corrected simulations with contributions to climate model emulator predictions allows compensating errors which would not be identifiable in the target variable signal itself to be identified.

Application of the method to a single model Perturbed Physics Ensembles (PPE) could also be investigated. Applying the climate model evaluation methods here to compare coefficients learnt by different ensemble methods with reanalysis could provide a useful tool for initial identification of processes that are improved by particular parameterisations. Additionally, the ability of the bias correction process to highlight potential compensating errors would be a useful tool in this context.

6.2.3 Constraining Future Uncertainty

A clear step for further evaluation of the Ridge-ERA5 future constraint would be a direct comparison with other methodologies for instance, ClimWIP (Knutti et al., 2017), REA (Giorgi and Mearns, 2002) - see Section 4.1 for further details. This kind of direct, quantitative comparison of the Ridge-ERA5 constraint would be an important step to place the constraint in context and to increase confidence in the conclusions that can be drawn. Additionally, confidence in the constraint could also be improved by the inclusion of additional CMIP6 models or models from CMIP5.

Other possibilities for extension would be to consider other future warming scenarios. Future analysis in this thesis has focused on *SSP585* as the most extreme future warming scenario in CMIP6 and thus the most challenging extrapolation case for Ridge-ERA5. The same process could easily be applied to constrain uncertainty ranges under other SSP future scenarios. One particular area of interest would be to apply the Ridge-ERA5 constraint to the Regional Aerosol Model Inter-comparison Project (RAMIP) (Wilcox et al., 2023). Given that anthropogenic aerosol forcings play a key role in future scenario uncertainty, RAMIP seeks to provide avenues to constrain that uncertainty. Experiment runs combine *SSP370* with anthropogenic aerosols from *SSP126*, a scenario in which aerosol contributions decline much faster. This results in a much quicker emergence of the climate change signal in temperatures in the Tropics. Ridge-ERA5 future constraints could be compared between this scenario and *SSP370* to attempt to quantify the impact of aerosol uncertainties, for example using aerosol optical depth as a predictor.

Building on suggestions in Section 6.2.1 to apply Ridge-ERA5 in other world regions, the future constraint method could also be applied to constrain regional temperature change

in the Southern Hemisphere and over ocean and coastal regions given a suitable predictive model. The current constraint on Northern Hemisphere land temperature anomalies is somewhat of an indirect constraint on regional climate sensitivity, this idea could be extended by changing the target predictor variable of the regression model in order to directly constrain quantities like global climate sensitivity. Other quantities which may be of interest for direct constraint included indices defining modes of variability or blocking indices given that blocking dynamics are known to be relatively poorly simulated in GCMs.

6.2.4 Analysing Extreme Events

A simple extension to the heatwave analysis carried out in Chapter 5 would be to consider more heatwave events. Analysing a greater volume of events from similar and more varied climatic regimes may offer greater insight into the link between physical mechanisms and the predictor variables in the Ridge regression models. As mentioned in previous results chapters, while some variables such as soil moisture have a relatively clear interpretation in terms of the climatic processes they represent in Ridge-ERA5, the meaning behind other variables such as monthly 850hPa specific humidity or the 850hPa wind variables is not so directly readable. By comparing SHAP value contributions to heatwaves in similar regions, known to be driven by similar drivers, interpretability of the Ridge-ERA5 predictor variables could be improved. This would develop Ridge-ERA5 into a more powerful tool for quick analysis of future heat extremes.

The Ridge-ERA5 model itself, or alternative ML methods, could also be trained with a more direct focus on extreme events specifically. Here, the methods applied seek to constrain the entire temperature distribution and, with extremes being by definition infrequent, performance is always likely to be relatively poorer for such events. A key challenge to overcome in achieving better extreme event performance is the scarcity of training data. One way to approach this problem would be to group data by climatic regime rather than by locality. For instance, grouping data from particularly dry regions together and training a model which specialises in predicting temperature anomalies under those conditions. One benefit to this approach is that anomalies which are extreme in one location may have been observed relatively frequently during training in another region and so the task to predict this 'extreme' case is no longer one of extrapolation.

Extreme event analysis in Chapter 5 is based on daily-mean temperature anomalies only, however there are of course other metrics that are useful for defining heat wave conditions. For example, modelling daily maximum temperatures would perhaps give a greater focus to extreme event processes with soil moisture and relative humidity anomalies likely to be particularly important. For night time heat extremes, daily minimum temperatures would be

a useful quantity to simulate. Here, predictor variables quantifying cloud cover and specific humidity would be key. Another important quantity in terms of human health impact is the wet bulb temperature. Again this would likely require adaptation of the predictor variable setup.

A Additional Figures

A.1 Coefficient Difference Maps

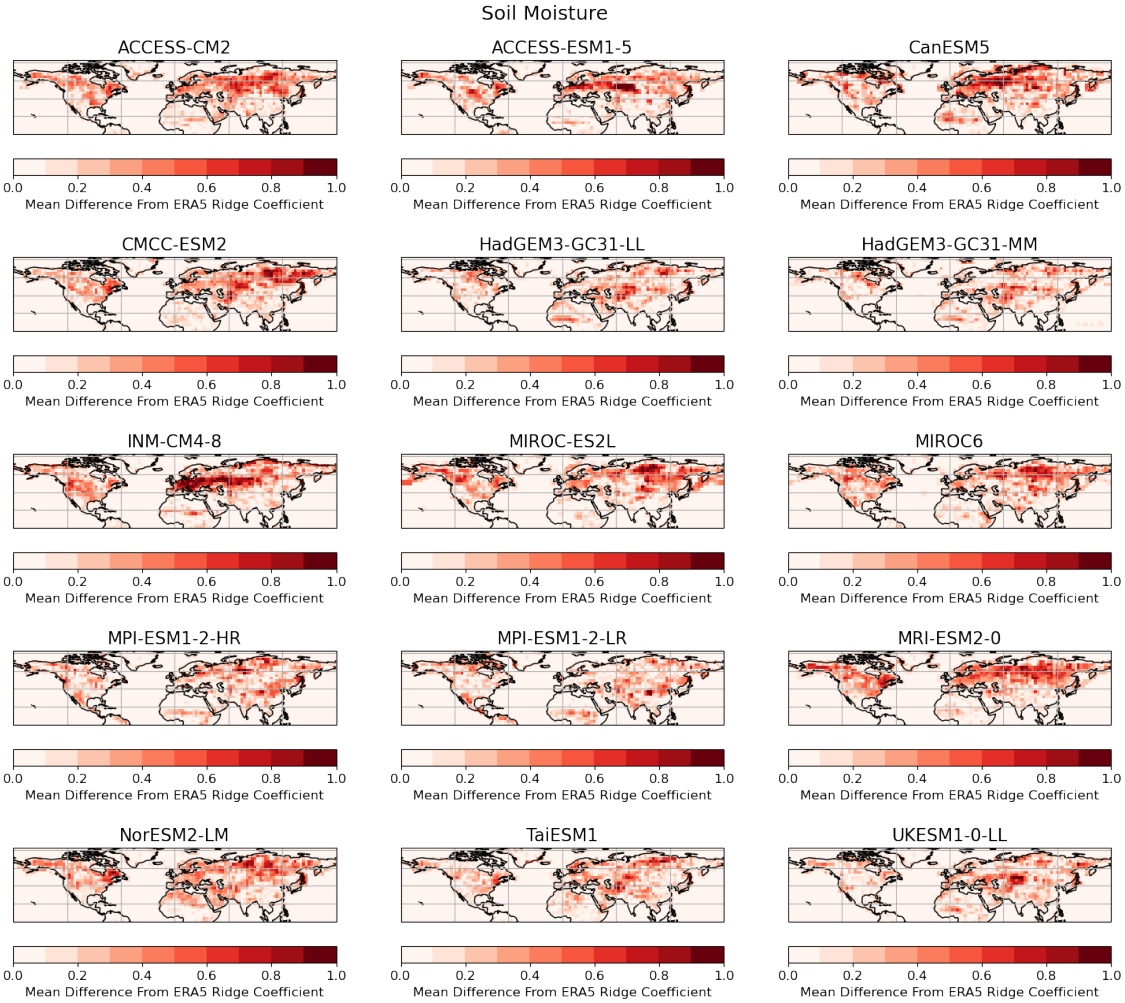

Figure A.1: Mean coefficient difference between Ridge coefficients learnt from CMIP6 models and those learnt from ERA5 for soil moisture.

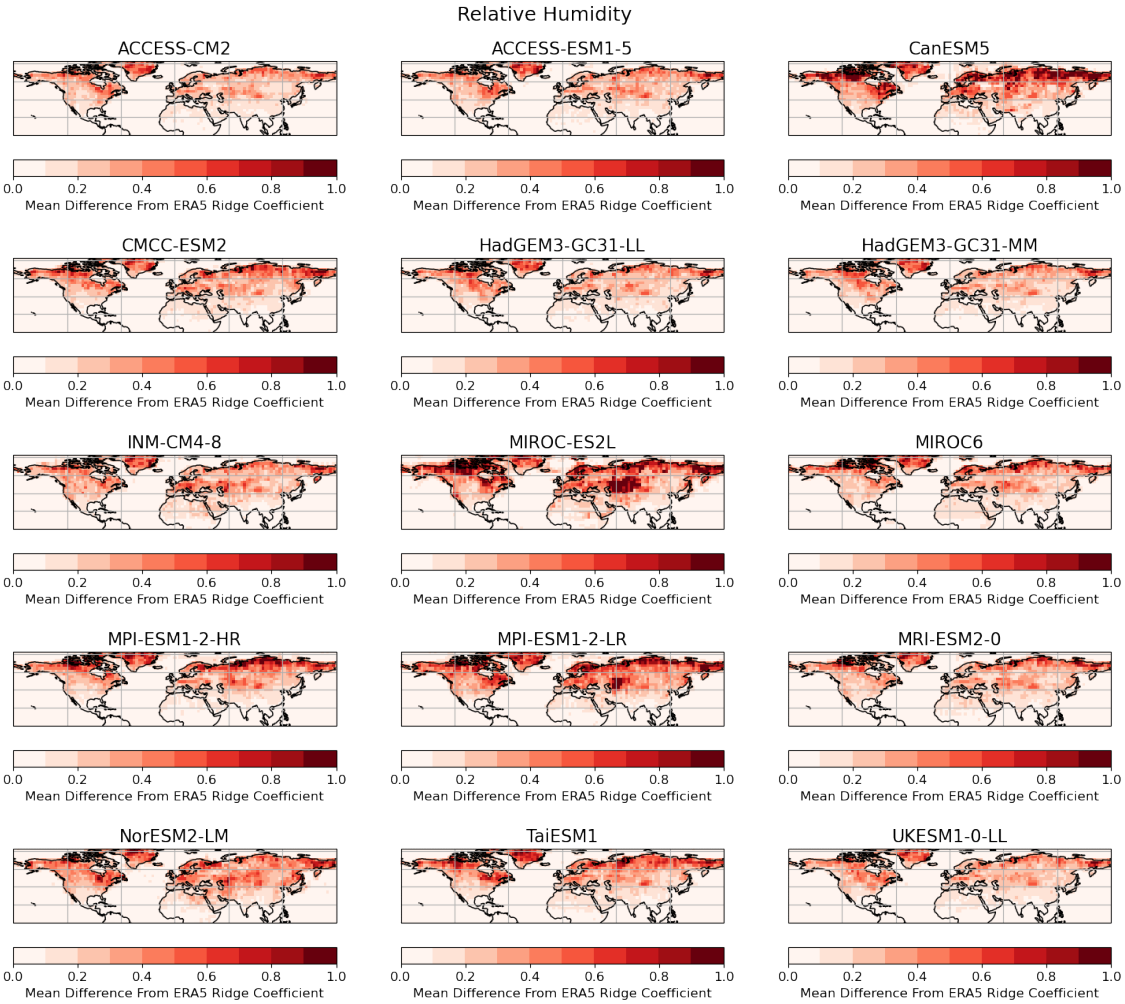

Figure A.2: Mean coefficient difference between Ridge coefficients learnt from CMIP6 models and those learnt from ERA5 for daily relative humidity.

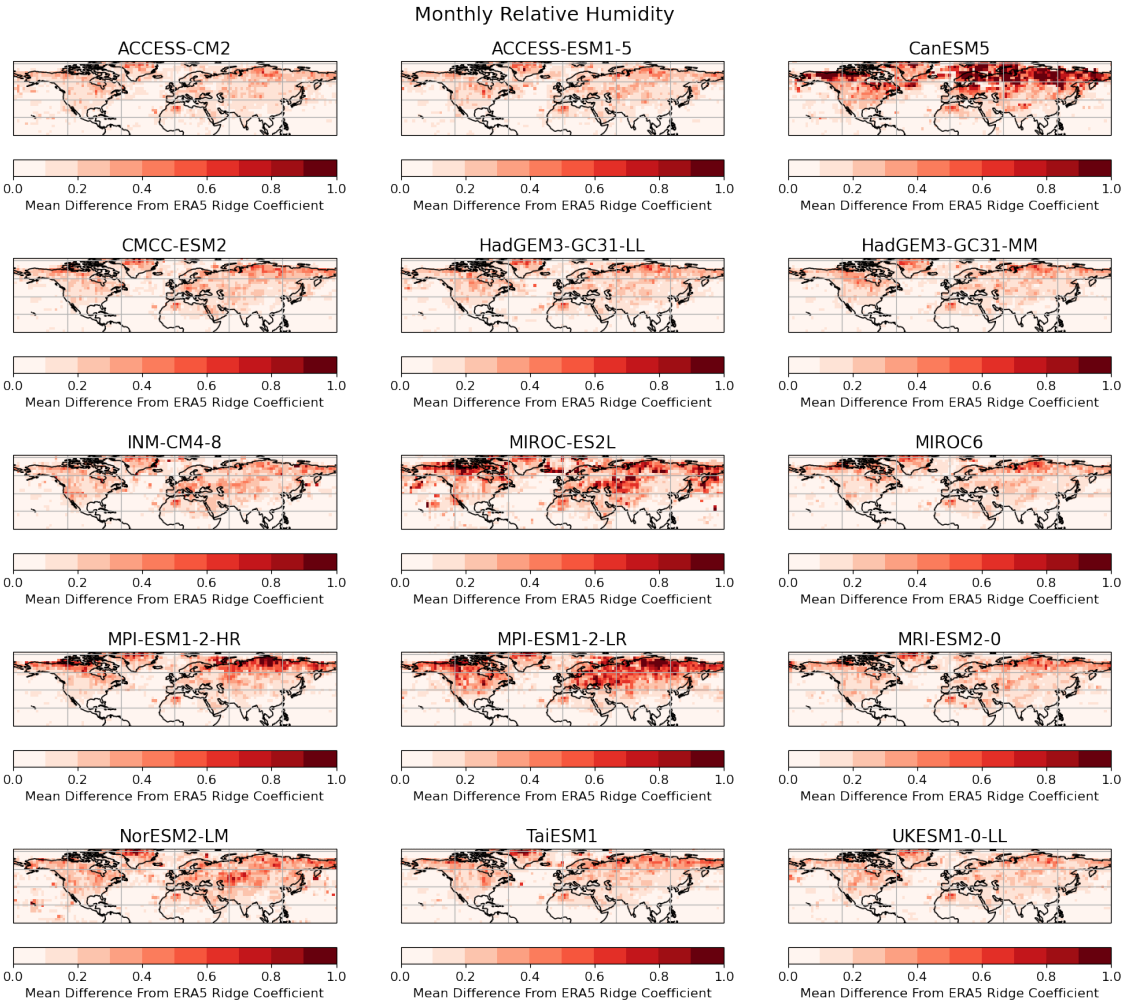

Figure A.3: Mean coefficient difference between Ridge coefficients learnt from CMIP6 models and those learnt from ERA5 for monthly relative humidity.

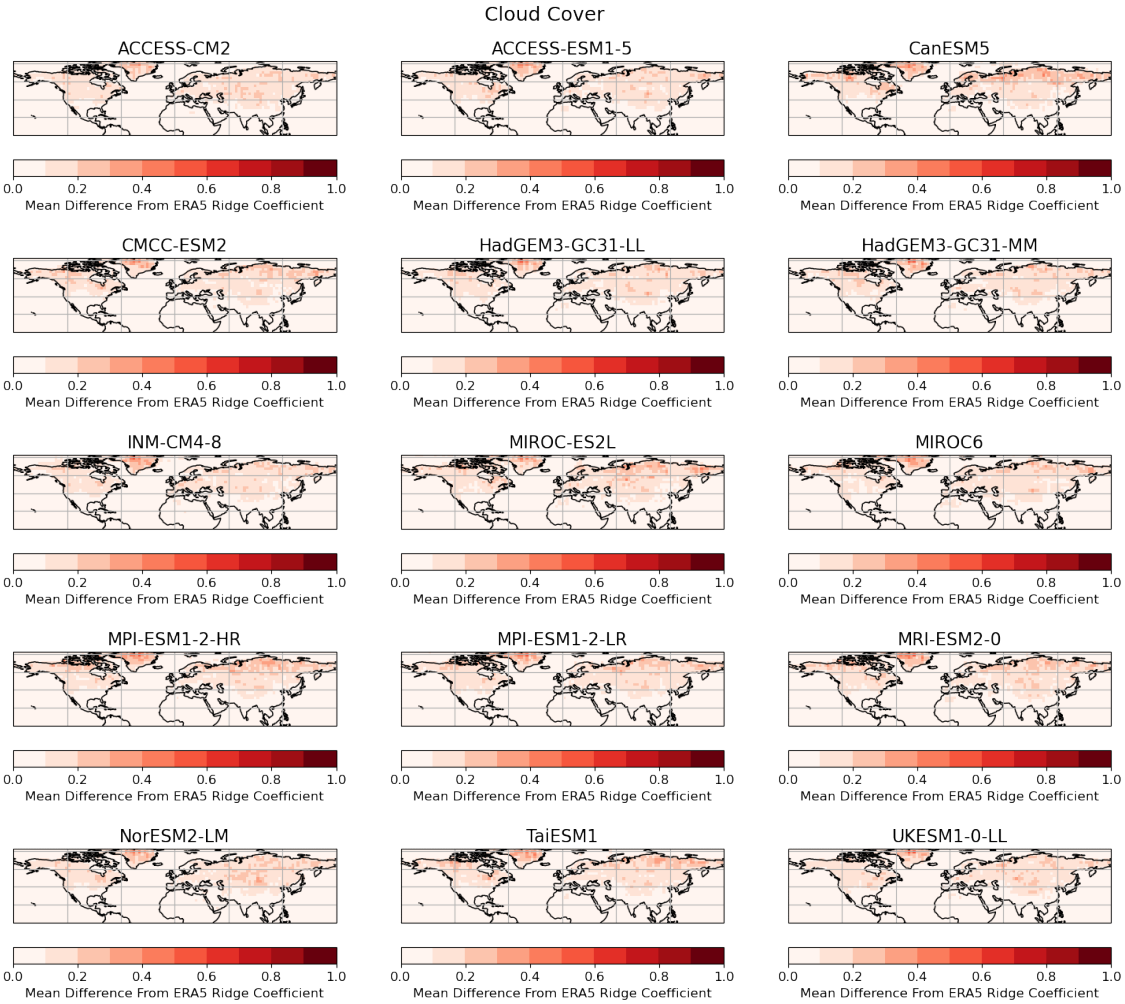

Figure A.4: Mean coefficient difference between Ridge coefficients learnt from CMIP6 models and those learnt from ERA5 for cloud cover.

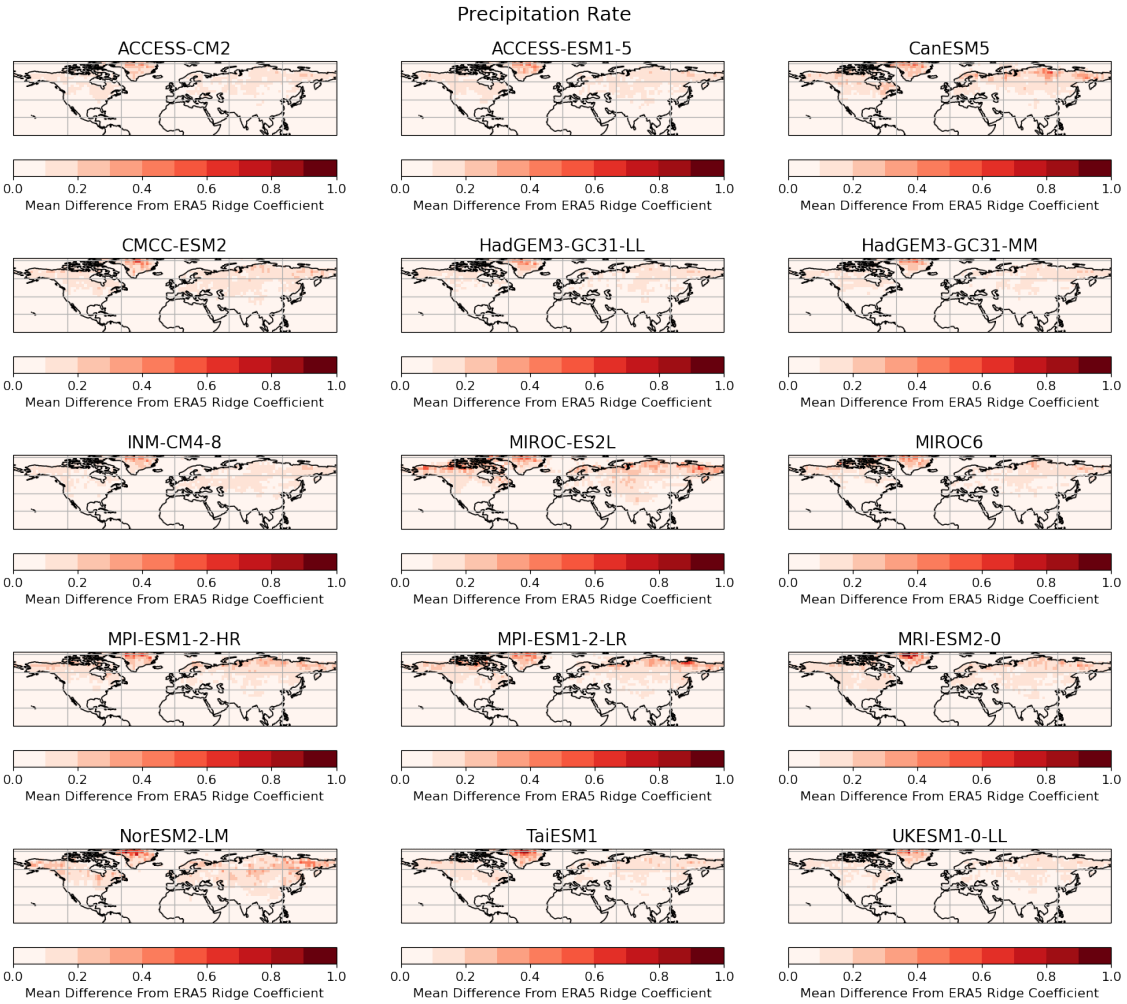

Figure A.5: Mean coefficient difference between Ridge coefficients learnt from CMIP6 models and those learnt from ERA5 for precipitation rate.

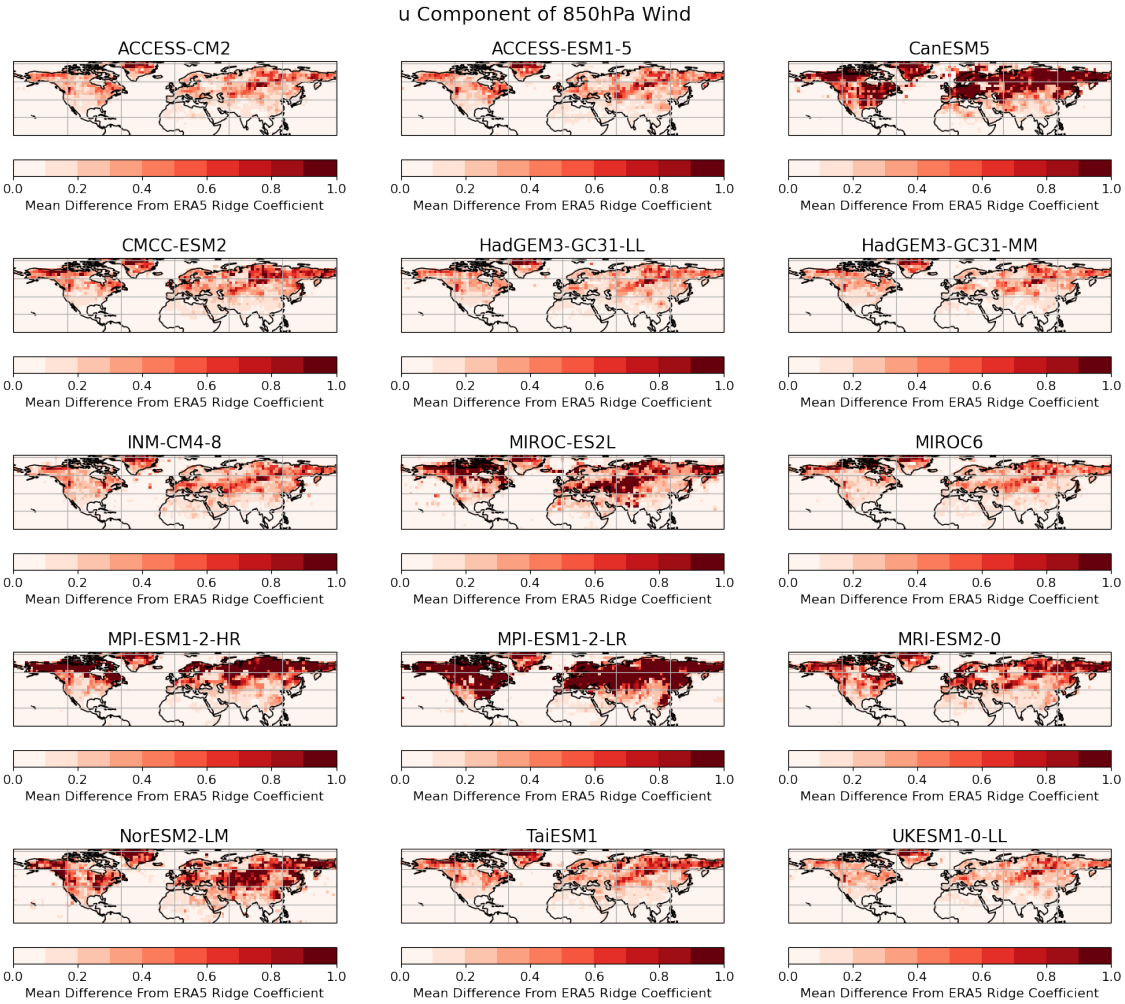

Figure A.6: Mean coefficient difference between Ridge coefficients learnt from CMIP6 models and those learnt from ERA5 for the u component of the 850hPa wind.

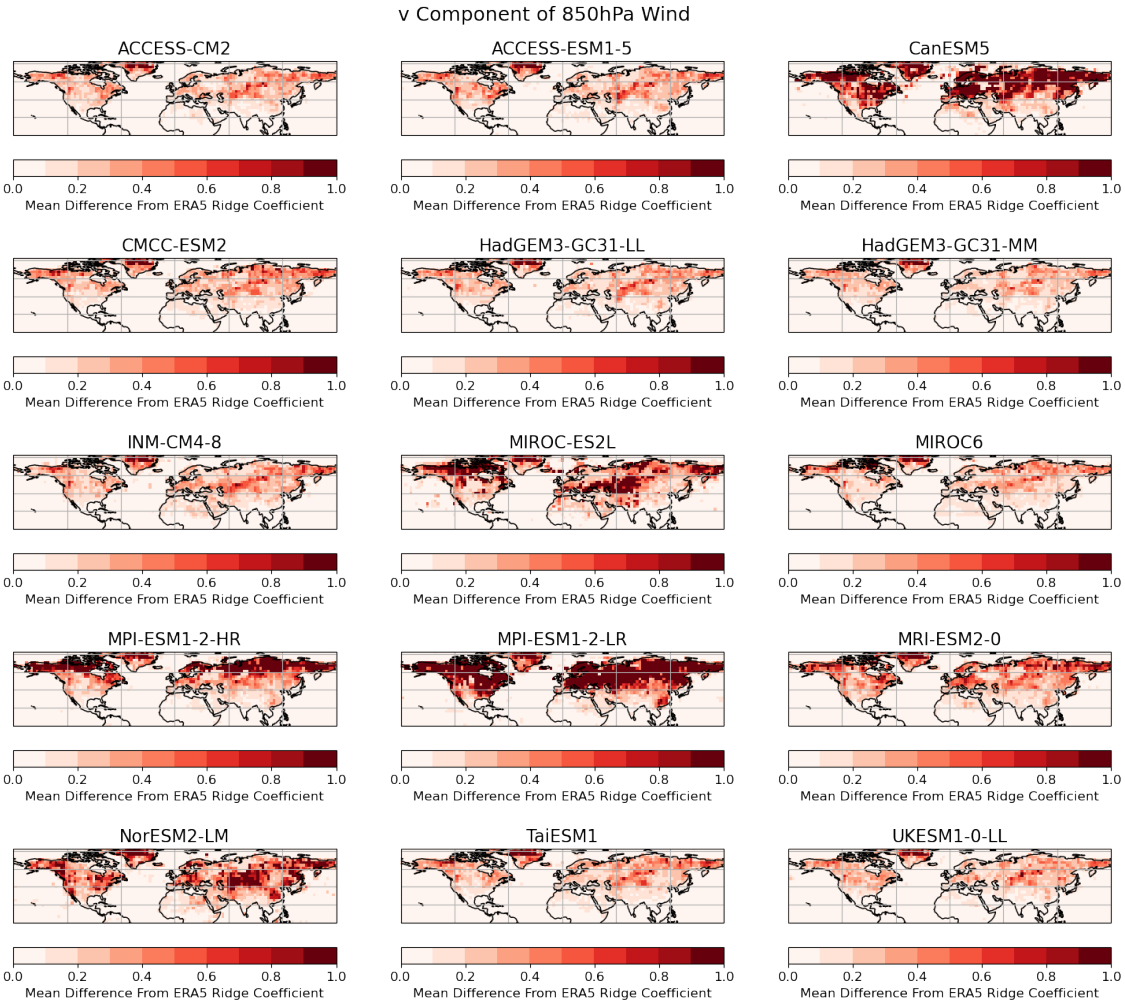

Figure A.7: Mean coefficient difference between Ridge coefficients learnt from CMIP6 models and those learnt from ERA5 for the v component of the 850hPa wind.

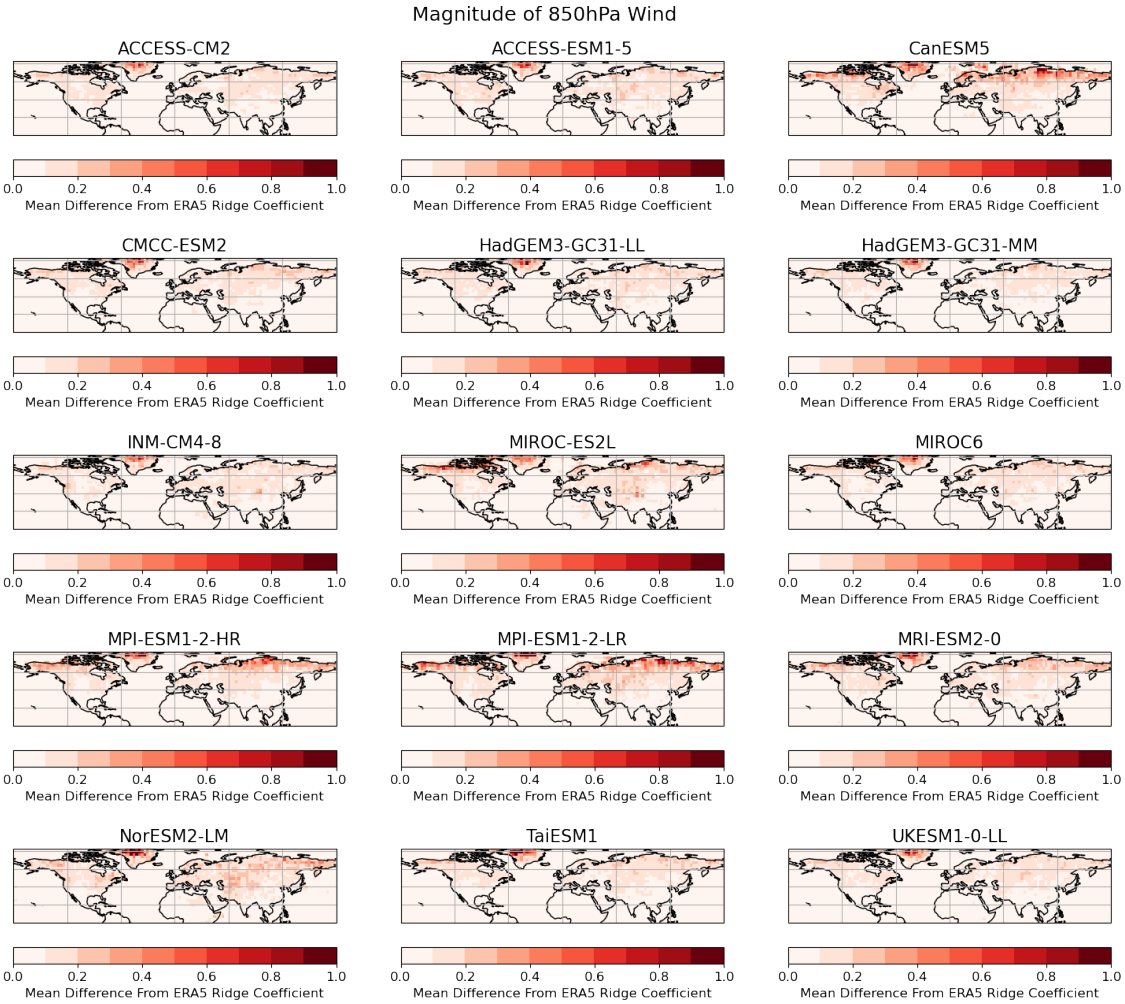

Figure A.8: Mean coefficient difference between Ridge coefficients learnt from CMIP6 models and those learnt from ERA5 for the magnitude of the 850hPa wind.

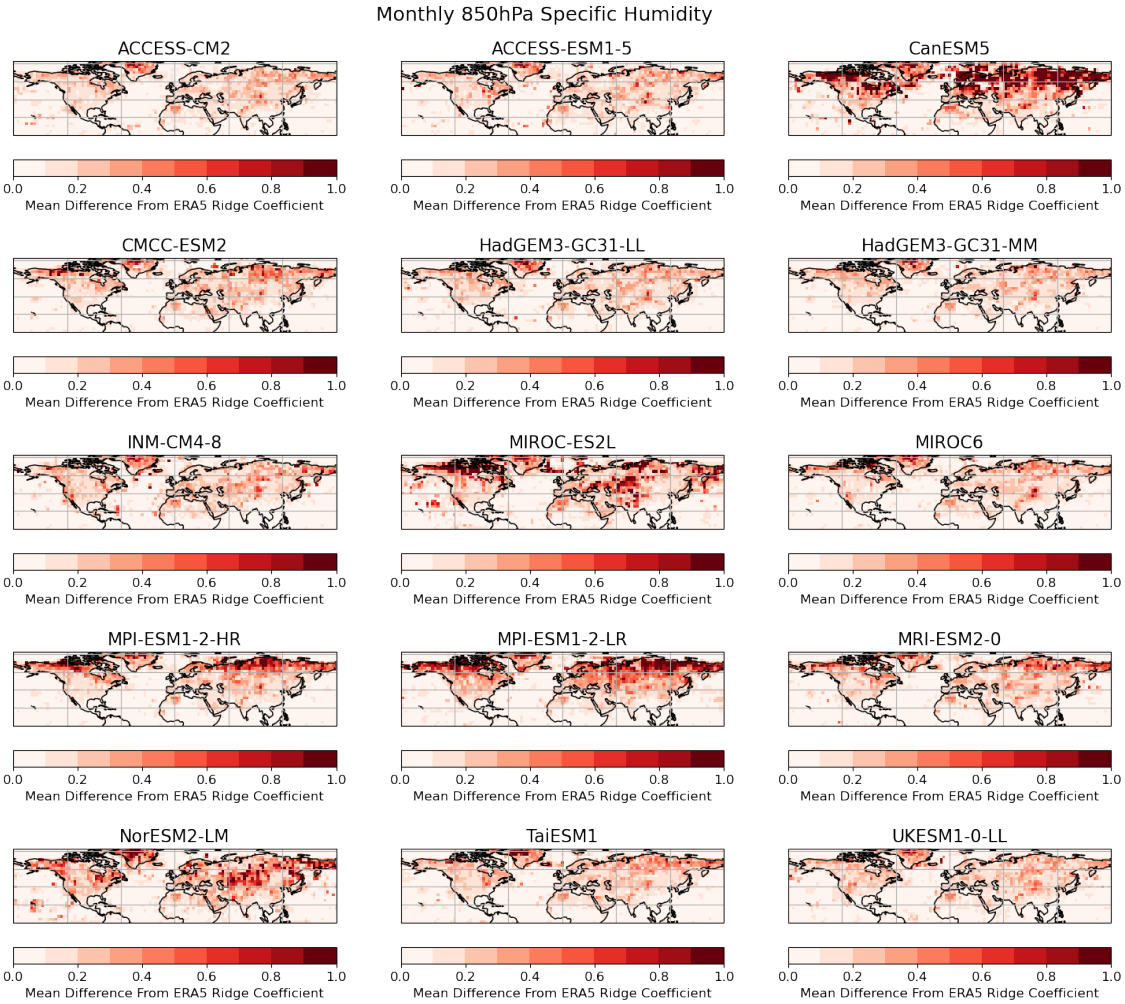

Figure A.9: Mean coefficient difference between Ridge coefficients learnt from CMIP6 models and those learnt from ERA5 for monthly 850hPa specific humidity.

References

Agbazo, M. N. and Grenier, P. (2020), 'Characterizing and avoiding physical inconsistency generated by the application of univariate quantile mapping on daily minimum and maximum temperatures over Hudson Bay', *International Journal of Climatology* **40**(8), 3868–3884.

Ahima, R. S. (2020), 'Global warming threatens human thermoregulation and survival', *The Journal of Clinical Investigation* **130**(2), 559–561.

Alexander, L. (2010), 'Extreme heat rooted in dry soils', *Nature Geoscience 2010 4:1* **4**(1), 12–13.

Ali, P. J. M. and Faraj, R. H. (2014), 'Data Normalization and Standardization: A Technical Report', *Machine Learning Technical Reports* **1**(1), 1–6.

Allen, M. R., Pauline Dube, O., Solecki, W., Aragón-Durand, F., Cramer France, W., Humphreys, S., Dasgupta, P., Millar, R., Dube, O., Solecki, W., Aragón-Durand, F., Cramer, W., Humphreys, S., Kainuma, M., Kala, J., Mahowald, N., Mulugetta, Y., Perez, R., Wairiu, M., Zickfeld, K., Zhai, P., Pörtner, H.-o., Roberts, D., Skea, J., Shukla, P., Pirani, A., Moufouma-Okia, W., Péan, C., Pidcock, R., Connors, S., Matthews, J., Chen, Y., Zhou, X., Gomis, M., Lonnoy, E., Maycock, T., Tignor, M. and Waterfield, T. (2018), Framing and Context, *in* V. Masson-Delmotte, P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J. Matthews, Y. Chen, X. Zhou, M. Gomis, E. Lonnoy, T. Maycock, M. Tignor and T. Waterfield, eds, 'Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change,', Australia, chapter Framing an.

Allen, M. R. and Stott, P. A. (2003), 'Estimating signal amplitudes in optimal fingerprinting, part I: Theory', *Climate Dynamics* **21**(5-6), 477–491.

Allen, M. R., Stott, P. A., Mitchell, J. F., Schnur, R. and Delworth, T. L. (2000), 'Quantifying the uncertainty in forecasts of anthropogenic climate change', *Nature 2000 407:6804* **407**(6804), 617–620.

Allen, M. R. and Tett, S. F. (1999), 'Checking for model consistency in optimal fingerprinting', *Climate Dynamics* **15**(6), 419–434.

Angélil, O., Perkins-Kirkpatrick, S., Alexander, L. V., Stone, D., Donat, M. G., Wehner, M., Shiogama, H., Ciavarella, A. and Christidis, N. (2016), 'Comparing regional precipitation and temperature extremes in climate model and reanalysis products', *Weather and Climate Extremes* **13**, 35–43.

Annan, J. D. and Hargreaves, J. C. (2017), 'On the meaning of independence in climate science', *Earth System Dynamics* **8**(1), 211–224.

Arshad, M., Ma, X., Yin, J., Ullah, W., Liu, M. and Ullah, I. (2021), 'Performance evaluation of ERA-5, JRA-55, MERRA-2, and CFS-2 reanalysis datasets, over diverse climate regions of Pakistan', *Weather and Climate Extremes* **33**, 100373.

Baede, A., Ahlonsou, E., Ding, Y. and Schimel, D. (2001), The Climate System: an Overview, *in* B. Bolin and S. P0llonais, eds, 'TAR Climate Change 2001: The Scientific Basis', pp. 87–98.

Ballard, T. and Erinjippurath, G. (2022), 'Contrastive Learning for Climate Model Bias Correction and Super-Resolution'.

Barnes, E. A., Slingo, J. and Woollings, T. (2012), 'A methodology for the comparison of blocking climatologies across indices, models and climate scenarios', *Climate Dynamics* **38**(11-12), 2467–2481.

Barnes, E. A., Toms, B., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C. and Anderson, D. (2020), 'Indicator Patterns of Forced Change Learned by an Artificial Neural Network', *Journal of Advances in Modeling Earth Systems* **12**(9).

Barriopedro, D., Fischer, E. M., Luterbacher, J., Trigo, R. M. and García-Herrera, R. (2011), 'The hot summer of 2010: redrawing the temperature record map of Europe.', *Science (New York, N.Y.)* **332**(6026), 220–224.

Barriopedro, D., García-Herrera, R., Ordóñez, C., Miralles, D. G. and Salcedo-Sanz, S. (2023), 'Heat Waves: Physical Understanding and Scientific Challenges', *Reviews of Geophysics* **61**(2), e2022RG000780.

Bartusek, S., Kornhuber, K. and Ting, M. (2022), '2021 North American heatwave amplified by climate change-driven nonlinear interactions', *Nature Climate Change 2022 12:12* **12**(12), 1143–1150.

Beer, E., Eisenman, I. and Wagner, T. J. (2020), 'Polar Amplification Due to Enhanced Heat Flux Across the Halocline', *Geophysical Research Letters* **47**(4), e2019GL086706.

Behera, S., Ratnam, J. V., Masumoto, Y. and Yamagata, T. (2013), 'Origin of extreme summers in Europe: The Indo-Pacific connection', *Climate Dynamics* **41**(3-4), 663–676.

Bengtsson, L. and Hodges, K. I. (2019), 'Can an ensemble climate simulation be used to separate climate change signals from internal unforced variability?', *Climate Dynamics* **52**(5-6), 3553–3573.

Berckmans, J., Woollings, T., Demory, M. E., Vidale, P. L. and Roberts, M. (2013), 'Atmospheric blocking in a high resolution climate model: Influences of mean state, orography and eddy forcing', *Atmospheric Science Letters* **14**(1), 34–40.

Beucler, T., Gentine, P., Yuval, J., Gupta, A., Peng, L., Lin, J., Yu, S., Rasp, S., Ahmed, F., O'Gorman, P. A., Neelin, J. D., Lutsko, N. J. and Pritchard, M. (2021), 'Climate-Invariant Machine Learning'.

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, Springer-Verlag.

Bjerknes, J. (1969), 'Atmospheric Teleconnection from the Equatorial Pacific', *Monthly Weather Review* **97**(3), 163–172.

Black, E., Blackburn, M., Harrison, G., Hoskins, B. and Methven, J. (2004), 'Factors contributing to the summer 2003 European heatwave', *Weather* **59**(8), 217–223.

Bosilovich, M., Lucchesi, R. and Suarez, M. (2015), 'MERRA-2: File Specification'.

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**(1), 5–32.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984), *Classification and Regression Trees*, Wadsworth, Pacific Grove, California.

Brient, F. and Schneider, T. (2016), 'Constraints on Climate Sensitivity from Space-Based Measurements of Low-Cloud Reflection', *Journal of Climate* **29**(16), 5821–5835.

Brunner, L., Lorenz, R., Zumwald, M. and Knutti, R. (2019), 'Quantifying uncertainty in European climate projections using combined performance-independence weighting', *Environmental Research Letters* **14**(12), 124010.

Brunner, L., Pendergrass, A. G., Lehner, F., Merrifield, A. L., Lorenz, R. and Knutti, R. (2020), 'Reduced global warming from CMIP6 projections when weighting models by performance and independence', *Earth System Dynamics* **11**(4), 995–1012.

Byrne, M. P. (2021), 'Amplified warming of extreme temperatures over tropical land', *Nature Geoscience 2021 14:11* **14**(11), 837–841.

Byrne, M. P. and O'Gorman, P. A. (2018), 'Trends in continental temperature and humidity directly linked to ocean warming', *Proceedings of the National Academy of Sciences of the United States of America* **115**(19), 4863–4868.

Caldwell, P. M., Bretherton, C. S., Zelinka, M. D., Klein, S. A., Santer, B. D., Sanderson, B. M., Caldwell, P. M., Bretherton, C. S., Zelinka, M. D., Klein, S. A., Santer, B. D. and Sanderson, B. M. (2014), 'Statistical significance of climate sensitivity predictors obtained by data mining', *Geophysical Research Letters* **41**(5), 1803–1808.

Caldwell, P. M., Zelinka, M. D. and Klein, S. A. (2018), 'Evaluating Emergent Constraints on Equilibrium Climate Sensitivity', *Journal of Climate* **31**(10), 3921–3942.

Cannon, A. J. (2016), 'Multivariate Bias Correction of Climate Model Output: Matching Marginal Distributions and Intervariable Dependence Structure', *Journal of Climate* **29**(19), 7045–7064.

Cao, L., Bala, G., Caldeira, K., Nemani, R. and Ban-Weiss, G. (2010), 'Importance of carbon dioxide physiological forcing to future climate change', *Proceedings of the National Academy of Sciences of the United States of America* **107**(21), 9513–9518.

Casado, M. J. and Pastor, M. A. (2012), 'Use of variability modes to evaluate AR4 climate models over the Euro-Atlantic region', *Climate Dynamics* **38**(1-2), 225–237.

Casado, M. J., Pastor, M. A. and Doblas-Reyes, F. J. (2009), 'Euro-Atlantic circulation types and modes of variability in winter', *Theoretical and Applied Climatology* **96**(1-2), 17–29.

Cassou, C., Terray, L. and Phillips, A. S. (2005), 'Tropical Atlantic Influence on European Heat Waves', *Journal of Climate* **18**(15), 2805–2811.

Ceppi, P. and Nowack, P. (2021), 'Observational evidence that cloud feedback amplifies global warming', *PNAS* **118**(30).

Chan, P. W., Catto, J. L. and Collins, M. (2022), 'Heatwave–blocking relation change likely dominates over decrease in blocking frequency under global warming', *npj Climate and Atmospheric Science 2022 5:1* **5**(1), 1–8.

Chan, S. C., Kendon, E. J., Fowler, H. J., Blenkinsop, S., Roberts, N. M. and Ferro, C. A. (2014), 'The Value of High-Resolution Met Office Regional Climate Models in the Simulation of Multihourly Precipitation Extremes', *Journal of Climate* **27**(16), 6155–6174.

Chantry, M., Christensen, H., Dueben, P. and Palmer, T. (2021), 'Opportunities and challenges for machine learning in weather and climate modelling: hard, medium and soft AI', *Philosophical Transactions of the Royal Society A* **379**(2194).

Chen, J., Brissette, F. P. and Leconte, R. (2011), 'Uncertainty of downscaling method in quantifying the impact of climate change on hydrology', *Journal of Hydrology* **401**(3-4), 190–202.

Chen, J., Li, C., Brissette, F. P., Chen, H., Wang, M. and Essou, G. R. (2018), 'Impacts of correcting the inter-variable correlation of climate model outputs on hydrological modeling', *Journal of Hydrology* **560**, 326–341.

Chen, R. and Li, X. (2023), 'Causes of the persistent merging of the western North Pacific subtropical high and the Iran high during late July 2022', *Climate Dynamics* **61**(5-6), 2285–2297.

Chen, Y. C. (2017), 'A tutorial on kernel density estimation and recent advances', *https://doi.org/10.1080/24709360.2017.1396742* **1**(1), 161–187.

Cho, D., Yoo, C., Im, J. and Cha, D. H. (2020), 'Comparative Assessment of Various Machine Learning-Based Bias Correction Methods for Numerical Weather Prediction Model Forecasts of Extreme Air Temperatures in Urban Areas', *Earth and Space Science* **7**(4), e2019EA000740.

Chou, S. C., Lyra, A., Mourão, C., Dereczynski, C., Pilotto, I., Gomes, J., Bustamante, J., Tavares, P., Silva, A., Rodrigues, D., Campos, D., Sueiro, G., Siqueira, G., Nobre, P., Marengo, J. and Chagas, D. (2014), 'Evaluation of the Eta Simulations Nested in Three Global Climate Models', *American Journal of Climate Change* **03**(05), 438–454.

Collins, M., Achutarao, K., Ashok, K., Bhandari, S., Mitra, A. K., Prakash, S., Srivastava, R. and Turner, A. (2013), 'Observational challenges in evaluating climate models', *Nature Climate Change* **3**(11), 940–941.

Conrick, R. and Mass, C. F. (2023), 'The influence of soil moisture on the historic 2021 Pacific Northwest heatwave', *Monthly Weather Review* **-1**(aop).

Cos, J., Doblas-Reyes, F., Jury, M., Marcos, R., Bretonnière, P. A. and Samsó, M. (2022), 'The Mediterranean climate change hotspot in the CMIP5 and CMIP6 projections', *Earth System Dynamics* **13**(1), 321–340.

Cox, P. M., Huntingford, C. and Williamson, M. S. (2018), 'Emergent constraint on equilibrium climate sensitivity from global temperature variability', *Nature 2018 553:7688* **553**(7688), 319–322.

Cox, P. M., Pearson, D., Booth, B. B., Friedlingstein, P., Huntingford, C., Jones, C. D. and Luke, C. M. (2013), 'Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability', *Nature 2013 494:7437* **494**(7437), 341–344.

Davini, P. and D'Andrea, F. (2016), 'Northern Hemisphere Atmospheric Blocking Representation in Global Climate Models: Twenty Years of Improvements?', *Journal of Climate* **29**(24), 8823–8840.

Davini, P. and D'Andrea, F. (2020), 'From CMIP3 to CMIP6: Northern Hemisphere Atmospheric Blocking Simulation in Present and Future Climate', *Journal of Climate* **33**(23), 10021–10038.

Deser, C., Phillips, A., Bourdette, V. and Teng, H. (2012), 'Uncertainty in climate change projections: The role of internal variability', *Climate Dynamics* **38**(3-4), 527–546.

Di Luca, A., Pitman, A. J. and de Elía, R. (2020), 'Decomposing Temperature Extremes Errors in CMIP5 and CMIP6 Models', *Geophysical Research Letters* **47**(14), e2020GL088031.

Diaz, H. F., Hoerling, M. P. and Eischeid, J. K. (2001), 'Enso variability, teleconnections and climate change', *International Journal of Climatology* **21**(15), 1845–1862.

Diffenbaugh, N. S., Singh, D., Mankin, J. S., Horton, D. E., Swain, D. L., Touma, D., Charland, A., Liu, Y., Haugen, M., Tsiang, M. and Rajaratnam, B. (2017), 'Quantifying the influence of global warming on unprecedented extreme climate events', *Proceedings of the National Academy of Sciences of the United States of America* **114**(19), 4881–4886.

Diro, G. T. and Sushama, L. (2020), 'Contribution of Snow Cover Decline to Projected Warming Over North America', *Geophysical Research Letters* **47**(1), e2019GL084414.

Doblas-Reyes, F. J., Sörensson, A. A., Almazroui, M., Dosio, A., Gutowski, W. J., Aalgeirsdóttir, G., Adhikary, B., Adnan, M., Ahrens, B., Amjad, M., Arias, P. A., Mohamed Azam, F., Cruz, A., Daron, J. D., Ruiz, L., Saeed Belgium, S., Saurral, R. I. and Schiemann, R. K. (2021), 2021: Linking Global to Regional Climate Change, *in* V. Masson-Delmotte, P. Zhai, A. Pirani, S. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. Matthews, T. Maycock, T. Waterfield, O. Yelekçi, R. Yu and B. Zhou, eds, 'Climate Change 2021: The Physical Science Basis.

Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change', Cambridge University Press, Cambridge, UK, pp. 1363–1512.

Domeisen, D. I., Eltahir, E. A., Fischer, E. M., Knutti, R., Perkins-Kirkpatrick, S. E., Schär, C., Seneviratne, S. I., Weisheimer, A. and Wernli, H. (2023), 'Prediction and projection of heatwaves', *Nature Reviews Earth & Environment 2022 4:1* **4**(1), 36–50.

Dong, B., Gregory, J. M. and Sutton, R. T. (2009), 'Understanding Land–Sea Warming Contrast in Response to Increasing Greenhouse Gases. Part I: Transient Adjustment', *Journal of Climate* **22**(11), 3079–3097.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., Mcclean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D. and Lautenbach, S. (2013), 'Collinearity: a review of methods to deal with it and a simulation study evaluating their performance', *Ecography* **36**(1), 27–46.

Drouard, M., Kornhuber, K. and Woollings, T. (2019), 'Disentangling Dynamic Contributions to Summer 2018 Anomalous Weather Over Europe', *Geophysical Research Letters* **46**(21), 12537–12546.

ECMWF (2016), Part IV Physical Processes, *in* 'IFS Documentation - Cy43r1'.

Edwards, P. N. (2011), 'History of climate modeling', *Wiley Interdisciplinary Reviews: Climate Change* **2**(1), 128–139.

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

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., Collins, W. D., Gier, B. K., Hall, A. D., Hoffman, F. M., Hurtt, G. C., Jahn, A., Jones, C. D., Klein, S. A., Krasting, J. P., Kwiatkowski, L., Lorenz, R., Maloney, E., Meehl, G. A., Pendergrass, A. G., Pincus, R., Ruane, A. C., Russell, J. L., Sanderson, B. M., Santer, B. D., Sherwood, S. C., Simpson, I. R., Stouffer, R. J. and Williamson, M. S. (2019), 'Taking climate model evaluation to the next level', *Nature Climate Change* **9**(2), 102–110.

Fan, X., Miao, C., Duan, Q., Shen, C. and Wu, Y. (2020), 'The Performance of CMIP6 Versus CMIP5 in Simulating Temperature Extremes Over the Global Land Surface', *Journal of Geophysical Research: Atmospheres* **125**(18), e2020JD033031.

Feng, M., McPhaden, M. J., Xie, S. P. and Hafner, J. (2013), 'La Niña forces unprecedented Leeuwin Current warming in 2011', *Scientific Reports 2013 3:1* **3**(1), 1–9.

Ferguglia, O., von Hardenberg, J. and Palazzi, E. (2023), 'Robustness of precipitation Emergent Constraints in CMIP6 models', *Climate Dynamics* **1**, 1–12.

Ferranti, L. and Viterbo, P. (2006), 'The European Summer of 2003: Sensitivity to Soil Water Initial Conditions', *Journal of Climate* **19**(15), 3659–3680.

Fink, A. H., Brücher, T., Krüger, A., Leckebusch, G. C., Pinto, J. G. and Ulbrich, U. (2004), 'The 2003 European summer heatwaves and drought –synoptic diagnosis and impacts', *Weather* **59**(8), 209–216.

Fischer, E. M. and Knutti, R. (2015), 'Anthropogenic contribution to global occurrence of heavy-precipitation and high-temperature extremes', *Nature Climate Change* **5**(6), 560–564.

Fischer, E. M. and Schär, C. (2010), 'Consistent geographical patterns of changes in high-impact European heatwaves', *Nature Geoscience* **3**(6), 398–403.

Fischer, E. M., Seneviratne, S. I., Vidale, P. L., Lüthi, D. and Schär, C. (2007), 'Soil Moisture–Atmosphere Interactions during the 2003 European Summer Heat Wave', *Journal of Climate* **20**(20), 5081–5099.

Fischer, E. M., Sippel, S. and Knutti, R. (2021), 'Increasing probability of record-shattering climate extremes', *Nature Climate Change 2021 11:8* **11**(8), 689–695.

Flato, G. M. (2011), 'Earth system models: an overview', *Wiley Interdisciplinary Reviews: Climate Change* **2**(6), 783–800.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C. and Rummukainen, M. (2013), Evaluation of Climate Models., *in* T. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, Bex V. and P. Midgley, eds, 'Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assess-ment Report of the Intergovernmental Panel on Climate Change', Cambrdige University Press, Cambridge, United Kingdom, pp. 741–866.

Foley, A. M. (2010), 'Uncertainty in regional climate modelling: A review', *Progress in Physical Geography* **34**(5), 647–670.

François, B., Vrac, M., Cannon, A. J., Robin, Y. and Allard, D. (2020), 'Multivariate bias corrections of climate simulations: Which benefits for which losses?', *Earth System Dynamics* **11**(2), 537–562.

Fulton, D. J., Clarke, B. J. and Hegerl, G. C. (2023), 'Bias Correcting Climate Model Simulations Using Unpaired Image-to-Image Translation Networks', *Artificial Intelligence for the Earth Systems* **2**(2).

Garcia-Herrera, R., Díaz, J., Trigo, R. M., Luterbacher, J. and Fischer, E. M. (2010), 'A Review of the European Summer Heat Wave of 2003', *http://dx.doi.org/10.1080/10643380802238137* **40**(4), 267–306.

Gettelman, A. and Rood, R. B. (2016), Essence of a Climate Model, *in* 'Demystifying Climate Models', Springer, Berlin, Heidelberg, pp. 37–58.

Giorgi, F. and Mearns, L. O. (2002), 'Calculation of Average, Uncertainty Range, and Reliability of Regional Climate Changes from AOGCM Simulations via the "Reliability Ensemble Averaging" (REA) Method', *Journal of Climate* **15**(10), 1141–1158.

Gleckler, P. J., Taylor, K. E. and Doutriaux, C. (2008), 'Performance metrics for climate models', *Journal of Geophysical Research: Atmospheres* **113**(D6), 6104.

Goodess, C. M. (2013), 'How is the frequency, location and severity of extreme events likely to change up to 2060?', *Environmental Science and Policy* **27**, S4–S14.

Graham, R. M., Hudson, S. R. and Maturilli, M. (2019), 'Improved Performance of ERA5 in Arctic Gateway Relative to Four Global Atmospheric Reanalyses', *Geophysical Research Letters* **46**(11), 6138–6147.

Gregorutti, B., Michel, B. and Saint-Pierre, P. (2017), 'Correlation and variable importance in random forests', *Statistics and Computing* **27**(3), 659–678.

Gudmundsson, L., Bremnes, J. B., Haugen, J. E. and Engen-Skaugen, T. (2012), 'Hydrology and Earth System Sciences Technical Note: Downscaling RCM precipitation to the station scale using statistical transformations-a comparison of methods', *Hydrol. Earth Syst. Sci* **16**, 3383–3390.

Haarsma, R. J., Roberts, M. J., Vidale, P. L., Catherine, A., Bellucci, A., Bao, Q., Chang, P., Corti, S., Fučkar, N. S., Guemas, V., Von Hardenberg, J., Hazeleger, W., Kodama, C., Koenigk, T., Leung, L. R., Lu, J., Luo, J. J., Mao, J., Mizielinski, M. S., Mizuta, R., Nobre, P., Satoh, M., Scoccimarro, E., Semmler, T., Small, J. and Von Storch, J. S.

(2016), 'High Resolution Model Intercomparison Project (HighResMIP v1.0) for CMIP6', *Geoscientific Model Development* **9**(11), 4185–4208.

Hall, A. (2004), 'The Role of Surface Albedo Feedback in Climate', *Journal of Climate* **17**(7).

Hall, A., Cox, P., Huntingford, C. and Klein, S. (2019), 'Progressing emergent constraints on future climate change', *Nature Climate Change 2019 9:4* **9**(4), 269–278.

Hall, A. and Qu, X. (2006), 'Using the current seasonal cycle to constrain snow albedo feedback in future climate change', *Geophysical Research Letters* **33**(3), 3502.

Hardwick Jones, R., Westra, S. and Sharma, A. (2010), 'Observed relationships between extreme sub-daily precipitation, surface temperature, and relative humidity', *Geophysical Research Letters* **37**(22).

Hargreaves, J. C. (2010), 'Skill and uncertainty in climate models', *Wiley Interdisciplinary Reviews: Climate Change* **1**(4), 556–564.

Hasselmann, K. (1997), 'Multi-pattern fingerprint method for detection and attribution of climate change', *Climate Dynamics* **13**(9), 601–611.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning*, 2nd edn, Springer.

Hauser, M., Orth, R. and Seneviratne, S. I. (2016), 'Role of soil moisture versus recent climate change for the 2010 heat wave in western Russia', *Geophysical Research Letters* **43**(6), 2819–2826.

Hawkins, E., Smith, R. S., Gregory, J. M. and Stainforth, D. A. (2016), 'Irreducible uncertainty in near-term climate projections', *Climate Dynamics* **46**(11-12), 3807–3819.

Hawkins, E. and Sutton, R. (2009), 'The potential to narrow uncertainty in regional climate predictions', *Bulletin of the American Meteorological Society* **90**(8), 1095–1107.

He, C., Zhou, T., Zhang, L., Chen, X. and Zhang, W. (2023), 'Extremely hot East Asia and flooding western South Asia in the summer of 2022 tied to reversed flow over Tibetan Plateau', *Climate Dynamics* **61**(5-6), 2103–2119.

Hegerl, G. and Zwiers, F. (2011), 'Advanced Review Use of models in detection and attribution of climate change', *Ltd. WIREs Clim Change* **2**, 570–591.

Held, I. M. and Soden, B. J. (2006), 'Robust responses of the hydrological cycle to global warming', *Journal of Climate* **19**(21), 5686–5699.

Herman, G. R. and Schumacher, R. S. (2018), 'Money doesn't grow on trees, but forecasts do: Forecasting extreme precipitation with random forests', *Monthly Weather Review* **146**(5), 1571–1600.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D. and Thépaut, J.-N. (2023*a*), 'ERA5 hourly data on pressure levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS)'.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D. and Thépaut, J.-N. (2023*b*), 'ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS)'.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S. and Thépaut, J. N. (2020), 'The ERA5 global reanalysis', *Quarterly Journal of the Royal Meteorological Society* **146**(730), 1999–2049.

Hirsch, A. L., Ridder, N. N., Perkins-Kirkpatrick, S. E. and Ukkola, A. (2021), 'CMIP6 MultiModel Evaluation of Present-Day Heatwave Attributes', *Geophysical Research Letters* **48**(22), e2021GL095161.

Ho, T. K. (1995), 'Random decision forests', *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR* **1**, 278–282.

Hodnebrog, O., Myhre, G., Samset, B. H., Alterskjær, K., Andrews, T., Boucher, O., Faluvegi, G., Fläschner, D., M Forster, P., Kasoar, M., Kirkeväg, A., Lamarque, J. F., Olivié, D., B Richardson, T., Shawki, D., Shindell, D., P Shine, K., Stier, P., Takemura, T., Voulgarakis, A. and Watson-Parris, D. (2019), 'Water vapour adjustments and responses differ between climate drivers', *Atmospheric Chemistry and Physics* **19**(20), 12887–12899.

Hoerl, A. E. and Kennard, R. W. (1970), 'Ridge Regression: Biased Estimation for Nonorthogonal Problems', *Technometrics* **12**(1), 55–67.

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J. C., Balaji, V., Duan, Q., Folini, D., Ji, D., Klocke, D., Qian, Y., Rauser, F., Rio, C., Tomassini, L., Watanabe, M. and Williamson,

D. (2017), 'The Art and Science of Climate Model Tuning', *Bulletin of the American Meteorological Society* **98**(3), 589–602.

Hua, W., Dai, A., Qin, M., Hu, Y. and Cui, Y. (2023), 'How Unexpected Was the 2022 Summertime Heat Extremes in the Middle Reaches of the Yangtze River?', *Geophysical Research Letters* **50**(16), e2023GL104269.

IPCC (2012), Summary for Policymakers, *in* Field, C.B., V. Barros, T. Stocker, D. Qin, D. Dokken, K. Ebi, M. Mastrandrea, K. Mach, G.-K. Plattner, S. Allen, M. Tignor and P. Midgley, eds, 'Managing the Risks of Extreme Events and Disasters to AdvanceClimate Change Adaptation', Cambridge University Press, Cambridge, UK, pp. 1–19.

IPCC (2013), Summary for Policymakers., *in* D. Qin, G.-k. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P. Midgley, eds, 'Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the ntergovernmental Panel on Climate Change', Cambridge University Press, Cambridge, UK.

IPCC (2021), Annex VII: Glossary, *in* J. Matthews, V. Möller, R. van Diemen, J. Fuglestvedt, V. Masson-Delmotte, C. Méndez, S. Semenov and A. Reisinger, eds, 'Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change', Cambridge University Press, pp. 2215–2256.

Iturbide, M., Gutiérrez, J. M., Alves, L. M., Bedia, J., Cerezo-Mota, R., Cimadevilla, E., Cofiño, A. S., Luca, A. D., Faria, S. H., Gorodetskaya, I. V., Hauser, M., Herrera, S., Hennessy, K., Hewitt, H. T., Jones, R. G., Krakovska, S., Manzanas, R., Martínez-Castro, D., Narisma, G. T., Nurhati, I. S., Pinto, I., Seneviratne, S. I., van den Hurk, B. and Vera, C. S. (2020), 'An update of IPCC climate reference regions for subcontinental analysis of climate model data: definition and aggregated datasets', *Earth System Science Data* **12**(4), 2959–2970.

Jackson, L. and Vellinga, M. (2013), 'Multidecadal to Centennial Variability of the AMOC: HadCM3 and a Perturbed Physics Ensemble', *Journal of Climate* **26**(7), 2390–2407.

Jiang, J., Liu, Y., Mao, J. and Wu, G. (2023), 'Extreme heatwave over Eastern China in summer 2022: the role of three oceans and local soil moisture feedback', *Environmental Research Letters* **18**(4), 044025.

Jiang, K., Pan, Z., Pan, F., Wang, J., Han, G., Song, Y., Zhang, Z., Huang, N., Ma, S. and Chen, X. (2022), 'Influence patterns of soil moisture change on surface-air temperature difference under different climatic background', *Science of The Total Environment* **822**, 153607.

Jiménez-de-la Cuesta, D. and Mauritsen, T. (2019), 'Emergent constraints on Earth's transient and equilibrium response to doubled CO2 from post-1970s global warming', *Nature Geoscience 2019 12:11* **12**(11), 902–905.

Jiménez-Esteve, B. and Domeisen, D. I. (2022), 'The role of atmospheric dynamics and large-scale topography in driving heatwaves', *Quarterly Journal of the Royal Meteorological Society. Royal Meteorological Society (Great Britain)* **148**(746), 2344.

Jin, S. (2020), 'Compositional kernel learning using tree-based genetic programming for Gaussian process regression', *Structural and Multidisciplinary Optimization* **62**, 1313–1351.

Jones, G. S., Stott, P. A. and Christidis, N. (2013), 'Attribution of observed historical near–surface temperature variations to anthropogenic and natural causes using CMIP5 simulations', *Journal of Geophysical Research: Atmospheres* **118**(10), 4001–4024.

Jones, M., Marron, J. and Sheather, S. (1992), 'Progress in Data-Based Bandwidth Selection for Kernel Density Estimation'.

Joshi, M. M., Gregory, J. M., Webb, M. J., Sexton, D. M. and Johns, T. C. (2008), 'Mechanisms for the land/sea warming contrast exhibited by simulations of climate change', *Climate Dynamics* **30**(5), 455–465.

Jung, T., Miller, M. J., Palmer, T. N., Towers, P., Wedi, N., Achuthavarier, D., Adams, J. M., Altshuler, E. L., Cash, B. A., Kinter, J. L., Marx, L., Stan, C. and Hodges, K. I. (2012), 'High-Resolution Global Climate Simulations with the ECMWF Model in Project Athena: Experimental Design, Model Climate, and Seasonal Forecast Skill', *Journal of Climate* **25**(9), 3155–3172.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A. and Kumar, V. (2019), 'Machine Learning for the Geosciences: Challenges and Opportunities', *IEEE Transactions on Knowledge and Data Engineering* **31**(8), 1544–1554.

Kashinath, K., Mustafa, M., Albert, A., Wu, J. L., Jiang, C., Esmaeilzadeh, S., Azizzadenesheli, K., Wang, R., Chattopadhyay, A., Singh, A., Manepalli, A., Chirila, D., Yu, R.,

Walters, R., White, B., Xiao, H., Tchelepi, H. A., Marcus, P., Anandkumar, A., Hassanzadeh, P. and Prabhat (2021), 'Physics-informed machine learning: case studies for weather and climate modelling', *Philosophical Transactions of the Royal Society A* **379**(2194).

Kettleborough, J. A., Booth, B. B., Stott, P. A. and Allen, M. R. (2007), 'Estimates of Uncertainty in Predictions of Global Mean Surface Temperature', *Journal of Climate* **20**(5), 843–855.

Kharin, V. V. and Zwiers, F. W. (2000), 'Changes in the Extremes in an Ensemble of Transient Climate Simulations with a Coupled Atmosphere–Ocean GCM ', *Journal of Climate* **13**(21), 3760–3788.

Kharin, V. V., Zwiers, F. W., Zhang, X. and Wehner, M. (2013), 'Changes in temperature and precipitation extremes in the CMIP5 ensemble', *Climatic Change* **119**(2), 345–357.

Klein Tank, A. M., Zwiers, F. W. and Zhang, X. (2009), 'Guidelines on Analysis of Extremes in a Changing Climate in Support of Informed Decisions for Adaptation', *WCDMP-No. 72, WMO-TD No. 1500* p. 56.

Knutson, T. (2017), Detection and Attribution Methodologies Overview, *in* D. Wuebbles, D. Fahey, K. Hibbard, D. Dokken, B. Stewart and T. Maycock, eds, 'Climate Science Special Report: Fourth National Climate Assessment, Volume I', Vol. I, pp. 443–451.

Knutti, R. (2010), 'The end of model democracy?', *Climatic Change* **102**(3), 395–404.

Knutti, R., Masson, D. and Gettelman, A. (2013), 'Climate model genealogy: Generation CMIP5 and how we got there', *Geophysical Research Letters* **40**(6), 1194–1199.

Knutti, R., Rugenstein, M. A. and Hegerl, G. C. (2017), 'Beyond equilibrium climate sensitivity', *Nature Geoscience 2017 10:10* **10**(10), 727–736.

Knutti, R. and Sedláček, J. (2012), 'Robustness and uncertainties in the new CMIP5 climate model projections', *Nature Climate Change 2012 3:4* **3**(4), 369–373.

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K. and Kiyotoshi, T. (2015), 'The JRA-55 Reanalysis: General Specifications and Basic Characteristics', *Journal of the Meteorological Society of Japan* **93**(1), 5–48.

Kornhuber, K., Osprey, S., Coumou, D., Petri, S., Petoukhov, V., Rahmstorf, S. and Gray, L. (2019), 'Extreme weather events in early summer 2018 connected by a recurrent hemispheric wave-7 pattern', *Environmental Research Letters* **14**(5), 054002.

Kotlarski, S., Szabó, P., Herrera, S., Räty, O., Keuler, K., Soares, P. M., Cardoso, R. M., Bosshard, T., Pagé, C., Boberg, F., Gutiérrez, J. M., Isotta, F. A., Jaczewski, A., Kreienkamp, F., Liniger, M. A., Lussana, C. and Pianko-Kluczyńska, K. (2019), 'Observational uncertainty and regional climate model evaluation: A pan-European perspective', *International Journal of Climatology* **39**(9), 3730–3749.

Kovats, R. S. and Kristie, L. E. (2006), 'Heatwaves and public health in Europe', *European Journal of Public Health* **16**(6), 592–599.

Lal, P., Shekhar, A., Gharun, M. and Das, N. N. (2023), 'Spatiotemporal evolution of global long-term patterns of soil moisture', *Science of The Total Environment* **867**, 161470.

Le Treut, H., Somerville, R., Cubasch, U., Ding, Y., Mauritzen, C., Mokssit, A., Peterson, T. and Prather, M. (2007), Historical Overview of Climate Change Science, *in* S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. Averyt, Tignor M. and H. Miller, eds, 'Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change', Cambridge University Press.

Li, C., Sinha, E., Horton, D. E., Diffenbaugh, N. S. and Michalak, A. M. (2014), 'Joint bias correction of temperature and precipitation in climate model simulations', *Journal of Geophysical Research: Atmospheres* **119**(23), 13,153–13,162.

Li, G., Xie, S. P., He, C. and Chen, Z. (2017), 'Western Pacific emergent constraint lowers projected increase in Indian summer monsoon rainfall', *Nature Climate Change* **7**(10), 708–712.

Li, J., Huo, R., Chen, H., Zhao, Y. and Zhao, T. (2021), 'Comparative Assessment and Future Prediction Using CMIP6 and CMIP5 for Annual Precipitation and Extreme Precipitation Simulation', *Frontiers in Earth Science* **9**, 687976.

Liu, H., Koren, I., Altaratz, O. and Chekroun, M. D. (2023), 'Opposing trends of cloud coverage over land and ocean under global warming', *Atmos. Chem. Phys* **23**, 6559–6569.

Liu, X., He, B., Guo, L., Huang, L. and Chen, D. (2020), 'Similarities and Differences in the Mechanisms Causing the European Summer Heatwaves in 2003, 2010, and 2018', *Earth's Future* **8**(4), e2019EF001386.

Loikith, P. C. and Broccoli, A. J. (2014), 'The Influence of Recurrent Modes of Climate Variability on the Occurrence of Winter and Summer Extreme Temperatures over North America', *Journal of Climate* **27**(4), 1600–1618.

Londhe, D. S., Katpatal, Y. B. and Bokde, N. D. (2023), 'Performance Assessment of Bias Correction Methods for Precipitation and Temperature from CMIP5 Model Simulation', *Applied Sciences 2023, Vol. 13, Page 9142* **13**(16), 9142.

Lorenz, R., Herger, N., Sedláček, J., Eyring, V., Fischer, E. M. and Knutti, R. (2018), 'Prospects and Caveats of Weighting Climate Models for Summer Maximum Temperature Projections Over North America', *Journal of Geophysical Research: Atmospheres* **123**(9), 4509–4526.

Lucas-Picher, P., Caya, D., Elía, R. and Laprise, R. (2008), 'Investigation of regional climate models' internal variability with a ten-member ensemble of 10-year simulations over a large domain', *Climate Dynamics* **31**(7-8), 927–940.

Lundberg, S. M. and Lee, S.-I. (2017), 'A Unified Approach to Interpreting Model Predictions', *Advances in Neural Information Processing Systems* **30**, 4768–4777.

Luo, M. and Lau, N. C. (2019), 'Amplifying effect of ENSO on heat waves in China', *Climate Dynamics* **52**(5-6), 3277–3289.

Lupo, A. R. (2021), 'Atmospheric blocking events: a review', *Annals of the New York Academy of Sciences* **1504**(1), 5–24.

Luterbacher, J., Dietrich, D., Xoplaki, E., Grosjean, M. and Wanner, H. (2004), 'European Seasonal and Annual Temperature Variability, Trends, and Extremes since 1500', *Science* **303**(5663), 1499–1503.

Mahfouf, J.-F. (1991), 'Analysis of Soil Moisture from Near-Surface Parameters: A Feasibility Study', *Cover Journal of Applied Meteorology and Climatology Journal of Applied Meteorology and Climatology* **30**(11), 1534–1547.

Manabe, S., Spelman, M. and Stouffer, R. (1992), 'Transient Responses of a Coupled Ocean-Atmosphere Model to Gradual Changes of Atmospheric CO2. Part II: Seasonal Response ', *Ice in the Climate System* **5**(2), 105–126.

Manabe, S. and Wetherald, R. T. (1980), 'On the Distribution of Climate Change Resulting from an Increase in CO2 Content of the Atmosphere ', *Cover Journal of the Atmospheric Sciences Journal of the Atmospheric Sciences* **37**(1), 99–118.

Maraun, D. (2016), 'Bias Correcting Climate Change Simulations - a Critical Review', *Current Climate Change Reports* **2**(4), 211–220.

Maraun, D., Shepherd, T. G., Widmann, M., Zappa, G., Walton, D., Gutiérrez, J. M., Hagemann, S., Richter, I., Soares, P. M., Hall, A. and Mearns, L. O. (2017), 'Towards process-informed bias correction of climate change simulations', *Nature Climate Change 2017 7:11* **7**(11), 764–773.

Marquardt Collow, A. B., Thomas, N. P., Bosilovich, M. G., Lim, Y. K., Schubert, S. D. and Koster, R. D. (2022), 'Seasonal Variability in the Mechanisms behind the 2020 Siberian Heatwaves', *Journal of Climate* **35**(10), 3075–3090.

Marron, J. S. (1988), 'Automatic smoothing parameter selection: A survey', *Empirical Economics* **13**(3-4), 187–208.

Masato, G., Hoskins, B. J. and Woollings, T. (2013), 'Winter and Summer Northern Hemisphere Blocking in CMIP5 Models', *Journal of Climate* **26**(18), 7044–7059.

Masson, D. and Knutti, R. (2011), 'Climate model genealogy', *Geophysical Research Letters* **38**(8), 8703.

Matsueda, M. (2009), 'Blocking Predictability in Operational Medium-Range Ensemble Forecasts', *SOLA* **5**(1), 113–116.

Mauritsen, T., Stevens, B., Roeckner, E., Crueger, T., Esch, M., Giorgetta, M., Haak, H., Jungclaus, J., Klocke, D., Matei, D., Mikolajewicz, U., Notz, D., Pincus, R., Schmidt, H. and Tomassini, L. (2012), 'Tuning the climate of a global model', *Journal of Advances in Modeling Earth Systems* **4**(3), 0–01.

McGuffie, K. and Henderson-Sellers, A. (2001), 'Forty years of numerical climate modelling', *International Journal of Climatology* **21**(9), 1067–1109.

Mckitrick, R., Mcintyre, S. and Herman, C. (2010), 'Panel and multivariate methods for tests of trend equivalence in climate data series', *Atmospheric Science Letters* **11**(4), 270–277.

McMichael, A. J., Woodruff, R. E. and Hales, S. (2006), 'Climate change and human health: present and future risks', *The Lancet* **367**(9513), 859–869.

Meehl, G. A., Senior, C. A., Eyring, V., Flato, G., Lamarque, J. F., Stouffer, R. J., Taylor, K. E. and Schlund, M. (2020), 'Context for interpreting equilibrium climate sensitivity and transient climate response from the CMIP6 Earth system models', *Science Advances* **6**(26).

Meehl, G. A. and Tebaldi, C. (2004), 'More intense, more frequent, and longer lasting heat waves in the 21st century', *Science (New York, N.Y.)* **305**(5686), 994–997.

Merrifield, A. L., Brunner, L., Lorenz, R., Medhaug, I. and Knutti, R. (2020), 'An investigation of weighting schemes suitable for incorporating large ensembles into multi-model ensembles', *Earth System Dynamics* **11**(3), 807–834.

Meyer, J., Kohn, I., Stahl, K., Hakala, K., Seibert, J. and Cannon, A. J. (2019), 'Effects of univariate and multivariate bias correction on hydrological impact projections in alpine catchments', *Hydrology and Earth System Sciences* **23**(3), 1339–1354.

Miralles, D. G., Gentine, P., Seneviratne, S. I. and Teuling, A. J. (2019), 'Land–atmospheric feedbacks during droughts and heatwaves: state of the science and current challenges', *Annals of the New York Academy of Sciences* **1436**(1), 19.

Miralles, D. G., Teuling, A. J., Van Heerwaarden, C. C. and De Arellano, J. V. G. (2014), 'Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation', *Nature Geoscience 2014 7:5* **7**(5), 345–349.

Muñoz-Díaz, D. and Rodrigo, F. S. (2005), 'Influence of the El Niño-Southern oscillation on the probability of dry and wet seasons in Spain', *Climate Research* **30**(1), 1–12.

Murphy, J., Harris, G., Sexton, D., Kendon, E., Bett, P., Clark, R., Eagle, K., Fosser, G., Fung, F., Lowe, J., McDonald, R., McInnes, R., McSweeney, C., Mitchell, J., Rostron, J., Thornton, H., Tucker, S. and Yamakazi, K. (2018), 'UKCP18 Land Projections: Science Report', *Met Office Report* .

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

Murphy, J., Sexton, D., Jenkins, G. and Booth, B. (2009), *UK Climate Projections Science Report: Climate Change Projections*, Met Office Hadley Centre.

Nakamura, N. and Huang, C. S. (2018), 'Atmospheric blocking as a traffic jam in the jet stream', *Science* **361**(6397), 42–47.

Neal, E., Huang, C. S. and Nakamura, N. (2022), 'The 2021 Pacific Northwest Heat Wave and Associated Blocking: Meteorology and the Role of an Upstream Cyclone as a Diabatic Source of Wave Activity', *Geophysical Research Letters* **49**(8), e2021GL097699.

Nijsse, F. J., Cox, P. M. and Williamson, M. S. (2020), 'Emergent constraints on transient climate response (TCR) and equilibrium climate sensitivity (ECS) from historical warming in CMIP5 and CMIP6 models', *Earth System Dynamics* **11**(3), 737–750.

NOAA (2018), 'State of the Climate: Global Climate Report for July 2018', *NOAA National Centers for Environmental Information* .

Nowack, P., Ceppi, P., Davis, S. M., Chiodo, G., Ball, W., Diallo, M. A., Hassler, B., Jia, Y., Keeble, J. and Joshi, M. (2023), 'Response of stratospheric water vapour to warming constrained by satellite observations', *Nature Geoscience 2023 16:7* **16**(7), 577–583.

O'Gorman, P. A. (2012), 'Sensitivity of tropical precipitation extremes to climate change', *Nature Geoscience 2012 5:10* **5**(10), 697–700.

O'Neill, B. C., Kriegler, E., Ebi, K. L., Kemp-Benedict, E., Riahi, K., Rothman, D., van Ruijven, B., van Vuuren, D. P., Birkmann, J., Kok, K., Levy, M. and Solecki, W. (2015), 'The roads ahead: Narratives for shared socioeconomic pathways describing world futures in the 21st century', *Global Environmental Change* **42**.

Otto, F. E., Massey, N., Van Oldenborgh, G. J., Jones, R. G. and Allen, M. R. (2012), 'Reconciling two approaches to attribution of the 2010 Russian heat wave', *Geophysical Research Letters* **39**(4).

Overland, J. E. (2021), 'Causes of the Record-Breaking Pacific Northwest Heatwave, Late June 2021', *Atmosphere 2021, Vol. 12, Page 1434* **12**(11), 1434.

Palmer, T. N. (2013), 'Climate extremes and the role of dynamics', *Proceedings of the National Academy of Sciences* **110**(14), 5281–5282.

Pan, B., Anderson, G. J., Goncalves, A., Lucas, D. D., Bonfils, C. J., Lee, J., Tian, Y. and Ma, H. Y. (2021), 'Learning to Correct Climate Projection Biases', *Journal of Advances in Modeling Earth Systems* **13**(10), e2021MS002509.

Pang, Y., Lin, J., Qin, T. and Chen, Z. (2022), 'Image-to-Image Translation: Methods and Applications', *IEEE Transactions on Multimedia* **24**, 3859–3881.

Park, T., Efros, A. A., Zhang, R. and Zhu, J. Y. (2020), 'Contrastive Learning for Unpaired Image-to-Image Translation', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **12354 LNCS**, 319–345.

Parzen, E. (1962), 'On Estimation of a Probability Density Function and Mode', *https://doi.org/10.1214/aoms/1177704472* **33**(3), 1065–1076.

Patz, J. A., Campbell-Lendrum, D., Holloway, T. and Foley, J. A. (2005), 'Impact of regional climate change on human health', *Nature 2005 438:7066* **438**(7066), 310–317.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Bertrand, T., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011), 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research* **2o**(85), 2825–2830.

Pennell, C. and Reichler, T. (2011), 'On the Effective Number of Climate Models', *Journal of Climate* **24**(9), 2358–2367.

Perkins-Kirkpatrick, S. E. and Lewis, S. C. (2020), 'Increasing trends in regional heatwaves', *Nature Communications 2020 11:1* **11**(1), 1–8.

Perkins, S. E. (2015), 'A review on the scientific understanding of heatwaves—Their measurement, driving mechanisms, and changes at the global scale', *Atmospheric Research* **164-165**, 242–267.

Pfahl, S., Wernli, H., Pfahl, S. and Wernli, H. (2012), 'Quantifying the relevance of atmospheric blocking for co-located temperature extremes in the Northern Hemisphere on (sub-)daily time scales', *GeoRL* **39**(12), L12807.

Philip, S. Y., Kew, S. F., Jan van Oldenborgh, G., Yang, W., Vecchi, G. A., Anslow, F. S., Li, S., Seneviratne, S. I., Luu, L. N., Arrighi, J., Singh, R., van Aalst, M., Hauser, M., Schumacher, D. L., Pereira Marghidan, C., Ebi, K. L., Bonnet, R., Vautard, R., Tradowsky, J., Coumou, D., Lehner, F., Wehner, M., Rodell, C., Stull, R., Howard, R., Gillett, N. and L Otto, F. E. (2021), 'Rapid attribution analysis of the extraordinary heatwave on the Pacific Coast of the US and Canada June 2021', *World Weather Attribution* .

Phillips, A. S., Deser, C. and Fasullo, J. (2014), 'Evaluating modes of variability in climate models', *Eos (United States)* **95**(49), 453–455.

Phillips, T. J. (1996), 'Documentation of the AMIP Models on the World Wide Web', *Bulletin of the American Meteorological Society* **77**(6), 1191–1196.

Pincus, R., Batstone, C. P., Patrick Hofmann, R. J., Taylor, K. E. and Glecker, P. J. (2008), 'Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models', *Journal of Geophysical Research: Atmospheres* **113**(D14).

Pirtle, Z., Meyer, R. and Hamilton, A. (2010), 'What does it mean when climate models agree? A case for assessing independence among general circulation models', *Environmental Science & Policy* **13**(5), 351–361.

Pithan, F. and Mauritsen, T. (2014), 'Arctic amplification dominated by temperature feedbacks in contemporary climate models', *Nature Geoscience 2014 7:3* **7**(3), 181–184.

Pithan, F., Shepherd, T. G., Zappa, G. and Sandu, I. (2016), 'Climate model biases in jet streams, blocking and storm tracks resulting from missing orographic drag', *Geophysical Research Letters* **43**(13), 7231–7240.

Pitman, A. J., Arneth, A. and Ganzeveld, L. (2012), 'Regionalizing global climate models', *International Journal of Climatology* **32**(3), 321–337.

Qasmi, S. and Ribes, A. (2022), 'Reducing uncertainty in local temperature projections', *Science Advances* **8**(41), 6872.

Qian, W. and Chang, H. H. (2021), 'Projecting Health Impacts of Future Temperature: A Comparison of Quantile-Mapping Bias-Correction Methods', *International Journal of Environmental Research and Public Health* **18**(4), 1–12.

Qiao, L., Zuo, Z., Xiao, D. and Bu, L. (2021), 'Detection, Attribution, and Future Response of Global Soil Moisture in Summer', *Frontiers in Earth Science* **9**, 882.

Quesada, B., Vautard, R., Yiou, P., Hirschi, M. and Seneviratne, S. I. (2012), 'Asymmetric European summer heat predictability from wet and dry southern winters and springs', *Nature Climate Change 2012 2:10* **2**(10), 736–741.

Räisänen, J. (2001), 'CO2-Induced Climate Change in CMIP2 Experiments: Quantification of Agreement and Role of Internal Variability', *Journal of Climate* **14**(9), 2088–2104.

Räisänen, J. (2007), 'How reliable are climate models?', *Tellus, Series A: Dynamic Meteorology and Oceanography* **59**(1), 2–29.

Ramon, J., Lledó, L., Torralba, V., Soret, A. and Doblas-Reyes, F. J. (2019), 'What global reanalysis best represents near-surface winds?', *Quarterly Journal of the Royal Meteorological Society* **145**(724), 3236–3251.

Rantanen, M., Karpechko, A. Y., Lipponen, A., Nordling, K., Hyvärinen, O., Ruosteenoja, K., Vihma, T. and Laaksonen, A. (2022), 'The Arctic has warmed nearly four times faster than the globe since 1979', *Communications Earth & Environment 2022 3:1* **3**(1), 1–10.

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S. and Thuerey, N. (2020), 'WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting', *Journal of Advances in Modeling Earth Systems* **12**(11), e2020MS002203.

Reichler, T. and Kim, J. (2008), 'How Well Do Coupled Models Simulate Today's Climate?', *Bulletin of the American Meteorological Society* **89**(3), 303–312.

Ribes, A., Boé, J., Qasmi, S., Dubuisson, B., Douville, H. and Terray, L. (2022), 'An updated assessment of past and future warming over France based on a regional observational constraint', *Earth System Dynamics* **13**(4), 1397–1415.

Ribes, A., Qasmi, S. and Gillett, N. P. (2021), 'Making climate projections conditional on historical observations', *Science Advances* **7**(4), 671–693.

Ribes, A. and Terray, L. (2013), 'Application of regularised optimal fingerprinting to attribution. Part II: Application to global near-surface temperature', *Climate Dynamics* **41**(11-12), 2837–2853.

Ribes, A., Thao, S. and Cattiaux, J. (2020), 'Describing the Relationship between a Weather Event and Climate Change: A New Statistical Approach ', *Journal of Climate* **33**(15), 6297–6314.

Ribes, A., Zwiers, F. W., Azaïs, J.-M. and Naveau, P. (2017), 'A new statistical approach to climate change detection and attribution', *Climate Dynamics* p. 48.

Robine, J. M., Cheung, S. L. K., Le Roy, S., Van Oyen, H., Griffiths, C., Michel, J. P. and Herrmann, F. R. (2008), 'Death toll exceeded 70,000 in Europe during the summer of 2003', *Comptes Rendus - Biologies* **331**(2), 171–178.

Robinson, P. J. (2001), 'On the Definition of a Heat Wave ', *Journal of Applied Meteorology and Climatology* **40**(4).

Rosenblatt, M. (1956), 'Remarks on Some Nonparametric Estimates of a Density Function', *https://doi.org/10.1214/aoms/1177728190* **27**(3), 832–837.

Röthlisberger, M. and Papritz, L. (2023), 'Quantifying the physical processes leading to atmospheric hot extremes at a global scale', *Nature Geoscience 2023* pp. 1–7.

Russo, S., Sillmann, J. and Fischer, E. M. (2015), 'Top ten European heatwaves since 1950 and their occurrence in the coming decades', *Environmental Research Letters* **10**(12), 124003.

Russo, S., Sillmann, J. and Sterl, A. (2017), 'Humid heat waves at different warming levels', *Scientific Reports 2017 7:1* **7**(1), 1–7.

Saha, S., Moorthi, S., Pan, H. L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., Liu, H., Stokes, D., Grumbine, R., Gayno, G., Wang, J., Hou, Y. T., Chuang, H. Y., Juang, H. M. H., Sela, J., Iredell, M., Treadon, R., Kleist, D., Van

Delst, P., Keyser, D., Derber, J., Ek, M., Meng, J., Wei, H., Yang, R., Lord, S., Van Den Dool, H., Kumar, A., Wang, W., Long, C., Chelliah, M., Xue, Y., Huang, B., Schemm, J. K., Ebisuzaki, W., Lin, R., Xie, P., Chen, M., Zhou, S., Higgins, W., Zou, C. Z., Liu, Q., Chen, Y., Han, Y., Cucurull, L., Reynolds, R. W., Rutledge, G. and Goldberg, M. (2010), 'The NCEP Climate Forecast System Reanalysis', *Bulletin of the American Meteorological Society* **91**(8), 1015–1058.

Sanderson, B. M., Knutti, R. and Caldwell, P. (2015*a*), 'A Representative Democracy to Reduce Interdependency in a Multimodel Ensemble', *Journal of Climate* **28**, 5171–5194.

Sanderson, B. M., Knutti, R. and Caldwell, P. (2015*b*), 'Addressing Interdependency in a Multimodel Ensemble by Interpolation of Model Properties', *Journal of Climate* **28**(13), 5150–5170.

Sanderson, B. M., Pendergrass, A. G., Koven, C. D., Brient, F., Booth, B. B., Fisher, R. A. and Knutti, R. (2021), 'The potential for structural errors in emergent constraints', *Earth System Dynamics* **12**(3), 899–918.

Sang, Y., Ren, H. L., Shi, X., Xu, X. and Chen, H. (2021), 'Improvement of Soil Moisture Simulation in Eurasia by the Beijing Climate Center Climate System Model from CMIP5 to CMIP6', *Advances in Atmospheric Sciences* **38**(2), 237–252.

Sansom, P. G., Stephenson, D. B. and Bracegirdle, T. J. (2020), 'On Constraining Projections of Future Climate Using Observations and Simulations From Multiple Climate Models', *Journal of the American Statistical Association* **0**(0), 1–12.

Santer, B. D., Taylor, K. E., Wigley, T. M., Johns, T. C., Jones, P. D., Karoly, D. J., Mitchell, J. F., Oort, A. H., Penner, J. E., Ramaswamy, V., Schwarzkopf, M. D., Stouffer, R. J. and Tett, S. (1996), 'A search for human influences on the thermal structure of the atmosphere', *Nature 1996 382:6586* **382**(6586), 39–46.

Santer, B. D., Thorne, P. W., Haimberger, L., Taylor, K. E., Wigley, T. M., Lanzante, J. R., Solomon, S., Free, M., Gleckler, P. J., Jones, P. D., Karl, T. R., Klein, S. A., Mears, C., Nychka, D., Schmidt, G. A., Sherwood, S. C. and Wentz, F. J. (2008), 'Consistency of modelled and observed temperature trends in the tropical troposphere', *International Journal of Climatology* **28**(13), 1703–1722.

Sato, T. and Nakamura, T. (2019), 'Intensification of hot Eurasian summers by climate change and land–atmosphere interactions', *Scientific Reports 2019 9:1* **9**(1), 1–8.

Schaller, N., Sillmann, J., Anstey, J., Fischer, E. M., Grams, C. M. and Russo, S. (2018), 'Influence of blocking on Northern European and Western Russian heatwaves in large climate model ensembles', *Environmental Research Letters* **13**(5), 054015.

Schär, C. and Jendritzky, G. (2004), 'Hot news from summer 2003', *Nature 2004 432:7017* **432**(7017), 559–560.

Schiemann, R., Athanasiadis, P., Barriopedro, D., Doblas-Reyes, F., Lohmann, K., Roberts, M. J., Sein, D. V., Roberts, C. D., Terray, L. and Vidale, P. L. (2020), 'Northern Hemisphere blocking simulation in current climate models: evaluating progress from the Climate Model Intercomparison Project Phase 5 to 6 and sensitivity to resolution', *Weather and Climate Dynamics* **1**(1), 277–292.

Schmittner, A., Latif, M. and Schneider, B. (2005), 'Model projections of the North Atlantic thermohaline circulation for the 21st century assessed by observations', *Geophysical Research Letters* **32**(23), 1–4.

Schneider, T. (2006), 'The General Circulation of the Atmosphere', *Annual Review of Earth and Planetary Sciences* **34**, 655–688.

Schneider, T., Lan, S., Stuart, A. and Teixeira, J. (2017), 'Earth System Modeling 2.0: A Blueprint for Models That Learn From Observations and Targeted High-Resolution Simulations', *Geophysical Research Letters* **44**(24), 12,396–12,417.

Schoetter, R., Cattiaux, J. and Douville, H. (2015), 'Changes of western European heat wave characteristics projected by the CMIP5 ensemble', *Climate Dynamics* **45**(5-6), 1601–1616.

Schubert, S. D., Wang, H., Koster, R. D., Suarez, M. J. and Groisman, P. Y. (2014), 'Northern Eurasian Heat Waves and Droughts', *Journal of Climate* **27**(9), 3169–3207.

Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., Mozaffari, A. and Stadtler, S. (2021), 'Can deep learning beat numerical weather prediction?'.

Schulzweida, U. (2022), 'CDO User Guide'.

Schumacher, D. L., Hauser, M. and Seneviratne, S. I. (2022), 'Drivers and Mechanisms of the 2021 Pacific Northwest Heatwave', *Earth's Future* **10**(12), e2022EF002967.

Scornet, E. (2021), 'Trees, forests, and impurity-based variable importance in regression', *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **59**(1), 21–52.

Screen, J. A., Bracegirdle, T. J. and Simmonds, I. (2018), 'Polar Climate Change as Manifest in Atmospheric Circulation'.

Seneviratne, S. I. and Hauser, M. (2020), 'Regional Climate Sensitivity of Climate Extremes in CMIP6 Versus CMIP5 Multimodel Ensembles', *Earth's Future* **8**(9), e2019EF001474.

Seneviratne, S., Zhang, X., Adnan, M., Badi, W., Dereczynski, C., Di Luca, A., Ghosh, S., Iskander, I., Kossin, J., Lewis, S., Otto, F., Pinto, I., Satoh, M., Vicente-Serrano, S., Wehner, M. and Zhou, B. (2021), Weather and Climate Extreme Events in a Changing Climate., *in* V. Masson-Delmotte, P. Zhai, A. Pirani, S. Connors, C. Pean, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. Matthews, T. Maycock, T. Waterfield, O. Yelekci, R. Yu and B. Zhou, eds, 'Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change', number Spain, Cambridge Uniersity Press, Cambridge, United Kingdom, pp. 1513–1766.

Seo, E., Lee, M. I., Schubert, S. D., Koster, R. D. and Kang, H. S. (2020), 'Investigation of the 2016 Eurasia heat wave as an event of the recent warming', *Environmental Research Letters* **15**(11), 114018.

Serreze, M. C., Barrett, A. P., Stroeve, J. C., Kindig, D. N. and Holland, M. M. (2009), 'The emergence of surface-based Arctic amplification', *The Cryosphere* **3**, 11–19.

Sexton, D. M., Murphy, J. M., Collins, M. and C, M. J. (2012), 'Multivariate probabilistic projections using imperfect climate models part I: Outline of methodology', *Climate Dynamics* **38**(11-12), 2513–2542.

Shapley, L. S. (1953), A Value for n-Person Games, *in* H. Kuhn and A. Tucker, eds, 'Contributions to the Theory of Games (AM-28), Volume II', Princeton University Press, pp. 307–318.

Sherwood, S. C., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., Hegerl, G., Klein, S. A., Marvel, K. D., Rohling, E. J., Watanabe, M., Andrews, T., Braconnot, P., Bretherton, C. S., Foster, G. L., Hausfather, Z., von der Heydt, A. S., Knutti, R., Mauritsen, T., Norris, J. R., Proistosescu, C., Rugenstein, M., Schmidt, G. A., Tokarska, K. B. and Zelinka, M. D. (2020), 'An Assessment of Earth's Climate Sensitivity Using Multiple Lines of Evidence', *Reviews of Geophysics* **58**(4), e2019RG000678.

Shih, Y. S. and Tsai, H. W. (2004), 'Variable selection bias in regression trees with constant fits', *Computational Statistics & Data Analysis* **45**(3), 595–607.

Shikwambana, L. (2022), 'Global Distribution of Clouds over Six Years: A Review Using Multiple Sensors and Reanalysis Data', *Atmosphere* **13**(9), 1514.

Shrestha, M., Acharya, S. C. and Shrestha, P. K. (2017), 'Bias correction of climate models for hydrological modelling – are simple methods still useful?', *Meteorological Applications* **24**(3), 531–539.

Sillmann, J., Kharin, V. V., Zhang, X., Zwiers, F. W. and Bronaugh, D. (2013), 'Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate', *Journal of Geophysical Research Atmospheres* **118**(4), 1716–1733.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*.

Sinclair, V. A., Mikkola, J., Rantanen, M. and Räisänen, J. (2019), 'The summer 2018 heatwave in Finland', *Weather* **74**(11), 403–409.

Sippel, S., Meinshausen, N., Fischer, E. M., Székely, E. and Knutti, R. (2020), 'Climate change now detectable from any single day of weather at global scale'.

Sippel, S., Meinshausen, N., Merrifield, A., Lehner, F., Pendergrass, A. G., Fischer, E. and Knutti, R. (2019), 'Uncovering the forced climate response from a single ensemble member using statistical learning', *Journal of Climate* **32**(17), 5677–5699.

Sousa, P. M., Trigo, R. M., Barriopedro, D., Soares, P. M. and Santos, J. A. (2018), 'European temperature responses to blocking and ridge regional patterns', *Climate Dynamics* **50**(1-2), 457–477.

Stefanon, M., D'Andrea, F. and Drobinski, P. (2012), 'Heatwave classification over Europe and the Mediterranean region', *Environmental Research Letters* **7**(1), 014023.

Stephenson, D. B. (2008), Definition, diagnosis, and origin of extreme weather and climate events, *in* H. F. Diaz and R. J. Murnane, eds, 'Climate Extremes and Society', Vol. 9780521870, Cambridge University Press, chapter Definition, pp. 11–23.

Stone, M. (1974), 'Cross-Validatory Choice and Assessment of Statistical Predictions', *Journal of the Royal Statistical Society: Series B (Methodological)* **36**(2), 111–133.

Stoner, A. M. K., Hayhoe, K. and Wuebbles, D. J. (2009), 'Assessing General Circulation Model Simulations of Atmospheric Teleconnection Patterns', *Journal of Climate* **22**(16), 4348–4372.

Stott, P. A. and Kettleborough, J. A. (2002), 'Origins and estimates of uncertainty in predictions of twenty-first century temperature rise', *Nature 2002 416:6882* **416**(6882), 723–726.

Stott, P. A., Tett, S. F., Jones, G. S., Allen, M. R., Mitchell, J. F. and Jenkins, G. J. (2000), 'External control of 20th century temperature by natural and anthropogenic forcings', *Science* **290**(5499), 2133–2137.

Stuecker, M. F., Bitz, C. M., Armour, K. C., Proistosescu, C., Kang, S. M., Xie, S. P., Kim, D., McGregor, S., Zhang, W., Zhao, S., Cai, W., Dong, Y. and Jin, F. F. (2018), 'Polar amplification dominated by local forcing and feedbacks', *Nature Climate Change 2018 8:12* **8**(12), 1076–1081.

Sutton, R. T., Dong, B. and Gregory, J. M. (2007), 'Land/sea warming ratio in response to climate change: IPCC AR4 model results and comparison with observations', *Geophysical Research Letters* **34**(2), 2701.

Talib, J., Müller, O. V., Barton, E. J., Taylor, C. M. and Vidale, P. L. (2023), 'The Representation of Soil Moisture-Atmosphere Feedbacks across the Tibetan Plateau in CMIP6', *Advances in Atmospheric Sciences* pp. 1–19.

Tan, J., Chen, S., Lee, C. Y., Dong, G., Hu, W. and Wang, J. (2021), 'Projected changes of typhoon intensity in a regional climate model: Development of a machine learning bias correction scheme', *International Journal of Climatology* **41**(4), 2749–2764.

Taylor, K. E. (2001), 'Summarizing multiple aspects of model performance in a single diagram', *Journal of Geophysical Research: Atmospheres* **106**(D7), 7183–7192.

Taylor, P. C., Cai, M., Hu, A., Meehl, J., Washington, W. and Zhang, G. J. (2013), 'A Decomposition of Feedback Contributions to Polar Warming Amplification', *Journal of Climate* **26**(18), 7023–7043.

Tebaldi, C. and Knutti, R. (2007), 'The use of the multi-model ensemble in probabilistic climate projections', *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* **365**(1857), 2053–2075.

Terink, W., Hurkmans, R. T. W. L., Torfs, P. J. J. F. and Uijlenhoet, R. (2010), 'Evaluation of a bias correction method applied to downscaled precipitation and temperature reanalysis data for the Rhine basin', *Hydrol. Earth Syst. Sci* **14**, 687–703.

Teutschbein, C. and Seibert, J. (2012), 'Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods', *Journal of Hydrology* **456-457**, 12–29.

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

Toreti, A., Belward, A., Perez-Dominguez, I., Naumann, G., Luterbacher, J., Cronie, O., Seguini, L., Manfron, G., Lopez-Lozano, R., Baruth, B., van den Berg, M., Dentener, F., Ceglar, A., Chatzopoulos, T. and Zampieri, M. (2019), 'The Exceptional 2018 European Water Seesaw Calls for Action on Adaptation', *Earth's Future* **7**(6), 652–663.

Toreti, A. and Naveau, P. (2015), 'On the evaluation of climate model simulated precipitation extremes', *Environmental Research Letters* **10**(1), 014012.

Toreti, A., Naveau, P., Zampieri, M., Schindler, A., Scoccimarro, E., Xoplaki, E., Dijkstra, H. A., Gualdi, S., Luterbacher, J., Naveau, P., Zampieri, M., Schindler, A., Scoccimarro, E., Xoplaki, E., Dijkstra, H. A., Gualdi, S. and Luterbacher, J. (2013), 'Projections of global changes in precipitation extremes from Coupled Model Intercomparison Project Phase 5 models', *Geophysical Research Letters* **40**(18), 4887–4892.

Träger-Chatterjee, C., Müller, R. W. and Bendix, J. (2013), 'Analysis of extreme summers and prior late winter/spring conditions in central Europe', *Natural Hazards and Earth System Sciences* **13**(5), 1243–1257.

Trenberth, K. E., Branstator, G. W., Karoly, D., Kumar, A., Lau, N. C. and Ropelewski, C. (1998), 'Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures', *Journal of Geophysical Research: Oceans* **103**(C7), 14291–14324.

Trenberth, K. E. and Shea, D. J. (2005), 'Relationships between precipitation and surface temperature', *Geophysical Research Letters* **32**(14), 1–4.

Ullah, S., You, Q., Chen, D., Sachindra, D., AghaKouchak, A., Kang, S., Li, M., Zhai, P. and Ullah, W. (2022), 'Future population exposure to daytime and nighttime heat waves in South Asia', *Earth's Future* .

Van De Velde, J., Demuzere, M., De Baets, B. and Verhoest, N. E. C. (2022), 'Impact of bias nonstationarity on the performance of uni-and multivariate bias-adjusting methods: a case study on data from Uccle, Belgium', *Hydrol. Earth Syst. Sci* **26**, 2319–2344.

Vellinga, M. and Wu, P. (2008), 'Relations between Northward Ocean and Atmosphere Energy Transports in a Coupled Climate Model', *Journal of Climate* **21**(3), 561–575.

Vidale, P. L., Lüthi, D., Frei, C., Seneviratne, S. I. and Schär, C. (2003), 'Predictability and uncertainty in a regional climate model', *Journal of Geophysical Research: Atmospheres* **108**(D18), 4586.

Wang, A., Kong, X., Chen, Y. and Ma, X. (2022), 'Evaluation of Soil Moisture in CMIP6 Multimodel Simulations Over Conterminous China', *Journal of Geophysical Research: Atmospheres* **127**(19), e2022JD037072.

Watson, P. A. (2019), 'Applying Machine Learning to Improve Simulations of a Chaotic Dynamical System Using Empirical Error Correction', *Journal of Advances in Modeling Earth Systems* **11**(5), 1402–1417.

Waugh, D. W. and Eyring, V. (2008), 'Quantitative performance metrics for stratospheric-resolving chemistry-climate models', *Atmos. Chem. Phys* **8**, 5699–5713.

Węglarczyk, S. (2018), 'Kernel density estimation and its application', *XLVIII Seminar of Applied Mathematics* **23**.

Wehner, M., Gleckler, P. and Lee, J. (2020), 'Characterization of long period return values of extreme daily temperature and precipitation in the CMIP6 models: Part 1, model evaluation', *Weather and Climate Extremes* **30**, 100283.

Wehrli, K., Guillod, B. P., Hauser, M., Leclair, M. and Seneviratne, S. I. (2019), 'Identifying Key Driving Processes of Major Recent Heat Waves', *Journal of Geophysical Research: Atmospheres* **124**(22), 11746–11765.

Weigel, A. P., Knutti, R., Liniger, M. A. and Appenzeller, C. (2010), 'Risks of Model Weighting in Multimodel Climate Projections', *Journal of Climate* **23**(15), 4175–4191.

Wenzel, S., Cox, P. M., Eyring, V. and Friedlingstein, P. (2014), 'Emergent constraints on climate-carbon cycle feedbacks in the CMIP5 Earth system models', *Journal of Geophysical Research: Biogeosciences* **119**(5), 794–807.

White, R. H., Anderson, S., Booth, J. F., Braich, G., Draeger, C., Fei, C., Harley, C. D., Henderson, S. B., Jakob, M., Lau, C. A., Mareshet Admasu, L., Narinesingh, V., Rodell, C., Roocroft, E., Weinberger, K. R. and West, G. (2023), 'The unprecedented Pacific Northwest heatwave of June 2021', *Nature Communications 2023 14:1* **14**(1), 1–20.

Wilcox, L. J., Allen, R. J., Samset, B. H., Bollasina, M. A., Griffiths, P. T., Keeble, J., Lund, M. T., Makkonen, R., Merikanto, J., O'Donnell, D., Paynter, D. J., Persad, G. G., Rumbold, S. T., Takemura, T., Tsigaridis, K., Undorf, S. and Westervelt, D. M. (2023),

'The Regional Aerosol Model Intercomparison Project (RAMIP)', *Geoscientific Model Development* **16**(15), 4451–4479.

Wilks, D. S. (2006), *Statistical Methods in the Atmospheric Sciences* , Elsevier Academic Press.

Willett, K. M., Gillett, N. P., Jones, P. D. and Thorne, P. W. (2007), 'Attribution of observed surface humidity changes to human influence', *Nature 2007 449:7163* **449**(7163), 710–712.

Williams, C. K. I. and Rasmussen, C. E. (1995), 'Gaussian Processes for Regression', *Advances in Neural Information Processing Systems 8 (NIPS 1995)* pp. 514–520.

Williamson, D. B. and Sansom, P. P. (2019), 'How Are Emergent Constraints Quantifying Uncertainty and What Do They Leave Behind?', *Bulletin of the American Meteorological Society* **100**(12), 2571–2588.

Woollings, T., Barriopedro, D., Methven, J., Son, S. W., Martius, O., Harvey, B., Sillmann, J., Lupo, A. R. and Seneviratne, S. (2018), 'Blocking and its Response to Climate Change', *Current Climate Change Reports 2018 4:3* **4**(3), 287–300.

Xie, S. P., Deser, C., Vecchi, G. A., Collins, M., Delworth, T. L., Hall, A., Hawkins, E., Johnson, N. C., Cassou, C., Giannini, A. and Watanabe, M. (2015), 'Towards predictive understanding of regional climate change', *Nature Climate Change 2015 5:10* **5**(10), 921–930.

Xu, Z., FitzGerald, G., Guo, Y., Jalaludin, B. and Tong, S. (2016), 'Impact of heatwave on mortality under different heatwave definitions: A systematic review and meta-analysis', *Environment International* **89-90**, 193–203.

Yang, J., Yin, P., Sun, J., Wang, B., Zhou, M., Li, M., Tong, S., Meng, B., Guo, Y. and Liu, Q. (2019), 'Heatwave and mortality in 31 major Chinese cities: Definition, vulnerability and implications', *Science of The Total Environment* **649**, 695–702.

Yiou, P., Cattiaux, J., Faranda, D., Kadygrov, N., Jézéquel, A., Naveau, P., Ribes, A., Robin, Y., Thao, S., van Oldenborgh, G. J. and Vrac, M. (2020), 'Analyses of the Northern European Summer Heatwave of 2018', *Bulletin of the American Meteorological Society* **101**(1), S35–S40.

Yoshikawa, Y. and Iwata, T. (2021), 'Gaussian Process Regression With Interpretable Sample-Wise Feature Weights', *IEEE Transactions on Neural Networks and Learning Systems* .

Yuan, S., Quiring, S. M. and Leasor, Z. T. (2021), 'Historical Changes in Surface Soil Moisture Over the Contiguous United States: An Assessment of CMIP6', *Geophysical Research Letters* **48**(1), e2020GL089991.

Zaitchik, B. F., Macalady, A. K., Bonneau, L. R. and Smith, R. B. (2006), 'Europe's 2003 heat wave: a satellite view of impacts and land–atmosphere feedbacks', *International Journal of Climatology* **26**(6), 743–769.

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., Klein, S. A. and Taylor, K. E. (2020), 'Causes of Higher Climate Sensitivity in CMIP6 Models', *Geophysical Research Letters* **47**(1), e2019GL085782.

Zhang, R., Sun, C., Zhu, J., Zhang, R. and Li, W. (2020), 'Increased European heat waves in recent decades in response to shrinking Arctic sea ice and Eurasian snow cover', *npj Climate and Atmospheric Science 2020 3:1* **3**(1), 1–9.

Zhang, S. and Chen, J. (2021), 'Uncertainty in Projection of Climate Extremes: A Comparison of CMIP5 and CMIP6', *Journal of Meteorological Research* **35**(4), 646–662.

Zhao, J., Meili, N., Zhao, X. and Fatichi, S. (2023), 'Urban vegetation cooling potential during heatwaves depends on background climate', *Environmental Research Letters* **18**(1).

Zhou, Y., Guan, H., Huang, C., Fan, L., Gharib, S., Batelaan, O. and Simmons, C. (2019), 'Sea breeze cooling capacity and its influencing factors in a coastal city', *Building and Environment* **166**, 106408.

Zhu, J., Poulsen, C. J. and Otto-Bliesner, B. L. (2020), 'High climate sensitivity in CMIP6 model not supported by paleoclimate', *Nature Climate Change 2020 10:5* **10**(5), 378–379.